ADVCOMP 2015

The Ninth International Conference on Advanced Engineering Computing and Applications in Sciences

ISBN: 978-1-61208-419-0

July 19 - 24, 2015

Nice, France

ADVCOMP 2015 Editors

Claus-Peter Rückemann, Westfälische Wilhelms-Universität Münster / Leibniz Universität Hannover / North-German Supercomputing Alliance, Germany

ADVCOMP 2015

Forward

The Ninth International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP 2015), held between July 19-24, 2015 in Nice, France, was a multi-track event covering a large spectrum of topics related to advanced engineering computing and applications in sciences.

With the advent of high performance computing environments, virtualization, distributed and parallel computing, as well as the increasing memory, storage and computational power, processing particularly complex scientific applications and voluminous data is more affordable. With the current computing software, hardware and distributed platforms effective use of advanced computing techniques is more achievable.

The goal of ADVCOMP 2015 was a forum to bring together researchers from the academia and practitioners from the industry in order to address fundamentals of advanced scientific computing and specific mechanisms and algorithms for particular sciences. The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The conference sought contributions presenting novel research in all aspects of new scientific methods for computing and hybrid methods for computing optimization, as well as advanced algorithms and computational procedures, software and hardware solutions dealing with specific domains of science.

The conference had the following tracks:

- Advanced methods in fusion physics
- Development of computing support
- Computing techniques
- Complex computing in application domains
- Computing applications in science
- Computing in Virtualization-based environments
- Advances in computation methods

Similar to previous editions, this event attracted excellent contributions and active participation from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the ADVCOMP 2015 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to ADVCOMP 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the ADVCOMP 2015 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that ADVCOMP 2015 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of Advanced Engineering Computing and Applications in Sciences. We also hope that Nice, France, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

ADVCOMP 2015 Chairs

ADVCOMP Advisory Chairs

Chih-Cheng Hung, Southern Polytechnic State University, USA
Juha Röning, Oulu University, Finland
Erich Schweighofer, University of Vienna, Austria
Paul Humphreys, University of Ulster, UK
Danny Krizanc, Wesleyan University, USA
Ivan Rodero, Rutgers University - Piscataway, USA
Ali Shawkat, CQ University of Australia - North Rockhampton, Australia
George Spanoudakis, City University London, UK
Vladimir Vlassov, KTH Royal Institute of Technology, Sweden
Jerry Trahan, Louisiana State University, USA
Dean Vucinic, Vrije Universiteit Brussel (VUB), Belgium
Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany
Wenbing Zhao, Cleveland State University, USA
Camelia Muñoz-Caro, Universidad de Castilla-La Mancha, Spain
Laurent Réveillère, Bordeaux Institute of Technology, France
Ewa Grabska, Jagiellonian University - Krakow, Poland

ADVCOMP Industry/Research Chairs

Jorge Ejarque Artigas, Barcelona Supercomputing Center (BSC-CNS), Spain
Helmut Reiser, Leibniz Supercomputing Centre (LRZ)-Garching, Germany
H. Metin Aktulga, Lawrence Berkeley National Lab, USA
Sameh Elnikety, Microsoft Research, USA
Umar Farooq, Amazon.com - Seattle, USA
Alice Koniges, Lawrence Berkeley Laboratory/NERSC, USA
Markus Kunde, German Aerospace Center & Helmholtz Association - Cologne, Germany
Peter Müller, IBM Zurich Research Laboratory- Rüschlikon, Switzerland
Simon Tsang, Applied Communication Sciences - Piscataway, USA
Anna Schwanengel, Siemens AG, Germany
Christoph Fuenfzig, Fraunhofer ITWM, Germany

ADVCOMP Publicity Chairs

Marie Lluberes, University of Puerto Rico at Mayagüez, USA
Sascha Opletal, University of Stuttgart, Germany
Álvaro Navas, Universidad Politecnica de Madrid, Spain
Iwona Ryszka, Jagiellonian University - Krakow, Poland

ADVCOMP 2015

Committee

ADVCOMP Advisory Chairs

Chih-Cheng Hung, Southern Polytechnic State University, USA
Juha Röning, Oulu University, Finland
Erich Schweighofer, University of Vienna, Austria
Paul Humphreys, University of Ulster, UK
Danny Krizanc, Wesleyan University, USA
Ivan Rodero, Rutgers University - Piscataway, USA
Ali Shawkat, CQ University of Australia - North Rockhampton, Australia
George Spanoudakis, City University London, UK
Vladimir Vlassov, KTH Royal Institute of Technology, Sweden
Jerry Trahan, Louisiana State University, USA
Dean Vucinic, Vrije Universiteit Brussel (VUB), Belgium
Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany
Wenbing Zhao, Cleveland State University, USA
Camelia Muñoz-Caro, Universidad de Castilla-La Mancha, Spain
Laurent Réveillère, Bordeaux Institute of Technology, France
Ewa Grabska, Jagiellonian University - Krakow, Poland

ADVCOMP Industry/Research Chairs

Jorge Ejarque Artigas, Barcelona Supercomputing Center (BSC-CNS), Spain
Helmut Reiser, Leibniz Supercomputing Centre (LRZ)-Garching, Germany
H. Metin Aktulga, Lawrence Berkeley National Lab, USA
Sameh Elnikety, Microsoft Research, USA
Umar Farooq, Amazon.com - Seattle, USA
Alice Koniges, Lawrence Berkeley Laboratory/NERSC, USA
Markus Kunde, German Aerospace Center & Helmholtz Association - Cologne, Germany
Peter Müller, IBM Zurich Research Laboratory- Rüschlikon, Switzerland
Simon Tsang, Applied Communication Sciences - Piscataway, USA
Anna Schwanengel, Siemens AG, Germany
Christoph Fuenfzig, Fraunhofer ITWM, Germany

ADVCOMP Publicity Chairs

Marie Lluberes, University of Puerto Rico at Mayagüez, USA
Sascha Opletal, University of Stuttgart, Germany

Álvaro Navas, Universidad Politecnica de Madrid, Spain
Iwona Ryszka, Jagiellonian University - Krakow, Poland

ADVCOMP 2015 Technical Program Committee

Witold Abramowicz, University of Economics - Poznań, Poland
Kenneth Adamson, University of Ulster, UK
H. Metin Aktulga, Lawrence Berkeley National Lab, USA
Sónia Maria Almeida da Luz, Polytechnic Institute of Leiria, Portugal / University of Extremadura, Spain
Daniel Andresen, Kansas State University, USA
Alina Andreica, Babes-Bolyai University, Romania
Sulieman Bani-Ahmad, Al-Balqa Applied University, Jordan
Roberto Beraldi, "La Sapienza" University of Rome, Italy
Mario Marcelo Berón, National University of San Luis, Argentina
Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany
Ateet Bhalla, Oriental Institute of Science and Technology, India
Muhammad Naufal bin Mansor, University Malaysia Perlis, Malaysia
Pierre Borne, Ecole Centrale de Lille - Villeneuve d'Ascq, France
Xiao-Chuan Cai, University of Colorado Boulder, USA
Christophe Calvin, CEA/DEN/DANS/DM2S, France
Kenneth P. Camilleri, University of Malta - Msida, Malta
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Mete Celik, Erciyes University, Turkey
Yeh-Ching Chung, National Tsing Hua University, Taiwan
Robert Clay, Sandia National Laboratories, USA
Marisa da Silva Maximiano, Escola Superior de Tecnologia e Gestão - Instituto Politécnico de Leiria, Portugal
Vieri del Bianco, Università dell'Insubria, Italy
Javier Diaz, Rutgers University, USA
Xing Cai, Simula Research Laboratory, Norway
Yves Caniou, Université de Lyon, France / University of Tokyo, Japan
Juan Carlos Dueñas López, Universidad Politécnica de Madrid, Spain
Cy Chan, Lawrence Berkeley National Laboratory, USA
Prabu Do, Wipro Technologies, USA
Jorge Ejarque Artigas, Barcelona Supercomputing Center (BSC-CNS), Spain
Sameh Elnikety, Microsoft Research, USA
Javier Fabra, University of Zaragoza, Spain
Simon G. Fabri, University of Malta - Msida, Malta
Umar Farooq, Amazon.com - Seattle, USA
Mehdi Farshbaf-Sahih-Sorkhabi, Azad University - Tehran / Fanavaran co., Tehran, Iran
Mohammad-Reza Feizi-Derakhshi, University of Tabriz, Iran
Dan Feldman, MIT, USA
Mikael Fridenfalk, Uppsala University, Sweden

Radu-Emil Precup, Politehnica University of Timisoara, Romania
Luciana Rech, Universidade Federal de Santa Catarina, Brazil
Helmut Reiser, LRZ, Germany
Michael Resch, HLRS - University of Stuttgart, Germany
Laurent Réveillère, Bordeaux Institute of Technology, France
Dolores Rexachs, Universidad Autónoma de Barcelona (UAB), Spain
Ivan Rodero, Rutgers University - Piscataway, USA
Alexey S. Rodionov, Institute of Computational Mathematics and Mathematical Geophysics - Siberian Division of the Russian Academy of Science, Novosibirsk, Russia
Juha Röning, Oulu University, Finland
Tomasz Rymarczyk, Net-art (Netrix Group), Poland
Iwona Ryszka, Jagiellonian University - Krakow, Poland
Maytham Safar, Focus Consultancy, Kuwait
Julio Sahuquillo, Universitat Politècnica de València, Spain
Francoise Sailhan, Cedric laboratory - Conservatoire National des Arts et Metiers (CNAM), France
Subhash Saini, NASA, USA
Jose Francisco Salt Cairols, Universitat de Valencia-CSIC, Spain
Rainer Schmidt, Austrian Institute of Technology, Austria
Bruno Schulze, National Laboratory for Scientific Computing - LNCC -Petropolis - RJ, Brasil
Erich Schweighofer, Vienna University, Austria
Kewei Sha, Oklahoma City University, USA
Ali Shahrabi, Glasgow Caledonian University, Scotland, UK
Ali Shawkat, CQ University of Australia - North Rockhampton, Australia
Jie Shen, University of Michigan, USA
George Spanoudakis, City University London, UK
Hari Subramoni, Ohio State University, USA
Saïd Tazi, INSA - Toulouse, France
Andrei Tchernykh, CICESE Research Center, Mexico
Parimala Thulasiraman, University of Manitoba, Canada
Jerry Trahan, Louisiana State University, U.S.A.
Simon Tsang, Applied Communications Sciences, USA
José Valente de Oliveira, Universidade do Algarve, Portugal
Ruud van der Pas, SPARC Microelectronics - Oracle, USA
Doru Vatau, University Politehnica of Timisoara, Romania
Vladimir Vlassov, KTH Royal Institute of Technology, Sweden
Dean Vučinić, Vrije Universiteit Brussel (VUB), Belgium
Zhonglei Wang, Intel, Germany
Zhi Wang, Florida State University, USA
Ted Willke, Intel Corporation, USA
Mudasser F. Wyne, National University, USA
Yinglong Xia, IBM, USA
Ping Xiong, Zhongnan University of Economics and Law, China
Chao-Tung Yang, Tunghai University, Taiwan

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A High-Order Relaxation Method for Condensed Explosives Detonation

Ming Yu
Key Laboratory for Computational Physics, Institute of Applied Physics and Computational Mathematics
Beijing, China
E-mail: yu_ming@iapcm.ac.cn

Zhibo Ma
Institute of Applied Physics and Computational Mathematics
Beijing, China
E-mail: ma_zhibo@iapcm.ac.cn

Abstract—**The paper gives a high-order precision and high resolution scheme for the governing equations of the detonation in condensed explosives. Based on the relaxation approximation, the nonlinear governing equations of condensed explosives detonation are transformed into linear relaxation systems, in which it can avoid solving Riemann problem and calculating the Jacobian matrix of nonlinear flux, and it is not necessary to split the source term of chemical reaction law. A fifth-order WENOM (Mapped Weighted Essentially Non-Oscillatory) reconstruction in space discretization and a fifth-order IMEX (IMplicit-EXplicit) scheme of linear multistep methods with monotonicity and TVB (Total Variation Boundedness) in time discrtiezation are utilized. The proposed method is applied to numerically simulate the steady structure of a one-dimensional planar detonation wave and the unsteady propagation of a one-dimensional spherically divergent detonation wave in explosives PBX-9502. The test cases demonstrate that the proposed method can obtain very satisfactory numerical results in terms of accuracy and resolution.**

Keywords-relaxation method; detonation wave; condensed explosives; high-order precision scheme; high resolution scheme

I. INTRODUCTION

The design of complex engineering devices that use high explosives to do useful and controlled work requires the capability to numerically simulate detonation with high fidelity [1]. In the past several decades, the Lagrangian method [1] is in the majority because of its nature character to treat with the interface of multimaterial. However, in the recent decade, more attention has been paid to the Eulerian method due to its following advantages: 1) to well preserve conservation of the total energy due to usually using finite volume discretization; 2) to commonly sharpen the discontinuity of detonation wave due to using high resolution scheme; 3) to easily construct high-order precision in temporal and spatial discretization; 4) to conveniently utilize small meshes to improve accuracy due to employing fixed space grids. Representative works show some fruits of Eulerian method [2]-[5]: using second-order Godunov scheme, adopting simple equation of state (usually perfect gas formulation) and chemical reaction model, employing split way to treat with the chemical source term.

The governing equations of the detonation in condensed explosives are nonlinear hyperbolic conservation system with strongly stiff reaction source term of chemical reaction and complex equation of state. It is the strong stiffness of reaction source term and the complexity of equation of state that brings enormous difficulty to numerically compute the detonation by high-order precision and high resolution scheme.

When strongly stiff source term is discretized, insufficient spatial/temporal resolution may cause an incorrect propagation speed of discontinuities. H. C. Yee [6][7] points out that the phenomenon of wrong propagation speed of discontinuities is connected with the smearing of the discontinuities caused by the discretization of the advection term. The smearing introduces a nonequilibrium state into the calculation, thus as soon as a nonequilibrium value is introduced in this manner, the source term turns on and immediately restores equilibrium, while at the same time shifting the discontinuity to a cell boundary. The analysis shows that the degree of wrong propagation speed of discontinuities is highly dependent on the overall amount of numerical dissipation contained by the numerical scheme. So, excellent shock-capturing scheme for detonation wave discontinuity must possess the high resolution, namely low numerical dissipation. At present, most high resolution schemes have utilized Riemann solver [8] based on simple equation of state, such as the perfect gas with gamma law [9]. However, unreacted solid component and gas product component of detonation in condensed explosives usually utilize some complex equation of state [10], such as Jones-Wilkins-Lee (JWL), HOM, BKW, Davis, extremely, SESAME data library, and so on, also, the temperature and pressure of mixing zone in chemical reaction needs to iterative operation when generally considering pressure and temperature as equilibrium state. Apparently, the high resolution scheme based on Riemann solver is difficult to construct numerically flux about the flow equations of detonation in condensed explosives. In order to capture exactly the shock discontinuity, besides the high resolution in spatial discretization, the high resolution in temporal discretization is very necessary. Hundsdorfer et al. [11] show that the unsplitting explicit and implicit scheme is more reliable: the advection term adopts explicit discretization, and the source term adopts implicit discretization.

Recently, developing a relaxation method is an effective strategy to numerically solve hyperbolic conservation system [12]-[15]. The main idea of the relaxation method is to transform the nonlinear hyperbolic conservation system into linear hyperbolic relaxation equations by means of relaxation approximation. When the relaxation rate tends to zero and

 ISBN: 978-1-61208-419-0

the subcharacteristic condition is satisfied, the solution of the relaxation equations converges to the solution of the original hyperbolic conservation system. In comparison with upwind schemes such as the Godunov scheme, relaxation method does not require the Riemann Solver and the computation of its Jacobians. These features make the relaxation method particularly suitable for those systems where the Riemann problem is difficult to solve or when it is not possible to perform analytical expression for Jacobians. The relaxation method is gradually applied to gasdynamics [16], shallow water motion [17], multimatierial and multicomponent flow [18], magnetohydrodynamic [19].

In this paper, the relaxation method is applied to numerically simulate the typical detonation problem about the condensed explosives. After the nonlinear governing equations of the condensed explosives are transformed into linear relaxation equations, an improved fifth-order WENOM [20] is utilized to spatially discretize and a fifth-order IMEX scheme of linear multistep methods with general monotonicity and boundedness properties is utilized to temporally discretize [11]. The numerical examples about one-dimensional detonation wave in explosives PBX-9502 demonstrate that our method has high accuracy and high resolution properties.

The paper is organized as follows. In Section II, we give the governing equations of detonation in condensed explosives. In Section III, we establish the relaxation equations for the governing equations of detonation. In Section IV, the numerical scheme for the relaxation equations is gives. In Section V several numerical tests are shown. Some conclusions are presented in Section VI.

II. GOVERNING EQUATIONS OF DETONATION IN CONDENSED EXPLOSIVES

The one-dimensional flow equations of detonation in condensed explosives under Eulerian frame are the following:

$$\frac{\partial \boldsymbol{u}}{\partial t}+\frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial r}=\boldsymbol{s}(\boldsymbol{u}) \tag{1}$$

where,

$$\boldsymbol{u}=\begin{bmatrix}\rho r^N\\ \rho v r^N\\ \rho E r^N\\ \rho\lambda r^N\end{bmatrix},\ \boldsymbol{f}(\boldsymbol{u})=\begin{bmatrix}\rho v r^N\\ (\rho v^2+p) r^N\\ (\rho E+p) v r^N\\ \rho v\lambda r^N]\end{bmatrix},$$

$$\boldsymbol{s}(\boldsymbol{u})=\begin{bmatrix}0\\ N r^{N-1} p\\ 0\\ \rho r^N R(\rho,p,\lambda)\end{bmatrix},$$

where ρ is density, v is velocity, E is total energy, λ is chemical reaction process, p is pressure, N is geometry factor ($N=0$ for plane, $N=1$ for cylinder, and $N=2$ for sphere), and R is chemical reaction rate where three-term Lee-Tarver reaction law is adopted [10]:

$$R=I(\eta_1-1-a)^n(1-\lambda)^y+G_1(1-\lambda)^{y_1}\lambda^{x_1}p^{z_1}+G_2(1-\lambda)^{y_2}\lambda^{x_2}p^{z_2}\ .$$

The unreacted solid component and gas product component of detonation in condensed explosives utilize JWL equation of state. On assumption that the pressure and temperature in the reaction mixing zone is in equilibrium, the state of mixing zone may be expressed as (subscript s denotes solid component and subscript g denotes gas product component):

$$\begin{cases} V=(1-\lambda)V_s+\lambda V_g\\ e=(1-\lambda)e_s+\lambda e_g-\lambda q\\ p=A_s\exp(-R_{1s}V_s)+B_s\exp(-R_{2s}V_s)+\dfrac{\omega_s C_{Vs}}{V_s}T\\ p=A_g\exp(-R_{1g}V_g)+B_g\exp(-R_{2g}V_g)+\dfrac{\omega_g C_{Vg}}{V_g}T\\ e_s=\dfrac{A_s}{\rho_0 R_{1s}}\exp(-R_{1s}V_s)+\dfrac{B_s}{\rho_0 R_{2s}}\exp(-R_{2s}V_s)+\dfrac{C_{Vs}}{\rho_0}T\\ e_g=\dfrac{A_g}{\rho_0 R_{1g}}\exp(-R_{1g}V_g)+\dfrac{B_g}{\rho_0 R_{2g}}\exp(-R_{2g}V_g)+\dfrac{C_{Vg}}{\rho_0}T \end{cases}$$

where $V=\rho_0/\rho$ is the relative volume, e is the internal energy per mass, T is temperature, and q is the specific heat for chemical reaction.

For condensed explosives, there are several huge numbers in the chemical reaction rate. For example, for high explosives PBX-9502, I=4.0×10^6, G_1=1100.0, G_2=30.0, so the source term for the chemical reaction rate is regarded as strongly stiff.

III. ESTABLISH OF RELAXATION EQUATIONS

By means of relaxation approximation, the governing equations about condensed explosives may be replaced by the following relaxation system:

$$\begin{cases}\dfrac{\partial \boldsymbol{u}}{\partial t}+\dfrac{\partial \boldsymbol{w}}{\partial r}=\boldsymbol{s}(\boldsymbol{u})\\ \dfrac{\partial \boldsymbol{w}}{\partial t}+\boldsymbol{A}^2\dfrac{\partial \boldsymbol{u}}{\partial r}=\dfrac{\boldsymbol{f}(\boldsymbol{u})-\boldsymbol{w}}{\varepsilon}\end{cases} \tag{2}$$

where $\boldsymbol{w}$ is a middle variable, $\boldsymbol{A}=diag[a_1,a_2,a_3,a_4]$ is a positive diagonal matrix, $0<\varepsilon\le 1$ is relaxation rate.

The linear characteristic of relaxation system (2) is utilized to construct simple and effective high resolution scheme. Papers [12]-[13] point out that the solutions of (2) approach the solutions of the original problem (1): $\boldsymbol{w}\to\boldsymbol{f}(\boldsymbol{u})$, when $\varepsilon\to 0$, and provided the following subcharacteristic condition holds:

$$-a_k\le\frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial \boldsymbol{u}}\le a_k\ (k=1,2,3,4)\ \text{ for all }\ \boldsymbol{u}\,.$$

The role of relaxation rate in numerical scheme may be analyzed [14] as follows.

Because $\boldsymbol{w}$ can converge to $\boldsymbol{f}(\boldsymbol{u})$, there is a Chapman-Enskog expansion:

$$\boldsymbol{w}=\boldsymbol{f}(\boldsymbol{u})+\varepsilon\boldsymbol{f}_1(\boldsymbol{u})+\varepsilon^2\boldsymbol{f}_2(\boldsymbol{u})+\cdots \tag{3}$$

 ISBN: 978-1-61208-419-0

Substituting (3) into (2) and collecting terms, a first-order approximation of system (2) can be obtained:

$$\frac{\partial \boldsymbol{u}}{\partial t}+\frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial r}=\boldsymbol{s}(\boldsymbol{u})+\varepsilon\frac{\partial}{\partial r}\left[\left(\frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial \boldsymbol{u}}\right)\boldsymbol{s}(\boldsymbol{u})\right]+\varepsilon\frac{\partial}{\partial r}\left\{\left[\boldsymbol{A}^2-\left(\frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial \boldsymbol{u}}\right)^2\right]\frac{\partial \boldsymbol{u}}{\partial r}\right\} \tag{4}$$

Thus, system (4) is dissipative if the subcharacteristic condition holds. It can be thought that introducing relaxation rate is equivalent to introducing numerical dissipation.

In practice, the elements of diagonal matrix in system (2) may be chosen as: $\boldsymbol{A}=max\left(\left|\partial \boldsymbol{f}(\boldsymbol{u})/\partial \boldsymbol{u}\right|\right)$ in the whole flowfield zone. Thus, $\boldsymbol{A}$ is a constant matrix, and the bigger $\boldsymbol{A}$ implies the bigger numerical dissipation.

IV. Solution of Relaxation Equations

Diagonalize the system (2) and holds:

$$\begin{cases}\dfrac{\partial(\boldsymbol{w}-\boldsymbol{A}\boldsymbol{u})}{\partial t}-\boldsymbol{A}\dfrac{\partial(\boldsymbol{w}-\boldsymbol{A}\boldsymbol{u})}{\partial r}=\dfrac{\boldsymbol{f}(\boldsymbol{u})-\boldsymbol{w}}{\varepsilon}-\boldsymbol{A}\boldsymbol{s}(\boldsymbol{u})\\ \dfrac{\partial(\boldsymbol{w}+\boldsymbol{A}\boldsymbol{u})}{\partial t}+\boldsymbol{A}\dfrac{\partial(\boldsymbol{w}+\boldsymbol{A}\boldsymbol{u})}{\partial r}=\dfrac{\boldsymbol{f}(\boldsymbol{u})-\boldsymbol{w}}{\varepsilon}+\boldsymbol{A}\boldsymbol{s}(\boldsymbol{u})\end{cases} \tag{5}$$

It can be found that the system (5) is constant linear hyperbolic law with characteristic lines $dr/dt=\pm\boldsymbol{A}$ and Riemann invariables $\boldsymbol{w}\pm\boldsymbol{A}\boldsymbol{u}$.

A semi-discrete finite difference scheme with uniform space sizes for the system (5) can be approximated into:

$$\begin{cases}\dfrac{d\tilde{\boldsymbol{u}}_i}{dt}=\boldsymbol{A}\dfrac{\tilde{\boldsymbol{u}}_{i+1/2}-\tilde{\boldsymbol{u}}_{i-1/2}}{\Delta r}+\dfrac{1}{\varepsilon}\left[\boldsymbol{f}(\boldsymbol{A}^{-1}\dfrac{\tilde{\boldsymbol{w}}_i-\tilde{\boldsymbol{u}}_i}{2})-\dfrac{\tilde{\boldsymbol{w}}_i+\tilde{\boldsymbol{u}}_i}{2}\right]-\boldsymbol{A}\boldsymbol{s}(\boldsymbol{A}^{-1}\dfrac{\tilde{\boldsymbol{w}}_i-\tilde{\boldsymbol{u}}_i}{2})\\ \dfrac{d\tilde{\boldsymbol{w}}_i}{dt}=-\boldsymbol{A}\dfrac{\tilde{\boldsymbol{w}}_{i+1/2}-\tilde{\boldsymbol{w}}_{i-1/2}}{\Delta r}+\dfrac{1}{\varepsilon}\left[\boldsymbol{f}(\boldsymbol{A}^{-1}\dfrac{\tilde{\boldsymbol{w}}_i-\tilde{\boldsymbol{u}}_i}{2})-\dfrac{\tilde{\boldsymbol{w}}_i+\tilde{\boldsymbol{u}}_i}{2}\right]+\boldsymbol{A}\boldsymbol{s}(\boldsymbol{A}^{-1}\dfrac{\tilde{\boldsymbol{w}}_i-\tilde{\boldsymbol{u}}_i}{2})\end{cases} \tag{6}$$

where $\tilde{\boldsymbol{u}}=\boldsymbol{w}-\boldsymbol{A}\boldsymbol{u}$, $\tilde{\boldsymbol{w}}=\boldsymbol{w}+\boldsymbol{A}\boldsymbol{u}$.

When the system (6) is spatially discretized, the numerical flux $\tilde{\boldsymbol{u}}_{i\pm1/2}$ and $\tilde{\boldsymbol{w}}_{i\pm1/2}$ may adopt a fifth-order mapped weighted essentially non-oscillatory (WENOM) [20]:

$$\tilde{\boldsymbol{u}}_{i+1/2}=w_1\tilde{\boldsymbol{u}}^{(1)}_{i+1/2}+w_2\tilde{\boldsymbol{u}}^{(2)}_{i+1/2}+w_1\tilde{\boldsymbol{u}}^{(2)}_{i+1/2},$$

$$\tilde{\boldsymbol{u}}^{(1)}_{i+1/2}=\frac{1}{3}\tilde{\boldsymbol{u}}_{i-2}-\frac{7}{6}\tilde{\boldsymbol{u}}_{i-1}+\frac{11}{6}\tilde{\boldsymbol{u}}_i,$$

$$\tilde{\boldsymbol{u}}^{(2)}_{i+1/2}=-\frac{1}{6}\tilde{\boldsymbol{u}}_{i-1}+\frac{5}{6}\tilde{\boldsymbol{u}}_i+\frac{1}{3}\tilde{\boldsymbol{u}}_{i+1},$$

$$\tilde{\boldsymbol{u}}^{(3)}_{i+1/2}=\frac{1}{3}\tilde{\boldsymbol{u}}_i+\frac{5}{6}\tilde{\boldsymbol{u}}_{i+1}-\frac{1}{6}\tilde{\boldsymbol{u}}_{i+2},$$

$$w_k=\alpha^*_k/(\sum_{l=1}^{3}\alpha^*_l),$$

$$\alpha^*_k=g_k(w^{(JS)}_k),$$

$$g_k(w)=\frac{w(\overline{w}_k+\overline{w}_k^2-3\overline{w}_k w+w^2)}{\overline{w}_k^2+w(1-2\overline{w}_k)},$$

$$w^{(JS)}_k=\alpha_k/(\sum_{l=1}^{3}\alpha_l),$$

$$\alpha_k=\frac{\overline{w}_k}{(\delta+\beta_k)^2},$$

$$\overline{w}_1=\frac{1}{10},\ \overline{w}_2=\frac{3}{5},\ \overline{w}_3=\frac{3}{10},$$

$$\beta_1=\frac{13}{12}(\tilde{\boldsymbol{u}}_{i-2}-2\tilde{\boldsymbol{u}}_{i-1}+\tilde{\boldsymbol{u}}_i)^2+\frac{1}{4}(\tilde{\boldsymbol{u}}_{i-2}-4\tilde{\boldsymbol{u}}_{i-1}+3\tilde{\boldsymbol{u}}_i)^2,$$

$$\beta_2=\frac{13}{12}(\tilde{\boldsymbol{u}}_{i-1}-2\tilde{\boldsymbol{u}}_i+\tilde{\boldsymbol{u}}_{i+1})^2+\frac{1}{4}(\tilde{\boldsymbol{u}}_{i-1}-\tilde{\boldsymbol{u}}_{i+1})^2,$$

$$\beta_3=\frac{13}{12}(\tilde{\boldsymbol{u}}_i-2\tilde{\boldsymbol{u}}_{i+1}+\tilde{\boldsymbol{u}}_{i+2})^2+\frac{1}{4}(3\tilde{\boldsymbol{u}}_i-4\tilde{\boldsymbol{u}}_{i+1}+\tilde{\boldsymbol{u}}_{i+2})^2.$$

When the system (6) is temporally discretized, the following ordinary differential equations can be obtained first:

$$\frac{d\boldsymbol{q}}{dt}=\boldsymbol{\mathcal{F}}(\boldsymbol{q})+\boldsymbol{\mathcal{S}}(\boldsymbol{q}) \tag{7}$$

where $\boldsymbol{q}=[\tilde{\boldsymbol{u}},\tilde{\boldsymbol{w}}]^T$, $\boldsymbol{\mathcal{F}}(\boldsymbol{q})$ denotes the discretization of the advection term in system (6), $\boldsymbol{\mathcal{S}}(\boldsymbol{q})$ denotes the discretization of the source term in system (6).

Then, a fifth-order IMEX scheme of linear multistep methods with general monotonicity and TVB [11] is adopted to solve the ordinary differential equation (7):

$$\begin{aligned}\boldsymbol{q}_n=&\frac{13553}{4096}\boldsymbol{q}_{n-1}-\frac{38121}{8192}\boldsymbol{q}_{n-2}+\frac{7315}{2048}\boldsymbol{q}_{n-3}\\&-\frac{6161}{4096}\boldsymbol{q}_{n-4}+\frac{2264}{8192}\boldsymbol{q}_{n-5}\\&+\frac{10306951}{5898240}\Delta t\boldsymbol{\mathcal{F}}_{n-1}-\frac{13656497}{2949120}\Delta t\boldsymbol{\mathcal{F}}_{n-2}\\&+\frac{1249949}{245760}\Delta t\boldsymbol{\mathcal{F}}_{n-3}-\frac{7937687}{2949120}\Delta t\boldsymbol{\mathcal{F}}_{n-4}\\&-\frac{3387361}{5898240}\Delta t\boldsymbol{\mathcal{F}}_{n-5}\\&+\frac{4007}{8192}\Delta t\boldsymbol{\mathcal{S}}_n-\frac{4118249}{5898240}\Delta t\boldsymbol{\mathcal{S}}_{n-1}\\&+\frac{768703}{2949120}\Delta t\boldsymbol{\mathcal{S}}_{n-2}+\frac{47849}{245760}\Delta t\boldsymbol{\mathcal{S}}_{n-3}\\&-\frac{725087}{2949120}\Delta t\boldsymbol{\mathcal{S}}_{n-4}+\frac{502321}{5898240}\Delta t\boldsymbol{\mathcal{S}}_{n-5}\end{aligned} \tag{8}$$

A third-order Runge-Kutta method [21] is used for starting procedure of this IMEX scheme.

Finally, the relaxation scheme with temporal-spatial fifth-order precision about the detonation flows in condensed explosives turns into the expression (8). It is worthy to indicate that the discretization procedure does not solve Riemann problem.

V. Numerical Example

In this section, the steady structure of one-dimensional planar detonation wave and the unsteady propagation of a

one-dimensional spherically divergent detonation wave in condensed explosives PBX-9502 are calculated. The JWL parameters and chemical reaction rate of PBX-9502 can be found in [22]. The values for Von Neumann spike are (cm-g-μs unit): $p_N = 0.375$, $V_N = 0.675$ and $u_N = 0.253$ respectively; and the corresponding values for Chapman-Jouguet state are: $p_{CJ} = 0.285$, $V_{CJ} = 0.753$, $u_{CJ} = 0.192$ and $D = 0.7665$.

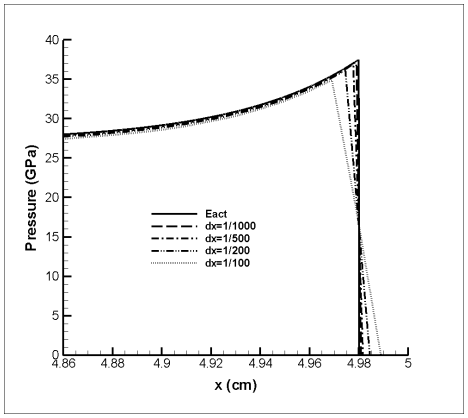
Several values of relaxation rate are tested, and here the results for relaxation rate $\varepsilon = 10^{-7}$ are shown.

A. Steady structure of 1D planar detonation wave

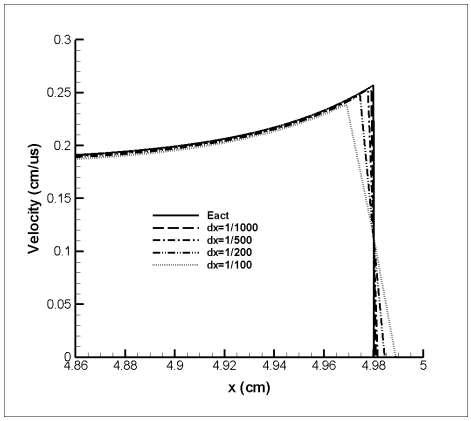
When the detonation arrives at the steady state, the distribution of physical variables in chemical reaction zone can be exactly obtained by means of the Rankine-Hugoniot relations of detonation wave.

The calculating length of explosives takes 5.0cm, and the explosives is initiated by the Chapman-Jouguet condition at its left hand side. The distributions of pressure, relative volume, velocity and mass fraction in chemical reaction zone are obtained, and comparisons are made with the exact solutions. Figure 1(a-d) gives the results where the mesh sizes are $\Delta x = 1/100$, $1/200$, $1/500$, $1/1000$ cm respectively. At the same time, the relation of the Chapman-Jouguet velocity and Von Neumann pressure to the mesh sizes is given in Figure 2(a-b). From Figure 1, the shock front of detonation wave is well resolved, and the spurious oscillation does not appear in the vicinity of the shock discontinuity. From Figures 1 and 2, when the mesh size is less than $1/500$ cm (about 50 meshes in the reaction zone), the calculating solutions agree well with the exact solutions.

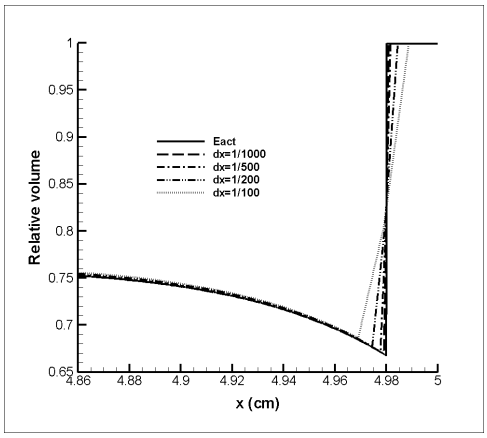
Figure 3(a-b) shows the change of pressure and velocity at several typical times on the course of unsteady propagation of the detonation, in which the discretized mesh is $\Delta x = 1/500$ cm and the corresponding time are: t=0.06, 0.12, 0.24, 0.48, 0.96, 1.44, 1.92, 2.40, 2.88, 3.36, 3.84, 4.32, 4.80, 5.28μs. From the results, the pressure grows much quickly, and reaches the steady state about 3.84μs after initiating CJ conditions. The change agrees well with the experimental results [10].



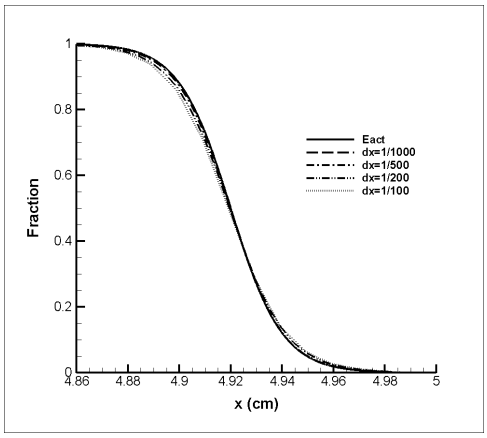
(a) Pressure profile under different mesh sizes



(b) Velocity profile under different mesh sizes

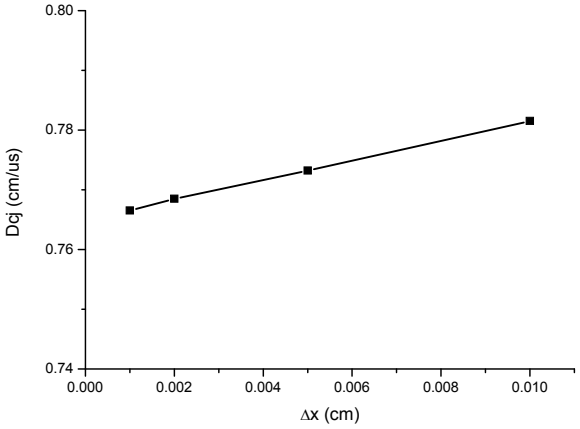


(c) Relative volume profile under different mesh sizes

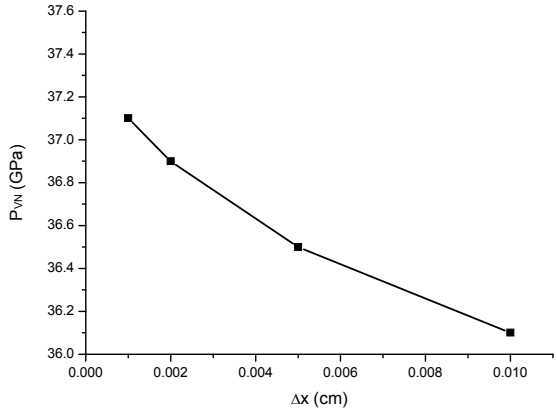


(d) Fraction profile under different mesh sizes

Figure 1. Distributions of physical variables in chemical reaction zone of PBX9502

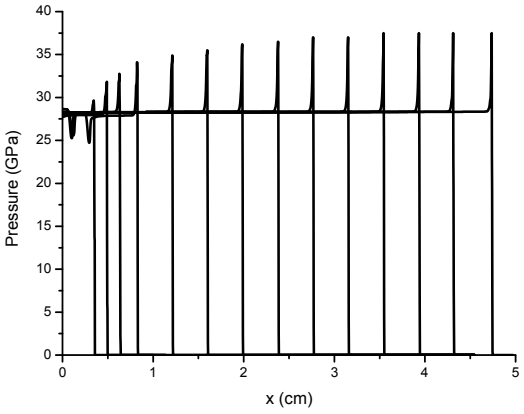


(a) Detonation CJ velocity

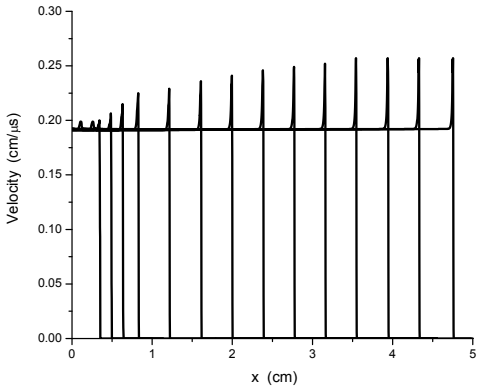


(b) Detonation CJ pressure

Figure 2. Relations of the CJ velocity and Von Neumann pressure to the mesh sizes in PBX9502



(a) Pressure



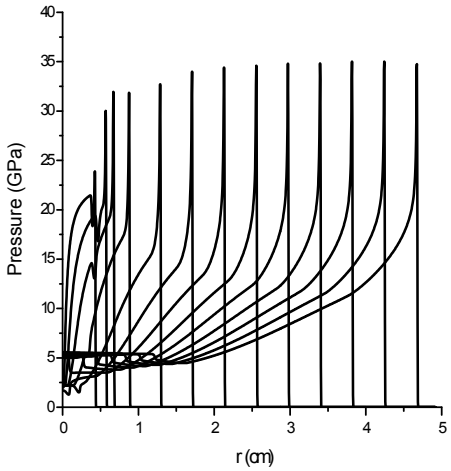
(b) Velocity

Figure 3. Pressure and velocity of planar detonation wave in PBX9502

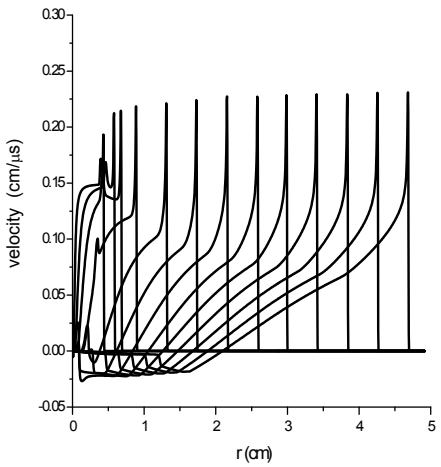
B. Unsteady propagation of 1D spherically divergent detonation wave

When a divergent detonation wave propagates in spherical way, the physical variables behind the shock front will sharply descend. A poor numerical scheme is usually unable to correctly treat with the effect of geometry factor to result in detonation extinguishing [1].

The calculating radius of spherical explosives takes 5.0cm, and the explosives is initiated by the CJ condition at the center. Figure 4(a-b) shows the change of pressure and velocity at several typical times on the course of unsteady propagation of the detonation wave, in which the discretized mesh is $\Delta r = 1/500$ cm and the corresponding time are: t=0.06, 0.12, 0.24, 0.48, 0.96, 1.44, 1.92, 2.40, 2.88, 3.36, 3.84, 4.32, 4.80, 5.28μs. From the results, the pressure and velocity grow along with increasing distance and reach quasi-steady state about 3.84μs after initiation, whose values are lower than the corresponding planar ones.



(a) Pressure



(b) Velocity

Figure 4. Pressure and velocity of spherically divergent detonation wave

VI. CONCLUSION

This paper presented the relaxation method for numerically simulating the detonation in condensed explosives, and a temporal-spatial fifth-order precision scheme is utilized to discretize the relaxation equations, which does not require solving Riemann problem and calculating the Jacobian matrix of nonlinear flux and splitting the source term of chemical reaction law. The calculating results for the steady structure of a one-dimensional planar detonation wave and unsteady propagation of a one-dimensional spherically divergent detonation wave in PBX-9502 demonstrate the high precision and high resolution of the present method. The present method will be generalized to two-dimensional detonation problems in condensed explosives.

ACKNOWLEDGMENT

This work was supported under Grant-11272064 of Natural Science Foundation of China.

REFERENCES

[1] C. L. Mader, Numerical modeling of explosives and propellants, 2nd edition, CRC Press, New York, 1998.

[2] W. D. Henshaw and D. W. Schwedeman, “An adaptive numerical scheme for high-speed reactive flow on overlapping grids”, Journal of Computational Physics, 19(2), 2003, pp. 420-447.

[3] A. K. Kapila, D. W. Schwedeman, and J. B. Bdzil, “A study of detonation diffraction in the Ignition-and-Growth Model”, Combustion Theory and Modeling, 11, 2007, pp. 781-822.

[4] J. W.Banks, D. W. Schwedeman and A. K. Kapila, “A study of detonation propagation and diffraction with compliant confinement”, UCRL-JRNL-233735, 2007.

[5] D. W. Schwedeman, A. K. Kapila, and W. D. Henshaw, “A study of detonation diffraction and failure for a model of compressible reactive flow”, UCRL-JRNL-M43735, 2010.

[6] H. C. Yee, D. V. Kotov, and B. Sjogreen, “Numerical dissipation and wrong propagation speed of discontinuties for stiff source terms”, Proceedings of ICCFD, Hawaii, 2011.

[7] H. C. Yee, D. V. Kotov, and Chi-Wang Shu, “Spurious behavior of shock-capturing methods: problems containing stiff source terms and discontuities”, Proceedings of ICCFD7, 2012.

[8] E. F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics, 2nd edition, Springer, Berlin, 1997.

[9] Shui Hongshou, Difference Methods for One-Dimensional Fluid Dynamics, National Defence Industry Press, 1998(In Chinese).

[10] Sun Chengwei, Applid Detonation Physics, Defense Industry Press, Beijing, 2000.12(In Chinese).

[11] W. Hundsdorfer and S. J. Ruuth, “IMEX extensions of linear multistep methods with general monotonicity and boundedness properties”, Journal of Computational Physics, 225, 2007, pp. 2016–2042.

[12] T. P. Liu, “Hyperbolic conservation laws with relaxation”, Comm. Math. Phys, 108, 1987, pp. 153-175.

[13] Jin S and Xin Z. “The relaxation schemes for systems of conservation laws in arbitrary space dimensions”, Communications of Pure and Applied Mathematics, 48, 1995, pp. 235-276.

[14] A. Chalabi, “Convergence of relaxation schemes for hyperbolic conservation laws with stiff source”, Mathematics of Computation, 68, 1999, pp. 955-970.

[15] T. Tang, “Convergence of MUSCL relaxing scheme to the relaxed scheme for conservation laws with stiff source terms”, Journal of Scientific Computing, 15(2), 2000, pp. 173-195.

[16] M. K. Banda and M. Seaid, “A class of the relaxation schemes for 2D Euler systems of gas dynamics”, ICCS 2002, LNCS 2329, pp. 930-939.

[17] A. I. Delis and T. Katsaounis, “Numerical solution of the two-dimensional shallow water equations by the application of relaxation methods, Applied Mathematical Modelling”, 29, 2005, pp. 754-783.

[18] S. B. Krishnamurthy and M. G. Gerritsen, “A variable relaxation scheme for multiphase, multicomponent flow”, Transaction Porous Medias, 71, 2008, pp. 345-377.

[19] M. K. Banda, “Non-oscillatory relaxation schemes for one-dimensional ideal magnetohydrodynamic equations”, Nonlinear Analysis: Real World Applications, 10, 2009, pp. 3345-3352.

[20] A. K. Henrick, T. D. Aslam, and J. M. Powers, ”Mapped weighted essentially non-oscillatory schemes: Achieving optimal order near critical points”, Journal of Computational Physics, 207, 2005, pp. 542–567.

[21] S. Gottlied, Chi-Wang Shu, and E. Tadmor, “Strong Stability Preserving High-Order Time Discretization Methods”, NASA/CR-2000-210093.

[22] C. M.Tarver and E. M. McGuire, “Reactive Flow Modeling of the Interaction of TATB Detonation Waves with Inert Materials”, The 12th Symposium (International) on Detonation, San Diego, California, 2002, pp. 641-649.

 ISBN: 978-1-61208-419-0

Uncertainty Quantification for Modeling and Simulation with Calibration

Ma Zhibo
Institute of Applied Physics and Computational Mathematics
Beijing, China
E-mail: mazhibo@iapcm.ac.cn

Yu Ming
Key Laboratory for Computational Physics, Institute of Applied Physics and Computational Mathematics
Beijing, China
E-mail: yu_ming@iapcm.ac.cn

***Abstract*—Calibration improves the consistency between simulation results and test data of a system, but it doesn't mean that the epistemic uncertainty of Modeling and Simulation (M&S) for subsystem is reduced, so propagation analysis with many uncertain inputs often leads to an overvaluation of uncertainty. As new system-level test is unavailable, it is unpractical to quantify M&S uncertainty with comparison between simulation results and test data. Taking advantage of the fact that calibration reduces the epistemic uncertainty of system-level simulation, we propose a method for Uncertainty Quantification (UQ), in which the uncertainty from comparison with existing system-level test data and the propagated uncertainty induced by additional cognitive defect for new system are used rationally. An example with virtual tests is displayed in which the method is demonstrated and validated.**

Keywords-uncertainty quantification; modeling & simulation; calibration; verification & validation; reliability certification.

I. INTRODUCTION

M&S needs to experience verification, validation and accreditation (VV&A) procedure to assess its credibility for intended use [1][2]. However, it is still difficult to detect and eliminate all drawbacks even if the verification and validation are adequately implemented. Additionally, owing to inevitable discretization errors, simulation results of complicated physical processes often have systematic errors. Calibration is then used to rectify errors and improve consistency between simulation results and test data. It is a long-standing case to predict the performance of a new engineering system with calibrated codes. When system-level test could not be fulfilled for a new system, it would be a great challenge for engineering design or reliability certification to quantify the uncertainty of prediction offered by M&S [3].

In many cases, a new system is only a modified version of its prototype. Some system-level test data for the prototype generally exist and could be used for calibration and uncertainty quantification [4]. The differences between a new system and its prototype are mainly caused by redesign or by state shift arising from long period stockpile, which may bring on recertification or assessment in engineering. In the case that system-level test is forbidden, numerical simulations for the modified design parameters or the additional engineering factors, and uncertainty quantification of the simulation results are consequently the main approaches to supply information for the recertification and assessment.

According to the concept of Verification and Validation (V&V), the parameter space of an engineering system and its environment may be divided into application domain and validation domain which corresponding to the new system and its prototype respectively. A complex system may be divided into an arbitrary number of progressively simpler hierarchy tiers [1]. Without system-level test of the new system, the uncertainty information of M&S in application domain may have two sources, one is obtained by extrapolation from the uncertainty in validation domain which is quantified by comparison between the simulation results and the existent system-level test data [4], the other is obtained by propagation of the M&S uncertainty from lower level tiers to system level tier [5].

The uncertainty quantification with single information source has been widely studied, such as the UQ method based on comparison and propagation. In a probability frame, Oberkampf and Roy offered a quantification method of M&S errors according to comparison between simulation results and the statistics of test data such as sample average and standard deviation [1]. Helton gave a discussion about sensitivity analysis and Monte Carlo sampling used for uncertainty propagation [6]. Liu et al. used non-intrusive polynomial chaos to quantify the propagation of parameter uncertainties in Jones-Wilkins-Lee equation-of-state (JWL-EOS) for explosive in a detonation system [7]. With the assumption that new system-level test can not be implemented, Ma et al. put forward a method to extrapolate the uncertainties from validation domain to application domain [4][8]. Up to now, it is still a choke point for UQ of M&S that how to fuse two kinds of information that comes from comparison and propagation, respectively.

Techniques of information fusion have become more and more important for reliability analysis as the data lack is just about a ubiquitous problem. Information from comparison and propagation are obtained from different cognitive approaches. It is necessary in engineering and rational in science to fuse them.

The uncertainties of M&S are mainly epistemic and are suitably represented or fused with interval theory. The UQ method should obey two basic principles, the unknown true value should be covered by the estimated uncertainty interval and the estimation of the uncertainty should be minimized

 ISBN: 978-1-61208-419-0

based on the available information [4]. If the estimation is only based on the extrapolated uncertainty originated by comparison, the additional cognitive defect of M&S for a new system may be neglected, and the true-value-covered principle may be violated on account of underestimation of the uncertainty. As a result, the risk to accept an unreliable system may be augmented. If the estimation is based on direct summation of the uncertainties from comparison and propagation, the estimated uncertainty may be irrationally magnified which may lead to violating the uncertainty-minimized principle and consequently increase the risk to reject a reliable system.

The paper is organized as follows. In Section 2, properties of M&S experienced calibration are analyzed. Section 3 offers a quantification method of total uncertainty of M&S based on information fusion, in which the basic component is an extrapolated uncertainty from comparison on system level, and an incremental component is the propagated uncertainty related to new system. Section 4 gives a comparison-based UQ methods needed in Section 3. An example to show these methods is displayed in Section 5 with a shock problem. Finally, we give a conclusion in Section 6.

II. PROPERTIES OF CALIBRATED M&S

There are two approaches to improve consistency between simulation results and test data. One is to enhance the cognitive ability by which the epistemic uncertainties in M&S are reduced. This is also an ideal approach for M&S development. The other is based on existent cognitive ability, to make artificially errors produced in M&S compensated with each other. Calibration of M&S depends mostly on the mechanism of error compensations. However, when M&S is used as prediction, the calibration only works well as the modification of the new system is not very large compared to the systems on which the calibration is made.

In this paper, the mathematical model is divided into an entity model and a physics model. The former represents the specific engineering system depicted by design parameters, such as material type, shape, size, mass, and initial or boundary conditions when the system works. The latter represents the abstract laws of hylic world, such as equations of state and constitution, turbulence model, detonation model, the universal conservation equations of mass, momentum, and energy etc. The uncertainties of physics models are usually greater than that of entity models, as dynamic measuring and converse reckon are generally involved to determine the parameters and forms of physics models.

Calibration is achieved by comparison with test data, in which the forms and parameters of models, methods and parameters of computation, knobs, and computer code are adjusted and then fixed. Knobs here are referred to the ad hoc parameters added to a model to simply obtain agreement with test data but lack definite physics significances or lack actual evaluating information [2]. Via sufficient verification and validation, knobs could be reduced but it is difficult to eliminate absolutely due to the existence of discretization errors and the deficiency in modeling and simulation for complex systems [9].

Generally, calibration is executed based on a range of entity models, to which we call calibration domain in the model space. Validation activities executed after calibration also have their validation domains, in which the M&S uncertainties may be quantified according to test data. After calibration is finished, the computer code and the parameters that need adjustment should be fixed for intended use. The fixation is usually relative and periodical, as the evaluation on parameters in physics model probably depend on methods and parameters of numerical computation under the expectation of good agreement with test data. The version of code and the parameters of physics models may vary with the development of M&S.

Comparison and propagation are two basic approaches to gain information of uncertainties and, from the point of methodology on cognition, they are pertaining to induction and deduction, respectively. As the former is based on practice and observation to apperceive the realities, information obtained from comparison is generally with an inherent credibility than that from propagation and it may dominate in information fusions when conflict occurs between them.

The characters in numerical simulations with calibration are summarized as follows:

- Uncertainties on system level can be effectively reduced by calibration. However, as the entity model departing from calibration domain, the error compensation may be gradually fading away and the simulation results for a new system may have lager deviations from the true values;
- Uncertainties obtained by comparison in validation domain could be extrapolated into application domain, while the extrapolated uncertainties do not include the uncertainties that introduced by additional cognitive defect of the M&S for a new system in application domain;
- Calibration can not reduce certainly the M&S uncertainties under system level, so the traditional propagation gives usually an overestimation of the M&S uncertainties on system level;
- The epistemic M&S uncertainties that come from comparison in validation domain should have a dominate weight than that come from propagation when information fusion is implemented;
- Without system-level test of new system, there are two independent information sources of M&S uncertainties for system level. One is that from the comparison in validation domain and the other is the additional uncertainty propagated from under system levels which induced by extra cognitive defects that M&S encounters. They can be fused based on interval theory and their additive property.

III. UQ OF M&S WITH CALIBRATION

The most important problem is to fuse information from comparison and propagation and to keep the UQ method observe the true-value-covered and uncertainty-minimized principles [4].

 ISBN: 978-1-61208-419-0

It is known that both aleatory and epistemic uncertainties can be quantified by test, but only the epistemic uncertainties can be reduced by test data. For nondeterministic M&S, the aleatory uncertainties that come from comparison and propagation respectively could be depicted by probability and be fused by Bayesian theory with the weight relative to their information quantities [10][11]. As the deterministic M&S only produces epistemic uncertainties, uncertainties from propagation should only be distributed little or zero weight in fusions when uncertainties from comparison exist.

In the case that system-level test of a new system is unavailable and no direct information from comparison could be used, we consider the following schemes for uncertainty quantification:

- Only use the extrapolated uncertainty from comparison in validation domain;
- Only use the uncertainty from propagation;
- Use both of above information.

The first scheme may leave out the additional uncertainties induced by the extra cognitive defects for new systems. The second scheme does not make use of the existing comparison to reduce the epistemic uncertainty and the uncertainty induced by propagation method and numerical computations are difficult to be quantified into a total M&S-uncertainty on system level. The third scheme has the most reasonable idea, but needs a proper design to avoid the disadvantages appearing in the former two schemes.

Based on the third scheme, we disassemble the uncertainty from propagation ($U_{application}^{propagation}$) to two parts. One refers to the propagated uncertainty for the systems in validation domain ($U_{validation}^{propagation}$). The other refers to an additional propagated uncertainty induced by an extra cognitive defects of new systems in application domain ($U_{\Delta}^{propagation}$). They have an approximate relationship

$$U_{application}^{propagation} \approx U_{validation}^{propagation} + U_{\Delta}^{propagation} \quad . \tag{1}$$

According to the three sources of M&S uncertainty, we have

$$U_{application}^{propagation} \approx \sum_{i=1}^{3} {}^{i}U_{application}^{propagation} \approx \sum_{i=1}^{3} \left({}^{i}U_{validation}^{propagation} + {}^{i}U_{\Delta}^{propagation} \right), \tag{2}$$

where ${}^{1}U^{propagation}$, ${}^{2}U^{propagation}$, ${}^{3}U^{propagation}$ represent the propagated uncertainties on system level that born from entity modeling, physics modeling and numerical computation, respectively.

In validation domain, as the information from comparison has an overwhelming weight than that from propagation, we neglect the contribution of $U_{validation}^{propagation}$ for information fusion. In application domain, the additional propagated uncertainty $U_{\Delta}^{propagation}$ has no corresponding uncertainty from comparison, so $U_{\Delta}^{propagation}$ must be counted wholly in the total M&S uncertainty.

Thus, the M&S uncertainty for a new system is composed of two components and they have an additive property, which is represented as

$$U_{application}^{M\&S} \approx U_{application}^{extrapolation} + U_{\Delta}^{propagation} \tag{3}$$

So, the steps of UQ for M&S may be arranged as follows:

- To calibrate the M&S with available test data and finish the fixation of concerned parameters and the computer code;
- To quantify M&S uncertainty based on comparison with test data in validation domain;
- To extrapolate uncertainties in validation domain to obtain M&S uncertainty ($U_{application}^{extrapolation}$) for the new system based on the relationship between uncertainties and parameters of the entity model;
- To quantify the additional propagated M&S uncertainty ($U_{\Delta}^{propagation}$) for the new system;
- To obtain the total M&S uncertainties of the new system by (3).

IV. UQ BASED ON COMPARISON

Test data may have aleatory uncertainties owing to the randomness in manufacture of physical models and in test measure. As these uncertainties are essentially not induced by deterministic M&S, it is necessary for a comparison-based UQ to build properly statistics to get rid of the impacts of the aleatory uncertainties on the quantification of epistemic uncertainties.

There are two cases when comparison is carried out:

- One simulation to one test (one-to-one);
- One simulation to many tests (one-to-many).

In the case of one-to-one, each physical model to undergo test must be measured and the results are used for M&S. The simulation results and the test data have one-to-one relationship. The difference between them after test error is recouped reflect directly the error of M&S for this physical model. If the number of test or simulation is n, we have M&S uncertainty as:

$$U^{m\&s} = MAX\left| y_i^{m\&s} - y_i^{test} \right| + \delta U^{test},\ i = 1,2\cdots n, \tag{4}$$

where $y_i^{m\&s}$ and y_i^{test} are results of M&S and test data of physical model i, respectively. U^{test} is uncertainty of the test data. As n is small δ is advised to be evaluated as $+1$ to evade the risk of underestimation for $U^{m\&s}$. But as n is great enough, δ could be evaluated as 0 or -1 to evade the risk of overestimation.

The case of one-to-many corresponds to the repeated tests, in which the mathematical modeling is based just on one set of parameters that evaluated generally from the average values of design for the physical models, and only one set of M&S result is output. Although all physical models are manufactured according to a same design, their test results may be stochastic owing to the random of manufacture and test measurement.

In order to screen the interference of aleatory uncertainty induced by these random factors, we suggest to dig out the M&S uncertainty from the difference between the M&S result and the average of the test data,

$$\bar{y}^{test} = \frac{1}{n}\sum_{i=1}^{n} y_i^{test} , \tag{5}$$

 ISBN: 978-1-61208-419-0

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(y_i^{test} - \bar{y}^{test} \right)^2 \right)^{1/2}, \quad (6)$$

$$U^{m\&s} = \left| y^{m\&s} - \bar{y}^{test} \right| + t_{(1-\beta)/2,\nu} \cdot \frac{s}{\sqrt{n}}. \quad (7)$$

In the following example, we assign the confidence β=0.95, and $t_{(1-\beta)/2,\nu}$ is the quantile of $(1-\beta)/2$ for t-distribution with the freedom of $\nu = n-1$.

When n is great enough and U^{test} is exactly estimated, the result from (4) and that from (5)-(7) should be accordant with each other.

V. AN EXAMPLE

A. Description of System

The system is simply composed of a 1-dimensional shock problem as shown in Figure 1, in which the left part is a high-pressure gas and the right part is a high-density metal with initial parameters of pressure, density, inner energy and velocity p_1, ρ_1, e_1, v_1 and p_2, ρ_2, e_2, v_2, respectively, where v_1, v_2, e_2, p_2 are zero. A left-marching expanding wave and a right-marching shock wave occur from the material interface. The metal is pushed by the high pressure of gas to move rightward. The response of interest for this system is the interface displacement D toward right when time is at 10 μs.

B. Design of Tests

Tests are needed in the procedures of calibration and validation to prove the UQ method is valid or not. As the exact solution of system response D^* exists, we use virtual tests to replace the real tests. It is needed to predefine a set of entity and physics models that express the real aging of the products with different stockpile time. In this paper, we call them as true aging models, which are only used to give virtual test data and are not necessary to be displayed in the context, as if in real tests the true parameters of the aging models are unknown.

The way to do the virtual test (called briefly test in the following text) is as follows. At first, we evaluate design parameters of the physical models according to the true aging models. Then change the design parameters to be random by adding virtual random variables (to simulate the random process of manufacture). The exact solution of each physical model is obtained based on the sampling value of the randomized model parameters. Finally, we get results that regarded as real test data from the exact solution plus a sampling value of another random variable that added to the exact solution (to simulate the random process of real test).

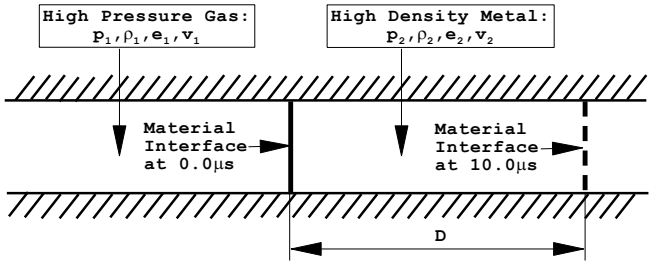


Figure 1. Entity model of the system

The parameters that need to add a virtual random variable are the design values ρ_1^*, ρ_2^*, e_1^*, γ_1^*, γ_2^* and the exact solution of system response D^*. Their virtual random variables $\tilde{\rho}_1$, $\tilde{\rho}_2$, $\tilde{e}_1$, $\tilde{\gamma}_1$, $\tilde{\gamma}_2$ and $\tilde{D}$ are supposed to follow normal distributions with zero-means and deviations of 4.5kg/m^3, 36.0kg/m^3, 0.2MJ/kg, 0.01, 0.2 and 0.1mm, respectively. The true parameters (unknowns in real tests) of physical models are formed as $\rho_1^{test} = \rho_1^* + \tilde{\rho}_1$, $\rho_2^{test} = \rho_2^* + \tilde{\rho}_2$, $e_1^{test} = e_1^* + \tilde{e}_1$, $\gamma_1^{test} = \gamma_1^* + \tilde{\gamma}_1$, $\gamma_2^{test} = \gamma_2^* + \tilde{\gamma}_2$. And the true test data are formed as $D^{test} = D^* + \tilde{D}$.

C. Calibration

In this system, two types of parameters are calibrated, namely numerical parameters and physics parameters. The numerical methods do not need to be calibrated as the physical process is not complicated. Like the sequence of V&V, numerical parameters should be calibrated before physics parameters. Calibration on numerical parameters is just based on the test data of fresh products considering the adequacy of test data and the least disturbance of aging models. And calibration on physics parameters is based on the test data of aged products.

The numerical parameters to be calibrated are artificial viscosity coefficients corresponding to the selected steps of space and time. The physics parameters to be calibrated are from aging models for e_1 and γ_2 of aged materials. All the calibrated numerical and physics parameters form a fixed association in M&S for intended use.

Calibration could be executed through following steps:

- Based on the demand analysis of M&S, determine numerical methods, numerical and physics parameters or knobs need to be calibrated, and the approach to get the reference solution for M&S (Here the reference solutions are test data);
- Choose fresh products to be tested and give their design values of physics parameters as $\rho_1^* = 2500.0 kg/m^3$, $e_1^* = 6.0 MJ/kg$, $\gamma_1^* = 3.0$, $\rho_2^* = 20000.0 kg/m^3$, $\gamma_2^* = 5.0$. Obtain ρ_1^{test}, e_1^{test}, γ_1^{test}, ρ_2^{test}, γ_2^{test} and the corresponding test data D^{test} for five physical models by plus the sampling values of their virtual random variables and the design values or exact solutions;
- Obtain the optimally calibrated numerical parameters through comparison between one numerical result $D^{m\&s}$ and five test data D^{test}=3.856, 3.852, 3.841, 4.010, 3.757 mm, such as

the artificial viscosity coefficients $\alpha = 1.5$ and $\beta = 0.06$ corresponding to the initial grid width $\Delta x = 0.1\,\text{mm}$ and time step $\Delta t = 0.0016\mu s$. The artificial viscosity model used is

$$q = \begin{cases} 0, & \dot{\varepsilon}_{kk} \geq 0 \\ \rho l \left|\nabla \cdot \vec{v}\right| \left(\alpha l \left|\nabla \cdot \vec{v}\right| + \beta c\right), & \dot{\varepsilon}_{kk} < 0 \end{cases},$$

where q is the viscous pressure, l is the grid size, c is the sound speed and $\nabla \cdot \vec{v}$ is the divergence of velocity. The numerical result corresponding to these optimal values is $D^{m\&s}$ =3.891mm;

- Obtain the optimally calibrated physics parameters based on the aging model and the comparison between numerical results and test data about stockpile time of 10 years, 30 years and 50 years. The aging models describe the changing of physics parameters are

$$e_1(t) = e_1(0)\left(1.0 + a_1 t + b_1 t^2\right), \quad (8)$$

$$\gamma_2(t) = \gamma_2(0)\left(1.0 + a_2 t + b_2 t^2\right), \quad (9)$$

where the time t is in "year" and the calibrated parameters are

$$\begin{cases} a_1 = -0.5\times10^{-4}, \\ b_1 = -1.5\times10^{-5}, \\ a_2 = \ 1.5\times10^{-4}, \\ b_2 = \ 3.0\times10^{-5}. \end{cases} \quad (10)$$

- Finish calibration by fixing the calibrated numerical methods and parameters.

D. Validation

In model space, the validation domain is defined as the stockpile time from 0 years to 50 years, in which the validation tests are for the stockpile time of 0 year, 10 years, 20 years, 30 years, 40 years, 50 years. For each stockpile time there are five repeated tests but only one numerical result.

By (7), in which β=0.95, and $t_{(1-\beta)/2,\nu} = t_{0.025,4}$=2.7764, uncertainty in validation domain are quantified as $U(t)$ =0.142, 0.128, 0.283, 0.152, 0.202, 0.257 mm for stockpile time t =0, 10, 20, 30, 40, 50 years.

E. Uncertainty Quantification in Application Domain

In model space, the application domain is defined as the stockpile time great than 50 years. There is no system-level test in this domain.

In order to quantify the first item in the right hand of (3), we have to determine the function showing uncertainty varies with the stockpile time. Here, a second order polynomial is used

$$U(t) = a_0 + a_1 t + a_2 t^2. \quad (11)$$

Based on uncertainties in validation domain, we have

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 10 & 100 \\ 1 & 20 & 400 \\ 1 & 30 & 900 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.142 \\ 0.128 \\ 0.283 \\ 0.152 \\ 0.202 \\ 0.257 \end{pmatrix}. \quad (12)$$

The minimum-norm solution of this over-determined equation is

$$(a_0, \ a_1, \ a_2) = (0.14, \ 0.0026, \ -1.38\times10^{-5})$$

With this solution, we get the extrapolated M&S uncertainty $U_{80\,year}^{extrapolation} = 0.261$mm for 80 years stockpile by (11). It is just the value of $U_{application}^{extrapolation}$ in (3).

Uncertainties of physics parameters for different stockpile time are listed in Table 1, in which the aleatory uncertainties are induced from manufacture and the epistemic uncertainties are induced from the cognitive defect of the statistical population average of physics parameters. The additional epistemic uncertainty of 80 years stockpile comparing to 0~50 years stockpile is 0.2 MJ/kg for ${}^{e_1}U$ and 0.2 for ${}^{\gamma_2}U$.

The sensitivities of system response D to each physics parameter are in Table 2.

For system of 80 years stockpile, the uncertainty propagated from the additional uncertainties of physics parameters is

$$\begin{aligned} U_{\Delta}^{propagation} &\approx \left|\partial D/\partial e_1\right| \times \Delta\left({}^{e_1}U\right) + \left|\partial D/\partial \gamma_2\right| \times \Delta\left({}^{\gamma_2}U\right) \\ &= 0.499\times(0.8-0.6) + 0.103\times(0.7-0.5) \\ &= 0.120(mm). \end{aligned}$$

The uncertainty of 80 years stockpile quantified by (3) is:

$$U_{80\,year}^{M\&S} \approx U_{80\,year}^{extrapolation} + U_{\Delta}^{propagation} = 0.381(mm). \quad (13)$$

If all the epistemic uncertainties are propagated, we get the uncertainty that comes from propagation as

$$\sum_{i=1}^{5} \left|\partial D/\partial x_i\right| \cdot \left({}^{x_i}U\right) \approx 0.496(mm).$$

Moreover, if the aleatory uncertainties are also propagated, the uncertainty from propagation will reach 0.928mm. However, it couldn't be regarded as total M&S uncertainty. From here, we see that the method depicted by (3)-(7) can reduce epistemic uncertainties through calibration and filter the aleatory uncertainties by properly defined statistics.

TABLE I. UNCERTAINTIES OF PHYSICS PARAMETERS

Uncertainty type	Aleatory		Epistemic	
Stockpile (Year)	0~50	80	0~50	80
${}^{\rho_1}U$ (kg/m^3)	13.50		5.00	
${}^{\rho_2}U$ (kg/m^3)	108.00		40.00	
${}^{e_1}U$ (MJ/kg)	0.60		0.60	0.80
${}^{\gamma_1}U$ (1)	0.03		0.01	
${}^{\gamma_2}U$ (1)	0.60		0.50	0.70

TABLE II. SENSITIVITY OF D TO PHYSICS PARAMETERS

$\partial D/\partial\rho_1$ $\left(\frac{mm}{kg/m^3}\right)$	$\partial D/\partial\rho_2$ $\left(\frac{mm}{kg/m^3}\right)$	$\partial D/\partial e_1$ $\left(\frac{mm}{MJ/kg}\right)$	$\partial D/\partial\gamma_1$ （mm）	$\partial D/\partial\gamma_2$ （mm）
1.085×10^{-3}	-1.356×10^{-4}	0.499	1.400	-0.103

F. Validity of the UQ Method

In practice, the system-level test is not available in application domain. We have specially arranged five new tests of 80 years stockpile here aims to prove the UQ method is valid. This procedure starts with uncertainty assessment by comparison the numerical results and test data as specified in (7). Then compare the assessment result with the prediction result in (13). If the latter is greater than the former just in a little amount, it may offer a positive evidence of validity for the UQ method in the sight of obeying the true-value-covered and uncertainty-minimized principles.

Implementation of the tests for 80 years stockpile is similar to the tests as narrated above, i.e., the values of physics parameters in physical models are obtained by the true aging models and sampling, the test data are obtained by exact solutions and sampling, the physics parameters for M&S are from (10) and their deviations from the true aging models agree with the uncertainty in Table 2.

The five test data D^{test} =3.455, 3.338, 3.216, 3.395, 3.174 mm and the numerical result is $D^{m\&s}$ =3.499mm, in which the parameters for M&S are ρ_1 =2500kg/m^3, e_1 =5.4MJ/kg, γ_1 =3.0, ρ_2 =20000kg/m^3, γ_2 =6.02. The assessed M&S uncertainty $U_{80\,year}^{M\&S}$ =0.331mm.

The result in (13) shows that the uncertainty (0.381mm) obtained by prediction with (3) is slightly greater than the uncertainty (0.331mm) obtained by assessment, from which the success of the UQ methods is exhibited.

VI. CONCLUSION

When system-level test is unavailable, the prediction by M&S and its uncertainty are the most important information for reliability certification or assessment, and propagation is the most imaginable UQ method. As the system becoming complicated and its hierarchy having more multiple tiers, the numerical errors and the great number of uncertain input factors will make it impractical to quantify the M&S uncertainty just by propagation. Based on the reality that the prototype of a new system generally has some test data and the awareness that the epistemic uncertainty of M&S could be reduced by calibration with existed test data, an UQ method is put forward to synthesize the uncertainties in validation domain and the propagated additional uncertainties. With an example the method is shown to observe the true-value-covered and uncertainty-minimized principles of UQ for M&S that is used as prediction.

ACKNOWLEDGMENT

This research is supported by the National Nature Science Foundation (Grant No. 11371066, 11272064) and CAEP (Grant No. 2012B0102010, 2013A0101004) of China.

REFERENCES

[1] W. L. Oberkampf and C. J. Roy, "Verification and validation in science computing," Cambridge University Press, 2010.

[2] C. J. Roy and W. L. Oberkampf, "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing," Computer Methods in Applied Mechanics and Engineering, vol. 200, no. 25-28, 2011, pp. 2131-2144.

[3] Z. B. Ma, Y. J. Ying and J. S. Zhu, "QMU certifying method and its implementation," Chinese Journal of Nuclear Science and Engineering, vol. 29, no. 1, 2009, pp. 1-9.

[4] Z. B. Ma, H. J. Li, J. W. Yin and W. B. Huang, "Uncertainty Quantification of Numerical Simulation for Reliability Analysis," Chinese Journal of Computational Physics, 2014, 31(4): pp. 424-430.

[5] W. Chen, L. Baghdasaryan, T. Buranathiti and J. Cao, "Model validation via uncertainty propagation," AIAA Journal, vol. 42, no. 7, 2004, pp. 1406-1415.

[6] J. C. Helton, "Conceptual and Computational Basis for the Quantification of Margins and Uncertaity," Sandia Report, SAND2009-3055.

[7] Q. Liu, R. L. Wang, Z. Lin and W. Z. Wen, "Uncertainty quantification for JWL EOS parameters in explosive numerical simulation," Chinese Journal of Explosion and Shock Waves, vol. 33, no. 6, 2013, pp. 647-654.

[8] Z. B. Ma, M. Zheng and J. W. Yin, "Quantification of uncertainties in detonation simulations," Chinese Journal of Computational Physics, vol. 28, no. 1, 2011, pp. 66-74.

[9] D. Higdon and M. kennedy, "Combining field observations and simulations for calibration and prediction," SIAM Journal of Scientific Computing, vol. 26, 2004, pp. 448-466.

[10] J. C. Helton, "Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty," Journal of Statistical Computation and Simulation, vol. 57, 1997, pp. 3-76.

[11] I. Babuska, F. Nobile and R. Tempone, "A systematic approach to model validation based on bayesian updates and prediction related rejection criteria," Computer Methods in Applied Mechanics and Engineering, vol. 197, no. 29-32, 2008, pp. 2517-2539.

 ISBN: 978-1-61208-419-0

A Cell-Centered Lagrangian Method Based on Local Evolution Galerkin Scheme for Two-Dimensional Compressible Flows

Ming Yu
Key Laboratory for Computational Physics, Institute of Applied Physics and Computational Mathematics
Beijing, China
E-mail: yu_ming@iapcm.ac.cn

Haibo Xu, Yutao Sun
Institute of Applied Physics and Computational Mathematics
Beijing, China
E-mail: xu_haibo@iapcm.ac.cn, syt@iapcm.ac.cn

***Abstract*—This paper presents a new cell-centered Lagrangian method for two-dimensional compressible flows. The main feature of the method is that the velocity and pressure at the cell vertex are computed using the local Galerkin evolution scheme for solving the linearized flow equations in terms of the bicharacteristic theory, and then the velocity and pressure are used to update the grid coordinates and evaluate the numerical flux across the cell interface. The local Galerkin evolution operator gives the solutions evolving for an infinite small time interval from the initial conditions and still can maintain the genuinely multidimensional nature of a hyperbolic system.**

Keywords-Lagrangian method; cell-centered scheme; local evolution Galerkin scheme; finite volume scheme

I. INTRODUCTION

In multimaterial flow simulation, a grid-staggered Lagrangian method is extensively adopted [1]. Recently, increased attention has been paid to the cell-centered Lagrangian method, in which the primary variables including the density, momentum (velocity) and total energy, are defined at the center of a cell. The cell-centered scheme is constructed by integrating directly the system of conservation laws on each moving cell with finite volume discretization, so it can well preserve the conservation of the momentum and total energy, and may not require the artificial viscosity and hourglass viscosity. In addition, it has synchronous time advancement among the flow governing equations. The idea of cell-centered scheme was firstly introduced by Godunov [2] in one-dimensional gas dynamics and then extended to multidimensional flows. The key point of multidimensional cases lies in the determination of the velocity at the cell vertex. There are several typical approaches to determine the vertex velocity of a cell [3]-[6].

Apparently, it is a good idea to construct the Riemann solver of the cell vertex directly from the characteristic property about multidimensional compressible fluid equations. To design this "genuinely multidimensional" numerical solver, the evolution Galerkin scheme [7]-[9] may be adopted, in which the exact integral equations from a general theory of bicharacteristics for the linear or linearized hyperbolic system were derived in terms of the primitive physical variables, and then the vertex solutions were obtained to determine the vertex velocity and evaluate the numerical fluxes across the cell interface. Usually, these integral solutions could be further approximated by approximate evolution operator in such a way that all of the infinitely many directions of propagation of bicharacteristics were explicitly taken into account. This vertex solver from the bicharacteristic theory essentially is a multidimensional Riemann solver or a generalization of the original idea of Godunov to multidimensional hyperbolic conservation laws. The idea has been studied extensively from theoretical as well as numerical point of view and applied to various science and engineering for the compressible fluid equations in the Eulerian formalism [7]-[9]. Traditionally, the evolution Galerkin operator gives the evolutive course within a certain time interval. In order to simplify the computations of the integral solutions and facilitate the semi-discrete finite volume scheme, the local evolution Galerkin operator is proposed by Sun and Ren [9], in which the solutions that are evolved for an infinitely small time interval from the initial condition in terms of the primitive variables are derived by means of a limit operation to let the evolution time approach to zero. The semi-discrete finite volume scheme decouples the temporal discretization and the spatial discretization while maintaining the genuine multidimensional nature of the original evolution Galerkin scheme.

The paper is organized as follows. In Section II, we give the compressible flows equations in the Lagrangian formulation. In Section III, the vertex solver to compute velocity and pressure by local evolution Galerkin operator is derived. In Section IV, the global description of the present algorithm is shown. In Section V, several numerical tests are shown. Some main conclusions are presented in Section VI.

II. NUMERICAL METHOD FOR COMPRESSIBLE LAGRANGIAN FLOW EQUATIONS

A. Governing equations of compressible flow

The governing equations of compressible flow without internal dissipation and external forces can express into the following integrals as the Lagrangian formalism:

$$\frac{d}{dt}\int_{\Omega(t)} \rho d\Omega = 0 \tag{1}$$

$$\frac{d}{dt}\int_{\Omega(t)} \rho \boldsymbol{u} d\Omega = -\int_{\partial\Omega(t)} p d\boldsymbol{l} \tag{2}$$

$$\frac{d}{dt}\int_{\Omega(t)} \rho E d\Omega = -\int_{\partial\Omega(t)} p\boldsymbol{u}\cdot d\boldsymbol{l} \tag{3}$$

$$\frac{d}{dt}\int_{\Omega(t)} d\Omega = -\int_{\partial\Omega(t)} \boldsymbol{u}\cdot d\boldsymbol{l} \tag{4}$$

 ISBN: 978-1-61208-419-0

where ρ is density, u and v are component velocity, p is pressure, E is specific total energy, $E = e + (u^2 + v^2)/2$, e is specific internal energy, and $\Omega(t)$ is a control volume with the boundary $\partial\Omega(t)$, $d\boldsymbol{l}$ is the differential length of the surface for the control volume.

For a given control volume W_c with the mass $m_c = \int_{\Omega_c} \rho d\Omega$ and the area $A_c = \int_{\Omega_c} d\Omega$, a definition about the average value of any physical variable f is $\bar{f}_c = \frac{1}{m_c}\int_{\Omega_c} \rho f d\Omega$. Thus, (1) becomes an algebraic equation $\bar{r}_c A_c = m_c = \text{const}$, and (2)-(4) can be written for these discrete unknowns in two-dimensional space with regard of a vector form:

$$\frac{d\bar{\boldsymbol{U}}_c}{dt} = -\frac{1}{m_c}\int_{\partial\Omega_c} \boldsymbol{H}\cdot\boldsymbol{n} dl \tag{5}$$

where $\bar{\boldsymbol{U}}_c = \left(-\tau_c, u_c, v_c, E_c\right)^T$, $\tau_c = A_c / m_c$, $\boldsymbol{n}$ is the outward unit vector normal to the boundary of the control volume, $\boldsymbol{H} = \left(u, p, 0, pu\right)^T \boldsymbol{i} + \left(v, 0, p, pv\right)^T \boldsymbol{j}$ is the tensor of fluxes.

Under Lagrangian coordinates, the control volume moves with the same velocity as the fluid particle, and the trajectory equations of any fluid particle is:

$$\frac{dx}{dt} = u\,,\ \frac{dy}{dt} = v \tag{6}$$

B. The finite volume scheme

For any nonoverlapping polygons cell with the number $N(c)$ of interfaces of the cell and a interface denoted by I_k, the flow governing equation may be written as:

$$\frac{d\bar{\boldsymbol{U}}_c}{dt} = -\frac{1}{m_c}\sum_{k=1}^{N(c)}\int_{I_k} \boldsymbol{H}_k\cdot\boldsymbol{n}_k dl \tag{7}$$

A suitable approach to solve (7) is the semi-discrete procedure that decouples the temporal discretization and the spatial discretization. Thus, the respective high order scheme about temporal and spatial discretization can be adopted independently. A second-order Runge-Kutta scheme for temporal discretization is following:

$$\begin{aligned}
\bar{\boldsymbol{U}}_c^{*} &= \bar{\boldsymbol{U}}_c^{n} - \frac{\Delta t}{m_c}\sum_{k=1}^{N(c)}\int_{I_k} \boldsymbol{H}_k\left(\boldsymbol{E}_{I,0}\boldsymbol{R}_c\bar{\boldsymbol{U}}_c^{n}\right)\cdot\boldsymbol{n}_k dl \\
\bar{\boldsymbol{U}}_c^{n+1} &= \frac{1}{2}\bar{\boldsymbol{U}}_c^{n} + \frac{1}{2}\left[\bar{\boldsymbol{U}}_c^{*} - \frac{\Delta t}{m_c}\sum_{k=1}^{N(c)}\int_{I_k} \boldsymbol{H}_k\left(\boldsymbol{E}_{I,0}\boldsymbol{R}_c\bar{\boldsymbol{U}}_c^{*}\right)\cdot\boldsymbol{n}_k dl\right]
\end{aligned} \tag{8}$$

where $\boldsymbol{R}_c$ is the reconstruction operator which transforms the cell averages of the conservative variables to their spatial distributions, and $\boldsymbol{E}_{I,0}$ is an approximate Galerkin evolution operator to compute the solution at time $t_n^{+} = t_n + 0$ on cell interface I_k.

The rest part of this section will give the procedures for the reconstruction and the numerical integration of the interface flux in (8), while the approximate Galerkin evolution operator will be discussed in Section III.

C. The reconstructions

Usually, the reconstruction is carried out in terms of the primitive physical variables $\boldsymbol{q} = (\rho, u, v, p)^T$ from the cell average data $\bar{\boldsymbol{q}}_c$. To obtain a spatially first-order scheme, a piecewise constant reconstruction is sufficient; and to obtain a spatially second-order scheme, a piecewise linear reconstruction is sufficient.

D. The numerical integration of the interface flux

In order to give the relation between variables at interface and variables at vertex and ensure the equivalent discretization between the numerical flux across interface and the numerical flux at vertex at the same time, the numerical integration to the interface flux in (8) may adopt the following midpoint rule [6]:

$$\begin{aligned}
&\sum_{k=1}^{N(c)}\int_{I_k} \boldsymbol{H}_k\left(\boldsymbol{E}_{I,0}\boldsymbol{R}_c\bar{\boldsymbol{U}}_c\right)\cdot\boldsymbol{n}_k dl \\
&= \frac{1}{2}\sum_{r=1}^{N(c)}\left[\boldsymbol{H}_r\left(\boldsymbol{E}_0\boldsymbol{R}_c\bar{\boldsymbol{U}}_c\right) + \boldsymbol{H}_{r+1}\left(\boldsymbol{E}_0\boldsymbol{R}_c\bar{\boldsymbol{U}}_c\right)\right]\cdot\boldsymbol{n}_{r,r+1}L_{r,r+1}
\end{aligned} \tag{9}$$

where $\boldsymbol{E}_0$ is the vertex solver from the local Galerkin evolution operator to compute the solution at the cell vertex at time $t_n^{+} = t_n + 0$, and r is the numbering of the vertices counterclockwise, $L_{r,r+1}$ denotes the length of an interface $[M_r, M_{r+1}]$ about the neighbouring vertices M_r and M_{r+1} and $\boldsymbol{n}_{r,r+1}$ denotes the outward unit vector normal to the interface $[M_r, M_{r+1}]$.

III. VERTEX SOLVER $\boldsymbol{E}_0$ BY THE LOCAL EVOLUTION GALERKIN OPERATOR

The central idea of the local evolution Galerkin operator is to compute the theoretical solutions along every bicharacteristic direction for a small time interval from the initial conditions about the hyperbolic equations, and then the theoretical solutions are made some approximate operations and limit operations to obtain the local approximate operator.

In order to derive the theoretical evolution Galerkin solutions about the nonlinear hyperbolic system, a suitable local linearization is usually utilized with regard to the primitive variables, so that the bicharacteristics are reduced to straight lines. For this purpose, we have:

$$\frac{d\boldsymbol{q}}{dt} + \boldsymbol{A}_1(\boldsymbol{q})\frac{\partial\boldsymbol{q}}{\partial x} + \boldsymbol{A}_2(\boldsymbol{q})\frac{\partial\boldsymbol{q}}{\partial y} = 0$$

where $\boldsymbol{q} = \begin{bmatrix}\rho\\ u\\ v\\ p\end{bmatrix}$, $\boldsymbol{A}_1(\boldsymbol{q}) = \begin{bmatrix}0 & \rho & 0 & 0\\ 0 & 0 & 0 & 1/\rho\\ 0 & 0 & 0 & 0\\ 0 & \rho c^2 & 0 & 0\end{bmatrix}$,

$\boldsymbol{A}_2(\boldsymbol{q}) = \begin{bmatrix}0 & 0 & \rho & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1/\rho\\ 0 & 0 & \rho c^2 & 0\end{bmatrix}$.

The flow equations are linearized by freezing the Jacobian matrices about a reference state $\tilde{\boldsymbol{q}} = (\tilde{\rho}, \tilde{u}, \tilde{v}, \tilde{p})$ at point $\tilde{P} = (\tilde{x}, \tilde{y}, \tilde{t})$. The linearized system with frozen constant Jacobian matrices can be written as:

$$\frac{d\boldsymbol{q}}{dt} + \boldsymbol{A}_1(\tilde{\boldsymbol{q}})\frac{\partial \boldsymbol{q}}{\partial x} + \boldsymbol{A}_2(\tilde{\boldsymbol{q}})\frac{\partial \boldsymbol{q}}{\partial y} = 0 \qquad (10)$$

Considering any unit vector denoted by $\boldsymbol{n}(\theta) = (n_x, n_y)^T = (\cos\theta, \sin\theta)^T$, $\theta \in [0, 2\pi]$, there is a matrix pencil $\boldsymbol{A}(\tilde{\boldsymbol{q}}, \theta) = n_x \boldsymbol{A}_1(\tilde{\boldsymbol{q}}, \theta) + n_y \boldsymbol{A}_2(\tilde{\boldsymbol{q}}, \theta)$, which has four real eigenvalues: $\lambda_1 = \tilde{c}$, $\lambda_{2,3} = 0$, $\lambda_4 = -\tilde{c}$, and four corresponding linearly-independent right eigenvectors $\boldsymbol{r}_1 = \left(-\tilde{\rho}/\tilde{c}, \cos\theta, \sin\theta, -\tilde{\rho}\tilde{c}\right)^T$, $\boldsymbol{r}_2 = \left(1, 0, 0, 0\right)^T$, $\boldsymbol{r}_3 = \left(0, \sin\theta, \cos\theta, 0\right)^T$, $\boldsymbol{r}_4 = \left(\tilde{\rho}/\tilde{c}, \cos\theta, \sin\theta, \tilde{\rho}\tilde{c}\right)^T$. The four right eigenvectors may construct a right eigenmatrix $\boldsymbol{R}$, and the characteristic variables can be define as $\boldsymbol{w} = \boldsymbol{R}^{-1}\boldsymbol{q}$.

Multiplying system (10) by $\boldsymbol{R}^{-1}$ from the left, an eigen-system can be obtained

$$\frac{d\boldsymbol{w}}{dt} + \boldsymbol{R}^{-1}\boldsymbol{A}_1\boldsymbol{R}\frac{\partial \boldsymbol{w}}{\partial x} + \boldsymbol{R}^{-1}\boldsymbol{A}_2\boldsymbol{R}\frac{\partial \boldsymbol{w}}{\partial y} = 0 \qquad (11)$$

Thus, (11) can be transformed into the following quasi-diagonalized system:

$$\frac{d\boldsymbol{w}}{dt} + \boldsymbol{\Lambda}_1\frac{\partial \boldsymbol{w}}{\partial x} + \boldsymbol{\Lambda}_2\frac{\partial \boldsymbol{w}}{\partial y} = \boldsymbol{s} \qquad (12)$$

where

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(-\frac{p}{\tilde{\rho}\tilde{c}} + u\cos\theta + v\sin\theta) \\ \rho - \frac{p}{\tilde{c}^2} \\ u\sin\theta - v\cos\theta \\ \frac{1}{2}(\frac{p}{\tilde{\rho}\tilde{c}} + u\cos\theta + v\sin\theta) \end{bmatrix},$$

$$\boldsymbol{s} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\tilde{c}(\sin\theta\frac{\partial w_3}{\partial x} - \cos\theta\frac{\partial w_3}{\partial y}) \\ 0 \\ \tilde{c}\sin\theta(\frac{\partial w_1}{\partial x} - \frac{\partial w_4}{\partial x}) - \tilde{c}\cos\theta(\frac{\partial w_1}{\partial y} - \frac{\partial w_4}{\partial y}) \\ \frac{1}{2}\tilde{c}(-\sin\theta\frac{\partial w_3}{\partial x} + \cos\theta\frac{\partial w_3}{\partial y}) \end{bmatrix},$$

$$\boldsymbol{\Lambda}_1 = diag(\lambda_{1,1}, \lambda_{1,2}, \lambda_{1,3}, \lambda_{1,4}) = \begin{bmatrix} -\tilde{c}\cos\theta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tilde{c}\cos\theta \end{bmatrix},$$

$$\boldsymbol{\Lambda}_2 = diag(\lambda_{2,1}, \lambda_{2,2}, \lambda_{2,3}, \lambda_{2,4}) = \begin{bmatrix} -\tilde{c}\sin\theta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tilde{c}\sin\theta \end{bmatrix}.$$

Equation (12) shows that each characteristic variable w_l (l=1,2,3,4) is evolved along the corresponding bicharacteristic curve:

$$\left(\frac{d\boldsymbol{z}}{dt}\right)_l = \left(\lambda_{1,l}(\theta), \lambda_{2,l}(\theta)\right)^T, l = 1, 2, 3, 4,$$

where $\boldsymbol{z} = (x, y)^T$, according to the relation

$$\frac{Dw_l}{Dt} = \frac{dw_l}{dt} + \lambda_{1,l}(\theta)\frac{\partial w_l}{\partial x} + \lambda_{2,l}(\theta)\frac{\partial w_l}{\partial y} = s_l.$$

Therefore, given the initial condition at time $\tilde{t}$, the solution of w_l at $P = (x, y, \tilde{t} + \tau)$ is:

$$w_l(x, y, \tilde{t} + \tau, \theta) = w_l[x - \lambda_{1,l}(\theta)\tau, y - \lambda_{2,l}(\theta)\tau, \tilde{t}] + \hat{s}_l(\theta) \qquad (13)$$

where

$$\hat{s}_l(\theta) = \int_{\tilde{t}}^{\tilde{t}+\tau} s_l[x - \lambda_{1,l}(\theta)(\tilde{t} + \tau - \xi), y - \lambda_{2,l}(\theta)(\tilde{t} + \tau - \xi), \xi]d\xi.$$

For any given angle θ, the four bicharacteristic curves from $P(x, y, \tilde{t} + \tau)$ denoted by $C_l(\theta)$ are depicted in Figure 1. The $C_1(\theta)$ or $C_4(\theta)$, for θ from 0 to 2π, generates a bicharacteristic cone or Mach cone, and the $C_2(\theta)$ or $C_3(\theta)$ is perpendicular to the bottom of the cone. The intersection point between $C_l(\theta)$ and the initial plane with $\tilde{P}(\tilde{x}, \tilde{y}, \tilde{t})$ is denoted by $Q_l(\theta)$. For $\theta \in [0, 2\pi]$, the $Q_1(\theta)$ and $Q_4(\theta)$ locate in the circle with the center point $\tilde{P}(\tilde{x}, \tilde{y}, \tilde{t})$ and the radius $\tilde{c}\tau$, and $Q_4(\theta) = Q_1(\theta + \pi)$, moreover, the $Q_2(\theta)$ and $Q_3(\theta)$ locate in the initial point $\tilde{P}(\tilde{x}, \tilde{y}, \tilde{t})$, and $Q_2(\theta) = Q_3(\theta)$. So, the expressions may hold: $Q_{1,4}(\theta) = (x \pm \tilde{c}\tau\cos\theta, y \pm \tilde{c}\tau\sin\theta, \tilde{t})$, $Q_{2,3}(\theta) = (x, y, \tilde{t})$.

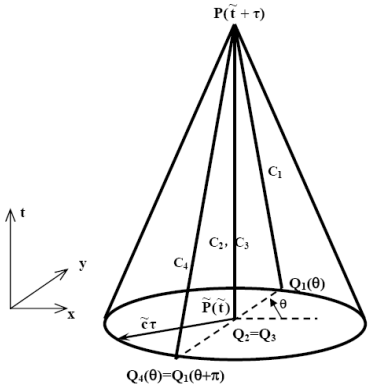


Figure 1. Bicharacteristic curves and bicharacteristic cone

Equation (13) can be also written in vector form as

$$\boldsymbol{w}(P, \theta) = \begin{pmatrix} w_1(Q_1(\theta)) \\ w_2(Q_2(\theta)) \\ w_3(Q_3(\theta)) \\ w_4(Q_4(\theta)) \end{pmatrix} + \begin{pmatrix} \hat{s}_1(\theta) \\ \hat{s}_2(\theta) \\ \hat{s}_3(\theta) \\ \hat{s}_4(\theta) \end{pmatrix}. \qquad (14)$$

Multiplying (14) with the right eigenmatrix $\boldsymbol{R}$ from the left and then integrating with respect to θ from 0 to 2π, it leads to:

$$\boldsymbol{q}(P)=\frac{1}{2\pi}\int_0^{2\pi}\left[\sum_{l=1}^{4}\boldsymbol{r}_l\left(w_l\left(Q_l(\theta)\right)+\hat{s}_l(\theta)\right)\right]d\theta\,.$$

The symmetries in characteristic variables and the source terms are used to obtain the solutions of the linearized hyperbolic system:

$$\begin{aligned}u(P)&=\frac{1}{2}u(\tilde{P})\\&+\frac{1}{2\pi}\int_0^{2\pi}\left[-\frac{p(Q_1)}{\tilde{\rho}\tilde{c}}\cos\theta+u(Q_1)\cos^2\theta+v(Q_1)\sin\theta\cos\theta\right]d\theta\\&+\frac{1}{2\pi}\int_0^{2\pi}\int_{\tilde{t}}^{\tilde{t}+\tau}S[\boldsymbol{z}+\tilde{c}(\tilde{t}+\tau-\xi)\boldsymbol{n}(\theta),\xi,\theta]\cos\theta\,d\xi\,d\theta\\&-\frac{1}{2\tilde{\rho}}\int_{\tilde{t}}^{\tilde{t}+\tau}\frac{\partial p(\boldsymbol{z},\xi)}{\partial x}d\xi\end{aligned}\tag{15}$$

$$\begin{aligned}v(P)&=\frac{1}{2}v(\tilde{P})\\&+\frac{1}{2\pi}\int_0^{2\pi}\left[-\frac{p(Q_1)}{\tilde{\rho}\tilde{c}}\sin\theta+u(Q_1)\cos\theta\sin\theta+v(Q_1)\sin^2\theta\right]d\theta\\&+\frac{1}{2\pi}\int_0^{2\pi}\int_{\tilde{t}}^{\tilde{t}+\tau}S[\boldsymbol{z}+\tilde{c}(\tilde{t}+\tau-\xi)\boldsymbol{n}(\theta),\xi,\theta]\sin\theta\,d\xi\,d\theta\\&-\frac{1}{2\tilde{\rho}}\int_{\tilde{t}}^{\tilde{t}+\tau}\frac{\partial p(\boldsymbol{z},\xi)}{\partial y}d\xi\end{aligned}\tag{16}$$

$$\begin{aligned}p(P)&=\frac{1}{2\pi}\int_0^{2\pi}\left[p(Q_1)-\tilde{\rho}\tilde{c}u(Q_1)\cos\theta-\tilde{\rho}\tilde{c}v(Q_1)\sin\theta\right]d\theta\\&-\frac{1}{2\pi}\tilde{\rho}\tilde{c}\int_0^{2\pi}\int_{\tilde{t}}^{\tilde{t}+\tau}S[\boldsymbol{z}+\tilde{c}(\tilde{t}+\tau-\xi)\boldsymbol{n}(\theta),\xi,\theta]d\xi\,d\theta\end{aligned}\tag{17}$$

where $S(\boldsymbol{z},t,\theta)=\tilde{c}\left[\frac{\partial u(\boldsymbol{z},t,\theta)}{\partial x}\sin^2\theta+\frac{\partial v(\boldsymbol{z},t,\theta)}{\partial y}\cos^2\theta\right]$

$-\frac{\tilde{c}}{2}\left[\frac{\partial u(\boldsymbol{z},t,\theta)}{\partial y}+\frac{\partial v(\boldsymbol{z},t,\theta)}{\partial x}\right]\sin 2\theta$.

For discretized grids, we assume that there are M control volumes with a common vertex $\boldsymbol{z}=(x,y)^T$, and θ_{ka} and θ_{kb} respectively are the starting and ending angles of the kth $k(k\le M)$ grid about the common vertex, thus (15)-(17) can be rewritten into:

$$\begin{aligned}u(P)&=\frac{1}{2}u(\tilde{P})\\&+\frac{1}{2\pi}\sum_{k=1}^{M}\int_{\theta_{ka}}^{\theta_{kb}}\left[-\frac{p(Q_1)}{\tilde{\rho}\tilde{c}}\cos\theta+u(Q_1)\cos^2\theta+v(Q_1)\sin\theta\cos\theta\right]d\theta\\&+\frac{1}{2\pi}\sum_{k=1}^{M}\int_{\theta_{ka}}^{\theta_{kb}}\int_{\tilde{t}}^{\tilde{t}+\tau}S[\boldsymbol{z}+\tilde{c}(\tilde{t}+\tau-\xi)\boldsymbol{n}(\theta),\xi,\theta]\cos\theta\,d\xi\,d\theta\\&+\frac{1}{2\tilde{\rho}}\int_{\tilde{t}}^{\tilde{t}+\tau}\frac{\partial p(\boldsymbol{z},\xi)}{\partial x}d\xi\end{aligned}\tag{18}$$

$$\begin{aligned}v(P)&=\frac{1}{2}v(\tilde{P})\\&+\frac{1}{2\pi}\sum_{k=1}^{M}\int_{\theta_{ka}}^{\theta_{kb}}\left[-\frac{p(Q_1)}{\tilde{\rho}\tilde{c}}\sin\theta+u(Q_1)\cos\theta\sin\theta+v(Q_1)\sin^2\theta\right]d\theta\\&+\frac{1}{2\pi}\sum_{k=1}^{M}\int_{\theta_{ka}}^{\theta_{kb}}\int_{\tilde{t}}^{\tilde{t}+\tau}S[\boldsymbol{z}+\tilde{c}(\tilde{t}+\tau-\xi)\boldsymbol{n}(\theta),\xi,\theta]\sin\theta\,d\xi\,d\theta\\&-\frac{1}{2\tilde{\rho}}\int_{\tilde{t}}^{\tilde{t}+\tau}\frac{\partial p(\boldsymbol{z},\xi)}{\partial y}d\xi\end{aligned}\tag{19}$$

$$\begin{aligned}p(P)&=\frac{1}{2\pi}\sum_{k=1}^{M}\int_{\theta_{ka}}^{\theta_{kb}}\left[p(Q_1)-\tilde{\rho}\tilde{c}u(Q_1)\cos\theta-\tilde{\rho}\tilde{c}v(Q_1)\sin\theta\right]d\theta\\&-\frac{1}{2\pi}\tilde{\rho}\tilde{c}\sum_{k=1}^{M}\int_{\theta_{ka}}^{\theta_{kb}}\int_{\tilde{t}}^{\tilde{t}+\tau}S[\boldsymbol{z}+\tilde{c}(\tilde{t}+\tau-\xi)\boldsymbol{n}(\theta),\xi,\theta]d\xi\,d\theta\end{aligned}\tag{20}$$

Equations (18)-(20) are the exact evolution Galerkin solutions for the linearized Lagrangian flow equations. For simplifying the computation of the integrals including pressure gradient term and source term, some approximate operations are required with the similar procedure [9]. And then, the local Galerkin evolution operator $\boldsymbol{E}_0$ about (18)-(20) at time $t_n^+=t_n+0$ is obtained to make the limit operations with $\tau\to 0$. From Figure 1, the effect of $\tau\to 0$ is to make $P\to\tilde{P}$ and $Q\to\tilde{P}$ and the length of the arc with two end points $Q(\theta_{ib})$ and $Q(\theta_{ie})$ tends to zero. Thus, we have $\boldsymbol{q}\left(Q(\theta)\right)\to\boldsymbol{q}_i$, for $\theta_{ib}\le\theta\le\theta_{ie}$, where $\boldsymbol{q}_i$ is the vector of the primitive variables at $\tilde{P}$ evaluated in terms of the reconstruction in the control volume containing the arc with two end points $Q(\theta_{ib})$ and $Q(\theta_{ie})$.

After approximate and limit operations, the analytical expressions of the vertex solver $\boldsymbol{E}_0$ by the local Galerkin evolution operator are the following:

$$\begin{aligned}u(P)&=\frac{1}{\pi}\sum_{i=1}^{N}\left[-\frac{p_i}{\tilde{\rho}\tilde{c}}(\sin\theta_{ie}-\sin\theta_{ib})\right.\\&\left.+u_i\left(\frac{\theta_{ie}-\theta_{ib}}{2}+\frac{\sin 2\theta_{ie}-\sin 2\theta_{ib}}{4}\right)-v_i\frac{\cos 2\theta_{ie}-\cos 2\theta_{ib}}{4}\right]\end{aligned}\tag{21}$$

$$\begin{aligned}v(P)&=\frac{1}{\pi}\sum_{i=1}^{N}\left[\frac{p_i}{\tilde{\rho}\tilde{c}}(\cos\theta_{ie}-\cos\theta_{ib})\right.\\&\left.-u_i\frac{\cos 2\theta_{ie}-\cos 2\theta_{ib}}{4}+v_i\left(\frac{\theta_{ie}-\theta_{ib}}{2}-\frac{\sin 2\theta_{ie}-\sin 2\theta_{ib}}{4}\right)\right]\end{aligned}\tag{22}$$

$$\begin{aligned}p(P)&=\frac{1}{2\pi}\sum_{i=1}^{N}\left[p_i(\theta_{ie}-\theta_{ib})\right.\\&\left.-\tilde{\rho}\tilde{c}u_i(\sin\theta_{ie}-\sin\theta_{ib})+\tilde{\rho}\tilde{c}v_i(\cos\theta_{ie}-\cos\theta_{ib})\right]\end{aligned}\tag{23}$$

It was found that the vertex solver $\boldsymbol{E}_0$ is able to take multidimensional effect into account in a natural way, and to consider the effect of the different sonic impedences to straightway apply to multimaterial flows, and to be fully competent for the structured or unstructured grids.

 ISBN: 978-1-61208-419-0

IV. DESCRIPTION OF THE PRESENT ALGORITHM

Step 1: Initialization

At time $t = t^n$, the geometrical coordinates x_i^n , y_i^n ($i = 1,2,\cdots,I$) of each vertex of each cell and the physical variables $\overline{\rho}_k^n$ ($\overline{\tau}_k^n$), $\overline{u}_k^n$, $\overline{v}_k^n$, $\overline{E}_k^n$, $\overline{p}_k^n$ ($k = 1,2,\cdots,K$) at center of each cell are known.

Step 2: Reconstruction

The physical primitive variables at each vertex of each cell are obtained by means of the formula in Subsection II.C.

Step 3: Vertex solver

The velocities u_i^n , v_i^n and pressure p_i^n ($i = 1,2,\cdots,I$) at each vertex of each cell are obtained by means of (21)-(23) for the local Galerkin evolution operator $\boldsymbol{E}_0$.

Step 4: Update of the geometrical quantities

The updated grids and the length and outward vector of each interface are achieved from the new coordinate data x_i^{n+1} , y_i^{n+1} ($i = 1,2,\cdots,I$) of each vertex of each cell.

Step 5: Update of the physical variables

The physical variables $\overline{\tau}_k^{n+1}$, $\overline{u}_k^{n+1}$, $\overline{v}_k^{n+1}$, $\overline{E}_k^{n+1}$ at center of the updated grids can be computed from (8), and then the corresponding $\overline{p}_k^{n+1}$ is obtained from the equation of state.

V. NUMERICAL RESULTS

A. Multimaterial Sod's shock tube problem

The initial conditions of two kinds of perfect gases with different adiabatic indexes are: $(\rho, u, p, \gamma) = (1, 0, 1, 7/5)$ in the left-hand side and $(\rho, u, p, \gamma) = (0.125, 0, 0.1, 5/3)$ in the right-hand side. The density solution at time $t = 0.2$ is shown in Figure 2 for the second-order scheme with CFL=0.8 under different meshes. It can be found that the smaller the grid used, the closer the numerical solution approaches to the exact solution, and there is not unphysical oscillation nearby the shock wave, the rarefaction fan is correctly described. An undershoot appears at the density discontinuity about the contact discontinuity, it is an indigenous property to Lagrangian method.

B. Sedov problem

A highly intense shock wave generated by a strong explosion propagates outward. The perfect gas with adiabatic index $\gamma = 5/3$ is initially at rest for $(\rho, u, p) = (1, 0, 0)$ but an energy spike is set as 182.09 at the center, and all the boundary conditions are solid walls. The 30×30 uniform Cartesian meshes in computational domain $[0,1.1] \times [0,1.1]$ are used with CFL=0.8. Figure 3(a-b) shows the calculated meshes and density contours by the second-order scheme at time t=1, and Figure 4(a-b) shows the density profile by the first-order and the second-order scheme at time t=1. It is found that the second-order scheme has an improved precision on the first-order scheme, and the second-order scheme still has excellent resolution and symmetry.

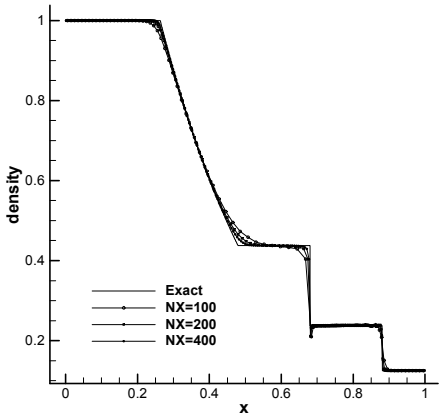
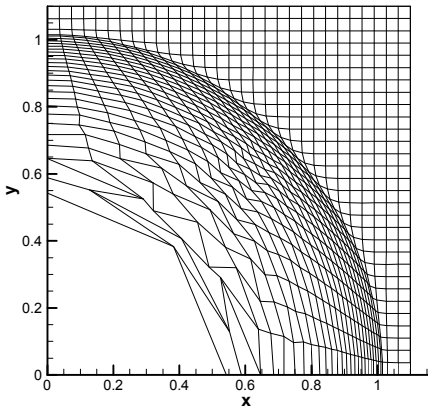
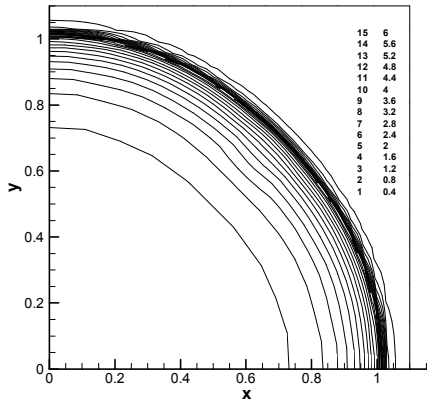


Figure 2. Comparison of numerical solution for Sod's shock tube with exact solution for second-order scheme



(a) Meshes variation

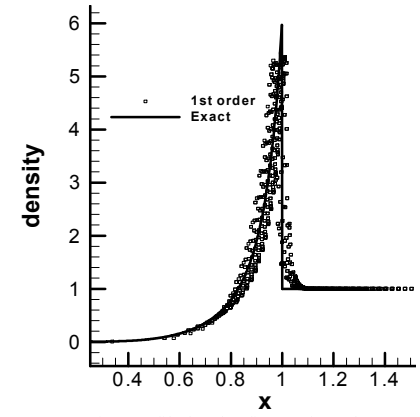


(b) Density contours

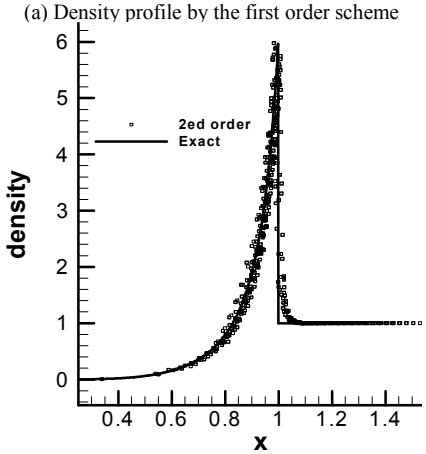
Figure 3. Solution of 2nd order scheme for Sedov problem at time t=1

C. Saltzman problem

A planar shock wave located initially at $x = 0$ moves rightward on the perfect gas with $\gamma = 5/3$, and the front state of shock wave is $(\rho, u, p) = (1, 0, 0)$. When the piston velocity at the left-hand is set as $u = 1$, the exact propagation velocity of shock wave should be $4/3$. A computational domain $[0,1.0] \times [0,0.1]$ on Cartesian coordinates with grid



(a) Density profile by the first order scheme



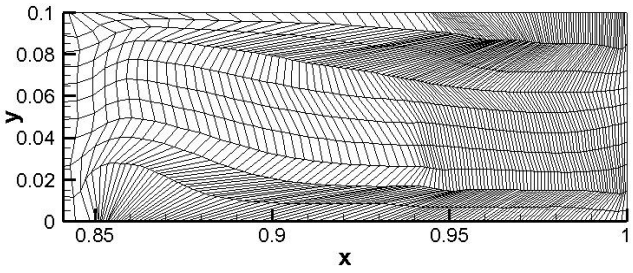
(b) Density profile by the second order scheme

Figure 4. Density solution for Sedov problem at time t=1

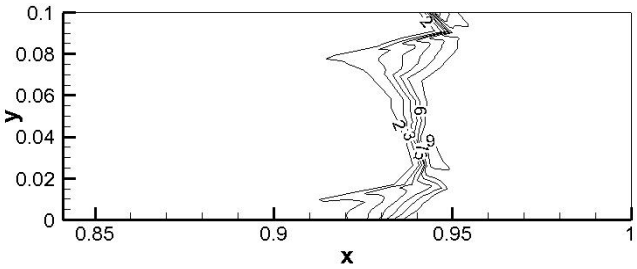
number 100×10 is taken, and a uniform meshes in y direction and a nonuniform meshes with the mapping $x = i\Delta x + (0.1 - j\Delta y)\sin(i\Delta x\pi)$ in x direction are used, meanwhile, the boundary condition at the left-hand is an invariable velocity and all the other boundaries are set as solid wall. Obviously, the shock wave will reflect on the right-hand wall at time t=0.75. The meshes and density contours at time t=0.84 are shown in Figure 5 about the reflected shock wave. It can be found that the one-dimensional property of the reflected shock wave can be well preserved. The robustness of this scheme is also powerfully demonstrated by the test case.

VI. CONCLUSION AND FUTURE WORK

A cell-centered Lagrangian method for 2D compressible flows is present on basis of the local evolution Galerkin scheme under semi-discrete finite volume framework where the vertex velocity is computed in a coherent manner with the numerical fluxes across the cell interface. The main feature of this method is the physical variables at vertex of a cell are computed by virtue of the bicharacteristics theory about the linearized flow equations, which is essentially a multidimensional Riemann solver taking "multidimensional effect" into account in a natural way. Our future most important works will be on the extension to arbitrary Lagrangian-Eulerian method.



(a) Meshes



(b) Density contours

Figure 5. Computational results about Salzman problem at t=0.84

ACKNOWLEDGMENT

This work was supported under Grant-11272064 of Natural Science Foundation of China.

REFERENCES

[1] M. Wilkins, "Calculation of Elastic-Plastic Flow, in Methods in Computational Physics", vol.3, Academic Press, New York, 1964, pp. 211-263.

[2] R. Ritchtmyer and K. Morton, Difference Methods for Initial-Value Problems, John Wiley, New York, 1967.

[3] J. Dukowicz and B. Meltz, "Vorticity Errors in Multidimensional Lagrangian Codes", Journal of Computational Physics, 99, 1992, pp. 115–134.

[4] C. Juan and S. Chi-Wang, "A High Order ENO Conservative Lagrangian Type Scheme for the Compressible Euler Equations", Journal of Computational Physics, 227, 2007, pp. 1567-1596.

[5] B. Després and C. Mazeran, "Lagrangian Gas Dynamics in Two Dimensions and Lagrangian Systems", Arch. Rational Mech. Anal., 178, 2005, pp. 327–372.

[6] P. Maire, R. Abgrall and J. Breil, "A Cell-Centered Lagrangian Scheme for Two-Dimensioanl Compressible Flow Problems", SIAM Journal of Scientific Computing, 29(4), 2007, pp. 1781-1824.

[7] M. Lukáčová-Medvid'ová, K. Morton and G. Warnecke, "Evolution Galerkin Methods for Hyperbolic Systems in Two Space Dimensions", Mathematics of Computation, 69(232), 2000, pp. 1355-1384.

[8] M. Lukáčová-Medvid'ová, J. Saibertová and G. Warnecke, "Finite Volume Evolution Galerkin Methods for Nonlinear Hyperbolic Systems", Journal of Computational Physics, 183, 2002, pp. 533–562.

[9] Y. Sun and Y.-X. Ren, "The Finite Volume Local Evolution Galerkin Method for Solving the Hyperbolic Conservation Laws", Journal of Computational Physics, 228, 2009, pp. 4945-4960.

 ISBN: 978-1-61208-419-0

A Patent Quality Classification System Using a Kernel-PCA with SVM

Pei-Chann Chang
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
iepchang@saturn.yzu.edu.tw

Jheng-Long Wu
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
vickeytsao@gmail.com

Cheng-Chin Tsao
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
vickeytsao@gmail.com

Meng-Hsuan Lin
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
syuan417@gmail.com

***Abstract*—Data mining (DM) approaches such as clustering and classification are employed in this paper to identify and classify the patent quality. We develop an effective and automatic patent quality classification system. First, the Self-organizing map (SOM) is used to cluster patents automatically into different quality groups with patent quality indicators instead of via expert identification. Then, the Kernel principal component analysis (kernel-PCA) is used to extract key indicator to improve classification performance. Finally, the Support vector machine (SVM) is used to build the quality classification model. The proposed classification model is applied to classify patent quality automatically in solar industries. Experimental results show that our proposed approach KPCA-SVM can improve the performance of the patent quality classification when compared with the traditional method. Another advantage is that the computational time is largely reduced.**

Keywords- patent quality classification, self-organizing maps, support vector machine, kernel principal component analysis, solar industries.

I. INTRODUCTION

An important issue of patent analysis is patent quality. The high quality patent information can ensure success for business decision-making process or product development [1-2]. This study reviewed the patent analysis approaches that can understand patent status like patent quality, novelty, litigation, trends and so on [3]. In addition, traditional patent analysis requires spending much time, cost and manpower. Therefore, the potential for high quality patent determining approach need to shorten the time for business or production. Recently, the self-organizing map (SOM) is used to analysis patent for patent trend [4] and regional innovation systems [5]. It can analyse patent current situation and trend, but it doesn't provide a solution for determine future patent quality. The future patent recognition is a key research for present time because patent impact on the industry. The future patent recognition is a key research for the time because patent impact on the industry need to response quickly. Therefore, support vector machine (SVM) forecasting model can solve patent classification problem for predict future unknown patent classification such as quality [6]. In this study, we propose a KPCA-SVM patent quality classification system that combines three data mining (DM) methods such as SOM, Kernel-PCA and SVM. The SOM is used to cluster patents into several groups for qualities according to their data characteristics. Then, the quality indictor can compute the quality levels for delimit patents quality on each groups. The kernel principle components analysis (kernel-PCA) is based on principle components analysis and used to transform patent data into new features set by nonlinear kernel mapping. SVM is used to build classification model for patent quality problem. This methodology helps experts rank and set values on patent quality in solar industry. Therefore, a trained model can evaluate unknown patents' quality, better enable engineers and product designers forecast patent potential for product development.

II. LITERATURE REVIEWS

A. Patnet Analysis

There are various tools utilized by organizations for analyzing patents. These tools are capable of performing wide range of tasks, such as forecasting future technological trends, detecting patent infringement and determining patent quality and so on [3]. Moreover, patent analysis tools can free patent experts from the laborious tasks of analyzing the patent documents manually and determining the quality of patents. The tools assist organizations in making decisions of whether or not to invest in manufacturing of the new products by analyzing the quality of the filed patents [2]. The eventually may result in imprecise recommendation of patents. However, the larger data and indicators for patent quality forecasting are needed.

B. Patnet Quality Indicator

The primary patent quality indicators are related to investment, maintenance, and litigation, which form a basis for assessing patent quality, when the evaluation focuses on the potential patents for business. One kind of indicator of patent quality is legal status (LS) that it can show which technologies are hot and which are not for business intelligence. The legal status and search tools on the internet are very sensitive that emphasis given to the issues related to the date of availability to the public of an Internet disclosure, its conformance and its possibly non-prejudicial nature [7]. The legal status change that suggestions on how these changes may be tracked are provided, specifically resolution

 ISBN: 978-1-61208-419-0

is also complicated by the "first to invent" concept in US patent law [8].

C. Self-Organizing Maps

The SOM is a two-layer neural network that maps multidimensional data on to a two dimensional topological grid. The data are group according to similarities and patterns found in the data set, using some form of distance measure which use the Euclidean distance. The results are displayed as nodes on the map, which can be divided into different clusters based upon the distances between the clusters. Since the SOM is unsupervised, no target outcomes are provided, and the SOM is allowed to freely organize itself, so the SOM is an ideal tool for exploratory data analysis. The authors [4] used SOM to identify patent trends that they analyze patent knowledge to identify research trends. They tested on patents from the United States Patent and Trademark Office (USPTO) and result both an overview of the directions of the trends and a drill-down perspective of current trends. Another algorithm using SOM that is evolving self-organizing map (ESOM), which features an evolving network structure and fast on-line learning. Their result shows that ESOM achieved better or comparable performance with a much shorter learning process [9].

D. Kernel Principle Component Analysis

Principal component analysis (PCA) is very useful to extract nonlinear features for many research applications. The Kernel-PCA is an extension of PCA using the kernel mapping before the Eigen-problem. Kernel-PCA is as a nonlinear alternative to classical PCA of combustion composition space is investigated. PCA is mathematically defined. The PCA is widely used in many field researches. The research [1] wants to identify the key impact factors using PCA and they selected a lot of variables according to first five components indicate. The Kernel-PCA is used to critical feature extraction in stock trading model and capture best performance compared to PCA, ICA and so on [10].

E. Support Vector Machinie

Support vector machine is a machine learning algorithm and widely used for classification problems [5]. This method aims to develop an optimal hyper-plane as a decision function using the maximum margin hyper-plane between class vectors on both sides of the hyper-plane. Support vector machine map input vectors into the high dimensional feature space via the non-linear mapping. An effective decision hyper-plane is developed to distinguish the correct training data. An approach is proposed integrated with a hybrid genetic-based support vector machine (HGA-SVM) model for developing a patent classification system [11]. But they are needed expert's knowledge to analysis. The authors claim that they use these models in real-world cases of patent classification rather than only use for International Patent Classification (IPC). The study integrated the honey-bee mating optimization algorithm with SVM (HBMOSVM) for patent document categorization. In their results show that the HBMOSVM could result in better patent documentation accuracy and better F-measure performance as an evaluation index than GASVM model in patents document categorization [12].

III. PROPOSED METHODOLOGY: KPCA-SVM PATENT QUALITY CLASSIFICATION SYSTEM

This study proposes an automatic patent quality classification system that integrated methodology as KPCA-SVM; the components used are SOM, kernel-PCA and SVM approaches. Fig. 1 shows that, first, we collect the patent data related industries from the patent database, use SOM approach to cluster patens into several groups and use patent quality indicators to compute potential quality on each group for quality identification; second, the kernel-PCA extracts key indicators into nonlinear feature space; finally, SVM forecasting model is used to build patent quality classification model using nonlinear feature space by kernel-PCA in order to predict quality of future patent who has potential effectiveness. Then, we evaluate patent quality classification system and forecast quality level for each paten by our proposed system. Our system is developed as follows:

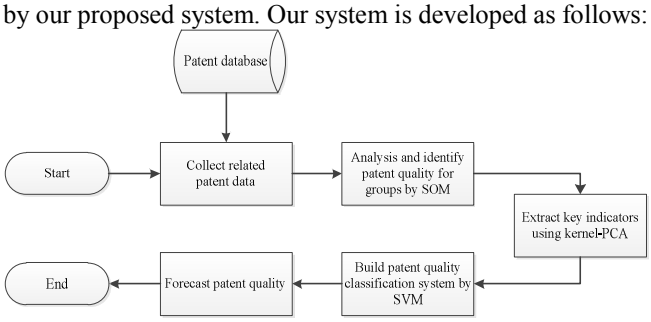


Figure 1. The system framework of KPCA-SVM.

A. Patent Quality Analysis and Identification by SOM with Quality Indicators

The SOM approach is adopted to cluster the patents, to identify patent quality and to explore hidden patterns among these patents. According to SOM, these patents will be split into several groups (clusters) and each cluster includes number of patents with a similar patent quality. This SOM analysis process continues until all input vectors are processed. Convergence criterion utilized here is in terms of epochs, which defines how many times all input vectors should be fed to the SOM for analysis. Details of the SOM algorithm are listed as follows:

- *Step 1:* Set-up the parameters in the SOM network.
- *Step 2*: Initialize each neuron weight $w_i = [w_{i1}, w_{i2}, .., w_{ij}]^T \in \Re^j$. In this study, neuron weights are initialized by drawing random samples from input dataset.
- *Step3:*Present an input pattern $x = [x_1, x_2, .., x_j]^T \in \Re^j$. In this case, the input pattern is a series of variables representing current patent status. Calculate the distance between pattern *x*, and each neuron weight w_i, and therefore, identify the winning neuron or best matching unit *c* such as

 ISBN: 978-1-61208-419-0

$$\|x - w_c\| = \min_i \{d_i\} \tag{1}$$

$$d_i = \sqrt{\sum_j (x_j - w_{ij})^2} \tag{2}$$

- *Step 4:* Adjust the weight of winning neuron *c* and all neighbor units.

$$w_i(t+1) = w_i(t) + h_{ci}(t)[x(t) - w_i(t)] \tag{3}$$

where *i* is the index of the neighbor neuron and *t* is an integer, the discrete time coordinate. The neighborhood kernel $h_{ci}(t)$ is a function of time and the distance between neighbor neuron *i* and winning neuron $ch_{ci}(t)$ defines the region of influence that the input pattern has on the SOM and consists of two parts: the neighborhood function $h(\|\cdot\|, t)$ and the learning rate function α^t,

$$h_{ci}(t) = h(\|r_c - r_i\|, t)\alpha^t \tag{4}$$

where *r* is the location of the neuron on two dimensional map grids. In this work we used Gaussian Neighborhood Function. The learning rate function $\alpha(t)$ is a decreasing function of time. The final form of the neighborhood kernel with Gaussian function is

$$h_{ci}(t) = \exp\left(\frac{\|r_c - r_i\|}{2\sigma^2(t)}\right)\alpha(t) \tag{5}$$

- **Step 5:** repeat step 3 and 4 until the convergence criterion is satisfied. Average value for each variable of each clustered group was calculated after the patent cases were clustered, and the average value for each variable of each group would be the basis when finding the most matching group for the new case. After the set of patent data has been processed by SOM, a new case can be categorized into a pre-defined group.

The patents of each group will compute the quality by quality indicators such as legal status. The quality levels of each group are calculated as follows:

$$Quality(Group_g) = \frac{1}{m \times n} \sum_{i=1, i \in Group_g}^{m} \sum_{j=1}^{n} q_{ij} \tag{6}$$

where q_{ij} denotes value of quality of *j*th variable of *i*th patent in *g*th group.

B. Extracting Key Patent Indicators by Kernel-PCA

In order to compute dot products of the form, we use kernel representation of the form.

$$K(x_i, x_j) = (\Phi(x_i), \Phi(x_j)) \tag{7}$$

which allows us to compute the value of the dot product in *F* without having to carry out the map Φ. A number of kernel functions exist as been chosen before we apply the algorithm. The representative Gaussian kernel function *K* is described as follows:

$$K(x_i, x_j) = e(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \tag{8}$$

Given a set of *m*-dimensional normalized patent indices $x_k \in R^m$, we compute the kernel matrix $K \in R^{N \times N}$ from two kernel methods,

$$K_{ij} = (\Phi(x_i), \Phi(x_j)) = [k(x_i, x_j)] \tag{9}$$

Carry out mean centering in the feature space for $\sum_{k-1}^{N} \tilde{\Phi}(xk) = 0$,

$$K' = K - C*K - K*C + C*K*C \tag{10}$$

C. Building Patent QualityClassification Model by SVM

The patent data of identified quality will be split into two datasets, i.e., training data set and testing data set. The new feature spaces of training data *training*(tr^k) and testing data *testing*(ts^k) are represented by the $\tilde{\Phi}(x)$ and α_i^k. For patent variables *x* in training period, we extract a nonlinear component via

$$\begin{aligned} Training\ (tr^k) &= (v^k, \tilde{\Phi}(x)) = \sum_{i=1}^{N} \alpha_i^k (\tilde{\Phi}(x_i), \tilde{\Phi}(x)) \\ &= \sum_{i=1}^{N} \alpha_i^k \tilde{K}(x_i, x) \end{aligned} \tag{11}$$

where $\tilde{\Phi}(x)$ is the mean centered.

The testing data is unknown data as well as the future data. We cannot directly use the testing data to compute the mean centering $\Phi(x)$ and eigenvalues α in PCA processes. In order to avoid this problem, we use the $\tilde{\Phi}(x)$ and α_i^k from the training data to extract a nonlinear component.

$$\begin{aligned} Testing\ (ts^k) &= (v^k, \tilde{\Phi}(x)) = \sum_{i=1}^{N} \alpha_i^k (\tilde{\Phi}(y_i), \tilde{\Phi}(x)) \\ &= \sum_{i=1}^{N} \alpha_i^k \tilde{K}(y_i, x) \end{aligned} \tag{12}$$

 ISBN: 978-1-61208-419-0

where y denotes the normalized variables $y_k \in R^m$ in testing data.

The input vector of SVM training employ the new nonlinear feature space *training*(tr^k) by kernel-PCA and the output vector of quality level is given by SOM with quality indicators. The trained model of SVM is then used to evaluate the testing data *testing*(tr^k) for patent quality forecasting. Thus, the proposed model can be applied to build the patent quality model for evaluating the potential of patents.

IV. EXPERIMENTAL RESULTS

In this paper, these patents were divided into seven groups in order to discriminating quality levels into seven degrees. The parameters of SOM including epochs and cluster are set up to 5,000 and 7, respectively. In SVM model, the kernel is radial basis function (RBF), while the cost is 256 and gamma is 0.25. The patent data of solar industry are collected from patent database of Thomson Innovation which has 60,000 patents in eleven patent offices. Seven different quality groups are used in SOM and its quality levels are sorted by legal status of quality indictor. The Table 1 shows that the highest legal status is in group 7, the second is in group 6 and the lowest is in group 1. The legal status on group 7 is 10.120 and the number of patents is 2,235. We observed that the higher of patent quality is, the less of patent number is. The other way around is for low patent quality patent.

TABLE I. THE LEGAL STATUS STATISTICS ON EACH GROUP

	Group						
	G1	*G2*	*G3*	*G4*	*G5*	*G6*	*G7*
No. of patent	22,431	6,249	9,796	4,564	11636	3,089	2,235
Avg. of legal status	1.90	2.15	2.35	3.41	3.53	9.06	10.10

Fig. 2 shows that the distribution on patent applications for 11 patent offices. The two larger shares of patents are US and CN. Their patents exist in different quality levels because their markets are the most important economic region attracting many patent applications.

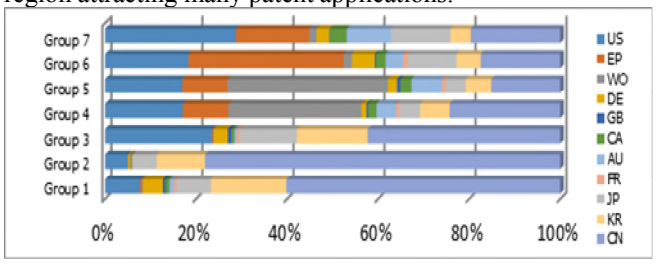


Figure 2. Distribution on patent applications for 11 patent offices

We focus on group 7 since the patent offices of US; CN; EP; JP and AU include 88% shares as shown in Fig. 3. Other offices are much less involved in solar industry. In addition, the patent applies in EP patent offices are high quality which are in Group 4, 5, 6 and 7. However, the quality levels are evenly spread in CN patent offices.

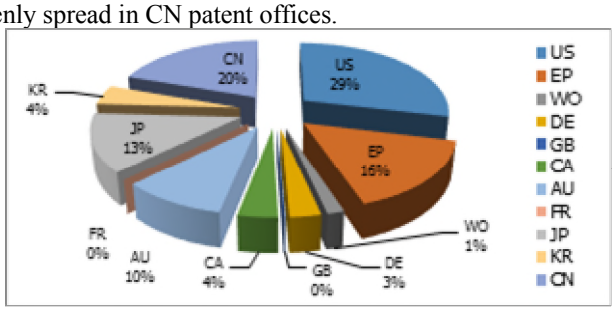


Figure 3. Distribution on patent applications on highest quality group (Group 7)

The results show the forecasting performance for KPCA-SVM and decision tree (DT). The optimal parameters were decided from the training data for kernel-PCA and SVM. Table 2 shows that our proposed model has best performance on three measure indicators and they are 99.71% on accuracy, 99.28% on average precision of all classes, and 99.31% on average recall of all classes.

TABLE II. FORECASTING PERFORMANCE FOR DIFFERENT CLASSIFICATION MODEL

Classifier	Accuracy	Precision	Recall
KPCA-SVM	99.91%	99.28%	99.31%
DT	96.86%	95.42%	95.08%

V. CONCLUSIONS

We proposed the KPCA-SVM patent quality system which combined SOM, kernel-PCA and SVM data mining approaches in solar industry. The experimental results showed that the proposed approach has a better performance when compared with other traditional approaches in terms of time consuming, cost and manpower. The proposed approach take a shorten time to determine patent quality and has 99.91% accuracy. In addition, the proposed system performs even better for a larger patent data and making fast and accurate recommendations. In the future research work, we will consider the relationship between patent quality and patent value creation. Therefore, different patent quality can be closely related to different patent values via accurate classification system.

REFERENCES

[1] Trappey, A.J.C., Trappey, C.V., Wu, C.Y. & Lin, C.L, A patent quality analysis for innovative technology and product development, Adv Eng Inform, 26(1), pp. 26-34, 2012.

[2] Trappey, A.J.C., Trappey, C.V., Wu, C.Y.W., Fan, C.Y. & Lin, Y.L., Intelligent patent recommendation system for Innovative design collaboration, J Netw Comput Appl, 36, pp. 1441-1450, 2013.

[3] Abbas, A., Zhang, L. & Khan, S.U., A literature review on the state-of-the-art in patent analysis, World Pat Inf, 37, pp. 3-13, 2014.

 ISBN: 978-1-61208-419-0

[4] Segev, A. & Kantola, J., Identification of trends from patents using self-organizing maps, Expert Syst Appl, 39(18), pp. 13235-13242, 2012.

[5] Hajek, P., Henriques, R. & Hajkova, V., Visualising components of regional innovation systems using self-organizing maps—Evidence from European regions, Technol Forecast Soc Change, 84,pp.197-214, 2014.

[6] Ercan, S. & Kayakutlu, G., Patent value analysis using support vector machines, Soft Comput, 18, pp. 313-328, 2014.

[7] Archontopoulos, E., Prior art search tools on the Internet and legal status of the results: a European Patent Office perspective, World Pat Inf, 26(2), pp. 113-121, 2004.

[8] Simmons, E.S. & Spahl, B.D., Of submarines and interference: legal status changes following citation of an earlier US patent or patent application under 35 USC §102 (e), World Pat Inf, 22(3), pp. 191-203, 2000.

[9] Deng, D. & Kasabov, N., On-line pattern analysis by evolving self-organizing maps, Neurocomputing, 51, pp. 87-103, 2003.

[10] Chang, P.C. & Wu, J.L., A critical feature extraction by kernel PCA in stock trading model, Soft Comput., DOI 10.1007/s00500-014-1350-5.

[11] Wu, C.H., Ken, Y. & Huang, T., Patent classification system using a new hybrid genetic algorithm support vector machine, Appl Soft Comput, 10(4) pp. 1164-1177, 2010.

[12] Chiu, C.Y., & Huang, P.T. Application of the honeybee mating optimization algorithm to patent document classification in combination with the support vector machine, Int. j. autom. smart technol, 3(3), pp.179-191, 2013.0.

 ISBN: 978-1-61208-419-0

Preference Elicitation in Decision Making Prioritisation Problems by Evolutionary Computing

Ludmil Mikhailov
Manchester Business School, University of Manchester
Manchester, United Kingdom
Email: ludi.mikhailov@manchester.ac.uk

***Abstract*—The paper investigates the application of evolutionary algorithms (EA) for solving a two-objective prioritisation problem. We propose two evolutionary computing approaches, based on single-objective and multi-objective EA. Our preliminary results from a Monte-Carlo simulation show that the multi-objective EA outperforms the single-objective solution approach with respect to accuracy and computational efficiency.**

Keywords- evolutionary computing, genetic algorithms, multiobjective optimisation, prioritisation methods.

I. INTRODUCTION

The assessment of weights of criteria and scores of alternatives is one of the most important tasks in the multicriteria decision-making. In the Analytical Hierarchy Process (AHP), proposed by Saaty [1], the values of weights and scores are assessed indirectly from comparison judgments. The elicitation process for both weights and scores is the same, so they are often called *priorities.*

The pairwise comparison process in the AHP assumes that the decision-maker can compare any two elements at a given hierarchical level and to provide a numerical value of the ratio of their importance. Comparing any two elements E_i and E_j, the decision-maker assigns a ratio a_{ij}, which represents a judgment concerning the relative importance of preference of the decision element E_i over E_j. If E_i is preferred to E_j then $a_{ij} > 1$, otherwise $0 < a_{ij} \le 1$.

A full set of ratio-scale judgments for a level with *n* elements requires *n(n-1)/2* comparisons. In order to derive a priority vector from a given set of judgments, Saaty constructs a positive reciprocal matrix $A = \{a_{ij}\}$ of the type

$$A = \begin{bmatrix} 1 & a_{12} & ... & a_{1n} \\ 1/a_{12} & 1 & ... & a_{2n} \\ ... & ... & ... & ... \\ 1/a_{1n} & 1/a_{2n} & ... & 1 \end{bmatrix}$$

and proposes the Eigenvector of this matrix as an estimation of the priority vector $w = (w_1, w_2, ..., w_n)^T$.

With the exception of the traditional Eigenvector prioritization method, all other methods for deriving priorities in the AHP are based on some optimization approach. The optimal prioritization methods, as the Goal programming, the Direct Least Squares, the Logarithmic Least Squares and the Fuzzy Preference Programming introduce an objective function, which measures the degree of approximation or the distance between the initial judgments and the solution ratios [2]. Thus, the problem of priority derivation is formulated as an optimization task of minimizing the objective function, subject to normalization and some additional constraints.

Despite the multicriteria nature of the requirements, regarding the properties of their solutions, all optimal prioritization methods optimize a *single* objective function. However, a single objective function cannot encompass and satisfy all requirements about the quality of solutions.

A new two-objective prioritization (TOP) method was proposed recently by the author [3], where the prioritisation problem is formulated as an optimization task for minimization of the Euclidean norm and the number of rank violations. The TOP method derives Pareto optimal solutions, which requires the application of efficient computational algorithms.

The paper investigates the application of evolutionary computing for solving the TOP problem. In order to eliminate the drawbacks of the numerical methods, we propose two evolutionary computing approaches. In the first one, the TOP problem is transformed into a single-objective one, which is then solved by a standard single-objective EA. The second approach applies a multi-objective EA for solving the TOP problem without such transformation.

In order to compare the solution approaches, we perform Monte-Carlo simulation experiments, by randomly generating a large number of pairwise comparison matrices. The paper presents some initial results from this simulation. Both computational approaches are also illustrated also by a numerical example.

The paper is organized as follows: Section II formulates the TOP problem; Section III discusses computational approaches to solving the problem; Sections IV and V provide some initial results from the simulation experiments, and Section VI concludes the paper.

II. THE TWO OBJECTIVE PRIORITISATION PROBLEM

Let $S = \{a_{ij} | j > i\}$ be a set of pairwise comparison judgments. The feasible set Q is defined as the set of all priority vectors $w = (w_1, ..., w_n)^T$, which satisfy the normalization and non-negativity constraints:

$$Q = \left\{ (w_1, ..., w_n) \middle| w_i > 0, \sum_{i=1}^{n} w_i = 1 \right\} \quad (1)$$

 ISBN: 978-1-61208-419-0

The accuracy of the each priority vector $w \in Q$, approximately satisfying the comparison judgments can be measured by the Total deviation criterion:

$$T(w) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (a_{ij} - \frac{w_i}{w_j})^2 . \quad (2)$$

This criterion is equivalent to the squared Euclidean distance for the upper triangular part of a Saaty's reciprocal comparison matrix.

The rank preservation properties of the solutions can be measured by the Number of Violations criterion [2]:

$$NV = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} v_{ij} \ ,. \quad (3)$$

where

$$v_{ij} = \begin{cases} 1, & if \ w_i > w_j \ and \ a_{ij} < 1, \ or \ \ w_i < w_j \ and \ a_{ij} > 1, \\ 1/2, & if \ w_i = w_j \ and \ a_{ij} \neq 1, \ or \ \ w_i \neq w_j \ and \ a_{ij} = 1, \\ 0, & otherwise. \end{cases}$$

The Number of Violations criterion (3) can be represented in the following compact form:

$$V(w) = \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| signum(a_{ij} - 1) - signum(\frac{w_j}{w_j} - 1) \right| ,$$

where the *signum* function is defined as:

$$signum(b) = \begin{cases} 1, & if \ b > 1 \\ 0, & if \ b = 0 \\ -1, & if \ b < 1. \end{cases}$$

The TOP problem is to find a feasible priority vector that 'simultaneously' minimizes the Total deviation and the Number of violations:

$$Minimize \ (T(w), V(w)) \quad (4)$$
$$subject \ to \ w \in Q \ ,$$

where $T : R^n \to R^1$ and $V : R^n \to R^1$ are real-valued objective functions.

Each feasible vector $w \in Q$ determines a unique value of the objective function vector $y = (T(w), V(w))$. Therefore, the feasible set Q in the space of decision variables can be transformed into a *payoff set* Y in the two-dimensional objective space. The payoff set represents a feasible region of the admissible values of $T(w)$ and $V(w)$, and can be considered as the image of the feasible set Q in the objective space.

The payoff set Y of the TOP problem consists in parallel line segments, as the function $V(w)$ takes non-negative discrete values in some range, and the function $T(w)$ is bounded.

III. COMPUTATIONAL APPROACHES FOR SOLVING THE TOP PROBLEM

A. Multiobjective Numerical Algorithms

Some classical multi-objective optimization (MOO) methods, which can be applied for finding Pareto optimal solutions to the TOP problem are the Weighting Method, the ε-Constraint Method, the Goal Programming Method or the Proper Equality Constraints method [4]. Generally, the main strength of classical MOO methods is their efficiency and ability to generate strong Pareto optimal solutions. However, these methods have some weaknesses in generating the Pareto optimal solutions, when specific problem knowledge is not available. Additionally, they cannot generate all Pareto optimal solutions with non-convex surfaces. From computational point of view, many optimization runs are required to obtain an approximation set of the Pareto optimal solutions [5].

Recently, the evolutionary algorithms have become an alternative to the classical methods for generating Pareto optimal solutions; since they can eliminate some of the drawbacks of the classical MOO methods.

B. Single-Objective Evolutionary Algorithms

Taking into account the specific properties of the TOP problem (4), we can transform it into a single-objective optimisation problem, which is easily solved by standard single-objective EA.

By associating weights *k* and *(1-k)* to both objective functions in (4), we obtain

$$J(w) = kT(w) + (1-k)V(w) ,$$

which is used as a fitness function of a single-objective EA. The value of the weight coefficient *k* is given by the user. This value represents his/her preferences with respect to the relative importance of those two objectives.

C. Multi-Objective Evolutionary Algorithms

Some multi-objective EA, as the Vector Evaluated Genetic Algorithm (VEGA), the Non-dominated Sorting Genetic Algorithm (NSGA), the Niched Pareto GA (NPGA), the Multi-objective Genetic Algorithm (MOGA) [5] and the Pareto Envelope-based Selection Algorithm (PESA II) [6]. However, it is well known that the presence of constraints scientifically affects the performance of multi-objective EA. Additionally, as opposed to the single objective case, the ranking of a population in the multi-objective case is not unique [7].

In order to assess the applicability of EA for solving the TOP problem, we perform a series of computational experiments by Monte-Carlo simulation. In our study we use the PESA-II, which has some advantages compared to other EA. PESA-II follows the standard procedures of an EA, but with the difference that two populations of solutions are maintained: an internal population (IP) of fixed size, and an external population (EP) of non-fixed but limited size. The internal population's job is to explore new solutions, and it achieves this by the standard EA processes of reproduction

 ISBN: 978-1-61208-419-0

and variation (i.e., recombination and mutation). The purpose of the external population is to store and exploit good solutions; it does this by maintaining a large and diverse set of the non-dominated solutions discovered during search [8].

An important advantage of PESA-II is that its niching policy uses an adaptive range equalization and normalization of the objective function values. This means that difficult parameter tuning is avoided, and objective functions that have very different ranges can be readily used.

IV. MONTE-CARLO SIMULATION

Monte-Carlo simulation experiments have been carried out, consisting in generation of comparison matrices with different dimensions and applying the single-objective and multi-objective evolutionary approaches. Initially random consistent pairwise matrices are generated; then they are perturbed by a user-driven parameter, denoted as *p*, which determines the degree of inconsistency.

The matrices for this comparison study are of dimensions *n*=3, 4, 5, 6, 7, 8 and 9. For every value of *n*, the parameter *p* takes 9 values, *p*=10, 20,…, 90 and each combination of {*n*, *p*} is replicated 30 times, which gives a total number of 810 *n*-dimensional matrices with different degrees of inconsistency. The overall number of generated random matrices is 5670.

The single-objective EA (a standard Genetic Algorithm) and PESA-II have been applied for solving the TOP problem for each pairwise comparison matrix, using the jMetal toolkit [9]. In the single-objective EA each chromosome is represented by a string of n components, associated with the n-dimensional priority vector w. The EA performs the basic genetic operators, which are roulette wheel selection, crossover with random mating and simple mutation.

Elitism has also been applied as an additional selection strategy, to make sure that the best performing chromosome always survives. The elitism has been realized by comparing the fitness of chromosomes from the current population and the fitness of the corresponding offspring. The fittest chromosome from the initial population survives for the next generation.

At the beginning of each cycle, all chromosomes are normalised, so that the values of their genes sum up to one. The stopping condition is the number of generations, which is selected to be equal to 100. The experimental results show that the single-objective EA converges to the optimal solution for less than 50 generation cycles.

The high-level pseudocode, showing the main steps in the PESA-II algorithm, is given in [8].

V. NUMERICAL RESULTS

Consider a problem with 5 comparison elements [8], where the DM provides the following pairwise comparison matrix A:

$$A = \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array} \begin{array}{c} \begin{array}{ccccc} E_1 & E_2 & E_3 & E_4 & E_5 \end{array} \\ \begin{bmatrix} 1 & 5 & 1/3 & 7 & 2/3 \\ 1/5 & 1 & 2 & 1/2 & 4 \\ 3 & 1/2 & 1 & 1/2 & 1/4 \\ 1/7 & 2 & 2 & 1 & 1/3 \\ 3/2 & 1/4 & 4 & 3 & 1 \end{bmatrix} \end{array}$$

The TOP problem is formulated as a single-objective one by the Weighting method and solved by the single-objective EA. The parameters of the EA are selected as follows:

- Population size = 100;
- Crossover probability=0.9;
- Mutation probability=0.01.

In this example, we have 5 Pareto optimal solutions, which are obtained by the EA. Due to the non-transitivity of the pairwise comparison problem, there are no priority vectors with less than two violations.

The results are given in Table 1.

TABLE I. PARETO OPTIMAL SOLUTIONS OBTAINED BY A SINGLE-OBJECTIVE EVOLUTIONARY ALGORITHM

w_1	w_2	w_3	w_4	w_5	T	V
0.309	0.113	0.096	0.135	0.347	50.048	2
0.344	0.075	0.109	0.111	0.361	39.806	3
0.345	0.058	0.140	0.071	0.386	29.717	4
0.363	0.067	0.146	0.058	0.367	26.962	5
0.381	0.076	0.139	0.062	0.342	26.720	6

PESA-II was applied for solving the same problem. A 30-bit Gray code was used to represent each of the five weights, giving a 150-bit binary chromosome. Uniform crossover was applied with probability 0.2 and a bit-flip per-gene mutation rate of 0.01 was used. It was found that PESA-II is not sensitive to these parameters, and other values give similar performance.

The values of the PESA-II parameter settings are:

IPsize=10; EPsize=100; Generations=50; pm=0.01; Pc=0.2; #grid-cells (niches)=100; representation=30-bits per weight.

The best values of the priority vectors obtained from 20 runs of PESA-II are shown in Table 2.

TABLE II. PARETO OPTIMAL SOLUTIONS OBTAINED BY PESA II

w_1	w_2	w_3	w_4	w_5	T	V
0.356	0.096	0.096	0.096	0.356	39.469	2
0.362	0.078	0.099	0.100	0.362	38.631	3
0.358	0.065	0.149	0.065	0.363	27.438	4
0.361	0.074	0.144	0.060	0.361	26.622	5
0.398	0.083	0.158	0.065	0.296	26.479	6

By comparing the values of T for each value of V, it is seen that PESA-II solutions outperform those obtained by the single-objective EA, with respect to the accuracy.

Regarding the computational efficiency, the average processing time of the single-objective EA for this example

 ISBN: 978-1-61208-419-0

is 1325 milliseconds, while the PESA II algorithm requires 622 milliseconds to find the optimal solution.

The performance comparison was obtained using an Intel-based PC with a Core2Duo T5500 CPU running at 1.66GHz and 2GB of physical memory. The tests were executed on Windows 7 with Java NetBeans IDE running in parallel.

The preliminary results from the Monte-Carlo simulation show that the multi-objective EA gives better accuracy than the single-objective EA, especially for high-dimensional and rather inconsistent pairwise comparison matrices.

Regarding the computation time, both approaches have rather similar performance for pairwise comparison matrices of lower dimension, n=3 and n=4. When the size of the matrices increases, PESA–II strongly outperforms the single-objective EA. The single-objective EA is particularly slower in inconsistent and non-transitive problems with many Pareto optimal solutions.

VI. CONCLUSIONS

The paper investigates the application of evolutionary algorithms for solving the TOP problem and shows that they are very good alternatives to the numerical multi-objective optimization methods. Two evolutionary approaches are applied for obtaining Pareto optimal solutions to the problem.

The numerical example and the preliminary results from a Monte-Carlo simulation experiment show that the multi-objective EA outperforms the single-objective EA with respect to the computational efficiency and accuracy of solutions.

REFERENCES

[1] T. Saaty, “A calling method for priorities in hierarchical structures,” Journal of Mathematical Psychology, vol. 15, pp.234-281, 1977.

[2] B. Golany and M. Kress, “A multicriteria evaluation of methods for obtaining weights from ratio-scale matrices”, European Journal of Operational Research, vol. 69, pp.:210-220, 1993.

[3] L. Mikhailov, “Deriving priorities from ratio-scale comparison judgements: A multiple objective approach”, In Proceedings of the 17th International Conference on Multiple Criteria Decision Making MCDM’2004, August 2004, Whistler, Canada.

[4] Y. Sawaragi, H. Nakyama, and T. Tanino, Theory of multiobjective optimization. Academic Press, Orlando, 1985.

[5] K. Deb, Evolutionary algorithms for multi-criterion optimization in engineering design. John Wiley & Sons, Chichester, UK, 1999.

[6] D. Corne, N. Jerram, J. Knowles, and M. Oates, “PESA-II: Region-based Selection in Evolutionary Multiobjective Optimization”, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO’2001), 2001, pp. 283-290.

[7] C. Fonseca and P. Fleming, “Multiobjective optimization and multiple constraints handling with evolutionary algorithms, part 1: A unified formulation”, IEEE Transactions on SMC vol. 28, pp. 26-37, 1998.

[8] L. Mikhailov and J. Knowles, “Priority Elicitation in the AHP by a Pareto Envelope-based Selection Algorithm”. In Multiple Criteria Decision Making for Sustainable Energy and Transportation Systems. (M Ehrgott, B Naujoks, T Stewart, and J Wallenius eds.), Springer's Lecture Notes in Economics and Mathematical Systems, vol. 634, pp. 249-258, Springer-Verlag, London 2010.

[9] J. Durillo, A. Nebro, F. Luna, B. Dorronsoro, and E. Alba, “jMetal: A Java Framework for Developing Multi-Objective Optimization Metaheuristics”, Tech. Rep. ITI-2006-10, Departamento de Lenguajes y Ciencias de la Computación, University of Málaga, December 2006. Available from http://jmetal.sourceforge.net (last accessed May 2015).

 ISBN: 978-1-61208-419-0

Using Application Oriented Micro-Benchmarks to Characterize the Performance of Single-node Hardware Architectures

R. Berrendorf, J. P. Ecker, J. Razzaq and S. E. Scholl
Computer Science Department
Bonn-Rhein-Sieg University of Applied Sciences
Sankt Augustin, Germany
e-mail: {rudolf.berrendorf, jan.ecker,
javed.razzaq, simon.scholl}@h-brs.de

F. Mannuss
EXPEC Advanced Research Center
Saudi Arabian Oil Company
Dhahran, Saudi Arabia
e-mail: florian.mannuss@aramco.com

***Abstract*—In this paper, a set of micro-benchmarks is proposed to determine basic performance parameters of single-node mainstream hardware architectures for High Performance Computing. Performance parameters of recent processors, including those of accelerators, are determined. The investigated systems are Intel server processor architectures as well as the two accelerator lines Intel Xeon Phi and Nvidia graphic processors. Results show similarities for some parameters between all architectures, but significant differences for others.**

Keywords–Performance benchmarks; Intel processors; Intel Xeon Phi; Nvidia graphic processors

I. INTRODUCTION

For resource-intensive computations in High Performance Computing (HPC) on a single node level the performance characteristics of the processor, memory and core-core / core-memory interconnect architecture are important to understand, to achieve good performance. Often HPC applications stress for most of their run time only parts of the hardware in compute intensive program kernels. Examples are compute bound problems, such as direct linear solvers [1] that are bound by the floating point capability of a system or memory bandwidth bound problems like the multiplication of a sparse matrix with a dense vector in iterative solvers [2]. Other application kernels may be bound differently.

This paper proposes a set of micro-benchmarks to characterize HPC hardware on a single-node level. The results of the micro-benchmarks are performance parameters related to performance bounds found in many computational kernels (see [3] for two of such parameters). These parameters often allow to draw conclusions for the (at least relative) performance of real applications or performance critical application kernels of certain classes that are bound by one or few of those parameters. Additionally, if carefully chosen, bottlenecks of architectures can be revealed. The benchmarks were chosen to allow conclusions on an application level, rather than to evaluate deep structures in a processor architecture with sophisticated low-level programs.

The proposed micro-benchmarks are applied to representatives of different classes of current hardware architectures. Results show similarities in performance between all architectures for some parameters (e.g., reaching near peak floating point performance for matrix multiply), but also significant differences between architectures (e.g., main memory latency and bandwidth). Consequently only certain application classes are suitable for a specific architecture.

The paper is structured as follows. The following section discusses related work. Then, current mainstream HPC hardware architectures are briefly described, focusing on their differences. Section IV contains a description of the proposed micro-benchmarks. Section V describes our experimental setup and finally in Section VI, the evaluation results are discussed, followed by a conclusion.

II. RELATED WORK

Benchmarks are widely used to evaluate systems concerning certain performance properties. The result of a benchmark should be usable as an indicator that can support a decision, e.g., whether this system is feasible for a certain task or not. A multitude of different benchmarks exists, dependent on the question to be answered.

The Top500 list [4] uses the High Performance Linpack [1] to rank (very) large parallel systems. This benchmark produces just a single value, the FLOP-rate (Floating Point Operations) per second for just one specific task, the direct solution of a very large linear system.

The widely used SPEC CPU benchmark [5] is a mix of several real world application programs for integer dominant computations or floating point dominant applications. Running the benchmark on a system produces one factor for each class. These numbers express a *relative* performance improvement compared to an older system.

Williams et al. introduced the roofline model [3] to describe the expectable performance space in a resource bound problem. The two resources in this model, evaluated in a two-dimensional chart, are computational density (operations per transferred byte) and peak floating point performance. This is an example where two limitating parameters on a system are used to show eligible perfomance values.

The NAS Parallel Benchmarks [6] are more application oriented benchmarks. These benchmarks consist of larger compute intensive kernels and where originally designed to stress large parallel computers. Each of these applications represents a different computing aspect. The applicications include, e.g., Conjugate Gradient (irregular memory access), Multi-Grid (long- and short-distance communication), or fast Fourier Transform (all-to-all communication). These benchmarks have been, amongst others, implemented in OpenMP [7] and recently in OpenCL [8]. With the OpenCL extension they can be used to measure recent accelerators, such as GPUs.

For a finer granularity, benchmarks that give individual results for several operation classes can be used. An example is the OpenMP micro-benchmark suite [9] [10] that gives a

 ISBN: 978-1-61208-419-0

TABLE I. OVERVIEW OF THE MICRO-BENCHMARKS.

Benchmark	Category	Application
Memory latency	Memory access	Single-thread latency to main memory
Memory bandwidth	Memory access	Bandwidth to main memory
Atomic update	Synchronization	Multi-threaded atomic update of a shared scalar variable
Barrier	Synchronization	Barrier operation of n threads
Reduction	Synchronization	Parallel reduction of n values to a single value
Communication	Communication	Data transfer bandwidth to/from an accelerator through PCI Express
DGEMM	Computation	Parallel dense matrix multiply (compute bound)
SPMV	Computation	Sparse matrix multiplied with a dense vector (memory bound)

developer a measure of how well basic constructs of OpenMP [11] map to a given system. If a developer knows important parameters that mainly determine the overall performance of an application in this programming model, he is able to estimate how well his own application will perform on the system using these basic constructs.

Other micro-benchmark suites, which aim for a finer granularity are proposed in [12], [13] and [14]. These benchmark-suites are based on OpenCL. Here, OpenCL is used to compare memory related issues, as well as low level floating point operations and real life applications on different hardware architectures, including accelerators.

III. Current Hardware Architectures

This section gives a very brief overview on current HPC processor architectures and memory technology. The chapter is partitioned into sections on mainstream HPC processor architectures, HPC accelerator architectures and finally memory technologies.

A. Processor Architectures

We concentrate on the Intel Xeon EP line of HPC relevant processors, as these processors are used in nearly all new systems in the Top500 list [4] of HPC computers. Intel's recent micro-architectures are Sandy Bridge (SB), and it's successor Ivy Bridge (IB). The last change in the architecture appeared late 2014 in the Haswell processors (HW). A detailed description of the architectures is given in the related literature of the manufacturer [15].

Processors nowadays have several cores. In HPC clusters multiprocessor nodes with 2 processors are often used. Keeping multiple core-private caches coherent is usually done in the hardware by cache coherence protocols. Keeping caches coherent costs latency, bandwidth and may also influence an architecture's scalability.

B. Accelerator Architectures

Certain application classes can be accelerated using special attached processors. Nvidia graphic processors (GPU) and Intel Many Integrated Core processors (MIC) of the Xeon Phi family are predominant in HPC [4].

A Nvidia GPU has a hierarchical design (CUDA architecture [16]) that differs from common CPUs. The execution units (SE, Streaming Processors) are organized in multiprocessors, called Streaming Multi-Processors (SM or SMX), a GPU has several of such multiprocessors. For example, the Kepler GPU has up to 15 SMX and 192 SE per SMX resulting in a total of 2880 SE in the largest device configuration. These execution units are always used by a group of 32 threads called a warp. Such an architecture leads to several aspects that have to be respected in performance critical programs, e.g., coalesced memory access [17].

An Intel Xeon Phi coprocessor [18] is compromised of multiple CPU-like cores. The current generation Xeon Phi Knights Corner (KNC) has between 57 and 61 of such cores, which are connected via a bi-directional ring bus. To achieve good performance on a Xeon Phi the application must use parallelism as well as vectorization. In [19] requirements for vectorization are specified for the usage of the Intel compiler, e.g., no jumps and branches in a loop.

Recent accelerators (i.e., GPU as well as Xeon Phi) are plug-in cards connected to the host through a PCI Express (PCIe) adapter. This adapter is often a severe bottleneck, because the transfer rate through a PCIe connection is significantly lower (8 GB/s for PCIe 2.0 (x16) and 16 GB/s for PCIe 3.0 (x16)) than for example memory transfer rates in a host system.

C. Memory Technologies

On one side DDR3 / DDR4 RAM, which is used in CPU-based systems is mostly optimized for a short latency time. On the other side, GDDR5 memory used in accelerators is optimized for bandwidth. This is important, as the performance of accelerators mainly comes from Single Instruction Multiple Data (SIMD) parallelism, where the same instruction is applied concurrently to multiple data items. These data items have to be fed to the functional units in parallel, needing high main memory bandwidth.

All processors of discussion, including recent GPUs, use caches to speed up memory accesses. While GPUs currently have a 2 level cache hierarchy, CPUs use 3 levels of caches with increasing sizes and latencies. Caches are only useful if data accesses initiated by the program instructions obey spatial or temporal locality [20].

IV. Proposed Micro-Benchmarks

We propose a set of 8 micro-benchmarks to determine performance critical parameters in single-node parallel HPC systems. Each single benchmark tests one specific aspect of a hardware architecture or parallel runtime system on that hardware. These aspects are performance critical for certain application classes. Table I gives an overview of the proposed set. One or a combination of these parameters are usually the performance bounds of an application. In real-life application it is possible, that a combination of these parameters occur with different factors / weights. It is up to the developer to use his knowledge of the application to weight these factors correctly. Nevertheless, if the application is truly dominated by one of these parameters the developer has an indication whether an architecture would be suitable for this application.

The presented set of micro-benchmarks was implemented in C with OpenMP for the use with Intel Procesors (including KNC). However, the OpenMP implementation could also be used for further architectures, like Power 8, ARM, or AMD Processors. Moreover, widely used C compilers like the Intel

 ISBN: 978-1-61208-419-0

icc or the GNU gcc support this programming approach. Recently the GNU gcc added support for OpenMP 4.0 constructs, which makes it possible to address future Intel Xeon Phi processors. For the usage with Nvidia accelerators the commonly used CUDA programming approach was chosen, as this is the programming model delivering the best performance on these GPUs. Porting the CUDA implementation for example to OpenCL should be straightforward as both programming platforms have similar concepts.

A. Memory Performance

Memory accesses are often the main performance bottleneck in applications. An example for that is an iterative solver working on large sparse matrices [2] or graph processing [21]. Memory performance itself is mainly influenced by memory latency and memory bandwidth as the key performance parameters. An indicator for a latency bound application are many accesses to different small data items (that are not cached). An indicator for a bandwidth bound application is a program (kernel) with low computational density, i.e., the ratio of the number of operations performed on data compared to the number of bytes ,which need to be transferred for that data, is low.

1) Memory Read Latency: Read latency can be determined by single threaded pointer chasing, i.e., a repeated read operation of type `ptr = *ptr` with a properly setup pointer table. If all accessed addresses are within an address space of size *S* (without associativity collisions in the cache) and *S* is smaller than a cache size then all accesses can be stored in this cache.

2) Memory Bandwidth: To measure main memory bandwidth the Stream benchmark [22] is commonly used. We adapted this freely available benchmark for the Xeon Phi using the OpenMP `target` construct [11] and for graphic processors using CUDA programming constructs [23].

B. Synchronization Performance

Synchronization between execution units (threads, processes, etc.) at certain points during the program execution is necessary to ensure parallel program correctness. However, synchronization is often a very performance critical operation [24], because serialization, e.g., atomic updates, or overall agreement, e.g., barrier between the execution units, is necessary. Moreover, reduction operations are another important and performance critical type of synchronization in real life applications.

1) Atomic Updates: In our atomic update benchmark all participating threads perform an atomic increment operation on a single scalar shared integer variable in parallel. As a side note, this operation also modifies the variable. Consequently, the coherence protocol initiates a cache line invalidation / update in a cache coherent multi-cache based system. The atomic increment operation is repeated several times during the benchmark by each thread.

2) Barrier: In the barrier benchmark, a barrier operation is carried out repeatedly. For multiprocessors the benchmark uses an OpenMP barrier pragma inside a parallel region. For the Xeon Phi, this is surrounded by a `target` region. The CUDA execution model [23] does not support a barrier synchronization as such, because this would violate the basic concept of warp independence. In CUDA, a program with global steps is implemented using a sequence of multiple kernels. Therefore, the kernel launch time (with an empty kernel) with the following synchronization to wait for the kernel finalization is the closest adequate comparison to a barrier.

3) Reduction: In the reduction benchmark, a vector with *n* elements of type **`double`** is reduced to one **`double`** value summing up all vector elements. For a reduction partial sums must be summed up in a synchronized way, which is additional work compared to a sequential implementation and needs some serialization between parallel entities. The program for the multiprocessors uses the OpenMP reduction clause in a parallel for-loop. On multiprocessors systems the vector is initialized in parallel, so that parts of the vector are split over different Non-Uniform memory Access [20] (NUMA) nodes in a NUMA system. It should be pointed out that such a distribution is done internally by the operating system. As CUDA does not provide reduction operations itself, the open source (CUDA-based) Thrust library [25] of Nvidia is used for this benchmark on the GPU systems.

C. Communication Performance

In the communication benchmark, we measured the transfer rate of a certain amount of data between a host and an accelerator device over PCI Express. This measurement is carried out for both directions (to and from the accelerator).

D. Programming Kernels

For many scientific application fields linear algebra operations are building blocks and often belong to the most time consuming parts of a program. Dependend on the problem origin, dense or sparse matrices are used. The following two evaluation benchmarks cover both matrix types and also stress different parts of a system (these are both performance limiting for many applications also outside linear algebra).

1) Compute bound application kernel: For dense matrix multiply, with a high computational density, many techniques are known (and applied inside optimized library functions), which allow to run this operation near the peak floating point performance. Consequently, if done the right way, dense matrix multiply evaluates in essence the floating point performance of a core / processor / multiprocessor system. This operation is well examined and implemented efficiently in the BLAS library [26] and vendor optimized libraries, like the Intel MKL [27] and Nvidia cuBLAS [28].

2) Memory bound application kernel: On the other side, a sparse matrix multiplied with a dense vector (SPMV) stresses almost only the memory system, as it has a low computational density. The operation is available for multiple storage formats [2] and is, at least for larger matrices, memory bandwidth limited and *not* compute bound. SPMV is also available in the vendor optimized libraries Intel MKL [27] and Nvidia cuSPARSE [29], both with a small selection of supported storage formats. The CSR format [2] is a general format with good/reasonable performance characteristics for many sparse matrices on CPU based systems. The ELL format is, for appropriate matrices (a small and ideally constant number of non-zero elements per row), a favorable storage format on GPUs [30]. It should be pointed out, that in this benchmark we are not interested in the best possible performance for a specific matrix. We rather want to relate the performance of different systems for this type of operation in a more general way.

 ISBN: 978-1-61208-419-0

TABLE II. SELECTED HARDWARE PARAMETERS OF THE SYSTEMS USED.

Parameter	**Processor Systems**			**Accelerator Systems**			
Architecture	SB	IB	HW	KNC	M2050 (Fermi)	K20 (Kepler)	K80 ($2\times$ Kepler)
Clock [GHz] (with TurboBoost)	2.6 (3.3)	2.7 (3.5)	2.6 (3.6)	1.053	1.15	0.706	0.560 (0.875)
Peak double prec. perf.[1][GFlops]; 1 proc.	20.8	21.6	33.17	16.8	-	-	-
Peak double prec. perf.[1][GFlops]; all proc.	332.8	518.4	929	1010.8	515	1170	$2\times$ 935
Theor. memory bandwidth [GB/s][2]	102.4	119.4	136	320	148	208	2×240
Main memory size [GB]	128	256	128	8	3	5	2×12
Degree of parallelism [3]	32	48	56	240	448	2496	2×2496

[1] In relation to baseclock
[2] ECC off for accelerators
[3] Including hyperthreads

V. Experimental Setup

In this section, we specify our parallel system test environment where the benchmarks were applied. Additionally, we discuss parameter settings of the benchmarks, because performance can be a parameterized function, e.g., dependent on the number of used threads or data items.

A. Test Environment

The used systems include the last generations of Intel server processors and for accelerators the Intel Xeon Phi KNC as a many-core architecture, as well as three recent Nvidia GPU architectures. These include the most actual systems in each class. The tested accelerators use PCIe 2 (x16) for KNC, M2050 and K20and PCIe 3 (x16) for K80. The new Nvidia K80 consists of two Kepler GPUs, which work as two single devices and have to be programmed seperately. Only one of the GPUs was used to perform the benchmarks. Otherwise this would have to be viewed as a multi-GPU setup and would not be comparable to the other acceelerators. Table II summarizes key hardware parameters of the systems used. The CPU based systems are all 2-way multiprocessor systems.

B. Test Parameters

The benchmark tests were executed with the following parameter settings:

- *Memory latency*: Variable size of the pointer table with a single threaded run.
- *Memory bandwidth*: Fixed large vector size of `STREAM_ARRAY_SIZE=40000000` and a repeat factor of `NTIMES=1000`.
- *Atomic update*: Variable number of threads according to the systems used.
- *Barrier*: Variable number of threads according to the systems used.
- *Reduction*: Variable vector size with a full parallel run.
- *Communication*: Variable size of the transferred data.
- *DGEMM*: Variable matrix size with a full parallel run.
- *SPMV*: Fixed test matrix according to the SPE10 problem [31], SPMV implementation of MKL and cuSPARSE, CSR and/or ELL format (dependent on the library).

VI. Results

In this section, we discuss the main results and concentrate on the interesting aspects. When performance data is plotted as a function of the number of threads, it is meant as number of thread blocks for GPUs, because the usage model for graphic processors differs from a multiprocessor system. On GPUs usually all stream processors of such a processor are used (with even much more concurrency in the application to hide latencies) instead of specifying the exact number of threads.

Figure 1 shows the results for the memory latency, with an access stride of 256 byte, in absolute times. Figures 2 shows these results in cycles relative to the respective base CPU/GPU clock. For all systems, levels of the same latency (induced by cache sizes of the different cache levels) and the huge difference to a main memory access (the last step to the right) are clearly visible. If only absolute times are considered, as expected, one can see that all accelarators have higher latency than the processor architectures and that the GPU based Nvidia accelerators are slower than a CPU based KNC. Moreover, there seems to be hardly any improvement between GPU generations. But, if relative latencies are considered, one can see that the GPUs improve over the generations quite significantly, as the base clock is much lower. Related to relative cycles, the newest K80 outperforms the KNC and gets even close to the CPUs in access to global/main memory. So, read latency seems to be limited by the lower base clock on the K80. Looking at the different cache levels, the measurements on the M2050 and K80 GPUs show three different levels in access time which can be explained by the L1/L2 caches and accesses to the main memory. On the K20 only two levels of similar access times are visible. This is induced by different versions of the the Kepler architecture. The K20 does not cache global memory accesses in the L1 cache, but this is done in the newer generation K80. On the CPU based systems one can see the smaller L1 and L2 caches, then the larger L3 cache and at last seen in a fourth step the access to the main memory. Access to the L1, L2, L3 caches is very fast, for L1 and L2 even on KNC. Altogether the processor systems still outperform the accelerators in latency time, although newer accelerator generations have improved. Therefore applications that are already latency bound have a severe problem on accelerator systems if they cannot hide this latency, e.g. by allowing multiple read requests to be open at the same time.

The memory bandwidth performance is shown in Figure 3. For the processor systems the default thread scheduling was used here, with variable numbers of threads. For graphic processors the usage model is different to a multiprocessor system, because usually all stream processors of such a processor are used instead of specifying the exact number of threads. The performance number(s) for GPUs are therefore given as a

 ISBN: 978-1-61208-419-0

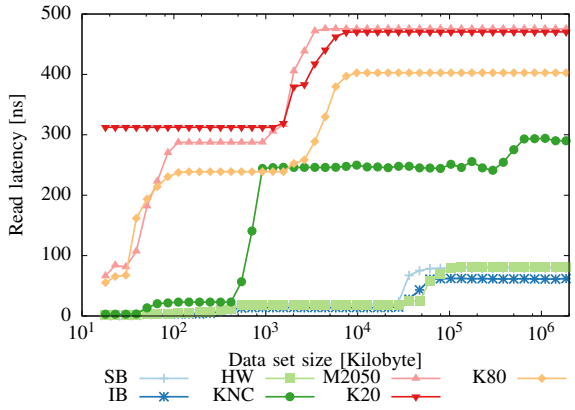


Figure 1. Memory latency results (absolute time).

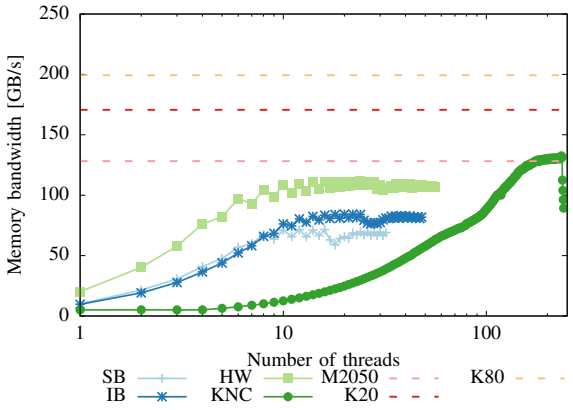


Figure 3. Memory bandwidth results.

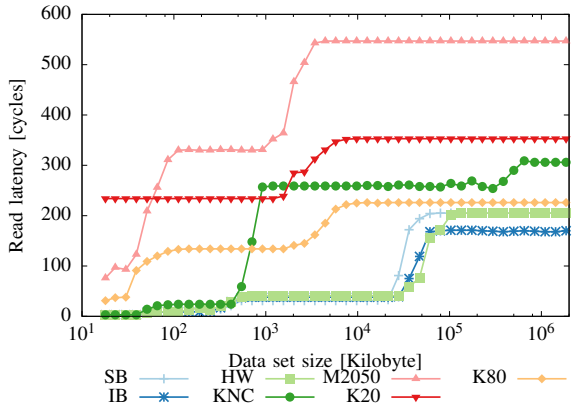


Figure 2. Memory latency results (relative cycles).

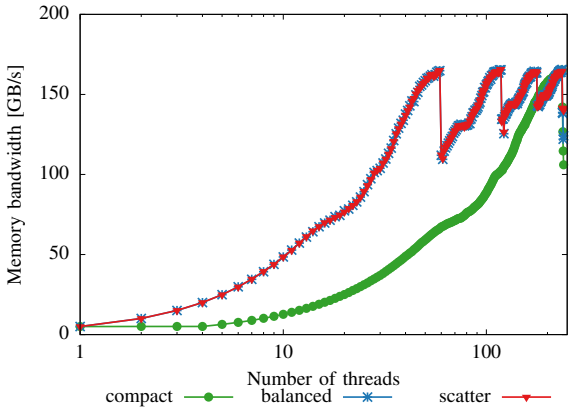


Figure 4. Memory bandwidth results on KNC, different thread mapping.

dashed line (for all stream processors used). In contrast to the results on latency, the accelerators perform better than the CPU systems. It is noteable that the KNC shows relativly bad performance here. Its bandwidth is comparable to the Haswell CPUs and the older Nvidia Fermi GPUs. Moreover the KNC is not able to reach its theoretical bandwidth at all, though it has the highest theoretical bandwidth of all tested systems. All processor systems almost reach their theoretical bandwith.

Thread mapping / binding can be an important aspect reaching good performance. A thread mapping defines how application threads are mapped to hardware threads, e.g., processors sockets, cores in a multi-core CPU, or hardware threads in a hyperthreaded core. Basic mapping strategies are to keep threads as close as possible in the hardware (compact; e.g., to exploit data locality between threads) or to spread as wide as possible (scattered; e.g., to exploit as much memory bandwidth as possible). Figure 4 shows the bandwidth test for the KNC with different thread mapping in OpenMP. A significant difference can be observed when different thread mappings are used. If the compact thread mapping is used (same as in Figure 4) bandwith increases steadily with increasing number of threads. The performance drop with the last four threads occurs, because at this point the last core with its four hardware threads is used in the application, but that performs a busy waiting for operating system tasks (communication with the host system). When scattered thread mapping is used, the impact of the four hardware threads can be seen. The performance increases until all cores are evenly utilized (one thread per core), then as soon as one core gets a second thread the performance drops and increases again steadily. Moreover, again the impact of the operating system core can be seen when all available threads of the KNC are used. Different to the KNC, changing the thread mapping for processor systems showed no difference at all for the benchmark.

Figure 5 shows the performance results for the atomic operation on the different systems. On the multiprocessors systems and KNC time increases linearly, proportional to the number of (competing) threads in use. Because the performance numbers show the normalized time for *one* operation, there is an increase in time *per operation* with the number of threads. This can be explained with the coherence and synchronization protocol, which is run by the processors / cores to ensure coherence and atomicity of such an operation. With more threads involved, the overhead increases [21]. For all three GPU systems the time is constant, which can be explained by the use of the unified L2 cache and the weak memory model without memory coherence. Moreover the performance improvement for atomic operations from Fermi (M2050) to Kepler (K20, K80) is clearly visible in this figure. For the KNC with compact thread mapping quite large fluctuations can be observed (note the logscale of the plot).

 ISBN: 978-1-61208-419-0

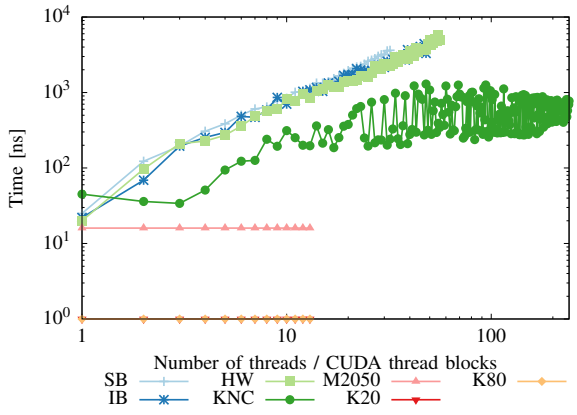


Figure 5. Atomic update results.

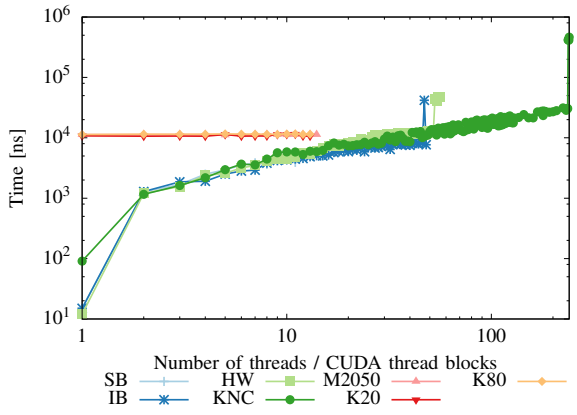


Figure 7. Barrier results.

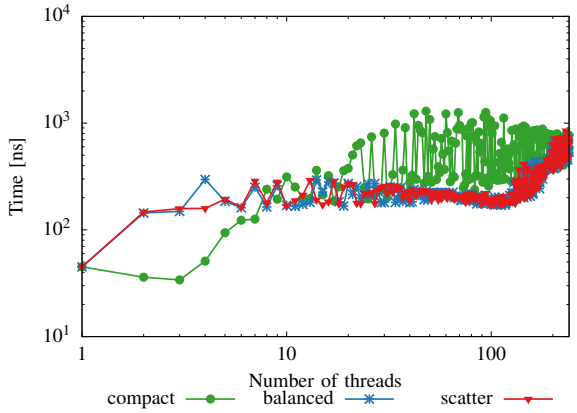


Figure 6. Atomic update results on KNC, different thread mapping.

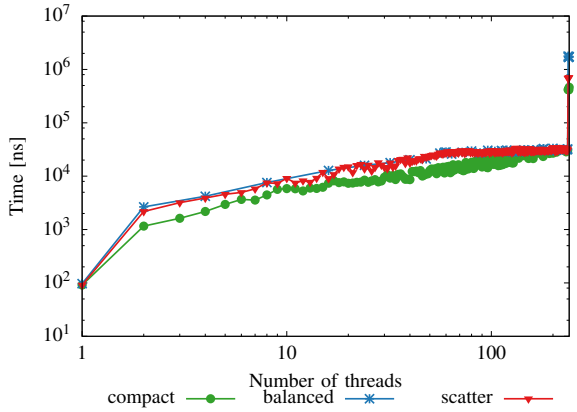


Figure 8. Barrier results on KNC, different thread mapping.

Figure 6 shows the atomic benchmark for KNC with different thread mappings. When scattered and balanced thread mapping is used the fluctuations become smaller, but still the changes of the performance with increasing number of thread is quite unsteady. An explanation for this could be the ring bus of the KNC. The cache of *all* cores in the KNC have to be kept coherent via this ringbus. Moreover, from the time when one core is popuated with all four hardware threads, time for an atomic update increases significantly. Again, changing the thread mapping for processor systems showed no difference at all for the benchmark.

Figure 7 illustrates performance results of the barrier test. The barrier synchronization on the Xeon Phi shows a similar behavior as on the multiprocessor systems. For the Nvidia accelerators the number of threads in the figure represents the number of used thread blocks (with 1024 threads per block). The figure shows that the kernel launch time is nearly constant and equal for M2050, K20 and K80. Further it does not depend on the number of blocks. Moreover, changing the thread mapping showed no significant differences on the multiprocessor systems. The impact of different thread mapping for the KNC can be seen in Figure 8. For balanced and scatter thread mapping, time increases steadily until all cores are populated with one thread. Afterwards, when more than one thread runs on a core, the releveance on the barrier execution time becomes less important. For compact thread mapping, the number of used cores increases steadily with the number of threads, so the time increases steadily, too.

For the reduction test, Figure 9 shows the parallel run time using all available parallelism on the system with increasing vector size. The M2050 card was limited by the available memory size and the largest vector size could not be used on this system. The GPUs are slower than the multiprocessors for a smaller number of elements and they are faster than the multiprocessors for large vectors which follows the usage model of GPUs.

Figure 10 shows data transfer rates in GB/s from the host to the attached accelerator and vice versa. The results show that, for reasonable large data sizes the bus is used efficiently on all systems (the theoretical data transfer rate of 8 GB/s for PCIe 2.0 and 16 GB/s for PCIe 3.0 minus protocol overhead). For the transfer back from the accelerator to the host there is a performance drop on the M2050 reaching only approx. 5 GB/s instead of nearly 7 GB/s on the other accelerators. The low bandwidth seems to be a problem with our combination of host system and accelerator card. The data transfer rates of the KNC are low, starting at smaller data sizes (e.g. below 2 GB/s for 80,000 bytes). The difference between host to device and device to host bandwidth could be due to the Direct Memory Access (DMA) initiator. When the DMA is initiated from the CPU it has better performance. Moreover the K80 does not reach the theoretical limit for the PCIe 3. This could perhaps

 ISBN: 978-1-61208-419-0

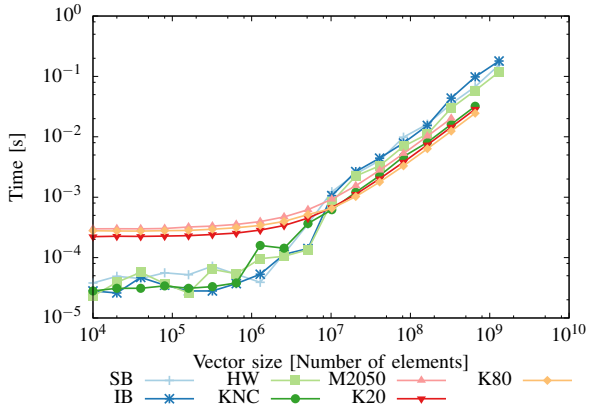


Figure 9. Reduction results.

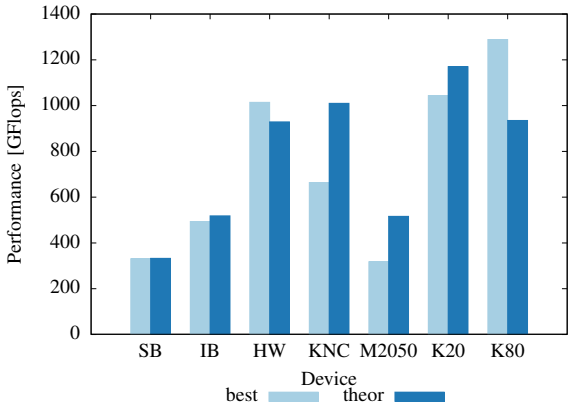


Figure 11. Dense Matrix Multiply results (DGEMM).

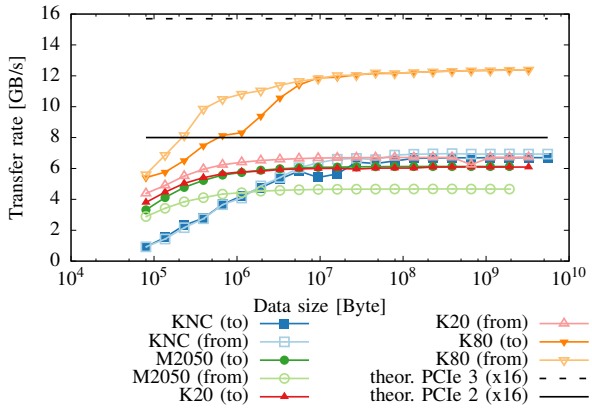


Figure 10. Communication performance results.

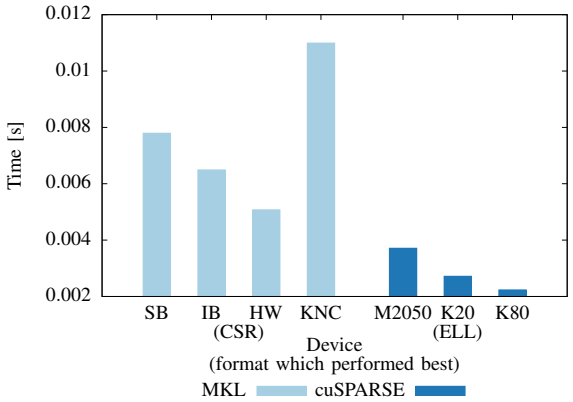


Figure 12. Sparse Matrix Vector Multiply results (SPMV).

be different when both GPUs are utilized.

In Figure 11, performance results for the dense matrix multiply operation are shown. Here, only the best performance, over all matrix sizes, is given. As expected the operation has better performance on the accelerators due to their better raw floating point performance. It can be seen that on the majority of the processor systems almost peak performance is reached. The Haswell processor even shows better performance than the given theoretical peak performance in Table II (related to the base clock). This can be explained by the intelligent turbo boost and temporal overclocking of these processors. Moreover the Haswell processors are the first processors which reach (nearly) one Teraflop performance. That makes it comparable to even recent accelerators. Haswell outperforms the older Fermi architechture and the KNC, which does not reach its theoretical performance at all. The Haswell results for matrix multiply are on nearly the same level as the recent Kepler architectures and get only beaten by the new K80. The K80 as well shows better performance than its given theoretical peak performance. Again, this can be explained by the use of turboboost.

For the SPMV operation shown in Figure 12 only the best results for a system and a format are given. All GPU systems perform well, compared to the multiprocessor and Xeon Phi systems. The K20 performs around 26% faster than the M2050 using ELL format. The low performance improvement of the K20 can be explained by the fact that the SPMV operation is memory bound and the memory bandwidth of the K20 is only around 25% higher compared to the M2050. Similar relations apply for K20 and K80. Surprisingly the KNC shows the weakest performance in this test although this card has the highest nominal memory bandwidth of all used systems. The reason for that was shown in the bandwidth results where the KNC did only reach half of its peak memory bandwidth.

VII. Conclusions

This paper introduced a set of benchmarks to determine important performance parameters of single-node parallel systems. One or a combination of these parameters are often performance limitating in parallel applications.

The benchmarks were applied to systems of the same basic architecture but different processor generations (Intel Haswell / Ivy Brige / Sandy Bridge) as well as to different architectures (CPU, two different accelerator architectures).

It was shown that some parameters (e.g., the memory related ones) show fairly different performance characteristics between the systems qualifying or disqualifying a system for certain application classes. In contrast, all systems showed a rather similar behavior for compute-dense problems reaching near-peak floating point performance that is quite comparable between accelerators and latest generation processors. Due to design decisions in the processor architecture graphic proces-

 ISBN: 978-1-61208-419-0

sors show a remarkable performance on some synchronization operations, operations that often limit the parallel performance.

In this paper we discussed only single-node parameters. An extension of this work would be to include cluster architectures, i.e., multiple-node architectures. Another extension could be to include multi-accelerator architechtures,e.g., using both GPUs of the K80. Further investigation would include the impact of different programming models such as OpenACC or OpenCL instead of CUDA on a GPU.

ACKNOWLEDGEMENTS

We would like to thank the CMT team at Saudi Aramco EXPEC ARC for their support and input. Especially we want to thank Ali H. Dogru for making this research project possible.

REFERENCES

[1] A. Petitet, R. Whaley, J. Dongarra, and A. Cleary, "HPL - a portable implementation of the high-performance Linpack benchmark for distributed-memory computers," http://www.netlib.org/benchmark/hpl/, Tech. Rep., 2008, version 2.0, [retrieved: Jun, 2015].

[2] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd ed. SIAM, 2003.

[3] S. Williams, A. Waterman, and D. Patterson, "Roofline: An insightful visual performance model for multicore architectures," Comm. ACM, vol. 52, no. 4, Apr. 2009, pp. 65–76.

[4] Top 500 List, http://www.top500.org/, [retrieved: Jun, 2015].

[5] SPEC CPU 2006, Standard Performance Evaluation Corporation, https://www.spec.org/cpu2006/, [retrieved: Jun, 2015].

[6] D. Bailey, E. Barszcz, J. Barton, D. Browning, R. Carter, L. Dagum, R. Fatoohi, S. Fineberg, P. Frederickson, T. Lasinski, R. Schreiber, H. Simon, V. Venkatakrishnan, and S. Weeratunga, "The NAS parallel benchmarks," NASA Ames Research Center, http://www.nas.nasa.gov/assets/pdf/techreports/1994/rnr-94-007.pdf, Tech. Rep., 1994, [retrieved: Jun, 2015].

[7] H. Jin, M. Frumkin, and J. Yan, "The openmp implementation of NAS parallel benchmarks and its performance," NASA Ames Research Center, http://www.nas.nasa.gov/assets/pdf/techreports/1999/nas-99-011.pdf, Tech. Rep., 1999, [retrieved: Jun, 2015].

[8] S. Seo, G. Jo, and J. Lee, "Performance characterization of the NAS parallel benchmarks in OpenCL," in The 2011 IEEE International Symposium on Workload Characterization (IISWC). IEEE, 2011, pp. 137–148.

[9] J. Bull and D. O'Neill, "A microbenchmark suite for OpenMP 2.0," SIGARCH Comput. Archit. News, vol. 29, no. 5, 2001, pp. 41–48.

[10] J. Bull, F. Reid, and N. McDonnell, "A microbenchmark suite for OpenMP tasks," in Proc. 8th Intl. Conference on OpenMP in a Heterogeneous World (IWOMP'12), 2012, pp. 271–274.

[11] OpenMP Application Program Interface, 4th ed., OpenMP Architecture Review Board, http://www.openmp.org/, Jul. 2013, [retrieved: Jun, 2015].

[12] P. Thoman, K. Kofler, H. Studtand, J. Thomson, and T. Fahringer, "Automatic OpenCL device characterization: Guiding optimized kernel design," in Euro-Par 2011. Springer-Verlag, 2011, pp. 438–452.

[13] A. Danalis, G. Marin, C. McCurdy, J. S. Meredith, P. C. Roth, K. Spafford, V. Tipparaju, and J. S. Vetter, "The scalable heterogeneous computing (shoc) benchmark suite," in Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units. ACM, 2010, pp. 63–74.

[14] X. Yan, X. Shi, and Q. Sun, "An opencl micro-benchmark suite for gpus and cpus," in 2012 13th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT). IEEE, 2012, pp. 53–58.

[15] Intel® 64 and IA-32 Architectures Optimization Reference Manual, Intel, http://www.intel.com/content/www/us/en/architecture-and-technology/64-ia-32-architectures-optimization-manual.html, Sep. 2014, [retrieved: Jun, 2015].

[16] Nvidia CUDA, https://developer.nvidia.com/cuda-zone, [retrieved: Jun, 2015].

[17] M. Wolfe, Understanding the CUDA Data Parallel Threading Model. A Primer, pgiinsider ed., PGI, https://www.pgroup.com/lit/articles/insider/v2n1a5.htm, Feb. 2010, (Updated December 2012), [retrieved: Jun, 2015].

[18] J. Jeffers and J. Reinders, Intel® Xeon Phi™Coprocessor High-Performance Programming. Morgan Kaufmann, 2013.

[19] M. Corden, Requirements for Vectorizable Loops, Intel, https://software.intel.com/en-us/articles/requirements-for-vectorizable-loops/, 2012, [retrieved: Jun, 2015].

[20] J. L. Hennessy and D. A. Patterson, Computer Architecture: A Quantitative Approach, 5th ed. Morgan Kaufmann Publishers, Inc., 2012.

[21] R. Berrendorf and M. Makulla, "Level-synchronous parallel breadth-first search algorithms for multicore- and multiprocessors systems," in Proc. Sixth Intl. Conference on Future Computational Technologies and Applications (FUTURE COMPUTING 2014), 2014, pp. 26–31.

[22] J. D. McCalpin, "Stream: Sustainable memory bandwidth in high performance computers," University of Virginia, Charlottesville, Virginia, Tech. Rep. TM-88, 1991-2007, [retrieved: Jun, 2015]. [Online]. Available: http://www.cs.virginia.edu/stream/

[23] Nvidia, CUDA C Programming Guide, pg-02829-001_v6.5 ed., http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf, Aug. 2014, [retrieved: Jun, 2015].

[24] R. Berrendorf, "A technique to avoid atomic operations on large shared memory parallel systems," Intl. Journal on Advances in Software, vol. 7, no. 7&8, 2014, pp. 197–210.

[25] Nvidia, Thrust, https://developer.nvidia.com/thrust, [retrieved: Jun, 2015].

[26] BLAS (Basic Linear Algebra Subprograms), http://www.netlib.org/blas/, [retrieved: Jun, 2015].

[27] Intel® Math Kernel Library, https://software.intel.com/en-us/intel-mkl, [retrieved: Jun, 2015].

[28] Nvidia cuBLAS, https://developer.nvidia.com/cublas, [retrieved: Jun, 2015].

[29] Nvidia cuSPARSE, https://developer.nvidia.com/cusparse, [retrieved: Jun, 2015].

[30] N. Bell and M. Garland, "Efficient sparse matrix-vector multiplication on CUDA," Nvidia Corp., Tech. Rep. NVR-2008-004, Dec. 2008.

[31] SPE Comparative Solution Project, Society of Petroleum Engineers, http://www.spe.org/web/csp/, [retrieved: Jun, 2015].

 ISBN: 978-1-61208-419-0

An Automatic Code Generator for Parallel Evolutionary Algorithms: Achieving Speedup and Reducing the Programming Efforts

Omar A. C. Cortes
and Eveline de Jesus V. Sá

Instituto Federal do Maranhão
Informatics Department
São Luis, MA, Brazil
Email: {omar,eveline}@ifma.edu.br

Jackson A. da Silva

Computer Engineering Department
Universidade Estadual do Maranhão
São Luis, MA, Brazil
Email: jackson.amarals@gmail.com

Andrew Rau-Chaplin

Dalhousie University
Faculty of Computer Science
Halifax, NS, Canada
Email: arc@cs.dal.ca

***Abstract*—Building parallel applications is not a trivial task, especially when these applications involve different kinds of evolutionary computation because two different fields have to be mastered. In order to overcome this problem, we propose an automatic code generator that automatically creates Java code for parallel evolutionary algorithms considering four models of parallelism: master-slave, island, cellular, and hierarchical. Furthermore, two evolutionary algorithms can be created: genetic algorithms and evolution strategies. A speedup experiment on a parallel genetic algorithm showed that good performance can be achieved using our generator. Moreover, we applied COCOMO and COCOMO II models in order to demonstrate that the cost for programming this kind of application can be considerably reduced in terms of effort, time and people when the generator is used. According to COCOMO II model, the generator may save about 6 months using 2 people.**

Keywords–Parallel Computing; Evolutionary Computation; Programming Effort; Code Generator.

I. INTRODUCTION

The popularity of multicore computers has increased the importance of building parallel applications. In fact, nowadays even cell phones take the benefits from parallel computing using multi-core architectures. Despite so, the parallel computing creates new challenges to programmers such as synchronization and the proper exploration of parallel algorithms. In addition, the lack of experience in developing parallel applications might impact directly in the software productivity, increasing significantly the effort on programming this kind of software.

One of many fields that can obtain advantages from parallel computing is the evolutionary computation. Doing so, we can combine the high performance environment provided by parallel architectures with the ability of solving complex problems using Evolutionary Algorithms (EA). Actually, making EAs faster by using parallel implementations has been one of the most promising choices in the area [1].

Evolutionary algorithms are search algorithms based on natural evolution [2]. They can be parallelized by using different models resulting in Parallel Evolutionary Algorithms (PEA). Even though other classifications might exist, they have been separated in four main categories: master-slave, island, cellular and hierarchical. In master-slave, there is a single population and the evaluation of fitness is distributed through the slaves. An island EA consists of two or more independent populations with occasional migration of individuals between sub-populations. In the cellular model, a predefined structure is held, such as a grid in a 2D case, then genetic operators are limited to a small neighborhood. Finally, the hierarchical model, also known as hybrid model, is a mix between island and either master-slave or cellular.

As we can see, each model introduces new complexities in developing a PEA, consequently increasing the effort of programming a parallel system based on EA. In this context, we built a tool called Java Parallel Evolutionary Algorithm Generator (JPEAG) which can generate both Parallel Genetic Algorithms (PGAs) and Parallel Evolution Strategies (PESs) in the four parallel models aforementioned. Moreover, from the user perspective, we show how much effort this tool can save. In order to do so, we applied two popular algorithm approaches named Constructive Cost Model (COCOMO) and COCOMO II [3], calculating metrics such as effort (people/month), time (months) and people.

To best of our knowledge, there are no other works which generate code for PAEs in a complete automatic way. There are some efforts toward creating parallel code such as the following authors: Passini [4], Hawick [5] and Rodrigues [6]; however, these works are based on general aspects of parallelization (synchronization and communication) where the programmer has to implement both his application and the PEA manually. Further, they do not show metrics for quantifying how much effort is saved. Also, our proposal is different from a framework because JPEAG can provide the programmers a entire runnable code. In terms of metric for measuring programming effort, COCOMO and COCOMO II have been used in many works such as [7], [8] and [9], for example. Furthermore, COCOMO II has been the bottom line for many researches whose aim is to improve the quality of the effort measurements such as [10][11][12].

For this sake, our paper is divided as follows: Section II introduces basic concepts on evolutionary algorithms, how JPEAG was built, how templates are used to produce the code, and how design patterns are presented in the generated code; Section III shows how COCOMO and COCOMO II compute and assess effort, time and people; Section IV presents the results considering the development of all PEAs; finally, Sec-

 ISBN: 978-1-61208-419-0

tion V draws the conclusions of this work.

II. The Code Generator

A. Evolutionary Algorithms

Basically, an evolutionary algorithm processes a population of individuals or solutions in a particular search space. All movements along the search space are done through genetic operators on either many iterations or until reaching some other stop criteria such as: there is no more evolution within the population; the algorithm reaches the known optimal; or, the solution is sufficiently close to the optimum considering an small error (commonly the error is 1×10^{-6}). Figure 1 shows a basic structure of an EA.

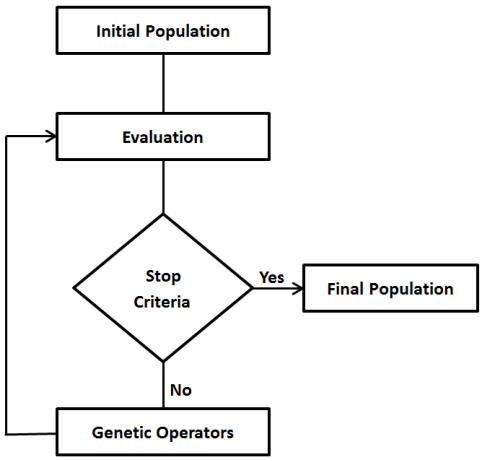


Figure 1. Basic structure of a EA

In the first step, the population is initialized at random normally using a uniform distribution. Then the population is evaluated in order to determine how each individual fits to the problem. The better the fitness, the stronger the individual within the population, thus might be higher the probability of an individual to be selected to undergo a genetic operators and, consequently, to go to the next iteration (generation). In fact, who goes to the next iteration depends on which kind of EA is being run. When the stop criteria is reached the final population containing the best solution is presented.

Typically, the genetic operators are selection, crossover and mutation. Selection is the process of choosing individuals to undergo a genetic operators or to go to next generation. Parents exchange information (genes) between themselves in order to create one or more offspring, in the crossover operator. Ideally, when two strong individuals exchange genes, in theory, the offspring is stronger than its parents [13], thus, strong individuals tend to spread its genes to next generations. On the other hand, this behavior can lead to a premature convergence of the solution because the population can end up trapped into a local optima. The mutation operator has the purpose of avoiding the premature convergence applying some modifications to one or more genes. In other words, the process of using genetic operators tends to improve the solution's quality as new generations carry on [14]. Taking those operators (crossover and mutation) into account, we can notice that several EAs share similar features. For instance, genetic algorithms and evolutionary strategies may use all those genetic operations. However, the sequence that these operators are used is different. Evolutionary strategy applies firstly the genetic operators then it uses the selection operator, whereas genetic algorithms select the individuals and afterward perform genetic operators. In addition, the individual representation can be different between EAs. For instance, evolutionary strategies need two vectors for representing an individual instead of only one of genetic algorithms. Details about these EAs can be seen in Herrara [13], Michalewicz [14] and Cortes [15].

B. Parallel Evolutionary Algorithms (PEA)

The main idea behind the parallel computing is to divide a problem into smaller pieces and solve them using different processing units. In this particular case, EAs are good candidates to be parallelized because they have an intrinsic parallelism [16], which can be explored in many different ways such as: to find out distinct solutions for the same problem, to explore several points in the search space at the same time, to distribute the evaluation function between two or more processors/threads, and to reduce the computation time to get a solution [26].

Regardless what it is being parallelized into an EAs, as previously mentioned, there are four basic models to explore the parallelism of EAs: master-slave, cellular, island and hierarchical. Different names might be used for these models; however, their characteristics remain the same.

Concerning the master-slave model, a PEA maintains a population in a master processor that delegates the function evaluation or the applications of genetic operators to slaves. Commonly, the parallelization is done distributing only the evaluation function among the available slaves. In the cellular model, all processors work on the same population, where each individual is placed into a grid, thereby genetic operators can be done only with their neighborhood in the grid. Independent populations are processed at the same time in the island model, introducing the concept of migration, where one or more individuals can be frequently exchanged between populations. This model also introduces new parameters such as the number of individuals being exchanged, how frequent the migration has to be done and the island topology which represents how islands can communicate each other. Researches such as Cantú-Paz [1] and Sakuray [18] indicate that the proper migration process contributes in the population diversity and enhance the quality of solutions. The cellular multi-individual is a hierarchical approach where each cell can contain two or more individuals, therefore being a combination between both cellular and island model.

In any parallel model, all communication between processor can be either synchronous or asynchronous. In the first one, if a processor wants to communicate with one or more processors, it has to wait until all of them be ready. On the other hand, in the asynchronous communication, the execution and the communication do not depend on the other processors, *i.e.*, if a processor wants to communicate with another one it sends the information and continues with its own execution. In our implementation of JPEAG, all communications are synchronous.

C. The Code Generator

Code generation can be defined as the technique in which we write programs that create another programs. According to

 ISBN: 978-1-61208-419-0

Herrington [19], creating code presents many benefits such as:

- Agile software development code generators go toward completion faster than a hand-made code, reducing the cost of development as well.
- Consistency code generators might maintain the standard in both design patterns and code conventions, avoiding breaches introduced by programmers.
- One point for gathering knowledge a change made in a definition file can be propagated to all files created previously, whereas programmers have to do so file-by-file in a hand-made process.

There are some strategies in order to implement a code generator. We are focused on templates which represent a predefined piece of software, *i.e.*, an unfinished code that may be completed using variables [20]. In other words, a software replaces some elements presented in a template file [21], where the substitution has to be performed by a template processor considering a set of inputs.

The approach based on templates was chosen because a significant amount of code for GAs and ESs are similar. Moreover, we also noticed some similar characteristics in parallel models, especially in the island and master-slave models. Thus, when these similarities were identified we could write the templates containing the common parts. In this context, the template approach allows to modify any part of the generated code changing only the proper template, therefore, becoming easier the software maintainability.

Figure 2 shows how components communicate each other in JPEAG. Its operation is described as follows: the interface is a web-based interface where users can configure all features of the PEAs, including the evaluation function (the programmer has to provide the evaluation function in terms of Java code); when all parameters are set, the user sends it to the core, which is responsible for selecting the proper template in the templates data base and use it to create the parallel EA code. The use of templates permit to create PEAs in different languages without the necessity of changing the application code; then, the parallel code is packed into a zip file and sent it back to the user via download because is a web-based application.

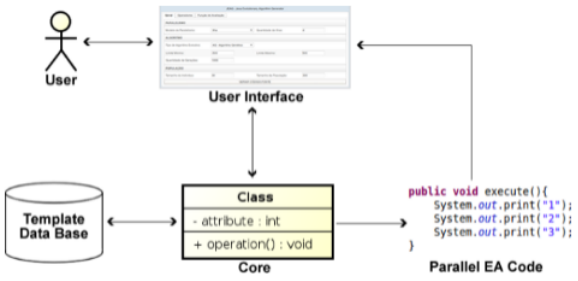


Figure 2. Code Generator Architecture

It is important to mention that the templates are processed by a framework called Apache Velocity [22], which implements an engine for template processing and defines a Velocity Template Language (VTL) for creating it. The main advantage of the Apache Velocity is to provide methods for processing templates and creating code for any textual language. Taking this into account, our templates are Java code mixed with VLT, where VLT instructions indicate where the JPEAG has to fill up the code according to the parameters previously defined in the graphic interface.

Figure 3 shows an example of a VLT code, where directives start with the character "#" and are executed when the template is processed.

```
#if ($parallelismModel == "Island")
     ${EAModel}${parallelismModel} ea = new
     ${EAModel}${parallelismModel}($islandNumber,
     $migrationRate, $migrationFrequency);
#end
```

Figure 3. Example of VLT code

The *#if* directive in the aforementioned example means that this part of code will be processed only if the user chose the island model. Variables begin with the character "$" and will be filled up according to the configuration done by the user in the graphic interface. As a result, the user will receive a pure Java code such as "$GAIslandea = newGAIsland(4, 5, 10);$", where the parameters in this particular example are the number of islands (4), the migration rate (5) and the migration frequency (10).

D. Design Patterns on PEA

The code received by the user via download as explained in the previous section, is built using design patterns that play an important role in the reuse of software because it tends to impact in programmer productivity, specially due to the similarity between operators in evolutionary algorithms. In our case, the code produced by the generator contains mainly two design patterns: strategy and observer.

The strategy pattern defines a group of algorithms encapsulating them by means of interfaces, allowing variations regardless it uses in the clients. This pattern is used, for instance, to hide the implementation of genetic operators and to assure that any changes in the operators do not affect other parts of the code [23]. Further, this pattern permits that genetic operators may be used in other kinds of EAs, for instance, in creating parallel hybrid algorithms. The stop criteria was implemented by using the strategy pattern in this work, as well.

The observer pattern defines a $1 - to - n$ relationship between objects. This relationship consists of one object being observed by many other ones with low coupling. When an observed object has its status modified all observers receive a notification being automatically updated [23]. This pattern is used for synchronization purposes between objects. Being specific, a process receives information about other processes, thus it can decide whether it have to wait for other processes, in case of migration for example, or carry on with its own execution.

Figure 4 presents a model with an example of the master-slave model for a parallel genetic algorithm, where we can see clearly how the observer interface is implemented by the *ParallelismMonitor* class, consequently controlling the communication between islands, while the strategy pattern interfaces are used to control the genetic operators and the stop criteria (classes which implement the pattern are omitted due to the lack of space).

 ISBN: 978-1-61208-419-0

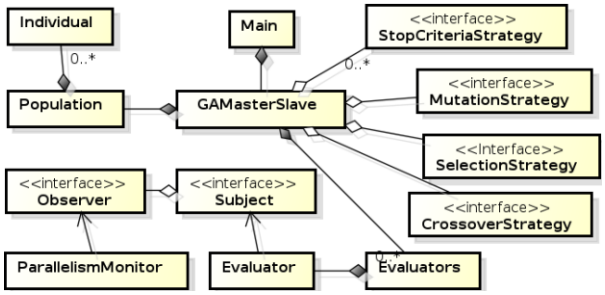


Figure 4. Master-Slave Patterns

III. Computing Programming Effort

A. COCOMO

COCOMO is a regression-based model used to estimate the programming effort in a software development project. The model can be divided into three sub-models: basic, intermediary and advanced. The basic one is presented in (1), where PM means Person/month (effort), A is a calibration factor, $KLOC$ represents the number of lines of code (in terms of 1000 lines or K) and B is a scale factor.

$$PM = A \times (KLOC)^B \tag{1}$$

$$time = C \times PM^D \tag{2}$$

$$p = \frac{PM}{time} \tag{3}$$

Also, the basic model can compute the required time for developing the software in months (2) and the number of required people (3) as well, where C and D are constants. The constants A, B, C, and D, which are originally from the model, can be seen in Table I, where $Project$ indicates the following features: (i) Organic is a project involving a small team with good experience and less than rigid requirements; (ii) Semi-detached is a project with a medium team with mixed experiences and a combination of rigid and less than rigid requirements; (iii) Embedded involves a set of tight constraints, being a combination of the organic and semi-detached project.

TABLE I. CONSTANTS FOR COCOMO - BASIC

Project	A	B	C	D
Organic	2.4	1.05	2.5	0.38
Semi-detached	3.0	1.12	2.5	0.35
Embedded	3.6	1.20	2.5	0.32

The intermediary model is similar to the basic one; however, it considers a multiplier effort (ME) as we can see in (4), where $n \in [1, 15]$ according to 15 different multipliers. In fact, ME is the product of all efforts that might be involved in the project.

$$PM = A \times (KLOC)^B \times \prod_{i=1}^{n}(ME_i) \tag{4}$$

Basically, what we do with the ME is multiply all of it according to the levels we have in the team. Which one we use depends on the requirement we have in the project. For example, a project might demand a high level of programming capability and a very high level of analyst capability. So, in this particular case the ME is computed as follows: $ME = 0.86 \times 0.71 = 0.6106$. All values in ME are predefined and can be seen in [24] and [25]. On the other hand, the constants A and B are different for the intermediary model as shown in Table II, while C and D remain the same from Table I.

TABLE II. CONSTANTS FOR COCOMO - INTERMEDIARY

Project	A	B
Organic	3.2	1.05
Semi-detached	3.0	1.12
Embedded	2.8	1.50

The advanced model is similar to the intermediary, however the calculation is done on each step of development, being adequate only for big projects.

B. COCOMO II

The main differences between COCOMO and COCOMO II are: (i) the first one uses 15 multiply efforts, whereas the last one uses 17 [27]; and, (ii) B can be computed based on a Scale Factor (SF) as presented in (5), where β is a constant equals to 0.91 as suggested in [24] and i varies from 1 to 5 according to predefined SFs. All these values can be obtained from COCOMO II manual and in [24].

$$B = \beta + 0.01 \times \sum_{i=1}^{5} SF_i \tag{5}$$

IV. Results

A. Speedup

In order to demonstrate that the software is usable and performs in parallel, we present an experiment based on the Griewank function [28], which is shown in (6), where x_i belongs to the range [-600,600] and n is equal 30, representing an individual with 30 real-coded genes.

$$f(x) = \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} cos(\frac{x_i}{\sqrt{i}}) \tag{6}$$

The experiment was conducted in an Intel i5 2.3 Ghz, 4GB of RAM with two physical cores and hyper threading (4 logical cores), using Ubuntu Linux, JDK 1.7.04. A Parallel Genetic Algorithm was built using the following configuration: selection operator is tournament; tournament size equals 10; heuristic crossover operator; probability of crossover equals to 0.8; uniform mutation; probability of mutation equals to 0.01; population size sets to 360; stop criteria equals to 2000 iterations. For the island model parameters are: migration rate sets to 5; migration frequency equals to 200 iterations; and topology is ring. We did not generate a cellular PEA because we did not have the proper architecture for its execution. The speedup is computed by $S_p = \frac{T_s}{T_p}$ where T_s represents the time for running the code in 1 thread and T_p is the required time for executing in p threads. This metric is known as weak speedup and was proposed by Alba [26] because the code is exactly the same regardless the number of threads.

 ISBN: 978-1-61208-419-0

Table III presents the speedup and the efficiency achieved by the Parallel Genetic Algorithm where we can observe that the island model got the best speedup being close to the ideal one. In terms of efficiency, the best one is reached using 2 threads as expected because the overhead caused by the parallel synchronization is smaller.

TABLE III. SPEEDUP AND EFFICIENCY ACHIEVED BY A GA

Master-Slave			
Threads	Time(ms)	Speedup	Efficiency
1	3071.774	-	-
2	2318.774	1.325	66.237
3	2158.129	1.423	47.445
4	2152.774	1.427	35.6728
Island			
Threads	Time(ms)	Speedup	Efficiency
1	2939.935	-	-
2	1516.29	1.939	96.945
3	1069.871	2.748	91.598
4	866.484	3.393	84.824
Hierarchical			
Threads	Time(ms)	Speedup	Efficiency
1	2940.42	-	-
2	1554.87	1.891	94.555
3	1101.58	2.67	88.975
4	942.29	3.12	78.01

B. COCOMO

The first result regards COCOMO basic considering the development of all possible outputs, *i.e.*, all parallel models and evolutionary algorithms summing up 5507 lines, can be seen in Table IV.

TABLE IV. NUMBER OF LINES OF CODE (LOC) PER MODEL AND ALGORITHM

Parallel Model/Algorithm	GA	EE
Master-Slave	602	589
Island	859	848
Cell	635	641
Cell Multi-Individual	675	658
Sub-Total	2771	2736
Total	**5507**	

Taking into account that the project is organic and using (1), (2) and (3), we can estimate the following results: $MP = 14.37$ (effort), $time = 6.9$ months, $p = 2.1$. In other words, developing all available models and algorithms would require 15 people/month of effort, 7 months of time and 3 people according to COCOMO basic model.

In the intermediary model, the following multiplication effort are considered as essential for developing all parallel models and evolutionary algorithms: RELY, CPLX, ACAP, AEXP, PCAP, LEXP, MODP and SCED [25]. Then, using the predefined values (4), and also taking into account the non-defined values as nominal inputs, we can calculate the values presented in Table V for different levels. For instance, considering parameter as very low level, a project would take 46.8 people/month, 11 months and 5 people to complete, whereas considering the level as very high those values are reduced to 13 people/month, 7 months and 2 people.

C. COCOMO II

In COCOMO II, we considered the following efforts: RELY, CPLX, RUSE, PVOL, ACAP, PCAP, AEXP, PLEX, LTEX, TOOL and SCED [24]. Further, we also considered all scale factors for computing B values using (5) as illustrated by Table VI.

TABLE V. COMPUTING THE EFFORT FOR THE COCOMO INTERMEDIARY MODEL

	Very Low	Low	Nominal	High	Very High
Total ME	2.4	1.5	1.0	0.8	0.7
PM	46.8	28.7	19.2	15.3	12.8
Time	10.8	8.9	7.7	7.1	6.6
p	4.3	3.2	2.5	2.2	1.9

TABLE VI. COMPUTING B AND ME FOR THE COCOMO II INTERMEDIARY MODEL

	Very Low	Low	Nominal	High	Very High	Extra High
B	1.23	1.16	1.10	1.04	0.97	0.91
ME	2.7	1.5	1.0	0.8	0.6	0.9

Taking into consideration that now we have six different Bs, we can calculate the table of effort for each one as presented in Table VII, where we can observe that as we increase both the level of scale factors and the multiplier factors, the effort tends to be lower, which is an expected behavior. It is important to notice that the extra high level is not completely filled up which causes an increase in the effort parameters, which is not desirable.

TABLE VII. COMPUTING EFFORT FOR COCOMO II INTERMEDIARY MODEL PER B

	Very Low	Low	Nominal	High	Very High	Extra High
ME	2.7	1.5	1.0	0.8	0.6	0.9
B = 1.23						
PM	71.02	38.05	25.88	20.11	15.90	22.26
Time	12.63	9.96	8.61	7.82	7.15	8.13
p	5.62	3.82	3.01	2.57	2.22	2.74
B = 1.16						
PM	63.74	34.15	23.23	18.05	14.27	19.98
Time	12.12	9.56	8.26	7.51	6.87	7.80
p	5.26	3.57	2.81	2.40	2.08	2.56
B = 1.10						
PM	57.24	30.67	20.86	16.21	12.82	17.95
Time	11.64	9.18	7.93	7.20	6.59	7.49
p	4.92	3.34	2.63	2.25	1.94	2.40
B = 1.04						
PM	51.39	27.54	18.73	14.55	11.51	16.11
Time	11.17	8.81	7.61	6.92	6.33	7.19
p	4.60	3.12	2.46	2.10	1.82	2.24
B = 0.97						
PM	46.14	24.72	16.81	13.06	10.33	14.46
Time	10.72	8.46	7.31	6.64	6.07	6.90
p	4.30	2.92	2.30	1.97	1.70	2.10
B = 0.91						
PM	41.43	22.19	15.10	11.73	9.28	12.99
Time	10.29	8.12	7.01	6.37	5.83	6.62
p	4.02	2.73	2.15	1.84	1.59	1.96

D. Discussion on Effort

As previously stated, we consider the software as an organic one, since it does not have a huge amount of lines. Thus, the result for COCOMO basic may be considered as an interesting estimation of effort, time and people if the team has some previous experience in the subjects. These results are similar to those one given by COCOMO II using a high level of multiplier efforts which would represents the necessity of learning the aforementioned subjects.

 ISBN: 978-1-61208-419-0

In terms of COCOMO II, we can observe that as we increase the level of the team we also decrease the effort parameters (people/month, months and pearson), as expected. Also, it gives a more precise measure due to the use of both multiplier efforts and scale factors, excepting when the extra high level is used in Multiplier Efforts (ME) because this particular level is not fully-filled, producing a little increase in the parameters. On the other hand, we can notice that parameters are more susceptible to the scale factors and multiplier effort making the measures more accurate in the all levels.

Doing an analysis of the person who developed the system, we can consider the following multiplier for COCOMO: RELY = high, CPLX =high, ACAP =high, AEXP =nominal, PCAP = high, LEXP = high, MODP = =high and SCED =nominal; and the following one for COCOMO II: RELY = high, CPLX = high, RUSE = high, PVOL = low, ACAP = high, PCAP = high, APEX = high, PLEX = high, LTEX = high, TOOL = high and SCED = Nominal. Either, we considered all scale factor as being in the maximum level.

TABLE VIII. ANALYSIS OF THE PERSON WHO DEVELOP THE PARALLEL EVOLUTIONARY ALGORITHMS

	COCOMO Basic	COCOMO Intermediary	COCOMO II (B = 0.91)
ME	-	0.76	0.67
PM	14.4	14.6	10.08
time	6.88	6.93	6.02
p	2.09	2.11	1.68

Thus, the results are shown in Table VIII, where COCOMO II presented a more precise estimation, considering that all knowledge was acquired during the disciplines in the master degree.

V. CONCLUSION

This paper presented a software whose main purpose is to reduce the programming effort when programming parallel evolutionary algorithms. A speedup test showed that it is possible to achieve a good speedup using the parallel models. Moreover, the models COCOMO and COCOMO II were used in order to support our hypotheses of saving effort. All in all, the JPEAG can save around one year of the time using 4 people if a low level of knowledge is held, or about 6 months and 2 people if the programmers has a high level of previous knowledge.

REFERENCES

[1] E. Cantú-Paz, "A Survey of Parallel Genetic Algorithms", Department of ComputerScience and Illinois Genetic Algorithms Laboratory, Universityof Illinois at Urbana-Champaign, 1998.

[2] A. E. Eiben and Smith, J. E., "Introduction to Evolutionary Computing", Berlin: Springer Verlag, 2003.

[3] B. Boehm, B. Clark, E. Horowitz, and C. Westland, "Cost Models for Future SoftwareLife Cycle Processes: COCOMO 2.0", Annals of Software Engineering, 1, 1995, pp. 57–94.

[4] F. Pasini and L. Dotti, "Code Generation for Parallel Applications Modelled with Object-Based Graph Grammars", Electronic Notes in Theoretical Computer Science, v. 184, 2007, pp. 113–131.

[5] K. A. Hawick and D. P. Playne, "Automated and parallel code generation for finite-differencing stencils with arbitrary data types". Procedia Computer Science, v. 1, n. 1, 2010, pp. 1795–1803.

[6] A. Rodrigues, F. Guyomarch, J-L. Dekeysera, and Y. Menach, " Automatic Multi-GPU Code Generation applied to Simulation of Electrical Machines", IEEE Transactions on Magnetics, v. 48, n. 2, 2012, pp. 831–834.

[7] W. Jiamthubthugsin and D. Sutivong, "Portfolio management of software development projects using COCOMO II". In Proceedings of the 28th international conference on Software engineering (ICSE). ACM, New York, NY, USA, 2006, pp. 889–892.

[8] T. N. Sharma. "Analysis of Software Cost Estimation using COCOMO II", International Journal of Scientific & Engineering Research, v. 2, Issue 6, 2011.

[9] L. L. Minku and X. Yao, "How to make best use of cross-company data in software effort estimation?". In Proceedings of the 36th International Conference on Software Engineering (ICSE). ACM, New York, NY, USA, 2014, pp. 446–456.

[10] L. V. Patiland, R. M. Waghmode, S. D. Joshi, and V. Khanna, "Generic model of software cost estimation: A hybrid approach", IEEE International on Advance Computing Conference (IACC), 2014, pp. 1379–1384.

[11] Z. Dan, "Improving the accuracy in software effort estimation: Using artificial neural network model based on particle swarm optimization", IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), 2013, pp. 180–185.

[12] L. V. Patil, N. M. Shivale, S. D. Joshi, and V. Khanna, "Improving the accuracy of CBSD effort estimation using fuzzy logic", International on Advance Computing Conference (IACC), 2014, pp. 1385–1391.

[13] F. Herrera, M. Lozano, and J. L. Verdagay, "Tack-ling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analisys", Artificial Intelligence Review, 4(12), 1998, pp. 265-319.

[14] Z. Michalewicz, "Genetic Algorithms + DataStructure = Evolution Programs", Springer-Verlag, New York, 3 edition, 1999.

[15] O. A. C. Cortes, R. H. C. Santana, M. J. Santana, and O. R. S. Mendez, "Anáise de Operadores de Recombinao em Estratgias Evolutivas Aplicados no Refinamento de um Sistema Nebuloso", In: Simpósio Brasileiro de Automática Inteligente, 2005.

[16] J. Yao, "Analysis of Scalable Parallel Evolutionary Algorithms", IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada. July, 2006.

[17] E. Alba and M. Tomassini, "Parallelism and Evolutionary Algorithms", IEEE Transactionson Evolutionary Computation, Vol. 6, No. 5, October, 2002.

[18] M. Sakuray, "Estudo da Influência dos Parâmetros de Algoritmos Paralelos da Computao Evolutiva no seu Desempenho em Plataformas Multinúcleos", PhD Thesis, Universidade Federal de Uberlândia, 2014.

[19] J. Herrington, "Code generation in action". Greenwich: Manning Publications, 2003.

[20] D. Lucrédio, "Uma Abordagem Orientada a Modelos para Reutilizao", Phd Thesis USP, So Carlos, SP, Brazil, 2009.

[21] D. Manolescu, M. Voelter, and J. Noble. Pattern Languages of Program Design 5. Reading: Addison-Wesley Professional, 2006.

[22] Apache Velocity Project Site - http://velocity.apache.org - accessed: Sep-9th-2014.

[23] H. Feng, K. Li-shaff, and C. Yu-ping, "A Generic Design Model for Evolutionary Algorithms", Wuhan University, *Journal of Natural Sciences*, v. 8, n. 1b, 2003, pp. 224–228.

[24] B. W. Boehm, "Software cost estimation with Cocomo II". Prentice Hall, Upper Saddle River, NJ, 2000.

[25] Y. Miyazaki and K. Mori, "COCOMO evaluation and tailoring". In Proceedings of the 8th International Conference on Software engineering, IEEE Computer Society Press, Los Alamitos, CA, USA, 1985, pp. 292–299.

[26] E. Alba, "Parallel Evolutionary Algorithms Can Achieve Super-Linear Performance", Information Processing Letters, v. 82, 2002, pp. 7-13.

[27] B. Steece and B. Boehm, "A constrained regression technique for cocomo calibration", Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement, 2008, pp. 213–222.

[28] Locatelli, M., "A Note on the Griewank Test Function", Journal of Global Optimization, v. 25, n. 02, 2003, pp. 169-174.

 ISBN: 978-1-61208-419-0

Automated Transformation of Multi-agent Protocols to Coloured Petri Nets

Ashwag Omar Maghraby
Computer Science Department
Umm Al-Qura university
Makkah, Saudi Arabia
e-mail: aomaghraby@uqu.edu.sa

***Abstract*— As multi-agent protocols are getting more and more complex, analyzing the behavior of such protocols is becoming increasingly important to ensure that they satisfy agents objectives and terminate correctly. This paper presents a tool for automated transformation from multi-agent protocols written in Lightweight Coordination Calculus language to high level Colored Petri nets models. This automation constructs a well-defined mathematical structure model that can be leveraged to formal analysis multi-agent protocol and used with the Standard Functional Programming language to automatically check whether the protocol is understandable and advantageous to the objectives of agents. The benefits of our approach consist in the new approach of analysing the MAS protocol and automatically validate key behavior properties of the MAS protocol to ensure that the protocol satisfies agents objectives.**

Keywords- Multi-agent protocol; Colored Petri net; Automated transformation; Protocol analysis.

I. INTRODUCTION

In a Multi-Agent System (MAS), two or more agents have to work together to find a final solution and satisfy their individual goals by exchanging messages following interaction protocols. An interaction protocol is a set of rules that direct the communication between several agents [1]. These protocols constrain the possible sequences of messages that may occur in agent interactions and describe how agents should react to messages received during interactions [2]. There are a finite number of messages in transmission and reception for each MAS protocol.

The need to understand, study and analyze MAS protocols properties is growing, as these protocols are becoming more complex due to the rich behavior introduced by concurrency, communication, and uncertainty. In fact, interactions between agents may be affected by different unexpected factors, for example, unexpected message, loss of messages or deviation in the message order.

Coloured Petri Nets (CPNs) [3] and CPN Tool [4] have been widely used to address these challenges. CPNs and CPN Tools provide a graphical representations and a mathematical formalism for the description, construction, execution, formal analyzing, and understanding of distributed and complex MAS protocols. CPNs ensure that a property is verified by all possible protocol executions [5].

In this paper, we propose an automatic transformation from multi-agent protocols written in Lightweight Coordination Calculus language (LCC) to high level CPNs models. This automation constructs formal and executable models of MAS protocols. It is used with the Standard Functional Programming language (SML is a general-purpose, modular, functional programming language with compile-time type checking and type inference [6]) to automatically validate key behavior properties of the protocol and to ensure that the protocol satisfies agents objectives and terminates correctly. This approach is divided into three main steps: (1) automated transformation LCC protocol to CPNs model; (2) construction of state space; (3) automated comparing of the agent's objectives properties and the behavioral properties of the LCC protocol.

The rest of this paper is organized as follows. Section II gives an overview of the CPNs, and how to use it to model agent protocol. Section III gives and overview of MAS protocol language (LCC). Section IV describes our approaches. Section V describes the automated transformation from the LCC protocol into an equivalent CPNs model. Section VI highlights the construction of state space approaches. Section VII details the verification approach and Section VIII gives an example of our approach.

II. CPNS AND USING CPNS TO MODEL MAS PRTOCOLS

CPNs used in a large variety of different areas such as MAS communication protocols. It provides a framework for the construction and analyzis of these protocols. A CPN model of a protocol describes the states of the protocol and the transitions between these states [3]. A brief introduction to CPNs is presented in this section.

A. CPNs

The CPN model consists of four elements [7][8]: data, place, transition, and arc which describe the net structure of the CPN model. An example of a CPN modelled in the CPN tool is depicted in Figure 1. This model has:

1) Three colour sets (Topic, Message and Role): A colour set can be a basic colour set (integer, string, real and Boolean) or a product of colour sets or a combination of other colour sets (a declared colour set from already declared colour sets). Colour sets are used to declare variables, other colour sets, functions, operations, constants and a place's inscription. A token is associated with a colour set and has data values (token colours) attached to it.

 ISBN: 978-1-61208-419-0

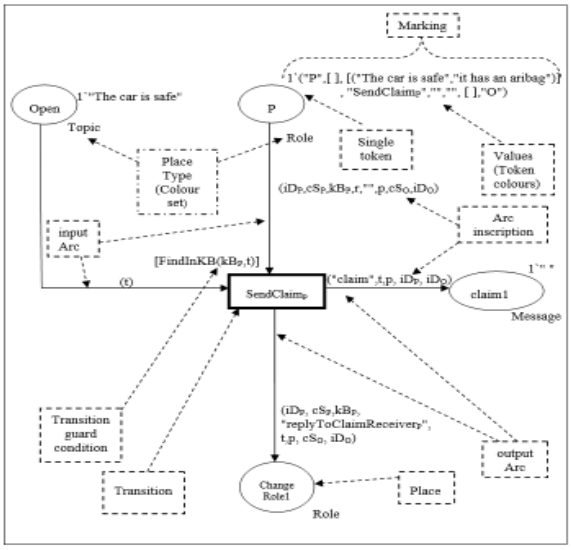


Figure 1. CPNs Model Elements Example

2) Four places (Open, P, claim1 and ChangeRole1): A place is a location (drawn as ellipse). It is used to hold data items (tokens). Tokens must match the place type (colour set). A place is associated with a marking, which indicates the number of stored tokens and the value (token colours) of these tokens. The state of the CPN model, at a particular moment, is represented by the set of markings of all the places.

3) One transition called SendClaim$_P$: A transition is an activity, which represents an event and is drawn as a rectangle. It is used to transform data between places. In practice, transition receives data from one or more places, checks its guard condition, executes its associated code segment, and sends the result to other places. A guard condition is a Boolean expression enclosed in square brackets that appears above the transition rectangle. A code segment is a computer program written in the CPN SML language (in the CPN Tool) or in the other kinds of notations, which has a well-defined syntax and semantic.

4) Four arcs: An arc is used to connect a place and a transition and to specify the data flow (the pre- and post-condition relation between transitions). An arc is associated with inscription, which is used to describe how the state of the modelled system changes. In the CPN Tool, an arc inscription is an expression that consists of CPN SML variables, constants and functions.

One of the key features of the CPN is its ability to construct large models in a hierarchical manner [8] by using subpages to build superpages. The subpages interact with each other and with the superpages through a set of transitions and a set of places. In practice, subpages used to model individual agent where superpages used to model communication protocol, which enables the message exchange among the agents of the protocol and produces a change of the protocol state.

B. Using CPNs to model MAS prtocols

There are a number of works using CPNs to model MAS protocols. In some related work, Calderon [9] developed a tool to transform UML–based systems of two large–scale UML systems [10] to CPN models (Design/CPN XML file) [3][8]. But the CPN models generated by the tool are not ready for analysis. The user needs to perform some manual work to get an executable CPN model and to be able to verify the correctness of the generated CPN. Another difference between our verification tool and Calderon's tool is that in the Calderons' approach, the dynamic behavior of the system is analyzing by running the Design/CPN tool simulator, while in our approach, the dynamic behavior of the system is analyzing by using state space techniques and the CPN SML language.

Suriadi's et al. [5] used the CPN Tool to model one case study of the Privacy Enhancing Protocols (PEPs) called the Private Information Escrow Bound to Multiple Conditions Protocol (PIEMCP) manually. Then, this work used the state space techniques, CPN SML language and session-data files to model validation and verification of the PIEMCP. The similarity between our verification approach and Suriadi's et al. approach [5] is that both use the state space techniques, CPN SML language and files (the session-data file in Suriadi's et al. approach and the properties file in our work). However, the main difference between our verification approach and this approach is that Suriadi's et al. approach generates a CPN model from a PIEMCP system model manually, while our approach generates a hierarchical CPN model from an LCC protocol by using a set of transformational rules automatically.

III. AGENT PROTOCOL DEVELOPMENT LANGUAGE

The Lightweight Coordination Calculus (LCC) [11] is a declarative, process calculus-based, executable specification language for choreography [12], which is based on logic programming and is used for specifying the message-passing behavior of MAS interaction protocols.

A. LCC Syntax

The abstract syntax of an LCC clause [11] is shown in Table 1. In an LCC framework, each of the N ≥ 2 agents is defined with a unique identifier *Id* and plays a *Role*. Each agent, depending on its *Role*, is assigned an LCC protocol.

An LCC protocol can be recursively defined as a sequential composition (denoted as *then*) or choice (denoted as *or*) of LCC protocols. In an LCC protocol, agents can change roles, exchange (receive or send) messages and exit the dialogue under certain constraints *C* (null ← C). Null represents an event (a do-nothing event) that does not involve role changing or message exchanging. A constraint is defined as a propositional formula specified over *term*s connected by *or* and *and* operators.

 ISBN: 978-1-61208-419-0

TABLE I. THE ABSTRACT SYNTAX OF LCC

	Meaning
Framework :=	{Clause,….}
Clause :=	Agent ::= Dn
Agent :=	a(Role, Id)
Dn :=	Agent | Message | null ← Constraint | Dn then Dn | Dn or Dn
Message :=	M => Agent | M => Agent ← Constraint | M <= Agent | Constraint ←M <= Agent
Constraint :=	Term | Constraint and Constraint | Constraint or Constraint
Role :=	Term
M :=	Term
Term:=	Constant (Argument,.....)
Id	Constant | Variable
Constant	Character sequence made up of letters or numbers beginning with a lower case letter
Variable	Character sequence made up of letters or numbers beginning with an upper case character
Argument	Term | Constant | Variable

Messages *M* are the only way to exchange information between agents. An agent can send a message *M* to another agent (M => Agent), and receive a message from another agent (M <= Agent). There are two types of constraints over the messages exchanged: pre-condition and post-condition. Pre-conditions (M => Agent ← C) specify the required conditions for an agent to send a message. Post-conditions (C ←M <= Agent) explain the states of the receiver after receiving a message.

B. LCC Examples

This is the simplest example of a persuasion protocol between two agents *P* and *O*. *P* and *O* have arguments for and against *Topic*. Agent *P* sends a *claim* message *Topic* and agent *O* receives this *claim* message *Topic*. A fragment of LCC protocol for the interchange in this argument is:

```
a(R1,P)::=
    claim(Topic) => a(R2, O)
    then         a(R3,P).
a(R2,O)::=
    claim(Topic) <= a(R1, P)
    then         a(R4,O).
```

This is read as: role *R1* of agent *P* sends a claim message to the role *R2* of agent *O* and role *R2* of agent *O* receives the claim message from role *R1* of agent *P*. Then *P* changes its role to *R3* and *O* changes its role to *R4*.

IV. AUTOMATED TRANSFORMATION APPROCH

Our automated transformation approach can answer the following question: Does the LCC MAS interaction protocol satisfy the agent's objectives (behavior properties) and terminates correctly? Three steps are needed to answer this question:

1) Transforming the LCC prtocol into an equivalent CPNs model. This step is processed in a fully automatic way*;*

2) Constructing the state space from the generated CPNs model.

3) Comparing the agent's objectives properties and the behavioral properties of the LCC protocol using CPN SML functions. A positive (negative) result indicates that a specific property is satisfied (unsatisfied).

The following sections discuss the details of each of these three steps.

V. STEP ONE: AUTOMATED TRANSFORMATION FROM LCC TO CPNS MODEL

Given the LCC interaction protocol as an input, the automated tool transforms the LCC protocol into an equivalent CPNs model using a set of transformational rules. In our approach, CPN model is described as CPNXML file. A CPNXML file [13] is an extended markup language (XML) document that describes the modelling elements of the CPN model.

We have developed a step-by-step technique that allows the user to automatically transform an LCC protocol into the CPNXML file by:

1) Declaring colour sets and functions.

2) Generating a CPN subpage for each LCC role. Each subpage represents a role behavior .

3) Connecting all the CPN subpages for each indivual agent by generating one CPN superpage. CPN superpage describes the interaction between roles, where the messages that are passed between two roles determine the interaction between the subpages of the two roles.

In practice, to automate the transformation process from an LCC protocol into CPNXML file we use 12 LCC-CPNXML tables (9 tables to generating CPN subpage and 3 tables to generate CPN superpage), where transitions and places are connected according to a set of transformation rules. The use of LCC-CPNXML tables makes the transformation faster and the resulting CPN model can be executed with data and analyzed, not only by our tool, but also by other users (using CPN Tool) since CPN has a comprehensible graphical representation. The following subsections give more details of the transformation process from an LCC protocol into CPNXML file.

A. Declaration of Colour Sets and Functions

Many communication between agents will be dialogues, and will specify more than two roles. In this approach, we use three different primary types of colour sets: *TOPIC*, which is used to model the main dialogue topic; *Message*, which is used to model messages arguments and *Role*, which is used to model role arguments.

Each agent has a knowledge base *KB* (private knowledge) and a commitment store *CS* (common knowledge). During the agent interactions, the agents take turns to send messages. Each agent makes his choice between possible messages depending on its *CS* and *KB*. In practice, the *CS* is continuously updated after sending or receiving each message by either adding to or subtracting from its argument. For that reason, we defined thirteen different basic functions, which are used to find, get, add or subtract an argument from either a *CS* or *KB* list. These functions are written in the CPN SML language [7].

 ISBN: 978-1-61208-419-0

The CPNXML format of the three types of colour sets and thirteen functions are saved in the Global Declaration file called "*CPNmainCode*". The user does not need to know about these colour set types or functions unless he/she needs to define new types or functions. For more information about how to define new CPN SML colour set types or functions, please read [4][8].

B. Generation of a CPN Subpage

Nine tables are used to automate the transformation process from LCC roles into CPN subpages. Space limitations prevent us from presenting each and every one of those tables instead we will discuss LCC *Message Sending Statement* table, which gives more details of the transformation process from an LCC message sending statement into CPNs model. The LCC message sending code is transformed into a high level CPN model by creating (as shown in Table II):

1) One new transition where the transition ID = unique identifier, the transition name= "Send" + Message name, and guard condition = LCC message Boolean conditions (line 1 to 7 of Table II);

2) One new place where the place ID = unique identifier, the place name = message name, place colour set type = Message and place (port) type= Out (line 8 to 19 of Table II);

3) One arc (output arc), which is used to connect the new transition to the new place, where the arc ID = unique identifier, the arc type= TtoP (output arc), the transition ID reference = the new transition ID, the place ID reference = the new place ID, the arc inscription = (Message arguments) (line 20 to 28 of Table II).

TABLE II. LCC-CPNC+XML TRANSFORMATION TABLE (SEND A MEESAGE)

LCC Code ***(Send a Message)***	Message(Topic) => a(RoleName(Arguments),AgentID) ← Conditions
CPNs Model	**CPNXML Structure**
Send message symbol [Boolean conditions] "Send"+ Message Name (Message arguments) Message Name Out Message	1. **<trans** id="ID1423689023"> 2. <text> ***"Send"+ message name*** </text> 3. <cond > 4. <text tool="CPN Tools "version="2.9.11"> 5. ***LCC Boolean conditions*** </text> 6. </cond> 7. **</trans>** 8. **<place** id="ID1423689035"> 9. <text> ***Message name*** </text> 10. <type id="ID1423689036"> 11. <text tool="CPN Tools" version="2.9.11"> 12. ***Message*** </text> 13. </type> 14. <initmark id="ID1423689037"> 15. <text tool="CPN Tools" version="2.9.11"/> 16. </initmark> 17. <port id="ID1424205036" type="**Out**"> 18. </port> 19. **</place>** 20. **<arc id="ID1423689049"** 21. **orientation="TtoP" order="1">** 22. <transend idref="***New transition ID***"/> 23. <placeend idref="***New place ID***"/> 24. <annot id="ID1423689050"> 25. <text tool="CPN Tools version="2.9.11"> 26. ***Message arguments*** </text> 27. </annot> 28. **</arc>**

TABLE III. LCC-CPNC+XML TRANSFORMATION TABLE (ROLE IN THE CPN SUPERPAGE)

LCC Code **(Role)**	a(RoleName(Arguments, Topic),AgentID)
LCC CPNs Model	**CPNXML Structure**
Role symbol Role Name Subpage Name	1. **<trans id="ID1414172135">** 2. <text> ***Role Name*** </text> 3. <subst subpage= ***"Corresponding subpage ID"*** 4. portsock= "(***socket ID, Port ID***) " 5. <subpageinfo id="ID1414172175" 6. name= ***Corresponding subpage Name*** > 7. </subpageinfo> 8. </subst> 9. **</trans>** 10. **<arc id="ID1423689049"** 11. **orientation="*PtoT*" order="1"** 12. <transend idref="***Substitution transition ID***"/> 13. <placeend idref="***Related socket ID***"/> 14. <annot id="ID1423689050"> 15. <text tool="CPN Tools version="2.9.11"> 16. ***Socket arguments (e.g. Role arguments ,*** 17. ***Message arguments)*** 18. </text> 19. </annot> 20. **</arc>**

TABLE IV. LCC-CPNC+XML TRANSFORMATION TABLE (DIAGLOUE TOIC)

LCC ***(Dialogue Topic Argument)***	a(RoleName(Arguments, Topic),AgentID)
LCC CPNs Model	**CPNXML Structure**
Dialogue Topic symbol Open Dialogue In Topic (Topic arguments)	1. **<place** id="ID1423689035"> 2. <text> ***OpenDialogue*** </text> 3. <type id="ID1423689036"> 4. <text tool="CPN Tools" version="2.9.11"> 5. ***Topic*** </text> 6. </type> 7. <initmark id="ID1423689037"> 8. <text tool="CPN Tools" version="2.9.11"/> 9. </initmark> 10. <port id="ID1424205036" type="***In***"> 11. </port> 12. **</place>** 13. **<arc id="ID1423689049"** 14. **orientation="*PtoT*" order="1">** 15. <transend idref="***New substitution transition ID***"/> 16. <placeend idref="***New place ID***"/> 17. <annot id="ID1423689050"> 18. <text tool="CPN Tools version="2.9.11"> 19. ***Topic arguments*** 20. </text> 21. </annot> 22. **</arc>**

C. Generation of a CPN Superpage

Three tables are used to automate the generation of one CPN superpage. Each LCC role is transformed into a high-level Petri net by creating (as shown in Table III):

1) One new substitution transition. (line 1 to 9 of *Table III*);

2) One or more arcs (line 10 to 20 of *Table III*);

3) If this role is the primary role (the first role in the LCC code, which is responsible for opening the dialogue), then:

a) Create one new place (line 1 to 12 of Table IV);

b) Create one arc (input arc) (line 13 to 22 of Table IV).

4) If this role is the agent's primary role, then:

a) Create one new place (line 1 to 12 of Table V);

b) Create one arc (input arc) (line 13 to 23 of Table Table V).

 ISBN: 978-1-61208-419-0

TABLE V. LCC-CPNC+XML TRANSFORMATION TABLE (AGENT'S STARTER ROLE ARGUMENTS)

LCC Code (Starter Role Arguments)	a(RoleName(Arguments, Topic),AgentID)
LCC CPNs Model	**CPNXML Structure**
Starter Role argument symbol Arguments initial values Agent ID In Role (Role Arguments)	1. <place id="ID1423689035"> 2. <text> *Agent ID_* </text> 3. <type id="ID1423689036"> 4. <text tool="CPN Tools" version="2.9.11"> 5. *Role* </text> 6. </type> 7. <initmark id="ID1423689037"> 8. <text tool="CPN Tools" version="2.9.11"/> 9. *Arguments initial values* 10. </initmark> 11. <port id="ID1424205036" type="*In*"> 12. </port> 13. </place> 14. <arc id="ID1423689049" 15. orientation="*PtoT*" order="1"> 16. <transend idref="*New substitution transition ID*"/> 17. <placeend idref="*New place ID*"/> 18. <annot id="ID1423689050"> 19. <text tool="CPN Tools version="2.9.11"> 20. *Role_arguments* 21. </text> 22. </annot> 23. </arc>

VI. STEP TWO: CONSTRUCTION OF STATE SPACE

The second step of our approach is to construct from the CPN model its state space (directed graph, which represents all possible executions of the CPN model). In the CPN Tool, state spaces can be constructed by:

1) Using the following CPN SML functions:
CalculateOccGraph( ) and *CalculateSccGraph( );*

2) Or, using the CPN State Space (SS) tool palette: For more information about using the state space tools see [14]. In our approach, the user needs to open the CPNXML file using the CPN Tool and construct the state space in a manual way using the CPN state space tool palette.

VII. STEP THREE: APPLYING VERIFICATION MODEL

The third step of our approach concerns:

1) The use of CPN Tool to observe and dynamic simulation and execution of the CPN model of MAS protocl (this will be done by the user);

2) The full state space analysis by applying a semi-automated verification model.

The verification model is carried out by checking four basic properties, which are independent of any dialogue (protocol) types:

1) Dialogue opening property: to check that the LCC protocol begins with a proper *Starting messge*;

2) Termination of a dialogue property: to determine if the LCC protocol terminates with a proper *Termination message*;

3) Turn taking between agents property: to guarantee that in the LCC protocol the turn-taking switches to the next agent after the current agent sends a message;

4) Message sequencing property: to check that the LCC protocol message exchange respects the gent messages expectation sequence.

In general, to verify each property, we use the following approach:

1) Create a new text file for each property and use the property name as the file name;

2) Extract the needed information from the state space graph and write this information in the property text file;

```
1.  Read&Save    SS=State Space information
2.  Read&Save    OpenDialogueMessages = information
3.  Call         CheckProperty1
4.  Input        (SS,OpenDialogueMessages)
5.  Extract      message1
6.  val checkODM =
7.      compare(OpenDialogueMessages,message1)
8.          if (checkODM ) then
9.              "Property 1(Dialogue opening) is Satisfied"
10.         else
11.             "Property 1(Dialogue opening) is not Satisfied"
12.  end      CheckProperty1
13.  Create&Save  Property1 result file
```

Figure 2. Property 1 as an SML Function

3) Get the information of a the gent messages expectation from the protocol expectation property file *(*this could be done in a fully automatic way or in a manual way*);*

4) Call the CPN SML property function, where the function inputs are the protocol expectation property file and the LCC protocol state space information (property text file);

5) Create a new text file (property result file) and write the CPN SML property function result in the property result file;

6) Repeat steps 1 to 5 for each property;

7) Present a report to the user indicating which properties are satisfied and which are unsatisfied.

Space limitations prevent us from presenting each and every one of those properties instead we will discuss Property-1 Dialogue Opening.

This property should guarantee that the LCC protocol will start if, and only if, a proposal agent sends a *Starting message*. Figure 2 shows the CPN SML specification of this property:

1) Line 1: Read the state space graph information from the Property1 text file and save this information in the state space informaiton (SS) variable.

2) Line 2: Read the Starting message information of a prtocol from protocol expectation property file and save this information in the *OpenDialogueMessages* variable.

3) Line 3: Call *CheckProperty1* function.

4) Line 4: CheckProperty1 function inputs are *SS* and *OpenDialogueMessages.*

5) Line 5: Extract the first message from the *SS (message1)*

6) Lines 6 and 7: Compare the first exchanged message in the state space graph with the *Starting message* where:

a) compare function is used to compare the first message;

 ISBN: 978-1-61208-419-0

b) checkODM variable represents the compare function result. It is considered true if the first message in the state space graph is the same as the *Starting message.*

7) Lines 8 to 11: Check the result of the comparison. A positive (negative) result indicates that Property 1 is satisfied (unsatisfied).

8) Line 13: Create a Property1 result file and write the result of *CheckProperty1* in this file.

Our verification method identifies only four basic properties, which are general properties that may be applied to several dialogues (protocols). However, if the user needs to verify different properties, the user needs to specify these properties and feed them to the generated CPNXML file manually.

VIII. EXAMPLE

An example of a MAS protocol is a persuasion dialogue (adapted from [15]), where a dialogue is presented as a game in which one participant (proponent 'P') attempts to persuade another participant (opponent 'O') to change their point of view about a particular topic 'T'. Our ***LCC/CPNProtocol*** Tool gets as input the LCC protocol of a persuasion dialogue (see Figure 3) and returns the corresponding CPN models (Space limitations prevent us from presenting each and every one of CPN models instead we will illustrate only two CPN models: *ClaimSender* in Figure 4 and *ClaimReceiver* CPN models in Figure 5) by using LCC-CPNXML tables. In practice, by using this tool, no additional programming is required.

User then can manually construct the state space of the generated CPN models using the SS tool palette in CPN Tools (see Figure 6). Then ***LCC/CPNProtocol*** Tool:

1) Gets agent's objectives properties expectation from user (*Figure 7 is an example of this properties where Starting Locutions file contains one message name claim, which is used to begin the persuasion dialogue*);

2) Automatically comparing the agent's objectives properties and the behavioral properties of the LCC protocol using CPN SML functions. To verify properties the following actions were performed:

a) Open the CPN model by using the CPN Tool;

b) Select the Evaluates a Text as ML Code(ML!) icon in the simulation tool palette and apply it to property page (*Figures 8 show the Dialogue opening property page after applying the ML! to it*);

3) Shows the verification result (see Figure 9).

IX. CONCLUSION AND FUTURE WORK

In this paper, we have presented a methodology to support formal validation of MAS protocol. The main idea is to automatically transform the LCC protocol into an equivalent CPN model using a set of transformational rules and then extracts four behavioral properties (Dialogue opening property, Termination of a dialogue property, Turn taking between agents property and Message sequencing property) of the LCC protocol from the CPN model state space. These properties can be used to check whether the protocol satisfies agents objectives and terminates correctly.

Our further work is targeted at investigating three questions: Can the user modify the available properties to suit their specific MAS protocol using our tool? Can our tool specify new properties in an automated manner? Can our tool take the new properties information from the user using a constrained form of natural language?

REFERENCES

[1] G. Chicoisne, "Dialogue between natural agents and artificial agents: An application to virtual communities", PhD thesis, National Institut of Polytechnique of Grenoble, pp. 71-74, 2004.

[2] M. Koji, S. Jin-Hua, and Q. Yasuhiro, "Study on Common Coordinate System by using Relative Position of Other Autonomous Robot", SICE Annual Conference, Japan, pp. 795-797, August 20-23, 2012.

[3] K. Jensen, "Coloured Petri Nets. Basic Concepts, Analysis Methods and Practical Use", Berlin, Springer Verlag, 1997.

[4] M. Westergaard and H. Verbeek, "*CPN Tools*" Eindhoven University of Technology, 2002, [retrieved 3, 2015], http://cpntools.org/.

[5] S. Suriadi, C. Yang, J. Smith, and E. Foo, "Modeling and Verification of Privacy Enhancing Security Protocols", 11th International Conference on Formal Engineering Methods ICFEM, Janeiro, Brazi, ICFEM, pp. 127-146, 2009.

[6] R. Milner, M. Tofte, R. Harper, and D. Macqueen, "The Definition of Standard ML" Cambridge, MA, USA, The MIT Press, revised edition, 1997.

[7] K. Jensen and L. Kristensen, "Coloured Petri Nets Modelling and Validation of Concurrent Systems" Berlin, Springer Verlag, 2009.

[8] K. Jensen, L. Kristensen, and L. Wells, "Coloured Petri Nets and CPN Tools for modelling and validation of concurrent systems" International Journal on Software Tools for Technology Transfer (STTT), 3rd ed., vol. 9, pp. 213–254, 2007.

[9] M. Eunice, "Model transformation support for the analysis of large-scale systems" Texas Tech University Electronic Theses and Dissertations, Master Thesis in Software Engineering, 2005.

[10] B. Bauer, J. Müller, and J. Odell, "Agent UML: A Formalism for Specifying Multiagent Interaction" Software Engineering and Knowledge Engineering, vol. 9, pp. 91-103, 2001.

[11] D. Robertson, "Multi-agent coordination as distributed logic programming" In DEMOEN, BART and LIFSCHITZ, VLADIMIR, Logic programming. Saint-Malo, France, 20th International Conference, pp. 416-430, 2004.

[12] R. Dijkman and M. Dumas, "Service-oriented Design: A Multi-viewpoint Approach" International Journal of Cooperative Information Systems, 4th ed., vol. 13, pp. 337-378, 2004.

[13] J. Billington et al. "The Petri Net Markup Language: Concepts, Technology, and Tools" The Netherlands, 24th International Conference, ICATPN 2003 Eindhoven, 2003.

[14] K. Jensen, S. Christensen, and L. Kristensen, "CPN Tools State Space Manual" University of Aarhus, Department of computer science, 2002, [retrieved 3, 2015].

[15] H. Prakken, "Coherence and flexibility in dialogue games for argumentation" Journal of logic and computation, 6th ed., vol. 15, pp. 1009-1040, 2005.

 ISBN: 978-1-61208-419-0

Agent P	Agent O
a(claimSender$_P$(KB$_P$,CS$_P$,CS$_O$,T,ID$_O$),ID$_P$) ::= claim(T) => a(claimReceiver$_O$(KB$_O$,CS$_O$, CS$_P$,ID$_P$),ID$_O$) ← addTopicToCS(T,CS$_P$) then a(replyToClaimReceiver$_P$(KB$_P$,CS$_P$,CS$_O$,T,ID$_O$), ID$_P$).	**a(claimReceiver$_O$(KB$_O$,CS$_O$,CS$_P$,ID$_P$),ID$_O$)** ::= claim(T) <= a(claimSender$_P$(KB$_P$,CS$_P$,CS$_O$,T,ID$_O$),ID$_P$) then a(replyToClaimSender$_O$(KB$_O$,CS$_O$,CS$_P$,T,ID$_P$),ID$_O$).

Figure 3. Two rols of LCC Protocol for Persuasion Dialogue

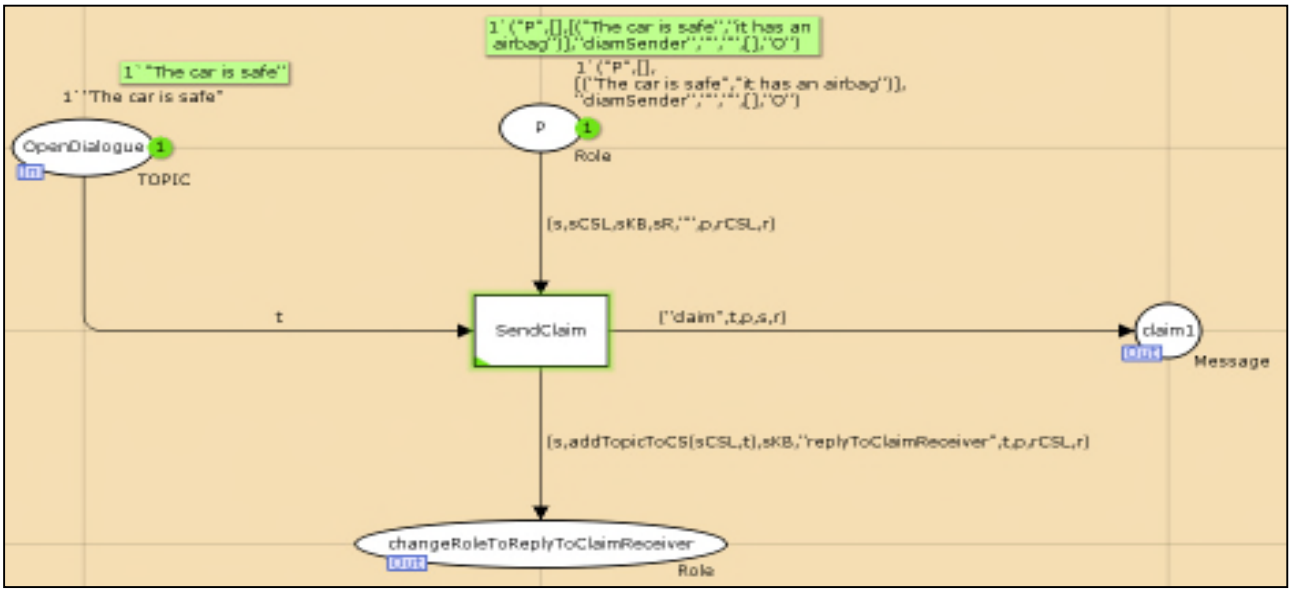


Figure 4. The claimSender$_P$ CPN Subpage

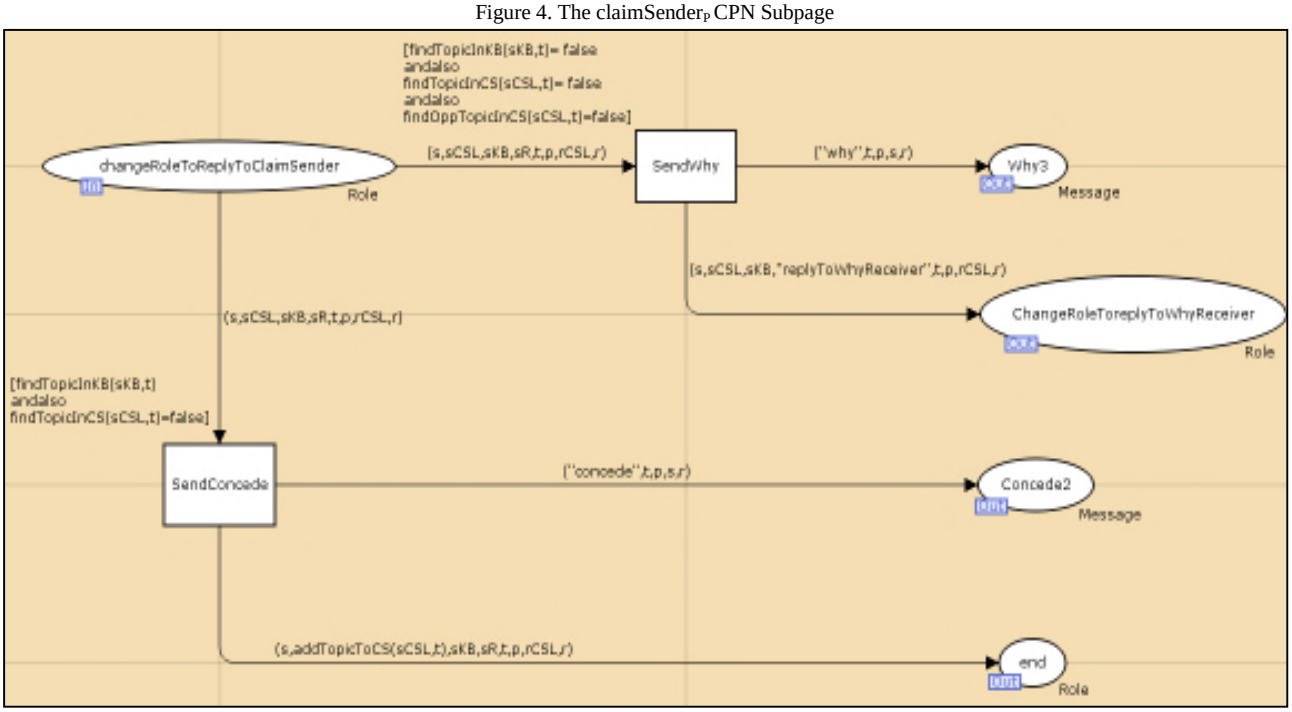


Figure 5. The claimReceiver$_O$ CPN Subpage

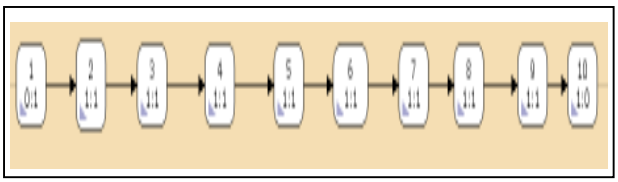


Figure 6. The State Space Graph

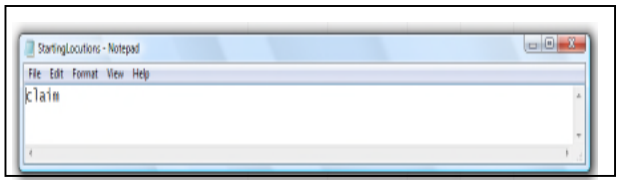


Figure 7. Starting Locutions file

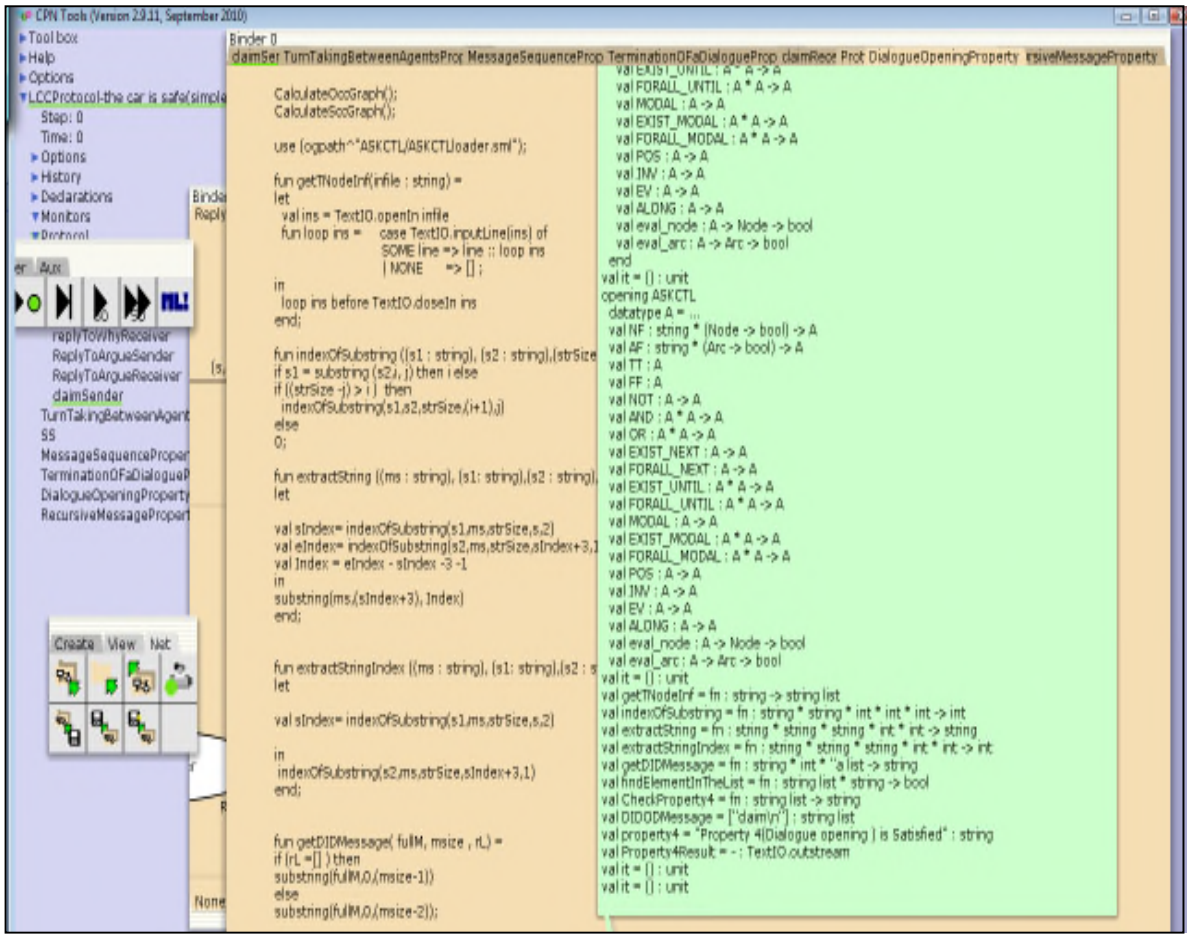


Figure 8. Dialogue Opening Property Page

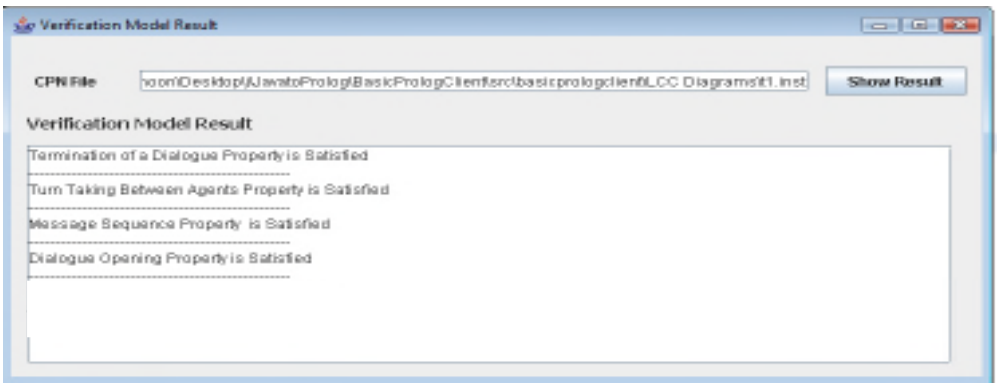


Figure 9. The Verification Result of the Five Basic Properties

 ISBN: 978-1-61208-419-0

Numerical Algorithms and Measurement Systems in Practical Implementation of Electrical Impedance Tomography

Tomasz Rymarczyk, Przemysław Adamkiewicz
Department of Research and Development
NET-ART, Lublin, Poland
Department of Computer Science, UCEA, Lublin, Poland
e-mail: tomasz@rymarczyk.com

Jan Sikora
Electrotechnical Institute, Warszawa, Poland
Lublin University of Technology, Lublin, Poland
e-mail: sik59@wp.pl

Abstract— **This paper presents a nondestructive method of brick wall dampness testing in real building structures and a new method of testing flood embankment dampness. We used a setup made of specially built laboratory models to determine the moisture level of test brick walls and flood embankments. The topological method and the gradient technique were used with the optimization approach. The finite element method was used to solve the forward problem. The proposed algorithm was initialized by using one step methods and topological sensitivity analysis. We constructed the forward model and we solved the inverse problem in order to visualize moisture inside objects. Practical examples of using the Electrical Impedance Tomography are also presented.**

Keywords-image reconstruction; inverse problem; level set method; finite element method; electrical impedance tomography.

I. INTRODUCTION

This paper presents a new method examining the flood embankment dampness and the brick wall dampness using the electrical impedance tomography [5][9][10]. Numerical methods of the shape and the topology optimization were based on the level set methods [2][3][4] and the gradient methods [15]. The discussed technique can be applied to the solution of inverse problems in the electrical impedance tomography. New algorithms to identify unknown conductivities were implemented. The purpose of the presented method is to obtain a better image reconstruction than gradient methods.

One of the major pathologies in historic buildings is the existence of dampness. Moisture transfer in walls of old buildings, which are in direct contact with the soil, leads to a migration of soluble salts responsible for many building problems. Building porous materials (e.g., brick or concrete), both natural and manufactured, have pores (like a sponge) and the moisture can be pulled up against gravity (capillary effect). Figure 1 shows an example of a damp historical wall. The dampness raising from the soil is a problem in old buildings, especially without adequate horizontal and vertical insulation of foundations. The moisture creates a danger not only to the walls, but also to human health. It promotes progress of rheumatic disorders and formation of fungus on the walls. Fungus can cause allergies and many other diseases. There are many different drainage systems (watertight barriers, injection of hydrofuge products, etc.). Regardless of the method, it is very important to continuously monitor the status of dampness during the drying process. In the case of historical buildings of great cultural importance, the use of destructive measurements is prohibited by conservation specialists. The traditional techniques used to deal with this kind of problem proved to be ineffective, justifying the need to find a new approach [1].

Figure 1. Historical damp wall. Examples of excessively damp brick walls in historical buildings.

Numerical methods of shape and topology optimization methods were based on the level set representation and the shape differentiation and there were topology changes possible during the optimization process [11][12]. Level set methods have been applied very successfully in many areas of the scientific modelling, for example in propagating fronts and interfaces [6][7][8]. Therefore, they are used to study shape optimization problems. Instead of using the physically driven velocity, the level set method typically moves the surfaces by solving the Hamilton-Jacobi equation (2). These approaches based on shape sensitivity include the boundary design of elastic.

In Section II, we present some information about electrical impedance tomography. Section III discusses

 ISBN: 978-1-61208-419-0

models of numerical methods. Section IV shows the numerical results and the paper concludes in Section V.

II. ELECTRICAL IMPEDANCE TOMOGRAPHY

The Electrical Impedance Tomography (EIT) is a technique of imaging the distribution of conductivity inside the tested object from measurements of the distribution of potentials on the object's surface. In Figure 2, the object is a brick wall with dampness rising from the ground.

The test results obtained by the nondestructive impedance tomography method are compared with the results obtained by numerical simulations. We prepared two prototype measuring systems. One is a low cost EIT tomography system containing 16 electrodes for measuring damp brick wall on one side. The second one is a full EIT system with 32 electrodes for testing on both sides of a wall.

The forward problem in EIT is described by Laplace's equation [9]:

$$\nabla \cdot (\gamma \nabla u) = 0, \qquad (1)$$

where γ denotes conductivity. Symbol u represents electrical potential.

III. MODELS

Nondestructive methods do not require that samples be taken from the wall, which is their advantage over the destructive methods.

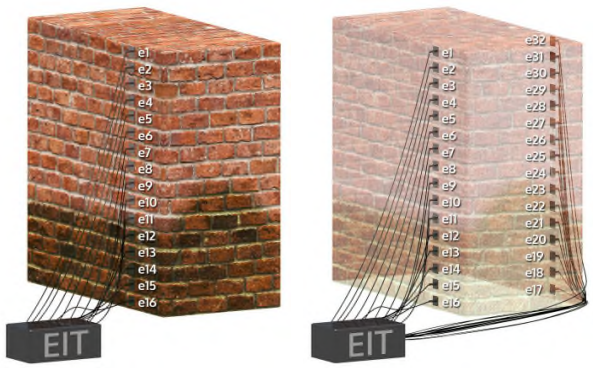


Figure 2. Measurement EIT systems with 16 electrodes and 32 electrodes on the damp brick wall.

Figure 3. Surface electrodes on the damp brick wall.

Among the nondestructive methods, the most popular are electric and nuclear methods, particularly, the electric resistance method, the dielectric method, the microwave method, and the neuronal method [1]. But, in the case of nondestructive methods, the dampness measuring instruments must be calibrated in order to determine the correlation between their indications and the weight concluded dampness of the tested material. In this paper, the electrodes can be easily attached to the tested object. The level set method and the gradient technique were based on shape and topology optimization to solve the inverse problem in the electrical impedance tomography. Such task can be considered as application of the electrical impedance tomography.

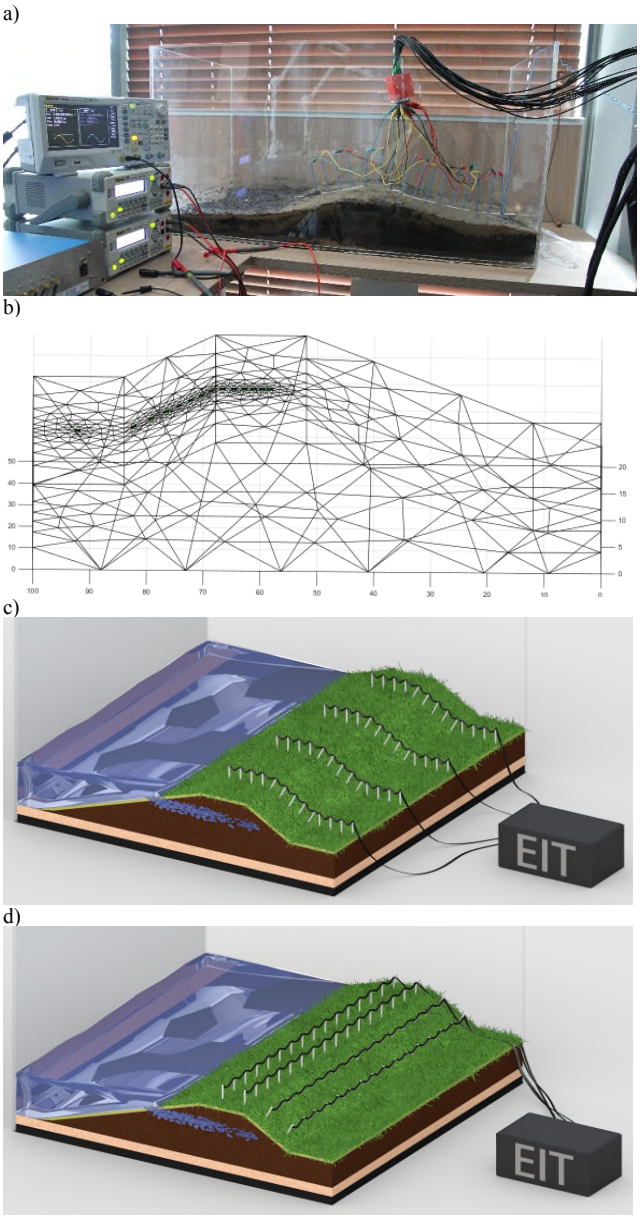


Figure 4. The geometrical laboratory models of the flood embankment: a) the laboratory model, b) the discretization model, c) the first model of measurement, d) the second model of measurements.

The object is a brick wall with dampness rising from the ground in Figure 2. Surface electrodes can be easily

attached to the wall (see Figure 3) by special conductive glue. We used an electrical impedance tomography device with multiplexer to make measurements.

The flood embankment system is given in Figure 4. The laboratory model (a) presents the measurement system with multiplexer and 16 electrodes. The discretization model (b) was based on the finite element method (using to solve the forward and inverse problem). The first model (c) collects measurements perpendicularly to the flood embankment. The second model (d) performs measurements parallel to the tested object.

IV. NUMERICAL METHODS

Numerical methods of the shape and the topology optimization were based on the level set representation and the shape differentiation and were made possible topology changes during the optimization process.

The motion is seen as the convection of values (levels) from the function ϕ with the velocity field $\vec{v}$. Such a process is described by the Hamilton-Jacobi equation [7]:

$$\frac{\partial \phi}{\partial t} + \vec{v} \cdot \nabla \phi = 0. \quad (2)$$

Here, $\vec{v}$ is the desired velocity on the interface, and is arbitrary elsewhere. Actually, only the normal component of $\vec{v}$ is needed ($v_n \equiv \vec{v} \cdot \vec{n} \equiv \vec{v} \cdot \nabla \phi / |\nabla \phi|$), so (2) becomes:

$$\frac{\partial \phi}{\partial t} + v_n |\nabla \phi| = 0. \quad (3)$$

Let λ be the adjoint function satisfying [5]:

$$-\Delta \lambda = u - u_m. \quad (4)$$

The material derivative $\dot{u}(x)$ is given by [5]:

$$\dot{u}(\vec{r}) \equiv \lim_{t \to 0} \frac{u_t(\vec{r} + t\vec{v}(\vec{r})) - u(\vec{r})}{t}, \quad (5)$$

where $(x,y) \in \Omega_t$. The shape derivative is following [5]:

$$u'(\vec{r}) \equiv \lim_{t \to 0} \frac{u_t(\vec{r}) - u(\vec{r})}{t} = \dot{u}(\vec{r}) - \vec{v}(\vec{r}) \cdot \nabla u(\vec{r}). \quad (6)$$

The steepest descent direction $\vec{v}$ is given by:

$$\vec{v} = -(\nabla u \cdot \nabla \lambda)\, \vec{n}. \quad (7)$$

In next step the level set function is updated:

$$\phi^{k+1} = \phi^k - \left(\nabla u^k \cdot \nabla \lambda^k\right) |\nabla \phi^k| \Delta t, \quad (8)$$

For the minimization problem, iterative coupling of the level set method and the topological gradient method have been proposed. Both methods are gradient-type algorithms, and the coupled approach can be cast into the framework of alternate directions descent algorithms. One step methods and topological algorithms were used to solve this problem.

V. RESULTS

The test results were obtained by using the EIDORS software [14] in Figure 5. We prepared two prototype measuring systems. The first of them is the EIT tomography system that contains 16 electrodes for measuring damp brick wall on one side.

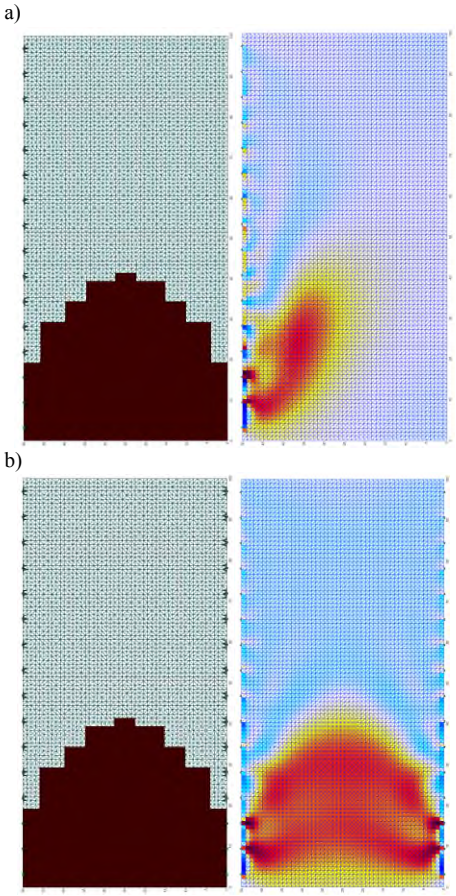


Figure 5. Geometrical model of the investigated dumped wall with (a) 16 and (b) 32 electrodes.

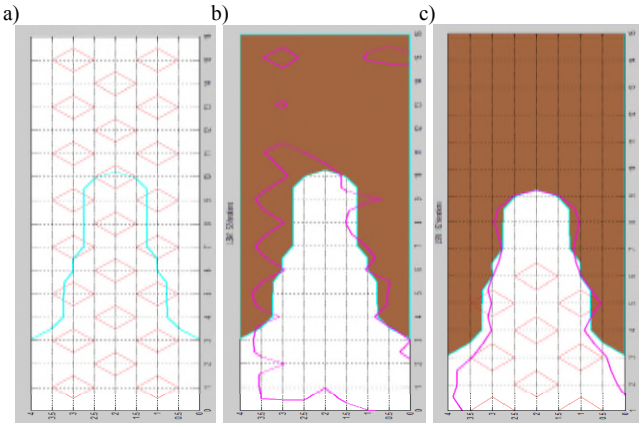


Figure 6. The image reconstruction by the level set method: (a) the original objects and the zero contour, (b) the reconstructed object with one side of the wall, (c) the reconstructed object with both side of the wall.

The latter is a full EIT system with 32 electrodes for testing on both sides of a wall. Figure 2 shows exemplary numerical reconstruction of moisture in the damp wall by

the Gauss Newton one step method. Figure 6 presents the image reconstruction by the level set method.

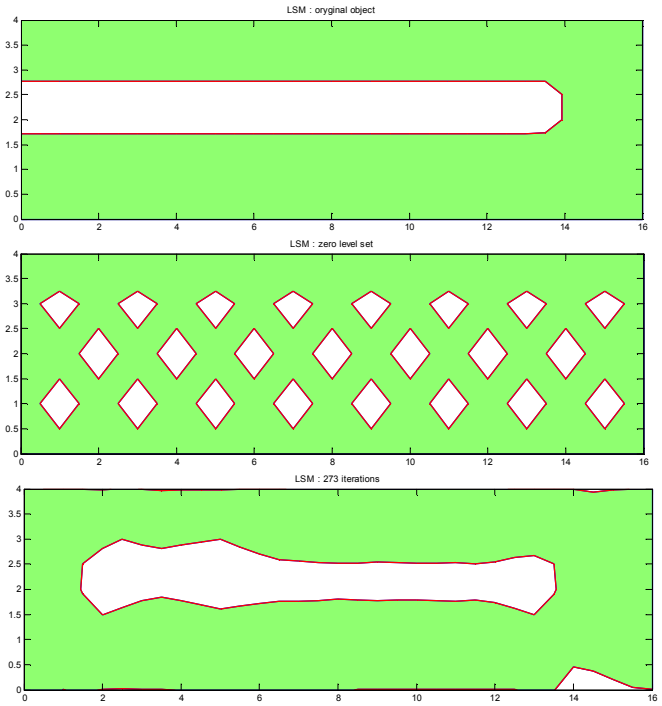


Figure 7: The image reconstruction - the original objects with the measurement system, the zero level set function, the reconstructed object.

Surface potential measurements are performed at different angles of projection whereby the information needed to determine an approximate distribution of conductivity inside the object is obtained. In the example reported below, the conductivity of searched objects is known. The representation of the boundary shape and its evolution during an iterative reconstruction process is achieved by the level set method. The forward problem was solved by the finite element method. Figure 7 shows the process of the image reconstruction with the zero level at initial step is represented by a lot of objects. The picture shows the original object and the reconstruction after 273 number of iterations. The process of reconstruction seems to be the correct one, because the region border is located nearly the real object edges.

VI. CONCLUSION AND FUTURE WORK

A new nondestructive method of the flood embankment dampness and the brick wall dampness was tested by the electrical impedance tomography. Numerical methods of the shape and the topology optimization were based on the gradient techniques and the level set representation. The presented methods have been applied very successfully in many areas of the scientific modelling. These approaches were based on sensitivity analysis. An efficient algorithm for solving the forward and inverse problems would also improve a lot of the numerical performances of the proposed methods. In modeling of the problem in the electrical impedance tomography, it is required to identify unknown conductivities from near-boundary measurements of the potential. Future work will be based on an implementation of artificial intelligence algorithms to solve the inverse problem.

REFERENCES

[1] J. Hoła, Z. Matkowski, K. Schabowic, J. Sikora, K. Nita, and S. Wójtowicz, "Identificationof Moisture Content in Brick Walls by means of Impedance Tomography", COMPEL, Vol. 31, Issue 6, 2012, pp. 1774-1792.

[2] S. Osher and R. Fedkiw, "Level Set Methods and Dynamic Implicit Surfaces", Springer, New York 2003.

[3] J. A. Sethian, "Level Set Methods and Fast Marching Methods", Cambridge Univeristy Press, 1999.

[4] G. Allaire, F. De Gournay, F. Jouve, and A. M. Toader, "Structural optimization using topological and shape sensitivity via a level set method", Control and Cybernetics, vol. 34, 2005, pp. 59-80.

[5] K. Ito, K. Kunish, and Z. Li, "The Level-Set Function Approach to an Inverse Interface Problem", Inverse Problems, vol. 17, no. 5, 2001, pp. 1225-1242.

[6] S. Osher and J. A. Sethian, "Fronts Propagating with Curvature Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations", Journal of Computational Physics, vol. 79, 1988, pp. 12-49.

[7] S. Osher and R. Fedkiw, "Level Set Methods: An Overview and Some Recent Results", Journal of Computational Physics, vol. 169, 2001, pp. 463-502.

[8] S. Osher and F. Santosa, "Level set methods for optimization problems involving geometry and constraints. Frequencies of a two-density inhomogeneous drum", Journal of Computational Physics, vol. 171, 2001, pp. 272-288.

[9] T. Rymarczyk "Using electrical impedance tomography to monitoring flood banks", International Journal of Applied Electromagnetics and Mechanics 45, 2014, pp. 489–494.

[10] T. Rymarczyk, P. Tchórzewski and J. Sikora, "Topological Approach to Image Reconstruction in Electrical Impedance Tomography", ADVCOMP 2014, ISBN: 978-1-61208-354-4, Rome, Italy, August 24 – 28,2014, pp. 42-45.

[11] J. Sokolowski and A. Zochowski, "On the topological derivative in shape optimization", SIAM Journal on Control and Optimization, vol. 37, 1999, pp. 1251–1272.

[12] C. Tai, E. Chung, and T. Chan, "Electrical impedance tomography using level set representation and total variational regularization", Journal of Computational Physics, vol. 205, no. 1, 2005, pp. 357–372.

[13] T. Rymarczyk, J. Sikora, and B. Waleska, "Coupled Boundary Element Method and Level Set Function for Solving Inverse Problem in EIT", Proc. 7th World Congress on Industrial Process Tomography, Sep. 2013, pp. 312-319.

[14] A. Adler and W. Lionheart, "Uses and abuses of EIDORS: An extensible software base for EIT", Physiol. Meas., 27, 2006, pp.:25-42.

[15] S.F. Filipowicz, T. Rymarczyk, J. Sikora, "Level Set Method for Inverse Problem Solution In Electrical Impedance Tomography", XII International Conference on Electrical Bioimpedance & V Electrical Impedance Tomography, 2004, pp.: 519-522.

 ISBN: 978-1-61208-419-0

Mapping Serial-Monadic Dynamic Programming onto CUDA-Enabled GPUs

Chao-Chin Wu, Kai-Cheng Wei, Jian-You Lin
Dept. of Computer Science and Information Engineering
National Changhua University of Education
Changhua, Taiwan
email: {ccwu, kcwei}@cc.ncue.edu.tw

Wei-Shen Lai
Department of Information Management
Chienkuo Technology University
Changhua, Taiwan
email: weishenlai@gmail.com

***Abstract*—With the advent of high performance computational power, processing particularly complex scientific applications and voluminous data is more affordable. One of the hot parallel processors is general-purpose graphics processing unit (GPU), which has been widely adopted to accelerate various time-consuming algorithms. This work demonstrates how to apply a more condensed data structure and the interblock synchronization to efficiently map the serial-monadic dynamic programming onto GPUs.**

Keywords-dynamic programming; parallel computing; graphics processing unit; CUDA; data dependence.

I. INTRODUCTION

Dynamic programming (DP) is a popular method used to solve complex problems. DP can be classified into four categories: (1) serial-monadic, (2) non-serial-monadic, (3) serial-polyadic, and (4) non-serial-polyadic. Recently, many efforts have studied how to map the DP problems onto emerging general-purpose graphics processing units (GPUs), where nVIDIA has introduced CUDA (Compute Unified Device Architecture) to ease the programming on their GPUs for various kinds of applications [1]. CUDA is a hardware and software coprocessing architecture for parallel computing enabling nVIDIA GPUs to execute programs written with C, C++, Fortran, OpenCL, and other languages.

Previously, we have investigated how to optimize the mapping of non-serial-polyadic DP problems onto CUDA-enabled GPUs, where the Optimal Matrix Parenthesization (OMP) problem was chosen as our study example. This work focused on how to optimize the mapping of serial-monadic DP problems onto nVIDIA GPUs and the 0/1 knapsack problem is adopted for this study.

Recently, Boyer and his colleagues proposed a DP approach with a compression mechanism to implement the 0/1 knapsack problem on a CUDA-enabled GPU [2]. The primary data structure used in their method, called the item selection table, is a 2-dimensional (2D) array and used to record if an item is selected or not for each capacity, C_i, where $0 < C_i < C$ and $C_i = i$, assuming the capacity of the knapsack is C. If there are N items in the problem, the dimension of the 2D array is $N \times C$. When N and C are both large integers, the 2D array requires a large amount of global memory space. To address this problem, they used one bit to represent if a certain item is selected or not. Next, each thread compressed the outcomes of every 32 stages into one integer, which is called the row-compression method. Furthermore, after analyzing the result values stored in the vectors, the row-compressed data, they found a large portion to the right hand side on the vector is filled with 1. On the other hand, the left hand side is filled with 0. Each thread recorded the indices of the boundaries of continuous 1's and 0's in the vector and then used the indices to replace the huge number of 1's and 0s' on both ends of the vector. In this way, the amount of the data to be transferred between the CPU and the GPU were reduced significantly. When the problem size becomes larger, the compression becomes more effective.

We observed two disadvantages of Boyer's method. First, although the compression method can reduce the amount of data transferred between CPU and GPU, if the item selection table is one-dimensional (1D) rather than 2D, the data can be minimized significantly. Second, the data in the shared memory cannot be reused because the Boyer's method invokes the same kernel one time for each stage of the DP problem, which can be addressed with the interblock synchronization. Based on the above observation, we propose a new approach to improve the performance of the knapsack problem on CUDA-enabled GPUs.

The remainder of the paper is organized as follows. Section II introduces the main idea of our proposed approach and details the key design issues. Section III gives the conclusion of our work.

II. PARALLEL APPROACH

The main idea of the new approach consists of two factors. First, the item selection table is 1D, the dimension is $1 \times C$. Adopting 1D structure not only can minimize the amount of data transferred between CPU and GPU but also can be stored in shared memory. Second, the interblock synchronization [3] is adopted to reuse the 1D item selection table in shared memory.

Although the 1D item selection table requires less memory space and can be fit into high-speed shared memory, the problem about this solution is its potentially exploitable parallelism is much less than that in the 2D one. The reason is explained as follows. To use dynamic programming with a 1D item selection table to solve the knapsack problem, items will be determined one by one whether one specific item is included in knapsacks of different capacities or not. One item will be considered during one stage of dynamic programming and one thread is assigned with one knapsack

 ISBN: 978-1-61208-419-0

with one specific capacity. If one thread determines that the current item is included in its assigned knapsack, it writes the current item ID to the corresponding field of the 1D table. Note that one item is processed during one stage of dynamic programming. To process the current item, one thread needs the information, produced by another thread, about how the previous item is included for another knapsack of one specific capacity that is related to the weight of the current item. Consequently, one thread cannot process the current item until the required information is written to the corresponding table field. There exists a read-after-write dependency for the above operations. Furthermore, the thread cannot update its assigned table field with the current item ID until every thread requiring the result, produced by the thread in the previous stage, has all finished his reading. There exists a write-after-read dependency. To enforce the correct data flow, two synchronization points should be inserted after each of the above two kinds of dependency. However, the second dependency occurs only for the 1D table and it can be eliminated totally if a 2D table is used instead because the results of two items for the same knapsack are placed on two different locations. To address the problem of the write-after-read dependency, we allocate multiple buffers to store several versions of the 1D table. With multiple buffers, one thread can write the result for the next item to one buffer even though the result for the current item on another buffer has not read by other threads. The maximum number of buffers is dependent on the shared memory space. In this way, only the read-after-write dependency needs a synchronization to enforce data consistency.

Because the required data for one thread might be in shared memory on other streaming multiprocessors, all the threads have to write the results of the current item selection to the global memory and all the thread blocks should participate a barrier synchronization followed also. Instead of invoking the same kernel as many times as the number of DP stages, as suggested by the CUDA programming guide, we adopt the interblock synchronization mechanism that requires to invoke the kernel only one time. The advantage of invoking the same kernel again is that the data in shared memory is stale and cannot be reused. Consequently, the results in shared memory have to be written to the global memory before return from the kernel and then retrieve the results from the global memory to shared memory after the kernel is invoked again. On contrast, invoking the kernel only one time allows the results in shared memory can be reused repeatedly for all the DP stages. That is, the results stored in shared memory can be read by all the threads on the same streaming multiprocessors even though we proceed to next DP stage. Using the interblock synchronization is able to significantly reduce the amount of data transferred between shared memory and the global memory. Moreover, reusing the results in shared memory can shorten the latency of accessing shared memory. However, we still need to write the results to the global memory because the threads of other blocks cannot access the shared memory on different streaming multiprocessor. There also exists a write-after-read dependency on global accesses. Therefore, the 1D table also has multiple copies to eliminate the write-after-read dependency.

We further analyze the data dependence between thread blocks. The blocks except the last one have to write data to the global memory to allow the subsequent blocks to read, as shown in Figure 1. On the other hand, the blocks except the first one have to read results from the global memory to local registers. Note that all threads have to write the final version of the 1D item selection table from shared memory to global memory during the last DP stage. The result 1D item selection table in global memory will then be transferred back to CPU.

We use CUDA version 3.2 to implement different algorithms, including our approach and Boyer's approach, for the 0/1 knapsack problem. They are run on a system consisting of one AMD Phenom 9850 CPU and one nVIDIA GEFORCE 460. Our approach, adopting a 1D table, outperforms the previous work, as shown in Figure 2.

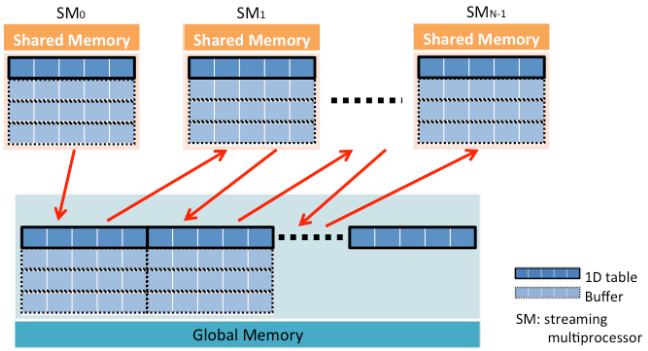


Figure 1. The addlocation of one-dimensional table and the buffers.

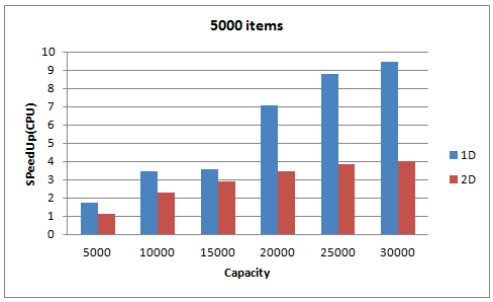


Figure 2. The speedup comparision between 1D and 2D tables.

In the legend of Figure 2, 1D and 2D represent our approach and the approach proposed by Boyer *et al.*, respectively. The number of items is 5000. The speedup is derived from dividing the execution time of the sequential CPU version by the execution time of one of the two GPU-based approaches. When the capacity is increased, the speedup is increased also. It is because either the 1D or the 2D Item Selection Table becomes larger, and the saved amount of global memory accesses becomes larger also. The size of 1D table is much smaller than that of 2D table. When using the 1D table, our approach can reduce the required

 ISBN: 978-1-61208-419-0

memory traffic between CPU and GPU significantly. The larger the capacity, the more memory traffic can be saved by our proposed 1D table, resulting in better performance.

III. CONCLUSION

This work introduced how to use 1-dimensional data structure and the explicit inter-block synchronization to map the knapsack problem, an application of serial-monadic dynamic programming, on to a CUDA-enabled GPU. The results showed the proposed approach outperforms the previous work reported by Boyer *et al*.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council, Taiwan, for financially supporting this research under Contract No. NSC103-2221-E-018-024.

REFERENCES

[1] NVIDIA GPU, http://www.nvidia.com/object/what-is-gpu-computing.html, retrieved: June 2015.

[2] V. Boyer, D. El Baz, and M. Elkihel, ”Solving knapsack problems on GPU”, Computers & Operations Research,vol. 39, no. 1, 2012, pp.42–47.

[3] C. C. Wu, K. C. Wei, and T. H. Lin, “Optimizing dynamic programming on graphics processing units via data reuse and data prefetch with inter-block barrier synchronization,” IEEE ICPADS, 2012 pp.45–52.

 ISBN: 978-1-61208-419-0

Exploring a Community Clustering Algorithm on Semantic Similarity in Large-Scale Social Network

Laizhong Cui, Yuanyuan Jin, Nan Lu
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, Guangdong, P. R. China
Email:cuilz@szu.edu.cn, jinyuanyuanzj@qq.com, lunan@szu.edu.cn

***Abstract*-This paper proposes a semantic similarity clustering algorithm on the cluster analysis of large-scale social network. By utilizing the semantic hierarchy of WordNet, the proposed method defines the key concept sets and the concept feature values for the community. In our method, the semantic relations between concepts of the community nodes are also constructed, which expands the application of clustering algorithms from text documents to social network. The cluster structures derived from the proposed algorithm are in concordance with peoples' judgments on a specific area, which will lead to the solution of the clustering problems in the social network of different areas. Compared with VSM and k-MEANS, the experiment results show that the proposed algorithm obtains more reasonable results, which validates its effectiveness.**

Keywords-semantic similarity; WordNet ontology; social network; community structure; clustering algorithm; key concept set

I. INTRODUCTION

Semantic similarity [1][2] is a concept with various definitions according to different areas. Taking the term "virus" as an example, the similarities of virus and its categories differ when it is taken to the biological area and computer area. This is caused by different definitions in these two areas. Therefore, understanding the definition other than the term itself becomes more and more important.

The WordNet ontology [3] is an online word sense mapping system, containing concept word sets from different areas and relationships among them with a sematic network structure. Based on WordNet, this paper extracts words and constructs a semantic IS-A relationship hierarchy and mixes concept word set into the standard hierarchical structure.

It is difficult to calculate similarity about different weighted features and classification learning from the long text and web document processing. The most popular methods for solving the problems are applying to vector space model (VSM) [4][5]. Most of these algorithms are costly in computing power, and some algorithms need some background knowledge, as well as manpower. Therefore, it is difficult to apply these algorithms in the unstructured social networks. Especially, when the nodes with similar meanings have few common words in the community structure, VSM will heavily affect the effect of cluster finding and result in serious deviation from actual structure. For example, the "sports community" and "badminton clubs" should belong to the topic of sports, but the VSM will return 0 in sematic similarity due to no common words.

The social networks are abstractions of complex systems and each node in the social network represents the individual unit in the complex system. The edges between nodes are relationships in the social network formed according to some certain rules. There are various types of social networks in the real world, such as social network, biological network, etc. Finding community structure is not a random selection in a large number of nodes with same properties, but a discrimination in nodes with different types, among which the nodes with same property are linked with more connections, while different types of nodes are sparsely connected. Finding community structures within a social network is an important step towards clustering analysis and research of the network.

Clustering is an important method in finding community structure. Through the clustering method, internal regulations and characteristics can be discovered. The clustering algorithms is capable of automatically generating the category number without adding manual annotation and training classifier. As an unsupervised machine learning

 ISBN: 978-1-61208-419-0

method, clustering has a higher flexibility and better automatic processing power. As the increasing tendency of community information dependence [6][7][8][9], people require intelligent information processing other than the processing of the word pattern or word sense. Therefore, the semantic similarity computing becomes one of the ways to solve community clustering problem. It is crucial in improving the effectiveness and accuracy of the clustering result, judging community structure correlations, classifying communities, and mining data.

This paper proposes a WordNet semantic network learning method. By utilizing the semantic hierarchy of WordNet, the proposed method defines the key concept sets [10] and the concept feature values for the community and then use them to define the semantic similarity. According to the semantic similarity, a community clustering method in social network is presented. The cluster structures derived from the proposed algorithm are in concordance with the judgments of peoples on a specific area, which will lead to the solution of the clustering problems in the social network of different areas. Compared with VSM and k-MEANS [11], the proposed algorithm discovers more reasonable results and shows its effectiveness.

The rest of the paper is organized as follows. Section II provides some relevant knowledge. In Section III, we propose a high performance cluster algorithm (CASN). The experiments and results are presented in Section IV. Finally, Section V concludes our work.

II. RELEVANT KNOWLEDGE

A. WordNet Ontology

WordNet is a widely used English words knowledge base and widely applied in natural language processing, semantic translation, which has attracted much international attention [12]. WordNet is organized by sematic relations. It uses synonym sets (synsets) to representative concepts. Keywords in synsets are bounded and the sematic relationships between synsets are also kept in the hierarchy. One word can be mapped into several synsets and one synset contains several words, which provides a way on representing semantic relationships into the relationships between the concept sets and synsets. WordNet semantic relationships mainly include: the parent and child, synonymous, antonym, is-a-part-of and containment, attribute properties, "leading to" relationships and so on. Based on the English WordNet, the Chinese WordNet is an ontology of the Chinese words and the concept word set, by using existing English-Chinese dictionary library to translate the word in English into Chinese and get the knowledge base. It also has the functions of the concept word, same-word and pan-word. The key concept word is the basic element of Chinese WordNet, and use a number of relation types to connect these concept words, which leads to a key concept word set.

B. Similarity Calculation

The calculation of similarities between any two words starts from mapping the words into the concept word sets that they belong to, and then calculates the similarities between each pair. Finally, according to these similarities of concepts word sets, the word similarities are achieved.

1) The concept word and similarity

For the convenience of knowledge sharing and reuse, WordNet clearly defines concepts of different areas and their relationships. Since there is no formal standard, the descriptions of the same problem in different areas will be different. Even in the same area, different ontologies may also have some heterogeneity. This will significantly affect the utilization of WordNet. The ontology mapping is one of the ways to resolve this heterogeneous problem, while the similarity calculation is a key part of the ontology mapping.

Definition 1 (concept word): Concept is an abstract description of the objects in real world. A concept word is defined as a triad $Con=\{N, A, R\}$, where N is the name, A is the attribute set, R is the relationship set. The name and attribute describe the internal characteristics of the concept, relationship set and express the relation between concept and external environment, and also reflect the external characteristics of concept. Concept words can be represented by instances and therefore be more specific in concept meaning.

Definition 2 (similarity): By the definition 1, it is known that the concept word has three important elements, including name, attribute and relationship. So the similarity

 ISBN: 978-1-61208-419-0

calculation also contains those three components. WordNet is seen as a semantic tree organized according to the hierarchical relationships and concept word relationships. In this paper, the similarity calculation is based on the word content similarity in WordNet, which is the semantics distance between them, in other words it is the path length of two semantics in the semantic tree. Similarity values range from 0 to 1. If two concept words are completely different, the similarity is 0. If they are identical, the similarity is 1. The similarity is calculated by (1), where $len(w_1, w_2)$ is the path length from the word w_1 to the word w_2.

$$SIM(w_1, w_2) = \frac{1}{1 + len(w_1 + w_2)} \quad (1)$$

a) The concept name similarity SIM_{cn}

Usually, the concept word in WordNet is a compound word and it is to determine its semantic distance directly. Therefore, the first several steps will be word segmentation and stopping-word removal to form a stopping-word table and key words extraction. Equation (1) is used to calculate key words similarities and get the summation to represent the similarity of concept name as (2).

$$SIM_{cn} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} SIM(c_{1i}, c_{2j}) \quad (2)$$

where, c_{1i} is the key word of concept c_1, $i \in [1,n]$. c_{2j} is the key word of concept c_2, $j \in [1,m]$.

b)The attribute similarity SIM_a

Attribute consists of attribute names (reflect attribute contents) and attribute domain (attributes' value ranges). Therefore, the attribute similarity calculation must include attribute name similarity and attribute domain similarity, which is described as follows:

$$SIM_a(a_1, a_2) = SIM_{an}(a_1, a_2) + \frac{1}{n} SIM_{ad}(a_1, a_2) \quad (3)$$

$SIM_{an}(a_1,a_2)$ is the attribute name similarity, which can use a similar concept name similarity calculation method to calculate the result. $SIM_{ad}(a_1, a_2)$ is the attribute domain similarity, and it mates calculate.

c) The relationship similarity SIMr

Relationship similarity reflects the connection degree between concept and external. The similarity calculation includes two parts: the relationship name similarity and the relationship association concept similarity, which is described as follows:

$$SIM_r(r_1, r_2) = \frac{1}{n} SIM_{rn}(r_1, r_2) + \frac{1}{n} SIM_{rc}(rc_1, rc_2) \quad (4)$$

$SIM_{rn}(r_1,r_2)$ is the relationship name similarity, achieved by similar calculation method as concept name similarity. $SIM_{rc}(rc_1,rc_2)$ is the association concept similarity, and it is calculated from the concept name similarity of association concept, where rc_1 represents the concept of associate relation r_1, and rc_2 represents the concept of associate relation r_2.

2) The semantic similarity

Semantic similarity calculation principle can be described as follows:

a) According to the relation between parent and child in WordNet, the further the distance between any two concept word nodes is, the smaller the semantic similarity is.

b) The higher density the concept word node locates at, the finer the local concepts are divided, which leads to a lower similarity.

As for two concept word nodes with the same distance, the deeper level it locates at, the more specific it will be, which means that the greater similarity is assigned.

The semantic similarity calculation formula is defined as follow:

$$SIM = \sigma + \alpha \times \bar{\varphi} + \beta \times \bar{\omega} \quad (5)$$

where, α and β are the weights of depth factor and density factor respectively, σ is the distance factor , φ is the density factor and ω is the depth factor.

C. The Key Concept Word Set and Concept Feature Value of Community

In the community clustering process, different community structure or different keyword density in community will lead to the distorted structure. Especially when the community data is in a high dimension, the quality, effect and the calculation speed of the clustering are significantly decreased. In order to improve the efficiency of the clustering, the dimension reduction method is a better choice.

 ISBN: 978-1-61208-419-0

At present, the dimension reduction methods mainly includes TF-IDF, information gain (IG), mutual information (MI), etc. [12][13], which are based on the lexical frequency statistics information. For the convenience of clustering operation, a structuring process for nodes in social network is required, which includes: establishing the community key concept word set and extracting concept feature values, and forming structured documents with key concept words. Similar with the text document processing, the words in structured documents can be divided into two classes: the function words and the content words. The function words are particle, which have no real meaning, while the content words are meaningful. According to the features of content words in the network, such as frequencies, positions and so on, weights are assigned to these words to obtain the concept feature values. This value is propositional to the frequency of the associated concept and if a concept appears in the title of the network, its concept feature value will be increased. When a certain concept feature value is greater than a given threshold, this word can be regarded as a key concept word.

At present, for clustering purpose, a document is always transformed into a noun list and the contribution of words' frequencies to the content of the document is ignored [14], which will lead to an unsatisfying performance. In this paper, the key concept word list illustrated in (6) is utilized, where a social network is regarded as a two-dimension array includes the concept words and their frequencies, to meet the requirement of clustering in the social network.

$$\sigma = \begin{cases} \sqrt{1-\frac{len^2}{\theta^2}}, & len < \theta \\ 0, & len \geq \theta \end{cases} \qquad \varphi = \frac{1}{\ln PN+1}$$

$$\omega = \begin{cases} \frac{\sqrt{|dep - E_d|}}{E_d}, & dep \geq E_d \\ -\frac{\sqrt{|dep - E_d|}}{E_d}, & dep < E_d \end{cases} \qquad (6)$$

In (7), w_i is the i-th concept appearing in community and f_i is the frequency associated with w_i. f_i is calculated by the frequency function, namely:

$$D = \{(w_1, f_1), (w_2, f_2), \ldots, (w_n, f_n)\} \qquad (7)$$

In (8), T_i is the feature value associated with the i-th concept word appearing in community; TF_i is the times which the first i concept word appearing in the community; m_i is the number of communities containing the first i concept word; M is the total number of the concept word in community. From (8), it is obvious that the feature value of a concept word is proportional to the frequency that the concept word appearing in sentence, and in inverse proportional to the number of communities containing the concept word.

$$T_i = TF_i log\left(\frac{M}{m_i}\right) \qquad (8)$$

III. HIGH PERFORMANCE CLUSTER ALGORITHM CASN

At present, no clustering algorithm could be generally applicable to the social network in revealing the complex structures that are represented by all kinds of multi-dimensional data sets. Generally, clustering algorithms can be classified into partitioning clustering, hierarchical clustering and density based clustering. The classic partitioning algorithms are vector space model clustering and k-neighborhood clustering [15], which are efficient for large data sets and applicable to Web document clustering applications. The hierarchical clustering algorithms use the association rules to split or cluster data in a hierarchical form to provide solution for hierarchical clustering. They are mostly applied in small data set.

To address the problems of predefined cluster number, initial value selection and local optimal issue, this paper proposes a clustering algorithm based on semantic similarity, called CASN (Clustering Algorithm of Social Network) to efficiently solve the community clustering problem in social networks.

A. Basic Ideas

The basic idea of our proposed social network community clustering algorithm is to define the node distance between community structures, which represents the similarity between community structures by the node semantic similarity (as shown in Figure 1). According to the similarities, the nodes are clustered one by one, and the closely related clusters will gather into a bigger cluster unit, which will grow in size gradually until all nodes form a cluster.

 ISBN: 978-1-61208-419-0

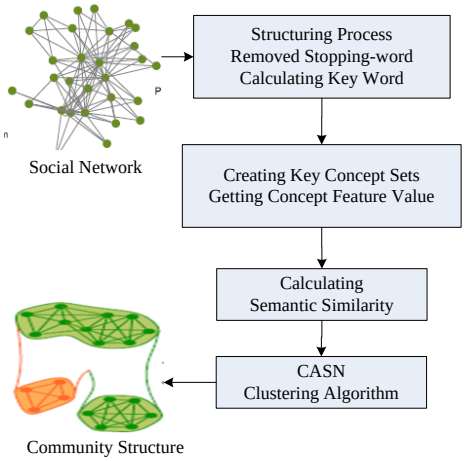


Figure 1. The basic idea of CASN

B. The details of the algorithm CASN

The first step of CASN includes the network structuration and feature extraction. The WordNet concepts and the semantic relationships among concepts are used to generate key concept sets and the concept feature values representing the community structure. Then the clustering algorithm based on the key concept set and the concept feature value is executed. Finally, the key concept set is used to express each clustered unit.

1) The key concept set extraction algorithm CASN-CSET

Clustering Algorithm of Social Network for Concept Set (CASN-CSET) algorithm scans through the identification of each network nodes to extract semantics and maps the extracted nouns into concept by WordNet. Each of the concept is initialized with an interpretation weight. The whole process is described in Figure 2.

Algorithm input: The network node identification document set *D*.

Algorithm output: The key concept of each network node in the document set *D*, *ConSET* [i].

```
i=0; continue=ture;    // i index network node document
/*each network node document in circulation processing D */
do
   file=nextfile (D); //take a network node document in order
   if (isnull(file))
        continue = false;
   else {
      titlewords = gettitle(file);
      word=first(titlewords); //extract the first semantic word
      titlewords.remove(word);
      /* remove the semantic word which have taken out and
        update titlewords */
   }
   while(isnotnull(word)) {
   conceptnode = lookupindexword(word, noun);
   if (isnotnull(conceptnode))
       ConSET[I].add(conceptnode,1);
       /*if concept nodes exist,then join ConSET[i],
         the weight is 1*/
      word = first(titlewords);
      titlewords. remove(word);
   }
   i ++ ;
   while(continue != false)
```

Figure 2. The key concept set extraction algorithm CASN-CSET

2) The concept feature value extraction algorithm CASN-FeaVAl

Clustering Algorithm of Social Network for Feature Value (CASN-FeaVAl) algorithm maps the semantic word into the concept word through the synonym and parent-child relationships in WordNet. Then, a small section of the concept words are selected to represent each document of the network structure. The whole process is described in Figure 3.

Algorithm input: Semantic word concept feature value array obtained from a normalization processing of the network structure document sets, including word segmentation, stemming, stopping and word-frequency calculation.

Algorithm output: Content feature value array *Feat[i]* representing the content of each network structure in the document set *D*.

```
i = 0;   // i index documents
/* circulation handling each network structure document */
do
   for each word in Feat[i] Do
      concept = mapintoconcept(word);
```

```
if (isnotnull(concept)) {
  if (concept in Feat[i])
    add cf to the orginal concepts weight;
    /* cf is concept frequency, the original concept
       feature valueue add concept frequency * /
  else
    Feat[i].add(concept, cf);
    hypernym = getdirecthypernym(concept)
  }
  if (isnotnull(hypernym))
    if (hypernym in Feat[i])
      add cf to the orginal hypernym 's weight;
    /* cf is concept frequency, the original
      concept feature valueue add concept frequency */
    else
     Feat[i].add (hypernym, cf);
    /* put the direct superior concept with concept
       feature valueue */
  endfor
  i ++;
  while (I !=Feat.length)
    i = 0;    /* inetwork node index*/
    /*circulation handleeach network node document */
    do
      for each concept in Feat[i] Do
      hypernym = getdirecthypernym(concept)
      if(isnotnull(hypernym)&&(hypernym in Feat[i]))
        /* reduce dimension */
      if (cf(hypernym) > cf(concept))
        Feat[i].remove(concept, cf)
      else
       Feat[i].remove(hypernym, cf)
      endfor
      i ++;
   while (I != Feat.length)
```

Figure 3. The concept feature value extraction algorithm CASN-FeaVAl

3) The clustering algorithm CASN

Algorithm input: The key concept set *ConSET[i]* and feature value *FeatVAL[i]* of each network structure in document set of network structure, Document number *n*, Cluster number *k*, weight coefficient of *he*, key concept set *kc*, and weight coefficients of concept feature value *cf*. The whole process is described in Figure 4.

Algorithm output: The clustering results *Clusters*, as well as the explanation for cluster results *Results*.

```
P = callsimilarity(ConSET, FeatVAL, kc, cf);
Clusters.initialize (n); .//cluster initialized to n cluster
Results.initialize(ConSET);
/* key concepts set ConSET initialized n cluster explain */
do
  findnearestcluster(c1, c2);
  Clusters. merge(c1, c2); // merger two cluster
  Results.merge(c1, c2);
  /* mark key concept set of explanation cluster and merger,
    and according to the similarity of concept in key concept
    set of cluster results */
  update();  // update similarity matrix
  n --;
  while(n>k) {
    callsimilarity (ConSET, FeatVAL, kc, cf)
    fndnearestcluster(c1, c2)
}
```

Figure 4. The clustering algorithm CASN

The similarity matrix is calculated based on the key concept set and its concept feature vector and its weight coefficient. Then two clusters *c1* and *c2* are found which their similarity degree are the biggest in the clusters.

IV. THE EXPERIMENT AND ANALYSIS

A. Experimental Data

The algorithm's experimental data are taken from 10 discussion groups of a Bulletin Board System (BBS). The discussion groups contain about more than 20000 discussion topics with a total number of 65000 entries.

B. Evaluation Standard

This paper adopts the NMI (Number Mutual Information) analysis method based on mutual information within clusters or categories. As stated by Deng and al. [14], this method can eliminate the influence on the final clustering result caused by the number of clusters. The closer the NMI values are, the better the clustering result is. The NMI value is

 ISBN: 978-1-61208-419-0

calculated as follows:

$$NMI = \frac{\sum_{h,l} n_{h,l} \lg\left(\frac{n n_{h,l}}{n_h n_l}\right)}{\sqrt{\left(\sum_h n_h \lg\left(\frac{n_h}{n}\right)\right)\left(\sum_l n_l \lg\left(\frac{n_l}{n}\right)\right)}} \quad (9)$$

where n_h is the number of data sample in categories h, n_l is the number of data sample in categories l, $n_{h,l}$ is the same data samples in both of the categories h and l, n is the total number of sample data.

C. The Experimental Result

The experiment uses the key concept set and concept feature value to calculate the similarity of the text. *kc* and *cf* are the parameters to adjust the key concept set and concept feature value. Figure 5 shows the effects of clustering by computing NMI under different ratio of *kc:cf*. We can see that when *kc:cf* =1:9, the clustering algorithm has the best performance.

In Figure 6, CASN is compared with other clustering algorithms, set kc:cf =1:9， the NMI value of CASN is about 0.6 when the cluster number is 10; meanwhile, NMI value of VSM algorithm is less than 0.5 but greater than 0.4, and K-means algorithm only get 0.4 on NMI; Even with the cluster number increasing, the NMI values of CASN algorithm are all near 0.6, which is superior to VSM and K-means. The testing results meet the requirements perfectly, which shows the validity and effectivity of the presented algorithms.

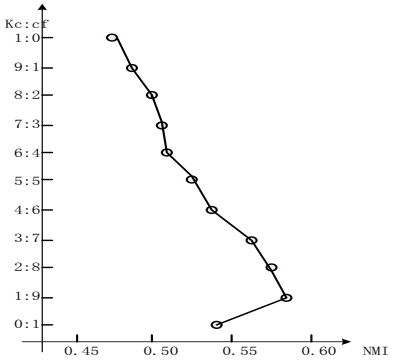


Figure 5. The effects of clustering by computing NMI different ratio of *kc:cf*

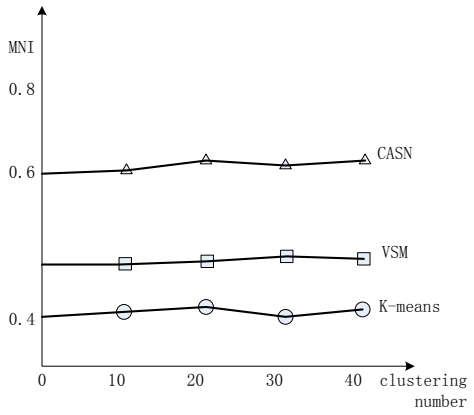


Figure 6. The effects of different clustering methods with *kc:cf* =1:9

V. CONCLUSION AND FUTURE WORK

This paper studied the community clustering algorithm based on semantic similarity under the social network scenario and proposed a feature value using WordNet semantic words which can construct community key concept set to express the community concept. Compared with clustering methods using space vector model SVM, the proposed algorithm shows a better performance. By introducing semantic relations between concept word set (also called synonyms set) and concepts to describe network nodes, our proposed algorithm reduces the dimension of feature lists representing community nodes, and therefore can be applied to the clustering analysis of the community structure. The method proposed in this paper deserves more research on some problems in the future. For example, the hierarchical relationships of the WordNet ontology is still not yet fully utilized. Some fuzzy concept word sets are hard to be defined and a better solution is also required to improve the clustering accuracy.

REFERENCE

[1] R. Zhang, “The Clustering Research On Terminology Definition,” Technology Terminology of China, pp. 14-19, Jan. 2011.

[2] S. Q. Zhao, T. Liu, and S. Li, “A Topical Document Clustering Method,” Journal of Chinese Information Processing, vol. 21, no. 2 pp. 58- 62, 2007.

[3] H. F. Zhu, W. L. Zuo, F. L. He, T. Peng, and W. Y. Ji, “A Novel Text Clustering Method Based on Ontology,” Journal of Jilin University (Science Edition), vol. 48, no.3, pp. 277-285, 2010.

[4] Q. Y. Yao, G. S. Liu, and X. Li, “VSM–based Text Clustering Algorithm” Computer Engineering, vol. 34, no. 9, pp. 39-43, 2008.

 ISBN: 978-1-61208-419-0

[5] M. W. Yuan and P. Jiang, “Compression algorithm for ontology based Vector Space Model Computer Engineering and Applications,” Computer Engineering and Applications, vol. 43, no. 10, pp. 12-17, 2007.

[6] Y. Li, H. Wang, and J. Yang, “Web document clustering algorithm based on semantic similarity,” Journal of hefei university of technology, vol. 32, no. 12, pp. 1846-1850, 2009.

[7] B. L. Yang and K. Y. Shao, “Web document clustering algorithm based on high performance feature selecting function,” Application Research of Computers, vol. 26, no. 2, pp. 546-551, 2009.

[8] P. Zhang, F. Yang, and S. Lu, “Clustering-Based Ontology Block Matching Approach,” Journal of Jilin University (Science Edition), vol. 49, no. 5, pp. 493-499, 2011.

[9] Y. J. Zhang, Y. P. Ren, L. C. Chen, and B. H. Xie, “Component clustering algorithm based on semantic similarity and optimization” Computer Engineering and Design, vol. 31, no.11, pp. 2531-2537, 2010.

[10] J. G. Sun, G. J. Huang, and J. L. LUO, “Modified concept similarity algorithm Computer Engineering and Applications,” Computer Engineering and Applications, vol. 45, no. 5, pp. 154-160, 2009.

[11] T. Kanungo, D. M. MountD, N. S. Netanyahu, C. D. Piatko, R. Siverman, and A. Y. Wu, “A Local Search Approximation Algorithm for K-Means Clustering,” Computational Geometry, vol. 28, no. 2-3, pp. 89-112, 2004.

[12] S. Y. Wu and Y. Y. Wu, “Chinese and English Word Similarity Measure Based on Chinese WordNet,” Journal Zheng Zhou Univ. (Nat. Sci. Ed.), vol. 42, no. 6, pp. 66-71, 2010.

[13] T. Lu, H. Wang, and H. L. Yao, “K-nearest neighbor Chinese text categorization algorithm based on center documents,” Computer Engineering and Applications, vol. 47, no. 2, pp. 127-132, 2011.

[14] D. M. Deng, J. Z. Long, and X. Z. Yin, “A clustering algorithm based on structured Web document,” Journal of Central South University (Science and Technology), vol. 41, no. 10, pp. 1871-1875, 2010.

[15] Y. Z. Qu, W. Hu, and G. Cheng, “Constructing Virtual Document for Ontology Matching;” Proc. of 15th International Conference on World Wide Web (WWW 06), pp. 23-31, 2006.

 ISBN: 978-1-61208-419-0

Where am I?
A fast multidimensional point location test and its applications

Tanja Clees, Martin Hüttemann, Igor Nikitin, Lialia Nikitina, Daniela Steffes-lai
Department of High Performance Analytics
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
Tanja.Clees| Martin.Huettemann |Igor.Nikitin|Lialia.Nikitina|Daniela.Steffes-lai@scai.fraunhofer.de

***Abstract*—We present our recent advances in RBF metamodeling of multidimensional data. A rapid point location test in multidimensional data cloud distinguishing the cases of interpolation and extrapolation is proposed. A linear program detecting a containment of the probe point in a convex hull of the dataset is formulated, simplex and interior point solution methods are tested in different dimensionalities and densities of the data cloud, extensions of the approach to nonconvex datasets and various acceleration strategies are implemented. The resulting software module is integrated in our optimization tool DesParO and applied to several real life problems from the fields of automotive industry and chemical engineering.**

Keywords-complex computing in application domains, automotive industry, chemical engineering, energy optimization.

I. INTRODUCTION

Numerical simulations define a mapping $y=f(x): R^n \rightarrow R^m$ from an n-dimensional space of simulation parameters to an m-dimensional space of simulation results. E.g. in automotive crash test simulation the dimensionality of simulation parameters x is moderate (n~10-30), while simulation results y are dynamical fields sampled on a large grid, typically containing $\sim 10^6$ nodes and ~100 time steps, resulting in values of $m \sim 10^8$. High computational complexity of crash test models restricts the number of simulations available for analysis (typically $Nexp < 10^3$), and this number shall be as small as possible. Metamodeling is an approximation technique allowing efficient representation of these large datasets for the purpose of data analysis, robust optimization and real time visualization. The metamodeling naturally involves in the analysis the uncertainties in optimization variables and other control parameters influencing the simulation. In practice a metamodeling with radial basis functions (RBF) is often used, i.e. representation of the form:

$$f(x)=\sum_{i=1..\ Nexp} c_i \Phi(|x-x_i|), \quad (1)$$

where x_i are the points with known function values $y_i = f(x_i)$. A suitable choice for the RBF is the multi-quadric function $\Phi(r)=(b^2+r^2)^{1/2}$, which provides non-degeneracy of interpolation matrix $\Phi_{ij}=\Phi(|x_i-x_j|)$ for all finite datasets of distinct points and all dimensions [1]. The result can be written in a form of weighted sum $f(x)=\sum_i w_i(x) y_i$, with the weights

$$w_i(x)=\sum_j \Phi^{-1}_{ij} \Phi(|x-x_j|). \quad (2)$$

RBF interpolation can be extended by adding polynomial terms, allowing reconstructing exactly polynomial (including linear) dependencies and generally improving precision of interpolation. Adaptive sampling and hierarchy of metamodels with appropriate transition rules are used for further precision improvement [2]. RBF metamodel is directly applicable for interpolation of high dimensional bulky data, e.g. complete simulation results can be interpolated at a rate linear in the size of data, and even faster in combination with PCA-based dimensional reduction techniques [3]. The precision can be controlled via the cross-validation procedure: the data point is removed, data are interpolated to this point and compared with the actual value at this point, which for an RBF metamodel leads to a direct formula [4]

$$err_i = f_{interpol}(x_i)-f_{actual}(x_i) = -c_i/(\Phi^{-1})_{ii}. \quad (3)$$

Metamodeling performed at controlled precision can replace simulation results or real experimental data in computationally intensive procedures, such as optimization, parameter studies, stochastic analysis (i.e. determination of probability distributions by Monte Carlo techniques). In our previous work [2-9] we use RBF metamodeling for solution of various applied problems. For correct metamodeling one should permanently control that the interpolated point is located inside the boundaries of the data cloud. In optimization process and data analysis it is necessary to avoid extrapolation or at least to warn the user about it, to ensure correct functionality of the metamodel. While it is straightforward for hypercube alike designs of experiments, the problem becomes non-trivial for more complex shapes. In this case, a bounding box provides a too loose estimator of the cloud.

The key contribution of this work is to provide metamodel with a precise and fast indicator whether a point belongs to a multidimensional data cloud. In general, testing whether a probe point belongs to a region is a classical (Where am I?) point location problem. We start with its

 ISBN: 978-1-61208-419-0

special type (Am I in a convex hull of data points?) and compare the efficiency of various algorithms. High dimensionality of the space (dim≥10) excludes the usage of standard tools (e.g. Qhull), since they perform direct segmentation of the region to simplices and the number of simplices explodes at high dimensions. For example, tessellation of an n-dimensional cube [10] produces the number of simplices $\geq (n+1)^{(n-1)/2}$, requiring memory [bytes] $\geq 4(n+1)^{(n+1)/2}$, which already at n=16 corresponds to 100 GB of memory.

Being reformulated as a linear program (LP), location test works also for high dimensions and can be implemented efficiently using state-of-the-art algorithms. Furthermore, restricting the test to a controlled neighbourhood of the probe point, the method can be extended to nonconvex clouds of data points. In the next sections we compare the performance of various algorithms for LP-based location test, discuss their extension to nonconvex data clouds and apply them to real life problems of chemical engineering and automotive industry.

II. LP-BASED LOCATION TEST

Let x_i be data points in R^n, conv$\{x_i\}$ – their convex hull, conv'$\{x_i\}$= conv$\{x_i\} \setminus \partial$ conv$\{x_i\}$ – the convex hull without its boundary, $x^* = \sum x_i$ /Nexp – center of mass, x – probe point. We consider non-degenerate datasets: conv'$\{x_i\} \neq \varnothing$. Let's define containment flag as Cflag(x)=0 iff (x_i-x) are contained in a half-space, i.e. there exists a separating hyperplane with a normal $v \neq 0$ satisfying $(x_i-x,v) \geq 0$ for all i, and Cflag(x)=1 otherwise, see Fig. 1a. In other words, x is located inside conv'$\{x_i\}$ when Cflag=1 and outside of conv'$\{x_i\}$ when Cflag=0. Practically, Cflag can be determined using the following linear program:

Algorithm LP(x,{x_i}):
find max(x*-x,v) at $(x_i-x,v) \geq 0$ for all i;
if solution is max=+∞ then Cflag=0;
else (v=0) Cflag=1.

For numerical solution of LP we compare two algorithms: simplex [11] and interior point [12], applying them to a randomly filled n-dim cube. The results of comparison are collected in Table 1. We see that for the given problem simplex method performs by factor 10^2-10^3 better than interior point method.

LP-algorithm can be accelerated by combining with two simple tests. Let's consider a (generally nonconvex) region Ω uniformly filled with random points $\{x_i\}$. Let BB=[min(x_i),max(x_i)] be bounding box of the dataset. Let $r=|x_i-x_j|$ be inter-point distance and <r> - its average. One should better use quasi-random (low discrepancy) sequences with narrow r-histograms, see Fig. 2. Let B be a ball around the probe point x with radius c<r>, where c is an empiric safety factor (e.g c=3 for rnd2D, c=2 for Sobol2D). If this ball does not contain points from $\{x_i\}$, then x is surely outside Ω, see Fig. 1b.

The following algorithms can serve as simple conservative containment tests:

Algorithm BBox(x,{x_i}):
if x∈BB then Cflag=1; else Cflag=0.

Algorithm BT(x,{x_i}):
find $\{x_i\} \cap B$;
if empty then Cflag=0; else Cflag=1.

Let's define a local convex hull as LCH=conv'($\{x_i\} \cap B$). Differently from the global convex hull (GCH) it considers only a small portion of data points and is computationally much faster:

Algorithm LCH(x,{x_i}):
call BBox(x,$\{x_i\}$);
if Cflag=1:
call BT(x,$\{x_i\}$);
if Cflag=1:
call LP(x,$\{x_i\} \cap B$).

If x is outside LCH (Cflag=0), it is also outside Ω. If x is inside LCH (Cflag=1), it is either inside Ω or at a distance ~<r> from its boundary, see Fig. 1c. LCH provides more tight location test then GCH and BT. Performance of LCH is slower than BBox/BT but much faster than GCH, since only a small portion of data points N/Nexp is contained in B. The number of data points passed to LCH can be additionally controlled by selecting N'>n nearest data points in B. Performance of BBox is O(n), BT is O(Nexp*n). LP-algorithms have theoretical worst case complexity exponential for simplex method and polynomial for ipopt, while in practice they show much better performance, especially at reduced N, see Table 1.

TABLE I. BENCHMARK OF SIMPLEX AND INTERIOR POINT METHODS IN LP-BASED LOCATION TEST*.

Nexp	dim	Simplex (ms)	Ipopt (ms)
100	10	0.020	18
250	10	0.106	42
500	10	0.470	77
100	20	0.096	36
250	20	0.181	67
500	20	0.260	258
100	30	0.070	54
250	30	0.206	310
500	30	0.546	266

*Resulting Cflag values for both methods are always identical.
Timing per solution @ 3 GHz Intel i7 CPU.

 ISBN: 978-1-61208-419-0

III. APPLICATIONS

The better performing method (simplex LCH) has been integrated in our software tool for design parameter optimization (DesParO [13]). It uses RBF metamodel to represent dependence between design parameters and optimization criteria. The graphical user interface allows to change interactively the parameters and to see immediately the variation of the criteria. Constraints can be set e.g. maximizing one objective and minimizing the other, in this way the constrained and multiobjective optimization problems can be investigated. A graphical representation of interdependencies between parameters and criteria allows to find most influencing parameters and most sensitive criteria. Also, the uncertainties of metamodeling found with cross-validation procedure are shown (the red bars under criteria sliders). Fig. 3 and Fig. 5 show screenshots of interface of DesParO tool in application to several industrial problems.

The first application is safety optimization in Audi B-pillar crash test. The model of B-pillar shown on Fig. 3 contains ten thousand nodes, 45 timesteps. Two parameters are varied representing thicknesses of two layers composing a part of a B-pillar, comprising 101 simulations. The purpose is to find a Pareto-optimal combination of parameters simultaneously minimizing the total mass of the part and crash intrusion in the contact area. Fig. 3 shows the optimization problem loaded in the DesParO Metamodel Explorer, where design variables (thicknesses1,2) are presented on the bottom image at the left and design objectives (intrusion and mass) at the right. First, the user imposes constraints on design objectives, trying to minimize intrusion and mass simultaneously, as indicated by red ovals. As a result, “islands” of available solutions become visible along the axes of design variables. Exploration of these islands by moving corresponding sliders shows an optimal configuration, shown on Fig. 3 (bottom). For this configuration both constraints on mass and intrusion are satisfied. For every criterion also its tolerance is shown corresponding to 1-sigma confidence limits, as indicated by horizontal bars under the corresponding slider as well as +/- errors in the value box. This indication allows satisfying constraints with 3-sigma (99.7%) confidence, as shown on the image. The Geometry Viewer, shown at the top of Fig. 3, allows to inspect the optimal design in full details. LCH algorithm has been used to indicate location of the probe point in the data cloud, interpolation (Cflag=1) and extrapolation (Cflag=0).

The second application is scatter analysis in Ford Taurus crash test simulation. As shown in our previous work [3], crash test simulations possess a random component, related to physical and numerical instabilities of the underlying simulation model. It can be triggered by microscopic variations of design variables (e.g. thicknesses of various parts in car body) and by the numerical process itself (e.g. propagation of round-off errors or by random scheduling in multiprocessing simulation). These microscopic sources are then amplified by inherent physical instabilities of the model related e.g. to buckling, contact phenomena or material failure. Stochastic analysis is used to track the sources of scatter, to reconstruct causal chains and to identify hidden parameters describing chaotic behavior of the model. The crash model shown on Fig. 4 contains 1 million nodes, 32 timesteps, 25 simulations. The simulation results have been processed by a method of temporal clustering [3], which decomposes the whole scatter in the model over a system of basis functions: $s=\sum_i \Psi_i c_i$. Every basis function Ψ_i is associated with elementary random process (bifurcation) and possesses a typical conic profile, originating from the corresponding bifurcation point and propagating forward in time. Fig. 4 right shows one of the major bifurcations, corresponding to a fold on the floor of the vehicle. In total, 15 bifurcation points have been identified, representing statistically independent sources of scatter. In this way the dimensionality of the problem is reduced to 15 variables (c-coefficients) completely describing stochastic behaviour of the model. One should only take care that reconstruction of scatter does not go beyond the boundary of simulated data cloud. The point location test by LCH algorithm at these values of dimensions (Nexp=25, dim=15) has typical performance 27 mks per query (inside BBox).

The third application is energy optimization for Polycarbonate production at Bayer MaterialScience AG. The purpose was to minimize consumption of various energy media, including electric power, gas, steam, water, etc. The optimization has been performed on the base of experimental measurements, collecting data from sensors on several production lines and comprising 1-year detailed records of plant performance. Optimization is performed in 10-dimensional parameter space sampled with ~8000 data points, using our software tool for design parameter optimization (DesParO). RBF interpolation has been used for continuous optimization, see Fig. 5. Optimization parameters (par1, par2, …) are displayed on the left, optimization objectives on the right: partial energy consumptions (E01, E02,…), total energy cost and production range used as a constraint. LCH algorithm has controlled location of point in the data cloud, ensuring applicability of the metamodel. Fig. 5 shows on the top an optimal point inside the data cloud (Cflag=1, interpolation), while the bottom image shows the middle point of bounding box located outside of the given data cloud (Cflag=0, extrapolation). Typical performance was 0.3 ms per query inside BBox, where complete LCH algorithm was involved; while outside BBox only O(n) part of the algorithm was active, showing an extremely fast performance of 20 ns per query.

Point location tests in all applications have been performed on 3 GHz Intel i7 CPU with 8 GB RAM.

IV. CONCLUSION

RBF metamodeling of multidimensional data supplemented by a rapid point location test for

 ISBN: 978-1-61208-419-0

distinguishing the cases of interpolation and extrapolation is presented. A linear program detecting a containment of the probe point in a convex hull of the dataset is formulated. Comparing simplex and interior point methods for solution of this particular linear program, we see that simplex method performs by factor 10^2-10^3 better than interior point one. A concept of local convex hull allows to extend the approach to nonconvex datasets, while simple geometrical containment tests are used to accelerate the algorithm. The resulting software module is integrated in our optimization tool DesParO and applied to real life problems from the fields of automotive industry and chemical engineering. In a typical problem with dimension 10 and number of data points ~8000 the performance of location test was 0.3 ms per query (inside BBox) and 20 ns per query (outside BBox).

REFERENCES

[1] M.D.Buhmann, Radial Basis Functions: theory and implementations, Cambridge University Press, 2003.

[2] G. van Bühren, T.Clees, N.Hornung, L.Nikitina, Aspects of adaptive hierarchical RBF metamodels for optimization, Journal of computational methods in sciences and engineering JCMSE 12 (2012), Nr.1-2, pp.5-23.

[3] T.Clees, I.Nikitin, L.Nikitina, C.-A.Thole, Analysis of bulky crash simulation results: deterministic and stochastic aspects, in N.Pina et al (Eds.): Simulation and Modeling Methodologies, Technologies and Applications, AISC 197, Springer 2012, pp.225-237.

[4] T.Clees, I.Nikitin, L.Nikitina, Nonlinear metamodeling of bulky data and applications in automotive design, in M. Günther et al (eds), Progress in industrial mathematics at ECMI 2010, Mathematics in Industry (17), Springer, 2012, pp. 295-301.

[5] T.Clees, I.Nikitin, L.Nikitina, S.Pott, Efficient Quantile Estimators for River Bed Morphodynamics, in Tören et al, Eds, Proc. of SIMULTECH 2013, 3rd International Conference on Simulation and Modeling Methodologies, Technologies and Applications, Reykjavik, Iceland 29 - 31 July, 2013, pp.737-743, SCITEPRESS, 2013.

[6] T. Clees, N. Hornung, I. Nikitin, L. Nikitina and D. Steffes-lai, Multi-objective Optimization and Stochastic Analysis in Focused Ultrasonic Therapy Simulation, in Tören et al, Eds, Proc. of SIMULTECH 2013, 3rd International Conference on Simulation and Modeling Methodologies, Technologies and Applications, Reykjavik, Iceland 29 - 31 July, 2013, pp.43-48, SCITEPRESS, 2013.

[7] Clees, T., Nikitin, I., Nikitina, L., Kopmann, R., Reliability analysis of river bed simulation models, in J.Herskovits, Ed., CDROM Proc. EngOpt 2012, 3rd Int. Conf. on Engineering Optimization, Rio de Janeiro, Brazil, no. 267.

[8] T. Clees, N. Hornung, I. Nikitin, L. Nikitina, D. Steffes-lai, S. Klimenko, Focused ultrasonic therapy planning: Metamodeling, optimization, visualization, J. Comp. Sci. 5 (6), Elsevier 2014, pp 891-897.

[9] T. Clees, I. Nikitin, L. Nikitina, S. Pott, Quasi-Monte Carlo and RBF Metamodeling for Quantile Estimation in River Bed Morphodynamics, in Obaidat, M.S. et al (Eds.), Simulation and Modeling Methodologies, Technologies and Applications, Advances in Intelligent Systems and Computing 319, Springer 2014, pp 211-222.

[10] A.Glazyrin, Lower bounds for the simplexity of the n-cube. Discrete Math. 312 (2012), no. 24, 3656-3662.

[11] G.Dantzig, Linear programming and extensions. Princeton University Press and the RAND Corporation, 1963.

[12] A.Wächter, Short Tutorial: Getting Started With Ipopt in 90 Minutes, IBM Research Report, 2009.

[13] DesParO: the tool for design parameter optimization, www.scai.fraunhofer.de/en/business-research-areas/high-performance-analytics/products/desparo.htm

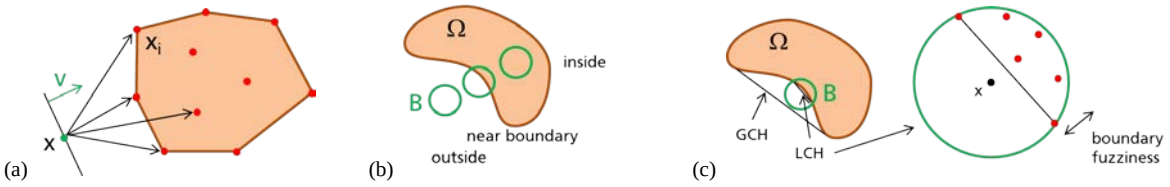


Figure 1. (a) to algorithm LP: data cloud, probe point and separating hyperplane; (b) to algorithm BT: data cloud and test ball; (c) to algorithm LCH: definition of local convex hull.

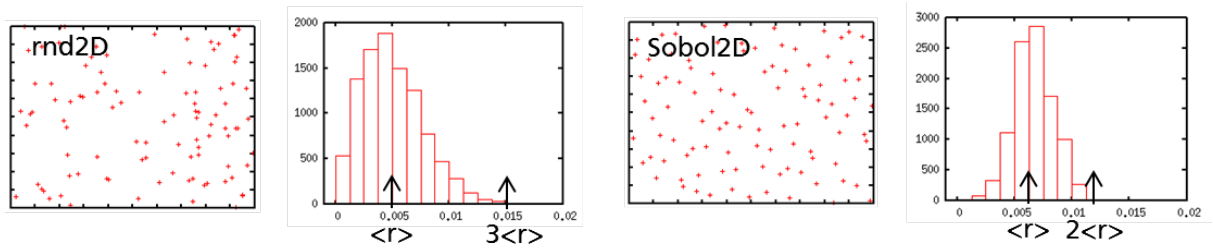


Figure 2. Pseudo-random (rnd2D) and quasi-random (Sobol2D) filling of a square and the corresponding r-histograms.

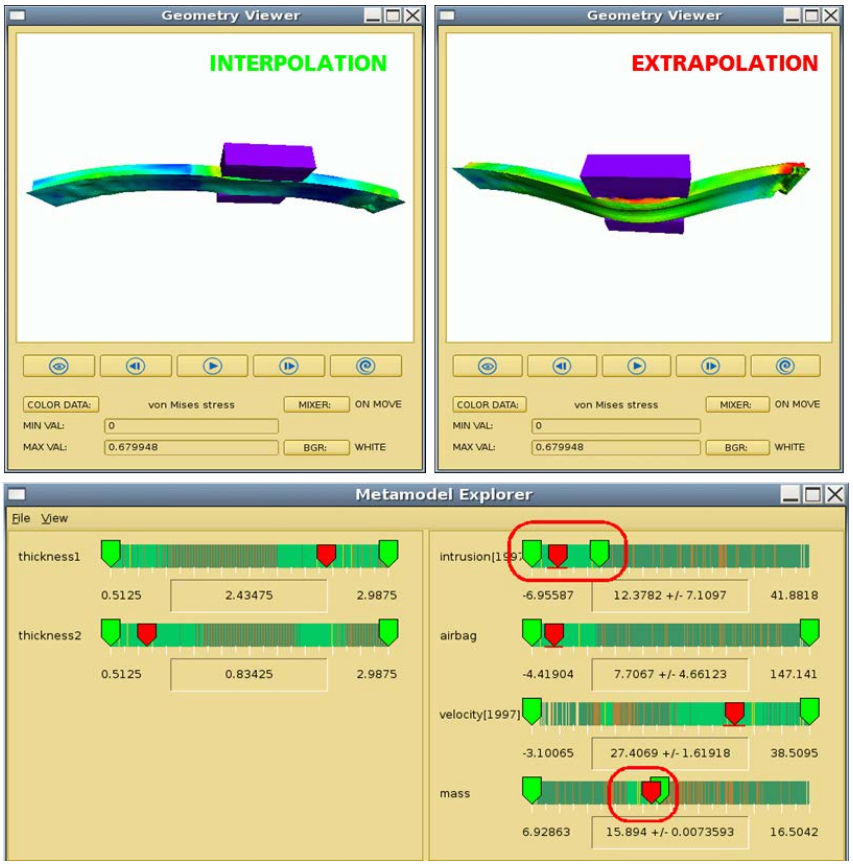


Figure 3. Metamodeling of B-pillar crash test simulation results. Point location test is used to distinguish cases of interpolation (point inside data cloud, in the image on the left) and extrapolation (point outside the data cloud, on the right). At the bottom: optimal design in DesParO Metamodel Explorer. Simulation model: courtesy of Audi.

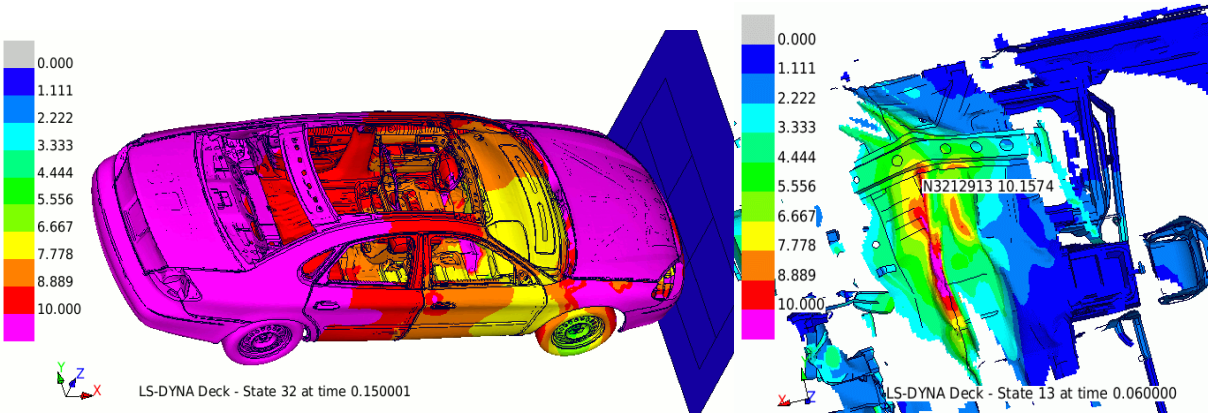


Figure 4. Scatter analysis in automotive crash test simulation. On the left: original scatter in mm. On the right: closeup to the main bifurcation, the source of scatter. Data courtesy of Ford.

 ISBN: 978-1-61208-419-0

Figure 5. Energy optimization for Polycarbonate production. Data courtesy of Bayer MaterialScience AG.

RBF-metamodel driven multiobjective optimization and its application in focused ultrasonic therapy planning

Tanja Clees, Nils Hornung, Igor Nikitin, Lialia Nikitina, Daniela Steffes-lai
Department of High Performance Analytics
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
Tanja.Clees|Nils.Hornung|Igor.Nikitin|Lialia.Nikitina|Daniela.Steffes-lai@scai.fraunhofer.de

***Abstract*—We consider bi-objective optimization problem from noninvasive tumor therapy planning. The therapy uses magnetic resonance tomography for the location of the target region and focused ultrasound for the destruction of tumor cells. Experimentally validated physical models are used to construct numerical simulation including nonlinear wave propagation, absorption in soft tissue, heat transfer and a hierarchical structure of the biological materials. The resulting cumulative thermal dose inside the target region should be maximized, providing a maximal level of tumor destruction, while the thermal dose outside the target region should be minimized, to decrease the influence to healthy organs. Metamodeling with radial basis functions is used for continuous representation of optimization objectives. The problem possesses nonconvex Pareto front. Detection of nonconvex Pareto fronts is especially difficult, this is a point where many simple algorithms fail. In this paper we consider different approaches to this problem: sequential linear programming (SLP), sequential quadratic programming (SQP) and generic 1- or 2-phase nonlinear programming (NLP). We show the ability of the algorithms to process such case and compare the efficiency of different approaches.**

Keywords-complex computing in application domains; medical computation and graphics; advanced computing in simulation systems; advanced computing for statistics and optimization.

I. INTRODUCTION

Focused ultrasonic therapy is a noninvasive therapy using magnetic resonance tomography for identification of tumor volume and focused ultrasound for the destruction of tumor cells. Numerical simulation becomes an important step for the therapy planning. Efficient methods for the focused ultrasonic simulation have been presented in paper [1]. It uses a combination of Rayleigh-Sommerfeld integral for near field and angular spectrum method for far field computations, which allows determining the pressure field in heterogeneous tissue. The bioheat transfer equation is used to determine the temperature increase in therapy region. Thermal dose is defined according to cumulative equivalent minutes metric (CEM, [2]) or Arrhenius model [3] as a functional of temperature-time dependence in every spatial point in therapy region. These methods have been accelerated by GPU based parallelization and put in the basis of software FUSimlib (www.simfus.de), developed by our colleagues at Fraunhofer Institute for Medical Image Computing.

3D visualization is used for interpretation of simulation results, in particular, for detailed inspection of MRT images (magnetic resonance tomography), corresponding material model and spatial distribution of the resulting thermal dose, see Fig. 1. Stereoscopic 3D visualization in virtual environments based on modern 3D-capable beamers with DLP-Link technology (Digital Light Processing), described in more details in [4] is especially suitable for this purpose. Such commonly available beamers do not require special projection screens and can turn every regular office to a virtual laboratory providing full immersion into the model space. We use 3D visualization software Avango (www.avango.org), an object-oriented programming framework for building applications of virtual environments. Our interactive application overlays three voxel models: original MRT sequence, material segmentation and resulting thermal dose. The user can mix the voxel models together, interactively changing their levels of transparency, set breathing phase, cut the model with a clipping plane, etc.

For continuous representation of optimization objectives from a discrete set of simulation results we use metamodeling with radial basis functions (RBF). It represents the interpolated function f(x) as a linear combination of special functions Φ() depending only on the distance to the sample points x_k:

$$f(x) = \sum_{k=1..Nexp} c_k \Phi(|x-x_k|) \qquad (1)$$

The coefficients c_k in (1) can be found from known function values in sample points $f(x_k)$ by solving a moderately sized linear system with a matrix $\Phi_{kn} = \Phi(|x_k-x_n|)$. A suitable choice for the RBF is the multi-quadric function $\Phi(r)=(b^2+r^2)^{1/2}$, which provides nondegeneracy of interpolation matrix for all finite datasets of distinct points and all dimensions [5]. RBF interpolation can be extended by adding polynomial terms, allowing reconstructing exactly polynomial (including linear) dependencies and generally improving precision of interpolation. Adaptive sampling and hierarchy of metamodels with appropriate transition rules are used for further precision improvement. RBF metamodel is directly applicable for interpolation of high dimensional bulky data, e.g. complete simulation results can be interpolated at a rate linear in the size of data, and even faster

 ISBN: 978-1-61208-419-0

in combination with PCA-based dimensional reduction techniques. The precision can be controlled via cross-validation procedure. So enhanced RBF metamodel is a part of our software tool for design parameter optimization DesParO [6-8].

The objective of therapy planning is a maximization of thermal dose inside the target zone (TDin) and minimization of thermal dose outside (TDout). As usual in multi-objective optimization, the optimum is not an isolated point but a hypersurface (Pareto front, [9]) composed of points satisfying a tradeoff property, i.e. none of the criteria can be improved without simultaneous degradation of at least one other criterion. Thus, for a two-objective problem, the Pareto front is a curve on the plot (TDin, TDout) bounding the region of possible solutions. Efficient methods have been previously developed for determining the Pareto front.

The simplest way is to convert multi-objective optimization to single objective one, by linearly combining all objectives into a single target function

$$t(x)= \sum w_i f_i(x) \qquad (2)$$

with user-defined constant weights w_i. Maximization of the target function (2) gives one point on Pareto front, while varying the weights allows to cover the whole Pareto front. In this way only convex Pareto fronts can be detected, because nonconvex Pareto fronts produce not maxima but saddle points of the target function.

There are methods applicable also for nonconvex Pareto fronts. Nondominated set algorithm (NDSA) finds a discrete analogue of Pareto front in a finite set of points. For two points f and g in optimization criteria space the first one is said to be dominated by the second one if $f_i \le g_i$ holds for all i=1..Ncrit. A point f belongs to nondominated set if there does not exist another point g dominating f. There is a recursive procedure [10] finding all nondominated points in a given finite set. The drawback of the algorithm is an extremely large number of samples necessary to populate multidimensional regions for good approximation of Pareto front.

Normal boundary intersection method (NBI) [11] provides a good heuristics for sampling of Pareto front. The idea is to find individual minima of objectives, to construct their convex hull, to sample it e.g. with Delaunay tessellation, to build normals in tesselation points and finally to intersect them with the boundary of par → crit mapping. The approach has problems e.g. at Ncrit>2, when non-Pareto points or not all Pareto points are covered, or if the number of minima >Ncrit, when several local Pareto fronts can be mixed together.

Meanwhile, practical applications just require an elementary algorithm performing a local improvement of current design towards the optimum. Being iterated such algorithm proceeds towards Pareto front. For definiteness, an improvement direction in the space of objectives can be fixed, e.g. every step all objectives are improved by a given increment. The algorithm stops when the further improvement in the given direction is not possible. Normally it happens when the solver reaches the Pareto front. Convex or nonconvex Pareto fronts can encounter and the algorithm should work equally efficient for both. The improvement can also stop on a non-Pareto boundary point. In this case it is allowed to return the other point on Pareto front, which does not necessarily belong to the original improvement direction.

In further sections we consider different approaches for this algorithm: sequential linear programming (SLP), sequential quadratic programming (SQP) and generic 1- or 2-phase nonlinear programming (NLP). We also consider a question of scalarization, i.e. a possibility to reformulate the multiobjective optimization problem as constrained optimization with a single objective, which allows to employ available NLP solvers for its solution.

II. USING SEQUENTIAL LINEAR PROGRAMMING

Linearizing the mapping y=f(x) using Jacoby matrix $J_{ij}=\partial y_i/\partial x_j$, let's consider a polyhedron of possible variations

$$\Pi_\varepsilon: \Delta y=J\Delta x\ ,\ \Delta y \ge \varepsilon > 0\ ,\ -\delta \le \Delta x \le \delta, \qquad (3)$$
$$x_{min} \le x+\Delta x \le x_{max}\ ,\ y_{min} \le y+\Delta y \le y_{max}$$

Here we require that all criteria Δy are improved, parameter variations Δx are bounded in a trust region $[-\delta,\delta]$ for linear approximation, while parameters and criteria satisfy bounding box or other polyhedral restriction in xy-space. By requiring in (3) that a maximally possible improvement of criteria in Π_ε is achieved, we formulate a linear program which can be solved e.g. by simplex method [12] and repeated sequentially:

Algorithm SLP:

Solve LP: max ε, s.t. $(\Delta x,\Delta y)\in\Pi_\varepsilon$

Repeat steps $x+\Delta x \rightarrow x$ until convergence.

The algorithm terminates at Pareto front, where no further improvements are possible.

Property: in general position LP-optimum is achieved in corners of polyhedron Π_ε.

E.g. $\Delta y=\varepsilon$ correspond to linear trajectories in y-space, $|\Delta x|=\delta$ correspond to linear trajectories in x-space. Therefore, the method tends to generate linear trajectories in certain projections.

SLP above is formulated for the case dim(x)=dim(y). At dim(x)<dim(y) multiobjective problem is ill defined, i.e. full dimensional regions in parameter space become Pareto equivalent. At dim(x)>dim(y) there are unstable directions from Ker(J): $J\Delta x=0$, i.e. there are Δx not influencing Δy. These directions can be suppressed by additional condition $J_\perp\Delta x=0$, where $J_\perp$ is orthogonal complement to J, constructed e.g. with Gram-Schmidt algorithm.

Example: let's consider a fold transform: $|y|=2|x|/(1+|x|^2)$ shown on Fig. 3 for 2D case. An upper right arc corresponds to a global Pareto front (PF) max y_1,y_2. There is also a

 ISBN: 978-1-61208-419-0

degenerate local PF at $y_{1,2}$=-0, corresponding to an image of $x_{1,2}$=-∞.

SLP algorithm generates trajectories shown by red lines on Fig. 3, in x-space in the left column and in y-space in the right column. The algorithm reconstructs correctly both global and local PF, shown by blue points on the images. The bottom closeups demonstrate piecewise linear trajectories described above. Particularly, there is a dashed linear trajectory in y-space tending to non-Pareto part of the boundary (nPF), which at a certain moment switches from $\Delta y=\varepsilon$ corner to $|\Delta x|=\delta$ corner, becomes curved and finally stops at PF.

III. USING SEQUENTIAL QUADRATIC PROGRAMMING

Polyhedron Π_0 is defined as above (with ε=0). Let v be a fixed search direction in y-space, ε is a constant. The following quadratic program [15] tries to perform $\Delta y=\varepsilon v$ steps if possible in Π_0:

Algorithm SQP:

Solve QP: min $||\Delta y-\varepsilon v||^2$, s.t. $(\Delta x,\Delta y)\in\Pi_0$

Repeat steps $x+\Delta x \rightarrow x$ until convergence.

Property: in general position QP-optimum can be achieved inside Π_0, in corners of Π_0 or on edges/faces of Π_0.

In the first case $\Delta y=\varepsilon v$ linear trajectories will be generated in y-space, in the second case $|\Delta x|=\delta$ linear trajectories will be generated in x-space, in the third case the trajectories become nonlinear.

IV. USING 1-PHASE NONLINEAR PROGRAMMING

Nonlinear target function in the form $t(x)=\sum w_i\, crit_i^p$ under certain conditions can detect nonconvex Pareto fronts. Here the target function is represented by a scaled L_p-norm with weights $w_i \geq 0$, $\sum w_i = 1$. Fig. 2 left shows level curve for 2D target function for different p. One has a straight line at p=1, a quadric at p=2, a superquadric at p>2 and a corner at p=∞.

Property: nonlinear target function can be used to detect nonconvex PF, if the curvature of the level curve exceeds the curvature of PF.

Also at higher dimensions, considering the level set (LS) tangent to PF, performing Taylor expansions of LS and PF: $z=u^TMu+o(u^2)$, where u,z are respectively parallel and normal components to a common tangent hyperplane to LS and PF, and requiring $z_{LS}\geq z_{PF}$, one can reformulate the property above as positive definiteness for the difference of curvature matrices M_{LS}-M_{PF}.

Note that L_∞ =max is applicable in any case (minmax method [13]), but the corresponding NLP will be nonsmooth. Practically, one can use large finite p, it is also convenient to normalize y_i in [0,1] and take a log of target function for numerical stability. In this way one achieves so called scalarization of multiobjective optimization, i.e. conversion of multiobjective problem to a single objective one. As a result, the problem becomes solvable with standard NLP-solvers, e.g. ipopt [14]. Here one can impose any additional constraints, e.g. require that $y(x) \leq c$. By putting $c=y_0$ one ensures that the result is better in all criteria than a starting point and finds only a corresponding segment of PF. One can also leave c=∞ and vary w_i to cover the whole PF.

Algorithm NLP1(c):

minimize $t(x)=\log \sum (w_i y_i)^p$, s.t. $y(x) \leq c$.

V. USING 2-PHASE NONLINEAR PROGRAMMING

The following algorithm combines the concepts of linear search from NBI and optimization of nonlinear target function. The first phase performs the linear search in a given direction v in y-space towards PF and the second phase tries to perform further improvement (if possible). The problem is solvable with two calls to ipopt.

Algorithm NLP2:

NLP2.1: maximize t, s.t. $y(x)=y_0+tv$; result y_1;

NLP2.2: call NLP1(y_1); result y_2.

Properties (see Fig. 2 right):

if $y_1 \in$ PF, phase 2 quits immediately;

if $y_1 \in$ non PF boundary, trajectory is bounced to PF.

In NLP2.2 not the whole PF is targeted, but a smaller part ΔPF possessing better criteria than y_1. Here one can use smaller p, while even for too curved PF the result y_2 will be still better than y_0 and y_1.

VI. APPLICATION IN FOCUSED ULTRASONIC THERAPY PLANNING

A generic workflow for ultrasonic therapy simulation has been described in our paper [16]. Numerical simulation with FUSimlib software uses 512 x 512 x 256 voxel grid. Ultrasound has been focused in the center of the target zone for the neutral breath state. The result after 10 seconds of exposure time (200 steps x 0.05sec) has a form of spatial distributions of pressure amplitude, temperature and thermal dose. Fig. 1 top-right shows a typical result for thermal dose on slice 97/256 near the focal point. The frequency of transducer is taken as optimization parameter controlling focused ultrasonic therapy simulation. The other one, initial particle speed, is proportional to an acoustic intensity emitted by the transducer [1]. As optimization objectives the thermal dose inside and outside the target zone have been defined as sums of the thermal dose over corresponding voxels, ∑TDin / ∑TDout. The variation range of optimization parameters was regularly sampled with 25 simulations, from which 16 fall in the region of interest, shown on Fig. 4 by red points. RBF metamodel constructed

 ISBN: 978-1-61208-419-0

on simulation results is used to oversample the region by green points, from which discrete method NDSA selects Pareto front, shown by blue points. We see that Pareto front is of nonconvex type. Magenta lines show application of three continuous methods described above. The trajectories generated by SLP and NLP2 coincide in every detail. Even bouncing from non-PF boundary works similar, although the mechanisms of this bouncing are different. NLP1 with p=8 and w_1=0.01,0.15,0.27,0.5,0.99 produces the other set of trajectories. Table I shows a summary of problem characteristics. SLP provides the best performance for the given application case. On the other hand, NLP is easier for integration with existing scalar solvers. In NLP class, NLP1 is faster than NLP2 for bounced trajectories, otherwise NLP2 is faster. Numerically NLP2 (with small p) is less singular than NLP1 (with large p) and therefore is more robust for detection of strongly curved Pareto fronts.

TABLE I. BI-OBJECTIVE OPTIMIZATION IN FOCUSED ULTRASONIC THERAPY PLANNING, PROBLEM CHARACTERISTICS

Parameter bounds: frequency 0.25…0.75 MHz ini.speed 0.23...0.282 m/s	**Timing per solution** @ 3GHz Intel i7:
Criteria bounds: ∑TDin 0…3000 eq.min ∑ TDout 0…6000 eq.min	SLP 7ms NLP1 16ms NLP2 13ms+12ms

VII. CONCLUSION

Several algorithms of continuous multiobjective optimization applicable for detection of nonconvex Pareto fronts have been discussed: sequential linear programming (SLP), sequential quadratic programming (SQP) and generic 1- or 2-phase nonlinear programming (NLP1,2). Scalarization, i.e. reformulation of the multiobjective optimization problem as constrained optimization with a single objective, allows to employ available NLP solvers for its solution. The algorithms have been applied to realistic test case in focused ultrasonic therapy planning. In the given problem SLP possesses the best performance, while NLP is easier for integration with existing scalar solvers. NLP1 is faster than NLP2 for bounced trajectories, otherwise NLP2 is faster. Numerically NLP2 is less singular than NLP1 and is therefore more robust for detection of strongly curved Pareto fronts. All these optimization methods provide real-time performance necessary for interactive planning of focused ultrasonic therapy.

REFERENCES

[1] J.Georgii et al, Focused Ultrasound - Efficient GPU Simulation Methods for Therapy Planning, in Proc. Workshop on Virtual Reality Interaction and Physical Simulation VRIPHYS, Lyon, France, 2011, J. Bender, K. Erleben, and E. Galin (Editors), Eurographics Association 2011, pp. 119-128.

[2] S.Nandlall et al, On the Applicability of the Thermal Dose Cumulative Equivalent Minutes Metric to the Denaturation of Bovine Serum Albumin in a Polyacrylamide Tissue Phantom. Proc. 8th Int. Symp. Therapeutic Ultrasound (AIP), 1113:205-209, 2009.

[3] J.A.Pearce, Relationships between Arrhenius models of thermal dose damage and the CEM 43 thermal dose. Energy-based Treatment of Tissue and Assessment V (Proceedings of SPIE), editor: Ryan, T.P., 7181, 2009.

[4] T. Clees et al, Focused ultrasonic therapy planning: Metamodeling, optimization, visualization, J. Comp. Sci. 5 (6), Elsevier 2014, pp 891-897.

[5] M.D.Buhmann, Radial Basis Functions: theory and implementations, Cambridge University Press, 2003.

[6] G. van Bühren et al, Aspects of adaptive hierarchical RBF metamodels for optimization, Journal of computational methods in sciences and engineering JCMSE 12 (2012), Nr.1-2, pp.5-23.

[7] T.Clees et al, Analysis of bulky crash simulation results: deterministic and stochastic aspects, in N.Pina et al (Eds.): Simulation and Modeling Methodologies, Technologies and Applications, AISC 197, Springer 2012, pp.225-237.

[8] T.Clees, I.Nikitin, L.Nikitina, Nonlinear metamodeling of bulky data and applications in automotive design, in M. Günther et al (eds), Progress in industrial mathematics at ECMI 2010, Mathematics in Industry (17), Springer, 2012, pp. 295-301.

[9] M.Ehrgott, X.Gandibleux (Eds.), Multiple criteria optimization: state of the art annotated bibliographic surveys, Kluwer 2002.

[10] H. T. Kung, F. Luccio, and F. P. Preparata. On finding the maxima of a set of vectors. Journal of the ACM, 22(4):469–476, 1975.

[11] G.Eichfelder, Parametergesteuerte Lösung nichtlinearer multikriterieller Optimierungsprobleme, Friedrich–Alexander-Universität Erlangen–Nürnberg, Dissertation 2006.

[12] G.Dantzig, Linear programming and extensions. Princeton University Press and the RAND Corporation, 1963.

[13] D.Müller-Gritschneder, Deterministic Performance Space Exploration of Analog Integrated Circuits considering Process Variations and Operating Conditions, Technische Universität München, Dissertation 2009.

[14] A.Wächter, Short Tutorial: Getting Started With Ipopt in 90 Minutes, IBM Research Report, 2009.

[15] R.Fletcher, Practical Methods of Optimization, Wiley 2000.

[16] T. Clees et al, Multi-objective Optimization and Stochastic Analysis in Focused Ultrasonic Therapy Simulation, in Proc. of SIMULTECH 2013, pp.43-48, SCITEPRESS, 2013.

 ISBN: 978-1-61208-419-0

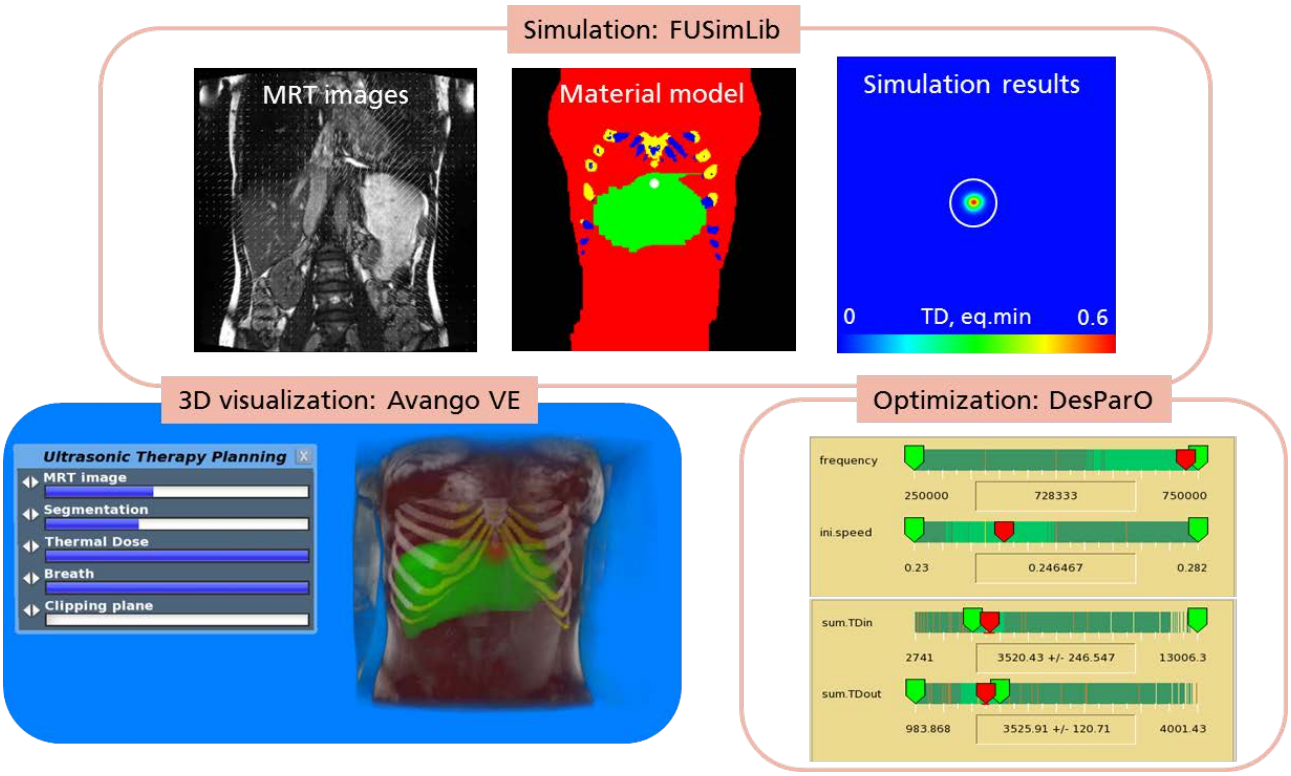


Figure 1. Focused ultrasonic therapy planning and its software components.

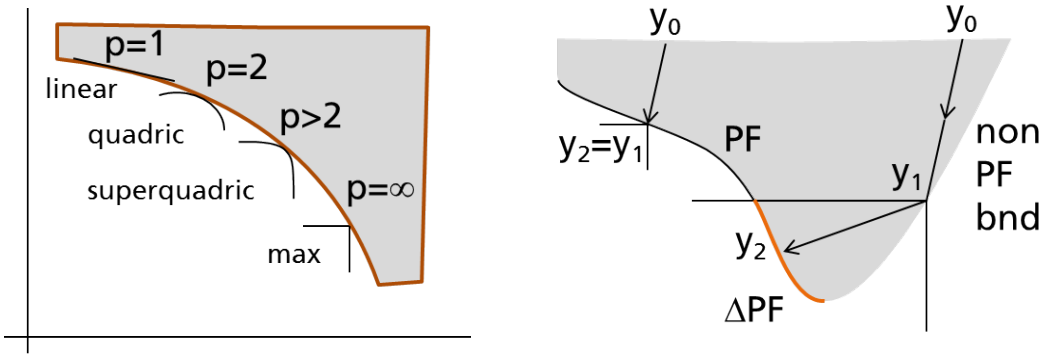


Figure 2. Scalarization of multiobjective optimization problem. On the left: algorithm NLP1; on the right: algorithm NLP2.

 ISBN: 978-1-61208-419-0

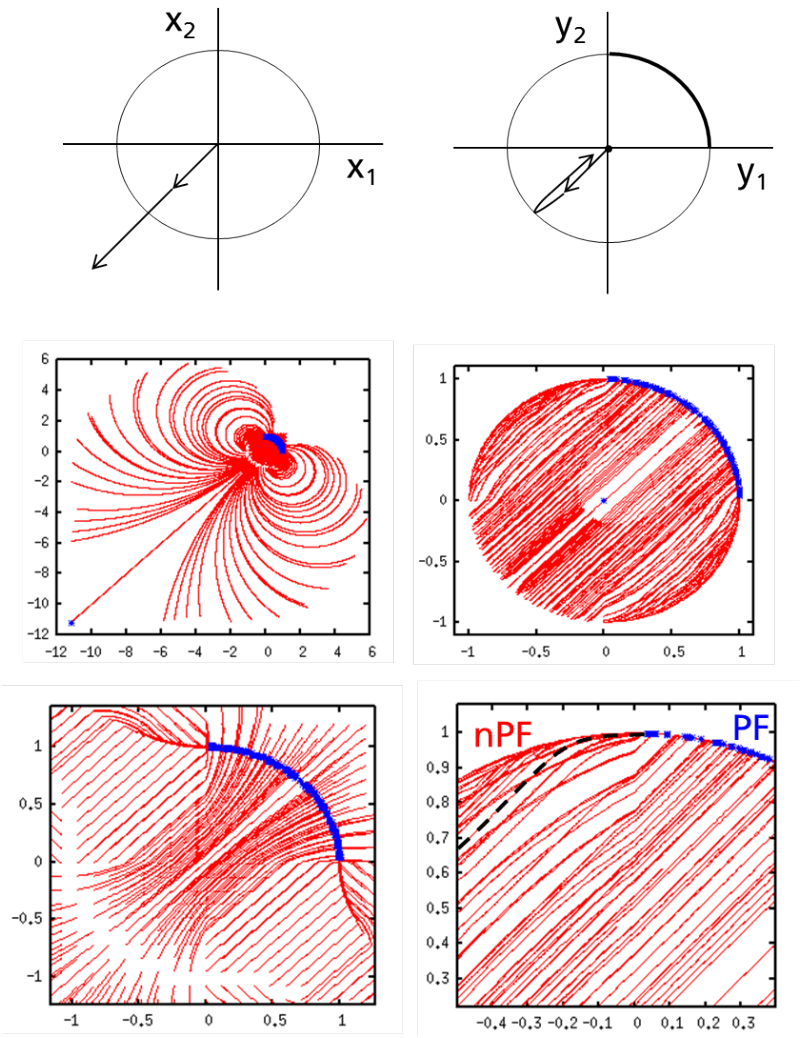


Figure 3. Pareto front detection for 2D fold transform.

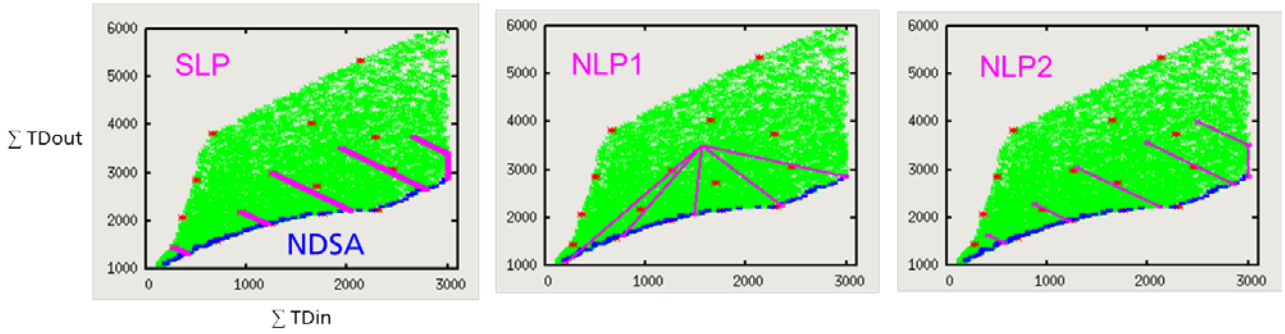


Figure 4. Nonconvex Pareto front in focused ultrasonic therapy planning, comparison of different methods.

A Lot Scheduling Problem on a Single Machine with Indivisible Orders

Wen-Hung Kuo and Dar-Li Yang
Department of Information Management
National Formosa University
Yunlin, Taiwan, R.O.C.
Email:{whkuo,dlyang}@nfu.edu.tw

***Abstract* — In this paper, a lot scheduling problem on a single machine with indivisible orders is studied. The objective is to minimize the total completion time of all orders. We use a binary integer programming approach to solve this problem. The binary integer programming approach with run time limit is considered as one heuristic method. As compared to a lower bound, it turns out the average performance of the method is really good.**

Keywords- lot scheduling; single machine; total completion time; indivisible order; integer programming

I. INTRODUCTION

Generally, there are two main production processes in a production system, that is, continuous production and batch production. Here, we are interested in batch production. In the literature, there are two categories of batch scheduling problems. One is batch scheduling with divisible batch sizes. Naddef and Santos [1] studied a single machine problem with batching jobs. The objective is to minimize the total completion times. They showed that the greedy algorithm solves the problem if jobs are all of one type. They also provided a heuristic for the problem with various job types. Coffman et al. [2] considered a single machine job shop in which subassemblies of two different types are made and then assembled into products. They provided an efficient algorithm for minimizing the total flow time of the products. Dobson et al. [3] considered batch jobs in the multiple-machine scheduling problem. The objective is to minimize the mean flow times. They proposed an efficient algorithm for computing the optimal solution for single product case. Hou et al. [4] studied a lot scheduling problem with orders which can be split. Orders are grouped into lots and then processed. The objective is to minimize the total completion time of all orders. They showed that this problem can be solved in polynomial time.

The other is batch scheduling with indivisible batch sizes. Shallcross [5] studied a problem of batching identical jobs on a single machine. He presented an algorithm to minimize the sum over all jobs of the batched completion times. Mosheiov et al. [6] addressed a classical minimum flow-time, single-machine, batch-scheduling problem. They introduced a simple rounding procedure for Santos and Magazine's solution [7], which guarantees optimal integer batches. Mor and Mosheiov [8] studied an identical parallel-machine scheduling problem with identical job processing times and identical setups. They showed that the solution is given by a closed form, consisting of identical decreasing arithmetic sequences of batch sizes on the different machines.

In a factory, products are usually made according to customers' orders. This production approach is called MTO (make to order). Since different orders may contain different quantities, two production strategies are applied in the batch production, especially when the lot size of the batch production is fixed. Also, in this particular situation, the production time of each lot is fixed no matter how many quantities in the lot. Therefore, from the viewpoint of efficiency, one order may be divided into several lots to fill up each lot. The study presented by Hou et al. [4] is based on this viewpoint. However, from the viewpoint of management, one order is not divided into different production lots because the products of the same order are finished at the same time and then delivered to the customer. Based on this viewpoint, in this paper, we study the problem given by Hou et al. [4] but orders are restricted to be indivisible.

The rest of the paper is organized as follows. In the second section, a description of the problem is given. Next, the integer programming formulation is provided. Computational experiments are given in the fourth section. Final section is the conclusion.

II. PROBLEM DESCRIPTION

There are n orders (O_i, i = 1, 2, …, N) to be grouped into lots and then be processed on a single machine. Each order has its own size (σ_i, i = 1, 2, …, N). The size of each order is no more than one lot's capacity (K). On top of that, every order is indivisible. It means products of each individual order have to be processed in the same lot. The orders in the

 ISBN: 978-1-61208-419-0

same lot have the same processing time (*t*). Therefore, all orders in the same lot have the same completion time.

The machine can handle at most one lot at a time and cannot stand idle until the last lot assigned to it has finished processing. The objective is to minimize the total completion time ($\sum C_{O_i}$) of all orders. Thus, using the three-field notation, this scheduling problem is denoted by $1/lot, indivisible/\sum C_{O_i}$.

III. INTEGER PROGRAMMING FORMULATION

The problem is conjectured to be NP hard [9]. Therefore, we use the following binary integer programming approach to solve this problem.

Let $X_{i[q]} = 1$ if the *i*th order is assigned to the *q*th lot, and 0 otherwise. Since the processing time of each lot is *t*, the completion times of the first lot, the second one, etc., are *t*, 2*t*, …, respectively. Thus, the total completion time of all orders is $t\sum_{q=1}^{N}\sum_{i=1}^{N} X_{i[q]}q$. Then, a binary integer programming (BIP) formulation to solve the proposed problem is developed as follows.

$$\text{Minimize} \quad t\sum_{q=1}^{N}\sum_{i=1}^{N} X_{i[q]}q \tag{1}$$

$$\text{Subject to} \sum_{q=1}^{N} X_{i[q]} = 1 \quad i = 1,2,...,N \tag{2}$$

$$\sum_{i=1}^{N} \sigma_i X_{i[q]} \le K \quad q = 1,2,...,N \tag{3}$$

$$X_{i[q]} \in \{0,1\} \quad i = 1,2,...,N,\ q = 1,2,...,N \tag{4}$$

The objective is to minimize the total completion time of all orders which is shown in (1). Equation (2) ensures that each order is only assigned to one lot. Equation (3) limits the total sizes of orders that are assigned to the same lot to the lot capacity (*K*). Finally, (4) guarantees that variable $X_{i[q]}$ is either 0 or 1.

IV. COMPUTATIONAL EXPERIMENTS

The above binary integer programming approach can solve the proposed problem, but it is time-consuming when it comes to a large problem. Considering the efficiency of the BIP, the run time limit of the BIP is set to 3600 seconds. Also, in order to evaluate the performance of the BIP, it is tested in the computational experiments which are conducted based on the following parameter set.

Order number *N* is equal to 20, 30, 40, 50, 60, 70, 80, 90, 100.

Lot capacity *K* is equal to 15, 30.

Order size σ_i is uniformly distributed over [1,5], [1,10] ($\sigma_i \overset{d}{=} \text{U}(1,5)$, $\sigma_i \overset{d}{=} \text{U}(1,10)$)

There are $9 \times 2 \times 2 = 36$ problem types. For each problem type, 30 test problems are generated. Each test problem is solved by BIP and LP, respectively. BIP and LP are solved by using a computer program coded in LINGO 11.0 with 4GB of memory available for working storage, running on a personal computer Intel(R) Core(TM) i7-2600 CPU @ 3.4GHz. To evaluate the performance of the computational results, we have to come up with a lower bound (LB) and then compare these percentage errors ($100*(BIP - LB)/LB$) in different test problems.

Obviously, one lower bound can be obtained from the solution of a variant of the original problem by changing the original problem to the one in which orders are divisible and can be processed in different lots. Therefore, we only need to change (4) as follows.

$$X_{i[q]} \ge 0 \quad i = 1,2,...,N,\ q = 1,2,...,N \tag{4'}$$

Then, since the problem becomes a Linear Programming (LP) problem, we take much less time to solve the problem than the original one. The lower bound is also tight because the solutions of the original problem and its variant can happen to be the same (integers).

The average and maximal percentage errors of each problem type for the BIP solutions and also the number of optimal solutions obtained within 3600 seconds are shown in the following table.

From Table 1, we have the following observations:

(1) For $N = 20$, the optimal solutions for all generated test problems can be found within 3600 seconds.

(2) The larger the lot capacity is or the smaller the order size range is, the more optimal solutions you can obtain within the run time limit.

(3) Average percentage errors of all problem types are less than 2.5, it means that the performance of the binary integer programming with run time limit is really good, especially, in the problem type with parameters *K* = 30 and $\sigma_i \overset{d}{=} \text{U}(1,5)$.

 ISBN: 978-1-61208-419-0

(4) Most maximal percentage errors of all problem types are less than 4.5, it implies that the BIP performs well in most test problems, even in the worst situations.

TABLE1. COMPUTATIONAL RESULTS.

	σ_i =1~5, K =15			σ_i =1~5, K =30			σ_i =1~10, K =15			σ_i =1~10, K=30		
	Error (%)			Error (%)			Error (%)			Error (%)		
N	avg	max	opt. no.	avg	max	opt. no.	avg	max	opt. no.	avg	max	opt. no.
20	0	0	30	0	0	30	0	0	30	0	0	30
30	0	0	30	0	0	30	1.05	9.23	26	0	0	30
40	0.38	3.44	26	0	0	30	1.30	10.08	25	0.15	4.46	29
50	1.08	3.04	13	0	0	30	1.31	10.86	24	0.13	3.93	29
60	1.38	2.77	7	0	0	30	1.34	7.52	23	0.28	3.19	27
70	1.59	2.37	2	0	0	30	1.06	10.24	25	0	0	30
80	1.23	2.38	6	0.56	1.71	18	1.49	8.01	22	0.09	2.63	29
90	0.91	2.54	10	0.57	1.44	15	1.30	7.27	23	0.18	3.15	28
100	1.47	2.06	1	0.86	1.32	3	2.46	6.55	17	0.08	2.47	29

σ_i : order size, K: lot capacity, N: order number

However, some of them in the problems with parameters K = 15 and $\sigma_i \overset{d}{=} \mathrm{U}(1,10)$ are greater than 10, even though their average percentage errors are less than 2.5. The performance of BIP in such problems is not robust. Therefore, it is worthwhile to come up with other heuristics with better performance.

V. CONCLUSION

In this paper, we studied a single-machine lot scheduling problem with indivisible orders. The problem is conjectured to be NP-hard. Therefore, a binary integer programming approach is given to solve the problem. Considering the efficiency of the BIP, the run time limit is set. Also, compared to the lower bound, it turns out the average performance of the BIP within run time limit is really good for all test problems. The maximal percentage errors of the BIP with run time limit are a little greater than 10 in one situation. Therefore, it is worthwhile to find other heuristics with better performance in the future.

ACKNOWLEDGEMENT

This research was supported in part by the National Science Council of Taiwan, Republic of China, under grant number NSC-102-2221-E-150-043-MY2.

REFERENCES

[1] D. Naddef and C. Santos, “One-pass batching algorithms for the one-machine problem,” Discrete Applied Mathematics, vol. 21, pp. 133–45, 1988.

[2] E. D. Coffman, A. Nozari, and M. Yannakakis, “Optimal scheduling of products with two subassemblies on single machine,” Operations Research, vol. 37, pp. 426–36, 1989.

[3] G. Dobson, U. D. Karmarkar, and J. L. Rummel, “Batching to minimize flow times on parallel heterogeneous machines,” Management Science, vol. 35, pp. 607–13, 1989.

[4] Y. T. Hou, D. L. Yang, and W. H. Kuo, “Lot scheduling on a single machine,” Information Processing Letters, vol. 114, pp. 718–722, 2014.

[5] D. F. Shallcross, “A polynomial algorithm for a one machine batching problem,” Operations Research Letters, vol. 11, pp. 213–218, 1992.

[6] G. Mosheiov, D. Oron, and Y. Ritov, “Minimizing flow-time on a single machine with integer batch sizes,” Operations Research Letters, vol. 33, pp.497–501, 2005.

[7] C. Santos and M. Magazine, “Batching in single operation manufacturing systems,” Operations Research Letters, vol. 4, pp. 99–103, 1985.

[8] B. Mor and G. Mosheiov, “Batch scheduling of identical jobs on parallel identical machines,” Information Processing Letters, vol. 112, pp. 762–766, 2012.

[9] M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-completeness. W. H. Freeman, New York, 1979.

 ISBN: 978-1-61208-419-0

Proposal of Clustering Approach Based on Structural Mechanics: An Application of Multi-Dimensional Truss

Kazuyuki Hanahara and Yukio Tada

Graduate School of System Informatics
Kobe University
Kobe, Japan
Email: {hanahara, tada}@cs.kobe-u.ac.jp

***Abstract*—We deal with an approach of clustering based on structural mechanics. Rupture of multi-dimensional truss due to a universal repulsive force is adopted as the process of clustering. The structural behavior of multi-dimensional truss is formulated. The feasibility of the proposed approach is examined and demonstrated by a number of calculation examples.**

Keywords–Clustering; Multi-Dimension; Truss; Tidal force.

I. INTRODUCTION

Clustering is one of the important processes for data management, especially in the case of pattern identification and recognition [1]. Several methods of clustering have been proposed [2][3]. There are typical approaches such as hierarchical algorithms e.g., the group average method and the Ward method and partitioning algorithms e.g., the k-means method and its families. Another approach such as based on PCA (Principal Compoment Analysis) has also been developed and studied [4]. These approaches having being developed so far are basically based on some mathematical or geometrical viewpoints. The authors see that these approaches are somewhat artificial, in the sense that the clustering processes are controlled by one or more mathematical parameters that are intentionally determined.

In the current study, we deal with an approach of clustering based on structural mechanics. Taking account of the mechanical characteristics of the target data set, a clustering of somewhat natural manner is considered to be possible. One of the significant problem is that structural systems dealt with in structural mechanics are two or three-dimensional entities, but the data set to be clustered can be an entity of higher dimensional space. In the case of truss structural system, however, it is possible to formulate the structural mechanical characteristics such as the stiffness matrix even in the case of a truss structure of four or higher dimensional space.

In this article, we regard the data elements to be clustered as the truss nodes. We develop the formulation of stiffness matrix of truss structure of general dimension. Rupture of the truss structure due to universal repulsive force is calculated; the obtained separated parts are recognized as the clusters.

In Section II, a general formulation of the nodal stiffness matrix of multi-dimensional truss is introduced. The developed clustering procedure is explained in Section III. Some preliminary example calculation results are demonstrated and discussed in Section IV and Section V gives the conclusion and future work.

II. MULTI-DIMENSIONAL TRUSS

A truss structure consists of a number of truss nodes and truss members connecting them. We denote the truss nodal positions vector as $\mathbf{X} = [\mathbf{x}_1^T, \cdots, \mathbf{x}_N^T]^T$, where $\mathbf{x}_n$ is a D-dimensional vector corresponding to the nth element of the data set to be clustered. Truss member connections are denoted as $\mathcal{C} = \{c_1, \cdots, c_M\}$, where $c_m = \{c_{m0}, c_{m1}\}$ and c_{m0} and c_{m1} correspond to the two truss nodes connected by the mth truss member.

A. Geometrical Relation

We denote the truss member lengths vector as $\mathbf{L} = [l_1, \cdots, l_M]^T$. Each of the member lengths is given as the Euclidean distance between the corresponding nodes expressed as

$$l_m = \left[(\mathbf{x}_{c_{m1}} - \mathbf{x}_{c_{m0}})^T(\mathbf{x}_{c_{m1}} - \mathbf{x}_{c_{m0}})\right]^{(1/2)} \tag{1}$$

For all of the truss member lengths and the truss nodal positions, (1) can be collected and expressed in the following form:

$$\mathbf{L} = \mathbf{L}(\mathbf{X}) \tag{2}$$

The total differential of (2) is given as

$$d\mathbf{L} = \frac{\partial \mathbf{L}}{\partial \mathbf{X}} d\mathbf{X} \tag{3}$$

which is obtained as the collection of the total differential of (1) expressed as

$$dl_m = \frac{\mathbf{x}_{c_{m1}} - \mathbf{x}_{c_{m0}}}{l_m} d\mathbf{x}_{c_{m1}} - \frac{\mathbf{x}_{c_{m1}} - \mathbf{x}_{c_{m0}}}{l_m} d\mathbf{x}_{c_{m0}} \tag{4}$$

B. Stiffness Matrix

In the case of linear elastic model with small deformation, the strain energy U of the entire truss under deformation is expressed as

$$U = \sum_{m=1}^{M} \frac{1}{2} k_m r_m^2 = \frac{1}{2} \mathbf{R}^T \mathbf{K}_L \mathbf{R} \tag{5}$$

where k_m and r_m are the stiffness and the elastic change in length of the mth truss member, $\mathbf{R} = [r_1, \cdots, r_M]^T$ is the member deformation vector and $\mathbf{K}_L = \mathrm{diag}[k_1, \cdots, k_M]$ is the member stiffness matrix. Since we deal with the case of small deformation, the nodal displacement vector $\mathbf{U} = [\mathbf{u}_1^T, \cdots, \mathbf{u}_N^T]^T$ and the member deformation vector have the following linear relation referring to (3):

$$\mathbf{R} = \frac{\partial \mathbf{L}}{\partial \mathbf{X}} \mathbf{U} \tag{6}$$

 ISBN: 978-1-61208-419-0

Substituting (6) for (5), we obtain

$$U = \frac{1}{2}\mathbf{U}^T \left(\frac{\partial \mathbf{L}}{\partial \mathbf{X}}\right)^T \mathbf{K}_L \frac{\partial \mathbf{L}}{\partial \mathbf{X}}\mathbf{U} = \frac{1}{2}\mathbf{U}^T \mathbf{K}_X \mathbf{U} \tag{7}$$

where

$$\mathbf{K}_X = \left(\frac{\partial \mathbf{L}}{\partial \mathbf{X}}\right)^T \mathbf{K}_L \frac{\partial \mathbf{L}}{\partial \mathbf{X}} \tag{8}$$

is the nodal stiffness matrix of the given truss structure.

III. Clustering Based on Rupture of Truss

The clustering is dealt with in terms of rupture of the truss structure corresponding to the given data elements, which experiences a kind of universal repulsive force.

A. Generating Member Connection

In the current study, we deal with two types of member connections among the truss nodes corresponding to the data elements. One is the full-connection type, where all of the combination of two nodes are connected by truss members. The other is a simplex-connection type, where truss members are connected to form appropriate simplices in the given dimensional space. Figure 1 shows such examples of member connections in 2D space. For the simplex-connection, we adopt the truss members in the order of length, from the shortest, among all the possible connections. The full-connection type is easy to generate; however, it is not natural from the viewpoint of truss structural system. On the other hand, the computational time to generate the simplex-type connection is not insignificant in the case of higher dimensional space; however, the obtained member connection is natural and more reasonable as a truss structural system.

Mechanical characteristics of a truss structure also depends on the stiffness of the members. We use the following relation to determine the stiffness of the truss members taking into account the weight values assigned to the data elements:

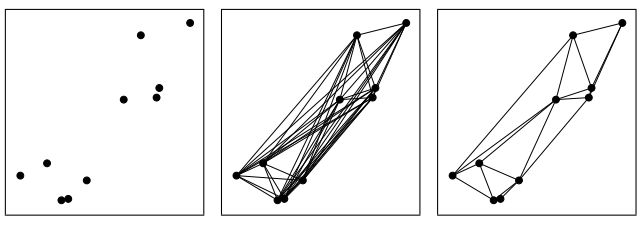
$$k_m = C_S \frac{1}{l_m^S} w_{c_{m0}} w_{c_{m1}} \tag{9}$$

where C_S is an adequate constant, S is the distance-evaluation parameter and $w_{c_{m0}}$ and $w_{c_{m1}}$ are the weight values of the data element nodes to be connected by the truss member m. The equation indicates that in the case of higher order of S, the connection strength of two data elements decreases rapidly in accordance with their distance.

B. Universal Repulsive Force

As the force to deform and rupture the truss structure, we introduce a universal repulsive force among the truss nodes corresponding to the data elements. The force between any two nodes is expressed as

$$\mathbf{f}_{ni} = C_R(\mathbf{x}_n - \mathbf{x}_i) d_{ni}^{R-1} w_n w_i, \quad d_{ni} = \|\mathbf{x}_n - \mathbf{x}_i\| \tag{10}$$



(a) Data elements (b) Full-type (c) Simplex-type

Figure 1. Two types of member connection. (2D case example)

where C_R is an adequate constant, w_n and w_i are the mass assigned to the two nodes corresponding to the weight values assigned to the data elements, and R is the parameter which denotes the nature of the repulsive force. On the basis of the introduced universal repulsive force between two nodes written as (10), the nodal force vector is obtained as follows:

$$\mathbf{F}_X = [\mathbf{f}_1^T, \cdots, \mathbf{f}_N^T]^T, \quad \mathbf{f}_n = \sum_{i=1}^{N} \mathbf{f}_{ni} \ (i \neq n) \tag{11}$$

It should be noted that the repulsive nodal force pattern for the case $R = 1$ is corresponding to the so-called tidal force, though the obtained force is not uni-directional but multi-directional.

C. Clustering Procedure

We deal with two-group clustering of the given data elements $\mathbf{x}_1, \cdots, \mathbf{x}_N$ having the weight values $w_1, \cdots, w_N$. The process is performed as follows:

Step 0 Generate truss member connections $\mathcal{C}$.

Step 1 Calculate the nodal stiffness $\mathbf{K}_X$ and the nodal force $\mathbf{F}_X$.

Step 2 Solve the stiffness equation

$$\mathbf{K}_X \mathbf{U} = \mathbf{F}_X \tag{12}$$

taking into account the condition of rigid body motion and obtain the nodal displacement $\mathbf{U}$.

Step 3 Calculate the member deformation $\mathbf{R}$ by (6) and obtain the magnitude of the member strain as follows:

$$\epsilon_m = |r_m / l_m| \quad (m = 1, \cdots, M) \tag{13}$$

Note that the obtained value immediately corresponds to the magnitude of the member stress, since we assume uniform structural material.

Step 4 Delete the truss member connections in the order of the magnitude of strain until the truss corresponding to the data set is separated into two parts.

Characteristic of the proposed clustering approach is determined by the type of member connection, the distance evaluation parameter S and the repulsive force parameter R. The constants C_S and C_R do not affect the clustering result.

IV. Example Calculations

Since this is a study still at a preliminary stage, we conduct example calculations in order to examine the feasibility of the proposed clustering approach. Influence on the clustering results of the types of truss member connection as well as the introduced two parameters is also discussed.

For each of the data sets to be clustered, the nth data element in D-dimensional space, $\mathbf{x}_n = [x_{n(1)}, \cdots, x_{n(D)}]^T$, is generated by the following equation for $i = 1, \cdots, D$ as

$$x_{n(i)} = \begin{cases} +\dfrac{D_G}{2} + D_R & (n = 1, \cdots, \dfrac{N}{2}) \\ -\dfrac{D_G}{2} + D_R & (n = \dfrac{N}{2} + 1, \cdots, N) \end{cases} \tag{14}$$

where N is the number of data elements, D_G is the assumed gap parameter between the two cluster centers and D_R is a random number. In the following calculation examples, the number of elements is $N = 50$, the gap parameter is $D_G = 0.6$ and the random number D_R is assumed to have the normal distribution of standard deviation 0.2.

 ISBN: 978-1-61208-419-0

A. Evaluation of Member Connection Type

First, we examine the difference between the results based on the two types of member connections. We use $S = 2$ and $R = 1$ a priori as the two parameters in the following examples. The distance evaluation parameter $S = 2$ is selected from the clustering point of view, which indicates that the thickness of truss member connection between two data elements becomes thinner in accordance with their distance. The repulsive force parameter $R = 1$ is selected because the value corresponds to a really existing repulsive force, that is the tidal force, although this is not unidirectional.

Figure 2 shows typical clustering results based on the two types of member connections. Figures (a) and (b) are the examples respectively based on the full-connection and the simplex-connection of truss members. Figures (a-1) and (b-1) are the same data elements to be clustered and Figures (a-2) and (b-2) are the clustering results. The data elements of the obtained major cluster are depicted as a filled circle (•) and the others are depicted as an empty circle (○). Both obtained results shown in Figures (a-2) and (b-2) are similar and considered to be acceptable; however, small difference is observed with the two data elements at the right-hand side of the center.

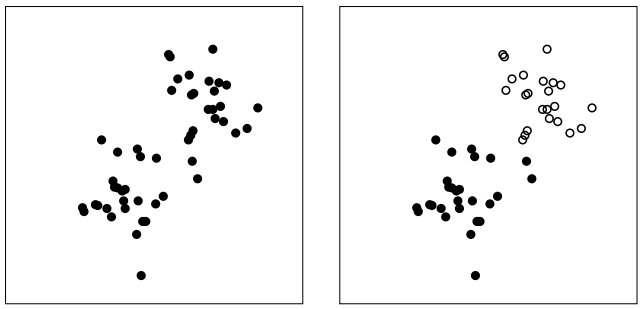
Figure 3 shows another clustering results based on the data elements shown in Figure (a-1). On the basis of the full-connection type truss, the first clustering result and the succeeding second clustering result respectively shown in Figures (a-2) and (a-3) are considered insufficient. The succeeding third clustering result shown in Figure (a-4) does not seem natural. On the basis the simplex-connection type truss, the first clustering result shown in Figure (b-1) can also be regarded as insufficient; however, the result having a cluster of single data element is unacceptable from the clustering point of view. The succeeding second result shown in Figure (b-2) is considered to be reasonable.

The examples shown in Figures 2 and 3 are typical results. Another clustering calculation examples also show similar tendency. In the following calculation examples, we use the simplex-connection type truss members for the clustering.

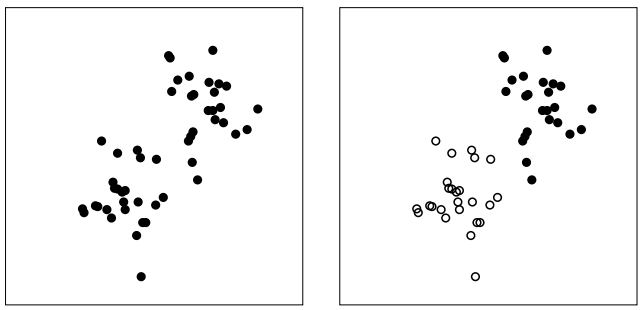
B. Evaluation of Two Introduced Parameters

We examine the influence of two parameters S and R. Another data set is adopted this time, since no significant difference with the parameters is observed in the clustering results based on the two data sets adopted in the previous examples. The case shown in Figure 4 is adopted as the reference. Figure (a) is the adopted data set for the parameter evaluation and Figure (b) is the clustering result based on $S = 2$ and $R = 1$. Clusters of this data set are comparably unclear; however, the clustering result shown in Figure (b) is considered to be reasonable.

Figure 5 shows the clustering results based on different values of S in the case of $R = 1$. Figures (a), (b) and (c) respectively based on $S = 0$, $S = 1$ and $S = 3$ exhibit different results. It can be observed for all the cases that the

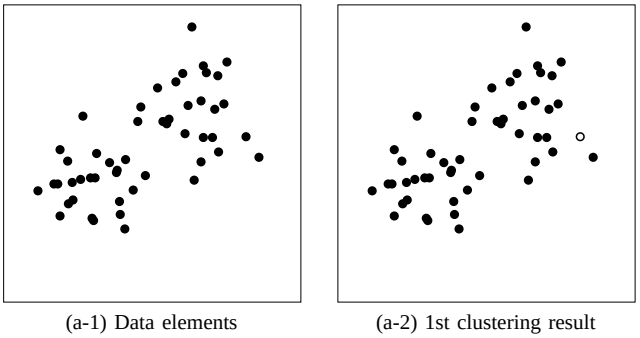


(a-1) Data elements (a-2) Clustering result
(a) Based on full-connection truss

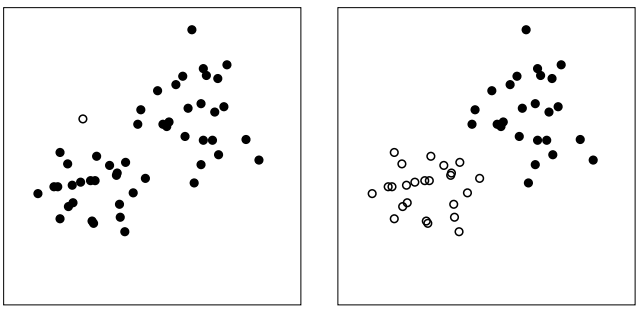


(b-1) Data elements (b-2) Clustering result
(b) Based on simplex-connection truss

Figure 2. Evaluation of connection-type based on data set A. ($S = 2, R = 1$)

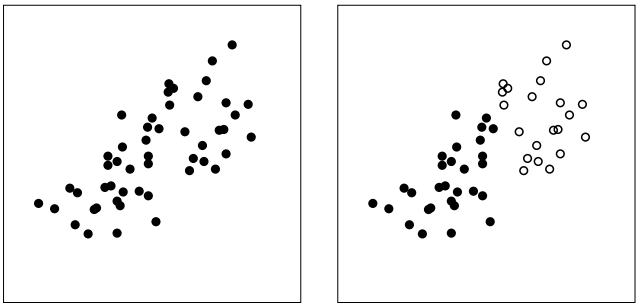


(a-1) Data elements (a-2) 1st clustering result (a-3) 2nd clustering result (a-4) 3rd clustering result
(a) Based on full-connection truss



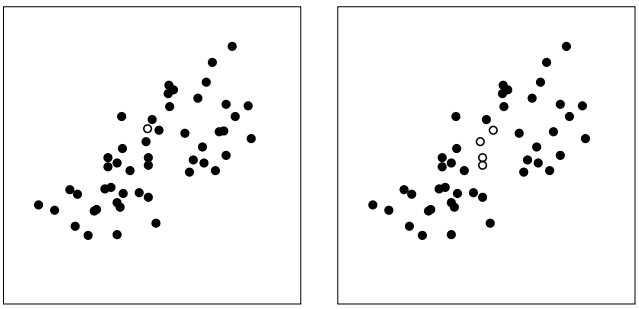
(b-1) 1st clustering result (b-2) 2nd clustering result
(b) Based on simplex-connection truss

Figure 3. Evaluation of connection-type based on data set B. ($S = 2, R = 1$)



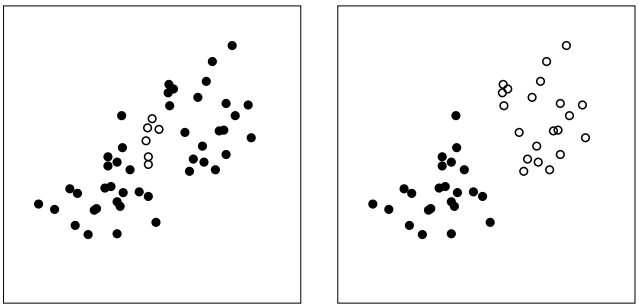
(a) Data elements (b) Clustering result

Figure 4. Reference clustering result based on data set C. ($S = 2, R = 1$)



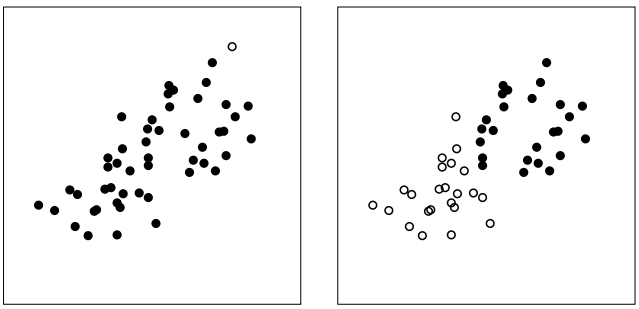
(a-1) 1st clustering result (a-2) 3rd clustering result

(a) $S = 0$



(b-1) 1st clustering result (b-2) 2nd clustering result

(b) $S = 1$

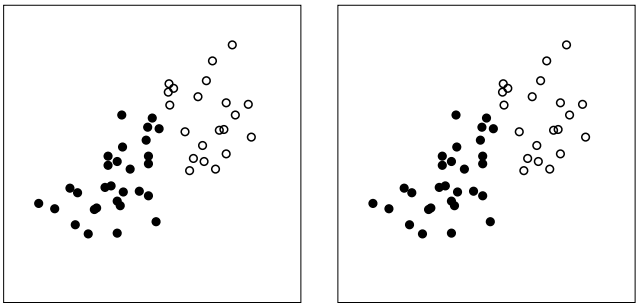


(c-1) 1st clustering result (c-2) 3rd clustering result

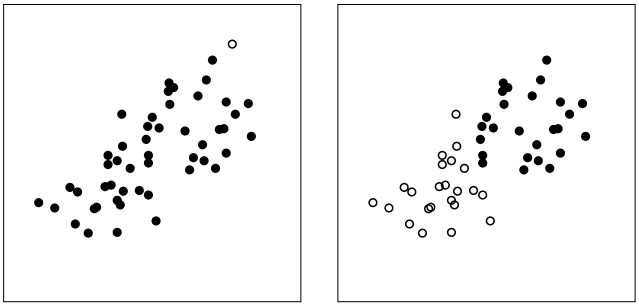
(c) $S = 3$

Figure 5. Evaluation of parameter S based on data set C. ($R = 1$)

first clustering results shown in Figures (a-1), (b-1) and (c-1) are insufficient. The second clustering result shown in Figure (b-2) in the case of $S = 1$ and the third clustering result shown in Figure (c-2) in the case of $S = 3$ are, however, considered to be reasonable results. The case $S = 0$ is clearly not acceptable



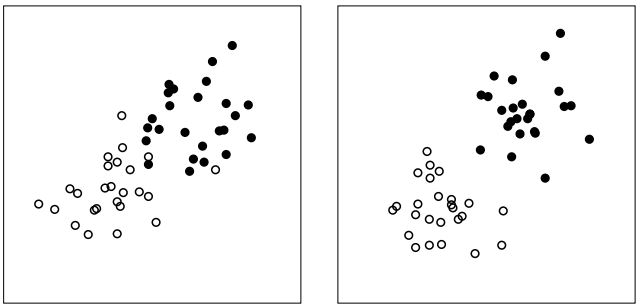
(a) $R = 0$ (b) $R = 2$



(c-1) 1st clustering result (c-2) 2nd clustering result

(c) $R = 3$

Figure 6. Evaluation of parameter R based on data set C. ($S = 2$)



(a) xy-plane (b) zw-plane

Figure 7. Four-dimensional example clustering result example based on expanded data set C. ($S = 2, R = 1$)

even for the third clustering result shown in Figure (a-2). Since all of the member stiffness values are assumed to be the same irrespective of their lengths in this case, the thickness of the assumed truss member becomes larger in accordance with the distance of the two data elements to be connected. This type of truss structural system is considered to be unreasonable from the clustering viewpoint.

Figure 6 shows the clustering results based on different values of R in the case of $S = 2$. As shown in the figure, the influence of different values of R is less significant than the case of S. The reference clustering result of $R = 1$ shown in Figure 4(b) is the same as the results of $R = 0$ and $R = 2$ respectively shown in (a) and (b) of Figure 6. Only the case of $R = 3$ shown in Figure 6(c) is slightly different.

Other calculation results that have been conducted so far exhibit similar tendencies. As a preliminary result, we conclude that the parameters determined a priori, $S = 2$ and $R = 1$, are considered to be appropriate, though the further

examination is required especially for the case of R.

C. Higher Dimensional Examples

Figure 7 shows an example clustering result of four dimensional data. The adopted data set in xy-plane is the same as the previous case shown in Figure 4(a), but it is expanded to z and w axes this time. In Figure 7, the clustering result plotted on xy-plane shown in (a) is slightly unreasonable but the result plotted on zw-plane shown in (b) demonstrates its adequateness.

V. Conclusion and Future Work

We proposed an approach of clustering based on structural mechanics, which is an application of multi-dimensional truss. The feasibility of the proposed approach was examined based on a number of calculation examples. As a preliminary result, we conclude that the clustering process based on the truss of simplex-type connection with the distance-evaluation parameter $S = 2$ and the repulsive force parameter $R = 1$ is considered to be adequate.

In the current study, the example artificial data sets are assumed to consist of only two clusters. In the case of a data set consisting of more clusters, iterative use of the proposed approach for the obtained clustering results is considered to be applicable. More detailed characteristics of the approach have to be studied with various patterns of data set examples. On the basis of the insights to be obtained, application of the approach to some practical problems has to be taken into consideration. These are considered to be part of the future work.

References

[1] C. M. Bishop, Pattern Recognition and Machine Learning. Springer-Verlag New York, 2006.

[2] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, Cluster Analysis, 5th Edition. John Wiley & Sons, 2011.

[3] F. Murtagh, “A Survey of Recent Advances in Hierarchical Clustering Algorithms,” The Computer Journal, vol. 26, 1983, pp. 354–359.

[4] B. McWilliams and G. Montana, “Subspace Clustering of High-Dimensional Data: A Predictive Approach,” Data Mining and Knowledge Discovery, vol. 28, 2014, pp. 736–772.

 ISBN: 978-1-61208-419-0

Design and Optimization of T-shaped Circulator Based on Magneto-Optical Resonator in 2D-Photonic Crystals

Victor Dmitriev, Gianni Portela and Leno Martins
Department of Electrical Engineering
Federal University of Para
Belem, Brazil
Email: victor@ufpa.br

***Abstract*—In this paper, we propose a development of a T-shaped circulator based on a 2D-photonic crystal, which has a simple and compact structure. This structure makes the non-reciprocal transmission of electromagnetic waves. Through a series of adjustments in the crystalline geometry and using the Nelder-Mead optimization method, we achieve a high level of isolation and low insertion losses.**

Keywords–Photonic crystals; Circulators; Optimization.

I. INTRODUCTION

Non-reciprocal components, such as isolators and circulators, are used in communications systems to reduce undesirable reflections that cause instability in generators and amplifiers, as well as loss of performance in these systems.

Different types of circulators based on Photonic Crystal (PhC) technology are known. Among them there are traditional three-port Y-circulators in PhCs with triangular lattice in optical region [1], in THz [2] and W-circulator [3], T-circulator in PhCs with square lattice [4][5]. All of them are based on resonance of the standing dipole mode of a Magneto-Optical (MO) resonator with a complex geometry. The proposed circulator presents a resonant cavity with a very simple structure.

Besides, when compared with the circulators presented by Wang et al. [4] and Jing et al. [5], the proposed one has the splitting factor about five times lower. Therefore, the scaling for operation at higher frequencies is much plausible.

The circulator consists of a square lattice of dielectric cylinders immersed in air. We consider a junction consisting of a resonator with MO material and three waveguides coupled to the resonator. This structure can operate in subTHz and THz frequency range and perform non-reciprocal transmission of electromagnetic waves.

The proposed device based on photonic crystal technology can be built with reduced dimensions, favoring an increase in the component integration density in communications systems. Due to strong dependence of parameters of photonic crystals with respect to geometry, adjustments were made in the crystal structure by using of an optimization technique.

This paper is organized as follows. In Section II, the optimization method is discussed and the optimal design is presented. In Section III, the device performance after using the optimization process is shown. After that, the conclusion is presented.

II. OPTIMIZATION PROCESS

The technique of optimization known as brute force is not appropriate for solution of our problem. The Nelder-Mead method, which is available in COMSOL [6], has been used for the geometry optimization. This algorithm [7] demonstrates a rapid convergence in comparison with other available methods.

Considering excitation in the three ports of the circulator, we look for its good transmission and isolation for a particular frequency band. The objective function was defined as S parameters of the circulator. Thus, optimized values for the radius and the position of the ferrite and dielectric cylinders comprising the resonant were obtained.

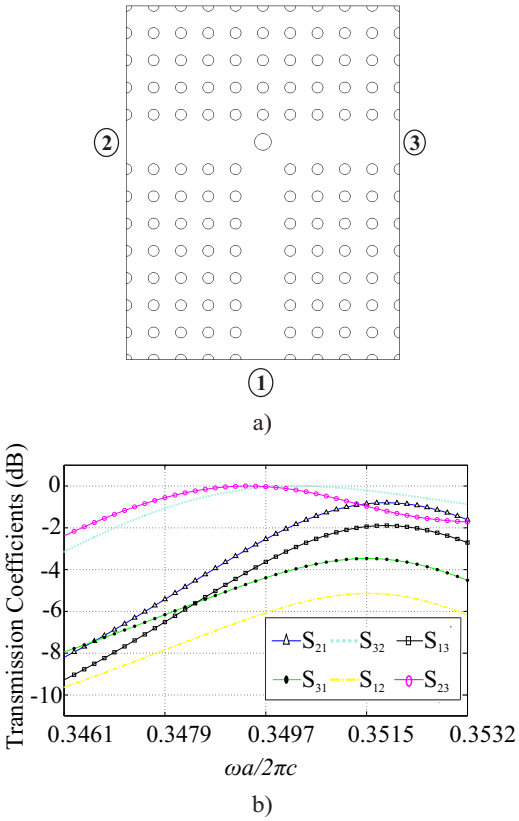


Figure 1. Before optimization process: a) Initial design. b) Frequency responses.

In Figure 1a, we show the structure of the crystal before the optimization process. The ferrite cylinder is positioned in

 ISBN: 978-1-61208-419-0

the center of the axis between the connecting waveguides. It is noticed that for this geometrical configuration, there is no non-reciprocal transmission, which leads us to seek change in the parameters of the cylinders near the resonator. There are also high losses of the structure for this geometrical arrangement, as it can be seen in Figure 1b. Then one realizes that must be applied to optimization for this problem.

From the values obtained using the optimization module, the final optimal design with the changes made in the crystal structure can be seen in Figure 2. The white cylinders are related to the periodic structure of the employed photonic crystal and each of them has radius equal to $0.2a$, where a is the lattice constant. For frequency $f = 100GHz$, $a = 1.065mm$.

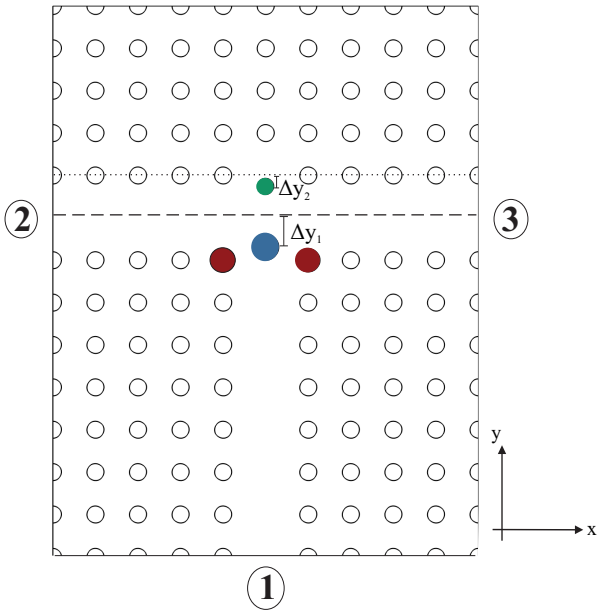


Figure 2. Optimal design.

After the changes made in the central structure, the radius of the blue cylinder remained equal to 0.30562a, but is displaced in relation to the axis of the waveguide between ports 2 and 3 (Δy_1) of 0.69086a. The green cylinder has a reduced radius 0.01249a and was moved vertically to the axis of the upper cylinders (Δy_2) in 0.2563a. The radii of the red cylinders were increased to 0.07439a.

III. Obtained Results

The frequency splitting of the rotating modes ω^+ and ω^- versus k/μ is shown in Figure 3. It is apparent that our circulator works with low parameter k/μ = 0.17, i. e. can be projected for THz region.

The resonant cavity is based on a nickel-zinc based ferrite rod inserted in the center of the device and the dipole modes are excited in this rod. The used ferrite is produced by Trans-Tech [8] and its product code is TT2-111. In order to obtain the magnetic permeability and permittivity of the employed ferrite, the following expressions were used in our simulations:

$$[\mu] = \mu_0 \begin{pmatrix} \mu & -ik & 0 \\ ik & \mu & 0 \\ 0 & 0 & \mu \end{pmatrix}; \ \epsilon = 12.5\epsilon_0. \qquad (1)$$

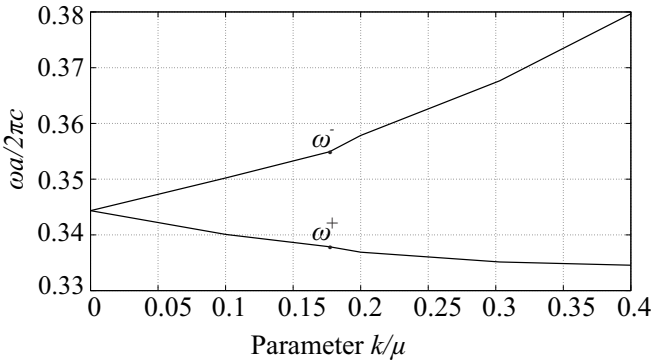


Figure 3. Frequency splitting of dipole modes excited in an MO resonator.

where the on-diagonal term μ and the off-diagonal term k are defined by the following expressions:

$$\mu = 1 + \frac{\omega_m (\omega_i + j\omega\alpha)}{(\omega_i + j\omega\alpha)^2 - \omega^2} \qquad (2)$$

$$k = \frac{\omega_m \omega}{(\omega_i + j\omega\alpha)^2 - \omega^2} \qquad (3)$$

The terms ω_m and ω_i are defined as:

$$\omega_m = \gamma M_0 \qquad (4)$$

$$\omega_i = \gamma H_0 \qquad (5)$$

In (2), (3), (4) and (5), M_0 is the saturation magnetization (398 kA/m), γ is the gyromagnetic ratio (2.33 x 10^5 rad/s per A/m), α is the damping factor (0.03175), ω is the radian frequency (rad/s), μ_0 is the free-space magnetic permeability (4π x 10^{-7} H/m) and $\mathbf{H}_0$ is the applied DC magnetic field (kA/m).

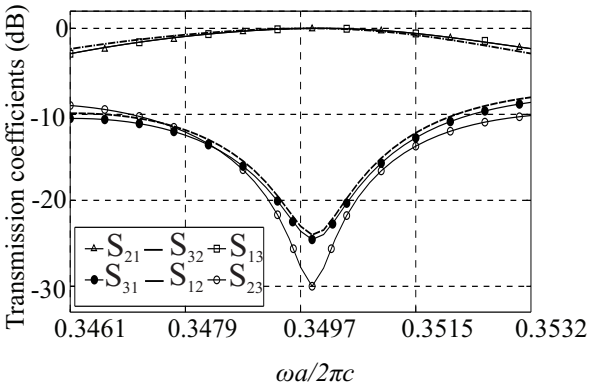


Figure 4. Frequency responses of T-circulator for excitation at ports 1, 2 and 3.

The device frequency response is shown in Figure 4. In the normalized central frequency $\omega a/2\pi c$ = 0.3499, the insertion losses are smaller than -0.05 dB, where: ω is the angular frequency (in radians per second); c is the speed of light in free space. In the frequency band located around 100 GHz, the bandwidth defined at the level of -15 dB of isolation is equal

 ISBN: 978-1-61208-419-0

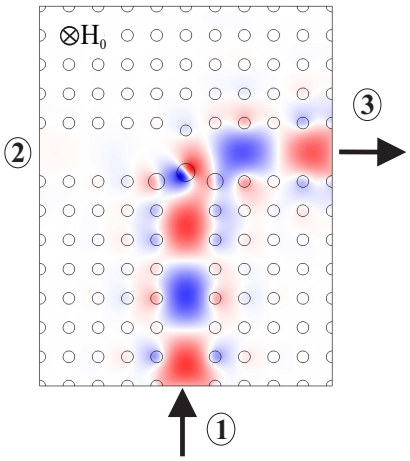


Figure 5. E_z-component of electromagnetic field for T-circulator at central frequency $f = 98.55GHz$ for excitation at port 1.

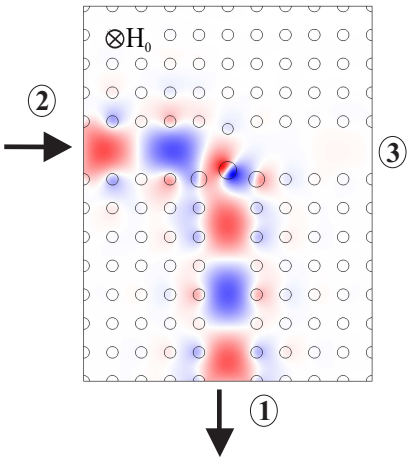


Figure 6. E_z-component of electromagnetic field for T-circulator at central frequency $f = 98.55GHz$ for excitation at port 2.

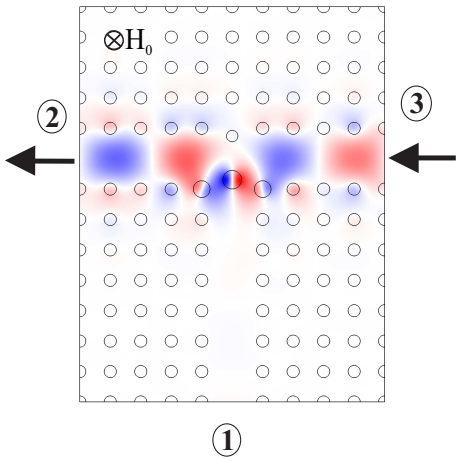


Figure 7. E_z-component of electromagnetic field for T-circulator at central frequency $f = 98.55GHz$ for excitation at port 3.

to 620 MHz for excitation at port 1, 680 MHz for excitation at port 2 and 730 MHz for excitation at the port 3.

The propagation of electromagnetic waves is given as follows: when the excitation is applied at port 1, there is signal transmission from this port to port 3, with isolation of port 2 due to the special alignment of the dipole mode, as can be seen in Figure 5. Similarly, when the input signal is applied in the port 2 (Figure 6), this is transferred to the port 1, with isolation of port 3 and in the port 3 (Figure 7), this is transferred to the port 2, with isolation of port 1. This case corresponds to the propagation in a counterclockwise direction. If the signal of the external DC magnetic field $\mathbf{H}_0$ is reversed, the propagation of signals is clockwise (1 to 2, 2 to 3 and 3 to 1).

In the cases illustrated in Figures 5 and 6, it can be seen that the stationary dipole mode excited in the resonant cavity is rotated by an angle of $45°$, which provides isolation of ports 2 and 3, respectively. On the other hand, in the case illustrated in Figure 7, it is shown that the stationary dipole mode suffers no rotation, making the input signal applied in port 3 is transferred to the port 2 with port 1 isolated.

Analysing intensity of the electric field in the T-junction, one can see that, the resonant cavity is formed by a central ferrite cylinder and two dielectric cylinders with increased diameters compared to other cylinders that comprising the photonic crystal.

IV. Conclusion

In this paper, we have presented a T-junction circulator with reduced dimensions. Several changes were made in the device initial design, in order to realize the isolation function. That is, to protect the signal source from stray reflections of not a ideally matched load. By using parameter optimization, we have obtained good characteristics of the circulator, namely, low insertion losses between the input and the output ports, high input isolation levels and a relatively wide operating frequency band.

Acknowledgment

This work was supported by the Brazilian agency CNPq.

References

[1] W. Smigaj, J. Romero-Vivas, B. Gralak, L. Magdenko, B. Dagens, and M. Vanwolleghem, "Magneto-optical circulator designed for operation in a uniform external magnetic field," Optics Letters, vol. 35, 2010, pp. 568–570, doi: 10.1364/OL.35.000568.

[2] F. Fan, S.-J. Chang, Y. H. C. Niu, and X.-H. Wang, "Magnetically tunable silicon-ferrite photonic crystals for terahertz circulator," Optics Communications, vol. 285, 2012, pp. 3763–3769, doi: 10.1016/j.optcom.2012.05.044.

[3] V. Dmitriev, M. Kawakatsu, and F. J. M. de Souza, "Compact three-port optical 2D photonic crystal-based circulator of W-format," Optics Letters, vol. 37, 2012, pp. 3192–3194, doi: 10.1364/ol.37.003192.

[4] Q. Wang, Z. Quyang, K. Tau, M. Lin, and S. Ruan, "T-shaped optical circulator based on coupled magneto-optical rods and a side-coupled cavity in a square-lattice photonic crystal," Physics Letters A, vol. 376, 2012, pp. 646–649, doi: 10.1016/j.physleta.2011.11.032.

[5] X. Jin, Z. Ouyang, Q. Wang, M. Lin, G. Wen, and J. Wang, "Highly Compact Circulators in Square-Lattice Photonic Crystal Waveguides," PLoS ONE 9(11): e113508, 2014, doi:10.1371/journal.pone.0113508.

[6] "COMSOL Multiphysics," www.comsol.com [accessed: 2015-05-18].

[7] J. Lagarias, J. Reeds, M. Wright, and P. Wright, "Convergence properties of the NelderMead simplex method in low dimensions," Siam Journal on Optimization, vol. 9, 1998, pp. 112–147, doi:10.1137/S1052623496303470.

[8] "Trans Tech," www.trans-techinc.com [accessed: 2015-05-18].

 ISBN: 978-1-61208-419-0

Takagi Sugeno Fuzzy Controller for Uncertain Single Link Manipulator

Umar Farooq[*], Jason Gu[#], Mohamed E. El-Hawary[#], Jun Luo[†] and Muhammad Usman Asad[‡]
[*]Department of Electrical Engineering, University of The Punjab Lahore-54590 Pakistan
[#]Department of Electrical and Computer Engineering, Dalhousie University Halifax, N.S., Canada
[†]School of Mechatronic Engineering and Automation, Shanghai University China
[‡]Department of Electrical Engineering, The University of Lahore, Lahore Pakistan
E-mails: engr.umarfarooq@yahoo.com, Jason.gu@dal.ca, elhawary@dal.ca, luojun@shn.edu.cn, usman.asad@ee.uol.edu.pk

***Abstract*—This paper presents the design of a fuzzy tracking controller for uncertain Single Link Manipulator (SLM) moving in the vertical plane. A Takagi-Sugeno (TS) fuzzy model of the uncertain nonlinear plant is constructed using sector nonlinearity approach and a set of operations point technique. The controller design for TS fuzzy plant is then carried out using Francis-Isidori-Byrnes (FIB) nonlinear regulation theory and parallel distributed compensation (PDC) technique. MATLAB simulations are performed to validate the designed controller for tracking constant and sinusoidal reference signals.**

Keywords- Uncertain single link manipulator; TS fuzzy model; Parallel distributed compensation; Linear matrix inequalities; Francis-Isidori-Byrnes nonlinear regulation theory; MATLAB/Simulink.

I. INTRODUCTION

Single Link Manipulators (SLM) are popular platforms to study control algorithms. Two types of SLM are often deployed by researchers to validate their control techniques. These include flexible joint and flexible link manipulators which can be made to operate either in horizontal or vertical plane. The vertical plane motion introduces an additional factor of gravity in the model. A variety of linear and nonlinear techniques are found in literature for the control of SLM. The design of H-∞ based Proportional-Derivative-Integral (PID) control is presented in [1] for tip regulation task in SLM. The method considers the model uncertainty as a result of neglecting high frequency modes and computes the gain space for PID controller using H-∞ optimization criterion. Real-time implementation results using a digital signal processor validates the proposed controller which is also found to outperform the Ziegler-Nichols-PID controller in terms of the transient performance and robustness. Back stepping method tuned by Genetic algorithm is used by Ali Sahab and Modabbernia [2] to control SLM. Through a series of virtual control inputs and control Lyapunov functions, convergence of tracking error is shown. A fitness function is formed to minimize the settling time and percentage overshoot. Based on this function, Genetic algorithm finds optimal gains for back stepping controller. The proposed algorithm is shown to perform better than robust control methods for stabilization and reference tracking tasks. An adaptive controller is proposed in [3] to control a SLM which adjusts the position and velocity gains based on the tracking error. The controller demands large bandwidth (as a function of error) during startup to provide fast response and bandwidth decreases as the error converges to zero which helps to eliminate the overshoot in system response. A notion of dynamic pole motion explains the system stability under the presented design scheme. The use of fuzzy logic in controlling a SLM is also addressed [4]-[7]. A two stage fuzzy controller is presented in [4] for tip position tracking in SLM. The first stage employs two fuzzy logic controllers with motor angle and its derivative being the inputs of first 81-rule base controller while the second controller processes tip angle and its derivative using a rule-base containing 49 rules. The outputs from these fuzzy logic controllers form input to a second stage fuzzy logic controller which generates pulse width modulated signal to drive the DC motor. Simulation and experimental results show the superior performance of the proposed controller in comparison to PID controller. The optimization of a fuzzy controller in terms of its scaling gains and membership functions is carried out using Genetic algorithm [5] which uses a weighted combination of conflicting objectives including the fast response and minimal overshoot as a fitness function. In addition, a command shaper is also integrated to modify the reference signal keeping in view the vibration modes. The command shaper is also tuned using genetic algorithm to give the optimal locations and amplitudes of impulses which are then convolved with desired reference signal to generate a modified reference signal.

This paper follows a model-based approach for the design of fuzzy logic controller for stabilization and tracking control of SLM moving in vertical plane. By assuming the parameters to be uncertain, a TS fuzzy plant model is constructed which exactly represents the original nonlinear dynamics in compact region formed from parameter bounds and operating region [8]. Francis-Isidori-Byrnes (FIB) nonlinear regulation theory and Parallel Distributed Compensation (PDC) technique [9][10] is used to design a controller for tracking constant and sinusoidal references. Controller part based on FIB is responsible for directing the system motion towards the steady state manifold and generates steady state input for forcing the system to stay there while PDC part ensures the system stability during convergence to steady state manifold. FIB part is designed after solving time varying matrix differential equations in

 ISBN: 978-1-61208-419-0

terms of fuzzy sets, while PDC part is designed using linear matrix inequality techniques where the existence of a symmetric positive definite matrix for all fuzzy sub-systems proves the system stability. MATLAB simulations are performed to show the effectiveness of the designed controller for SLM. It is found that controller has remained successful in tracking the reference trajectories with good transient performance. The contribution of the paper lies in constructing a fuzzy model for uncertain single link manipulator based on the idea of set of operation point's technique. The stabilization and tracking of the resulting model is achieved using exact output regulation theory.

We start by constructing the TS fuzzy plant model in Section II. Controller design is presented in Section III followed by simulation results in Section IV. Conclusions are drawn in Section V.

II. TS FUZZY MODEL OF SLM

The dynamics of a SLM consisting of a rod with a circular disc at one end and moving in the vertical plane can be described by the following differential equation:

$$\left(\frac{1}{3}ml^2+Ma^2+M\left(l+a\right)^2\right)\ddot{\theta}+b\dot{\theta}+g\left(M\left(a+l\right)+\frac{ml}{2}\right)\sin\theta=\tau \tag{1}$$

Where m is the mass of the rod, l is the length of the rod, M is the mass of the circular disc, a is the radius of the disc, b is the coefficient of viscous friction at the pivot, θ is the angle of the link from vertical, g is the acceleration due to gravity and τ is the torque provided by the DC motor for reference tracking purposes. The numerical values of these parameters are listed in Table 1 where mass of the circular disc and the damping coefficient are assumed to be uncertain.

By defining the state vector to be $\mathbf{x}=\begin{bmatrix} x_1=\theta & x_2=\dot{\theta}\end{bmatrix}^T$ the system in (1) can be represented as:

$$\begin{aligned}\dot{\mathbf{x}}&=\begin{bmatrix}0 & 1\\ f_{21}\left(x_1,M\right) & f_{22}\left(b,M\right)\end{bmatrix}\mathbf{x}+\begin{bmatrix}0\\ g_2\left(M\right)\end{bmatrix}u\\ y&=\begin{bmatrix}1 & 0\end{bmatrix}\mathbf{x}\end{aligned} \tag{2}$$

Where u represents the torque to be generated by the motor, $y=x_1$ denotes the system output and the nonlinear functions are given as:

$$f_{21}\left(x_1,M\right)=-\frac{g\left(M\left(a+l\right)+\frac{ml}{2}\right)}{\left(\frac{1}{3}ml^2+Ma^2+M\left(l+a\right)^2\right)}\cdot\frac{\sin x_1\left(t\right)}{x_1\left(t\right)} \tag{3}$$

$$f_{22}\left(b,M\right)=-\frac{b}{\left(\frac{1}{3}ml^2+Ma^2+M\left(l+a\right)^2\right)} \tag{4}$$

$$g_2\left(M\right)=\frac{1}{\left(\frac{1}{3}ml^2+Ma^2+M\left(l+a\right)^2\right)} \tag{5}$$

By assuming the angular displacement of the link to lie in the range $-\frac{\pi}{2}\le x_1\left(t\right)\le\frac{\pi}{2}$, we can define the following compact region covering the parametric uncertainties and operating range as:

$$D=\left\{\left(b,M,x_1\right)\in\mathbb{R}^3:0.01\le b\le 0.05,0.01\le M\le 0.1,-\frac{\pi}{2}\le x_1\le\frac{\pi}{2}\right\} \tag{6}$$

TS fuzzy model of the system in (2) can be constructed so as to exactly reproduce the plant dynamics over the compact region (6) by finding the extreme values of the nonlinear functions (3)-(5) for this region. This result is based on the following property and will ensure the stabilization of the plant over the compact region by using PDC controller:

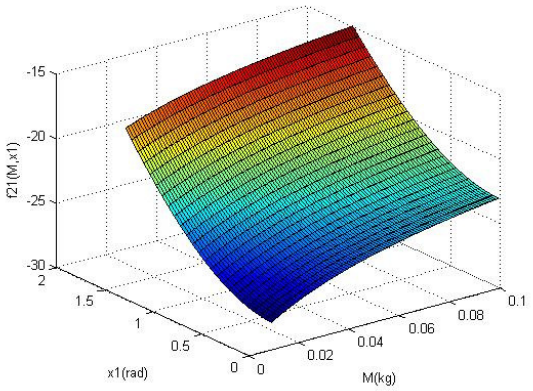
Property 1: Let $I_p\subset\mathbb{R}^p$ and $I_q\subset\mathbb{R}^q$ be compact subsets and $I=I_p\times I_q$. Let $f:I\subset\mathbb{R}^t\to\mathbb{R}$ be a continuous function with $t=p+q$. If for some given $p_0\in I_p$, $M=\max_{q\in I_q}\left\{f\left(p_0,q\right)\right\}$, and $m=\min_{q\in I_q}\left\{f\left(p_0,q\right)\right\}$; then $M\le\max_{(p_0,q)\in I}\left\{f\left(p_0,q\right)\right\}$, and $m\ge\min_{(p_0,q)\in I}\left\{f\left(p_0,q\right)\right\}$.

The variation of the functions (3)-(5) over the compact region (6) is depicted in Fig. 1 and the extreme values are found to be:

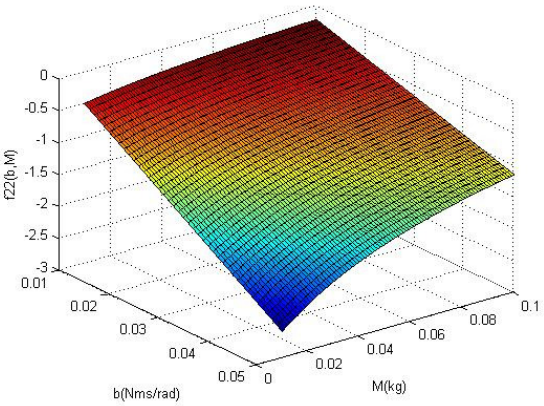
$$f_{21,\min}=-28.0339 \tag{7}$$
$$f_{21,\max}=-14.7767 \tag{8}$$
$$f_{22,\min}=-2.5949 \tag{9}$$
$$f_{22,\max}=-0.2343 \tag{10}$$
$$g_{2,\min}=33.6965 \tag{11}$$
$$g_{2,\max}=51.8977 \tag{12}$$

We can now define the following fuzzy sets with universe of discourse being the extreme values in (7)-(12) which will enable us to build the TS fuzzy plant model:

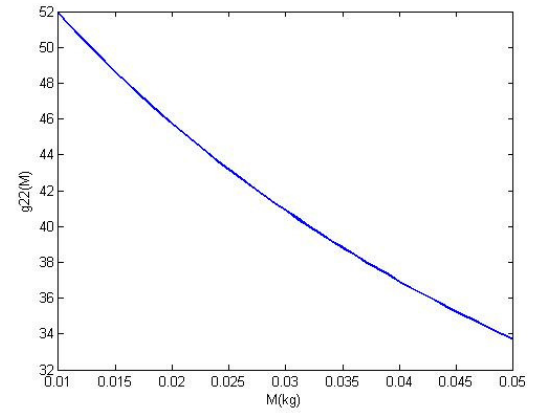
$$\begin{aligned}M_1&=\begin{cases}\dfrac{f_{21}\left(M,x_1\right)-f_{21,\min}}{f_{21,\max}-f_{21,\min}}, & x_1\left(t\right)\ne 0\\ 0, & x_1\left(t\right)=0\end{cases}\\ M_2&=1-M_1\end{aligned} \tag{13}$$



(a)



(b)



(c)

Figure 1. Plot of functins (3)-(5) over compact region (6) (a) $f_{21}\left(x_1^+, M\right)$ (b) $f_{22}\left(b, M\right)$ (c) $g_2\left(M\right)$

$$N_1 = \frac{f_{22}\left(b,M\right) - f_{22,\min}}{f_{22,\max} - f_{22,\min}} \qquad (14)$$
$$N_2 = 1 - N_1$$

$$O_1 = \frac{g_2\left(M\right) - g_{2,\min}}{g_{2,\max} - g_{2,\min}} \qquad (15)$$
$$O_2 = 1 - O_1$$

Based on the fuzzy sets (13)-(15), we define the following plant rules:

Rule 1: IF f_{21} is M_2 AND f_{22} is N_2 AND g_2 is O_2
THEN $\dot{\mathbf{x}} = \mathbf{A}_1\mathbf{x} + \mathbf{B}_1 u$

Rule 2: IF f_{21} is M_2 AND f_{22} is N_2 AND g_2 is O_1
THEN $\dot{\mathbf{x}} = \mathbf{A}_2\mathbf{x} + \mathbf{B}_2 u$

Rule 3: IF f_{21} is M_2 AND f_{22} is N_1 AND g_2 is O_2
THEN $\dot{\mathbf{x}} = \mathbf{A}_3\mathbf{x} + \mathbf{B}_3 u$

Rule 4: IF f_{21} is M_2 AND f_{22} is N_1 AND g_2 is O_1
THEN $\dot{\mathbf{x}} = \mathbf{A}_4\mathbf{x} + \mathbf{B}_4 u$

Rule 5: IF f_{21} is M_1 AND f_{22} is N_2 AND g_2 is O_2
THEN $\dot{\mathbf{x}} = \mathbf{A}_5\mathbf{x} + \mathbf{B}_5 u$

Rule 6: IF f_{21} is M_1 AND f_{22} is N_2 AND g_2 is O_1
THEN $\dot{\mathbf{x}} = \mathbf{A}_6\mathbf{x} + \mathbf{B}_6 u$

Rule 7: IF f_{21} is M_1 AND f_{22} is N_1 AND g_2 is O_2
THEN $\dot{\mathbf{x}} = \mathbf{A}_7\mathbf{x} + \mathbf{B}_7 u$

Rule 8: IF f_{21} is M_1 AND f_{22} is N_1 AND g_2 is O_1
THEN $\dot{\mathbf{x}} = \mathbf{A}_8\mathbf{x} + \mathbf{B}_8 u$

Where

$$\mathbf{A}_1 = \mathbf{A}_2 = \begin{bmatrix} 0 & 1 \\ -28.0339 & -2.5949 \end{bmatrix}$$
$$\mathbf{A}_3 = \mathbf{A}_4 = \begin{bmatrix} 0 & 1 \\ -28.0339 & -0.2343 \end{bmatrix}$$
$$\mathbf{A}_5 = \mathbf{A}_6 = \begin{bmatrix} 0 & 1 \\ -14.7767 & -2.5949 \end{bmatrix} \qquad (16)$$
$$\mathbf{A}_7 = \mathbf{A}_8 = \begin{bmatrix} 0 & 1 \\ -14.7767 & -0.2343 \end{bmatrix}$$
$$\mathbf{B}_1 = \mathbf{B}_3 = \mathbf{B}_5 = \mathbf{B}_7 = \begin{bmatrix} 0 \\ 33.6965 \end{bmatrix}$$
$$\mathbf{B}_2 = \mathbf{B}_4 = \mathbf{B}_6 = \mathbf{B}_8 = \begin{bmatrix} 0 \\ 51.8977 \end{bmatrix}$$

Using the singleton fuzzification, product inference engine and average defuzzification technique, TS fuzzy plant model can be given as:

 ISBN: 978-1-61208-419-0

TABLE I. PLANT PARAMETERS

Parameter	Value
m	0.2 Kg
l	0.5 m
a	0.01 m
M	[0.01, 0.1] Kg
b	[0.01, 0.05] Nms/rad
g	9.8 m/s^2

$$\dot{\mathbf{x}} = \sum_{i=1}^{8} \alpha_i\left(\mathbf{z}(t)\right)\left(\mathbf{A}_i\mathbf{x} + \mathbf{B}_i u\right)$$
$$y = \sum_{i=1}^{8} \alpha_i\left(\mathbf{z}(t)\right)\mathbf{C}_i\mathbf{x} \tag{17}$$

$$\alpha_i\left(\mathbf{z}(t)\right) = \frac{\rho_i\left(\mathbf{z}(t)\right)}{\sum_{i=1}^{8}\rho_i\left(\mathbf{z}(t)\right)} \tag{18}$$

$$\rho_i\left(\mathbf{z}(t)\right) = M_i\left(M, x_1\right)\times N_i\left(b, M\right)\times O_i\left(M\right) \tag{19}$$

Where $\rho_i\left(\mathbf{z}(t)\right)$ and $\alpha_i\left(\mathbf{z}(t)\right)$ are the firing and normalized firing strengths of the 'i^{th}'rule respectively which contains the fuzzy sets M_i, N_i and O_i. This degree of belongingness is determined based on the scheduling vector, $\mathbf{z}(t) = \left[f_{21}\left(M, x_1\right) \quad f_{22}\left(b, M\right) \quad g_2\left(M\right)\right]$.

III. TS FUZZY CONTROLLER DESIGN

TS fuzzy controller for SLM is designed based on Francis-Isidori-Byrnes (FIB) nonlinear regulation theory which guarantees exact tracking through the control law (20) subject to the solution of the differential equations (21):

$$u(t) = -\mathbf{K}\left(\mathbf{x}(t) - \boldsymbol{\pi}\left(\mathbf{w}(t)\right)\right) - \boldsymbol{\gamma}\left(\mathbf{w}(t)\right) \tag{20}$$

$$\frac{\partial\boldsymbol{\pi}\left(\mathbf{w}(t)\right)}{\partial\mathbf{w}(t)}\mathbf{s}\left(\mathbf{w}(t)\right) = \mathbf{f}\left(\boldsymbol{\pi}\left(\mathbf{w}(t)\right), \mathbf{w}(t), \boldsymbol{\gamma}\left(\mathbf{w}(t)\right)\right)$$
$$\mathbf{0} = \mathbf{h}\left(\boldsymbol{\pi}\left(\mathbf{w}(t)\right), \mathbf{w}(t)\right) \tag{21}$$

Where $\boldsymbol{\pi}\left(\mathbf{w}(t)\right)$ is the steady state zero error manifold, $\boldsymbol{\gamma}\left(\mathbf{w}(t)\right)$ is the steady state input, $\mathbf{f}$ is the system dynamics, $\mathbf{s}$ forms the exosystem to be tracked, $\mathbf{h}$ is the tracking error and $\mathbf{K}$ is the stabilizing gain. It is shown in [9] that these nonlinear equations can be exactly solved in terms of fuzzy sets with time varying degree of membership. We introduce the following fuzzy exosystem which will serve the purpose of reference signal generation:

$$\dot{\mathbf{w}} = \sum_{i=1}^{2}\beta_i\left(f_{21}(t)\right)\mathbf{S}_{ji}\mathbf{w}$$
$$y_{ref} = \sum_{i=1}^{2}\beta_i\left(f_{21}(t)\right)\mathbf{Q}_i\mathbf{w} \tag{22}$$

Where $\mathbf{S}_{1i}$ and $\mathbf{S}_{2i}$ denote the constant and sinusoidal reference state matrices respectively while $\mathbf{Q}_i$ is the reference output vector. β_i is the normalized firing strength for the 'i^{th}'rule of fuzzy exosystem.

$$\mathbf{S}_{11} = \mathbf{S}_{12} = \begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}, \mathbf{w}(0) = \begin{bmatrix}1 & 0\end{bmatrix}^T \tag{23}$$

$$\mathbf{S}_{21} = \mathbf{S}_{22} = \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}, \mathbf{w}(0) = \begin{bmatrix}1 & 0\end{bmatrix}^T \tag{24}$$

$$\mathbf{Q}_1 = \mathbf{Q}_2 = \begin{bmatrix}x_{1ref} & 0\end{bmatrix}^T \tag{25}$$

Using (17) and (22), the tracking error can be given as:

$$e(t) = \sum_{i=1}^{8}\alpha_i\left(\mathbf{z}(t)\right)\mathbf{C}_i\mathbf{x} - \sum_{i=1}^{2}\beta_i\left(f_{21}(t)\right)\mathbf{Q}_i\mathbf{w} \tag{26}$$

The above error will be converged asymptotically to zero using the control law (27):

$$u(t) = -\sum_{i=1}^{8}\alpha_i\left(\mathbf{z}(t)\right)\mathbf{K}\left(\mathbf{x}(t) - \boldsymbol{\Pi}\left(\mathbf{w}(t)\right)\right) - \boldsymbol{\Gamma}\left(\mathbf{w}(t)\right) \tag{27}$$

Where $\boldsymbol{\Pi}(t)$ and $\boldsymbol{\Gamma}(t)$ are updated as a result of the solution of following time varying matrix equations:

$$\dot{\boldsymbol{\Pi}}(t) = \sum_{i=1}^{8}\alpha_i\left(\mathbf{z}(t)\right)\mathbf{A}_i\boldsymbol{\Pi}(t) + \sum_{i=1}^{8}\alpha_i\left(\mathbf{z}(t)\right)\mathbf{B}_i\boldsymbol{\Gamma}(t) - \boldsymbol{\Pi}(t)\sum_{i=1}^{2}\beta_i\left(f_{21}(t)\right)\mathbf{S}_{ji}$$
$$\mathbf{0} = \sum_{i=1}^{8}\alpha_i\left(\mathbf{z}(t)\right)\mathbf{C}_i\boldsymbol{\Pi}(t) - \sum_{i=1}^{2}\beta_i\left(f_{21}(t)\right)\mathbf{Q}_i \tag{28}$$

The control objective of tracking the constant and sinusoidal references by SLM leads to the following mappings:

$$x_1(t) = w_1(t)$$
$$\dot{x}_1(t) = \dot{w}_1(t) \Rightarrow x_2(t) = w_2(t) \tag{29}$$

$$\dot{x}_2(t) = M_1\left(x_1(t)\right)\left(f_{21,\max}x_1(t) + f_{22n}x_2(t) + g_{2n}u(t)\right) + M_2\left(x_1(t)\right)\left(f_{21,\min}x_1(t) + f_{22n}x_2(t) + g_{2n}u(t)\right) \tag{30}$$

Where f_{22n} and g_{2n} denote the nominal function values. Using (28)-(30), we find the following steady state zero error manifold and steady state input matrices:

$$\mathbf{\Pi}(t)=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \tag{31}$$

$$\mathbf{\Gamma}^1(t)=-\frac{1}{g_{2n}}\begin{bmatrix}M_1(w_1(t))f_{21,\max}+M_2(w_1(t))f_{21,\min} & f_{22n}\end{bmatrix} \tag{32}$$

$$\mathbf{\Gamma}^2(t)=-\frac{1}{g_{2n}}\begin{bmatrix}M_1(w_1(t))f_{21,\max}+M_2(w_1(t))f_{21,\min}+1 & f_{22n}\end{bmatrix} \tag{33}$$

Where $\mathbf{\Gamma}^1(t)$ and $\mathbf{\Gamma}^2(t)$ govern the steady inputs for steady state zero error manifold corresponding to reference state matrices $\mathbf{S}_{1i}$ and $\mathbf{S}_{2i}$ respectively. The other part of the control law will ensure the stabilization of the equilibrium point. We will use PDC technique to design the stabilizing controller for SLM model (17) which will share the same fuzzy sets as that of plant to weight the control gains of fuzzy sub-systems. The 'i^{th}' control rule will be defined as:

Con. Rule i: IF f_{21} is M_i AND f_{22} is N_i AND g_2 is O_i
THEN $u_{Ki}(t)=-\mathbf{K}_i\mathbf{x}(t)$

The net stabilizing control gain is found as:

$$u_K(t)=-\sum_{i=1}^{8}\eta_i(\mathbf{z}(t))\mathbf{K}_i\mathbf{x}(t) \tag{34}$$

Using (17) and (34), the closed loop dynamics can be given as:

$$\dot{\mathbf{x}}(t)=\sum_{i=1}^{8}\sum_{j=1}^{8}\alpha_i(\mathbf{z}(t))\eta_j(\mathbf{z}(t))(\mathbf{A}_i-\mathbf{B}_i\mathbf{K}_j)\mathbf{x}(t) \tag{35}$$

To ensure the closed loop system stability, the following Lyapunov inequality must hold:

$$(\mathbf{A}_i-\mathbf{B}_i\mathbf{K}_j)^T\mathbf{P}+\mathbf{P}(\mathbf{A}_i-\mathbf{B}_i\mathbf{K}_j)<0,\forall i,j\le 8 \tag{36}$$

Where $\mathbf{P}$ is a symmetric positive definite matrix. The above inequalities can be cast as LMIs whose solution can return the control gains for fuzzy sub-systems. By pre- and post-multiplying (36) with $\mathbf{P}^{-1}$ and re-defining, $\mathbf{P}=\mathbf{P}^{-1}$ and defining $\mathbf{Q}_i=\mathbf{K}_i\mathbf{P}$, we obtain the following LMIs with the inclusion of decay rate constraint:

$$\begin{aligned}&\mathbf{P}>0\\ &\mathbf{A}_i\mathbf{P}+\mathbf{P}\mathbf{A}_i^T-\mathbf{B}_i\mathbf{Q}_i-\mathbf{Q}_i^T\mathbf{B}_i^T+2\lambda\mathbf{P}<0,\forall i\le 8\\ &\mathbf{A}_i\mathbf{P}+\mathbf{P}\mathbf{A}_i^T+\mathbf{A}_j\mathbf{P}+\mathbf{P}\mathbf{A}_j^T-\mathbf{B}_i\mathbf{Q}_j\\ &-\mathbf{Q}_j^T\mathbf{B}_i^T-\mathbf{B}_j\mathbf{Q}_i-\mathbf{Q}_i^T\mathbf{B}_j^T+4\lambda\mathbf{P}\le 0,\forall i<j\le 8\end{aligned} \tag{37}$$

The solution of LMIs will give $\mathbf{P}$ and $\mathbf{Q}_i$ matrices from which the control gains can be determined as:

$$\mathbf{K}_i=\mathbf{Q}_i\mathbf{P}^{-1},\forall i=1-8 \tag{38}$$

The above set of 37 LMIs (37) is solved using LMI toolbox of MATLAB with $\lambda=1$ and following control gains and symmetric positive definite matrix are found:

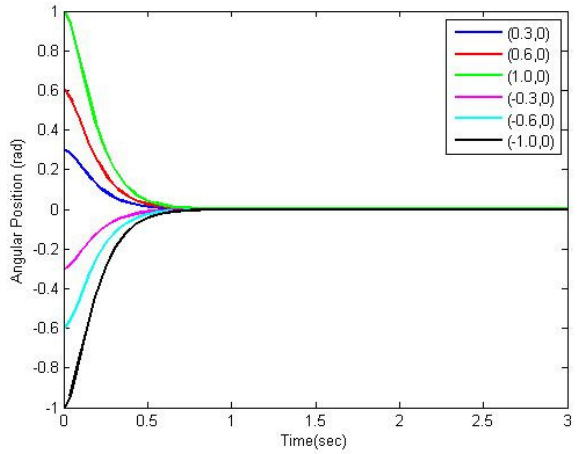
$$\begin{aligned}\mathbf{K}_1&=\begin{bmatrix}2.4249 & 0.5923\end{bmatrix}\\ \mathbf{K}_2&=\begin{bmatrix}1.5788 & 0.3869\end{bmatrix}\\ \mathbf{K}_3&=\begin{bmatrix}2.4692 & 0.6691\end{bmatrix}\\ \mathbf{K}_4&=\begin{bmatrix}1.5175 & 0.4275\end{bmatrix}\\ \mathbf{K}_5&=\begin{bmatrix}2.6174 & 0.5716\end{bmatrix}\\ \mathbf{K}_6&=\begin{bmatrix}1.8637 & 0.3917\end{bmatrix}\\ \mathbf{K}_7&=\begin{bmatrix}2.7258 & 0.6453\end{bmatrix}\\ \mathbf{K}_8&=\begin{bmatrix}1.8669 & 0.4385\end{bmatrix}\end{aligned} \tag{39}$$

$$\mathbf{P}=\begin{bmatrix}0.5153 & -2.1941\\ -2.1941 & 11.2771\end{bmatrix} \tag{40}$$

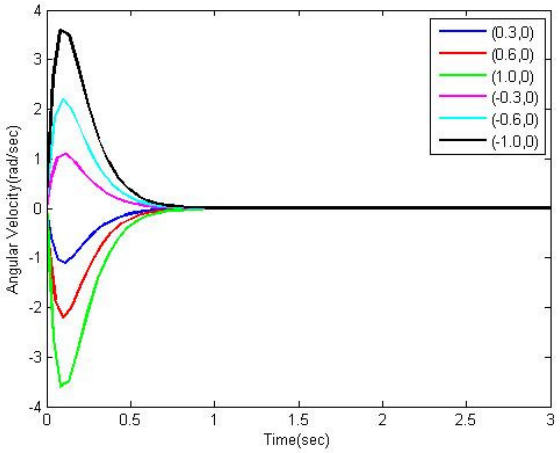
IV. SIMULATION RESULTS

The designed controller is simulated in MATLAB/Simulink environment for stabilization and tracking control of SLM. We select M =0.05Kg and b =0.03Nms/rad from the compact region for simulation purpose. The stabilization result is depicted in Fig. 2 for various initial conditions. Note that the reference generator for the stabilization $\mathbf{w}(t)$ has zero initial conditions. It can be seen that controller has remained successful to stabilize the plant. The step response of the controller is shown in Fig. 3. A set of constant reference points are also generated and controller is found to track these set points offering no overshoot, zero steady state error and less than 1sec settling time as evident from Fig. 4. Square wave reference tracking by the controller is shown in Fig. 5. It should be noted that the steady state input for all these reference signals is computed as: $u_{ss}(t)=-\mathbf{\Gamma}^1(t)\mathbf{w}(t)$. Performance of the controller for sinusoidal reference signals is also evaluated. Perfect tracking is achieved as seen from simulation results in Fig. 6, where the tracking error converges to zero within 1sec. Note that the steady state input in this case is generated as: $u_{ss}(t)=-\mathbf{\Gamma}^2(t)\mathbf{w}(t)$. For the purpose of comparison, a pole placement controller is designed for the same transient performance as offered by fuzzy logic controller $(T_s=0.6s,\xi=1)$. The comparison result in the form of tracking error is depicted in Fig. 7 when both the controllers are made to track the sinusoidal reference signal with angular position varying in the range [-1,1] rad. It can be seen that steady state error exists in case of pole placement controller while fuzzy logic controller exactly tracks the input signal after a transient.

 ISBN: 978-1-61208-419-0

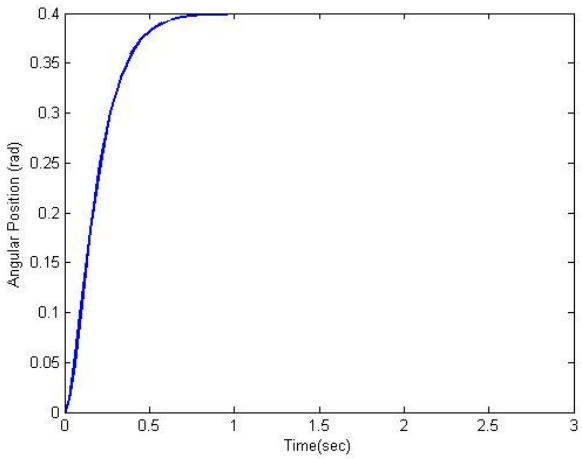


(a)

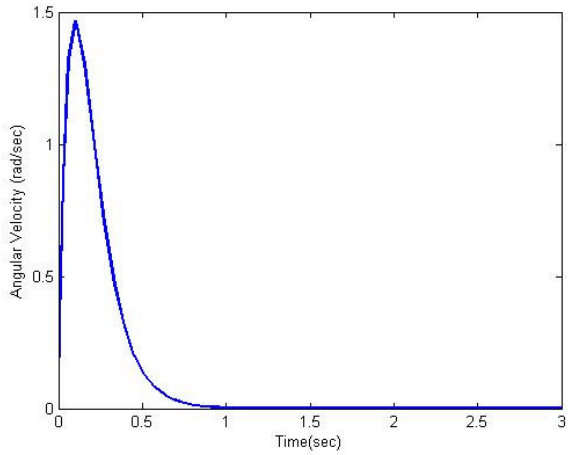


(b)

Figure 2. Stabilization of SLM for various initial conditions (a) Angular position (b) Angular velocity

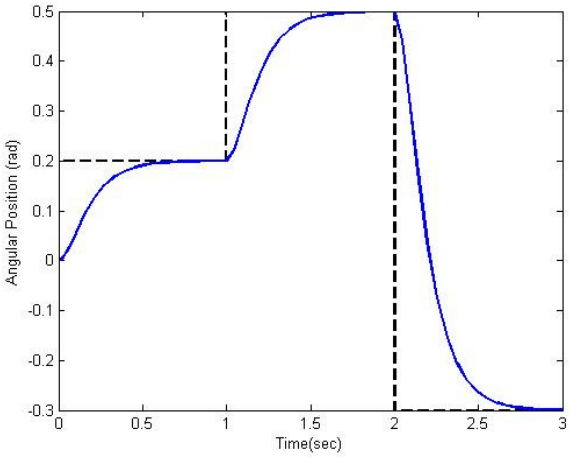


(a)

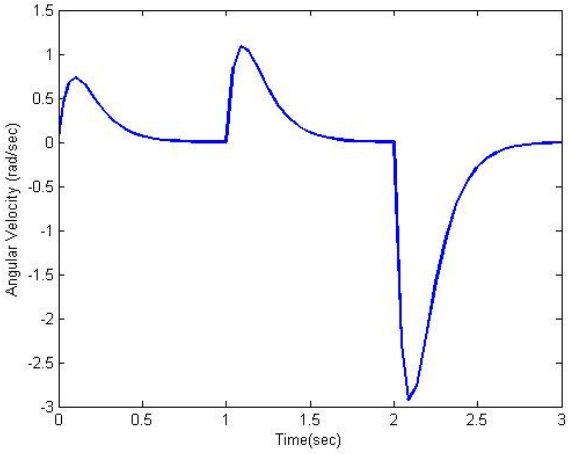


(b)

Figure 3. Step response (a) Angular position (b) Angular velocity

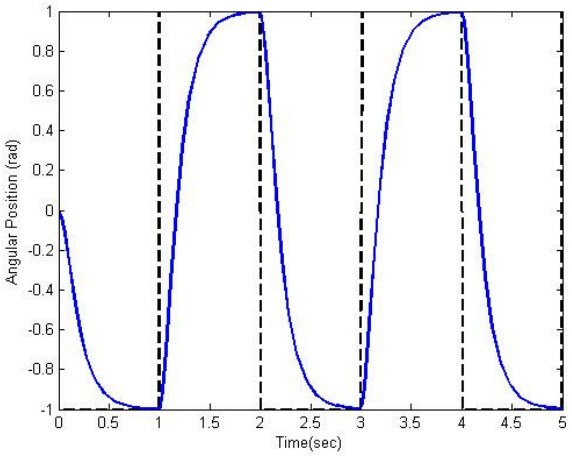


(a)

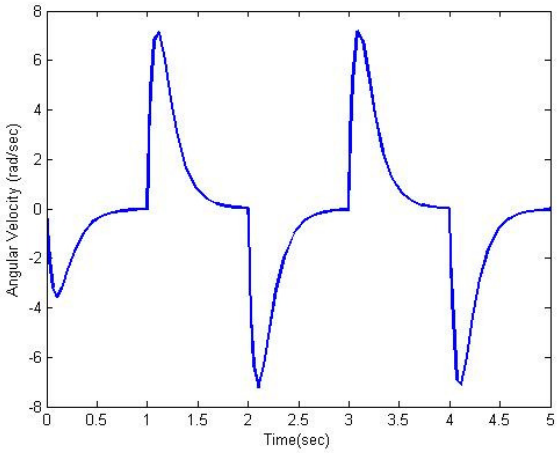


(b)

Figure 4. Reference tracking (a) Angular position (b) Angular velocity

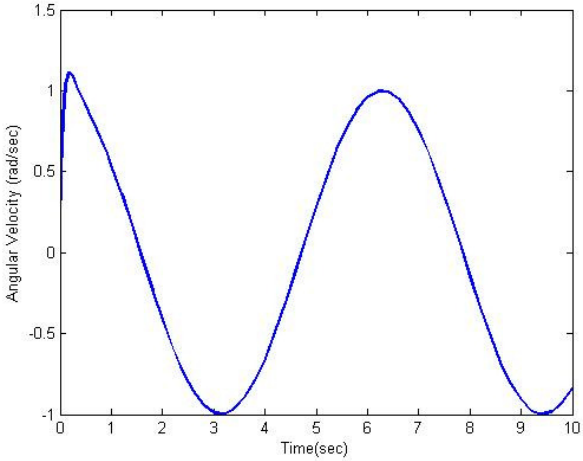


(a)

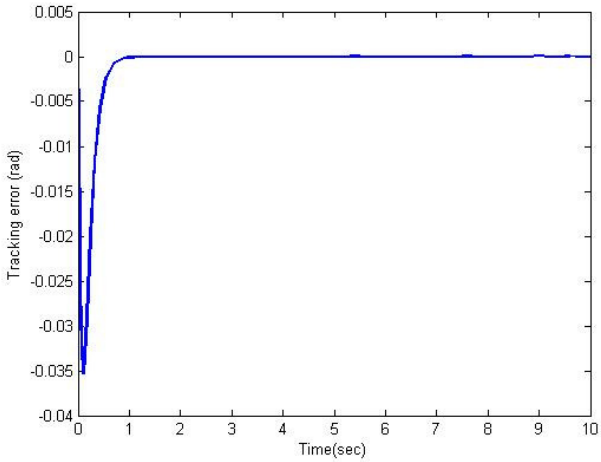


(b)

Figure 5. Square wave tracking (a) Angular position (b) Angular velocity

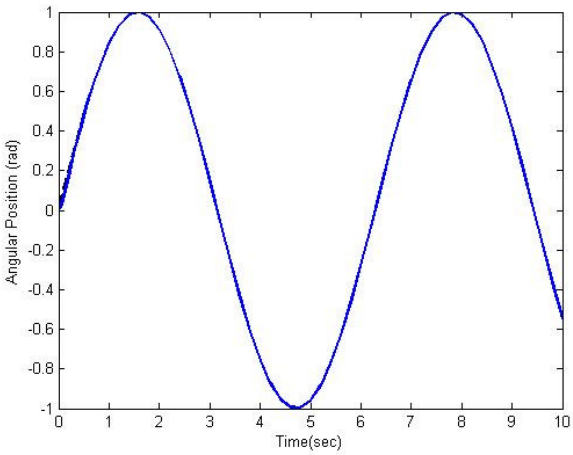


(b)



(c)

Figure 6. Sine wave tracking (a) Angular position (b) Angular velocity (c) Tracking error



(a)

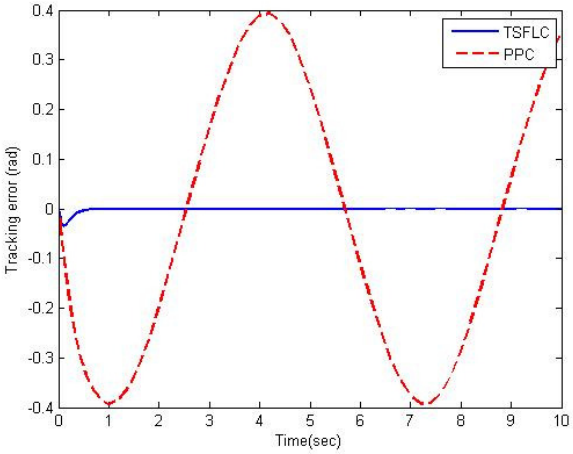


Figure 7. Comparison of TS FLC and PPC for Sine wave tracking

V. CONCLUSIONS

TS fuzzy model of uncertain single link manipulator is derived using set of operations point technique. For the purpose of demonstration, mass of the disc and friction coefficient are assumed as uncertain parameters. For exact output regulation, a PDC controller in conjunction with FIB theory is designed. MATLAB simulations are then performed to validate the designed controller for tracking constant and time varying trajectories. A comparison with pole placement controller is also drawn. Future work involves the design of estimation law for immeasurable scheduling vector.

REFERENCES

[1] M. T. Ho and Y. W. Tu, "Position control of a single link flexible manipulator using H∞ based PID control," IEE Proceedings of Control Theory, vol. 153, no. 5, 2006, pp. 615-622.

[2] A. R. Sahab and M. R. Modabbernia, "Backstepping method for single link flexible joint manipulator using genetic algorithm," International Journal of Innovative Computing, Information and Control, vol. 7, no. 7(B), 2011, pp. 4161-4170.

[3] K. Y. Song, M. M. Gupta and N. Homma, "Design of an error based adaptive controller for a flexible robot arm using dynamic pole motion approach," Journal of Robotics, vol. 2011, pp. 1-9.

[4] I. H. Akyuz, Z. Bingul and S. Kizir, "Cascade fuzzy logic control of a single link flexible joint manipulator," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 20, no. 5, 2012, pp. 713-726.

[5] M. S. Alam and M. O. Tokhi, "Hybrid fuzzy logic control with genetic optimization for single link flexible manipulator," Engineering Applications of Artificial Intelligence, vol. 21, no. 6, 2008, pp. 858-873.

[6] M. A. Ahmad, M. Z. M. Tumari and A. N. K. Nasir, "Composite fuzzy logic control approach to a flexible joint manipulator," International Journal of Advanced Robotic Systems, vol. 10, no. 58, 2013, pp. 1-9.

[7] M. R. Kandroodi, M. Mansouri, M. A. Shoorehdeli and M. Teshnehlab, "Control of flexible joint manipulator via reduced rule based fuzzy control with experimental validation," ISRN Artificial Intelligence, vol. 2012, 2012, pp. 1-8.

[8] M. P. A. Santim, M. C. M. Teixeira, W. A. de Souza, R. Cardim and E. Assuncao, "Design of a Takagi-Sugeno fuzzy regulator for a set of operation points," Mathematical Problems in Engineering, vol. 2012, 2012, pp. 1-17.

[9] J. A. Meda-Campana, J. C. Gomez-Mancilla and B. Castillo-Toledo, "Exact output regulation for nonlinear systems described by Takagi-Sugeno fuzzy models," IEEE Transactions on Fuzzy Systems, vol. 20, no. 2, 2012, pp. 235-247.

[10] H. O. Wang, K. Tanaka, M. F. Griffin, "An approach to fuzzy control of nonlinear systems: stability and design issues," IEEE Transactions on Fuzzy Systems, vol. 4, no. 1, 1996, pp. 14-23.

 ISBN: 978-1-61208-419-0

Selection of Wavelet Decomposition Levels for Vibration Monitoring of Rotating Machinery

Hocine Bendjama, Daoud Idiou, Kaddour Gherfi, Yazid Laib
Welding and NDT Research Centre
(CSC)
Algiers, Algeria
e-mails: hocine_bendjama@daad-alumni.de, ddidiou@yahoo.com, gkaddour2@yahoo.fr, yaziddl@yahoo.fr

***Abstract*—The vibration signal of a rotating machine always carries the dynamic information of the machine. Its analysis is very useful for the condition monitoring and fault diagnosis. Many signal analysis methods are able to extract useful information from vibration data. In this paper, bearing fault diagnosis is performed using Wavelet Transform (WT) and Parseval's theorem. The WT is used to decompose the original signal into several signals in order to obtain multiple data series at different resolutions. The fault can be detected from a given level of resolution. For this purpose, Parseval's theorem is used as an evaluation criterion to select the optimal level. Associated to envelope analysis, it allows clear visualization of fault frequencies. Vibration signals from a pilot scale are used to demonstrate the usefulness of the proposed method. The results of the application in inner and outer races bearing diagnosis are satisfactory.**

Keywords-vibration; fault diagnosis; wavelet transform; Parseval's theorem; bearing.

I. INTRODUCTION

Fault diagnosis is extremely important task in process monitoring. During the two past decades, various monitoring methods have been developed, such as dynamics, vibration, tribology and non-destructive techniques [1][2]. The vibration signal analysis is essential in improving condition monitoring and fault diagnosis of rotating machinery, because it always carries the dynamic information of the system. Effective utilization of the vibration signals depends upon the effectiveness of the applied signal processing techniques. A wide variety of techniques have been introduced such as: time domain and frequency domain [3][4]. Unfortunately, they are not suitable for non-stationary signal analysis [5]. In order to solve this problem, Wavelet Transform (WT) has been developed. The WT, also called time-frequency analysis, is a kind of variable window technology, which uses a time interval to analyze the frequency components of the signal. This makes the application of the WT for non-stationary signal processing an area of active research over the past decade. An overview of the WT used in vibration signal analysis was provided in [6][7][8].

The original signal using WT can be decomposed into approximations and details versions with different frequency bands by using a successive low-pass and high-pass filtering. The decomposed levels will not change their information in the time domain [9]. However, useful information can be contained in some sub-bands. So, the fault can be detected from a given level of resolution. This is based on a choice of an indicator to determine the optimal level where failure can occur. The selection of the most reliable indicator has been studied by several authors. Prabhakar et al. [10] selected the periodic impulses of bearing faults in time domain based on low and high frequency nature of decomposed levels. Similar analyses were carried out by Purushotham et al. [11] in order to extract the periodic impulses from the time signals using discrete wavelet transform at Mel-frequency scales. Chinmaya and Mohanty [12] used the sidebands of the gear meshing frequencies as an evaluation criterion for gear faults diagnosis. Djebala et al. [13] analyzed the vibration of faults inducing periodical impulsive forces by selecting the kurtosis as indicator.

In this work, the measured vibration signals are decomposed using the Daubechies wavelet. Clearly, useful information is contained in some decomposition levels. In order to extract useful information, the energy distribution is established by Parseval's theorem. The latter is used as principal criterion to select the optimal level of resolution. The proposed method is evaluated using the vibration measurements obtained from accelerometer sensors. The aim of this method is to provide a solution of bearing fault diagnosis.

The remainder of this paper is structured as follows. Section II presents the experimental rig used. Section III describes the fault diagnosis method. Results and discussion are presented in Section IV. Finally, the main conclusions are outlined in Section V.

II. EXPERIMENT DATA ACQUISITION

Vibrations caused by defective bearing elements account for the vast majority of problems with rotating machinery. Each element such as inner race or outer race has a characteristic rotational frequency. With a fault on a particular element, an increase in the vibration energy at this element rotational frequency may occur. The monitoring of these elements has a primary importance for the correct operation of the machine.

The experimental measurements presented in this paper are entirely based on the vibration data obtained from the

 ISBN: 978-1-61208-419-0

Case Western Reserve University Bearing Data Centre [14]. As shown in Figure 1, the motor is connected to a dynamometer and torque sensor by a self-aligning coupling. The vibration signals were collected from an accelerometer mounted on the motor housing at the drive end of the motor. The vibration data was obtained from the experimental system under the four different operating conditions: (1) normal condition; (2) with inner race fault; (3) with outer race fault; and (4) with ball fault. The data is sampled at a rate of 12 kHz and the duration of each vibration signal was 10 seconds. More details about experimental setup were reported in [14].

The bearings used in this study are deep groove ball bearings manufactured by SKF. Faults were introduced to the test bearings using electro-discharge machining method. The defect diameters of the three faults were the same: 0.018, 0.036, and 0.053 mm. The motor speed during the experimental tests is 1797–1720 rpm. Each bearing was tested under the four different loads: 0, 1, 2, and 3 horse power (hp).

In order to evaluate the proposed method, the data measured under 0-load (0 hp) at rotation speed of 1797 rpm (30 Hz) including the faults on the inner and outer races were used. The original signal is divided into segments of samples that each sample covered 4096 data points.

Figures 2a, 2b and 2c represent respectively a vibration signal collected at 1797 rpm from the normal state, inner race fault and outer race fault.

The fault frequency can be calculated from the geometry of the bearing and element rotational speed. Frequencies associated with defective inner and outer races are as follows:

$$f_{IR} = (n/2) f_r \left(1 + (d/D)\cos\alpha\right) \quad (1)$$

$$f_{OR} = (n/2) f_r \left(1 - (d/D)\cos\alpha\right) \quad (2)$$

where, f_r is the rotational frequency, d the ball diameter, D the pitch diameter, n the number of balls and α the contact angle.

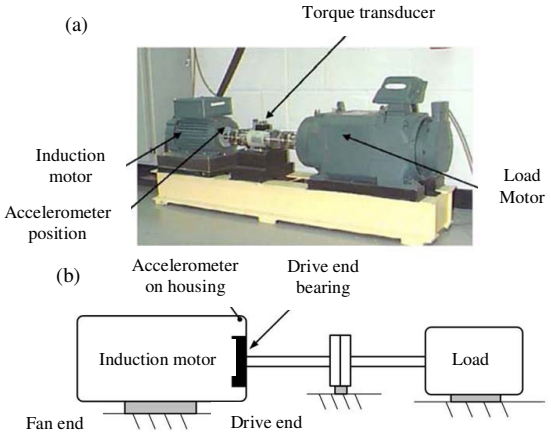


Figure 1. (a) Bearing test rig and (b) its schematic description [15].

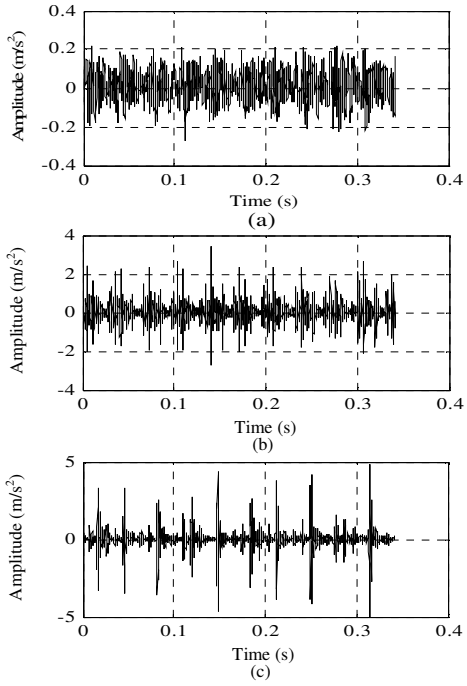


Figure 2. Vibration signals of: a) normal state, b) inner race fault and c) outer race fault.

The fault frequencies of inner race and outer race are calculated, respectively, according to (1) and (2), which are 162 Hz and 107 Hz.

III. FAULT DIAGNOSIS METHOD

In this section, a diagnosis method, which consists of two approaches, namely, the WT and Parseval's theorem, is described to monitor the bearing inner and outer races.

A. Wavelet Transform

The WT is one of the most important methods in signal analysis. It is a time-frequency analysis technique. Due to its strong capability in time and frequency domain, it is applied recently by many researchers in rotating machinery. The WT decomposes a signal in both time and frequency in terms of a wavelet, called mother wavelet (3). The mother wavelet must be compactly supported and satisfied with the admissibility condition (4).

$$\psi(t) = 1/(\sqrt{a})\,\psi((t-b)/a) \quad (3)$$

$$\int_{-\infty}^{+\infty} |\hat{\psi}(w)|^2 / |w| dw < \infty \quad (4)$$

where $\hat{\psi}(w)$ is the Fourier transformation of $\psi(t)$.

Two variations of the WT exist: Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). They are described below: let $s(t)$ be the original signal, the CWT of $s(t)$ is defined as:

$$CWT(a,b) = 1/(\sqrt{|a|}) \int_{-\infty}^{\infty} s(t)\psi^*((t-b)/a)\,dt \quad (5)$$

where * denotes complex conjugate, a and b are the dilation (scaling) and translation (shift) parameters, respectively.

The DWT is derived from the discretization of the CWT by discrete values of a and b. The DWT is given by:

$$DWT(j,k) = 1/(\sqrt{2^j}) \int_{-\infty}^{\infty} s(t)\psi^*((t-2^j k)/2^j)\,dt \quad (6)$$

where a and b are replaced by 2^j and $2^j k$, j is an integer.

The DWT can be regarded as a multiresolution analysis technique [16], as illustrated in Figure 3. The DWT analyzes the signal at different scales or resolutions. It employs two sets of functions, called scaling functions and wavelet functions [16][17], which are associated with low pass (L) and high pass (H) filters, respectively. The discrete signal is convolved with L and H, resulting in two vectors *A1* and *D1* on a first level. The vector *A1* is called approximation and the vector *D1* is called detail. The application of the same transform on the approximation *A1* causes it to be decomposed further into approximation *A2* and detail *D2* on a second level. Finally, the signal is decomposed at the expected level.

The selection of the appropriate wavelet is very important in signals analysis. There are many functions available can be used, such as Haar, Daubechies, Meyer, and Morlet functions [18][19]. In the present study, we use the Daubechies wavelet to identify the inner and outer races bearing frequencies.

B. *Parseval's Theorem*

The Parseval's theorem refers to the result that the sum of square of a function is equal to the sum of the square of its transform.

In the wavelet domain, the Parseval's theorem can be defined as the energy of a function in the time domain is equal to the sum of all energy concentrated in the different decomposition levels. This can be described by [20]:

$$\sum_{1}^{N} |s(t)|^2 = \sum_{1}^{N} |A_m(t)|^2 + \sum_{1}^{m}\sum_{1}^{N} |D_m(t)|^2 \quad (7)$$

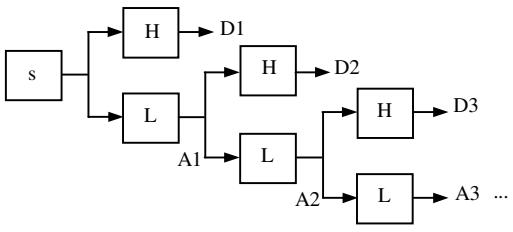


Figure 3. Principal of DWT decomposition.

where N is the number of samples and m is the maximum level of wavelet decomposition. The left-hand term of (7) represents the total energy of the signal $s(t)$, the first and the second term on the right denote respectively, the total energy of the approximation in the level m and the total energy of the detail from level 1 to m.

The time domain information will not be lost when the signal is decomposed. In order to extract the maximum information in the different resolution levels, the energy distribution of the approximation and the detail of the signal is calculated. It is given by:

$$P_a = \frac{\|A\|^2}{N_m} \quad (8)$$

$$P_d = \frac{\|D_m\|^2}{N_m} \quad (9)$$

where $\|\ \|$ denotes the norm operator.

IV. MONITORING RESULTS

The proposed method is applied to the diagnosis of the SKF bearing with inner race fault and outer race fault. The motor runs at a speed of 1797 rpm (30 Hz).

The multiresolution analysis is applied by using the Daubechies wavelet of order 4 (db4). Here, level 4 decomposition is employed to extract approximations and details coefficients from vibration signals. The result of db4 decomposition is given in Figures 4 and 5, respectively.

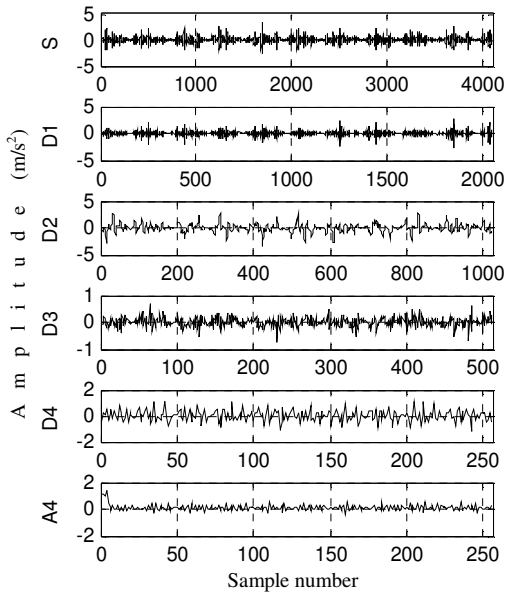


Figure 4. Wavelet decomposition of inner race fault.

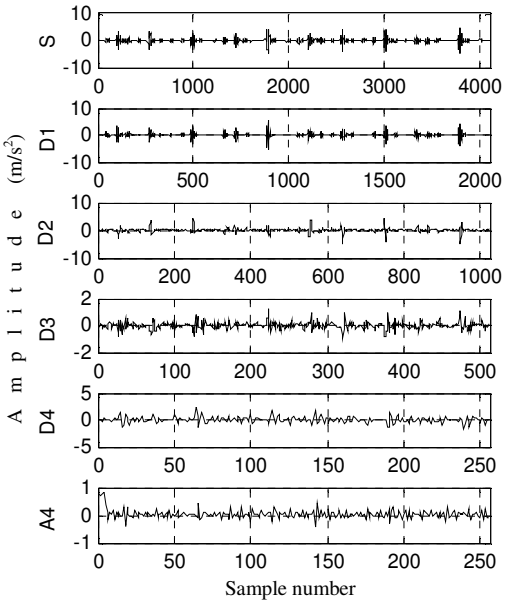


Figure 5. Wavelet decomposion of outer outer fault.

The objective of the proposed method is to demonstrate the effectiveness of the energy distribution as principal criterion for selecting the optimal decomposition level. The level having the largest value indicates the desired level.

The energy distribution of each level is shown in Figure 6. The decomposition levels 1 to 4 represent the detailed version and the levels 5 stand for the approximated version of the signal. The figure shows the obvious difference between levels. From this figure, it can be seen that the energy distribution using db4 occurs in the second level for each fault. So, our choice is attached to the detail *D2*.

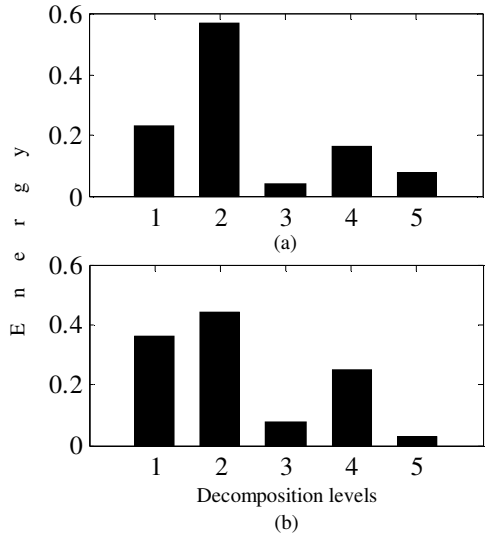


Figure 6. Energy distribution: a) inner race and b) outer race.

In order to diagnose the inner race fault and the outer race fault from the selected level, we use envelope analysis. Figures 7a and 8a show respectively the selected decomposition level (*D2*) of inner race fault and outer race fault. It is clear that this level shows the shocks generated by the considered faults.

Figures 7b and 8b illustrate respectively the envelope spectrum of *D2* of inner race fault and outer race fault. The frequency spectra clearly show many frequency components, at the rotation frequency (30 Hz), also at the characteristic frequencies of the inner race (162 Hz) and the outer race (107 Hz) and their harmonics, which indicates a defective bearing.

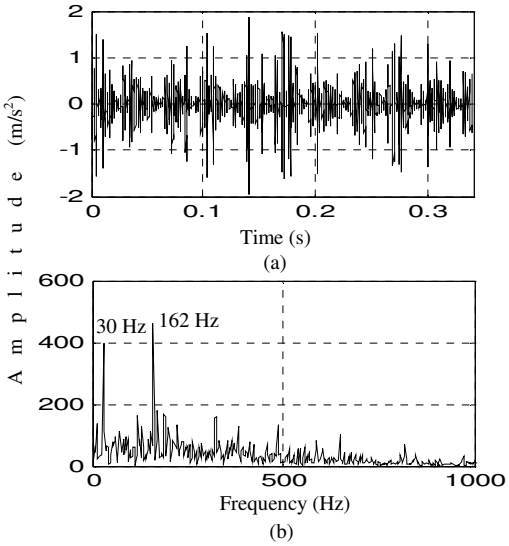


Figure 7. (a) Selected level of inner race fault and (b) its envelope spectrum.

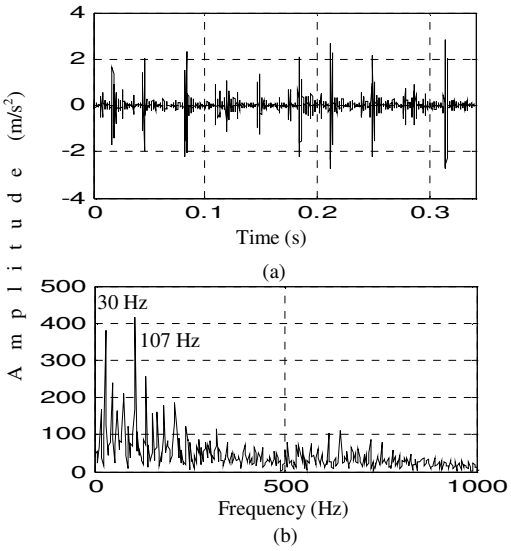


Figure 8. (a) Selected level of outer race fault and (b) its envelope spectrum.

 ISBN: 978-1-61208-419-0

V. CONCLUSION AND FUTURE WORK

This paper presented a method for improving the bearing fault diagnosis based on WT and Parseval's theorem. It is adapted to obtain multiple data series at different resolutions by wavelet decomposition and calculate the energy distribution using Parseval's theorem in order to select the optimal decomposition level, for a possible diagnosis. A case study on SKF bearing diagnosis with defective inner race and outer race has shown that this method can greatly improve the accuracy of diagnosis. Hence, the proposed method is a successful approach for vibration monitoring. It remains to test its application on a signal containing other types of faults.

REFERENCES

[1] J. Altmann, Application of discrete wavelet packet analysis for the detection and diagnosis of low speed rolling-element bearing faults, Ph.D. thesis, Monash University, Melbourne, Australia, 1999.

[2] H. Yang, Automatic fault diagnosis of rolling element bearings using wavelet based pursuit features, Ph.D. thesis, Queensland University of Technology, Australia, 2004.

[3] S. Seker and E. Ayaz, "A study on condition monitoring for induction motors under the accelerated aging processes," IEEE Power Engineering, vol. 22, no. 7, 2002, pp. 35-37.

[4] K. Shibata, A. Takahashi, and T. Shirai, "Fault diagnosis of rotating machinery through visualisation of sound signal," Mechanical Systems and Signal Processing, vol. 14, 2000, pp. 229-241.

[5] J. D. Wu and C. H. Liu, "Investigation of engine fault diagnosis using discrete wavelet transform and neural network," Expert Systems with Applications, vol. 35, 2008, pp. 1200-1213.

[6] G. K. Singh and S. A. S. Al Kazzaz, "Isolation and identification of dry bearing faults in induction machine using wavelet transform," Tribology International, vol. 42, 2009, pp. 849-861.

[7] H. Bendjama, S. Bouhouche, and M. S. Boucherit, "Vibration monitoring for fault diagnosis in rotating machinery using wavelet transform," Proc. of the International Conference on Advanced Computer Theory and Engineering, ASME Press, Dec 2011, pp. 167-170, ISBN: 978-0-7918-5993-3.

[8] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," Signal Processing, vol. 96, 2014, pp. 1-15.

[9] Z. L. Gaing, "Wavelet-based neural network for power disturbance recognition and classification," IEEE Transactions on Power Delivery, vol. 19, 2004, pp. 1560-1568.

[10] S. Prabhakar, A. R. Mohanty, and A. S. Sekhar, "Application of discrete wavelet transform for detection of ball bearing race faults," Tribology International, vol. 35, 2002, pp. 793-800.

[11] V. Purushotham, S. Narayanan, and S. A. N. Prasad, "Multi-fault diagnosis of rolling bearing elements using wavelet analysis and hidden Markov model based fault recognition," NDT&E International, vol. 38, 2005, pp. 654-664.

[12] K. Chinmaya and A. R. Mohanty, "Monitoring gear vibrations through motor current signature analysis and wavelet transform," Mechanical Systems and Signal Processing, vol. 20, no. 1, 2006, pp. 158-187.

[13] A. Djebala, N. Ouelaa, and N. Hamzaoui, "Detection of rolling bearing defects using discrete wavelet analysis," Meccanica, vol. 43, 2008, pp. 339-348.

[14] K. A. Loparo, Bearings vibration data set, Case Western Reserve University, Available from: http://www.eecs.cwru.edu, 2003, [retrieved: January, 2010].

[15] Y. Huang, C. Liu, X. F. Zha, and Y. Li, "A lean model for performance assessment of machinery using second generation wavelet packet transform and Fisher criterion", Expert Systems with Applications, vol. 37, 2010, pp. 3815-3822.

[16] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Trans Pattern Anal Machine Intelligence, vol. 11, no. 7, 1989, pp. 674-693.

[17] N. Lu, F. Wang, and F. Gao, "Combination method of principal component and wavelet analysis for multivariate process monitoring and fault diagnosis," Industrial & Engineering Chemistry Research, vol. 42, 2003, pp. 4198-4207.

[18] C. K. Chui, An introduction to wavelets, Academic Press, New York, 1992.

[19] I. Daubechies, "Orthonormal bases of compactly supported wavelets," Communication on Pure and Applied Mathematics, vol. 41, 1988, pp. 909-996.

[20] A. M. Gaouda, M. M. A. Salma, M. R. Sultan, and A.Y. Chikhani, "Power quality detection and classification using wavelet-multiresolution signal decomposition," IEEE Transaction on Power Delivery, vol.14, no.4, 1999, pp. 1469-1476.

 ISBN: 978-1-61208-419-0

Application of Copulas in Analysis of Drought and Irrigation

Milan Cisty, Lubomir Celar and Anna Becova
Department of Land and Water Resources Management
Slovak University of Technology
Bratislava, Slovakia
Email: milan.cisty@stuba.sk, lubomir.celar@stuba.sk, anna.becova@stuba.sk

***Abstract*—This paper presents probabilistic analysis of the occurrence of irrigation needs. We have conducted a joint analysis of the severity and duration of the most demanding potential annual irrigation periods by a bivariate copula methodology. The characteristics of these periods are derived from both temperature and precipitation. The maximum annual length of the potential irrigation periods and the corresponding rainfall deficit were inferred from basic climatic variables such as inputs for a two-dimensional probability analysis by a copula methodology. The results of this work indicate the suitability of the proposed methodology for an analysis of irrigation needs with greater benefits than in the case of the usual one-dimensional analysis of individual climatic variables. A case study with the aim of testing the methodology was accomplished in southwest Slovakia, where a frequency analysis of the need for irrigation was estimated. The results indicate, e.g., that every second year, a period can be expected in which temperatures above 25°C occur and which lasts one month with a moisture deficit of about 30 mm. Even more significant periods of drought can be expected, for example, with a 5 or 10-year return period. These phenomena cause significant damage to agriculture yields in the territory investigated, so a requirement for irrigation structures in this area is indicated by the proposed methodology.**

Keywords–drought; irrigation; copula; precipitation; temperature

I. INTRODUCTION

Water scarcity and droughts have a direct impact on the inhabitants and various economic sectors of a region which use and depend on water, such as agriculture, tourism, industry, energy or transport. Quantifying the expected probability characteristics of droughts assists in the planning and management of water resources, such as the design and maintenance of irrigation systems. The issue of how to characterize a drought is often dealt with through the help of various drought identification indices [1]. The numerous indices of drought that may be mentioned include, for example, the decile index (DI) [2], the percentage of normality (PN), the standardized precipitation index (SPI) [3], the Palmer PDSI index [4], and the effective drought index (EDI) [5]. Among the above–mentioned drought indices, the standardized precipitation index (SPI) is most frequently used.

Research on the probabilistic characterizations of droughts was formerly conducted using a univariate analysis [1], [6]. However, drought is a multidimensional phenomenon characterized by, for example, its severity, duration and intensity, so it is necessary to examine the properties of dry episodes using multidimensional methods [7]. For this reason, the traditional drought risk assessment based on univariate frequency analyses may lead to erroneous or incomplete conclusions about the occurrence of drought events [8]. Over the last decade, copulas have emerged as a method for addressing multivariate problems in several disciplines. Probabilistic analysis using a copula method has various positive features; the main one is that it does not assume that the variables have the same types of probability distribution functions [9]. Copulas have been adopted for hydrological studies of multivariate flood frequency analyses [10]–[14] and rainfall frequency analyses [15]–[17].

In this paper, we have chosen a different approach, which is intended for an analysis of irrigation needs and is oriented towards the point of view of the necessity for the construction of an irrigation system in a given area. Various drought indices used in previous studies are designed for the identification of drought months, not for the identification of the necessity to irrigate, which should be analyzed on a timescale of days or weeks, not months. We applied a novel approach and directed our research towards an analysis of the severity and duration of the most demanding annual potential irrigation periods.

Although previous studies have used multivariate analysis, they usually only investigated the lack of precipitation, - e.g., the duration, severity or intensity of dry periods [3], [6], [13]. In the present paper, both the temperature and precipitation are included in the analyses as will be explained in the methodology part. To evaluate the expected occurrence of periods with an increased need for irrigation, a two–dimensional analysis has been applied to the distribution of two variables, which together characterize expectations about the occurrence of episodes which require irrigation. The first variable is the length of the maximal potential irrigation period mentioned, i.e., the maximum number of consecutive summer days in a year. The second variable is the rainfall deficit during this time interval.

In Section 2 of the paper, a description of the area and data studied follows; then in Section 3, assessments of both one–dimensional and bivariate probability distribution functions are described. The results are presented in tabular and graphic forms in Section 4, and the paper ends with the conclusions (Section 5) from the research presented.

II. STUDY AREA AND DATA

The analysis was carried out on an agricultural area in Slovakia with a warm and relatively dry climate–the area of the Danubian Lowland, namely, its central part around the municipality of Hurbanovo (Figure 1). The weather in this area

 ISBN: 978-1-61208-419-0

is a transition between oceanic and terrestrial influences. The annual average temperature of a substantial part of the lowland ranges between 9°C and 10°C. In terms of precipitation, it is the driest part of Slovakia with an average annual rainfall of 550 mm to 650 mm.

The analyses were accomplished using climatic data from the period 1930–2013. In the analyses, daily temperature and precipitation data were used.

Figure 1. Study area location

Basic climatic variables (temperature and precipitation) were used to determine the two derived and actually used variables which, in this paper, characterize dry and hot periods requiring irrigation. As already mentioned, such a period is defined by its duration and the rainfall deficit with respect to the normal period (1960–1990). For each year, the hot and dry periods that lasted the longest were identified. The duration was derived from the number of consecutive days with temperatures above 25°C. The hot period identified was extended by precipitation-free days before and after it. In the following, this variable is referred to as the maximum annual length of the potential irrigation period. Although plants have, to a certain extent, the ability to adapt to periods with a lack of moisture, a long duration of this period usually requires irrigation if a reduction in yields is to be prevented, especially if these periods occur in the important growth stages of plants.

III. METHODOLOGY

The procedure used to achieve the objective of this paper, e.g., a joint analysis of the two variables introduced in the previous text, was the following: 1) the preparation of the datasets of the variables investigated; 2) a verification of the dependence and relationships between the variables; 3) the identification of one-dimensional distributions of the selected variables; 4) identification of the expected class of the copulas forming a two-dimensional probabilistic dependence; 5) the determination of the copula parameters, e.g., the fitting and evaluation of the best suitable copula; and 6) the specification of the return periods with critical temperature and rainfall-deficit characteristics.

A. Specification of the one-dimensional probability distribution functions

One-dimensional distributions are required to determine the probabilistic characteristics of the individual variables that are used to describe the properties of dry and hot periods with potential irrigation requirements. In the context of this paper, they serve as the means to determine the so-called marginal functions needed in defining a two-dimensional probability.

The fitting process can be divided into three steps: 1) selecting an appropriate probability distribution function; 2) determining its parameters; 3) verifying the quality of the fitting by the appropriate statistical characteristics.

A preliminary selection of the candidate probability distribution functions was performed based on a data analysis employing descriptive statistics and graphic techniques as well as on the existing literature on the probability distribution fittings of the variables describing a drought [1], [18]–[20].

The parameters of the probability functions were determined using the maximum likelihood method (MLE) [21].

The quality of the selection of the type of distribution functions and its fitting could be evaluated by the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) [21]. The quality of the fitting is also verified in the paper using the Kolmogorov-Smirnov test and the Anderson-Darling test [21].

B. Joint probability distribution specification using copulas

When modelling with copula operators, it is first necessary to conduct certain tests of the relationship between the variables under study. The application part of this paper uses the Kendall correlation test and a multivariate test of independence based on an empirical copula process which was proposed by [22] and is often used to test independence in copula modelling (for example, [23] or [24]).

The advantages of copulas for constructing multivariate distributions lie in the fact that the multidimensional modelling of a distribution can be decomposed into a separate determination of the one-dimensional marginal functions of the variables examined and a separately conducted search of the dependencies between them using copulas [25].

The essence of modelling a two-dimensional relationship between two variables by means of copulas is based on Sklar's theorem (1959), which mathematically justifies the intuitive principle specified in the previous paragraph.

For two variables, Sklar's theorem [25] states that if $F_{X,Y}(x, y)$ is a joint distribution function of bivariate random variables (X, Y) with marginal distributions $F_X x$ and $F_Y y$ respectively, then there exists a copula function $C(.)$ such that:

$$F_{X,Y}(x, y) = C(F_x(x), F_Y(y)) \quad (1)$$

If both $F_X(x)$ and $F_Y(y)$ are continuous distributions, then this copula is unique for the particular joint distribution.

To perform probabilistic analyses of a drought, it is necessary to select an appropriate copula function on the basis of certain principles. There are many varieties of copulas which, based on common features, belong to several classes. Among the most widely used are included, for example, elliptical copulas, the Ali-Mikhail-Haq (AMH) copula, the Clayton, Frank, Galambos, Gaussian, Gumbel-Hougaard, Joe and Plackett copulas [25].

The choice of the proper copula is based on various factors, such as the scope of the dependence to be described by the copula. In this paper, we selected one parametric copula, specified by parameter Θ.

For each choice it is necessary to optimally determine the Θ copula parameter. In this paper, we used the values of the em-

 ISBN: 978-1-61208-419-0

pirical marginal one-dimensional distribution functions so that the choice of the family of the parametric marginal distribution does not affect the search for the Θ copula parameter. This includes finding the ranks of the individual values of the data and scaling them to the interval (0, 1). The actual calculations were performed using the copula and acopula packages for the R language [26], [27]. There are several methods of copula fitting such as the maximum likelihood method, the inversion of Kendall's τ, and the inversion of Spearman's ρ. This work uses the maximum likelihood method.

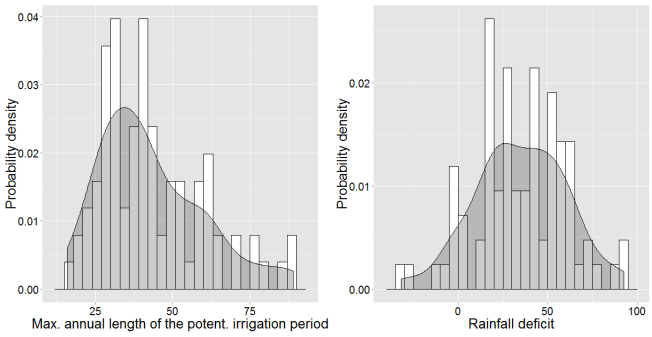


Figure 2. Histogram and kernel-estimated nonparametric probability distribution functions

The best fitting of a copula model to describe the relationship between both characteristics of a drought was verified using multiple criteria, namely by application of the AIC and the BIC. There are several alternatives for testing the goodness of fit, and active research is being done in this field. We used the test referenced in [28], which also provides an overview of the other tests for this purpose. In this procedure a nonparametric empirical copula was computed and compared with the values of the parametric copulas. The parametric copula that was closest to the empirical copula was defined as the most appropriate choice [29].

IV. Results

The analyzed data, which are the two time series derived from the temperatures and precipitation at the Hurbanovo climatological measuring station, have been described in the "Study area and data" section. Figure 2 contains a histogram with kernel–estimated nonparametric probability distributions of both variables. On that basis, it could be expected that, due to the different shapes of the kernel distribution function, these two variables will be described by different parametric probability distribution functions. This means that the joint probability of these variables should be determined by a copula methodology (as opposed to standard multivariate probability distributions, which assume the same distribution for all jointly evaluated variables).

TABLE I. The Coefficients of the Correlation Between the Variables Studied

Correlation coefficient	
Pearson	0.392
Kendall	0.287
Spearman	0.399

The two data series examined were successfully tested for independence using a nonparametric test based on an empirical copula according to [22], in which the output statistics of Harald Cramér are used [30]. The Mann-Kendall test for correlation was also carried out with the alternative hypothesis that the true τ is greater than zero, which was accepted (for the null hypothesis of the equality of the correlation to zero) on the basis of the p-value = 6.37810^{-05}.

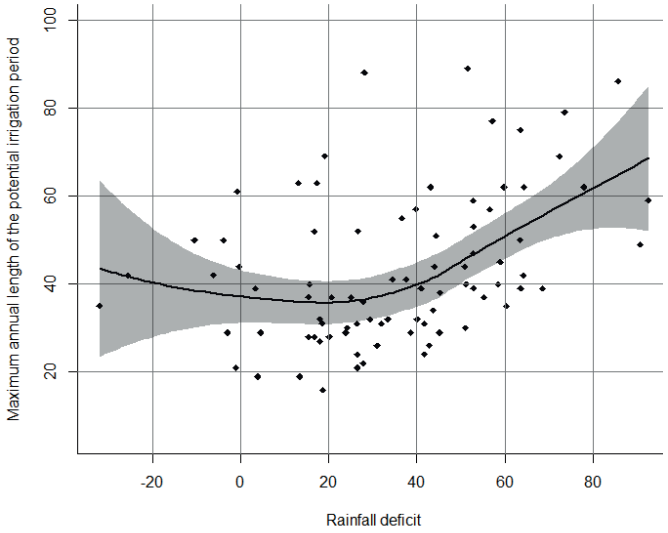


Figure 3. Data compared by a nonparametric Loess regression line

TABLE II. Evaluation of the Distribution Function for the Rainfall Deficit in the Maximum Potential Annual Irrigation Period

Evaluation of the rainfall deficit fitting						
	Statistical indicator				p-value	
Distribution:	KS Test	AD Test	AIC	BIC	AD Test	KS Test
GEV	0.10	1.34	799	804	0.22	0.30
Normal	**0.06**	**0.23**	**784**	**789**	**0.98**	**0.92**
Log-Logistic	0.07	0.32	786	791	0.92	0.79
Cauchy	0.09	1.51	817	822	0.17	0.52
Gumbel	0.10	1.34	799	804	0.22	0.30

The values of the correlation are shown in Table 1; in Figure 3, the two variables are plotted with a graphic representation of the non-linear dependence using a Loess curve with a confidence level of 0.95. Both the table and the picture show that the relationship between the variables is not very strong but can be detected.

The results of this testing could be summarized as follows: the simultaneous probability of these two variables will not be equal to their product, but should be modelled using joint bivariate probability distributions (copulas).

The next step in the analysis is to determine the one-dimensional probability distribution functions of both variables studied. Based on the above reasons, the probability distribution functions according to Table 2 were selected as the candidate probability distributions for the rainfall deficit. Table 2 also includes an evaluation using the methods described in the methodological section of this article. Table 3 serves the same purpose, except it is for the "maximum annual length of the potential irrigation period" variable. In both tables, the bold typeface indicates the selected distribution based on the values of the statistical indicators and p-values; a normal distribution was chosen for the rainfall deficit, and the logarithmic Pearson type III distribution was chosen for the length of the maximum potential annual irrigation period.

To construct an associated probability distribution function, several one-parameter copulas were evaluated (Gumbel, Clay-

 ISBN: 978-1-61208-419-0

ton, Frank, Ali-Mikhail-Haq, Joe, Normal, t-copula, Plackett and Hussler-Reiss). The results and evaluation of the copula function fitting calculations are shown in Table 4.

The main criterion for the selection of a suitable copula operator from Table 4 was the p-value obtained from the goodness-of-fit test according to [28], which uses a parametric bootstrap and is mentioned in the methodological part about fitting the copula function. The copula parameter was determined in the test using the inverse Kendall's τ method. The values of the AIC and BIC served as the auxiliary criteria. The table shows that the Gumbel, Joe, and Husler-Reiss copulas are the admissible copula functions. On the basis of the highest p-value, the Joe copula was selected to describe the dependence.

TABLE III. EVALUATION OF THE DISTRIBUTION FUNCTION FOR THE LENGTH OF THE MAXIMUM POTENTIAL ANNUAL IRRIGATION PERIOD

Evaluation of the rainfall deficit fitting						
	Statistical indicator				p-value	
Distribution:	KS Test	AD Test	AIC	BIC	AD Test	KS Test
Gumbel	0.08	0.39	700	705	0.86	0.72
Pearson III	0.08	0.55	701	706	0.70	0.59
Gamma	0.08	0.55	701	706	0.70	0.59
Logn w.3p	0.06	0.28	700	705	0.95	0.95
Log-P.III	**0.06**	**0.28**	**705**	**705**	**0.95**	**0.88**

TABLE IV. EVALUATION OF FITTING THE COPULA OPERATOR

Copula class	AIC	BIC	θ parameter	p-value
Gumbel	-12.88	-10.45	1.400	0.120
Clayton	-4.10	-1.67	0.800	0.001
Frank	-12.08	-9.65	2.757	0.049
AMH	-8.13	-5.70	0.915	0.013
Joe	**-12.59**	**-10.16**	**1.719**	**0.229**
Normal	-11.88	-9.46	0.434	0.027
t-copula	7.36	-4.94	0.434	0.041
Plackett	-11.58	-9.15	3.723	0.041
Husler-Reiss	-13.11	-10.69	1.074	0.137

In water resources management, the results of a probabilistic analysis are usually expressed using the concept of return periods. These correspond to the long-term average time between two successive occurrences of a certain event. For the problem considered in this paper, the return values of the two variables which are useful in assessing the need for irrigation and for dimensioning the components of an irrigation project were evaluated.

A multivariate analysis of a phenomenon in comparison with a one-dimensional analysis has a distinctive feature in that a combination of variables with different values can lead to the same joint probability and, consequently, to the same return period. In Figure 4, this feature is expressed by means of a contour plot. The chart in this figure shows the return periods of the two variables studied simultaneously in this paper, using the contour lines of the return periods. Some selected results of the joint analysis are also shown in Table 5. The probability of variable values occurring simultaneously can be expressed by the following relationship [15]:

$$T_{x,y} = \frac{1}{1 - F_{(x)} - F_{(y)} + C\left(F_{(x)}.F_{(y)}\right)} \quad (2)$$

where $F_{(x)}$ and $F_{(y)}$ are one-dimensional probability distribution functions, and $C(F_{(x)}, F_{(y)})$ represents a joint distribution function based on the Joe copula.

TABLE V. JOINT RETURN PERIODS OF SOME SELECTED VALUES OF THE VARIABLES INVESTIGATED

Probability	One-dimensional return period	Length of the potential irrigation period	Rainfall deficit	Joint return period
	years	*days*	*mm*	*years*
0.5	2	39	35	4
0.8	5	55	56	9
0.9	10	66	67	20
0.95	20	78	76	39

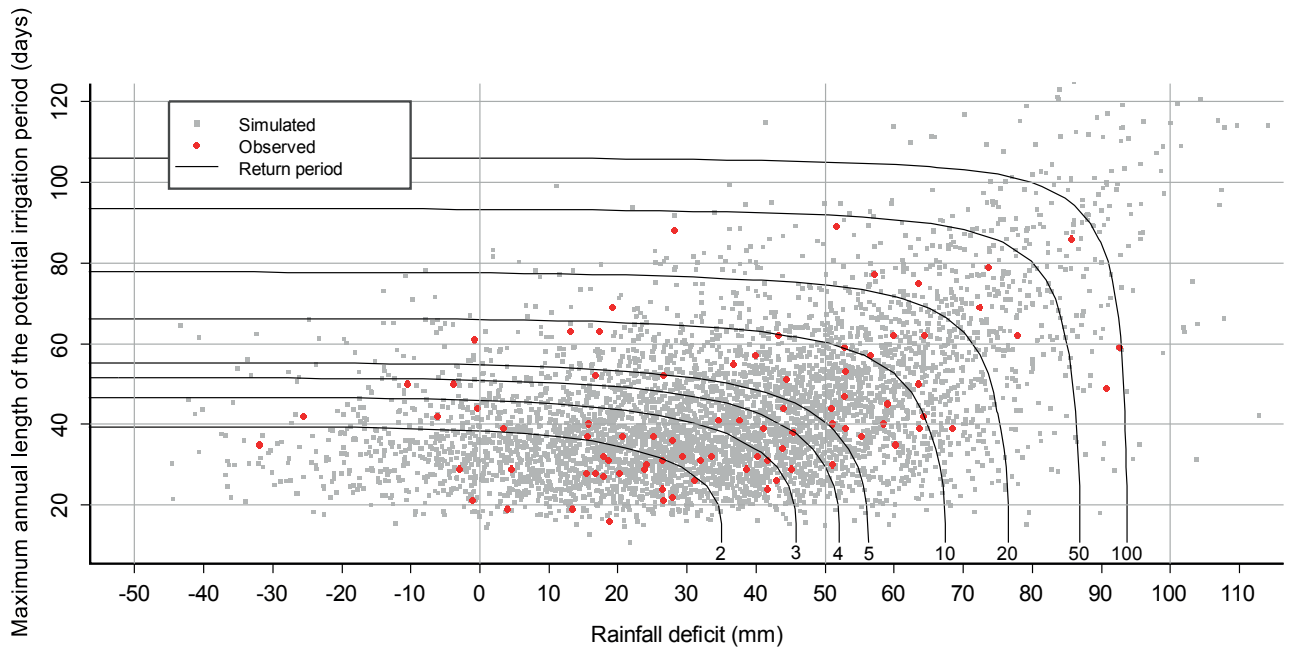


Figure 4. Joint return period of the variables studied for the simultaneous attainment of the corresponding values expressed by the return period contours (in years)

 ISBN: 978-1-61208-419-0

V. CONCLUSION

The results of this work (Table 5, Figure 4) indicate that in the context of the case study accomplished in south-west Slovakia, the need for irrigation occurs very often. Every second year, for example, a period can be expected in which temperatures above 25°C occur, and a dry period usually lasts one month with a moisture deficit of about 30 mm. Months of the growing season with rainfall totals smaller than 50 mm are considered to be those with irrigation needs. A precipitation of 80 mm in such a period (which would be needed to maintain this limit) occurs with a probability in the upper quartile, i.e., it is very rare. Even more significant periods of drought can be expected, for example, with a 5 or 10-year return period. These phenomena result in significant damage to agriculture yields, which, as is often declared in the domestic water management community, are greater than the investment needed for the reliable maintenance or reconstruction of irrigation systems.

ACKNOWLEDGMENT

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0665/15.

REFERENCES

[1] P. Angelidis, F. Maris, N. Kotsovinos, and V. Hrissanthou, "Computation of drought index SPI with alternative distribution functions," Water Resour Manag, vol. 26, 2012, pp. 2453–2473.

[2] A. Shaban, "Indicators and aspects of hydrological draught in Lebanon," Water Resour Manag, vol. 23, 2009, pp. 1875–1891.

[3] F. Yusof, F. Hui-Mean, J. Suhaila, and Z. Yusof, "Characterisation of drought properties with bivariate copula analysis," Water Resour Manag, vol. 27, 2013, pp. 4183–4207.

[4] W. Palmer, "Meteorological drought." US Department of Commerce, Weather Bureau, Tech. Rep. 45, 1965, [retrieved: June, 2015]. [Online]. Available: http://www.ncdc.noaa.gov/temp-and-precip/drought/docs/palmer.pdf

[5] S. Morid, V. Smakhtin, and M. Moghaddasi, "Comparison of seven meteorological indices for drought monitoring in Iran," Int J Climatol, vol. 26, 2006, pp. 971–985.

[6] J. Santos, M. Portela, and I. Pulido-Calvo, "Regional frequency analysis of droughts in Portugal," Water Resour Manag, vol. 25, 2011, pp. 3537–3558.

[7] J. Zhai, B. Liu, H. Hartmann, B. Su, T. Jiang, and K. Fraedrich, "Dryness/wetness variations in China during the first 50 years of the 21st century," Hydrol Earth Syst Sci Discuss, vol. 6, 2009, pp. 1385–1409.

[8] L. Vergni, F. Todisco, and F. Mannocchi, "Analysis of agricultural drought characteristics through a two–dimensional copula," Water Resour Manag, vol. 29, 2015, pp. 2819–2835.

[9] G. Salvadori and C. D. Michele, "Frequency analysis via copulas: Theoretical aspects and applications to hydrological events," Water Resour Res, vol. 40, 2004, W12511.

[10] A. Favre, S. E. Adlouni, L. Perreault, N. Thiemonge, and B. Bobée, "Multivariate hydrological frequency analysis using copulas," Water Resour Res, vol. 40, 2004, pp. 1–20.

[11] M. J. Reddy and P. Ganguli, "Bivariate flood frequency analysis of upper Godavari River flows using Archimedean copulas," Water Resour Manag, vol. 26, no. 14, 2012, pp. 3995–4018.

[12] J. Shiau and R. Modarres, "Copula-based drought severity-duration-frequency analysis in Iran," Meteorol Appl, vol. 16, 2009, pp. 481–489.

[13] N. Bezak, M. Miko, and M. Šraj, "Trivariate frequency analyses of peak discharge, hydrograph volume and suspended sediment concentration data using copulas," Water Resour Manag, vol. 28, 2014, pp. 2195–2212.

[14] S. Kao and R. Govindaraju, "A bivariate frequency analysis of extreme rainfall with implications for design," J Geophys Res-Atmos, vol. 112, 2007, D13119.

[15] C. D. Michele and G. Salvadori, "A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas," J Geophys Res-Atmos, vol. 108, 2003, pp. 1–9.

[16] L. Zhang and V. Singh, "Bivariate rainfall frequency distributions using Archimedean copulas," J Hydrol, vol. 332, 2007, pp. 93–109.

[17] U. Rauf and P. Zeephongsekul, "Analysis of rainfall severity and duration in Victoria, Australia using non-parametric copulas and marginal distributions," Water Resour Manag, vol. 28, 2014, pp. 4835–4856.

[18] X. Lana, M. Martnez, A. Burgueño, C. Serra, J. Martín-Vide, and L. Gómez, "Distributions of long dry spells in the Iberian peninsula, years 1951-1990," Int J Climatol, vol. 26, 2006, pp. 1999–2021.

[19] Q. Zhang, V. Singh, Y. Chen, and M. Xiao, "Regional frequency analysis of droughts in China: A multivariate perspective," Water Resour Manag, vol. 29, 2015, pp. 1767–1787.

[20] S. Nadarajah, "A bivariate distribution with gamma and beta marginals with application to drought data," J Appl Stat, vol. 36, 2009, pp. 277–301.

[21] A. Hayter, Probability and Statistics for Engineers and Scientists. Cengage Learning, 2006.

[22] C. Genest and B. Rémillard, "Test of independence and randomness based on the empirical copula process," Test, vol. 13, 2004, pp. 335–369.

[23] X. Wang, M. Gebremichael, and J. Yan, "Weighted likelihood copula modeling of extreme rainfall events in Connecticut," J Hydrol, vol. 390, 2010, pp. 108–115.

[24] T. Bacigal, V. Jagr, and R. Mesiar, "Non-exchangeable random variables, Archimax copulas and their fitting to real data," Kybernetika, vol. 47, 2011, pp. 519–531.

[25] R. Nelsen, An introduction to copulas. Springer, 2006.

[26] T. Bacigal, "R Package to handle Archimax or any user-defined continuous copula construction: acopula. In: Bustince H, Fernandez J, Mesiar R, Calvo T (eds) Aggregation functions in theory and in practise," Advances in intelligent systems and computing, vol. 26, 2013, pp. 75–84.

[27] I. Kojadinovic and J. Yan, "Modeling multivariate distributions with continuous margins using the copula R package," J Stat Softw, vol. 34, 2010, pp. 1–20.

[28] C. Genest, B. Rémillard, and D. Beaudoin, "Goodness-of-fit tests for copulas: A review and a power study," Mathematics and Economics, vol. 44, 2009, pp. 199–213.

[29] C. Genest and A. Favre, "Everything you always wanted to know about copula modeling but were afraid to ask," J Hydrol Eng, vol. 12, 2007, pp. 347–368.

[30] H. Cramér, "On the composition of elementary errors," Scandinavian Actuarial Journal, vol. 11, 1928, pp. 141–180.

 ISBN: 978-1-61208-419-0

Towards Coordinated Task Scheduling in Virtualized Systems

Jérémy Fanguède, Alexander Spyridakis and Daniel Raho
Virtual Open Systems
Grenoble - France
Email: {j.fanguede, a.spyridakis, s.raho}@virtualopensystems.com

***Abstract*—Task scheduling is one of the key subsystems of an operating system. Generally, by providing fairness in terms of processor time allocated to tasks, the task scheduler can guarantee a low latency and high responsiveness to applications. In this paper, we demonstrate that some problems can occur in virtualized environments, in relation to standard task scheduler implementations and the way that tasks and virtual cores are scheduled. More precisely, there is a need to implement a communication channel between virtualized schedulers in the virtual machines and the host task scheduler, particularly when full-virtualization techniques are used, which could lead to latency issues and loss of responsiveness in virtual machines, especially when processors execute excessive workloads. After having analyzed the potential problems in virtual machines, experiments were done with real world and benchmarking applications. For testing, a Linux-based system and two different task schedulers were used, with a benchmark suite especially designed for virtualized environments where application responsiveness and latency can be measured. As an experimental platform, an ARM embedded system was used; this system is almost equivalent to general-purpose systems in terms of task scheduling.**

***Keywords*—KVM/ARM; embedded virtualization; coordinated scheduling; embedded systems; task scheduling; CFS; BFS**

I. INTRODUCTION

Virtualization technology offers a way to increase efficiency and adaptability both in general purpose and embedded systems, but to get an efficient virtualization solution, latency of virtual machines and responsiveness of applications should be guaranteed at a reasonable level. For instance, an interactive application launched in a virtual machine should not have much worse performance in terms of responsiveness and latency than one executed in a host machine in the same conditions.

In previous work, we have already experimented with this objective in mind, specifically for storage-I/O, and the implementation of Virtual-BFQ [1] [2], a Linux I/O scheduler based on the BFQ scheduler [3]. The work described in this paper, instead targets process scheduling, so that it could be used as a complementary approach.

In this paper, we provide the following contributions:

A. Contributions of this paper

We highlight that in virtualized environments there are latency problems with task scheduling, where a missing link between the guest and the host scheduler can affect performance negatively. In fact, there is a need to implement a coordinated communication channel between schedulers in virtual machines and the host task scheduler. As a consequence, latency of a guest operating system is higher, especially in a system with many CPU-bound tasks. This results in degraded responsiveness of applications in virtual machines, compared to similar conditions for non-virtualized systems. To show this problem, through experimentation, we use two different Linux task schedulers.

Then, experimental results are reported, these results confirm that, in virtualized environments, when a process requires a high portion of the processor's time in both the guest and host system, the latency and the responsiveness of the guest application is not guaranteed.

An ARM-based embedded system was used to run the experiments, Kernel-based Virtual Machine (KVM) and Quick EMUlator (QEMU) is the virtualization solution used, which is among the most popular solutions in embedded virtualization.

B. Organization of this paper

The paper is organized as follows. In section II, a description of the two task schedulers used is provided. Then in Section III latency problems and the lack of responsiveness is highlighted. After describing the benchmark suite and the experimentation methods in section IV, the results are reported in Section V. Finally, in SectionVI, possible solutions are detailed in order to solve the issue highlighted.

II. LINUX TASK SCHEDULERS

The task scheduler, also named process or CPU scheduler, is the part of an operating system that decides which task runs when, and on which core. The job of a scheduler is to share the CPU time between processes that require CPU resources, to pick a suitable task to run next if required, and to balance processes between the different CPUs in a multi-core system.

Two Linux task schedulers were used, CFS [4], which stands for *Completely Fair Scheduler* and is the default scheduler of the Linux kernel, and BFS [5], which stands for *Brain Fuck Scheduler* and is a popular alternative.

By default, Linux can handle real-time and non real-time policies, which are implemented by the selected scheduler. Both CFS and BFS schedulers implement their own non real-time and share the same real-time policies. By extension, with the term CFS or BFS we refer to both scheduling policies of these schedulers, as well as the whole of their implementation.

 ISBN: 978-1-61208-419-0

BFS, which is not part of the Linux mainline kernel [6], could be considered as an alternative, it is designed for desktop interactivity on machines with few cores [5], and its source code has a smaller footprint and is by design simpler. For these reasons, this scheduler was also selected for investigation.

A. The Completely Fair Scheduler

The default Linux kernel scheduler, named *Completely Fair Scheduler* [4], is modular and permits to use different policies for different tasks. Linux has two main types of scheduling policies: a real-time one for real-time task and a normal one named *fair policy* for all other tasks.

Among the real time scheduling, Linux distinguishes three policies: SCHED_FIFO, a first-in, first-out policy; SCHED_RR, a round robin policy; and SCHED_DEADLINE, a policy implementing the earliest deadline first algorithm (since kernel v3.14).

And within the fair scheduling policies: SCHED_NORMAL, the default Linux time-sharing policy, and SCHED_BATCH, a policy for "batch" processes.

Linux defines the static priority of a task by a value, which ranges from 0 to 99, and the real-time scheduling class uses values from 1 (lowest priority) to 99 (highest priority). Processes using the fair scheduling class have necessarily a static priority of 0. In order to determine which thread (or process) should be run next, the Linux scheduler maintains a list of runnable processes for each possible static priority, and it selects the head of the list with the highest static priority. In other words, a thread, with a higher static priority than the current running thread which becomes runnable, will necessarily preempt the current process.

For the fair scheduling class, the kernel uses a priority called dynamic priority, which from a user's point of view is also better known as the *nice value*, and it ranges from -20 (highest priority) to +19.

CFS is used as the default Linux scheduler since kernel version v2.6.23, it replaced the old scheduler: *O(1)*. And implements a completely fair algorithm (hence the name). The algorithm is based on the concept of an ideal multi-tasking processor. With such a processor, each runnable task would run at the same time, sharing the processor power. Of course this behavior is not possible, but an equivalent behavior, would be to run each runnable task for an infinitesimal amount of time with full processing power. Due to task switching cost, CFS, only approximates this behavior.

For that purpose, CFS stores the runtime value of each task in a variable called *vruntime* (stands for virtual runtime) and tries to keep all *vruntime* values the closer to each other. So the runnable task which has the lower *vruntime* value is chosen to be the next task to run. The priority of a task (the dynamic priority, i.e., the nice value) influences the way vruntime is increased.

To handle interactive tasks, CFS doesn't use complex heuristics. In fact, the concept of fair scheduling is enough to maximize interface performance. For example, consider a processor-bound task (e.g., an encryption calculation, a video encoder, etc.) and a I/O-bound task (e.g., a terminal, a text editor, etc.), which will be the interactive task. In that situation, the scheduler should give to the interactive task a larger share of the processor time to enhance the user experience. In fact, this is what CFS will do: CFS wants to be fair, so each time the interactive task become runnable, CFS will see that this task consumed significantly less processor time than the CPU-bound task. So the interactive task will preempt the other, and will be executed until its runtime reaches the value of the processor-bound task or be blocked from an I/O request.

B. BFS - The Alternative

BFS is an alternative to CFS, it was written by *Con Kolivas*. It is not in the mainline kernel and is available as source code patches [6].

BFS focuses on a simplistic design (about 2.5 times fewer lines of code than CFS) and aims for excellent desktop interactivity and responsiveness on personal computers with a reasonable amount of cores [5]. It uses a single work-queue, *O(n)* look-up for all cores unlike CFS, and implements the *earliest eligible virtual deadline first* algorithm for non real-time policies.

BFS, like CFS, provides real-time task policies: SCHED_FIFO and SCHED_RR, and also two others policies for normal tasks: SCHED_ISO and SCHED_IDLEPRIO. The first, SCHED_ISO (for isochronous) is designed to provide "near real-time" performance to unprivileged users. And SCHED_IDLEPRIO scheduling policy can be used to run tasks only when the CPU would be idle otherwise.

The design of BFS makes it efficient when the number of running processes is small (inferior than the number of CPUs), which is normally, according to its author [5], a common use case for a desktop computer.

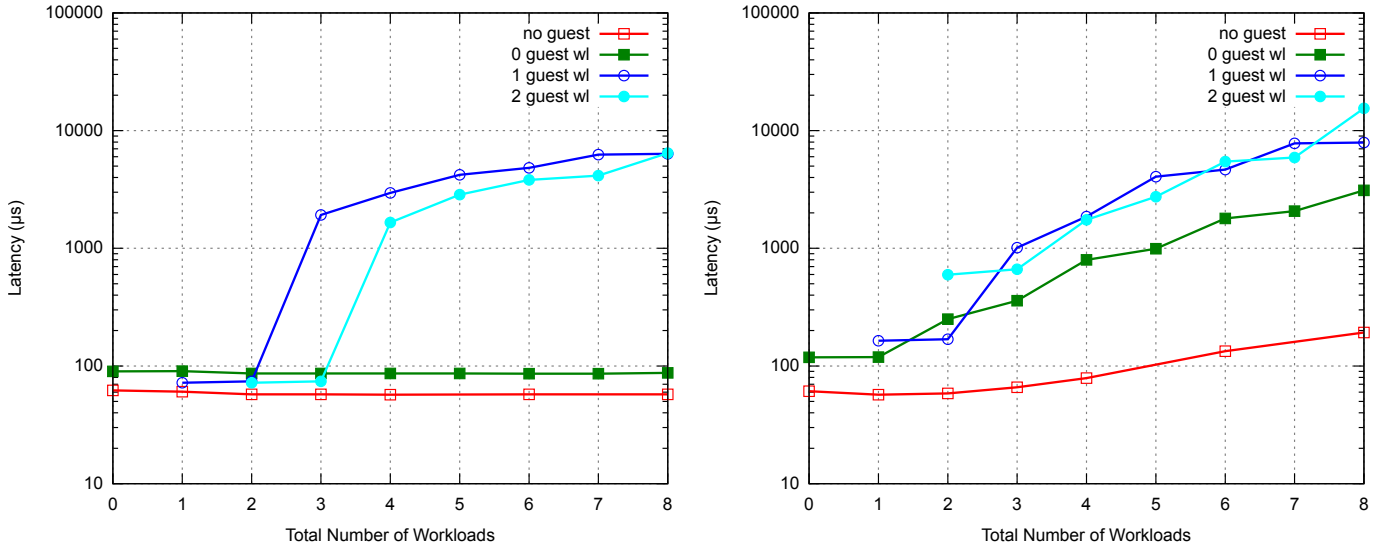
III. Potential Problems in Virtualized Environments

In a virtualized environment a guest system is seen, from the host scheduler, as just one, or more additional jobs to schedule, without any awareness from the host of the fine-grained requirements of the corresponding guest scheduler. For example, a new spawned task in the guest system could be scheduled in a different way by the guest scheduler, but this information is not visible on the host side. Under certain conditions, that could lead to undesired behavior.

To highlight the problem we can consider a system, with two physical CPUs and a guest with one virtual CPU. Two CPU-bound workloads are launched in the host (one per CPU) and one in the guest (one per virtual-CPU). In this situation, the task scheduler will share fairly the processor time between the vCPU thread (which runs a workload) and the two workloads in the host, since these three tasks are quite similar in terms of CPU time demand.

When an interactive task is started in the guest system, the guest scheduler will *detect* this new task and assign a substantial amount of the vCPU time compared to the work-load running in the same guest. On the host side though, the scheduler sees only three processes that request a large amount of CPU time for only two CPUs. So, the host scheduler has absolutely no reasons to privilege the vCPU thread compared

 ISBN: 978-1-61208-419-0



(a) CFS

(b) BFS

Figure 1. Latency results

to other processes (workloads). Additionally, the latency of this interactive task will probably be higher than in a host system with the same number of workloads (aside from the constant overhead of KVM/QEMU). This problem persists for whatever value the priority of the interactive task in the guest is set to (could be a real-time one), since the priorities and policies are not made aware to the host system.

IV. Experiment Method and Benchmark Suite

To highlight the problem described above, we set-up a benchmark suite in order to measure, in particular, the latency of the system. We use the tool, *cyclictest*, which is usually used to measure latency on a *Real Time Linux* (i.e. patched with *rt* patches) [7]. Generally, *cyclictest* is used to measure the latency of real-time thread/process (schedule with SCHED_FIFO or SCHED_RR), but we can also use it with *normal* (SCHED_NORMAL) threads. For each latency measurement *cyclictest* is run twice, each one with a 100000 loop, which means that the latency provided by the benchmark is the average of 200 thousands measurements. The following command line is used "cyclictest -q -n -l 100000 -h 5000" to generate the results, and the latency histogram is also retrieved (*-h* option) in order to analyze in more detail.

The second kind of benchmark measures the *start-up time* of an application. We simply measure how long it takes from when an application is launched to when an application is ready. This benchmark gives an idea of the *responsiveness* of an application. The start-up time is measured with hot caches, to avoid any I/O perturbations. For each configuration (i.e. number of workload in the host and guest), 100 measurement iterations are performed, and the average, as well as the standard deviation are retrieved.

As workload, we use a simple program that does an *infinite loop*, and therefore has a very low memory footprint.

V. Experimental Results

We executed our experiments on a Samsung Chromebook equipped with an ARMv7-A Cortex-A15 processor (dual-core, 1.7 GHz) and 2 GB of RAM. Both the host and the guest run upstream Linux v3.17 with the *PREEMPT* configuration option enabled.

A. Latency

In order to measure latency, we used the *cyclictest* tool and the number of workloads is kept the same as in the start-up time test. The result of this experiment is shown in Figure 1, where latency is measured in microseconds and represented in a logarithmic scale on axis Y.

For the host and guest system we employ up to 8 and 2 workloads respectively. Axis X corresponds to the total number of workloads, i.e., host plus guest workloads. The output of the results are four different curves:

no guest: No virtual machine, serves as reference, the application is launched in the host

N guest wl: With N workloads in the guest, the application is launched in the guest. With N ranges from **0** to **2**.

We can notice that, with the CFS scheduler (Figure 1a), as soon as there are more workloads than physical cores (total of two cores in the system, critical curves are *1 guest wl* and *2 guest wl*) and with at least one workload in the guest, latency increases significantly. By adding more workloads, this behavior persists until values are not suitable for interactive usage. This kind of result confirms the issue highlighted in Section III, where an interactive application in a virtualized system can have an extremely high latency.

One can also point out that the latency is better with *2 guest workloads* than with *1 guest workload* when the total number of workloads is high. This behavior is perfectly explainable

 ISBN: 978-1-61208-419-0

due to the difference in the number of workloads in the host. For instance, in the specific case of 4 total workloads, when we have *1 guest workload* the host system *sees* four main processes requesting a high amount of CPU time for only two CPUs, but when we have *2 guest workloads*, there are only three processes that still share two CPUs. In the latter case, the process corresponding to the vCPU has more CPU time: this could lead, depending on the efficiency of the guest scheduler, to a better latency compared to the former case.

With BFS (Figure 1b), results are less obvious, but we can still notice the difference between virtualized and normal environments, and between the curves of 1 or 2 guest workloads and the curve of 0 guest workloads.

Although our objective is not to purely compare the two schedulers, which has already been done [8], we can remark that even with no virtual machines (curve *no guest*), latency with BFS increases steadily, contrary to CFS. This is probably due to the fact that BFS is not designed to be efficient when the number of running tasks is higher than the number of physical cores [5].

We can also analyze the histogram provided by the *cyclictest* results to compare the distribution of latency. Figure 2 shows the two latency histograms on a virtual machine without any workload. We can notice that even if the average value is slightly lower with CFS, the BFS case exposes more converged values with a lower maximum.

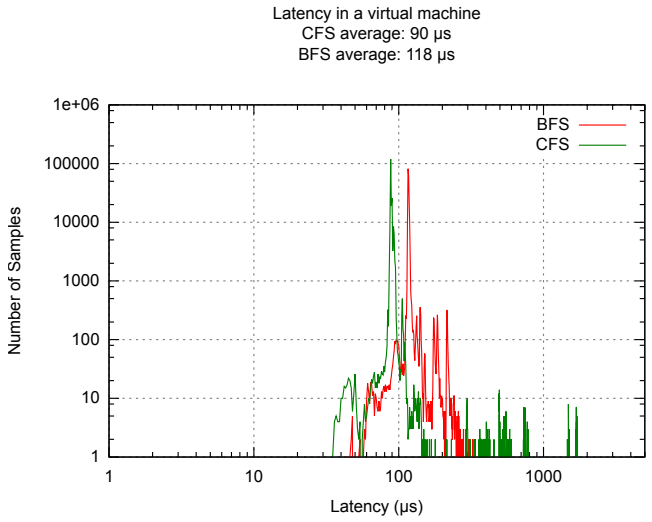


Figure 2. Guest Latency compared between BFS and CFS

In Figure 3, two cases are compared for CFS latency measured in a virtual machine. Both test cases have the same amount of CPU-bound workloads, but distributed in a different manner. In the first case all workloads reside in the host, while in the second, one of the workloads is reserved for the guest. Although the distribution of samples for low latency is quite similar for both cases, in the case where one of the workloads is in the guest, we still observe a significant amount of samples in the range of 200 to 5000 μs. This is different from the first case, where almost all samples are around the 100 μs mark.

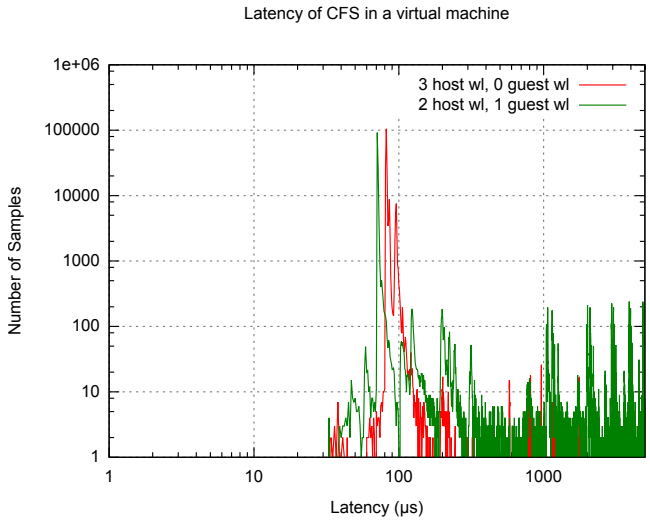


Figure 3. CFS latency in a virtual machine, compared between 3 host workloads and 2 host and 1 guest workloads

B. Start-up Time

Next, we measure the start-up time of an application. We choose the *xterm* application because its start-up time can be easily measured. In addition, this application was also selected to measure performance of the BFQ and Virtual-BFQ I/O scheduler [1] [2] [3].

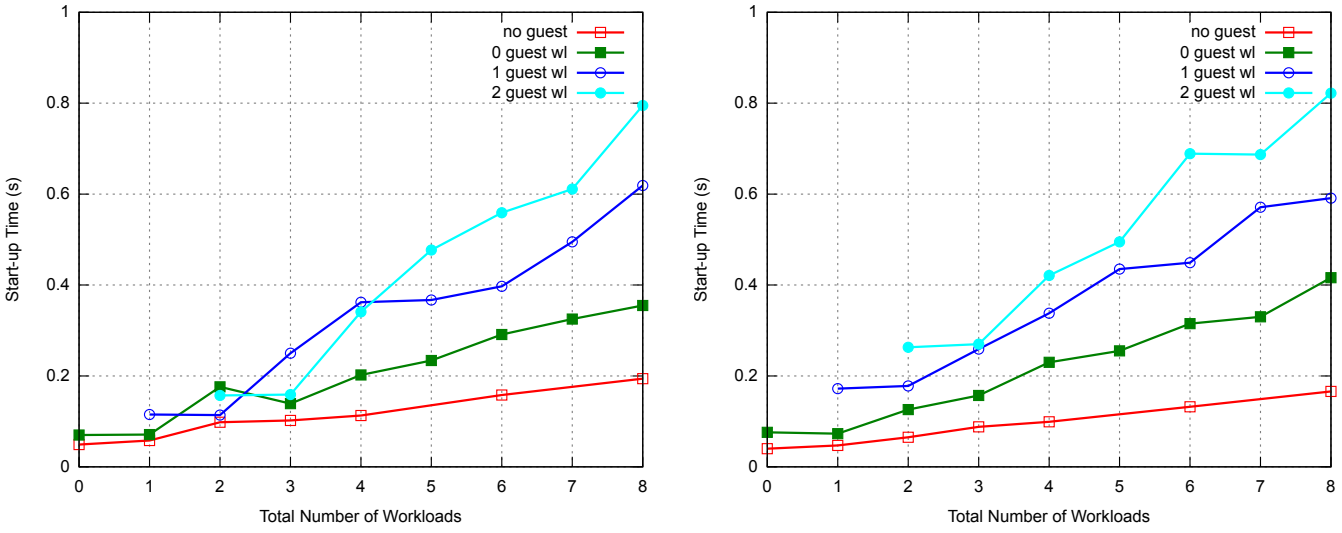
As we can see in this Figure 4a, which represents the startup time measured with the CFS scheduler, the curve corresponding to a measurement in the host (*no guest*) has a slightly positive constant slope. This increase is not unexpected because CFS tries to guarantee only fairness: an increase in the number of CPU-bound can negatively affect the start-up time of a new application. Curve *0 guest wl*, corresponds to the case in which there is no workload in the guest, but only in the host. We can see that this curve almost follows curve *no guest*, where a constant overhead is observed.

In view of the problem highlighted above, the critical scenarios are the ones corresponding to the curves *1 guest wl* and *2 guest wl*, more particularly when the number of workloads in the host is equal or greater than number of physical cores (in our case 2). In fact, when vCPU threads are allowed to use all available cores, the results are acceptable as the start-up time remains quite low (case *1 guest wl* with a total workload of *1* and *2*, and with *2 guest wl* with *2* and *3* total workloads). To summarize our test case results, when the number of workloads in the host is higher than two, the start-up time increases significantly.

With the BFS scheduler (Figure 4b), although the appearance of the curves seems quite different, we have the same behavior: higher start-up times when there are too many workloads.

To sum up, our results are coherent both for start-up times as well as latency. Moreover they clearly prove that, in scenario where a workload is present in both guest and host, the responsiveness of an application in the guest can not be guaranteed.

 ISBN: 978-1-61208-419-0



(a) CFS

(b) BFS

Figure 4. Start-up time results

VI. SOLUTIONS

Some solutions, can be developed to address the above problem. Two possible solutions are described below, a simple static prioritizing, and a coordinated solution that enables a communication between scheduler of the host and guests systems.

A. Static prioritizing

A straightforward solution could be a static prioritizing scheme, by simply increasing the priority of the QEMU vCPU threads, or by changing the scheduling policy to a real-time one. This solution will allow QEMU to not be interfered by other tasks in the host system (if there are no other real-time threads). This method will result in a better latency, in particular a reduction of the maximum latency [9]. With this solution though, the guest is always privileged even when it doesn't execute an interactive program. This solution can be useful for simple use cases, i.e, when a guest system which executes soft real-time applications needs to be prioritized compared to other guests or applications.

B. Coordinated scheduling

Instead of prioritizing QEMU threads statically, another solution could be to *boost* these threads only when it is necessary, i.e, temporary increasing the priority or changing the scheduler policy, when the guest system requests it. It is a sort of dynamic prioritizing with a coordinated scheduling mechanism: the guest kernel detects when it needs higher priority, and informs the host system about it. This *co-scheduling* mechanism was already implemented successfully for Virtual-BFQ [2], therefore the communication mechanism could be equivalent to the one developed for this storage I/O scheduler.

This type of solution has already been implemented and evaluated, especially to make KVM a real-time hypervisor [10] [11]. Such attempts mainly focused to run a real-time Linux OS as a guest, thus when a guest executes a real-time thread it informs the host of its current scheduling policy and priority, the host system then has to pass on this policy and priority to the affected QEMU thread.

In order to extend this coordinated scheduling mechanism also to non real-time applications, a mechanism to detect interactive application in the guest system is needed. Heuristic algorithms have to be added for this purpose.

The communication mechanism between the host and guest scheduler, is a crucial part, it needs to be fast or at least not too frequent. The solution chosen in the Virtual-BFQ [2] I/O scheduler is to use, a special ARM instruction, *HVC*, that results in a hypervisor *trap*. Moreover, the cost of calling this instruction, around 2000 CPU cycles, is not very expensive and can fit the requirement of a task scheduling coordinated mechanism.

VII. CONCLUSION AND FUTURE WORKS

In virtualized environments, we highlighted that the host task scheduler could fail to achieve full system low latency, and thus to preserve responsiveness when the system is loaded with CPU-bound programs in certain conditions. The behavior of an interactive application inside a guest will be masked by other processes requiring a lot of CPU time in the host, and the attempts of the guest scheduler to enhance the responsiveness of this application may be useless. This issue mostly occurs when the number of CPU-bound processes is higher than the number of physical cores.

We are currently designing a solution which implements a coordinated scheduling mechanism between the host and guests schedulers, and we have promising results. The target of this approach is ARM embedded systems with the KVM hypervisor. Besides, we also plan to extend tests with more complex scenarios including more than one virtual machines.

 ISBN: 978-1-61208-419-0

ACKNOWLEDGMENT

This research work has been supported by the Seventh Framework Programme (FP7/2007-2013) of the European Community under the grant agreement no. 610640 for the DREAMS project.

REFERENCES

[1] A. Spyridakis, and D. Raho. "On Application Responsiveness and Storage Latency in Virtualized Environments," CLOUD COMPUTING 2014, The Fifth International Conference on Cloud Computing, GRIDs, and Virtualization, 2014, pp. 26-30.

[2] A. Spyridakis, D. Raho, and J. Fanguède. "Virtual-BFQ: A Coordinated Scheduler to Minimize Storage Latency and Improve Application Responsiveness in Virtualized Systems," International Journal on Advances in Software, vol 7 no 3 & 4, 2014, pp. 642-652.

[3] P. Valente and M. Andreolini, "Improving application responsiveness with the BFQ disk I/O scheduler," Proceedings of the 5th Annual International Systems and Storage Conference (SYSTOR'12), June 2012, p. 6.

[4] "CFS scheduler," [retrieved: June 2014]. Available: http://lwn.net/Articles/230501/

[5] C. Kolivas, "BFS FAQ," [retrieved: June 2014]. Available: http://ck.kolivas.org/patches/bfs/bfs-faq.txt.

[6] C. Kolivas, "BFS Patches," [retrieved: June 2014]. Available: http://ck.kolivas.org/patches/bfs/3.0/.

[7] "Cyclictest," [retrieved: June 2014]. Available: https://rt.wiki.kernel.org/index.php/Cyclictest

[8] T. Groves, J. Knockel, and E. Schulte. "Bfs vs. cfs scheduler comparison," 2009.

[9] R. Ma, F. Zhou, E. Zhu, and H. Guan, "Performance Tuning Towards a KVM-based Embedded Real-Time Virtualization System," Journal of Information Science and Engineering 29.5, 2013, pp. 1021-1035.

[10] J. Kiszka, "Towards linux as a real-time hypervisor," In Proceedings of the 11th Real-Time Linux Workshop, 2009, pp. 215-224.

[11] Aichouch, Mehdi, J-C. Prevotet, and Fabienne Nouvel. "Evaluation of an RTOS on top of a hosted virtual machine system," In Design and Architectures for Signal and Image Processing (DASIP), 2013 Conference on. IEEE, 2013, pp. 290-297.

 ISBN: 978-1-61208-419-0

PLC and Its Applications : A Wireless and Automatic Pet-Feeding System for Rabbits

Hsin-Ching Chiang
Dept. of Medical Informatics, Chung Shan Medical University
Taiwan 40201, ROC
chivialk9913@gmail.com

Tzu-Fang Sheu
Dept. of Computer Science and Communication Engineering, Providence University
Taiwan 40201, ROC
tfsheu@gmail.com

Hsiao-Ping Lee
Dept. of Medical Informatics, Chung Shan Medical University
Taiwan 40201, ROC
shopping.lee@gmail.com

Yi-Hsin Chen
Dept. of Computer Science and Communication Engineering, Providence University
Taiwan 40201, ROC
cin0213@gmail.com

***Abstract*—In this paper we design a Supervisory Control and Data Acquisition system, which includes newly designed feeding device, programmable logic controller, a graphical human machine interface programmed with Visual Basic, plus, the Internet communication module. By using our system, the users only need a computer or a hand-held device that can connect to the Internet to achieve the goal of remote controlling and monitoring the feeding device.**

Keywords- programmable logic contorller (PLC); wireless; supervisory control and data acquisition (SCADA)

I. INTRODUCTION

According to Taiwan Taipei Rabbit Society Associations statistics, the number of people who keep a rabbit is up to 420,000 in 2010 in Taiwan [1]. The ranking of the most popular pets are led by dogs followed by cats and rabbits at the third place. The number of rabbit owners has increased rapidly around the world in the last few years. Since 2005, people begin to celebrate the "Worldwide Rabbit Love and Appreciation Day" in August [2]. However, Taipei Rabbit Society Associations warns whomever keeps a rabbit, need to take good care of it, since they are not as strong as cats or dogs.

Based on the statistics, the quantity of rabbits dying from abandoning and bad care is up to 17,000 per year. That is to say, nearly 67 percent of total death is caused by misconception [1]. Most of the abandoned rabbits die of heart paralysis or shock, due to their timid nature. However, holding a rabbit as a pet is not too difficult, since they only need to be fed twice a day. This also includes the supply with a sufficient amount of water. The purpose of the automatic feeder we proposed in this paper is to support people to maintain the basic needs of their pet.

The remote control and monitor system of this feeder is established by using Internet, Global System for Mobile Communication (GSM) and integrated Programmable Logic Controller (PLC). The whole system contains two parts, a remote control subsystem and a supervisory subsystem. Within the GSM covering range, complex logical actions can be controlled, instead of just operating an electrical switch.

In our study, a traditional PLC is used to control the hardware, and a simple Supervisory Control and Data Acquisition (SCADA) application software which can be used on mobile device and PC. Eventually, with the four wireless data transmission's major systems (Wideband Code Division Multiple Access (W-CDMA) [8], Code Division Multiple Access (CDMA) [9], General Packet Radio Service (GPRS) [10] and Personal Handy-phone System (PHS) [11]) cooperating and taking action, will turn the network transmit protocol (TCP/IP) into wireless operation. We will introduce the hardware we use in Section 2 and the prototype machine in Section 3.

II. MATERIALS

A. PLC

PLC is an electronic device with digital movement, its hardware design can be basically divided into two kinds, medium-large sized and miniature sized. Instead of mechanical equipment, the control function is reached by using integrated circuits and digital and analogy I\O modules [3]. PLC possesses order, timing, counting, computing, data process and communication, etc. The commands are stored in EPROMs. With the scientific and technological progress, the enhancement of CPU and single chip function, also improve PLC's abilities, functions and attachments can be offered by it thereupon increase.

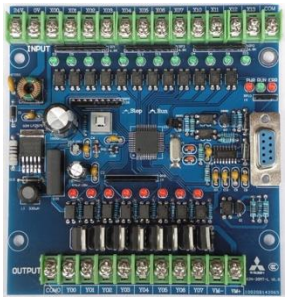


Figure 1. Mitsubishi PLC (FX2N-20MT-L)

 ISBN: 978-1-61208-419-0

1) Mitsubishi's PLC Communication Protacal

The PLC device we chose is one of the Mitsubishi FX2N series which is shown in Figure 1. It is built as a master-slave-system. The PLC device is the slave, other device connected to the PLC is the master, for instance, the server that receive command from the client. In a master-slave-system the master sends commands to the slave systems. These slaves only send feedbacks only if receiving a request from the master. Hence data collisions are avoided. PLC device set up link through RS232, and PLC's Wi-Fi attachment, so we can let users control and manage the system in a browser or software by using a personal computer or mobile device. All functions can be controlled remotely like set up, daily arrange, food fall/stop, etc.

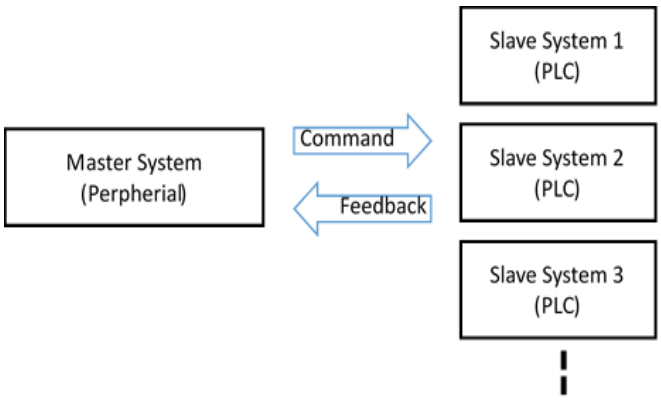


Figure 2. Fatek's PLC communication message format

2) Mitsubishi's PLC Communication Message Format

For possible kinds of messages, commands or feedbacks, the message can be split into 5 - 7 fields due to its command. All of these fields uses ASCII 16 bit code. In all cases the beginning letter is STX (02H). Field CMD are used to classify the command type (Table 1.), for example read/write, forced ON/OFF, etc. After asking data from Y0~Y7 (0A-00) and Y10~Y17(0A-01), which is shown in Figure 6. The slave system will pass the message back along with the data to the master system as shown in Figure 7.

Function	CMD	Assign to	Explain
Read	0	X,Y,M,X,T,C,D	Read Data
Write	1	X,Y,M,X,T,C,D	Write Data
Force ON	7	X,Y,M,X,T,C	Force Node ON
Force OFF	8	X,Y,M,X,T,C	Force Node OFF

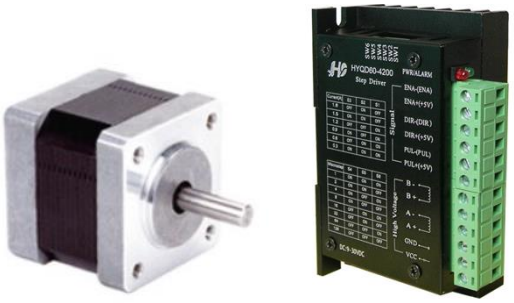
Table 1. PLC CMD code

The Device address is divided into two fields, and need to be read from the back, for example, Figure 10 is a command of forcing node Y1 (0501) to turn on, so the device address will be 0105. If the command is asking for data writing the fifth field will be the data that needs to be written, as shown in Figure 9, which is trying to write a value 3586. Next to it is the end-message code ETX (03H), let the system know the where the command ended. The last field SUM CHEK = CMD + DEVICE ADDRESS + BYTES + ETX, and only use the last two words, for example 30H+30H+30H+41H+30H+30H+32H+03H='166'H, so the check sum will be '66'H. For writing and force function command, the PLC will answer YES (06H) or NO (15H).

B. Stepper Motor And Motor Driver

Stepper motor is divided by the steps they need to complete a full rotation. The motor we use is a simplest two-phase stepper motor, it rotates 1.8 degree (±5%) every step. So, there will be 360/1.8 = 200 steps per full rotation. That is why the motor can finish simple but high accurate rotate/stop positioning. Advantages of stepper motors are low cost, high reliability, high torque at low speeds and a simple, rugged construction that operates in almost any environment.

After triggered by users, PLC will send the operate command with demanded motor speed and rotate quantity to the motor driver. The driver will converts the PLC command signals into electric pulse, which is provided by external power source, to energize the motor windings. The stepper motor then converts digital pulses into mechanical shaft rotation.



(a) (b)

Figure 3. (a) 35mm Series Motor (b) Motor driver

III. METHOD

A. Prototype Machine

The prototype machine is consists of PLC, motor driver, stepper motor, 50W external power source and home-made feeder shell. The PLC will send the signal to motor driver, the driver will then control the stepper motor to rotate the component which is attach on it. After stepping a few times, the pet food will fall into the bowl. The quantity of the food can be controlled by counting the motor's steps or timing the rotation.

Considering rabbits chewing habit, especially when they are hungry and it contains tempting smell of the food inside, the material of feeder's shell should avoid using plastic or anything that they should not been eating, so we decided to use wood. For the function of regular feeding,

 ISBN: 978-1-61208-419-0

PLC's timer can easily finish the job, and for the quantity control, it is mentioned before. Beside our first prototype machine, we are considering using graphical user interface attachment for the ease of use. Therefore, after finishing our first prototype machine, we will take user survey into account and add new attachments in the future work.

B. *Combining Network*

For a typical industrial control system, the network fieldbus should be able to transmit real-time control messages and non-real-time maintenance messages. Nowadays, with the advantages of IEEE 802.15.4 standard's diversified network architectures, high reliability [5] and the advancement of wireless technology, a wireless network can be integrated into an existing fieldbus system. Due to this development, there are improvements in terms of mobility and flexibility and consequently in the ability to support applications [6] [7].

Moreover, considering the requirement of remote control, besides using normal RJ45 to connect self-home network station, we use Wi-Fi attachment (Figure 4) to control and monitor PLC from PC without virtual line. Other like WSN (Wireless Sensor Network) [4], can also achieve the same goal.

Figure 4. PLC Wi-Fi attachments (upper one for PLC, other one for PC)

C. *System Working Process*

The working process of our device is shown in Figure 5. After complete the start-up diagnose, users can connect to the system through virtual IP address, which is provided by the Wi-Fi attachment, to monitor and control. When the commands are sent, user-side will be waiting for the system to reply. Meanwhile, if the connection is good, the system will start process the command and send back a message whether it is success or not. After comparing the Checksum, the message will be displayed on user's device. Otherwise, the system will report error or timeout.

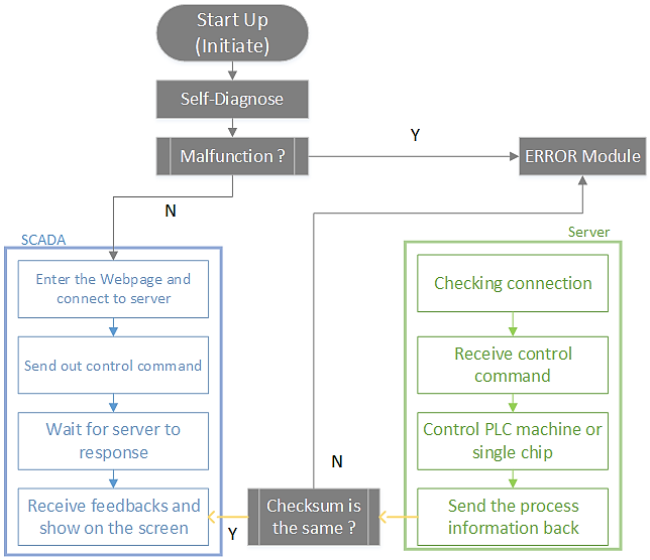


Figure 5. PLC's working process

D. *Webpage Graphical Human Machine Interface*

Graphical Human Machine Interface focus on using words, numbers, and pictures to transmit or reveal the relevant joint state and the value in data register. To control the device from far side, the simplest way is to write a server end program, then through a third-party software to achieve remote controlling, but people will need to pay for the software. With the advance of mobile devices and the wireless data transmission, a simple webpage or software can achieve the same goal, also, the computers have various types of browsers, it is much far more convenient than using other commercial software. To browse from web, programmers must build a server station which can connects to the Wi-Fi attachment first, it can control and process all PLC required function in the server, then sent the data back and shown in the webpage.

IV. CONCLUSIONS & FUTURE WORKS

In this paper, we have finished the build of our first prototype, the basic mechanism device. And our future work is to achieve the goals we mentioned earlier. We plan to find volunteered users to test the system's practicability and ease of use in our future work for system improvement. The final goal is to combine it with normal feeder and grass feeder. Once the system has been developed, it will be possible to lower the rate of pet abandoning and increase pet owners' willingness to actively take care of their new friends by knowing that : "Pets are our friends, not burdens".

After it's finished, we hope our research provides a good inspiration and make our idea fit for any other pets. If this idea can spread, by lowering the manufacture cost, every pet owners can enjoy the convenience of our system.

 ISBN: 978-1-61208-419-0

REFERENCES

[1] "Taipei Rabbit Society Association", Available from: http://www.loverabbit.org/candy/index.asp, 2015/03/06.

[2] "Worldwide Rabbit Love and Appreciation Day", Available from: http://www.vgr1.com/wrlad/, 2015/02/13.

[3] Chun-Chiang, M., "Implementation of PLC Graphical User Interface its Applications", ChienkuoTechnology University. Master of Technology: 85, 2014.

[4] Chung-Ming, O., T. Cheng-Ya, Z. Jing-Ran, Y. Wen-Yuan and T. Shang-Chun, "Intelligent pet monitor system with the internet of things", Machine Learning and Cybernetics (ICMLC), 2011 International Conference on, 10-13 July 2011.

[5] Krasinski, P., B. Pekoslawski and A. Napieralski, "IEEE 802.15.4 wireless network application in real-time automation systems", Mixed Design of Integrated Circuits and Systems (MIXDES), 2013 Proceedings of the 20th International Conference, 20-22 June 2013.

[6] Willig, A., K. Matheus and A. Wolisz, "Wireless Technology in Industrial Networks", Proceedings of the IEEE, vol. 93 (6): p. 1130-1151, 2005.

[7] Xiaolong, L., S. Munigala and Z. Qing-An, "Design and Implementation of a Wireless Programmable Logic Controller System", Electrical and Control Engineering (ICECE), 2010 International Conference, 25-27 June 2010.

[8] "Wideband Code Division Multiple Access", Available from : https://en.wikipedia.org/wiki/W-CDMA_%28UMTS%29, 2015/06/27.

[9] "Code division multiple access", Available from : https://en.wikipedia.org/wiki/Code_division_multiple_access, 2015/07/09.

[10] "General Packet Radio Service", Available from : https://en.wikipedia.org/wiki/General_Packet_Radio_Service, 2015/03/28.

[11] "Personal Handy-phone System", Available from : https://en.wikipedia.org/wiki/Personal_Handy-phone_System, 2015/01/29.

STX	CMD	DEVICE ADDRESS	BYTES	ETX	SUM CHECK
CHR(2)	0	00A0	02	CHR(3)	66
02H	30H	30H 30H 41H 30H	30H 32H	03H	36H 36H

Figure 6. communication form – read data

STX	Y7~Y0	Y17~Y10	ETX	SUM CHECK
	35	86		D9
02H	33H 35H	38H 34H	03H	44H 39H

Figure 7. communication form – PLC return

Y7	Y6	Y5	Y4	Y3	Y2	Y1	Y0	Y17	Y16	Y15	Y14	Y13	Y12	Y11	Y10
0	0	1	1	0	1	0	1	1	0	0	0	0	1	0	1
3				5				8				6			

Figure 8. communication form – node data

STX	CMD	DEVICE ADDRESS	BYTES	WRITE	ETX	SUM CHECK
CHR(2)	1	00A0	02	3586	CHR(3)	3D
02H	31H	30H 30H 41H 30H	30H 32H	33H 35H 38H 36H	03H	33H 44H

Figure 9. communication form – write data

STX	CMD	DEVICE ADDRESS	ETX	SUM CHECK
CHR(2)	7	0105	CHR(3)	00
02H	37H	30H 31H 30H 35H	03H	33H 44H

Figure 10. communication form - force ON

Dual-OS Infrastructure for Mixed-Criticality Systems on ARMv8 Platforms

Alexander Spyridakis, Petar Lalov, Daniel Raho
Virtual Open Systems
Grenoble - France
Email: {a.spyridakis, p.lalov, s.raho}@virtualopensystems.com

***Abstract*—ARM devices are proliferating the mobile and embedded market segments, and with the introduction of Virtualization Extensions (ARMv7-A), and the latest 64-bit architecture changes (ARMv8), ARM is expected to expand further in the networking and server markets. At the same time, Mixed-Criticality use cases for In-vehicle (IVI) and In-flight (IFI) infotainment are of increased interest, where a feature rich Operating System (OS) is required for multimedia applications, while at the same time legacy real time operating systems are still needed for time critical applications. In this paper, we propose and test a Dual-OS environment for Mixed-Criticality systems using ARM devices, by exploiting latest architecture changes and software advancements. The technologies tested and covered in this paper, which enable a Dual-OS environment, include the TrustZone Security Extensions, ARMv7/v8 Virtualization Extensions, as well as the ARM Trusted Firmware (ATF) software infrastructure. Feasibility tests and latency/performance metrics were acquired on ARMv7/v8 platforms including Versatile Express and the Juno development boards.**

Keywords–Mixed-Criticality; ARM; embedded-virtualization; Dual-OS; real-time; Linux; KVM; GPOS and RTOS

I. INTRODUCTION

Real-time systems have pre-defined timing constraints and are deterministic in nature, thus a Real-Time Operating System (RTOS) has the ability to execute tasks with low latency, which are guaranteed to be completed on a predetermined deadline [1] [2]. On the other hand a General Purpose Operating System (GPOS), for example Linux, is targeting best performance instead of providing latency guarantees.

In the context of automotive, some subsystems are time critical, e.g., Electronic Stability Control (ESC) or Adaptive Cruise Control (ACC), while others are tied to multimedia services which are of low priority, with less to no criticality concerns. In such a Mixed-Criticality use case, there is no definite Operating System that can meet all needed characteristics, e.g., strict determinism, low latency, performance, portability, certifiability, feature richness, ease of development and maintenance.

Additionally, providing more performance by just increasing the clock frequency of the CPU is no longer feasible, instead the trend in current System on Chip (SoC) solutions, is to increase the number of cores and lower power dissipation. A continuous growth of multi-core platforms is being observed over the years, which essentially creates incentives for an efficient overcommitment of available resources and the combination of hardware for multiple purposes, which results in a lower total cost. For this reason, with the abundance of multi-core SoCs, it is no longer cost efficient to have multiple hardware instances with different software, instead the use of different Operating Systems in the same hardware platform is desired.

A. Contributions of this paper

First, we highlight the concept of a Dual-OS environment on modern ARM platforms and how a GPOS, such as Linux, can co-exist with other Operating Systems by using the TrustZone technology along with a novel firmware/monitor layer to handle interrupts, context switches and shared resources. Then, by leveraging Linux and KVM on ARM, we show how a host/guest OS can interface with an isolated RTOS running in the secure world. Additionally, we show how an efficient coordination scheduling mechanism can be implemented, to dynamically change scheduling policies triggered by synchronous and asynchronous events.

Finally, experimental results are reported, where the latency overhead of the ATF firmware layer is measured, as well as the communication overhead between a KVM guest and a TrustZone isolated bare-metal binary. The selected hardware platform for testing and benchmarking is ARM's latest 64-bit development board called Juno.

B. Organization of this paper

In Section II, we describe the overall architecture of a Dual-OS infrastructure on a modern ARM platform, and describe how the TrustZone security extensions can be combined with Virtualization Extensions to run two isolated Operating Systems with the option of also adding further Virtual Machines. Additionally, in Section III a list of previous related work is provided, along with how this paper contributes further to the concept. Then, in Section IV we document the ATF firmware layer that is responsible for handling the secure and non-secure world context switches, as well a basic description of its components. In Section V, we provide details on the ATF modifications needed for experimental measurements, along with actual results from the Juno and Versatile Express development boards. Finally, we conclude the work in this paper and list further possible directions.

II. DUAL-OS USE CASE IN IVI

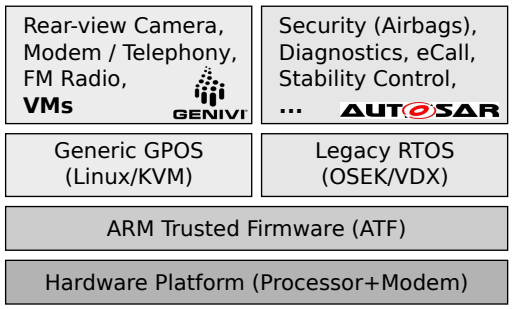
The target use case, similar to [3], is the deployment of two Operating Systems, one RTOS and a GPOS on a single multi-core hardware platform. In an automotive or even aerospace use-case, the medium of travel has an important placement

 ISBN: 978-1-61208-419-0

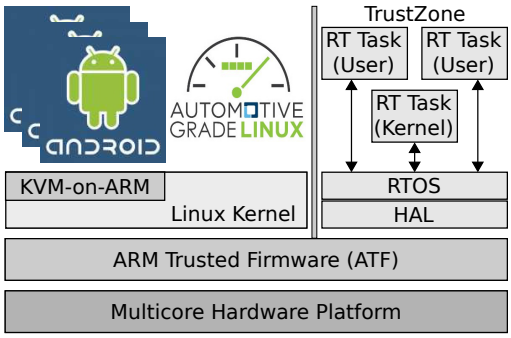
in the Internet of Things (IoT) arena. More specifically in the automotive example, high speed mobile communication enables the car to take the role of a gateway for connected objects.

Moreover, modern technology standards for modems such as 4G/4G+, enable the connection of multiple automotive devices. Applications that interface with the Long Term Evolution (LTE) software stack [5] will need to share resources both in the GPOS and RTOS, which will require a scheme of LTE virtualization either as direct device assignment or hardware assisted virtualization for the target device. Other protocols that are related to this use-case are the Controller Area Network (CAN) bus, as well as the high bandwidth communication IEEE 802.1 AVB Ethernet bus [4], which will also have to be accessible by the multiple actors (OSes) in the system.

Finally, similar to [6], this particular use case depends on the functionality of the GPOS to utilize the Virtualization Extensions of latest ARM architectures, for the instantiation of Virtual Machines completely isolated from the RTOS. This is possible with Linux and KVM on ARM, which allows the kernel to also act as a full-fledged hypervisor. Figure 1 shows the generic architecture of this use case.



(a) Automotive Dual-OS use case



(b) GPOS on KVM & RTOS in TrustZone

Figure 1. GPOS and RTOS on an Automotive Platform

For this mentioned automotive use case, the latest ARMv8-A architecture iteration, is targeting the full spectrum of high-end performance in embedded platforms, but at the same time keeping power needs to a minimum. The most significant change in this new architecture is the transition to 64-bit computing, while preserving 32-bit compatibility for legacy systems. As it was the case with its ARMv7-A predecessors, it provides Virtualization Extensions that enable its efficient usage for virtualization needs, as well as an isolated execution environment called TrustZone for secure computing. In general ARMv8-A brings the most advanced ARM features extended for 64-bit environments, which makes this architecture an ideal selection for virtualization and Dual-OS use cases.

A. Security Extensions

In modern CPU architectures, execution is split in multiple operational modes, with different security aspects and a fine-grained granularity to various instructions. The most obvious use-case for this paradigm is the separation of the kernel-space to user-space execution, where in user-space the permission rights for specific actions are reduced, while in kernel-space most instructions are available and the kernel has almost total control of the hardware.

In x86, this scheme is implemented with protection rings, where 4 different execution mode rings are available and the kernel/user -space code is placed in the most/least privileged mode respectively. For ARM devices these execution modes are called Supervisor and User mode, and newer architectures introduce additional modes such as the Secure Monitor and after-mentioned Hypervisor mode.

The ARM Security Extensions (a.k.a TrustZone) [7] [8], is a system-wide security approach for numerous client to server use-cases, including mobile devices, general purpose computers and enterprise systems. It can be utilized also as means to implement digital rights management, Bring Your Own Device scenarios and secure transactions. TrustZone is a core part of latest Cortex-A processors, although a complete implementation can be extended to the whole platform with specific TrustZone compatible devices/blocks, including secure memory, peripherals, accelerators, etc.

In essence, TrustZone adds a ”Secure” context to the available modes plus the addition of the Secure Monitor which is the most privileged CPU execution level. With this addition two instances of each mode (with the exception of Hypervisor mode which can only be non-Secure) can co-exist together, completely isolated from each other, where they are subject to the central authority of the Secure Monitor Exception Level 3 (EL) [9]. Practically, this means that with TrustZone you can have a Secure Supervisor or User mode (S-EL1/S-EL0) along with the previous non secure instances of these (EL1/EL0). With this set of features TrustZone allows the deployment of General Purpose Operating Systems such as Linux, together with Trusted Execution Environments (TEE). The secure modes have the same features as the normal ones, while operating in an isolated memory space. Finally, the Secure Monitor has utmost authority of all modes handling the world switch (context) between them.

B. Virtualization Extensions

With ARM attempting to break into new markets, but also trying to keep its dominance in existing segments, since ARMv7-A (Cortex-A15, A7, etc.) they have added a number of new features in the ARM architecture in order to facilitate virtualization, usually referred to as the ARM Virtualization Extensions [10].

A new processor mode is introduced, called Hypervisor mode, which allows each guest to have access to its own privileged mode; the processor’s state can be switched between

 ISBN: 978-1-61208-419-0

guests, allowing the processor to be virtualized without expensive binary patching techniques, and with very few traps being necessary. The Virtualization Extensions also allow certain instructions to be set up to trap to the hypervisor if that is necessary to support certain guest features.

The ARM Virtualization Extensions also include functionality to assist with the virtualization of memory for guests. For this, the Large Physical Address Extensions [11], besides an updated page table format also include the possibility to set up a second stage of memory translation to be used by the hypervisor.

The above described extensions are sufficient to fully virtualize the CPU and memory for any number of guests, and also trap any accesses I/O devices so they can be emulated by software. These satisfy the requirements to implement an efficient native virtualization solution on newer ARM processors.

III. Related Work

The concept of Dual-OS in embedded systems has been explored previously, dealing mostly with ARM devices by leveraging TrustZone. Yoshinori Endo et al [18], propose a generic architecture of dual operating systems in automotive called Dependable Autonomous hard Real-time Management (DARMA), with an implementation based on an SH-4 RISC processor.

For ARM platforms, Daniel Sangorrin et al [20], give details on a thin Secure Monitor layer called SafeG and provide examples of scheduling and device sharing between the two Operating Systems on an ARMv6 platform. Finally, Soo-Cheol Oh et al [19], implement their own solution called ViMoExpress based on a Cortex-A8 processor and an LCD virtualization feature.

For this paper, the added contribution is the application of the concept on latest ARMv8 platforms, with the possibility to run 64-bit or legacy 32-bit software, as well as the combination of multiple guest operating systems by utilizing KVM and the virtualization extensions of the ARM architecture.

IV. ARM Trusted Firmware

For a Dual-OS environment we need a firmware layer that will make use of the TrustZone security extensions to isolate resources to their secure and non-secure equivalents. For our firmware needs, which also fit the Juno hardware platform, ARM Trusted Firmware is selected.

ATF [12] is a secure world software implementation for ARMv8-A platforms, provided as a reference firmware infrastructure. Additionally, ATF is meant to be a modular framework for handling the boot procedure, interrupt management and world switching on all available SoC cores. Modularity is key for portability and maintainability, as well as covering any separation concerns in the firmware level, resulting in easier development/testing and certification. At its current state ATF is a standardized EL3 Runtime Firmware for all 64-bit ARMv8-A platforms, with plans to extend it further to cover older ARMv7-A architectures. Although ATF is designed as a firmware solution to run multiple OSes in parallel, it can still be used in cases where only one operating system is used, without a strict requirement of a Secure-OS. The top-level architecture of ATF can be summarized in Figure 2, the functionality overview is also listed below:

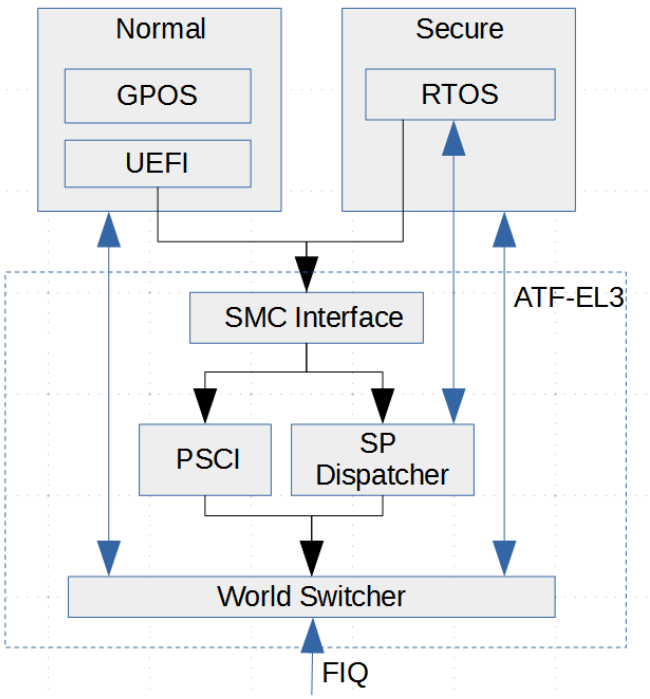


Figure 2. ATF architecture

- Secure world Initialization (e.g., exception vectors, interrupt handling, registers, etc.).
- Support for the newer Generic Interrupt Controllers found in latest ARM devices, which are also virtualization aware and compatible with TrustZone.
- Proper initialization of the Normal world, typically in AArch64 EL2 mode, which is also required for KVM initialization by the kernel.
- Handles Secure Monitor Call (SMC) requests from booted Operating Systems for PSCI power management features such as, booting secondary cores, hotplug and shutdown/reset events.
- Secure-EL1 Payload Dispatcher for handling world switching and interrupt routing.
- Option to replace the Trusted Boot Firmware adapted for the needs of the target platform.
- Memory isolation from secure/normal world based on the features provided by TrustZone.

Figure 3 depicts an overview of the boot procedure in ATF, with the execution sequence between the blocks/modules that ATF consists of. Every stage of the ATF Boot Loader has a dedicated purpose during initialization system boot and any of the following listed BLs can be replaced by custom implementations according to the target platform needs and requirements.

 ISBN: 978-1-61208-419-0

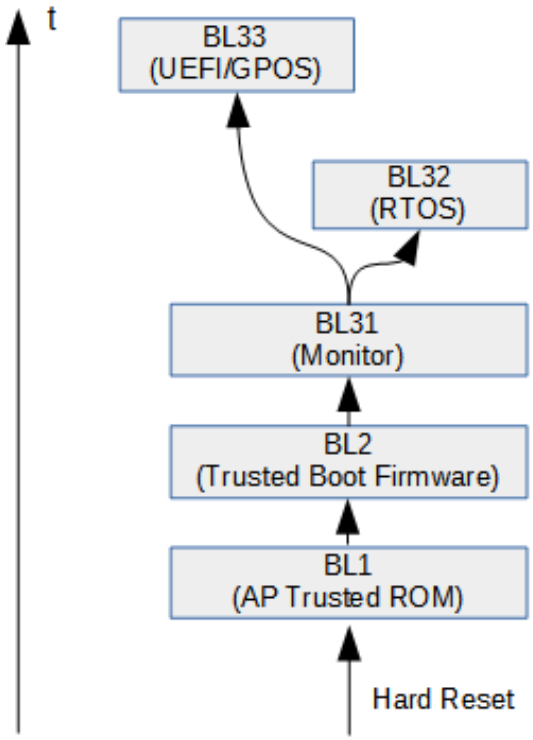


Figure 3. Top-level ATF execution flow

- BL1 - AP Trusted ROM - basic initial boot up procedure in EL3, which subsequently calls BL2.
- BL2 - Trusted Boot Firmware - responsible to pass execution to further BL3-x modules.
- BL3-1 - EL3 Runtime Firmware - Exception routing to dedicated handlers in the corresponding world (world-switch).
- BL3-2 - Secure-EL1 Payload (optional) - runs in SEL-1 and could be an RTOS or another secure bare-metal binary
- BL3-3 - Non-trusted Firmware - any non-secure firmware/software, e.g., u-boot, UEFI (Tianocore EDK2), GPOS, etc.

A. Interrupt management & world switching

In general, exception routing and world switching is implemented at the BL3-1 level, and since an actual Interrupt Service Routine is not part of the EL3 core logic, instead, for the final handling of interrupts, ATF defines interfaces for the retransmission of exceptions to the end destination OS. More specifically, FIQ routing, is managed by the Secure-EL1 Dispatcher layer, while SMC exceptions are standardized in the EL3 runtime service framework based on the SMC Calling Convention PDD. Finally, the EL3 runtime service interface, is complemented by the Power State Coordination Interface (PSCI).

Secure-EL1 Dispatcher service, as part of the EL3 runtime service, is a link interface between BL3-2 (Secure OS/Payload) and BL3-1 (ATF). It is responsible for processing the entry/exit requests for the target secure software (e.g., RTOS), and is designed in a way to provide a common calling convention between these two layers. This service is implemented according to the needs/requirements of the deployed secure software.

PSCI [13] is responsible for the power management of all available SoC cores, where it also supersedes the old mechanism of waking up secondary cores, also known as the CPU holding pen. With PSCI an Operating System can signal the firmware layer to power up/down available cores, through the use of SMC/Hypervisor Call (HVC) instructions. This new paradigm considerably simplifies power management for an OS, as instead of having to implement low level target specific power routines, the OS can rely on the firmware layer.

B. ATF modifications for Dual-OS

By default ATF is using time-triggered signals to change between the secure/non-secure payloads. For this purpose, the internal architected ARM timer [14] is programmed to fire as a secure interrupt and signal a world switch, this interrupt is configured as a Fast Interrupt Request (FIQ) and is handled according to the current context. If at the time of the received FIQ the non-secure world (GPOS) is active, execution will be immediately directed to EL3 for a world switch to the secure payload (RTOS). On the other hand if the secure payload has the context, the interrupt is by default serviced by itself, without the need of EL3 interception.

For experimental measurements, we opted for an event-triggered implementation, where the world switch is triggered by the SMC instruction. That way each world can yield CPU resources deterministically, and give back the context when this is desired. Modifications on the Secure-EL1 Dispatcher service were needed for this behavioral change, which allows us to have a more fine-grained control for latency measurements.

V. Experimental results

Experimental results are split in two sections. First we deploy ATF with two separate world switch triggering methods, and we measure the overall latency introduced when switching between the Secure and non-Secure worlds. Subsequently, QEMU is used to create a KVM accelerated virtual machine and measure the guest exit overhead, that results when an SMC or HVC instruction is called from the guest.

By using the Performance Monitor Unit (PMU), which is found on all recent ARM CPUs, we can precisely measure programmatically the latency between two points in execution. The PMU [15] provides a number of counters and registers for gathering statistics on the operations executed in the processor. Among a set of configurable event and performance counters the PMU provides a 64-bit cycle counter, which is incremented on every clock cycle. For the following measurements the PMU cycle counter (PMCNT) is reset to zero before starting the process of a world/context switch and is stopped after execution is resumed in the respective world.

A. ATF world switch latency

As described in Section IV-B, ATF by default is using time-triggered signals to issue a world switch. For a more deterministic approach ATF was modified so that a world

 ISBN: 978-1-61208-419-0

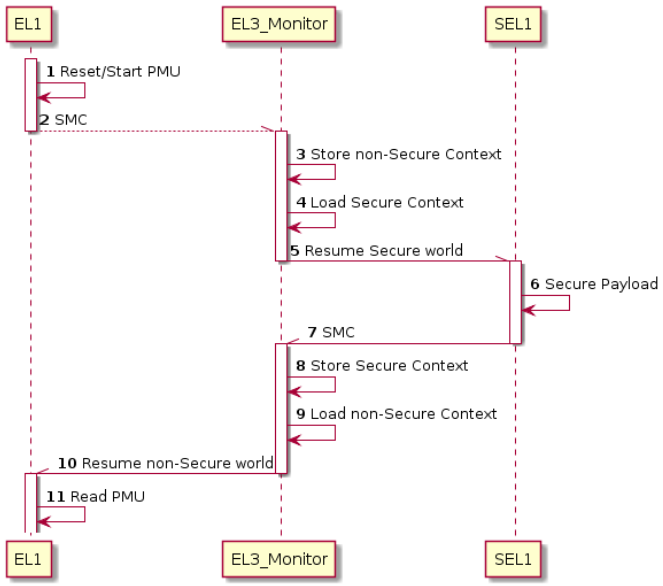


Figure 4. Execution flow of world switch measurement

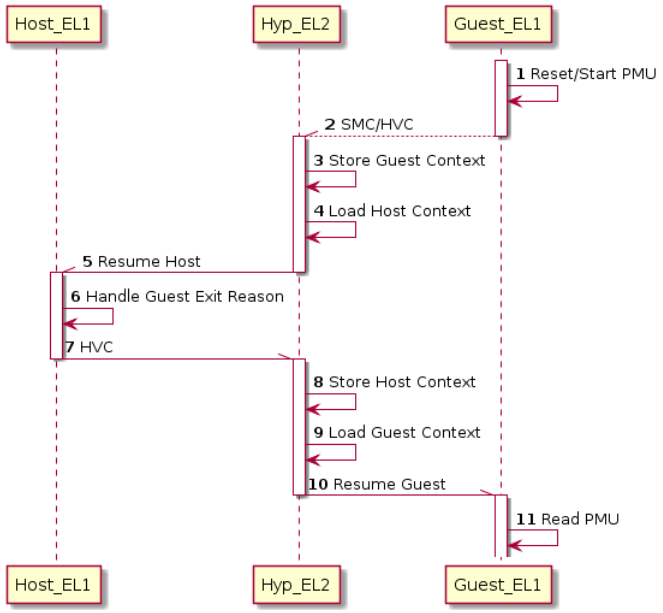


Figure 5. Execution flow of guest/host context switch measurement

TABLE I. LATENCY FOR A FULL WORLD-SWITCH IN CLOCK CYCLES

Average	Minimum	Maximum	Std. deviation
3716,66	3520	3998	95,245

switch can occur programmatically when issuing the SMC instruction.

In the case of Linux, we implemented a small module which at the time of its loading it would reset the PMU cycle counter and immediately issue an SMC instruction. This Secure Monitor Call is then trapped by ATF, which saves the non-Secure context (Linux) and then proceeds to restore and resume the Secure payload. On the Secure side, the SMC is re-issued immediately to trap back to ATF and finally resume execution in Linux.

Figure 4 depicts how latency was acquired in relation to the execution flow. Measurements were repeated 1000 times and results are reported in Table I. The reported results show that a full world switch (non-Secure to Secure and back) is in the range of just under 4000 cycles. This eventually means that the Secure payload, at any point of execution, will receive control of the system in less than 2000 cycles, with actual values in time depending on the clock frequency of the processor.

B. Virtual machine context switch latency

For guest to host latency measurements a similar approach to the world switch measurement is made. The difference is that for the guest we use a thin bare-metal program instead of a full Linux guest. The guest executes a loop where the PMU cycle counter is reset and immediately an SMC or HVC instruction is issued. After guest execution is resumed the PMU cycle counter register is saved to memory and the process is repeated for 1000 times. Before the guest is terminated results are reported to the user.

From the host side, KVM will not allow the guest to execute an SMC or HVC instruction, as they are considered ”sensitive” instructions and are immediately trapped by the hypervisor. An example of this interaction is how KVM wakes up secondary guest cores through the use of PSCI, in which the guest will call HVC with the proper argument. KVM traps the guest, checks the provided argument and decides if the HVC instruction was meant to be a PSCI wake up event. In our case, as PSCI is not used, KVM will inject an abort exception and the guest will be halted. For this reason KVM needs to be modified in order to resume execution to the guest without aborting. A similar usage pattern of HVC instructions, has already been highlighted as means to implement a Storage I/O co-ordination scheduling approach, between a Linux/KVM host and a guest system, with significant improvements in latency and responsiveness of guest applications [16] [17].

Once again Figure 5 highlights the execution flow of the measurement and results are reported in Table II. It is interesting to note that this measurement was replicated in both an ARMv7-A target (Versatile Express TC1 - Cortex-A15) and an ARMv8-A platform (Juno board - Cortex-A53). Furthermore the ARMv8-A Virtualization Extensions with KVM provide a way to run legacy ARMv7 guests, but results were not affected by this scenario. Finally, results show a full host/guest context switch (guest to host and back) of around 1500 cycles for both ARMv7 and ARMv8 platforms. Coupled with the world switch latency results, it means that if a guest needs to be interfaced and communicate with the Secure software, a latency of at least 5500 - 6000 cycles is expected on this firmware implementation.

Summarizing the results, with a reference CPU clock of 1GHz the average latency for a full world switch (non-Secure to Secure and back) is in the range of 4 μs. A similar path for a guest to host context switch (and back to guest) is around 1,5 μs.

 ISBN: 978-1-61208-419-0

TABLE II. LATENCY FOR A FULL GUEST CONTEXT-SWITCH IN CLOCK CYCLES

Type	Average	Minimum	Maximum	Std. deviation
SMC - ARMv8	1497,88	1475	1906	27,766
HVC - ARMv8	1492,62	1461	1932	39,697
SMC - ARMv7	1647,58	1612	2429	66,130
HVC - ARMv7	1679,46	1655	3115	50,140

VI. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we proposed the use of a Dual-OS infrastructure to efficiently leverage ARM multi-core hardware platforms in Mixed-Criticality use cases. Through architecture technologies such as TrustZone and a proper firmware layer like ATF, two completely independent and isolated OS instances can be run simultaneously on the same hardware resources. Additionally, by using Linux as the GPOS and KVM on ARM, we can extend the number of concurrent Operating Systems even further.

Experimentally, with the ARMv8 Juno development platform, we highlighted the latency overhead of a secure/non-secure world context switch and interrupts handled by ATF, as well as the equivalent latency when switching from a guest OS to the host with KVM (including measurements with ARMv7 Versatile Express). Such measurements show the feasibility of the Dual-OS concept, with an average world switch latency of around 4 μs and guest to host switch of 1,5 μs on our reference platforms.

Future work, will include a communication interface between GPOS/RTOS and VMs, in order to implement a complete TEE solution and a fine-grained co-scheduling policy in the system. Finally, the ATF firmware layer is going to be replaced by a fully redesigned custom bare-metal software that will be targeting automotive certification.

ACKNOWLEDGMENTS

This research work has been supported by the Seventh-Framework Programme (FP7/2007-2013) of the European Community under the grant agreement no. 610640 for the DREAMS project.

REFERENCES

[1] J. Altenberg, “Using the Realtime Preemption Patch on ARM CPUs,“ 11th Real-Time Linux Workshop (RTLW 09), Sep 2009, pp. 229-236.

[2] K. Koolwal, “Myths and Realities of Real-Time Linux Software Systems,“ 11th Real-Time Linux Workshop (RTLW 09), Sep 2009, pp 13-18.

[3] M. Hamayun, A. Spyridakis, and D. Raho, “Towards Hard Real-Time Control and Infotainment Applications in Automotive Platforms,“ 10th annual workshop on Operating Systems Platforms for Embedded Real-Time applications (OSPERT 2014), July 2014, pp 39-44.

[4] R. Kreifeldt, “AVnu Alliance White Paper: AVB for Automotive Use,“ [retrieved: June 2015]. [Online]. Available: http://www.avnu.org.

[5] “Long Term Evolution Protocol Overview,“ [retrieved: June 2015]. [Online]. Available: https://rt.wiki.kernel.org/index.php/Cyclictest

[6] H. Mitake, Y. Kinebuchi, A. Courbot, and T. Nakajima, “Coexisting Real- time OS and General Purpose OS on an Embedded Virtualization Layer for a Multicore Processor,“ Proceedings of the 2011 ACM Symposium on Applied Computing (SAC 11), 2011, pp. 629-630.

[7] “ARM TrustZone,“ [retrieved: June 2015]. [Online]. Available: http://www.arm.com/products/processors/technologies/trustzone/index.php

[8] “ARM Security Technology,“ [retrieved: June 2015]. [Online]. Available: http://infocenter.arm.com/help/topic/com.arm.doc.prd29-genc-009492c/PRD29-GENC-009492C_trustzone_security_whitepaper.pdf

[9] “ARM Exception Levels,“ [retrieved: June 2015]. [Online]. Available: http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0438d/CHDCGIBF.html

[10] “ARM Virtualization Extensions,“ [retrieved: June 2015]. [Online]. Available: http://www.arm.com/products/processors/technologies/virtualization-extensions.php

[11] “ARM LPAE Architecture,“ [retrieved: June 2015]. [Online]. Available: http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0438d/CHDCGIBF.html

[12] “ARM Trusted Firmware,“ [retrieved: June 2015]. [Online]. Available: https://github.com/ARM-software/arm-trusted-firmware

[13] “Power State Coordination Interface,“ [retrieved: June 2015]. [Online]. Available: http://infocenter.arm.com/help/topic/com.arm.doc.den0022c/DEN0022C_Power_State_Coordination_Interface.pdf

[14] “Generic Timer architecture,“ [retrieved: June 2015]. [Online]. Available: http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0438d/BABBJGCH.html

[15] “Performance Monitor Unit,“ [retrieved: June 2015]. [Online]. Available: http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0363e/CHDCBAAH.html

[16] A. Spyridakis, and D. Raho, “On Application Responsiveness and Storage Latency in Virtualized Environments,” CLOUD COMPUTING 2014, The Fifth International Conference on Cloud Computing, GRIDs, and Virtualization, 2014, pp. 26-30.

[17] A. Spyridakis, D. Raho, and J. Fanguède, “Virtual-BFQ: A Coordinated Scheduler to Minimize Storage Latency and Improve Application Responsiveness in Virtualized Systems,” International Journal on Advances in Software, vol 7 no 3 & 4, 2014, pp. 642-652.

[18] Endo, Yoshinori, et al, “In-Vehicle Multimedia Platform Based on Darma (Dual Os Approach),” Proc. 7th World Congress on Intelligent Systems, Turin, Italy. 2000.

[19] O. Soo-Cheol, K. Koh, C. Kim, K. Kim, and SeongWoon Kim, “Acceleration of dual OS virtualization in embedded systems,” In Computing and Convergence Technology (ICCCT), 2012 7th International Conference on, pp. 1098-1101. IEEE, 2012.

[20] D. Sangorrin, S. Honda and H. Takada, “Reliable device sharing mechanisms for Dual-OS embedded trusted computing,” TRUST’12 Proceedings of the 5th international conference on Trust and Trustworthy Computing, pp. 74-91.

 ISBN: 978-1-61208-419-0

A Fuzzy-Genetic Algorithm Method for the Breast Cancer Diagnosis Problem

Abir Alharbi and Fairouz Tchier
Mathematics Department, King Saud University, Riyadh, Saudi Arabia
e-mails: {abir, ftchier}@ksu.edu.sa

***Abstract*—The computer-aided medical diagnosis of complex systems, such as breast cancer is an important medical problem. In this paper, we focus on combining two major methodologies, namely, the fuzzy-based systems and the evolutionary genetic algorithms to find a computer aided diagnosis system that will aid physicians in an early diagnosis of breast cancer in Saudi Arabia. Our results show that the fuzzy-genetics approach produces systems that attain high classification performance, with simple and well interpretive rules and a good degree of confidence.**

Keywords-Breast cancer diagnosis problem; Fuzzy systems; Genetic algorithms; Rule-based system; Computer-aided diagnosis.

I. INTRODUCTION

In medical science, diagnosis of a disease is a complicated problem and confirming a diagnosis is difficult even for medical experts. This has given rise to computerized aided diagnostic tools, intended to aid the physician in making primary medical decisions. A major area for such computerized tools is in the domain of breast cancer; to know early on whether the patient under examination exhibits the symptoms of a benign, or a malignant case helps to determine a suitable treatment for the cancer. The automatic diagnosis should attain the highest possible performance, which means they must correctly classify cases with a good degree of confidence. Moreover, it would be desirable for such diagnostic systems to be well interpreted by the physicians.

In this research paper, an automated diagnosis system for breast cancer is designed by combining two methodologies, namely, the fuzzy rule based systems and the genetic algorithms. Medical diagnosis is a decision-making problem that commonly has uncertainty involved; therefore, fuzzy set theory has emerged in this field. The major advantage of fuzzy systems is the simple interpretation; however, finding good fuzzy systems is a hard task. This is where the role of genetic algorithms comes up in tuning the parameters of the fuzzy systems, based on a database of training cases. There are several different examples of the application of fuzzy systems and evolutionary algorithms in the medical domain, such as applying them to the Wisconsin Breast Cancer Diagnosis Data (WBCD) in USA [4], or applying them on pathogenesis of acute sore throat conditions in humans [5], or combining with wavelets, as in [20]. In our paper, we describe the fuzzy-genetic approach, which we developed for the Saudi breast cancer data consisting of 260 patients.

In Sections II and III, we provide a brief overview of fuzzy systems and genetic algorithms respectively. Then, in Section IV, we describe the fuzzy-genetic approach, which is developed in this work for the Saudi breast cancer data discussed in Section V. In Section VI, we discuss the parameters settings in our approach and show the results of our best system evolved, and finally, we present our concluding remarks and future work in Section VII.

II. FUZZY SYSTEMS

Fuzzy logic is a computational method manipulating information in a way that resembles human logical reasoning processes [23][24]. A fuzzy variable is characterized by its fuzzy variable A and the membership functions of these variables; with a membership value $\mu_A(u)$, to a given real value u(R). A fuzzy inference system is a rule-based system that uses fuzzy logic, rather than Boolean logic [26][27]. The structure includes four main components: a fuzzifier, which translates crisp (real-valued) inputs into fuzzy values, an inference engine which applies a fuzzy reasoning mechanism to obtain a fuzzy output, a defuzzifier, that translates the output back into a crisp value, and a knowledge base, containing both an ensemble of fuzzy rules (the rule base), and a group of connection membership functions (the database); see Figure 1.

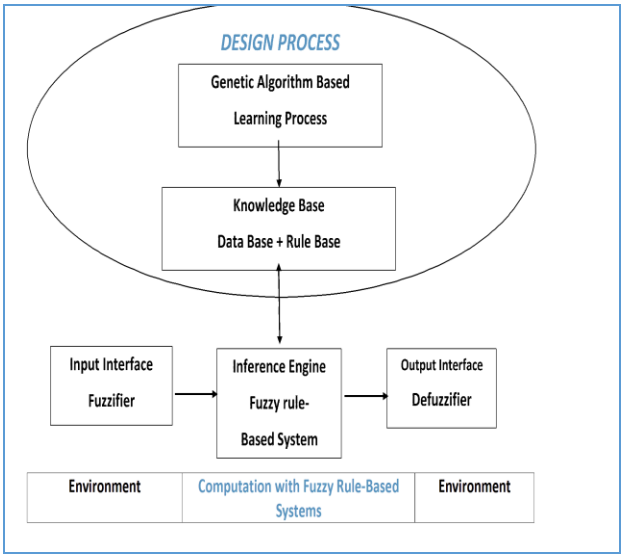


Figure 1. Basic structure of a fuzzy-Genetic system

 ISBN: 978-1-61208-419-0

Moreover, the decision-making process is performed in the inference engine using the rules contained in the rule base. A fuzzy rule has the form "IF antecedent THEN consequent", where the antecedent is a fuzzy-logic expression composed of one or more simple fuzzy expressions connected by fuzzy operators, and the consequent is an expression that assigns fuzzy values to the output variables. The inference engine performs the learning phase, where it evaluates all the rules in the rule base and combines the weighted consequents of all relevant rules into a single fuzzy set using the aggregation operation [16][28]. An example of a fuzzy rule in our case would be: if (v_1 is Low) and (v_2 is Low) then (output is benign), where v_1 and v_2 are variables given in the data set.

Using the direct fuzzy model with knowledge from a human expert, the fuzzy simulation identifies the parameters of a fuzzy inference system, so that a desired decision can be made. This task is difficult when the problem space is complex and very large; thus, motivating us to use genetic algorithms to produce fuzzy models. In the literature, there are several approaches to fuzzy modelling based on neural networks [10][12], genetic algorithms [1][6][8], and other hybrid methods [25]. Selection of relevant variables and adequate rules is critical for obtaining a good accurate classification system. One of the major problems in fuzzy simulation is that the amount of computation grows exponentially with the number of variables.

III. Genetic Algorithm

A Genetic Algorithm (GA) is a search heuristic that mimics the process of natural selection. Genetic algorithms are used to generate solutions to optimization and search problems. They belong to the larger class of evolutionary algorithms, used to generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection based on a relative fitness, and crossover [13]. Genetic algorithms are usually applied to spaces which are too large to be exhaustively searched and have applications in bioinformatics, industry, medical, science, engineering, chemistry, computational mathematics [3], and many other fields.

The genetic algorithm method is an iterative procedure that involves a population representing the search space for solutions to the problem, as individuals, each one represented by a finite string of symbols, called the genome. The basic procedure proceeds as follows: an initial population of individuals is generated at random or heuristically. In every evolutionary step (generation), the individuals in the current population are decoded and evaluated according to a fitness function that describes the optimization problem in the search space. To form a new population (the next generation), individuals are selected according to their fitness, a fitness function is a particular type of objective function that is used to measure how close the given individual is to achieving the set aims of the problem [18]. Many selection procedures are available, one of the simplest being fitness-proportionate selection, where individuals are selected with a probability proportional to their relative fitness. This ensures that the expected number of times an individual is chosen is approximately proportional to its relative performance in the population. Thus, high-fitness individuals stand a better chance to reproduce and bring new individuals to the population, while low-fitness will not. Genetic algorithms are stochastic iterative processes, which are not necessarily guaranteed to converge, and the stopping condition may be specified as a maximal number of generations or a chosen level of the fitness.

IV. Fuzzy-Genetic Algorithms

Since GAs are used to search large complex search spaces and are able to give optimal and near-optimal solutions on numerous problems; therefore, fuzzy–genetic algorithms can be considered as an optimization process where the parameters of a fuzzy system constitute the search space. Many researchers investigated the application of evolutionary techniques in the domain of fuzzy modelling [1][5][4], where the tuning of fuzzy inference systems involved in control tasks were done by genetic algorithms. Fuzzy–genetic modelling has been applied to many domains [6][8][11][19], branching into many areas as electric engineering, chemistry, telecommunications, biology, geophysics and medicine. The GA can be used to tune the knowledge contained in the fuzzy system by finding membership function values. An initial fuzzy system is defined by an expert; then, the membership function values are encoded in a genome, and a genetic algorithm is used to find systems with high performance. GAs often overcomes the local-minima problem seen in other gradient descent-based optimization methods [18]. GAs can be applied in different stages of the fuzzy system parameters search depending on several conditions, like the availability of a priori knowledge, the size of the parameter, and the availability and completeness of input/output data. The fuzzy parameters used to define targets for genetic fuzzy modelling are: structural parameters, connective parameters, and operational parameters.

In many cases, the available information about the system is composed almost exclusively of input/output data, and specific knowledge make up the system structure. In such a system, evolution has to deal with the simultaneous design of rules, membership functions, and structural parameters. Structure learning permits to specify other criteria related to the interpretability of the system, such as the number of membership functions and the number of rules, while, the strong interdependency among the parameters involved in this form of learning may slow down the convergence of the genetic algorithm. Both connective and structural parameters simulation [1][11] are viewed as rule base learning processes with different levels of complexity. In most GA applications, the main approaches for evolving such rule systems are the Michigan approach [1], the Pittsburgh approach [13] and the iterative rule learning approach [11].

 ISBN: 978-1-61208-419-0

V. BREAST CANCER DATA BASE

Breast cancer is known as one of the most common cancers type affecting the female population. It is one of the major causes of death among women and a true emergency for health care systems of industrialized countries. One of the epidemiological studies conducted by AlDiab et al. [2] reported that the incidence of breast cancer in Saudi Arabia was 19.8% of all the female cancers detected in Saudi Arabia. Researchers in the field [7][21] have shown that breast cancer is the second most common malignancy for women in Saudi Arabia. Nevertheless, there is a paucity of detailed published epidemiologic data. An earlier report according to Saudi National Cancer Registry, mentioned an increasing proportion of breast cancer among women of different ages from 10.2% in 2000 to 24.3% in 2005 [7]. The presence of a breast mass is an alert, but it does not always indicate a malignant cancer. Fine needle aspiration $(FNA)^2$ of breast masses is a cost-effective, non-traumatic, and mostly non-invasive diagnostic test that obtains information needed to evaluate malignancy. The medical diagnosis data of breast cancer used in this study is from patients in Saudi Arabia. The database is similar to the WBCD dataset of the University of Wisconsin Hospital [17], where diagnosis of breast masses is based solely on an FNA test [15]. Nine visually assessed characteristics of an FNA sample considered relevant for diagnosis are identified, and were assigned an integer value between 1 and 10. The diagnostics in the database were done by specialists in the field, and the database itself consists of 260 cases, with each entry representing the classification for a certain ensemble of measured values, (Case number, [v_1, v_2 , v_3 , … , v_9 , Diagnostic: Benign or Malignant). The measured variables are as follows: v_1 is clump thickness, v_2 is uniformity of cell size, v_3 is uniformity of cell shape, v_4 is marginal adhesion, v_5 is single epithelial cell size, v_6 is Bare nuclei, v_7 is bland chromatin, v_8 is normal nucleoli and v_9 is mitosis.

Basically, an initial fuzzy rule base is defined by an expert, for example a fuzzy rule in this case would be: if [v_1 *is Low*] and [v_7 *is Low*] then (output is benign). Therefore, each of the nine variables (v_1-v_9) has two parameters P and d, defining the start point and the length of the membership function, respectively. Then, the GA fine-tunes the membership functions. Also for the antecedents: the i^{th} rule has the form if (v_1 is M_1^i) and (v_7 is M_7^i) then (output is benign), Where M_j^i represents the membership function applicable to variable v_j , M_j^i can take on the values: 1 (*Low*), 2 (*High*). The GA is also used to find either the rule consequents, or other subset rules to be included in the rule base. As the membership functions are fixed this approach lacks the flexibility to modify substantially the system behaviour. One of the major disadvantages of knowledge tuning is its dependency on the initial setting of the knowledge base. Furthermore, as the number of variables and membership functions increases, large dimensionality decreases the system's performance. Evolutionary structure modelling is done by encoding within the genome an entire fuzzy system using the Pittsburgh approach. The fuzzy system computes a continuous appraisal value of the malignancy of a case, based on the input values. According to the fuzzy system's output the threshold unit then outputs a benign or malignant diagnostic. In order to evolve the fuzzy model, as seen in Figure 2, we must set some preliminary parameters in the fuzzy–genetic system itself encoding.

VI. FUZZY-GENETIC PARAMETERS

All previous knowledge about the problem and about the existent rule-based models gives us valuable information for our choices of fuzzy parameters. Since all the labels have semantic meaning, for each label, at least one element of the space should have a membership value equal to one. Hence, a Low membership value of 0.8 entails a High membership value of 0.2, and for each element the sum of all its membership values should be equal to one. The parameter settings are set as in the following.

A. The Fuzzy-Genetic System Parameter Settings

- Number of input membership functions: is set to two, (Low and High).
- Number of output membership functions: is two singletons for the benign and malignant diagnostic cases.
- Number of rules: is fixed to three.
- Antecedents of rules: is found by the genetic algorithm.
- Consequent of rules: the algorithm finds rules for the benign diagnostic; the malignant diagnostic is an else condition.
- Rule weights: the learning is done by letting active rules have a weigh of value 1, and the else condition has a weight of 0.25.
- Input membership function values: is found by the genetic algorithm.
- Output membership function values: following the database, we used a value of 2 for benign and 4 for malignant.

We applied the Pittsburgh-style-structure learning, using a genetic algorithm to search for three parameters, namely, the genome (encoding relevant variables), input membership function values, and antecedents of rules: Relevant variables are searched for implicitly by letting the algorithm choose non-existent membership functions as valid antecedents; in such a case the respective variable is considered irrelevant. To evolve the fuzzy inference system, we used a genetic algorithm with a fixed population size of 50 individuals. The algorithm terminates when the maximum number of generations is reached at 300, or when the increase in fitness of the best individual over five successive generations falls below a certain threshold, set at 2×10^{-6}. Our fitness function F is set to the classification performance, computed as the percentage of cases correctly classified, given by

$$F = Fr - \alpha\, Fc \quad (1)$$

where $\alpha = 0.1$, *Fr*, the ratio of correctly diagnosed cases, which is the most important measure of performance, and *Fc* measures the confidence, penalizing systems with large number of low appraisal value cases i.e., cases that are diagnosed with low confidence. The crossover between the

two chosen parents genome is done at a single point randomly chosen with probability 0.8 to produce the new generation offspring. The selection operator of parent's genome is set to the stochastic uniform selection method, and the mutation done on the new offspring has probability 0.01. Hence, the experiment starts by finding from a population of 50 genomes of length 45, where the first 18 bits represent the parameters of the membership functions (P_i, d_i) of each v_i and the remaining 27 bits are the output function M_j^i for each v_i in the three rule base system showing Low or High or irrelevant. Table I shows the parameters encoding to form a single individual's genome. The GA runs throughout the generations to find the best genome in this population. The best genome is the one which classifies correctly the largest number of the 260 cases given in the data set. After all 300 generations (repeated 50 times), the genetic algorithm found the optimum genome; hence, it found the best diagnostic system with three rules given in Figure 2.

TABLE I. PARAMETER ENCODING IN A GENOME

Parameter	values	Total number of bits (45)
P	1-8	9
d	1-8	9
M	0-2	27

Database

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
P	2	5	8	4	6	3	4	5	4
d	5	3	1	2	1	6	3	2	1

Rule base

Rule 1 : if (v_3 ***is Low) and (*** v_7 ***is Low) and (***v_8 ***is Low) and (***v_9 ***is Low) then (output is benign)***

Rule 2 : if (v_1 ***is Low) and (***v_2 ***is Low) and (v4 is Low) and (v5 is High) and (v9 is Low) then (output is benign)***

Rule 3 : if (v1 is Low) and (v4 is Low) and (v6 is Low) and (v8 is Low) then (output is benign) else (output is malignant)

Figure 2. The best evolved fuzzy-gnetic diagnostic system with three rules which exhibits an overall classification rate of 97.33%.

B. Results

The solution scheme we present for the Saudi breast cancer diagnosis problem consists of a fuzzy system model and a threshold unit. The fuzzy system computes a continuous appraisal value of the malignancy of a case, based on the input values. The threshold unit then outputs a benign or malignant diagnostic according to the fuzzy system's output. In order to evolve the fuzzy model, we must set the fuzzy system parameters and the genetic algorithm encoding according to the previous discussion in part A. The evolutionary performed experiments fall into a learning category, in accordance with the data partitioning into two distinct sets: training set and testing set, Training set contains 50% of the database cases and the testing set contains the remaining 50% of the cases. Fifty evolutionary runs were performed, all of which found systems whose classification performance exceeds 95%. MATLAB Genetic Toolbox [29] was modified to implement the fuzzy-genetic algorithm and to generate the results. Taking into account the performance classification rate, the best diagnostic system with three rules stated in details in Figure 2 is the top one over all 50 evolutionary runs. It obtained 98.3% correct classification rate over the benign cases, 96.2% correct classification rate over the malignant cases, and an overall classification rate of 97.33%. The performance value denotes the percentage of cases correctly classified. Three such performance values are shown in Table II: the performance over the training set, the performance over the test set, and the overall performance on the entire database. Figure 3 shows a close up of the plot of the best fitness value over the generations, which scored on average 254 cases accurate out of the 260 data cases. Figure 4 shows the best individual (45 parameters) for the evolved fuzzy three rule diagnostic system described in Figure 2. In Figure 5, we can see the average distances between individuals for the evolved fuzzy three-rule system throughout the generations. Figure 6 shows the best, worst, and mean fitness scores reached by the evolved fuzzy three-rule system during the procedure.

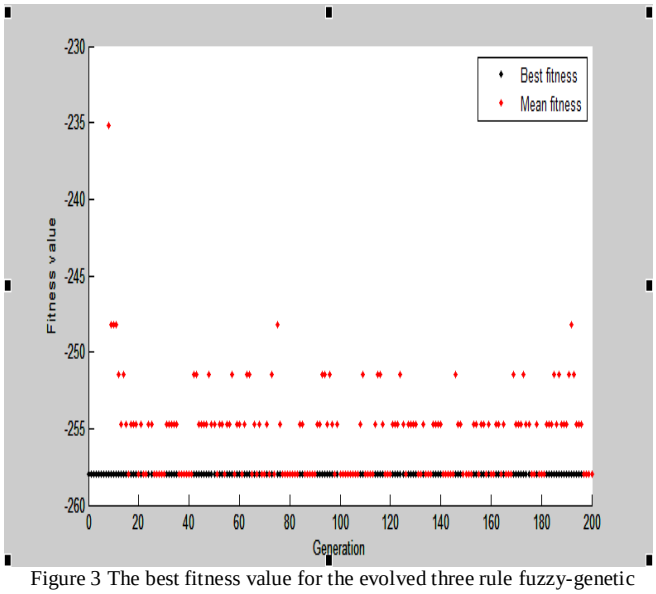


Figure 3 The best fitness value for the evolved three rule fuzzy-genetic system.

The proposed fuzzy system described in this paper performs very well and reached comparable results similar to work done on the WBCD data by Andres et al. [4], and Setiono [22] in terms of both performance and simplicity of rules as seen in Table III. It is worth noting that [4] had 699

cases in the WBCD dataset from patients in USA and they used a different fitness function denoted $F=Fc-0.05Fv-0.01Fe$, such that Fc, the number of correctly diagnosed cases, Fv measures the linguistic integrity (interpretability), and Fe adds selection pressure towards systems with low quadratic error. Moreover, Setiono [22] used an application of neural networks that involves Boolean rule bases extracted from trained neural networks. Table III shows the classification performance values obtained by these different approaches, looking very close in terms of accuracy and in time efficiency.

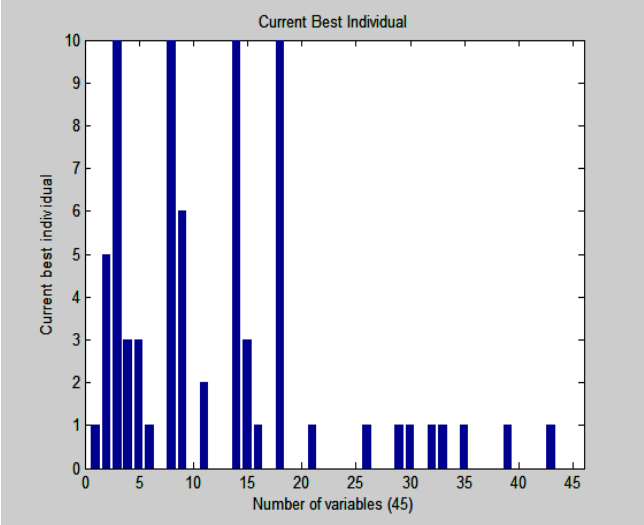


Figure 4. Current best individual in the three rule fuzzy-genetic system.

Following these steps and obtaining the results complete the fuzzification phase; it is time for the inference engine to compute the truth value of each rule, by applying the fuzzy 'and' operator to combine the antecedent clauses in a fuzzy manner. This results in the output truth value, which is a continuous value which represents the rule's degree of activation inference. Thus, a rule is not merely either activated or not, but in fact is activated to a certain degree represented by a value between 0 and 1. The inference engine now goes on to apply the aggregation operator and combining the continuous rule activation values to produce a fuzzy output with a certain truth. Then, the defuzzifier works to produce the final continuous value of the fuzzy system; this latter value is the value that is passed on to the threshold unit. For our best three rule fuzzy system we calculate the membership values for each 260 patients and with the "and" function we get the appraisal value in the range [3,5] . We chose to place the threshold value at 3, with inferior values classified as benign, and superior values classified as malignant. Hence, a value of 2.42 is classified as benign, which is correct; but, it is among the closest to the threshold value, and its confidence is low. Most other cases result in an appraisal value that lies close to one of the extremes (i.e., close to either 2 or 4). Thus, in a sense, we can say that we are somewhat less confident where this case is concerned, with respect to most other entries in the database. Moreover, the appraisal value can accompany the final output of the diagnostic system, serving as a confidence measure. This demonstrates a prime advantage of fuzzy systems, namely, the ability to output not only a (binary) classification, but also a measure representing the system's confidence in its output. For our best three rule system presented here, only 13 cases out of 260 were diagnosed with low confidence.

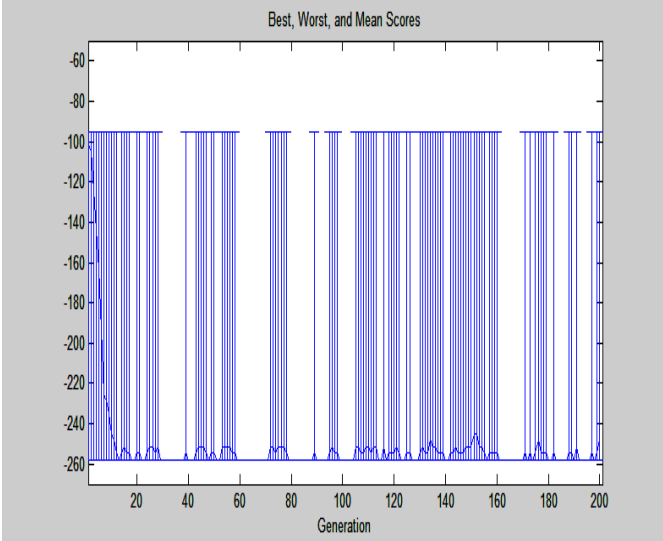


Figure 5. The best, worst and mean fitness scores for the three rule fuzzy-genetic system.

TABE II. RESULTS OF 50 EVOLUTIONARY RUNS, DIVIDED ACCORDING TO THE THREE CATAGORIES

Training/test 50/50	Performance		
	Training set	Test set	Overall
	97.70 %	96.91%	97.33%

TABLE III. COMPARING OUR RESULTS FOR A THREE RULE BASE SYSTEM WITH OTHER APPROACHES

This work	Andres, Pena and Sipper [4]	Setiono [22]
97.33 %	97.80 %	97.14 %

VII. CONCLUSION AND FUTURE WORK

In this paper, we applied a combined fuzzy-genetic approach to the Saudi breast cancer diagnosis database. Our evolved three rules system exhibits both high classification performance and a good confidence measure. Our results suggest that the fuzzy-genetic approach could be highly effective on medical diagnosis problems and may help in designing computer-aided software to obtain an early diagnosis and reduce treatment expenses, which are considered to be among the highest sanitary priorities in many countries. Our future work will involve finding more rule bases and making comparisons. We also plan to apply

 ISBN: 978-1-61208-419-0

the fuzzy-genetic approach to other complex real-world diagnosis problems and extend our work to data from all over the Middle East. We will also try alternative fuzzy logic approaches, such as neuro-fuzzy networks or fuzzy-Petri with the evolutionary genetic algorithm method. In addition, we will explore another promising area combining genetic algorithm with neural networks such as adaptive neuro-fuzzy inference systems.

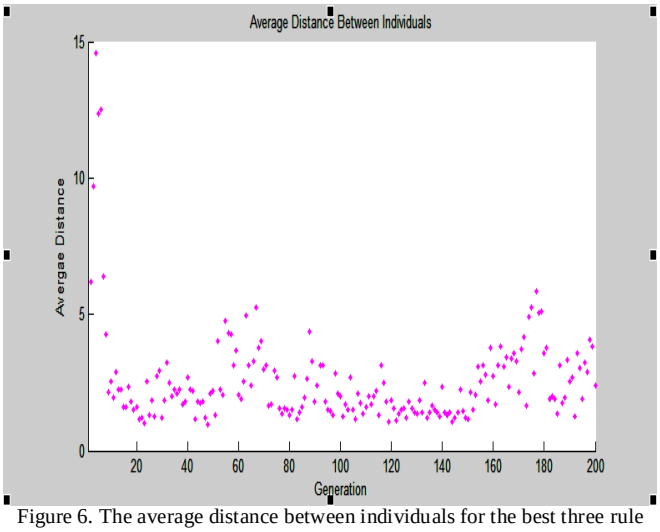


Figure 6. The average distance between individuals for the best three rule fuzzy-geneic system.

ACKNOWLEDGMENT

This research project was supported by a grant from the "Research Centre of the Female Scientific and Medical Colleges", Deanship of Scientific Research, King Saud University.

REFERENCES

[1] J. T. Alander, "An indexed bibliography of genetic algorithms with fuzzy logic", Fuzzy Evolutionary Computation, Dordrecht: Kluwer, 1997, pp. 299–318.

[2] A. R. AlDiab, S. Qureshi, K. A. AlSaleh, F. H. AlQahtani, A. Aleem, V. F. Qureshi, and M. R. Qureshi,. "Studies on the Methods of Diagnosis and Biomarkers Used in the Early Detection of Breast Cancer in the Kingdom of Saudi Arabia", World Journal of Medical Science, 2013, pp. 72-88.

[3] A. Alharbi, W. Rand, and R. Rolio, "Understanding the Semantics of Genetic Algorithms in Dynamic Environments A case Study Using the Shaky Ladder Hyperplane-Defined Functions", Workshop on Evolutionary Algorithms in Stochastic and Dynamic Environments, 2007.

[4] C. Andres, P. Reyes, and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis", Artificial intelligence in Medicine, vol. 17, Elsevier, 1999, pp. 131-155.

[5] C. J. Carmona, V. Ruiz-Rodado, M. J. del Jesus, A. Weber, M. Grootveld, P. González, and D. Elizondo, "A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans", Information Sciences, vol. 298, 2015, pp. 180-197.

[6] O. Cordon, F. Herrera, and M. Lozano. "On the combination of fuzzy logic and evolutionary computation: a short review and bibliography", Fuzzy Evolutionary Computation, Kluwer, 1997, pp. 33–56.

[7] S. M. El-Akkad, M. H. Amer, G. S. Lin, R. S. Sabbah, and J. T. Godwin, "Pattern of cancer in saudi Arabia referred to King Faisal Hospital Cancer", vol. 58, 1986, pp. 1172-1178.

[8] H. Heider and T. Drabe,"Fuzzy system design with a cascaded genetic algorithm. IEEE International Conference on Evolutionary Computation, 1997, pp. 585–588.

[9] F. Herrera, M. Lozano, and J. L. Verdegay, "Generating fuzzy rules from examples using genetic algorithms", Fuzzy Logic and Soft Computing, World Scientific, 1995, pp. 11–20.

[10] J. R. Jang and C. T. Sun, "Neuro-fuzzy modeling and control", Proceedings of the IEEE, vol. 83 (3), 1995, pp. 378-406.

[11] C. L. Karr, "Genetic algorithms for fuzzy controllers", A I Expert, vol. 6(2), 1991, pp. 26–33.

[12] B. Kovalerchuk, E. Triantaphyllou, J. F. Ruiz, and J. Clayton, "Fuzzy logic in computer-aided breast cancer diagnosis" Artif. Intell. Med., vol. 11 (1), 1997, pp. 75–85.

[13] J. R. Koza, Genetic Programming, MIT Press, 1992.

[14] M. A. Lee and H. Takagi, "Integrating design stages of fuzzy systems using genetic algorithms", IEEE International Conference on Fuzzy Systems, 1993, pp. 612–617.

[15] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming", Mathematical Programing Technical Report, 1994, pp.94-101.

[16] J. M. Mendel, "Fuzzy logic systems for engineering: a tutorial", Proceeding of the IEEE, vol. 83, 1995, pp. 345-377.

[17] C. J. Merz and P. M. Murphy, "UCI repository of machine learning-databases", [retrieved; Oct. 2014] http : // www. ics.uci.edu / ~mlearn /MLR epository, 1996.

[18] Z. Michalewicz, Genetic Algorithms+Data Structures= Evolution Programs, 3rd edition, Springer-Verlag, 1996.

[19] S. Muthukrishnan, "GFS: Adaptive Genetic Fuzzy System for medical data classification B Dennis", Applied Soft Computing, 2014, Elsevier.

[20] T. Nguyen, A. Khosravi, D. Creighton, "Classification of healthcare data using genetic fuzzy logic system and wavelets", Expert Systems with Applications, vol. 42, Issue 4, 2015, pp. 2184-2197.

[21] K. Ravichandran, N. A. Hamdan, and A. R. Dyab, "Population based survival of female breast cancer cases in Riyadh Region, Saudi Arabia", Asian Pacific Journal of Cancer Prevention, vol. 6, 2005, pp. 72-76.

[22] R. Setiono, "Extracting rules from pruned neural networks for breast cancer dianosis", Artificial Intelligence in Medicine, 1996, pp. 37-51.

[23] F. Tchier, "Relational Demonic Fuzzy Refinement", Journal of Applied Mathematics, 2014, pp. 17-21.

[24] F. Tchier, "Fuzzy Demonic refinement", International Conference on Basic and Applied Sciences Regional Annual Fundamental Science Symposium 2013, Johor, Malaysia.

[25] P. Vuorimaa, "Fuzzy self-organizing map", Fuzzy Sets Syst, vol. 66, 1994, pp. 223-231.

[26] R. R. Yager and D. P. Filev, Essentials of Fuzzy Modeling and Control, Wiley, 1994.

[27] R. R. Yager, L. A. Zadeh, Fuzzy Sets, Neural Networks, and Soft Computing, New York, Van Nostrand Reinhold, 1994.

[28] L. A. Zadeh, "Fuzzy sets", Information and Control, vol. 8, no. 3, 1965, pp. 338-353.

[29] Matlab optimization Tool box user guide [retrieved: Jan 2015] from http://www.mathworks.com/help/optim/index.html.

 ISBN: 978-1-61208-419-0

Numerical Simulation of Ocean Ice Dynamics using Hybrid FE/FV Methods

Sridhar Palle, Shahrouz K. Aliabadi
Northrop Grumman Center for High Performance Computing,
Jackson State University, USA
Email: sridhar.palle@jsums.edu, shahrouz.k.aliabadi@jsums.edu

***Abstract*— Hybrid Finite Element/Finite Volume (FE/FV) methods have been employed to study oceanic circulation in a simplified domain consisting of both non-moving and moving ice floes. Our hybrid FE/FV flow solver combines the merits of both finite element and finite volume methods, and is highly sophisticated, robust and is a first of its kind approach extended for studying ocean ice dynamics and dispersion. Sea ice dynamics is one of the key components of ocean circulation models. The ultimate goal through the present project is to develop a highly accurate ice dynamics model that can be used to predict ice-edge positions, velocities for offshore operations, short term forecast for waterways, and also for long term global climatic studies. Preliminary results show that hybrid FE/FV methods can be successfully extended for studying ocean ice dynamics, through coupled implementation of automatic mesh movement. Movement of an isolated ice floe through the ocean and also waves impacting multiple dynamic ice floes are successfully simulated while maintaining the mesh integrity.**

Keywords-Shallow water; Hybrid FE/FV; Ice dynamics; Mesh movement.

I. INTRODUCTION

Sea ice plays a crucial role in the Arctic region impacting navigational shipping routes, military, costal guard applications, weather forecasting, and also offshore drilling platforms. With release of greenhouse gases into atmosphere and associated global warming, the Arctic region has become that much more important and accurate studies of Arctic Ocean ice dynamics have become highly imperative. Arctic region can be classified arbitrarily as central Arctic where sea ice is continuous and also into a marginal ice zone (with individual ice floes) which is an interfacial region between Open ocean and Frozen Central Ocean. Marginal Ice Zone (MIZ) is of particular interest due to its proximity to shipping routes and also as a threat to offshore drilling structures. However, modelling sea ice dynamics in MIZ, where individual ice floes are of arbitrary shapes and much more mobile and fluid, is a highly complex and challenging task.

MIZ region has received significant attention over the past few decades and literature in this area is thoroughly discussed in recent reviews by Squire et al. [1][2]. Within the MIZ region, researchers focussed on either continuum ice models, where MIZ is assumed to have certain rheological properties a priori (like a granular material), or on accurately modelling individual ice floes [2]. Within the second group, most of the studies are still limited to theoretical works, numerical models, and recently to spectral methods and Laplace transforms [3][4]. Direct numerical simulation studies in the MIZ region are relatively scarce due to the huge challenges involved in simulating individual ice floes requiring mesh movement, simulating complex wave-floe, floe-floe interactions and also due to the computing power required in realistic simulations of large regions of MIZ. The present work is the first of its kind to the author's knowledge in simulating sea ice dynamics using hybrid finite element/volume (FE/FV) methods. Due to the highly challenging nature of the problem, the present study is being conducted in a systematic way by employing the hybrid FE/FV methodology to ocean ice dynamics with varying degrees of complexity. As a first step, in this work, circular waves impacting both non-moving and moving ice-floes in simplistic oceanic conditions are simulated. Presently, only translation motion is implemented. Section II details the governing equations. Numerical methods that are used to solve the governing equations are discussed in Section III. Results are presented in Section IV and Section V highlights the conclusions and future work.

II. GOVERNING EQUATIONS

A. Shallow Water Equations

The Shallow Water Equations (SWEs) [5] are derived by depth-averaging the Reynolds averaged Navier-Stokes equations for a column of fluid with mass and momentum conservation. In SWEs, it is assumed that vertical motions are negligible and that pressure is hydrostatic. The depth and velocity of fluid moving in the domain $\mathbf{x}(x,y)\in\Omega$ with boundary $\partial\Omega=\partial\Omega_g+\partial\Omega_h$ during the time interval $t\in(0,\mathrm{T})$ in non-conservation form can be described by,

$$\frac{\partial H}{\partial t}+\nabla\cdot\ H\mathbf{u}\ =\dot{n} \tag{1}$$

$$\frac{\partial \mathbf{u}}{\partial t}+\ \mathbf{u}\cdot\nabla\ \mathbf{u}=-g\nabla H-\nabla\left(\frac{p_a}{\rho_0}+gZ\right)+\frac{\nu}{H}\nabla\cdot\nabla\ \mathbf{u}H \tag{2}$$

 ISBN: 978-1-61208-419-0

where $H, \dot{n}, \mathbf{u}, g, p_a, \rho_0, \nu$ and Z are the water depth, net source term, velocity, gravity, surface pressure, fluid density, kinematic viscosity and surface elevation, respectively. Figure 1 demonstrates the terminology to describe *H*, *h*, and Z.

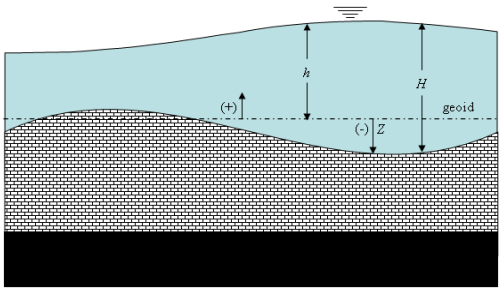


Figure 1. Shallow water problem description.

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\,\mathbf{u} = -g\nabla H - \nabla\left(\frac{p_a}{\rho_0} + gZ\right) + \nu\,\nabla\cdot\nabla\,\mathbf{u} + \frac{\nu}{H}\mathbf{u}\,\nabla\cdot\nabla\,H \tag{3}$$

Here, we expanded the viscous term from conservation form to non-conservation form. In our hybrid method we store the variable H at the node and the velocity $\mathbf{u}$ at the element center. Using linear interpolation function will result in constant gradient and zero Laplacian for working variable H. As a result, we can drop the last term in (3). Also, while our solver includes the capabilities for studying wind stress, tide, and Coriolis forces [5] they have been omitted in the present study and also from above equations.

B. Linear Elasticity Equations for Mesh Moving

Mesh moving equations [6] from linear elasticity are described below:

$$\nabla.\sigma = F \tag{4}$$

$$\sigma = \lambda\,\nabla.X\;I + 2\mu\varepsilon \tag{5}$$

$$\varepsilon = \frac{1}{2}\left[\nabla X + \nabla X^{T}\right] \tag{6}$$

where σ is the stress tensor, ε is the strain tensor, F is the body force per unit volume, λ and μ_l are the lame parameters, and X is the displacement vector.

III. NUMERICAL METHOD

A. Hybrid FE/FV Methodology

The time discretization of Eq. (3) using backward difference will yield

$$\frac{\alpha_1\mathbf{u} + \alpha_0\mathbf{u}^n + \alpha_{-1}\mathbf{u}^{n-1}}{\Delta t} + \mathbf{u}\cdot\nabla\,\mathbf{u} - \nu\,\nabla\cdot\nabla\,\mathbf{u} = -g\nabla H - \nabla\left(\frac{p_a}{\rho_0} + gZ\right) \tag{7}$$

where both $\mathbf{u}$ and H are unknowns at time step $n+1$. For first order time accurate scheme, $\alpha_1 = 1.0$, $\alpha_0 = -1.0$ and $\alpha_{-1} = 0.0$ and for second order time accurate scheme $\alpha_1 = 1.5$, $\alpha_0 = -2.0$ and $\alpha_{-1} = 0.5$. The hybrid FE/FV scheme evolves by perturbing H such that

$$H \to H + H' \tag{8}$$

where H' is very small compared to H. The time discretized momentum equation will lead to (4)

$$\frac{\alpha_1\tilde{\mathbf{u}} - \alpha_1\mathbf{u}}{\Delta t} + \frac{\alpha_1\mathbf{u} + \alpha_0\mathbf{u}^n + \alpha_{-1}\mathbf{u}^{n-1}}{\Delta t} + \mathbf{u}\cdot\nabla\,\mathbf{u} - \nu\,\nabla\cdot\nabla\,\mathbf{u} = -g\nabla H - g\nabla H' - \nabla\left(\frac{p_a}{\rho_0} + gZ\right). \tag{9}$$

Here, we introduced $\tilde{\mathbf{u}}$ which is the final velocity at time $n+1$ in the iterative nonlinear scheme. In this context, $\mathbf{u}$ will be intermediate velocity field during the nonlinear iteration and the balance between $\tilde{\mathbf{u}}$ and $\mathbf{u}$ will be enforced through the gradient of H'. This process is similar to the projection methods commonly used to solve incompressible Navier Stokes equations [7][8][9][10]. Note that as $H' \to 0$, then $\nabla H' \to 0$ which ensures $\mathbf{u} \to \tilde{\mathbf{u}}$. We use the fractional time splitting method to first compute an intermediate velocity from Eq. (3) which is the predictor step and then use the results in the correction phase described as:

$$g\frac{\Delta t}{\alpha_1}\nabla H' = \mathbf{u} - \tilde{\mathbf{u}} \tag{10}$$

We can observe that Equation (3) used in the predictor phase is time discretized momentum equation in its original form. Clearly, the predictor phase satisfies consistency criteria and conserves the momentum. To derive the continuity wave equation, we multiply Eq. (10) by H and then take the divergence to obtain

$$g\frac{\Delta t}{\alpha_1}\nabla\cdot\,H\nabla H' = \nabla\cdot\,H\mathbf{u} - \nabla\cdot\,H\tilde{\mathbf{u}} \tag{11}$$

Since the last term in Eq. (11) includes the final velocity at time step $n+1$, we can replace it by its equivalent in

 ISBN: 978-1-61208-419-0

continuity equation. The time discretization of continuity equation with perturbed H is

$$\nabla\cdot\ H\tilde{\mathbf{u}}\ =\dot{n}-\frac{\alpha_1 H+\alpha_1 H'+\alpha_0 H^n+\alpha_{-1}H^{n-1}}{\Delta t} \tag{12}$$

We combine (11) and (12) and obtain time-discretized wave equation. The results can be written as

$$\begin{aligned} &H'+\frac{\Delta t}{\alpha_1}\nabla\cdot\ H'\mathbf{u}\ -\frac{\Delta t^2}{\alpha_1^2}\nabla\cdot\ C^2\nabla H' \\ &=-\frac{\Delta t}{\alpha_1}\left[\frac{\alpha_1 H+\alpha_0 H^n+\alpha_{-1}H^{n-1}}{\Delta t}\ +\nabla\cdot\ H\mathbf{u}\ -\dot{n}\right]. \end{aligned} \tag{13}$$

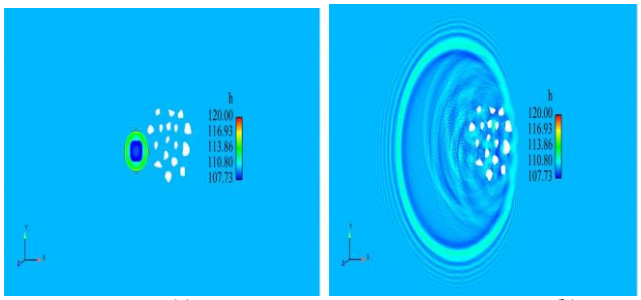
where $C=\sqrt{gH}$ is the wave speed. It can be seen that the right hand side of (13) is weighted by time discretized continuity equation. Therefore, as $H'\to 0$, (13) will yield zero residual for continuity equation. Clearly, Eq. (13) satisfies consistency criteria and conserves the mass. We use the cell-centered finite volume method (FV) to solve the momentum equation for the intermediate velocity and the node-based Galerkin finite element method (FE) to solve the wave equation and also for the elasticity equations. From velocity and water depth, forces acting on the individual ice floes are calculated which are used to solve the linear elasticity equations for mesh displacement using finite element method. In our deployment, the velocity unknowns are put at the cell centers and water depth variable is put at the mesh vertices. This deployment makes it convenient to compute the gradients of water depth using local finite element basis function, which is required in solving the momentum equations. Previous numerical results have shown that our hybrid implementation is super convergent in terms of the spatial convergence rates [7][8][9]. Unlike our previous compressible/incompressible flow solvers, the present hybrid FE/FV has not yet been parallelized. For realistic simulations of large regions of MIZ however, the flow solver will eventually be parallelized in future studies.

IV. RESULTS

To test our hybrid FE/FV methodology, non-moving ice floes were first studied, followed by moving ice-floe simulations. In the present simplistic study, idealized conditions are maintained by ignoring wind, tide, and Coriolis forces. Additionally, the ocean bed elevation was assumed to be flat (Z = 0).

A. Non-moving Ice-Floes

Wave-ice interactions in a 10 Sq-Km ocean domain with simplified initial/boundary conditions are studied utilizing our hybrid FE/FV flow solver. At the far open ocean boundary, symmetric boundary conditions are applied on all four sides. Ice floes are initially assumed to be rigid, non-moving, and wave effects are analysed on both uniform and also non-uniform randomly placed ice floes. Different simplified artificial forcing mechanisms are imposed to study wave effects on ice floes. Figure 2 below shows circular tsunami type waves impacting rigid, non-uniform, randomly placed ice floes, where water height is plotted at different non-dimensional times. Present hybrid FE/FV extensions combine the merits of both finite element and finite volume methods and are particularly suitable for high aspect ratio grids around ice floes and also for solving incompressible flows. As shown in Figure 2, wave features are well resolved bouncing back after impacting the ice floes. Validation and benchmark comparisons of our hybrid FE/FV methodology for shallow water equations can be found in the work of Aliabadi et al. [9].



(a) (b)

Figure 2. Circular waves impacting randomly distributed non-uniform, non-moving ice floes at different non-dimensional time's.

B. Moving Ice-Floes (Automatic Mesh Movement)

Having tested the flow solver on non-moving ice floes, automatic mesh movement was implemented in the flow solver by solving linear elasticity equations. Figure 3 below shows the grid around an isolated circular ice floe moving with a given constant velocity. As it can be seen from the figure, mesh refinement around the circular floe is well maintained as it moves in the ocean domain.

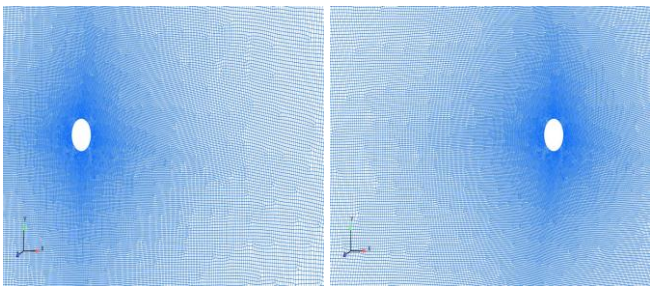
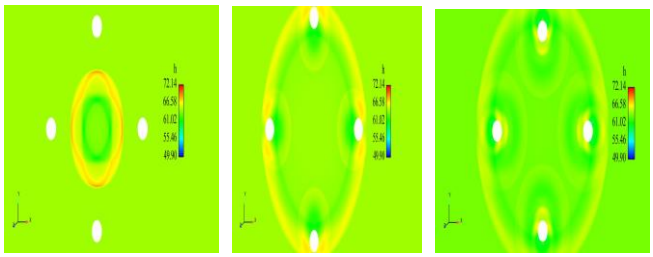


Figure 3. Isolated circular ice floe moving with a given constant velocity.

Circular waves impacting multiple circular ice floes are shown in Figure 4, where zoomed in portion of the ocean domain around the ice floes is shown. It can be seen that the waves sway the ice floes back and forth as they pass through.



(a) (b) (c)

Figure 4. Circular waves impacting uniformly placed movable ice floes at different non-dimensional time's (a) t = 2, (b) t = 5, (c) t = 6

V. CONCLUSIONS AND FUTURE WORK

Hybrid FE/FV methods have been successfully extended for studying ocean ice dynamics in idealized conditions. Automatic mesh movement was also successfully implemented, with well resolved and captured wave features and also grid refinement around the ice floes. Present hybrid FE/FV simulations are the first of its kind in direct simulation of individual ice floe-wave interactions. Mesh moving studies that are presently being conducted are based on the movement of a single and also multiple ice floes perturbed using different forcing mechanisms. Ongoing and potential future work in this highly challenging area is described below (not in any particular order).

- Moving isolated ice floes with different shapes (circular, square, non-uniform) and sizes will be conducted to more thoroughly understand the influence of shape/size on the ice motion through impinging waves.
- While experimental data in large MIZ regions is severely limited, select recent works have focussed on calculating the drift velocity for individual ice floes through laboratory experiments [11][12] and also by theoretical studies [13][14][15]. Single floe studies are crucial for thoroughly analysing the kinematic response of ice floes with characteristic lengths comparable or lower than the impact wavelengths. Isolated ice-floe impacts with offshore structures are observed to be one of the common ice-structure interaction events [11][16]. However, literature and guidance on the impact forces, based on which off-shore structures can be designed, appears to be severely limited. Results from the present study will therefore be compared with available data. Ice floe dynamics studies, such as in the present work, can play a significant role in estimating the forces from single/multi-year ice impacts on offshore platforms.
- Multiple ice-floe studies need to account for floe-floe interactions requiring constitutive relationships.
- Apart from the ocean waves, ice mobility in the MIZ, can also be influenced by wind stress, current, pressure and tides which will be incorporated in future studies.
- Realistic simulation of large regions in the MIZ will require tremendous computing power. Therefore, parallelizing the hybrid FE/FV shallow water flow solver is an important goal for future studies.

ACKNOWLEDGMENT

This work is funded by US Army Research Office (ARO).

REFERENCES

[1] V.A. Squire, J.P. Dugan, P. Wadhams, P.J. Rottier, and A.K. Liu, "Of Ocean Waves and Sea Ice," Annu. Rev. Fluid Mech, 27, 1995, pp. 115–168.

[2] V.A. Squire, "Of ocean waves and sea-ice revisited," Cold Reg. Sci. Technol, 49, 2007, pp. 110–133.

[3] M.H. Meylan, "Spectral solution of time-dependent shallow water hydroelasticity," J. Fluid Mech, 454, 2002, pp. 387–402.

[4] M.H. Meylan, C. Hazard, and F. Loret, "Linear time-dependent motion of a two-dimensional floating elastic plate in finite depth water using the Laplace transform," Proceedings 19th International Workshop on Water Waves and Floating Bodies, 2004, Cortona, Italy.

[5] C.B. Vreugdenhil, "Numerical Methods for Shallow Water Flow," Kluwer Academic Publishers, Dordrecht, Netherlands, 2010.

[6] K. Stein, T. Tezduyar, and R. Benney, "Mesh Moving Techniques for Fluid-Structure Interactions with Large Displacements," Transactions of the ASME, 70, 2003, pp.58-63.

[7] S. Tu, S. Aliabadi, "Development of a Hybrid Finite Volume/Element Solver for Incompressible Flows," International Journal of Numerical Methods in Fluids, 55, 2007, pp. 177-203.

[8] S. Tu, S. Aliabadi, R. Patel, and M. Watts, "An Implementation of the Spalart-Allmaras DES Model in an Implicit Unstructured Hybrid Finite Volume/Element Solver for Incompressible Turbulent Flow," International Journal of Numerical Methods in Fluids, 59, 2009, pp. 1051-1062.

[9] S. Aliabadi, M.K. Akbar, and R. Patel, "Hybrid Finite Element / Volume Method for Shallow Water Equations," Int. J. Num. Meth. Fluids, 83 (13), 2010, pp. 1719-1738.

[10] J.L. Guermond, P. Minev, and J. Shen, "An Overview of Projection Methods for Incompressible Fows," Computer Methods in Applied Mechanics and Engineering, 195, 2006, pp. 6011–6045.

[11] D.J. McGovern, and W. Bai, "Experimental Study on Kinematics of Sea Ice Floes in Regular Waves," Cold Regions Science and Technology, 103, 2014, pp. 15-30.

[12] G. Huang, A.L. Wing-Keung, and H. Huang, "Wave-induced Drift of Small Floating Objects in Regular Waves," Ocean Eng, 38, 2011, pp. 712–718.

[13] R. Grotmaack, and MH. Meylan, "Wave Forcing of Small Floating Bodies," Journal of Waterway, Port, Coastal, and Ocean Engineering, ASCE 132 (3), 2006, pp. 192–198.

[14] V.W. Harms, "Steady Wave-drift of Modeled Ice Floes," Journal of Waterway, Port, Coastal, and Ocean Engineering, ASCE 113 (6), 1987, pp. 606–622.

[15] H.H. Shen, and Y. Zhong, "Theoretical Study of Drift of Small Rigid Floating Objects in Wave Fields," J. Waterw. Port Coast. Ocean Eng, 127 (6), 2001, pp. 343–351.

[16] G.W. Timco, "Isolated Floe Impacts," Cold Reg. Sci. Technol, 68, 2011, pp. 35–48.

 ISBN: 978-1-61208-419-0

A Characteristic Adaptive Wavelet Method for Aerosol Dynamic Equations

Qiang Guo and Dong Liang

Department of Mathematics and Statistics
York University
Toronto, Ontario, M3J 1P3, Canada
Emails: dliang@yorku.ca; pangpangguo@gmail.com

***Abstract*—In this paper, a characteristic adaptive wavelet method is developed for solving aerosol dynamic equations. The proposed method combines the adaptive multi-resolution technique and the characteristic method to obtain the fully adaptive multi-resolution scheme, in which the solution is represented and computed in dynamically evolved wavelet bases along the characteristic curves. It overcomes numerical dispersions and can use large time steps. The efficiency and accuracy of the new algorithm is verified by numerical experiments. The developed characteristic adaptive wavelet algorithm in the paper has great applications in the modelling of aerosol dynamics.**

Keywords–Aerosol dynamic equation; the characteristic technique; adaptive multi-resolution method; wavelet.

I. INTRODUCTION

Aerosols are now clearly identified as an important factor in many environmental aspects of climate and radiative forcing processes, as well as in the health effects of air quality [7][8][11][15][17]. The aerosol dynamics with respect to size distribution is a nonlinear partial differential and integral equation.

Numerical methods have been proposed to solve the aerosol dynamic equations such as sectional method [9], moment method [1][14], modal method [18], finite element method [16], and stochastic approach [5], etc. The modal and moment approaches have the high numerical efficiency but only applied to some particular cases. When the distribution function is required, sectional methods are popular technique in aerosol dynamic modelling. But the treatment of condensation by sectional approaches usually leads to extra numerical diffusion, while smears the steep fronts. Sandu and Borden [16] developed a framework of finite element methods for numerical solutions of the aerosol dynamic equations, Liang et al. [13] developed the characteristic finite element methods for aerosol dynamic equations, and Liang et al. [12] developed a splitting wavelet method for solving the general aerosol dynamic equations on time, particle size and vertical spatial coordinate.

The size of atmospherical aerosols spans order of magnitude and the mechanisms for different size regions are totally different, so the aerosol size distribution is highly uneven distributed, such as multiple lognormal distributions in some regions. Thus, the most important problem encountered in the solutions of aerosol dynamic equations is how to efficiently solve the equations in size and time since the aerosol distributions vary very sharply in the size direction. Another problem is to approximate the advection process caused by the condensation growth term.

Multiresolution methods have been recognized to be important adaptive techniques in the applications to solutions of Partial Differential Equations (PDEs). For many real problems, solutions often exhibit localized singular features, such as sharp peaks. Uniform basis function space is not a practical option since high resolution is only needed in small regions. For improving the accuracy, the localization property of the wavelets both in space and in frequency makes the adaptivity efficiently [2][3][6][10].

In the paper, a characteristic adaptive wavelet method is developed to solve the aerosol dynamic equations, in which the time derivative and the condensation advection are transferred to the directional derivative along the characteristics and then discretized by the difference along the characteristics. For approximating size distribution, the differential systems of equations in time variable are obtained based on the wavelet bases. Owe to the advantage of characteristics method, we can refine the adaptive wavelets at the next time step along the characteristic curves. Adaptive space refinement strategy can follow the flow of solution over time. It reduces temporal errors and eliminates the excessive numerical dispersion. Compared with the uniform mesh method, the characteristic adaptive wavelet method has higher computational efficiency. Numerical experiments show the excellent performance of the developed algorithm in simulating aerosol dynamics.

The paper is arranged as follows. The mathematical model of aerosol dynamic system is presented in Section 2. In Section 3, the characteristic adaptive wavelet scheme is proposed for the aerosol dynamic equations. Numerical experiments are given in Section 4. Finally, we address some conclusions in Section 5.

II. AEROSOL DYNAMIC EQUATIONS

The aerosol dynamic equations can be described as [7]

$$\begin{aligned}\frac{\partial n(v,t)}{\partial t} &= -\frac{\partial[G(v)n(v,t)]}{\partial v} - n(v,t)\int_{V_{min}}^{V_{max}} \beta n(w,t)dw \\ &\quad +\frac{1}{2}\int_{V_{min}}^{v-V_{min}} \beta n(v-w,t)n(w,t)dw, \qquad (1)\end{aligned}$$

with boundary and initial conditions

$$n(V_{min},t) = 0, \quad t \in (0,T], \qquad (2)$$

$$n(v,0) = n_0(v), \quad v \in \Omega. \qquad (3)$$

 ISBN: 978-1-61208-419-0

where $t > 0$ is the time, v is the aerosol particle volume, and $T > 0$ is the time period. $n(v,t)$ is the number concentration distribution associated with particles volume v at time t. In practice, one assumes that the particle population has a nonzero minimal volume and a finite maximal volume, i.e., in a finite volume interval $[V_{min}, V_{max}]$, where V_{min} and V_{max} are chosen as lower and upper limits of the aerosol volume respectively. The condensation growth rate $G(v)$ is defined as the rate of change of the volume of a particle of volume v and $G(v) = \sigma_0 v$ will be considered in this paper due to the important application of linear growth rate. Coagulation of aerosol particles occurs through a variety of mechanisms such as Brownian motion, turbulent diffusion, etc. The coefficient β is the coagulation kernel.

III. The Characteristic Adaptive Wavelet Scheme

A. The Characteristic Method

For treating the condensation advection efficiently, we first propose the characteristic semi-discretization scheme in time. Denote the number of time steps by the positive integer Q and the time level by $t^q = q\Delta t$, $q = 0, 1, \cdots, Q$, where Δt is the time step size. For any particle size x at time $t = t^{q+1}$, the characteristics curve $X(x,t;\tau)$ passing through (x,t) satisfies:

$$\begin{cases} \frac{dX}{d\tau}(x,t^{q+1};\tau) = a\sigma_0 X(x,t^{q+1};\tau), \\ X(x,t^{q+1};t^{q+1}) = x. \end{cases} \tag{4}$$

where τ is the characteristics direction. Let $\hat{x}$ be the intersection point of tracking back along the characteristic curve from the point (x,t^{q+1}) to time level $t = t^q$.

For the aerosol dynamic equations in logarithmic coordinates, the characteristic semi-discretization scheme is defined as

$$\begin{aligned} \frac{n(x,t^{q+1}) - n(\hat{x},t^q)}{\Delta t} &= -\sigma_0 n(x,t^{q+1}) \\ &+ \frac{\beta}{2}\int_0^{aln(e^{x/a}-1)} \frac{e^{(y-b)/a}}{a} n(x^*,t^q)n(y,t^q)dy \\ &- \beta n(x,t^q)\int_0^1 \frac{e^{(y-b)/a}}{a} n(y,t^q)dy \end{aligned} \tag{5}$$

with initial value $n(x,0) = n^0(x)$ and the boundary condition $n(x,t) = 0$.

Let $\tilde{V}^{q+1}(\Omega)$ be wavelet space at $t = t^{q+1}$ defined in the next section. Then, the characteristic wavelet scheme is to find $n(x,t^{q+1}) \in \tilde{V}^{q+1}(\Omega)$ such that

$$\begin{aligned} &\big((1+\sigma_0\Delta t)n(x,t^{q+1}),\xi(x)\big) = (n(\hat{x},t^q),\xi(x)) \\ &+\tfrac{\Delta t\beta}{2}\left(\int_0^{aln(e^{x/a}-1)} \tfrac{e^{(y-b)/a}}{a_2} n(x^*,t^q)n(y,t^q)dy, \xi(x)\right) \\ &-\Delta t\beta\left(n(x,t^q)\int_0^1 \tfrac{e^{(y-b)/a}}{a} n(y,t^q)dy, \xi(x)\right) \end{aligned} \tag{6}$$

with $n(x,0) = n_0(x)$.

B. The Characteristics Adaptive Wavelet Algorithm

In this section, we shall construct an adaptive multiresolution scheme of Haar wavelets for (6).

Haar wavelets, which are Daubechies wavelets of order 1 (see [4]), consist of piecewise constant functions and are therefore the simplest orthonormal wavelets with a compact support. Because of the advantages of Haar wavelets, we will apply Haar wavelets as the basis functions in the scheme (6). Let $\psi(x)$ be the Haar wavelet, and the corresponding scaling function $\phi(x)$ The adaptive space $\tilde{V}^{q+1}(\Omega)$ composed by the Haar scaling functions is the cell-average multiresolution representation, where $J_0 \le j \le J$ and J_0 is the coarsest resolution level and J is the highest resolution level. The scaling coefficients $c_{j,k}^{q+1}$ are cell-average values. Find $\tilde{n}^{q+1}(x) \in \tilde{V}^{q+1}(\Omega)$ with

$$\tilde{n}^{q+1}(x) = \sum_{(j,k)\in\tilde{\Lambda}^{q+1}} c_{j,k}^{q+1}\phi_{j,k}(x) \tag{7}$$

in the scheme (6), where $\tilde{\Lambda}^{q+1}$ is the index set of scaling functions at $t = t^{q+1}$. Once $\hat{\Lambda}^q$ is determined, which is the final index set at $t = t^q$, we initialize $\tilde{\Lambda}^{q+1}$ from $\hat{\Lambda}^q$ by tracking along the characteristics.

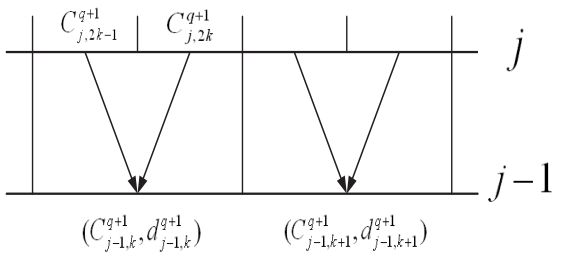


Figure 1. The operator P_j^{j-1}.

To estimate the cell-averages and detail information at level $j-1$ from the ones of the level j, we use the multiresolution transform P_j^{j-1}, see Figure 1, as

$$c_{j-1,k}^{q+1} = \frac{1}{\sqrt{2}}\left(c_{j,2k-1}^{q+1} + c_{j,2k}^{q+1}\right), \tag{8}$$

$$d_{j-1,k}^{q+1} = \frac{1}{\sqrt{2}}\left(c_{j,2k-1}^{q+1} - c_{j,2k}^{q+1}\right). \tag{9}$$

Then, we have $\tilde{n}^{q+1}(x)$

$$\begin{aligned} \tilde{n}^{q+1}(x) = \sum_{(j,2k),(j,2k-1)\in\tilde{\Lambda}^{q+1}} &[c_{j-1,k}^{q+1}\phi_{j-1,k}(x) \\ &+ d_{j-1,k}^{q+1}\psi_{j-1,k}(x)]. \end{aligned} \tag{10}$$

A threshold parameter ϵ is prescribed for the adaptive procedure

$$\epsilon_j = \epsilon/2^{j-J_0}, \quad J_0 \le j \le J-1.$$

If $|d_{j-1,k}^{q+1}| < \epsilon_j$, we reduce $\phi_{j,2k-1}$ and $\phi_{j,2k}$ from the space $\tilde{V}^{q+1}(\Omega)$; while if $d_{j-1,k}^{q+1}$ is big, then we add $\phi_{j+1,4k-3}$, $\phi_{j+1,4k-2}$ $\phi_{j+1,4k-1}$ and $\phi_{j+1,4k}$, based on operator P_j^{j+1}, see Figure 2. The space after adjustment is called as $\hat{V}^{q+1}(\Omega)$ and the corresponding index set is $\hat{\Lambda}^{q+1}$.

The important feature of the characteristic adaptive wavelet algorithm is that it adjusts the approximation wavelet space

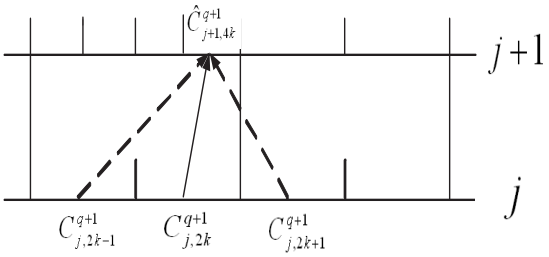


Figure 2. The operator P_j^{j+1}.

at next time level by tracking back along the characteristics, which is further refined or coarsened by the adaptive wavelet technique. The highly accurate approximation can be obtained by the new algorithm even large time step sizes are used, while other classic algorithms need to use very small time steps.

IV. Numerical simulation

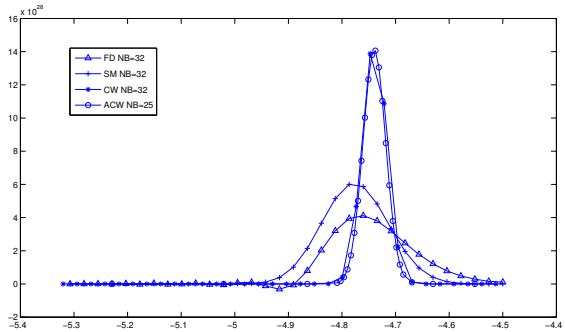
In this section, numerical examples are taken to illustrate the performance of the characteristic adaptive wavelet algorithm.

Example 1.

In this example, we consider the condensation process with initial single mode distribution. The initial distribution is a log-normal distribution on the volume domain $[1 \times 10^{-16}m^3, 1 \times 10^{-13.5}m^3]$ described by

$$n_0(v,t) = \frac{N_0}{3\sqrt{2\pi}\ln\sigma}\exp\left(-\frac{\ln^2(v/v_g)}{18\ln^2\sigma}\right)\frac{1}{v} \qquad (11)$$

with the volume concentration $N_0 = 5 \times 10^{10}$particles$/m^3$, the geometric average volume $v_g = 1 \times 10^{-15}m^3$ and the standard deviation $\sigma = 1.05$. We choose coarsest level $J_0 = 2$ and highest level $J = 6$ for our method.

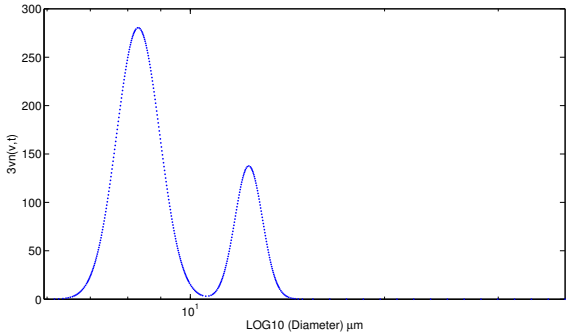


(a) $\sigma_0 = 1.8/h$ and $\Delta t = 0.05h$

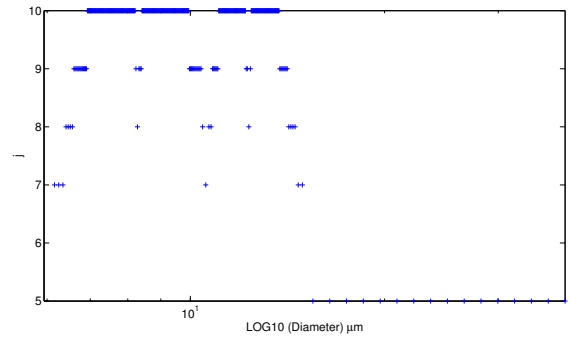
Figure 3. Comparison of number distribution $3vn(v,t)$ by different methods.

We take the numerical experiment of our Characteristic Adaptive Wavelet (CAW) method with threshold parameter $\epsilon = 10^{-2}$ and compare with other methods including upstream Finite Difference (FD) method, Sectional Method (SM) in [9] and the Characteristics Wavelet (CW) method, where the number of the bins NB is 32 for the last three methods. Figure 3 shows the numerical computations after 1h for different growth rates and time step sizes, where the vertical coordinate represents numerical distribution $3vn(v,t)$ and the horizontal coordinate is the logarithmic diameter of the aerosol particles.

From Figure 3, where the growth rate σ is $1.8/h$, numerical distributions obtained by the characteristic wavelet method and our characteristic adaptive wavelet algorithm are excellent. but both of the finite difference method and sectional method suffer from numerical diffusion greatly. Moreover, our characteristic adaptive wavelet method is only with the number of the bins $NB = 25$, however the characteristic wavelet method requires the number of the bins $NB = 32$. The numerical solutions by our algorithm with less number of bins show much better numerical efficiency even with a large time step and growth rate.



(a) Aerosol number distribution



(b) Adaptive mesh

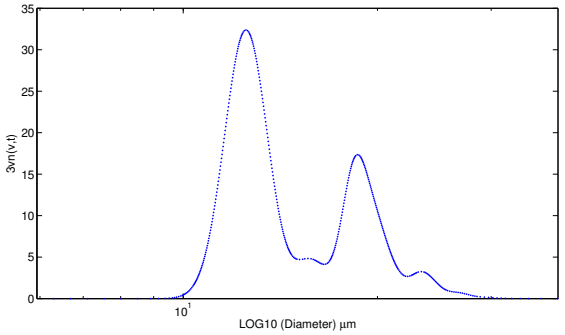
Figure 4. Aerosol distribution for linear condensational growth and constant coagulation kernel.

Example 2.

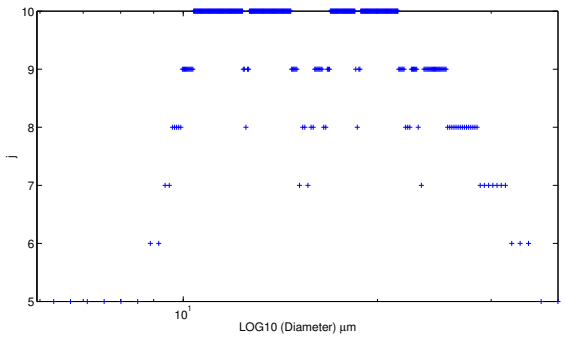
In this example, we consider aerosol dynamic systems evolving both condensation and coagulation processes with the initial two-modal log-normal distribution.

The volume domain is $[10^2\mu m^3, 10^{4.5}\mu m^3]$. Take time step $\Delta t = 0.1h$, $J_0 = 5$ and $J = 10$. Figures 4 and 5 show the numerical number densities of aerosol distribution and their corresponding adaptive bases at times $T = 0.1h$ and $40h$ with $\sigma = 0.03/h$ and coagulation kernel $\beta_0 = 0.01\mu m^3/h$. The horizontal coordinate represents logarithm of particle diameter. The vertical coordinate represents the number distribution $3vn(v,t)$ and resolution level separately in the left figures and right figures. In Figure 4, we can see that the multiresolution

bases are centered at the places where the two peaks of the two-modal log-normal distributions located. Since the number condensations at the larger particle size region are very small, it's good enough to describe the number condensation with coarsest resolution level.



(a) Aerosol number distribution



(b) Adaptive mesh

Figure 5. Aerosol distribution for linear condensational growth and constant coagulation kernel.

As time goes on, shown in Figure 5, the number condensation moves forward along the size direction, meanwhile number condensation of smaller particles decreases and that of larger particles increases because coagulation is the process whereby two particles collide and form larger particle, as a result, higher resolution level bases are adaptively added to capture the change of the distribution at the larger particle size.

V. Conclusion and Future Work

A new characteristic adaptive wavelet algorithm was developed for solving the aerosol dynamic equations. The considered model is a nonlinear partial differential and integral equation with hyperbolic part from the condensation term.

Using the multiresolution technique, the computational bases are reduced by deleting non-significant wavelet coefficients while keeping the desired accuracy. The adaptive space refinement strategies are simplified and we refine the adaptive bases at the next time step along the characteristic curves, which save computational time and memory.

We demonstrated numerically the efficiency of the characteristic adaptive wavelet algorithm for different tests of the condensation process and the joint effect of condensation and coagulation processes. The method exhibited good shape and high accuracy even when large time steps are used in computations, which has great applications in the modelling of aerosol dynamics.

The proposed characteristic adaptive wavelet method can be further extended to solve aerosol spatial transport problems in atmosphere, where the characteristic adaptive wavelet technique can efficiently treat the transport process. Developing fast and adaptive algorithms for the general aerosol dynamic process will be our another interested work.

Acknowledgment

The work was supported partially by Natural Sciences and Engineering Research Council of Canada.

References

[1] J. R. Brock and J. Oates, "Moment simulation of aerosol evaporation", J. Aerosol. Sci., vol. 18, 1987, pp. 59-64.

[2] A. Cohen, Numerical Analysis of Wavelet Methods, Amsterdam; Boston: Elsevier, 2003.

[3] A. Cohen, W. Dahmen, and R. Devore, "Adaptive wavelet methods for elliptic operator equations: convergence rates", Math. Comput., vol. 70, 2000, pp. 27-75.

[4] I. Daubechies, Ten lectures on wavelets, Philadelphia, PA: Society for Industrial and Applied Mathematics, 1992.

[5] E. Debry, B. Sportisse, and B. Jourdain, "A stochastic approach for the numerical simulation of the general dynamics equation for aerosols", J. Comput. Phys., vol. 184, 2003, pp. 649-669.

[6] J. L. Diaz Calle, P. R. B. Devloo, and S. M. Gomes, "Wavelets and adaptive grids for the discontinuous Galerkin method", Numer Algorithms, vol. 39, 2005, pp. 143-154.

[7] S. K. Friedlander, Smoke, Dust, and Haze: Fundamentals of Aerosol Dynamics, Oxford University Press, USA, 2000.

[8] Y. Gao, X. Liu, C. Zhao, and M. Zhang, "Emission controls versus meteorological conditions in determining aerosol concentrations in Beijing during the 2008 Olympic Games", Atmospheric Chemistry and Physics, vol. 11, 2011, pp.12437-12451.

[9] F. Gelbard, Y. Tambour, and J. H. Seinfeld, "Sectional representations for simulating aerosol dynamics", J. Colloid. Interf. Sci., vol. 76, 1980, pp. 541-556.

[10] M. Holmström, and J. Waldén, "Adaptive wavelets methods for hyperbolic PDEs", J. Sci. Comput., vol. 13, 1998, pp. 19-49.

[11] M. Z. Jacobson, Fundamentals of Atmospheric Modelling, Cambridge University Press, Cambridge, 1999.

[12] D. Liang, Q. Guo, and S. L. Gong, "A new splitting wavelet method for solving the general aerosol dynamics equation", J. Aerosol. Sci., vol. 39, 2008, pp. 467-487.

[13] D. Liang, W. Wang, and Y. Cheng, "An efficient second order characteristic finite element method for nonlinear aerosol dynamic equations", Int. J. Numer. Methods Engng., vol. 80, 2009, pp. 338 - 354.

[14] R. McGraw, "Description of aerosol dynamics by the quadrature method of moments", Aerosol Sci. Tech., vol. 27, 1997, pp. 255-265.

[15] U. Nopmongcol, et al., "Modeling Europe with CAMx for the air quality model evaluation international initiative (AQMEII)", Atmospheric Environment, vol. 53, 2012, pp.177-185.

[16] A. Sandu and C. Borden, "A framework for the numerical treatment of aerosol dynamics", Appl. Numer. Math., vol. 45, 2003, pp. 475-497.

[17] J. Seinfeld and S. Pandis, Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, Second Edition, Wiley-Intersciences, New York, 2006.

[18] E. R. Whitby and P. H. McMurry, "Modal aerosol dynamics modeling", Aerosol. Sci. Tech., vol. 27, 1997, pp. 673-688.

 ISBN: 978-1-61208-419-0