INFOCOMP 2023

The Thirteenth International Conference on Advanced Communications and Computation

ISBN: 978-1-68558-073-5

June 26 - 30, 2023

Nice, France

INFOCOMP 2023 Editors

Claus-Peter Rückemann, Westfälische Wilhelms-Universität Münster (WWU);
DIMF, Germany, LUH, Germany

INFOCOMP 2023

Forward

The Thirteenth International Conference on Advanced Communications and Computation (INFOCOMP 2023), held between June 26th and June 30th, 2023, continued a series of events dedicated to advanced communications and computing aspects, covering academic and industrial achievements and visions.

The diversity of semantics of data, context gathering and processing led to complex mechanisms for applications requiring special communication and computation support in terms of volume of data, processing speed, context variety, etc. The new computation paradigms and communications technologies are now driven by the needs for fast processing and requirements from data-intensive applications and domain-oriented applications (medicine, geoinformatics, climatology, remote learning, education, large scale digital libraries, social networks, etc.). Mobility, ubiquity, multicast, multi-access networks, data centers, cloud computing are now forming the spectrum of de factor approaches in response to the diversity of user demands and applications. In parallel, measurements control and management (self-management) of such environments evolved to deal with new complex situations.

We take here the opportunity to warmly thank all the members of the INFOCOMP 2023 technical program committee, as well as all the 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 who dedicated much of their time and effort to contribute to INFOCOMP 2023. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the INFOCOMP 2023 organizing committee for their help in handling the logistics of this event.

We hope that INFOCOMP 2023 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of communications and computation.

INFOCOMP 2023 Chairs

INFOCOMP 2023 Steering Committee

Claus-Peter Rückemann, Westfälische Wilhelms-Universität Münster (WWU) / DIMF / Leibniz Universität Hannover, Germany
Nicola Calabretta, Eindhoven University of Technology, Netherlands

INFOCOMP 2023 Publicity Chairs

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

INFOCOMP 2023
Committee

INFOCOMP 2023 Steering Committee

Claus-Peter Rückemann, Westfälische Wilhelms-Universität Münster (WWU) / DIMF / Leibniz Universität Hannover, Germany
Nicola Calabretta, Eindhoven University of Technology, Netherlands

INFOCOMP 2023 Publicity Chairs

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

INFOCOMP 2023 Technical Program Committee

Vicki H. Allan, Utah State University, USA
Mohammad AlMasri, Nvidia Corporation, USA
Daniel Andresen, Kansas State University, USA
Vijayan K. Asari, University of Dayton, USA
Marc Baaden, CNRS, France
Bernhard Bandow, GWDG - Göttingen, Germany
Christine Bassem, Wellesley College, USA
Raoudha Ben Djemaa, MIRACL, Sfax, Tunisia
Tekin Bicer, Argonne National Laboratory, USA
Julien Bigot, Maison de la Simulation / CEA, France
David Boehme, Lawrence Livermore National Laboratory, USA
Radouan Boukharfane, KAUST, Saudi Arabia / MSDA | UM6P, Morocco
Abbas Bradai, University of Poitiers, France
Stephanie Brink, Lawrence Livermore National Laboratory, USA
Paolo Burgio, University of Modena and Reggio Emilia, Italy
Xiao-Chuan Cai, University of Colorado Boulder, USA
Nicola Calabretta, Eindhoven University of Technology, Netherlands
Jian Chang, Bournemouth University, UK
Jieyang Chen, Oak Ridge National Laboratory, USA
Albert M. K. Cheng, University of Houston, USA
Enrique Chirivella Pérez, Universitat de Valencia, Spain
Noelia Correia, Center for Electronics Opto-Electronics and Telecommunications (CEOT) | University of Algarve, Portugal
Tiziano De Matteis, ETH Zurich, Switzerland
Iman Faraji, Nvidia Inc., Canada
Josué Feliu, Universitat Politècnica de València, Spain
Francesco Fraternali, University of California, San Diego, USA
Hans-Hermann Frese, Gesellschaft für Informatik e.V., Germany
Steffen Frey, Visualization Research Center - University of Stuttgart, Germany
Marco Furini, University of Modena and Reggio Emilia, Italy
Jason Ge, Snark AI Inc, USA

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Franca Giannini, IMATI-CNR, Italy
Barbara Guidi, University of Pisa, Italy
Anqi Guo, Boston University, USA
Önder Gürcan, CEA LIST, France
Bing He, Georgia Institute of Technology, USA
Nikhil Hegde, Indian Institute of Technology Dharwad, India
Enrique Hernández Orallo, Universidad Politécnica de Valencia, Spain
Mert Hidayetoglu, University of Illinois at Urbana-Champaign, USA
Md Shafaeat Hossain, Southern Connecticut State University, USA
Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek, Hannover, Germany
Thomas Hupperich, University of Münster, Germany
Mohamed Assem Ibrahim, William & Mary, USA
Sergio Ilarri, University of Zaragoza, Spain
Ali Jannesari, Iowa State University, USA
Yunhan Jia, Bytedance Inc., China
Eugene B. John, The University of Texas at San Antonio, USA
Izabela Karsznia, University of Warsaw, Poland
Alexander Kipp, Robert Bosch GmbH, Germany
Felix Klapper, Leibniz Universität Hannover, Germany
Zlatinka Kovacheva, Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, Sofia, Bulgaria
Christian Köhler, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Germany
Manfred Krafczyk, Institute for Computational Modeling in Civil Engineering (iRMB) - TU Braunschweig, Germany
Nane Kratzke, Lübeck University of Applied Sciences, Germany
Navjot Kukreja, University of Liverpool, UK
Sonal Kumari, Samsung Research India-Bangalore (SRI-B), India
Julian M. Kunkel, University of Reading, UK
Stephen Leak, NERSC User Engagement, USA
Seyong Lee, Oak Ridge National Laboratory, USA
Yiu-Wing Leung, Hong Kong Baptist University, Kowloon Tong, Hong Kong
Hongbo Li, Latitude AI, USA
Peizhao Li, Brandeis University, USA
Shigang Li, ETH Zurich, Switzerland
Yanting Li, Shaoguan University, China
Walter Lioen, SURF, Netherlands
Jiyao Li, Utah State University, USA
Jinwei Liu, Florida A&M University, USA
Hui Lu, SUNY Binghamton, USA
Sandeep Madireddy, Argonne National Laboratory, USA
Sumit Maheshwari, Microsoft, USA
Adnan Mahmood, Macquarie University, Australia / Telecommunications Software & Systems Group, WIT, Republic of Ireland
Antonio Martí-Campoy, Universitat Politècnica de València, Spain
Artis Mednis, Institute of Electronics and Computer Science, Latvia

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

Simulation of Pipeline Transport of Carbon Dioxide with Impurities

Mehrnaz Anvari
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
email: Mehrnaz.Anvari@scai.fraunhofer.de

Anton Baldin
PLEdoc GmbH and Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
email: Anton.Baldin@scai.fraunhofer.de

Tanja Clees
University of Applied Sciences Bonn-Rhein-Sieg and Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
email: Tanja.Clees@scai.fraunhofer.de

Bernhard Klaassen
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
email: Bernhard.Klaassen@scai.fraunhofer.de

Igor Nikitin
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
email: Igor.Nikitin@scai.fraunhofer.de

Lialia Nikitina
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
email: Lialia.Nikitina@scai.fraunhofer.de

Sabine Pott
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
email: Sabine.Pott@scai.fraunhofer.de

***Abstract*—The transport of carbon dioxide through pipelines is one of the important components of Carbon dioxide Capture and Storage (CCS) systems that are currently being developed. If high flow rates are desired, a transportation in the liquid or supercritical phase is to be preferred. For technical reasons, the transport must stay in that phase, without transitioning to the gaseous state. In this paper, a numerical simulation of the stationary process of carbon dioxide transport with impurities and phase transitions is considered. We use the Homogeneous Equilibrium Model (HEM) and the GERG-2008 thermodynamic equation of state to describe the transport parameters. The algorithms used allow to solve scenarios of carbon dioxide transport in the liquid or supercritical phase, with the detection of approaching the phase transition region. Convergence of the solution algorithms is analyzed in connection with fast and abrupt changes of the equation of state and the enthalpy function in the region of phase transitions.**

***Index Terms*—simulation and modeling; mathematical and numerical algorithms and methods; advanced applications; carbon dioxide; capture and storage; pipeline transport.**

I. Introduction

To reduce greenhouse gas emissions into the atmosphere, Carbon dioxide Capture and Storage (CCS) systems are currently being developed. Typically, such systems consist of 3 parts: (1) capturing carbon dioxide (CO_2) at its source; (2) transporting CO_2 through pipelines to special storage sites; (3) and finally injecting it into wells, when underground storage is used. In this paper, we focus on the second part of the aforementioned process. It is generally required that CO_2 be in the liquid or supercritical phase during transport in order to increase the density and mass flows. It is essential to avoid the transition of fluid phase to gas, which leads to cavitation and destruction of the pipeline during transportation. To ensure reliable operation of the CO_2 pipeline, both an extensive experimental base and stable numerical simulation of the transportation process are required. At the same time, for a long-term planning, it is sufficient to simulate a stationary process of the transportation, with CO_2 in a 1-phase state and an indication of a possible phase transition, in order to prevent it.

The pioneering work [1] has considered in detail the stationary process of transporting pure CO_2 through a pipeline and pumping it into an underground storage, taking into account phase transitions. In that and in subsequent papers, the importance of taking into account impurities that have a strong influence on the parameters of the transportation process even at low concentrations, has been pointed out. The papers [2]–[9] considered the process of CO_2 transport, both stationary and dynamic. Papers [1]–[8] consider a Homogeneous Equilibrium Model (HEM), in which different phases of a fluid are homogeneously mixed and have the same speed, pressure, temperature and chemical potential. In papers [4]–[6], [8], [9], phase split is also considered, i.e., when the phases are geometrically separated, and phase slip, i.e., when the phases have different speeds. Also, in works [4], [6], [8] the formation of a solid phase of CO_2 (dry ice) is considered. In the works [5], [6], [8], [9], fast transient processes occurring during depressurization of a pipe are considered, together with the related experiments. The economic aspects of pipeline CO_2

 ISBN: 978-1-68558-073-5

transport have been considered in papers [10]–[13].

In this paper, we describe a stationary simulation of the CO_2 transport process with the possibility of considering impurities, phase transitions, several sources with different composition, and networks of complex topology. Simulations of this type have extended the capabilities of our software MYNTS [14]–[18]. The system provides an open, freely configurable and user-friendly specification of modeling, defined as a list of variables and equations. An open Python code for workflow procedures is also provided. The main calculations are performed in a fast C++ solver. The system also has a Graphical User Interface (GUI) with the ability to edit networks and scenarios. This architecture allows to formulate and quickly solve very large network problems, as well as the ability to model different energy carriers and couple different energy sectors.

For problems of stationary transportation of fluids, we implement standard pipe transport equations with friction terms by Nikuradse [19], Hofer [20] and spatial discretization of type [21]. The GERG equation of state [22], [23], which is currently the ISO standard [24], is used to accurately model the thermodynamics of fluids, in particular CO_2 with impurities and phase transitions. Additionally, we have developed an algorithm for detecting the proximity to the region of phase transitions. A number of numerical experiments were carried out to test the developed algorithms. Based on them, it is shown that the fast, sometimes abrupt, behavior of the system in the presence of phase transitions affects the convergence properties of the numerical algorithms used for the solution. In the scenarios we have considered, the divergence, if it occurs, is entirely localized in the region of phase transitions. On the other hand, scenarios without phase transitions are converging, which makes it possible to solve them with detection of proximity to the region of phase transitions.

Section II reviews the physics of phase transitions applied to CO_2 with impurities. Section III discusses the transport equations used. In Section IV, we describe numerical experiments, with particular attention paid to the questions of convergence of iterative processes. Finally, in Section V, we summarize our results.

II. Physics of phase transitions

Phase transitions occur in slightly different ways for pure substances and their mixtures. Figure 1a shows the phase transition for pure CO_2. At a constant temperature, the pressure decreases starting in the region of the liquid state. There is a line of phase transitions on the diagram. When the pressure decreases, the process proceeds until it intersects with this line, after that the pressure decrease stops until all the fluid passes from the liquid state to the gaseous state. At the same time, Figure 1c shows that during this process, the average density changes from large values, typical for the liquid phase, to small values, typical for a gas. Figure 1b shows what happens in the case of a mixture, here 95% CO_2, 3% N_2, 2% O_2. Now, the 2-phase state corresponds not to a line, but to a region on (T, P)-diagram. The boundary of this region is called the

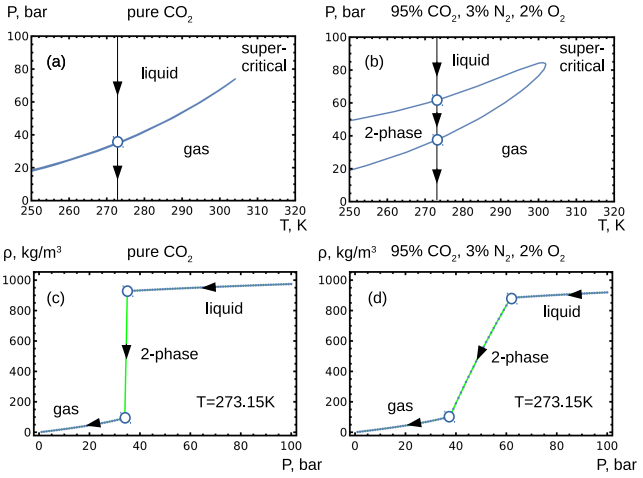

Fig. 1. Phase transitions at fixed temperature: (a),(c) – for pure CO_2; (b),(d) – for CO_2 with impurities.

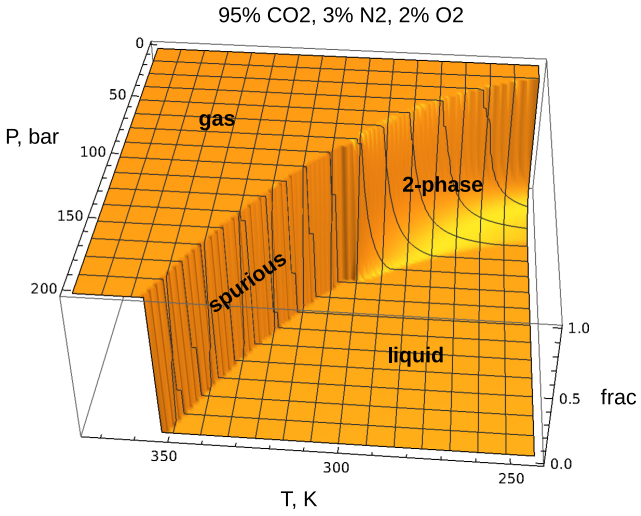

Fig. 2. Fraction of gaseous phase as a function of pressure and temperature.

Vapour-Liquid Equilibrium (VLE) diagram, or *phase envelope*. When the pressure decreases, the point enters this region and the fluid also passes from the liquid state to the gaseous state, but here the pressure continues to decrease. Figure 1d shows that in the 2-phase state, the density decreases in the same way as for pure substance, but at a decreasing pressure.

The 3D diagram in Figure 2 shows the behavior of $frac$-value, which varies in the interval $[0, 1]$ and measures the fraction of the gaseous phase in the fluid. Here, one can also see the region where the phase transition occurs, which proceeds continuously for mixed compositions. Also, this diagram has a jump on a line starting from the critical point, however this transition is spurious. Above the critical point, gas and liquid do not really differ from each other, but according to the scheme of description, it is required to make a transition from gas to liquid somewhere. Although the quantity $frac$ has a formal jump here, the physically measurable quantities have

 ISBN: 978-1-68558-073-5

no jumps on this line.

Interestingly, this surface resembles the surfaces considered in the theory of functions of a complex variable. Namely, if we take this surface, as well as the $1-frac$ surface and join them together, we get an object that looks like a Riemann surface for a complex square root. The similarity is not accidental, in both cases there is a 2-sheeted surface without the possibility of continuously separating the sheets from each other.

For the thermodynamical description of the fluid, the GERG equation of state and its accompanying implementation [22]–[24] is used. Technically, it is delivered as a software library where one can access a variety of functions describing the fluid state. In addition to the already mentioned phase envelope and $frac$-value, we use the Equation Of State (EOS) and energy functions

$$z = z(T, P, x),\ W = W(T, P, x), \tag{1}$$

where T is absolute temperature, P is pressure, x is a vector describing fluid composition, $W = (H, U, G, A)$ is a vector describing molar energies of different types: enthalpy, internal energy, Gibbs energy, Helmholtz energy, respectively. Compressibility factor z enters in the gas law $P = \rho RTz/\mu$, where R is the universal gas constant, ρ is the mass density, μ is the molar mass.

As a parameter important for the user, the $frac$-value or a conservative algorithm based on $frac$-values in the vicinity of the solution can be used to detect the proximity of phase transitions:

Algorithm (proximity-alarm):

```
given (T0,P0,x,dT,dP,val)
for T in (T0-dT,T0,T0+dT)
  for P in (P0-dP,P0,P0+dP)
    if frac(T,P,x)!=val return true
return false.
```

The algorithm considers a 3x3 grid created by $(\pm dP, \pm dT)$-variations, and if $frac$ differs from the user-specified val at least at one point, triggers a proximity alarm. This simple algorithm is applied to every node in the network. It has the advantage that it works even in the networks with many fluid compositions, i.e., variable x-values. Alternative algorithms based on the construction of the phase envelope produce many diagrams for different compositions, which complicates the analysis. At the same time, this algorithm has one drawback, it can produce a false alarm when approaching a spurious line. In this case, the user can visually control the solution trajectory on the (T, P)-diagram by constructing a phase envelope for the local network segment with constant x. The development of other algorithms for automatic detection of phase transitions that work for the variable composition of the fluid in the network is in our future plans.

III. Pipe transport equations

A pressure drop in the pipe in the stationary case is described by the equation:

$$dP/dL = -\lambda\rho v|v|/(2D) - d(\rho v^2)/dL - \rho g\, dh/dL, \tag{2}$$

where L is the running length along the pipe, v is the speed of the fluid, D is the internal diameter of the pipe, g is the gravitational acceleration, and h is the height. On r.h.s. the first term is usually dominant, describing the contribution of the friction force, defined in terms of the dimensionless friction coefficient $\lambda(k/D, Re)$ using the Nikuradse [19] formula or the more accurate Hofer [20] formula. Here, k is the pipe roughness, $Re = 4|Q_m|/(\pi\mu_{visc}D)$ is Reynolds number, where μ_{visc} is the dynamic viscosity and $Q_m = \rho v\pi D^2/4$ is the mass flow constant along the pipe. Further r.h.s. includes the convective and gravitational terms.

For discretization purposes, we consider a short pipe segment of length L and integrate the equation over it. Expressing the velocity in terms of the mass flow, and keeping only the leading first term for illustration, we get $dP/dL = c_1/\rho$, where c_1 is constant. When integrating, we replace the variable density ρ by the average $\bar\rho = (\rho_1 + \rho_2)/2$ over the end points of the segment, i.e., $P_2 - P_1 = c_1 L/\bar\rho$. As an alternative, we multiply the original equation by P, use the gas law $P/\rho = RTz/\mu$, replace the variables T and z with the end averages and, thereby, we get $(P_2^2 - P_1^2)/2 = c_1 LR\bar T\bar z/\mu$, in a more familiar quadratic form for gas dynamics [21]. To find the optimal pipe subdivision, the number of segments is increased until the solution stops changing, up to a given tolerance.

Temperature profiles are described by the equation

$$dH/dL = -\pi D c_h (T - T_s)\mu/Q_m, \tag{3}$$

according to which the enthalpy change in a segment of the pipe is equal to the heat exchange with the soil or other environment. Here, c_h is the heat transfer coefficient, T_s is the soil temperature. Note that when the heat exchange is switched off $c_h = 0$, the process described by this formula is isoenthalpic $dH = 0$, and the temperature change is related to the pressure change by the well-known formula $dT = \mu_{JT} dP$, where $\mu_{JT} = -(\partial H/\partial P)_T/(\partial H/\partial T)_P$ – Joule-Thomson coefficient. The equation can also be modified by introducing kinetic and gravitational terms.

For discretization, in the form $dH/dL = c_2(T - T_s)$ with constant c_2, the variable temperature T is replaced by the constant T_x, which can be taken as the end average $\bar T$ or the value of the outflow temperature T_{out}, which better represents the case of longer segments. After integration, we get $H_2 - H_1 = c_2 L(T_x - T_s)$. Further, in an iterative solution process in which the pressure profile and fluid composition are kept constant, the enthalpy values can be linearized using the formula $H(T^{i+1}) = H(T^i) + c_p(T^i)(T^{i+1} - T^i)$, where the superscripts indicate the number of iterations and $c_p = (\partial H/\partial T)_P$ is the isobaric molar heat capacity, also calculated by the GERG software library.

Next, we will consider in more detail the process of convergence of the iterations used for the solution. In our previous work [18], the architecture of MYNTS system has been described. Due to software-technical reasons, the solution was divided into 2 parts: (1) *Pressure-Massflow (PM)-iterations*,

 ISBN: 978-1-68558-073-5

solved by a sparse non-linear Newtonian solver; and (2) *mix-iterations*, solved by a sparse linear solver. PM iterations determine the pressure, density and mass flow, by solving a relatively small nonlinear system. This system, however, has strong numerical instabilities associated with nearly zero Jacobi matrix eigenvalues and requires special stabilization measures [17]. Mix iterations solve a large linear system defining a multicomponent fluid composition, determine temperature and call external modules, such as GERG that would otherwise be called too often in a fully coupled system. After the temperature linearization described above, all mix equations of the system at each iteration become linear, their solution can be produced by a sparse linear solver such as *Pardiso*. Further, these two processes are iterated, while using an additional stabilization algorithm *weighted relaxation* [18], the result of the combined PM-mix-iteration $h(x)$ is replaced by a weighted average $x_{i+1} = wh(x_i) + (1-w)x_i$.

Among the modeling limitations, it should be mentioned that the GERG module does not consider the solid phase and derives equilibrium conditions for the liquid and gaseous phases under the HEM assumptions. The transport equations considered here treat 2-phase solutions as 1-phase, with the values of thermodynamic parameters calculated by the GERG module in the *total* system, which also means calculations within the HEM framework.

At the end of this section, it is worth to mention a general point regarding the simulation of static and dynamic types. Often, the user assumes the uniqueness of the solutions obtained in the simulations. In general, this may not be the case. Existence and uniqueness theorems for solutions are formulated only in rare cases. So, for example, they are guaranteed for the PM subsystem under the conditions of generalized resistivity [14]. Being combined with the mix system, the uniqueness of the solution is not guaranteed. Theoretically imaginable is the situation when there are two stationary solutions, one 1-phase, the other 2-phase, and it may happen that the stationary solver finds the first one, but in reality the second one will be realized. Consideration of dynamic simulation can decide which solution the trajectory will go to when integrating from a given initial state. But even for a dynamic solver, saddle points, bifurcations of the solution are possible, where, with a small variation, the solution can go in one direction or the other. Questions about the uniqueness of stationary solutions and the stability of dynamic solutions must be investigated in the practical analysis of simulation results.

IV. Numerical experiments

To test the implemented algorithms, we use a pipe segment with parameters taken from [1]. In our experiments, different scenarios are considered, see Table I. In the first scenario, a small flow is set, at which no phase transitions occur. The entire pipe is filled with liquid or supercritical fluid. In the second scenario, a larger flow is set, the pressure drops more strongly, and a phase transition occurs in the system. Both

TABLE I
Parameters of test scenarios

parameter	symbol [units]	value
total pipe length	$L_{tot}[km]$	150
pipe internal diameter	$D[m]$	0.5
pipe roughness	$k[mm]$	0.5
heat transfer coefficient	$c_h[W/(m^2K)]$	4
fluid composition	$x(CO_2, N_2, O_2)$	(0.95,0.03,0.02)
inlet pressure	pset [bar]	100
outlet norm.vol.flow, scen1	qset1 $[10^3 m^3/h]$	200
outlet norm.vol.flow, scen2	qset2 $[10^3 m^3/h]$	310

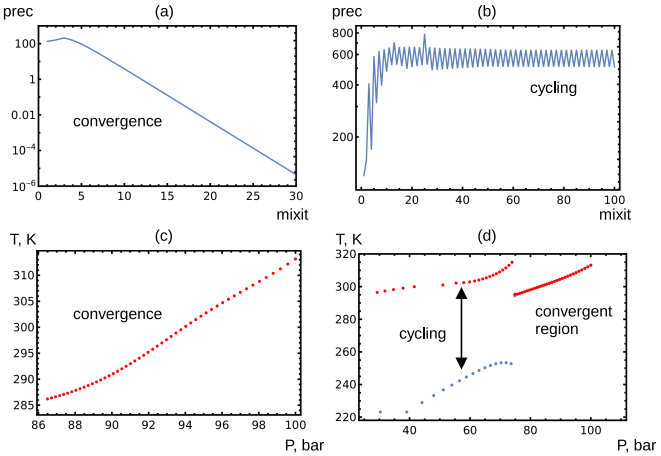

Fig. 3. (a),(c) – convergent iterations for scenario without phase transitions; (b),(d) – cycling iterations for scenario with phase transitions, red color - iteration 100, blue color – iteration 99.

scenarios use a mixture of 95% CO_2, 3% N_2, 2% O_2. The pipe is laid horizontally with $h = 0$.

Figure 3 shows the convergence characteristics for our test scenarios, left column for scen1, right column for scen2. The dimensionless precision parameter $prec = \max(res_i/norm_i)$ is defined as the maximum of the residuals of the equations divided by the normalizing value, for each equation its own. For the Kirchhoff equation of conservation of flow, the friction law in quadratic form, and the gas law expressed with respect to density, the normalization factors $norm = (1kg/s, 100bar^2, 1kg/m^3)$ are chosen, respectively. In our system, the equations and their normalizing factors can be freely configured by the user. For a purely 1-phase solution scen1 shown in Figure 3(a) and (c), the value of $prec$ decreases exponentially with the number of iterations and the solution procedure converges. For scen2, as seen in Figure 1(b) and (d), the procedure has cycling. In more detail, we see that there is a converging region for the 1-phase and a part of the 2-phase state, after which a temperature jump occurs, and oscillations are observed in the remaining pipe segment.

Along with the two main scenarios, we ran a number of additional simulations with small qset variations around the specified values. Simulations show stability of the effects, convergence in the 1-phase solution, and divergence in the 2-phase solution. The reason for this divergence is that EOS

 ISBN: 978-1-68558-073-5

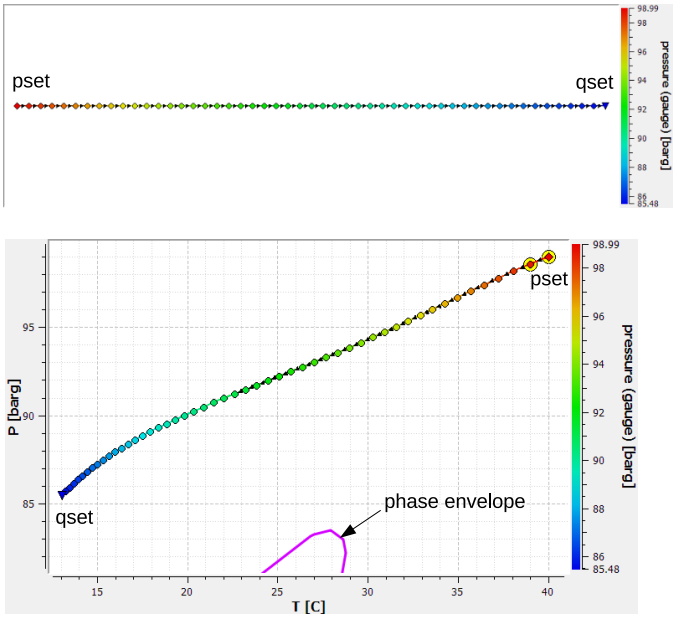

Fig. 4. Screenshot of MYNTS GUI for scenario without phase transitions.

and the enthalpy function receive large derivatives in the phase transition region. These functions are actually jump-like for a pure substance and formally continuous for a mixture, but at a low concentration of impurities, the derivatives are still large.

A prototype example of such instability is the logistic map: $x_{i+1} = rx_i(1-x_i)$, which characterizes the behavior of simple iterations near the root $x = 1-1/r$. When r rises from 1, and passes the value 3, the absolute value of the r.h.s. derivative of the logistic map equation exceeds 1, which is a critical value for the convergence of simple iterations. Below this value, the iterations converge. Above it, limit cycles appear, first with a multiplicity of 2, then they double, and finally the system goes to chaos.

Qualitatively, the same effects happen in our case. The stabilization algorithm used in principle helps to overcome such divergences, but for an ever higher derivative it becomes less and less effective. We are going to explore this problem in more detail in our future work. In order to overcome the divergence, we can try to adjust the weight parameter in the stabilizing algorithm. The dynamic solver behaves in much the same way as weighted relaxation with a low weight; with a decrease in the integration step, the stability of the integration also increases. As shown in Figure 1, high derivatives only occur for EOS in the form $\rho(T, P)$, changing variables to $P(T, \rho)$ could also be a solution of the problem.

At the same time, within the framework of the set technical task, it is required to consider only those scenarios in which there are no phase transitions and also there are no divergences associated with them. For such solutions, it is required to determine the proximity of the solution to the region of phase transitions. That can be done using the proximity-alarm algorithm described above.

Figure 4 shows the screenshots for scen1 solution in MYNTS GUI. At the top, there is the pipe geometry with the pressure profile shown in color. At the bottom, there is the solution on the (T, P)-plane, where a part of the phase envelope is also shown. The yellow disks show the proximity-alarm triggered in the given node for the values $dT = 1K$, $dP = 1bar$. The first 2 nodes near pset appear to be close to the spurious line on the phase diagram. The alarm in them can be canceled, because they are located top-right to the phase envelope, in the supercritical region. In general, this visual criterion is difficult to automate, since phase envelopes can have a more complex appearance than in the figures of this paper. Further, the figure shows how the solution trajectory passes at a safe distance from the phase envelope, providing the required CO_2 transport without phase transitions.

V. Conclusion

In this paper, we have considered a numerical simulation of the stationary process of CO_2 transport with impurities and phase transitions. We have developed the algorithms that allow to solve scenarios of CO_2 transport in the liquid or supercritical phase and to detect the approaching of the phase transition region. We have analyzed a convergence of the solution algorithms in connection with fast and abrupt changes of the equation of state and the enthalpy function in the region of phase transitions.

The performed numerical experiments show that the scenarios with a single CO_2 phase converge. For the obtained temperature and pressure profiles, a conservative algorithm for detecting the proximity of phase transitions can be applied, giving the solution to the technical problem posed. At the same time, divergences can occur in scenarios with phase transitions due to the abrupt change of thermodynamic parameters. Questions about the possible suppression of these divergences as well as improved detection of phase transitions are the subject of our further work.

Acknowledgments

The work has been supported by Fraunhofer research cluster CINES. We acknowledge support from Open Grid Europe GmbH in the development and testing of the software.

References

[1] M. Nimtz, M. Klatt, B. Wiese, M. Kühn, and H.-J. Kraautz, “Modelling of the CO2 process- and transport chain in CCS systems – Examination of transport and storage processes”, Chemie der Erde – Geochemistry, vol. 70, suppl. 3, 2010, pp. 185-192.

[2] S. Liljemark, K. Arvidsson, M. T. P. Mc Cann, H. Tummescheit, and S. Velut, “Dynamic simulation of a carbon dioxide transfer pipeline for analysis of normal operation and failure modes”, Energy Procedia, vol. 4, 2011, pp. 3040-3047.

[3] M. Chaczykowski and A. J. Osiadacz, “Dynamic simulation of pipelines containing dense phase/supercritical CO2-rich mixtures for carbon capture and storage”, International Journal of Greenhouse Gas Control, vol. 9, 2012, pp. 446-456.

[4] P. Aursand, M. Hammer, S. T. Munkejord, and Ø. Wilhelmsen, “Pipeline transport of CO2 mixtures: Models for transient simulation”, International Journal of Greenhouse Gas Control, vol. 15, 2013, pp. 174-185.

[5] L. Raimondi, “CO2 Transportation with Pipelines - Model Analysis for Steady, Dynamic and Relief Simulation”, Chemical Engineering Transactions, vol. 36, 2014, pp. 619-624.

 ISBN: 978-1-68558-073-5

[6] M. Drescher et al., “Towards a Thorough Validation of Simulation Tools for CO2 Pipeline Transport”, Energy Procedia, vol. 114, 2017, pp. 6730-6740.
[7] B. Chen, H. Guo, S. Bai, and S. Cao, “Optimization of process parameters for pipeline CO2 transportation with impurities”, IOP Conf. Series: Earth and Environmental Science, vol. 300, 2019, 022002.
[8] M. Vitali et al., “Risks and Safety of CO2 Transport via Pipeline: A Review of Risk Analysis and Modeling Approaches for Accidental Releases”, Energies, vol. 14, 2021, 4601.
[9] L. Raimondi, “CCS Technology - CO2 Transportation and Relief Simulation in the Critical Region for HSE Assessment”, Chemical Engineering Transactions, vol. 91, 2022, pp. 43-48.
[10] S. T. McCoy and E. S. Rubin, “An engineering-economic model of pipeline transport of CO2 with application to carbon capture and storage”, International Journal of Greenhouse Gas Control, vol. 2, 2008, pp. 219-229.
[11] X. Luo, M. Wang, E. Oko, and C. Okezue, “Simulation-based Techno-economic Evaluation for Optimal Design of CO2 Transport Pipeline Network”, Applied Energy, vol. 132, 2014, pp. 610-620.
[12] V. E. Onyebuchi, A. Kolios, D. P. Hanak, C. Biliyok, and V. Manovic, “A systematic review of key challenges of CO2 transport via pipelines”, Renewable and Sustainable Energy Reviews, vol. 81, part 2, 2018, pp. 2563-2583.
[13] H. Lu, X. Ma, K. Huang, L. Fu, and M. Azimi, “Carbon dioxide transport via pipelines: A systematic review”, Journal of Cleaner Production, vol. 266, 2020, 121994.
[14] T. Clees et al., “MYNTS: Multi-phYsics NeTwork Simulator”, in Proc. of SIMULTECH 2016, International Conference on Simulation and Modeling Methodologies, Technologies and Applications, pp. 179-186, SciTePress, 2016.
[15] T. Clees, I. Nikitin, and L. Nikitina, “Making Network Solvers Globally Convergent”, Advances in Intelligent Systems and Computing, vol. 676, 2018, pp. 140-153.
[16] A. Baldin, T. Clees, B. Klaassen, I. Nikitin, and L. Nikitina, “Topological Reduction of Stationary Network Problems: Example of Gas Transport”, International Journal On Advances in Systems and Measurements, vol. 13, 2020, pp. 83-93.
[17] A. Baldin et al., “Principal component analysis in gas transport simulation”, in Proc. of SIMULTECH 2022, International Conference on Simulation and Modeling Methodologies, Technologies and Applications, pp. 178-185, SciTePress, 2022.
[18] A. Baldin et al., “On Advanced Modeling of Compressors and Weighted Mix Iteration for Simulation of Gas Transport Networks”, Lecture Notes in Networks and Systems, vol. 601, pp. 138-152, 2023.
[19] J. Nikuradse, “Laws of flow in rough pipes”, NACA Technical Memorandum 1292, Washington, 1950.
[20] P. Hofer, “Error evaluation in calculation of pipelines”, GWF-Gas/Erdgas, vol. 114, no. 3, 1973, pp. 113-119 (in German).
[21] J. Mischner, H. G. Fasold, and K. Kadner, System-planning basics of gas supply, Oldenbourg Industrieverlag GmbH, 2011 (in German).
[22] O. Kunz and W. Wagner, “The GERG-2008 wide-range equation of state for natural gases and other mixtures: An expansion of GERG-2004”, J. Chem. Eng. Data, vol. 57, 2012, pp. 3032-3091.
[23] W. Wagner, Description of the Software Package for the Calculation of Thermodynamic Properties from the GERG-2008 Wide-Range Equation of State for Natural Gases and Similar Mixtures, Ruhr-Universität Bochum, 2022.
[24] ISO 20765-2: Natural gas – Calculation of thermodynamic properties – Part 2: Single-phase properties (gas, liquid, and dense fluid) for extended ranges of application, International Organization for Standardization, 2015.

 ISBN: 978-1-68558-073-5

Prehistorical Archaeology Discipline's Contextualisation Facts and Workflow Logic: Complements-Components Blueprints for the Creation of Efficient Coherent Multi-disciplinary Conceptual Knowledge-based Discovery

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU), Germany
Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF), Germany
Leibniz Universität Hannover, Germany
Email: ruckema@uni-muenster.de

***Abstract*—This paper presents the results of the methodological discovery and parallelisation of workflow logic of prehistorical archaeology discipline's contextualisation, based on coherent multi-disciplinary conceptual knowledge. The goal is the creation of efficient, flexible, and sustainable contextualisation workflows, also providing efficient parallelised frame conversion. Implementations and realisations are enabled by the latest versions of the prehistory-protohistory and archaeology conceptual knowledge reference implementation and the component reference implementations framework. The paper provides the results on archaeological / prehistorical facts, universal contexts, and logical and formal entities, factual, conceptual, and procedural complements, components, and results required for exemplary practical hard criteria and fact-based contextualisation by the disciplines and even for consequent creative historico-cultural exploitation. Future research will address the creation and further development of a conceptual knowledge reference implementation and a component reference framework for coherent multi-disciplinary conceptual contextualisation, enabling multi-disciplinary equal footing with contributions from all scientific disciplines for example for prehistorical archaeology knowledge integration, contextualisation and analysis with prehistorical and archaeological knowledge resources.**

Keywords–Prehistory and Archaeology; Discipline's Facts and Workflow Logic; Fact-based Contextualisation; Historico-cultural Interpretation; CKRI and CRI Framework.

I. Introduction

The need to implement workflows of many disciplines beyond 'manual operation' has been continuously increasing over the last decades. In practice, we can also see a strong motivation for sustainable Knowledge Resources (KR) creation and development and efficient employment of resources, e.g., with high performance computation and storage.

After reading, writing, and arithmetic are established and accepted as general competences, the capabilities of achieving efficient analysis and contextualisation solutions are becoming 'state-of-the-art' increasingly important personal competences in the sciences. Efficiently and sustainably organising and de-isolating knowledge complements, the results of research, within a discipline is at least as important as its short term analysis. These organisation processes cannot refer only to the data if they should be useful for reuse of knowledge and insight. Therefore, such organisation can in no ways be seen technically or being task of third parties without risk of losing the competence on fact-based methods, insight, and interpretation. Additionally, organisation of its knowledge complements is a core scientific matter of any discipline and closely associated with the methods employed, with the ongoing analysis processes, and with the further interpretation potential. In accordance with best practice, any scientists dealing with methodological workflows in a discipline, e.g., when applying a method, should know and practice themselves the way steps can be created, organised, and implemented, e.g., the algorithms, symbolic representation, and the structure and computation related characteristics. This practice is especially relevant with all knowledge complements or in other words the non-technical aspects of workflows in prehistorical archaeology. Scientific work, including state-of-the-art practices in archaeological disciplines and humanities, comprise of a number of essential principles, including further continuous employment of valuation methods for new factual knowledge and insights, re-contextualisation and resources development, and consequent fact-based contextualisation, analysis, and interpretation. When done properly, the tasks including contextualisation allow to practice equal footing with contributing scientific disciplines. Numerous surveys and studies were conducted for archaeological and prehistorical cases and multi-disciplinary contexts during the last decades, e.g., specific object groups' contextualisation [1] and discovery [2] and providing factual knowledge for interpretation, including historico-cultural contexts. The Prehistory and Archaeology Knowledge Archive (PAKA) is continuously collecting [3] new knowledge and insight. This research delivers the respective blueprints resulting from previously unpublished contexts and workflows and efficient workflow implementations proven sustainable over many years and widely reusable.

The rest of this paper is organised as follows. Section II presents the fundaments and state-of-the-art methodological implementations and realisations employed. Section III presents the results of the contexts and workflow logic for processes in prehistorical archaeology, factual/conceptual and procedural complements. Section IV delivers the discipline's results, efficiency results, and discussion for the presented contexts and workflow logic cases. Section V summarises lessons learned, conclusions, and future work.

II. Fundaments and Previous Work

Two major practical reference implementations were deployed for full implementations, realisations, and continuous further developments: the latest versions of the prehistory-protohistory and archaeology Conceptual Knowledge Reference Implementation (CKRI) [1] and the Component Reference Implementations (CRI) framework [4] for conceptual knowledge-based context integration, complements processing, and geoscientific visualisation. CKRI provides the universal knowledge framework, including multi-disciplinary contexts of natural sciences and humanities [5]. CRI provides the required component groups and components for the implementation

 ISBN: 978-1-68558-073-5

and realisation of all the procedural modules. The reference implementations are based on the fundamental methodology of knowledge complements [6], considering that many facets of knowledge, including prehistory, need to be continuously acquired and reviewed [7]. Creating contextualisation requires to coherently integrate multi-disciplinary knowledge and to enable symbolic representations. Realisations need to integrate a wide range of components as required from participating disciplines, e.g., for dynamical processing, geoprocessing, spatial contextualisation. Prehistoric object groups and contexts are taken from the latest edition of PAKA, which is in continuous development for more than three decades [8], and from The Natural Sciences KR (NatSciKR), all released by DIMF [3]. The PAKA and The NatSciKR support Factual, Conceptual, Procedural, Metacognitive, Structural (FCPMS) knowledge complements [9] and enable seamless coherent multi-disciplinary conceptual knowledge integration for workflow procedures. systematical and methodological approaches based on CKRI. CKRI references are illustrated for demonstration via the multi-lingual Universal Decimal Classification (UDC) summary [10] released by the UDC Consortium under Creative Commons license [11].

III. Discipline's Contexts and Workflow Logic

Prehistorical archaeology discipline's resulting contexts and workflow logic are often matter of multi-disciplinary long-term research, which requires universal context identification and assignment to contributing scientific disciplines.

A. Resulting Factual and Conceptual Complements Blueprint

The discipline's factual and conceptual knowledge complements and major logical and formal entities resulting from the long-term surveys and practical implementations are given in Table I. Employed resources are High Resolution (HR) Digital Elevation Model (DEM) data, e.g., (Space) Shuttle Radar Topography Mission (SRTM) data [12], updates [13], and further satellite data. Common DEM can be supplemented by local Light Detection And Ranging (LiDAR) data for special features and resolutions. DEM data for spatial contexts is used via Network Common Data Form (NetCDF) [14], developed by the University Corporation for Atmospheric Research (UCAR/Unidata), National Center for Atmospheric Research (NCAR). KR and complement implementations in contributing contexts and disciplines are PAKA and NatSciKR [3] accompanied by HR Digital Chart of the World (DCW) [15], and Global Self-consistent Hierarchical High-resolution Geography (GSHHG) [16]. The symbolic representation of the contextualisation can be done with a wide range of methods, algorithms, and available components, e.g., via LX Professional Scientific Content-Context-Suite (LX PSCC Suite) [17] deploying the Generic Mapping Tools (GMT) and integrated modules [18] for visualisation. The GMT suite application components are used for handling the spatial data, applying the related criteria, and for the visualisation. For sustainability we also consequently employ xyz files in GMT, e.g., Point of Interest (PoI) and Point of Discovery (PoD) contexts. Signatures and Colour Palettes (CPT) can also be flexibly integrated via GMT. Mostly all contexts and object groups are in continuous development, based on their structural implementations. Practically all contexts are dealt with employing the CKRI and its facets and operation facilities. Many properties of the contexts, e.g., chorological and chronological properties, can be addressed using international standards, e.g., for georeferencing and time. The consequent knowledge approach enables a wide range of workflow creation and analysis, in the scenarios discussed here ranging from fact-based contextualisation to consequent fact-based historico-cultural interpretation. The results allow even further consequent creative historico-cultural exploitation.

B. Resulting Procedural Complements Blueprint

The discipline's resulting procedural complements (the knowledge complements) and the corresponding workflow implementation resulting from the long-term surveys and practical implementations are given in Table II. The implementations are designed for end-user deployment by members of every responsible discipline dealing with their major logical and formal entities. The matrix shows context / object groups, required logical and formal workflow entities (major processing groups pre, main, post), examples of their symbolic representation, structure and procedure implementations. The table confirms that all contexts and object groups are in continuous further development, including the implementations of knowledge complements, e.g., factual, conceptual, procedural, and structural, which is a major achievement for scientific best practice and sustainability. The characteristics include the contexts addressed with CKRI and georeferencing, as well as the potential of mostly all contexts can be deployed in workflow parallelisation. The table especially lists excerpts of embarrassingly (E) and loosely (L) parallelisation features. The components of the workflow blueprint allow very high flexibility for fact-based methods and context integration of scientific, fact-based symbolic representation, e.g., the symbolic representation of archaeological, prehistorical contexts requires the employment of different geographic projections, e.g., geospherical orthographic, isometrical, and equal area. Projections can be flexibly implemented via GMT [19] and via PROJ [20]. Besides the implemented components we already named: The workflow allows processing usable for most disciplines, Area of Interest (AoI) calculations, regular expression patterns for context structures, e.g., via Perl Compatible Regular Expressions (PCRE) [21]. Attributations not applicable (n.a.) are marked accordingly. Workflow output, e.g., frames and visualisation can be created for many common structures, e.g., Joint Photographic Experts Group (JPG), Portable Network Graphics (PNG), and Portable Document Format (PDF), as well as Motion/Moving Pictures Expert Group, version 4 (MP4). Transformation can also be done for Keyhole Markup Language generation.

Multi-dimensional or sequences of view, e.g., focus dependent views for knowledge dimensional computation per object, are implemented via OpenMP [22] and specifications [23], e.g., . Job parallel procedures, e.g., knowledge objects and resources localities, are supported by respective modular solutions [24].

IV. Discipline's Results and Workflow Efficiency

A. Discipline's Workflow: Parallelisation and Results

Table III shows the scalability of the example workflow procedure for parallelised processing parts (pre, timing; main, parallelisation; post, batch) of the coherent multi-disciplinary conceptual knowledge. The results are referring to a scenario of a set of 1440 frames created in parallel for 4 k canvas size for a 60 s sequence with a rate of 24 FPS (Frames Per Second).

 ISBN: 978-1-68558-073-5

Table I. Prehistorical Archaeology Discipline's Contexts and Logical / Formal Entities: Resulting Factual / Conceptual Contextualisation Matrix Implemented for Major Complements and Components (Excerpt).

Context / Discipline / Object Group	*Logical / Formal Entities*	*Symb. Repr. (Example)*	*Structure Impl. (Example)*	*In Dev.*	*CKRI*	*Georef.*	*Parallelisation E*	*L*
Factual / Conceptual Domain (Focus Complements: FCPMS)								
Hybrid	(Spatial) structure	Signature / CPT	netCDF, GMT, LX PSCC	✓	✓	✓	✓	✓
		HR DEM	netCDF, GMT, LX PSCC	✓	✓	✓	✓	✓
		SRTM	netCDF, GMT, LX PSCC	✓	✓	✓	✓	✓
		LiDAR	netCDF, GMT, LX PSCC	✓	✓	✓	✓	✓
Point	Singular structure	Signature / Symbol	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Prehistorical archaeology		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Settlements		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Ritual places		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Notable objects		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Geophysics		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Impact craters		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Planetology		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Plate tectonics features		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Geology		NatSciKR, PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Mineral resources		NatSciKR, PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Pedology		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Soil characteristics		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Volcanology		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Volcanological features		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Speleology		NatSciKR, PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Caves		NatSciKR, PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Oceanography		NatSciKR, GSHHG	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Bathymetry features		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Hydrology		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Mobility, transport		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Pre-modern trackways		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Linguistics		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Open field names		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Geography		NatSciKR	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Humanities, administrative		NatSciKR, DCW	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
...	...	...	...	...	...	...	...	...
Line / Polygon	Linear structure	Signature	xyz, GMT, LXPSCC	✓	✓	✓	✓	✓
Prehistorical archaeology		PAKA	xyz, GMT, LXPSCC	✓	✓	✓	✓	✓
...	...	...	...	...	...	...	...	...
Polygon	Areal structure	Signature	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Prehistorical archaeology		PAKA	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
...	...	...	...	...	...	...	...	...
Bathymetry features		GSHHG, DEM	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓
Administrative features		DCW	xyz, GMT, LX PSCC	✓	✓	✓	✓	✓

The architecture choosen for this realisation is an efficient 36-core-based Central Processing Unit (CPU) (Intel Xeon), which is taking into account that we commonly use 36 cores for many basic global approaches, e.g., considering 360 degrees of a global model. Precondition for parallelisation is sufficient memory for parallel use of integrated resources. Considering the employed resources, especially SRTM / NetCDF and KR, 128 GB RAM (Random Access Memory) for 36 parallel processes is comfortable when data limits are cut to the limits required for the algorithms with the range of a few hundred kilometres area per object entity.

Table III. Scalability of Discipline's Workflow (Example Runs, Parallelised Processing KR and Context Resources).

Threads (Cores)	*Wall Time*				
	Pre, Timing	*Main, Parallel*	*Post, Batch*	$\Sigma Pre, Main, Post$	
1	1145 s	2581175 s	84972 s	2667292 s	$\approx 741\,h$
18	526 s	143668 s	4759 s	143668 s	$\approx 40\,h$
36	262 s	71833 s	2386 s	74481 s	$\approx 21\,h$

The parallel instances are allowed for 90 GB HDD (Hard Disk Drive) space and separate 50 GB SSD (Solid State Disk) space for highly volatile data of parallel instances. Wall and compute times, especially of multi-dimensional workflow results, can greatly be reduced from the integrated parallelisation, which makes the procedural solution highly scalable. The wall times for thread numbers confirm the high scalability when implementations of the workflow are using higher numbers of threads. Many practical workflows may contain some parts which cannot be reasonably parallelised. This is especially true for scientific tasks with a certain complexity. Anyhow, the percentage of non parallelised parts is very low with CKRI and the CRI framework. However, individual instances may show non-linear characteristics due to instance content and references, e.g., different satellite data, different data types, and different knowledge complements. For large sets, hundreds up to thousands of CPU cores were employed, so parallelised wall times per object can be very reasonably reduced from days to hours or even minutes, e.g., for warning and tracking systems.

The following results from the above discipline workflow show an excerpt of eight frames from a large frame sequence for calculated Areas of Interest (AoI) contexts in top views (Figure 1). Ellipsoid is World Geodetic System 84 (WGS-84). Projection for frames is Lambert Azimuthal

Table II. Blueprint of Prehistorical Archaeology Discipline's Workflow Logic: Resulting Procedural Contextualisation Matrix, From FC (Table I), implemented for major complements and components, including parallel frame conversion (excerpt).

Context / Discipline / Object Group	*Logical / Formal Entities*	*Symb. Repr. (Example)*	*Struc. / Proc. Impl. (Example Complement / Environment)*	*In Dev.*	*CKRI*	*Georef.*	*Parallel. E*	*Parallel. L*
Procedural Domain (Focus Complements: FCPMS)								
Selection, preparation (KR)	Pre-processing	Pre-routines	CKRI, PCRE, … / LX PSCC	✓	✓	(✓)	(✓)	(✓)
Context resources	Pre-processing	Pre-routines	netCDF, CKRI, PCRE, … / LX PSCC	✓	✓	(✓)	(✓)	(✓)
Sequence	Pre-proc., timing structure	Parameter	[FCPMS] / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Procedure modules	Main processing	Main-routines	[FCPMS)], …	✓	✓	(✓)	(✓)	(✓)
Contextualisation Scenario	Integration	Hybrid	… / GMT, LX PSCC …	✓	✓	(✓)	(✓)	(✓)
Observer path	Path / project	Line	xyz / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Observer track	Track / project	Line	xyz / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
AoI	Selection, cut	Area	netCDF, xyz / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Sampling	Resampling	[Raster]	netCDF / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Canvas mapping	Basemap	[Mapping]	netCDF / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Gridding	Grid operations, …	[Grid]	netCDF / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Illumination	Height	Singular	netCDF / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Math operations	Calculation	[Algorithm]	… / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Triangulation	Calculation	[Algorithm]	… / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Regression	Calculation	[Algorithm]	… / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Colour	Colourisation	[Sequence]	CPT,/ GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Filtering	Selection, select	[Decimation]	… / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Movie module	Iteration, …	Parameter	… / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Proc. of knowledge compl.	Calculation	[Algorithm]	CKRI, PCRE, … / LX PSCC	✓	✓	(✓)	(✓)	(✓)
Spatial proc. of preh. ctxts.	Selection, calculation	[Algorithm]	CKRI, PCRE / GMT, …, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Events	Symbolic, functional	Symb. repr.	CKRI, PCRE / GMT, …, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Arbitary symbols	Symbolic, functional context	Symb. repr.	[vector graphics] / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Degenerated ellipses	Azimuthal	Area	xyz / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Range	Azimuthal	Area	xyz / GMT, LX PSCC	✓	✓	(✓)	(✓)	(✓)
Projection	Geospherical, orthographic	[Algorithm]	… / GMT, PROJ, LX PSCC	✓	✓	(✓)	(✓)	(✓)
…	…	…	…	…	…	…	…	…
Resource usage (main proc.)	Main proc., parallelisation	Frame, view	… / OpenMP, GNU parallel, LX PSCC	✓	✓	n.a.	✓	✓
	On-scratch processing	Various	… / OpenMP, GNU parallel, LX PSCC	✓	✓	n.a.	✓	✓
	Model reduction frame, anim.	Various	… / GMT, (LX PSCC)	✓	✓	n.a.	✓	✓
	Live frame control	Various	JPG, PNG, PDF / (LX PSCC)	✓	✓	n.a.	✓	✓
Transform., symbolic repr.	Post-processing, batch	Post-routines	Scales, KML, … / LX PSCC	✓	✓	n.a.	(✓)	(✓)
Visualisation, analysis	Post-processing, interactive	Image, Video	PNG, MP4, … / LX PSCC	✓	✓	n.a.	(✓)	(✓)

Equal Area. The resolution is drastically reduced for use in this publication. Generated representations include integrated CKRI references, projection of topographic and bathymetric results, and further knowledge for respective areas, based on the coherent conceptual knowledge. The frame sequence of symbolic representations enable to contextualise named factual data (CKRI: UDC:551.2…,551.21,550.3,(23);"62" and UDC:167/168…;51… referring to CKRI: UDC:711.…,692,903,902 for 150 km radii) [1].

Major multi-disciplinary results are the shown insights regarding the details of prehistoric settlement infrastructures / Holocene maars for which we find larger numbers of prehistorical settlements were set and used in the volcanic regions Eifel (DE) and Auvergne (FR) areas than in the other areas, all of which can be precisely assigned and further contextualised. Ongoing analysis and discussion of the multitude of resulting historico-cultural meanings will be given in later publications.

B. Discipline's Workflow: Frame Conversion Benchmark

The number of parallel cores used for the making of individual frames can be efficiently controlled. The parallel processing itself does not depend on OpenMP. Table IV gives the dimensions of canvas sizes for an excerpt of common formats, represented by pixel (p) scales. The given formats High Definition (HD), Ultra High Definition (UHD), Ultra Extended Graphics Array (UXGA), Extended Graphics Array (XGA), and Super XGA Plus (SXGA+) are commonly used in resources development and practical high resolution workflows.

Table IV. Canvas Sizes and Formats Used in Practical Case Scenario Implementations (Table II, excerpt).

Canvas Size (p)	*Format*	
Format 16:9 (e.g., 24×13.5 cm)		
7680 × 4320	UHD-2	8 k
3840 × 2160	UHD	4 k
1920 × 1080	HD	
Format 4:3 (e.g., 24×18 cm)		
1600 × 1200	UXGA	
1400 × 1050	SXGA+	
1024 × 768	XGA	

The conversion of frames can be done in parallel using GraphicsMagick [25]. GraphicsMagick includes Gnu's Not Unix (GNU) libgomp [26] of the GNU Offloading and Multi-Processing Project (GOMP). Table V shows the frame conversion benchmark results for different canvas sizes as used in the parallel implementations of practical case scenarios. The results compare number of threads, iterations, user time, total time, iterations per second, iterations per CPU, speedup, and Karp-Flatt result. The conversion uses a common 128 × 128

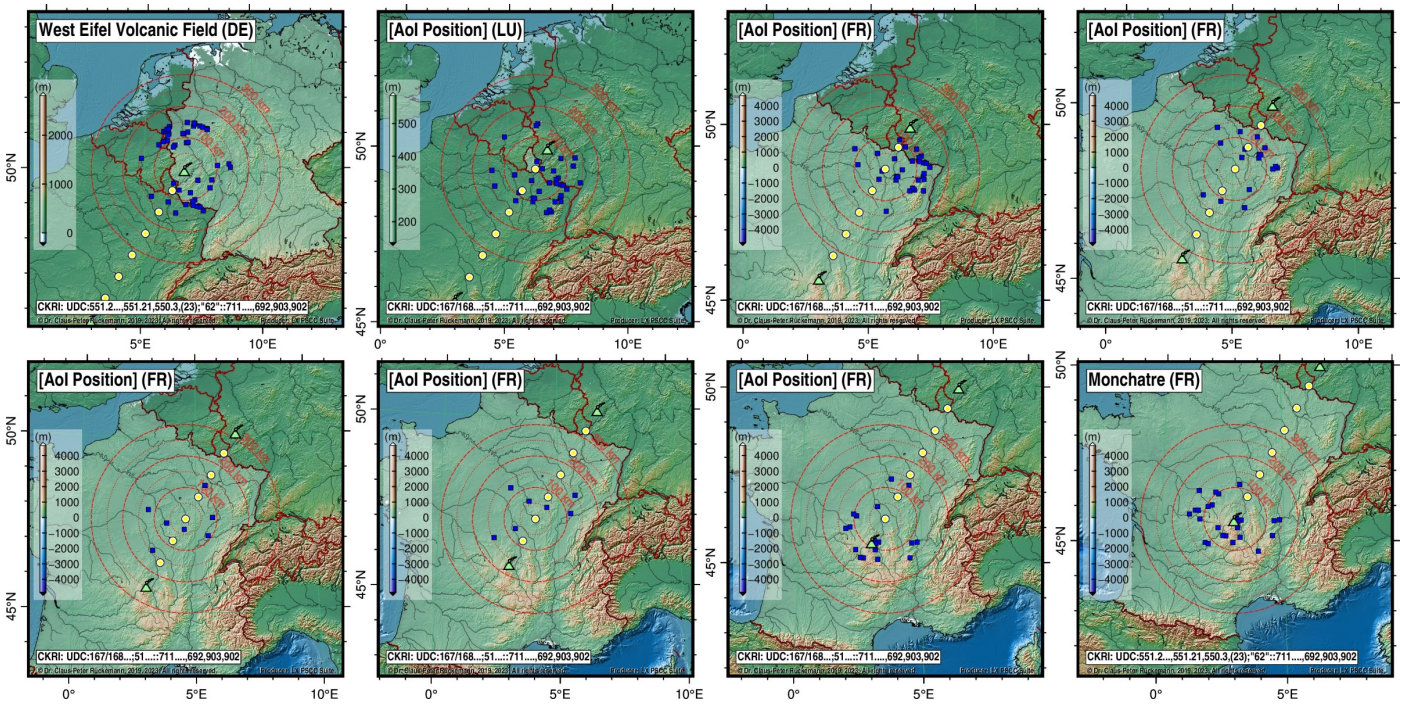

Figure 1. Discipline workflow results of prehistoric settlement infrastructures in factual and historico-cultural chorological and chronological contexts with a volcanological features group (maars, Holocene-historical) and satellite data based on the coherent conceptual knowledge integration and discovery (excerpt).

Table V. Frame conversion benchmark results for canvas sizes used in parallel implementations of practical multi-disciplinary case scenario workflows of archaeological / prehistorical contextualisation (Table II, excerpt).

Threads	*Iterations*	*User Time*	*Elapsed Time*	*Iterations / s*	*Iterations / CPU*	*Speedup*	*Karp-Flatt*
			7680×4320 (UHD-2)				
1	2	10.56 s	10.563899 s	0.189	0.189	1.00	1.000
18	26	138.39 s	10.094287 s	2.576	0.188	13.60	0.019
36	41	220.25 s	10.067891 s	4.072	0.186	21.51	0.019
			3840×2160 (UHD)				
1	8	10.62 s	10.625725 s	0.753	0.753	1.00	1.000
18	104	141.39 s	10.075150 s	10.322	0.736	13.71	0.018
36	166	233.24 s	10.056526 s	16.507	0.712	21.92	0.018
			1024×768 (XGA)				
1	82	10.08 s	10.078310 s	8.136	8.135	1.00	1.000
18	1191	179.99 s	10.007169 s	119.015	6.617	14.63	0.014
36	1856	358.52 s	10.001333 s	185.575	5.177	22.81	0.017

granite texture pattern iteration for standardisation. The benchmark uses the Karp-Flatt metric [27], which is a measure of code parallelisation in parallel processor systems. The resulting implementation is very scalable and can use practical workflow parallelisations from small canvas sizes up to defined sizes even beyond UHD-2. Sizes of UHD are very appropriate for many HR scenarios with commonly available technical infrastructures while being relatively efficient with resources.

V. Conclusion

This paper presented the results achieved for the methodological discovery and parallelisation of workflow logic of contextualisation in prehistorical archaeology, based on coherent multi-disciplinary conceptual knowledge. The implemented workflows employed the latest versions of the prehistory-protohistory and archaeology CKRI and the CRI framework. The implemented and realised contextualisation workflows proved efficient, flexible, and sustainable. The presented contexts, entities, and workflow implementations provide solid fact-based fundaments for contextualisation and consequent fact-based historico-cultural interpretation, procedures, which should be deployed by members of the contributing disciplines.

Ongoing, the reference implementations and procedures will be extended for generation of symbolic representation for advanced multi-dimensional knowledge models. Future research will address the creation and further development of the prehistory-protohistory and archaeology CKRI and the CRI framework for coherent multi-disciplinary conceptual contextualisation, enabling multi-disciplinary equal footing with contributions from all scientific disciplines, e.g., natural sciences, soil science, and linguistics, especially supporting new advanced methods in prehistorical archaeology for knowledge integration, contextualisation, and analysis.

 ISBN: 978-1-68558-073-5

Acknowledgements

This ongoing research is supported by scientific organisations and individuals. We are grateful to the "Knowledge in Motion" (KiM) long-term project, Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF), for partially funding this research, implementation, case studies, and publication under grants D2022F1P05308, D2020F1P05228, and D2022F2P05355 and to its senior scientific members and members of the permanent commission of the science council, especially to Dr. Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek (GWLB) Hannover, to Dipl.-Biol. Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, for fruitful discussion, inspiration, and practical multi-disciplinary contextualisation and case studies. We are grateful to Dipl.-Geogr. Burkhard Hentzschel and Dipl.-Ing. Eckhard Dunkhorst, Minden, Germany, for prolific discussion and exchange of practical spatial, UAV, context scenarios, and measurement support. We are grateful to Dipl.-Ing. Hans-Günther Müller, Göttingen, Germany, for providing specialised, manufactured high end computation, storage, and visualisation solutions. We are grateful to The Science and High Performance Supercomputing Centre (SHPSC) for long-term support. / DIMF-PIID-DF98_007; URL: https://scienceparagon.de/cpr .

References

[1] C.-P. Rückemann, "Faceting the Holocene-prehistoric Inventory of Volcanological Features Groups Towards Sustainable Multi-disciplinary Context Integration in Prehistory and Archaeology Based on the Methodology of Coherent Conceptual Knowledge Contextualisation," International Journal on Advances in Intelligent Systems, vol. 15, no. 3&4, 2022, pp. 115–129, ISSN: 1942-2679, LCCN: 2008212456 (Library of Congress), URL: http://www.iariajournals.org/intelligent_systems [accessed: 2023-06-18].

[2] C.-P. Rückemann, "Advanced Contextualisation Reference Implementation Frameworks in Practice: Coherent Multi-disciplinary Conceptual Knowledge-Spatial Context Discovery Results from the Holocene-prehistoric Volcanological Features and Archaeological Settlement Infrastructure Surveys," in Proceedings of The Fifteenth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2023), April 24 – 28, 2023, Venice, Italy. XPS Press, Wilmington, Delaware, USA, 2023, pp. 98–103, ISSN: 2308-393X, ISBN-13: 978-1-68558-079-7, URL: https://www.thinkmind.org/articles/geoprocessing_2023_2_150_30100.pdf [accessed: 2023-06-18].

[3] "The Prehistory and Archaeology Knowledge Archive (PAKA) license," 2023, (release 2023), Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF): All rights reserved. Rights retain to the contributing creators.

[4] C.-P. Rückemann, "Component Framework Implementation and Realisation for Development and Deployment of a Coherent Multi-disciplinary Conceptual Knowledge-based Holocene-prehistoric Inventory of Volcanological Features Groups and Faceting," International Journal on Advances in Intelligent Systems, vol. 15, no. 3&4, 2022, pp. 103–114, ISSN: 1942-2679, LCCN: 2008212456 (Library of Congress), URL: http://www.iariajournals.org/intelligent_systems [accessed: 2023-06-18].

[5] C.-P. Rückemann, "The Information Science Paragon: Allow Knowledge to Prevail, from Prehistory to Future – Approaches to Universality, Consistency, and Long-term Sustainability," The International Journal "Information Models and Analyses" (IJ IMA), vol. 9, no. 3, 2020, pp. 203–226, Markov, K. (ed.), ISSN: 1314-6416 (print), Submitted accepted article: November 18, 2020, Publication date: August 17, 2021, URL: http://www.foibg.com/ijima/vol09/ijima09-03-p01.pdf [accessed: 2023-06-18].

[6] C.-P. Rückemann, "From Knowledge and Meaning Towards Knowledge Pattern Matching: Processing and Developing Knowledge Objects Targeting Geoscientific Context and Georeferencing," in Proc. GEOProcessing 2020, November 21–25, 2020, Valencia, Spain, 2020, pp. 36–41, ISSN: 2308-393X, ISBN-13: 978-1-61208-762-7.

[7] R. Gleser, Zu den erkenntnistheoretischen Grundlagen der Prähistorischen Archäologie. Leiden, 2021, (title in English: On the Epistemological Fundaments of Prehistorical Archaeology), in: M. Renger, S.-M. Rothermund, S. Schreiber, and A. Veling (Eds.), Theorie, Archäologie, Reflexion. Kontroversen und Ansätze im deutschsprachigen Diskurs, (in print).

[8] C.-P. Rückemann, "Information Science and Inter-disciplinary Long-term Strategies – Key to Insight, Consistency, and Sustainability: Conceptual Knowledge Reference Methodology Spanning Prehistory, Archaeology, Natural Sciences, and Humanities," International Tutorial, DataSys Congress 2020, Sept. 27 – Oct. 1, 2020, Lisbon, Portugal, 2020, URL: http://www.iaria.org/conferences2020/ProgramINFOCOMP20.html [accessed: 2023-06-18].

[9] C.-P. Rückemann, "Coherent Knowledge Solutions From Prehistory to Future – Towards Coherent Multi-disciplinary Knowledge Reference Implementation Blueprints for Industrial Learning: Insight from Consistent Coherent Conceptual Integration of Prehistory, Archaeology, Natural Sciences, and Humanities," ML4I – Machine Learning for Industry Forum 2021; High-Performance Computing Innovation Center (HPCIC) and Data Science Institute (DSI), at Lawrence Livermore National Laboratory (LLNL), Aug. 10–12, 2021, Livermore, U.S.A., (Invited Speech), URL: http://www.llnl.gov [accessed: 2023-06-18], 2021.

[10] "Multilingual Universal Decimal Classification Summary," 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php [accessed: 2023-06-18].

[11] "Creative Commons Attribution Share Alike 3.0 license," 2012, URL: http://creativecommons.org/licenses/by-sa/3.0/ [accessed: 2023-06-18], (first release 2009, subsequent update 2012).

[12] C. L. Olson, J. J. Becker, and D. T. Sandwell, "SRTM15_PLUS: Data fusion of Shuttle Radar Topography Mission (SRTM) land topography with measured and estimated seafloor topography," (NCEI Accession 0150537), National Centers for Environmental Information (NCEI), NOAA, 2016.

[13] B. Tozer, D. T. Sandwell, W. H. F. Smith, C. Olson, J. R. Beale, and P. Wessel, "Global Bathymetry and Topography at 15 Arc Sec: SRTM15+," Earth and Space Science, vol. 6, no. 10, Oct. 2019, pp. 1847–1864, ISSN: 2333-5084, DOI: 10.1029/2019EA000658.

[14] "Network Common Data Form (NetCDF)," 2023, DOI: 10.5065/D6H70CW6, URL: http://www.unidata.ucar.edu/software/netcdf/ [accessed: 2023-06-18].

[15] P. Wessel, "DCW for GMT 6 or later," 2022, URL: http://www.soest.hawaii.edu/pwessel/dcw/ [accessed: 2023-06-18].

[16] P. Wessel, "GSHHG," 2017, URL: http://www.soest.hawaii.edu/pwessel/gshhg/ [accessed: 2023-06-18].

[17] "LX Professional Scientific Content-Context-Suite (LX PSCC Suite) license," 2023, (release 2023), Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF): All rights reserved. Rights retain to the contributing creators.

[18] P. Wessel, W. H. F. Smith, R. Scharroo, J. Luis, and F. Wobbe, "The Generic Mapping Tools (GMT)," 2020, URL: http://www.generic-mapping-tools.org/ [accessed: 2023-06-18].

[19] "GMT - Generic Mapping Tools," 2023, URL: http://gmt.soest.hawaii.edu/ [accessed: 2023-06-18].

[20] PROJ contributors, PROJ Coordinate Transformation Software Library, Open Source Geospatial Foundation, 2023, URL: https://proj.org/ [accessed: 2023-06-18].

[21] "Perl Compatible Regular Expressions (PCRE)," 2023, URL: https://www.pcre.org/ [accessed: 2023-06-18].

[22] L. Dagum and R. Menon, "OpenMP: an industry standard API for shared-memory programming," Computational Science & Engineering, (IEEE), vol. 5, no. 1, 1998, pp. 46–55.

[23] OpenMP Architecture Review Board, "OpenMP API 5.1 Specification," Nov. 2020, URL: https://www.openmp.org/wp-content/uploads/OpenMP-API-Specification-5-1.pdf [accessed: 2023-06-18].

[24] "GNU Parallel," 2023, URL: http://www.gnu.org/s/parallel [accessed: 2023-06-18].

[25] "GraphicsMagick Image Processing System," 2023, URL: http://www.graphicsmagick.org/ [accessed: 2023-06-18].

[26] "GNU libgomp – GNU Offloading and Multi-Processing Project (GOMP)," 2023, URL: https://gcc.gnu.org/projects/gomp/ [accessed: 2023-06-18].

[27] A. H. Karp and H. P. Flatt, "Measuring Parallel Processor Performance," in Communications of the ACM, 33, vol. 5, 1990, pp. 539–543, DOI: 10.1145/78607.78614.

 ISBN: 978-1-68558-073-5

Governance-Centric Paradigm: Overcoming the Information Gap between Users and Systems by Enforcing Data Management Plans on HPC-Systems

Hendrik Nolte
GWDG
Göttingen, Germany
hendrik.nolte@gwdg.de

Julian Kunkel
University of Göttingen
Göttingen, Germany
julian.kunkel@gwdg.de

***Abstract*—Along with the increase in available compute power of High-Performance Computing (HPC) systems and the success of novel data-driven methods, the amount of data processed and the user groups increase as well. This gave rise to two big challenges: The traditional interaction scheme of users with modern HPC systems becomes more and more unsuited to deal with large data sets and many independent tasks working on these data sets. This highly manual way can quickly lead to unreproducible results and data loss due to missing backups since it is stored fragmented on multiple storage tiers. Similarly, domain-specific data management systems have been established to ease the burden of data and process management of particularly inexperienced users. These systems, however, only offer a very rigid, and tool-specific interaction scheme. This resulted in a gap between these two user groups, which even hinders large-scale cooperations across different domains. In this paper, we introduce the Governance-Centric interaction paradigm, a novel, and holistic concept which allows us to enforce data management plans to bridge this gap.**

***Index Terms*—data management, high-performance computing, provenance, reproducibility, IO performance.**

I. Introduction

Data-driven methods gained a lot of momentum in recent years and their success lead to adoptions in a wide variety of scientific domains. Also, many sciences are data-intense such as climate/weather. These data-driven projects have a few things in common. First, they require large data sets to be able to derive results with good statistics. Second, these large data sets often have large storage requirements due to their size. Third, these data sets often consist of millions of small files which are typically organized in storage within a few flat namespaces. However, these methods are not only data-intensive, but processing all of these data sets in a reasonable amount of time requires large compute resources. Therefore, researchers have started to utilize High-Performance Computing (HPC) clusters to serve those tasks.

There are various challenges when handling and processing this data. **1) Performance**: these iterative procedures on these small files lead to heavy loads on the storage system, which can overload, particularly the metadata servers, and can lead to large performance degradation due to storage bottlenecks. **2) Data management**: fulfilling the FAIR principles [1], i.e., making data findable, accessible, interoperable, and reusable is challenging. For instance in order to make data findable a naming scheme is mandatory - coming up with a naming scheme for files/objects created and then actually following it. Sharing data with other researchers often comes as an afterthought. It is a reasonable assumption, that most projects do neither strictly follow the FAIR principles nor their Data Management Plan (DMP) if there was one defined at the beginning of a project. It can be expected that this issue will only be exacerbated by the increasing complexity and heterogeneity of the employed storage systems in the compute continuum. **3) Integration of compute and data handling**: Computing on the HPC system feels a bit archaic. Users have to manually define many system settings, for instance, file names define what storage to use. Meanwhile, the complexity of the tiered storage systems in modern HPC systems has drastically increased. There exists no way to define and enforce data governance, which is homogeneously applicable across all of the disparate storage tiers. **4) Reproducibility**: Being able to understand the lineage of data and how to reproduce certain outputs is important for trust in the scientific results. However, as execution on HPC systems are usually scripts that are invoked manually, on binaries created specifically for the given supercomputer, it is tricky to reproduce results.

We do not list usability as an independent key challenge on its own explicitly, as it is primarily a function of data management and integration.

One promising solution for 2) would be to use new or established Data Management Systems (DMS) in these data-intensive projects. These systems could provide a unified namespace across a tiered and distributed storage architecture by offering a single point for data copies to reside. However, this requires tight integration of the tools in HPC systems. Additionally, since there will always be a gap between a remote DMS and a HPC system, ensuring reliable information within the DMS which originates from a HPC system is an unsolved problem. In this article, we systematically discuss an overarching next-generation concept to integrate DMS into data-intensive HPC workflows and create a user-friendly unified infrastructure for compute and storage that we believe

 ISBN: 978-1-68558-073-5

HPC should be. This includes the following contributions:

- the current user interaction paradigms with HPC systems are discussed and classified
- the involved components are discussed
- the prevailing gap between different user groups is identified using a layer model
- the novel governance-centric paradigm is proposed

The remainder of this paper is structured as follows: in Section II, the related work is discussed, leading to the discussion of the prevailing interaction paradigms in Section III. Based on this, the novel governance-centric paradigm is presented in Section IV, which is followed by the conclusion in Section V.

II. Context and Related Work

In the following, we describe the state-of-the-art in our four challenges.

A. Performance

Usually, HPC clusters provide at least two parallel file systems, providing file access via Portable Operating System Interface I/O (POSIX-IO) semantics, the relaxed Message Passing Interface I/O (MPI-IO) [2] semantics or the close-to-open semantics used by the Network File System (NFS), to name just a few. All of these semantics require dedicated metadata servers to empower parallel file systems, like Lustre [3], or the General Parallel File System (GPFS) [4]. These metadata servers handle all metadata operations using special data structures called inodes to handle these metadata operations. If an inode represents a folder on such a filesystem, it contains a list of all inodes located in this folder. Depending on the actual operation, which should be done, it might be necessary to also read additional information from each inode within a folder, for example, the file permissions, or the ownership. The cost for these metadata operations scales linearly with the number of files stored within a single folder. However, if the list of inodes stored in a single inode becomes too long, indirect inodes have to be used. This behavior can be triggered if those inode lists are inlined within a small data block within the inode itself. This is typically done to avoid lookups on the storage servers holding the actual data of a file, which would otherwise increase the latency of such a metadata operation drastically. These indirect inodes, potentially even consisting of multiple layers, lead to an even worse performance degradation. Therefore, having too many files within a single folder has a huge performance penalty. However, this can often be observed in machine learning projects, e.g., if there a tens of thousands of small images in a single folder whose name encodes the particular target, e.g., a folder called *cats* containing many small images of cats.

Although there is a varying amount of overhead necessary in the different semantics and filesystems, they all share the problem of bottlenecking when exposed to this described small file IO. Current mitigation strategies consist of either providing a multi-tier storage system, where each tier is optimized to handle certain workloads, or meeting a specific cost-to-capacity ratio. This option leads to increased complexity and requires the users to manually move and stage data to the correct tiers to achieve optimal performance while ensuring that cold data is not piling up on fast and expensive storage, which is not backed up. In addition, novel storage concepts, like object storages, are being integrated into HPC cluster, which supports flat namespaces by design. These are already common in cloud environments, with prominent standards like Amazon's S3, or Openstack Swift. However, their REST-based interfaces entail additional overhead, both, on the communication layer, and also on the application layer, since the file handling drastically differs from the well-established POSIX-IO compatible file systems.

B. Data Management

There are already some established tools that try to abstract and simplify the interaction with complex and heterogeneous HPC systems. One of these tools is *VIKING* [5] which is used specifically for molecular dynamics simulations and provides a user-friendly web interface to run *NAMD* [6] or *Gromacs* [7] among others.

Similarly, *XNAT* [8] is a DMS specifically built for neuroimaging data. It allows the organization of data within a hierarchical structure consisting of projects, subjects, and experiments. Analysis can be done and solely controlled from within the web interface using *Docker* container on a dedicated *Docker-Swarm* or *Kubernetes* cluster.

However, these tools only provide a very restrictive compute model or require a lot of manual steps by the users to allow for larger flexibility.

C. Integration of Compute and Data Handling

In order to integrate HPC systems into XNAT, DAX [9] was developed. It supports the execution of preconfigured tasks, called *Spiders* on a batch system, but does not support direct access to the data or the tasks by the users on the HPC system.

In cloud environments, the integration of compute and data is well established and is commonly implemented in any Infrastructure-as-a-Service offering. The entire approach can even be found on Hadoop systems, where tasks and data were transparently integrated either by batch jobs accessing data via the Hadoop distributed file system [10] or by interactive tasks using the YARN [11] resource manager.

Today, similar approaches can be seen with Jupyter-Hub deployments in HPC centers [12], which, however, do not lift the burden of managing tiered storage systems of the user.

Therefore, data and compute are currently considered separate parameters, a user has to individually and manually manage and optimize. Particularly, the integration into a user's overarching experiment is lacking, since no globally defined data governance can be homogeneously enforced across all storage tiers.

D. Reproducibility

Since mostly data-parallelism is assumed in these data-intensive workloads, the question of reproducibility of HPC jobs is reduced to the problem of provenance auditing while

 ISBN: 978-1-68558-073-5

the deterministic execution of a HPC job is completely neglected. There is related work for provenance auditing on HPC systems. Two often used approaches are either monitoring system calls like *PASS* [13] to create audit trails. Following a similar approach, *LPS* [14] has drastically reduced the runtime overhead, but is not completely transparent due to the use of a dedicated *Library Wrapper*. *ReproZip* [15] also continues the idea of audit trails of system calls to automatically build packages to re-run an experiment.

A different way for lineage recording is provided by *Data Pallets* [16]. Here, all processes run within containers where all write access to the storage devices is intercepted and transparently redirected to data containers. Hereby, all data containers are automatically annotated with reliable provenance recordings.

Although provenance tools for HPC systems have evolved, they currently lack integration of containers, and awareness of DMS, i.e. if a problem data management is done, input data can be linked, and must not be archived along each and every single execution. In addition, they commonly lack the overall awareness of a workflow. Therefore, there is a gap between these node-local and hardware-close tools and the higher level interaction a user wants to have with a HPC system.

III. Overarching Concept of Integrating Data Management Tools into HPC Workflows

There are a number of challenges that a user typically faces when accessing an HPC system. For instance, a HPC system should reduce the complex and heterogeneous storage architecture to a unified namespace to offer users a quick and easy overview of their data. In addition, the data management system (DMS) should optimize the usage of a tiered storage system to provide maximal performance during compute and uses durable and low-cost storage for cold data. It should also adhere to the FAIR principles and perform transparent provenance auditing to ensure reproducibility. All data within a flat namespace should be searchable by domain-specific, semantic metadata, ideally even with respect to globally enforced policies for data access.

First, we identify and discuss the abstract components and their features when interacting with a storage and compute infrastructure, such as an HPC system. Then we describe the status quo as an archetype for the standard interaction paradigm and our envisioned user-friendly and data-centric flow.

A. Components

The components necessary when handing data and compute are as follows:

- Resources - these are raw storage, compute and network infrastructures such as compute nodes and object/file systems and their interconnect. They come with their own specification, i.e., what resource they actually provide and their characteristics.
- Resource management (compute) - this layer manages the usage of the resources by assigning compute jobs to available compute resources satisfying the requirements for the respective (parallel) jobs.
- Resource management (storage) - this abstract concept defines where to store certain data and provides the respective space on a storage system.
- Job specification - defines the scope of a compute job together with it's requirements and specification such that it can be executed.
- Program - a code that can be executed on the compute infrastructure, e.g., binary program or script.
- Software landscape - the ecosystem and environment provided by the platform that allows to prepare programs on the system.
- Workflow specification - defines how to execute jobs in order to achieve the overall data-processing goal.
- Data management plan - defines for any data inputs and data products the policies, data handling and such to enable the FAIR principles while the data sovereignity of the user is preserved.
- User interface - allows the user to interact with the system, e.g., to manage and interact with some or all of the above components and to upload/download data.
- Client - the computer system of the user, where the user interface(s) are accessed.

The way, how one can fulfill the previously specified requirements and implement the components depends on the way a user wants to interact with it and the system providing these capabilities, and the data flow involved. For example, either, a user connects to the HPC system, and uses it as the central contact point, or the interface of the DMS is used to manage the data processing on the HPC system remotely.

B. Interaction Paradigms

On the most extreme scale, one can argue that there are three different kinds of users. For instance, tech-affine people who want to natively work on the HPC in a traditional command-line approach, users that utilize state-of-the-art compute-centric tools, or those, who ideally only want to work with the interface of their domain-specific DMS.

1) Traditional Paradigm: In the traditional approach, the user interface is a shell (such as bash) on a login node of the cluster and the client is an SSH-enabled program that the user runs on their Desktop/Laptop. There exists no data management plan, the user thinks about how to manage data, therewith, manually performs the resource management for storage, identifies how to map output data to files (influenced by the applications) and directory structures, and utilizes the available parallel file systems. Also, the user manually prepares programs s/he wants to use by downloading the necessary codes on the machine and ensuring it works with the system architecture and software environment that is deployed on the HPC system. The wider software landscape on the HPC system was prepared by data center staff but libraries can be extended by the users in order to create meaningful programs. In most cases, workflows are not explicitly specified but manually invoked. The resource management of the

 ISBN: 978-1-68558-073-5

compute resources is provided by tools such as Slurm. The job specifications are (bash) scripts that are invoked - they define the compute requirements. Such scripts are submitted to Slurm which decides how to map and schedule them on the available compute resources. These steps are basically manually set up, requiring a scientist to think about how the experiment should be conducted and then documented (if at all) in a lab notebook or scripts that do some of the work. Potentially, workflow tools such as Snakemake are utilized to specify dependencies between tasks and to automate dependencies between tasks. This is not only error-prone, but any change to the environment requires the user to modify the experimental setup and perform the steps again. We consider this the most typical interaction with the HPC system archaic.

2) Compute-Centric Paradigm: In a Compute-centric approach, a user would connect to the HPC frontend as usual. In the simplest form, a user would delegate the job of maintaining a data catalog and staging the selected input data to a DMS tool. This workflow is depicted in Figure 1.

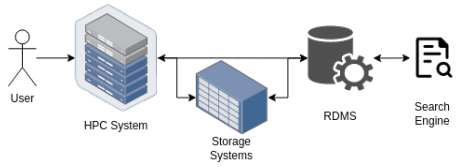

Fig. 1. HPC-centric flow

Here, in order to get access to the requested input data, a user would formulate a domain-specific, semantic search query and send this request to the DMS. Usually, a DMS would use a dedicated database or search engine to filter the requested data. The data is loaded into the running code of the user. This could either require a dedicated data transfer to a pre-configured storage target, or the HPC system and the DMS are already working on the same storage system. When looking at different established systems, compare, the user is typically responsible for lineage recording and enforcing reproducibility. Therefore, these solutions typically only assist users to manage and organize their data but do not free them from the burden of efficient IO and working in agreement with good scientific practice.

3) Use-Case-Centric Paradigm: The opposite way to integrate a DMS into an HPC workflow, is to use the DMS as the user frontend, see Figure 2.

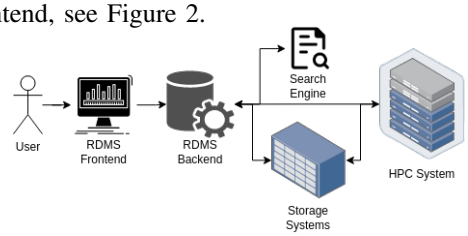

Fig. 2. DMS-centric flow

This DMS-provided user interface can be used to query and select input data, define a compute job, and submit this job to an HPC system, without the need to extra login to the HPC system or transfer data explicitly. This functionality requires a communication channel, between a remote DMS and an HPC system. Additionally, the DMS needs to be able to work with the individual resource manager, of each HPC system. The advantage of this approach lies in the capability to perform transparent lineage recording and can guarantee reproducibility. That is, because the DMS have complete control over the input data and the processes which run on them, using thorough logging methods is enough to ensure reproducibility. Similar to the HPC-centric use case, storage tiering is hard to support. In some existing implementations [17] data staging is explicitly required for each task execution.

C. Data Flow

These two scenarios also differ in their data path, i.e. in the storage systems involved in the data management and data processing.

1) Compute-Centric: In the HPC-centric use case, a user accesses data through their respective, native interface, e.g. through the library functions of their respective programming language, or as an input parameter of their program which they want to run. That means, that only data is available which is directly accessible from the HPC system, and data transfers, e.g., for better performance, have to be done manually. In

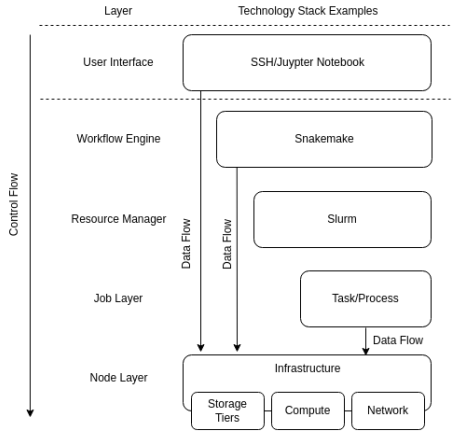

Fig. 3. Compute-centric flow

Figure 3 a layered diagram is shown which shows the possible data staging strategies. Within the user interface, e.g. ssh or a jupyter notebook, a user can explicitly copy/stage data on non-node-local storage. Within this layer, this has to be deliberately done. Assuming that a workflow engine, like snakemake, is used, data can be staged on non-node-local storage in a more automated and transparent way in the form of a dedicated workflow step. These two options can be considered asynchronous data staging since this will not lead to stalling times on the compute infrastructure. Synchronous data staging happens on the Job Layer, where a process first has to access data on a slow storage tier and stage it, and in this case even on a very fast node-local storage tier, before it can continue to process the data. Since the node-local storage is typically only available during the resource reservation, which is managed by the resource manager, e.g. Slurm, a user has to ensure that the overall process run on that node, does not only stage data, but also archives it once it is done. The entire data staging and IO optimization is therefore solely the user's responsibility.

 ISBN: 978-1-68558-073-5

2) DMS-Centric: Within the DMS-centric approach, a user is generally not interested to access the data directly, e.g. with a suitable library into self-written code. Instead, they are rather interested to have the complexity of running their job abstracted away. For this strategy, a layered diagram is

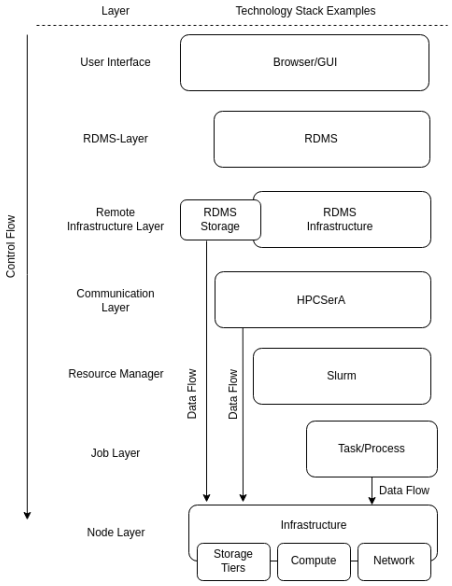

Fig. 4. DMS-centric flow

shown Figure 4, depicting the data flow. Here, a user access the DMS via a browser, or a DMS-specific graphical user interface (GUI). Within this interface, a user triggers the execution of a workflow, or a single analysis step on the selected input data. Often, these DMS are deployed in a cloud environment and have their own storage layer. Depending on whether this storage tier can be integrated into the HPC system, there are different strategies for data access. Either one can asynchronously or synchronously fetch data from the DMS within a dedicated data mover process and stage it either node-local, in the case of a synchronous data transfer, or non-node-local in the case of an asynchronous data transfer. For this purpose, the data mover process would either be granted access to the DMS storage with respect to the user's permissions, or the DMS can provide an endpoint, for instance, a REST endpoint, from which the process can fetch the required data. Since generally there is a communication layer, like HPCserA, required, to access the resources of an HPC system from the outside, this can also be used to asynchronously fetch data from the DMS. Lastly, a dedicated mover process can also fetch the data synchronously from the DMS on the compute node itself. This means, that the entire data movement and staging strategy is solely in the hands of the admins of the DMS, where the corresponding functionality is implemented and configured.

D. Control Flow

Similar to the aforementioned data flow, there also exists a control flow, as can be seen by the left arrow in Figure 3 and Figure 4. The control flow is initially triggered by the user within the user interface and is from there passed down to the final task running on a node. From this upmost layer, the control path goes down via an optional workflow layer to the resource manager where the tasks get mapped on the actual hardware, in case of the HPC-centric view. In the DMS-centric view, the control flow gets even more abstracted, since the user input recorded by the user interface has to first pass through DMS layer, where the user request gets initially processed and mapped on the DMS infrastructure. Since this DMS system is completely disjunct from the HPC system, a dedicated communication layer, like HPCSerA, is required to bridge those two systems. On the HPC system, the individual tasks are again mapped to the nodes in the infrastructure layer via the resource manager.

E. Analysis

To summarize the previous discussion about the different user interaction paradigms, Table I compares the characteristics of the individual components – ignore the column Governance-Centric for now.

The responsibility for one of the defined components and features are either the user, i.e., a manual process, semi-automatic - thus aiding the user (potentially following a specification), or fully automated. User-specific means it depends on the skill of the user. The resulting differences in the degree of automation can be illustrated best if we look at the interfaces a user can use to interact with the processes and data. Traditionally, only ssh connections are supported, whereas at least for the process interaction in the compute-centric paradigm, some interactions may take place via a web interface. In the use case-specific paradigm typically only a web interface is available that hides the HPC system and all internal processes.

The resource handling requires a lot of manual interaction of the users in the traditional and the compute-centric concept, while it is completely automated in the use case-centric approach based on configurations provided by the admins of the specific system.

A similar pattern can be seen in the characteristics of the task-related components. Here, the user experience with the HPC system evolves from a completely manual interaction to a partially automated or guided system in the compute-centric context, where already some low-level programming tools for workflow orchestration, data selection, and containers for dependency management are used. However, the program and data management rely still on manual work and are therefore potentially error-prone. On the other side, the use case-centric system fully automates these steps. Again, all task-related interactions are fully automated by the use case-specific system.

These intrinsic characteristics have different advantages and disadvantages. The traditional HPC usage paradigm relies heavily on manual work by the user to achieve a reasonable performance. Also the data management, the integration of storage and compute, and therefore the overall reproducibility is very much exposed to user errors. However,

 ISBN: 978-1-68558-073-5

TABLE I
COMPARISON OF DIFFERENT HPC USER INTERACTION PARADIGMS

Characteristics	Traditional	Compute-Centric	**Governance-Centric**	Use Case-Centric
Resources (Compute)	Auto	Auto	Auto	Auto
Resources (Storage)	Manual	Manual	Auto	Auto
Res. Mgmt (Compute)	Semi-Auto	Semi-Auto	Auto	Auto
Res. Mgmt (Storage)	Manual	Manual	Auto	Auto
Job spec	Manual	Semi-Auto	Semi-Auto	Auto
Program	Manual	Manual	Semi-Auto	Auto
Software land	Provided	Provided/User-Container	Provided/User-Container	Provided
Workflow spec	Manual	Semi-Auto	Semi-Auto	Auto
DMP	Manual	Manual	Semi-Auto	Tool-specific
User interface	SSH	SSH+Web	Web+SSH	Web
User interface (Data)	SSH	SSH	Web+SSH	Web
Client	SSH	SSH+Browser	Browser+SSH	Browser
Performance	User-specific	User-specific	++	+
Data management	-	-	++	Tool-specific
Integration	–	0	++	++
Reproducibility	–	+	++	Tool-specific
Flexibility	++	++	+	–

this enables the highest level of flexibility. The compute-centric paradigm improves this by utilizing the discussed semi-automated components and hereby improves the integration and reproducibility. The use case-specific systems will most likely have reasonable, but not custom-made, configurations to achieve good performance - here the interaction with data is challenging as the upload/download via Web-frontend limit performance. The current challenge is to unify the concepts of the compute-centric and the use case-centric paradigms and combine the advantages of these worlds.

IV. GOVERNANCE CENTRIC ARCHITECTURE

Our goal is to expand upon the existing concepts to provide a novel, unified view of processes and data in order to improve the user experience on HPC systems. Here, the metrics for the user experience, i.e. performance, data management, integration, reproducibility, and flexibility, all basically boil down to the question of where the data is located and how they are linked. This has to be tackled simultaneously in two directions: first, an additional integration layer above the resource manager (compare Figure 3 and Figure 4) is required. This layer has to provide an integrated and unified namespace to the users. Secondly, an information flow, which is directed in the opposite direction as the control flow in Figure 3 and Figure 4, is required. Although this can be achieved with available auditing tools, see Section II-D, there is no concept for a tool that processes the incoming information and hereby makes it actionable. The advantage of an actionable information flow compared to the existing systems is that the information is utilized to create a desired, predefined state, and not just create yet another piece of data a user has to manage manually.

To this end, we propose the governance-centric interaction paradigm that aids the users and automizes the integration of data and compute. In Table I, we have identified the required degree of automation to bridge the gap between the compute-centric and the use-case-centric paradigms. Resource management should be fully automated to achieve the highest degree of integration of data and compute. The task-specific components should be semi-automated to guide the user in managing the data, working reproducibly, and ensuring that a predefined, ideal state is reached while allowing as much flexibility as possible. Similarly flexible should be the interface, to allow interaction from both user groups.

We believe professional and proper data management requires users to define an *experimental description* at the beginning of a project consisting of a workflow linking data sets and compute tasks and a data management plan for the respective input/output data. Then the user has to initially *modify their tasks*, e.g., job scripts, to allow linking of the tasks and their data products to the workflow and also to generate descriptive metadata for the data sets. Building upon the previous discussion, the goal is to not only use the workflow as an abstract concept that users may informally follow but rather enforce its usage. The *implications of the design* are that the HPC system can exploit the information to perform many previously manual tasks automatically and fulfill our goals.

For instance, to automatically receive and process information about input data and artifacts created during task execution, or enforcing archival/deletion policies defined in the DMP. To unify both user groups, the ingest of results into a DMS along with all required metadata including lineage information has to be one of the supported features.

To explain this idea in more detail, the experimental description shall be a user-defined and machine-readable workflow description that contains information about the data flow, the tasks which process these data sets and create artifacts, and further optional information like access policies or the IO intensity of each task. This means that, for every task a user wants to schedule via the resource manager, this task has to be linked to a specific workflow step within the experimental description at job submission time. Thus, each and every submission of a job on a HPC system becomes one concrete invocation of the abstract task description within the experimental workflow linked to data in the DMP.

A. Experimental Description

Specifically, in data-driven projects, it is common that there is not a single task, but that the entire processing consists

 ISBN: 978-1-68558-073-5

of multiple steps which are concatenated into a workflow. Therefore, a user has to provide a simple graph, compare Figure 5, connecting input and output data via tasks as a workflow description. This workflow could represent a weather prediction, where each cycle represents the simulation of the next hour (in the future); Dataset2 is the initial conditions while Dataset1 holds the model. Manual steps in the workflow are explicitly annotated as they require data to be accessible.

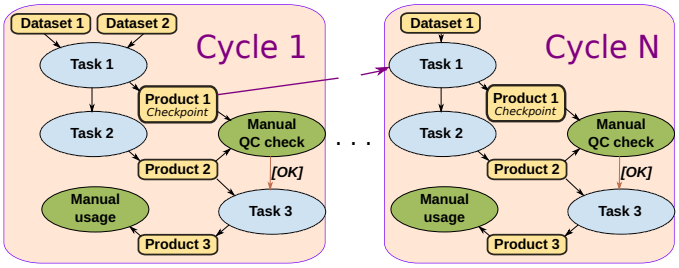

Fig. 5. High-level view of a workflow.

Within this workflow definition, general policies can be defined, e.g. where and when data should be archived, how long artifacts should be kept on hot storage if a manual inspection is required, what accompanying metadata are required, or if input data can be altered. The archiving of data should explicitly support remote DMS as a target, to integrate HPC systems with remote DMS, and similarly, data within a remote DMS should also serve as possible input data. The required data mover tools have therefore to be integrated as dependencies into the DMP. In addition, the users can provide information about the expected IO profile, aiding the proposed DMP tool to find the best storage tier based on heuristics configured by the HPC admins. Based on further metrics, like the available bandwidth of a remote DMS and the HPC system, or the amount of data, the DMP tool can also determine, whether data should be staged synchronously, i.e. during compute time, or asynchronously, i.e. as a dedicated, dependent step before the compute task starts.

Describing Datasets: The user can and should add further information to the data sets which are getting processed or created. To improve the findability a user should provide domain-specific metadata, or define a task to extract those. The required, and optional domain-specific metadata fields can be defined in the DMP. For instance, in our figure, Product 2 may be characterized by the date/time of the weather prediction and each product could be tagged with the model configuration settings. This can even ensure a homogeneous metadata quality across a larger group working on a joint project. In addition, the data life cycle should be defined, i.e. what is the retention time, what are the deletion policies. To meet the data governance policies required by the user, additional aspects such as access control must be defined to prevent unwanted data leakage.

B. Modifying Tasks

Compute jobs on HPC systems are dispatched to the actual resources using resource managers such as Slurm. On this level, a job has to be prepared, annotated, and linked with the workflow. The user annotations should specify the task within the defined workflow, which is to be executed. In addition, a user can further restrict and specify the input data. Here, the largest change compared to the traditional HPC interaction paradigm becomes apparent: Instead of working with explicit files, a user rather works with datasets defined by metadata. For instance, a user specifies the input data either based on domain-specific metadata or simply due to the link in the DMP. Therefore, the actual storage location is abstracted from the user. The actual data directory, which a program still needs to specify in API calls, can be automatically exported via environment variables or generated via support tools in the job script. Before reserving dedicated compute resources, the proposed DMP tool decides to use either synchronous or asynchronous data staging and stages the data respectively.

One key requirement in science is reproducibility. In a first step, this requires at least sufficient provenance information to allow for retrospective comprehensibility of the lineage of the resulting artifacts. One important element to retrospectively comprehend an HPC job is the run script used for batch processing. This can be automatically archived along with the artifacts by the proposed DMP tool. However, these batch scripts, which contain the actual compute job to be run in the form of a shell script, can have multiple ambiguities. One simple example of this would be the execution of an interpreted script, e.g. a *Python* script. Here, one would have a simple line within the batch-script which would look similar to:`$ python my_script.py`
The challenge within this call is to track differences between multiple invocations of this script, where the content of *my_script.py* has been changed. For this, three different high-level modi are proposed.

The recommended way is to use a Git repository, where changes in code are properly tracked. In this case the DMP tool checks in the directory of the script and saves the information about the Git repository and the used Git commit hash so that this information can be stored in the metadata of the created data products. The DMP tool will create a dedicated sidecar file for this metadata in the output directory. The usage of version control systems can, and should, also be part of the required specifications within the DMP.

Alternatively, if no Git repository is set up, the batch script is parsed and untracked dependencies in the user namespace, like a Python script, are tried to be identified and archived along with the artifacts. Since this method is potentially more error-prone compared to a proper version control system, like Git, it is not recommended, but should still offer a better chance for retrospective comprehensibility when compared to other strategies. These dependencies will be listed in the before-mentioned sidecar file, and are archived alongside it. One important distinction to make is the use of containers. In this case, the container image should be archived and linked to the sidecar file. Of course, utilizing provenance-specific tools, as discussed in Section II and translating them to the required standard in the sidecar file, is also an option to explore.

 ISBN: 978-1-68558-073-5

The third option is that users compose the sidecar file by adding code to the batch script, where this provenance information are provided. This can be added to the directory where the output, which should be archived, resides. This file will also override information that was automatically tried to extract in the previous step.

C. Implications of the Design

This presented design has different positive implications on the user experience that we summarize in Table I.

a) Integration: First of all, the abstraction of files on storage towards more high-level data sets achieves the tight integration of storage and compute. Abstracting the storage from the data will motivate users to use proper metadata management systems and establish a data catalog, instead of encoding information into file paths.

b) Performance: Since users are only working with datasets and not with a filepath anymore, HPC admins can configure data placement strategies, therefore relieving this burden from the users and optimizing the performance.

c) Reproducibility: Since compute tasks, their input, and resulting data are strongly linked with each other, the lineage of artifacts is much more comprehensible and less subjected to user errors. Utilizing further tools like containers and a version control system will ensure full reproducibility, which is integrated into this paradigm by design because this is just another policy in the defined data governance, which will be enforced for the users.

d) Enforcing DMP: Although the general idea of using data management plans in HPC is far from new, the novel advantage of this particular tool is that it can be enforced. There are various ways to achieve this goal. A naive approach compatible with existing systems is to use a cronjob that reads in the workflow and task definition files, which a user has provided, and compares the specified, desired state of the storage systems of the user against the actual state at hand. If new output data are detected and the required sidecar file for the necessary metadata is available, the output data is handled as specified. However, if the required sidecar file is not, or only with insufficient content provided the user will be reminded to provide the missing information after a specified grace time. Similarly, if data is detected which can not be matched to the dataset specification in the workflow definition, an error or warning can be raised to the user as such unclassified data shall not exist. Thus, the DMP becomes actionable and hereby ensures a homogeneous system state in sync with the experimental description and user expectations.

V. Conclusion and Future Work

In conclusion, we introduced the governance-centric interaction paradigm, which by design integrates storage and compute for the users. It relies on a researcher to define an experimental description and a DMP for the datasets at the beginning, something that good scientific practice requires anyhow. This will allow the system to perform various tasks on behalf of the user and increase overall automatization. Ultimately, the burden of performance optimization can be shifted partially from each user to data center operators. Furthermore, the abstraction of files to data sets allows the seamless integration of a DMS. Since this paradigm seamlessly links data to compute tasks, it ensures retrospective comprehensibility and reproducibility by design.

We are in the process of developing tools and an environment where this vision is implemented. In future work, this concept will be evaluated on specific use cases. In addition, synthetic benchmarks will be used to evaluate the proposed concept of storage and compute integration with other tools offering a unified namespace across a tiered storage system.

ACKNOWLEDGEMENT

We gratefully acknowledge funding by “Nationales Hochleistungsrechnen” and BMBF under 01—S22093A.

References

[1] M. D. Wilkinson *et al.*, “The fair guiding principles for scientific data management and stewardship,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[2] P. Corbett *et al.*, “Overview of the mpi-io parallel i/o interface,” *Input/Output in Parallel and Distributed Computer Systems*, pp. 127–146, 1996.

[3] P. Schwan *et al.*, “Lustre: Building a file system for 1000-node clusters,” in *Proceedings of the 2003 Linux symposium*, vol. 2003, 2003, pp. 380–386.

[4] F. B. Schmuck and R. L. Haskin, “Gpfs: A shared-disk file system for large computing clusters.” in *FAST*, vol. 2, no. 19, 2002, pp. 231–244.

[5] V. Korol *et al.*, “Introducing viking: A novel online platform for multiscale modeling,” *ACS omega*, vol. 5, no. 2, pp. 1254–1260, 2019.

[6] M. T. Nelson *et al.*, “Namd: a parallel, object-oriented molecular dynamics program,” *The International Journal of Supercomputer Applications and High Performance Computing*, vol. 10, no. 4, pp. 251–268, 1996.

[7] D. Van Der Spoel *et al.*, “Gromacs: fast, flexible, and free,” *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[8] D. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, “The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data,” *Neuroinformatics*, vol. 5, pp. 11–33, 2007.

[9] R. L. Harrigan *et al.*, “Vanderbilt university institute of imaging science center for computational imaging xnat: A multimodal data archive and processing environment,” *NeuroImage*, vol. 124, pp. 1097–1101, 2016.

[10] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system,” in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. Ieee, 2010, pp. 1–10.

[11] V. K. Vavilapalli *et al.*, “Apache hadoop yarn: Yet another resource negotiator,” in *Proceedings of the 4th annual Symposium on Cloud Computing*, 2013, pp. 1–16.

[12] M. Milligan, “Interactive hpc gateways with jupyter and jupyterhub,” in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, 2017, pp. 1–4.

[13] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. I. Seltzer, “Provenance-aware storage systems.” in *Usenix annual technical conference, general track*, 2006, pp. 43–56.

[14] D. Dai, Y. Chen, P. Carns, J. Jenkins, and R. Ross, “Lightweight provenance service for high-performance computing,” in *2017 26th International Conference on Parallel Architectures and Compilation Techniques (PACT)*. IEEE, 2017, pp. 117–129.

[15] F. Chirigati, R. Rampin, D. Shasha, and J. Freire, “Reprozip: Computational reproducibility with ease,” in *Proceedings of the 2016 international conference on management of data*, 2016, pp. 2085–2088.

[16] J. Lofstead, J. Baker, and A. Younge, “Data pallets: containerizing storage for reproducibility and traceability,” in *High Performance Computing: ISC High Performance 2019 International Workshops, Frankfurt, Germany, June 16-20, 2019, Revised Selected Papers 34*. Springer, 2019, pp. 36–45.

[17] S. Bingert, C. Köhler, H. Nolte, and W. Alamgir, “An api to include hpc resources in workflow systems,” in *INFOCOMP 2021: The Eleventh International Conference on Advanced Communications and Computation*, 2021, pp. 15–20.

 ISBN: 978-1-68558-073-5

Accuracy of Simulation of Wireless Technology Using MATLAB and NS-3

David Newell, Philip Davies, Russell Wade, Andrew Yearp, Ben Lister
Faculty of Science & Technology
Bournemouth University
Bournemouth, UK
dnewell@bournemouth.ac.uk, daviesp@bournemouth.ac.uk
rwade@bournemouth.ac.uk, ayearp@bournemouth.ac.uk

Mak Sharma
School of Computing & Digital Technology
Birmingham City University
Birmingham, UK
Mak.Sharma@bcu.ac.uk

***Abstract*---Performance predictions of wireless IEEE 802.11 specifications have been obtained using two well-known simulators - MATLAB and Network Simulator 3 (NS-3). Benchmarking was done by comparison to laboratory performance measured practically using an experimental method based on technical features claiming to contribute to higher bandwidth. The findings show that both simulators' predictions match the specifications at zero line-of-sight between transmitter and host. Technical features were confirmed to broadly increased data rates approximately to specification. However, accuracy of the simulators was inversely correlated with propagation distance. In certain cases, the claim of higher bandwidth by the latest amendment was not borne out in practice over distances greater than 10m. In conclusion, NS-3 was more often better correlated with measured values than MATLAB. In future, better software modelling of ray tracing and beam forming, physical techniques employed in cellular simulators that better model propagation effects would be expected to improve accuracy.**

Keywords-WIFI, Wireless technology, MATLAB, NS-3.

I. INTRODUCTION

The IEEE 802.11 standard is the de-facto solution for wireless network access. Previous workers have considered discrepancies between theoretical and physical performance under various conditions [1]-[3].

Little work has investigated variance between theoretical, practical and simulated performance measurements. With such practical variances being observed, the possibility of a similar discrepancy with the predictions of 802.11 simulations arises. The lack of research into these variances thus draws into question the validity of simulated performance measurements and hence the indicated benefits of any proposed enhancements demonstrated through simulations. The IEEE standard features an evolving series of enhancements each of which is defined within the specification amendments with the effect on the performance theoretically attainable. To evaluate the effectiveness of proposed enhancements, several academics have utilized simulation software [4]-[6]. The results of simulations are interpreted as an indication of the improvements attainable in practical applications.

The 802.11n amendment, published in 2009 and ratified in IEEE 802.11-2012 [7], specifies a number of enhancements to improve on throughput, range and reliability. Physical layer (PHY) enhancements include advanced modulation techniques, utilization of Multiple-Input Multiple-Output (MIMO) antennas, wider channels and operation in the 2.4GHz or optional 5GHz frequency bands. Medium Access Control (MAC) layer enhancements consist of frame aggregation and block acknowledgements to increase MAC layer efficiency.

The 802.11ac amendment was published in 2013 and ratified in IEEE 802.11-2016 [8]. It specifies PHY layer enhancements for 5GHz exclusivity, increased number of MIMO streams, Multi-user MIMO, wider channels and further advancements to modulation techniques. The MAC layer enhancements build on previous frame aggregation techniques.

The structure of the paper is as follows. Section 1 is the Introduction. Section 2 is a discussion of the selection of a method and the design choices made. Section 3 is an analysis of the findings. Section 4 is the evaluation and draws conclusions. Finally, section 5 is a list of the references.

II. METHOD

Innovative features in 802.11 amendments were identified - modulation, convolutional coding, channel widths, guard intervals, MIMO, spatial streams, beam-forming, frame aggregation and block acknowledgements. Each innovation is responsible for a notable increase in PHY or MAC level data rates. Additionally the medium-specific variables of Signal to Noise Ratio (SNR) and obstructions must be accounted for. Each experiment must isolate an individual variable as best as possible within the environment in which the experiments are performed, the conditions under which the experiment is conducted and the measure of performance. Then repeated in MATLAB and NS-3 simulation, the results collated into datasets. Experiments measured the downstream data rate from access point to client, over distinct distances.

NS-3 [9] and MATLAB [10] were used for simulations. NS-3 is an open-source real-time network simulator,

 ISBN: 978-1-68558-073-5

supporting numerous technologies e.g. Ethernet, Wi-Fi, Worldwide Interoperability for Microwave Access (WiMAX), Long Term Evolution (LTE) and 5G. MATLAB supports a variety of technologies using 'Toolboxes'.

In order to simulate the 802.11 specification and physical layer communication effects (attenuation, noise, interference etc.) in MATLAB, two add-on packages were used, viz., Communications System Toolbox for Radio Frequency (RF) modelling and Wireless Local Area Network (WLAN) System Toolbox for modeling the 802.11 MAC and PHY layers.

MATLAB simulations used an 802.11ac simulation script [15]. Ten simulations were scripted, five for 802.11n and five for 802.11ac. The channel model within the simulation was configured to replicate the RF environment as closely as possible. All configured parameters were equal for both High Throughput (HT) and Very High Throughput (VHT) simulations.

MATLAB was used with wlanTGacChannel and wlanTGnChannel objects configured with a delay profile of Model-D. The model was found to be the most representative of the practical results recorded. The key factor in this decision was the breakpoint distance of 10 meters i.e. Line Of Sight (LOS) transmission for $\leqslant$ 10m and Non Line Of Sight (NLOS) for > 10m, produced results to most closely match those observed.

The carrier frequency was set to channel 44, 5.20GHz. Noise was introduced through the Adds White Gaussian Noise (AWGN) Channel object, with the initial signal to noise ratio being set to 97dB. The simulation was configured to take into account path loss and shadowing RF propagation effects.

NS-3 simulations used scripts modified for HT [16] and VHT [17]. Six scripts were created - three for 802.11n and three for 802.11ac. Each script calculated the data rate achieved at each distance for all parameter combinations, with each script doing so for 1, 2 or 3 spatial streams.

NS-3 provides a variety and ever growing number of wireless propagation models. Each is built around a unique set of equations and up to seven may be chained to produce a single complex propagation model. A number of models were evaluated for suitability. Based on the results of each, and considering recommendations towards path loss models for this use case by [18], the decision was made to implement the Log Distance Propagation Model alongside the Nakagami Propagation Loss Model.

The key disadvantage to these models, however, was the lack of accounting for complex RF effects, such as shadowing. Thus a degree of variance was expected in the results. Chaining the Shadowing Loss propagation model resulted in an unrealistic simulation of wireless performance, with a greater degradation of signal quality over distance being exhibited. Hence the shadowing loss propagation model was not implemented. In order to model shadowing loss to some degree, the Random Propagation Loss Model was implemented using a random integer between 0 and 5. The result was a random signal loss between 0 and 5dBm. For configuration of the propagation model, the logarithmic power was set to 3.00. Due to constraints imposed by the execution time of each simulation, it was not possible to evaluate powers to a greater precision than 0.25.

As shown in Table 1, measurements and simulations were made using a range of tools. Directional data rate was measured Access Point (AP) to client using a constant stream of uniformly formatted packets and frames. The Internet Performance Working Group (IPERF) [11] network benchmarking tool was chosen as it supports User Datagram Protocol (UDP), packet loss, logging and streaming.

For the monitoring of noise and SNR, SDRSharp (SDR#) [12] for Windows was chosen for basic functionality and simplistic interface of SDR#. For accurate identification of the received signal strength (dBm) the inSSIDer [13] tool was chosen. Combined, these tools provide the necessary measurements to monitor local RF conditions using the Software Defined Radio (SDR).

TABLE 1. SUMMARY OF TOOLS FOR PRACTICAL MEASUREMENTS AND SIMULATIONS.

Software	Usage	Description
IPERF	Practical	UDP Datagram generation
SDR#	Practical	Spectral Analysis
inSSIDer	Practical	Identify Signal Strength
WLAN Toolbox	Simulation	MATLAB simulation
Comms Toolbox	Simulation	MATLAB RF simulation

For the monitoring of noise and SNR, SDR# [12] for Windows was chosen for basic functionality and simplistic interface of SDR#. For accurate identification of the received signal strength (dBm) the inSSIDer [13] tool was chosen. Combined, these tools provide the necessary measurements to monitor local RF conditions using a software defined radio.

The infrastructure in Figure 1 consists of the Cisco Aironet 2702i with 2504 WLAN controller. A TP-Link T9UH adapter provided support for up to 3x3:3 MIMO. As a result of access point and network adapter limitations, the experiments were limited to a maximum of 3x3:3 MIMO and 80MHz channel widths. For spectral analysis the HackRF SDR was chosen for 20MHz bandwidth and $\leqslant$ 6GHz tuning range [14]. All equipment was switched through the C3650-24PS switch in order to provide a 1Gbps link.

 ISBN: 978-1-68558-073-5

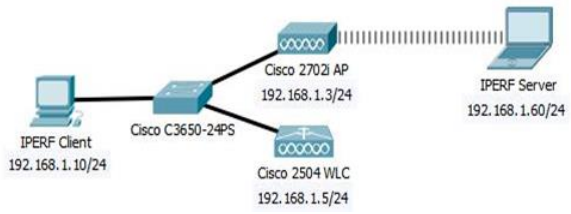

Fig. 1. Wireless Network Infrastructure for Practical Measurements

III. FINDINGS

A. Modulation Schemes

Practical measurements matched theoretical and simulated increases in data rates because of increases in Orthogonal Frequency-Division Multiplexing (OFDM) symbol density. HT and VHT exhibited in data rate variations of 3.5% and 9.6% respectively over four distances. NS-3 consistently predicted optimistic data rates with 35%, 22% and 14.5% average over prediction for Binary Phase Shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK) and 16-Quadrature Amplitude Modulation (QAM). For VHT, performance did not match that predicted for larger distances. Both simulations suggest logarithmic compared to observed linear performance due to pessimistic channel modelling or environmental effects producing gains. NS-3 showed minimal variation between predicted and measured performance with a 0% and 6% variation.

B. Spatial Streams

Theoretical data rates were not exhibited in reality. The improvement in data rate over lesser numbers of spatial streams decreased. In the case of VHT, the data rates for 2 and 3 spatial streams decreased below that of 1 spatial stream at 10 and 15 meters.

For HT, MATLAB correctly predicted the degradation of performance for 1 spatial stream across all distances. However, for 3 spatial streams the prediction was not representative of reality. For VHT, whilst the prediction for the degradation of performance over distance was pessimistic, the simulation did correctly predict that beyond 10 meters, 1 spatial stream would retain the data rate such that it would exceed the data rates of both 2 and 3 spatial streams. The predictions were, in general, not indicative of the true performance. NS-3 was noted to be the most accurate at predicting HT performance over all distances, especially for 1 and 2 spatial streams. It correctly predicted that over all four distances the performance of the increased number of spatial streams would not degrade below that of the lower numbers of spatial streams. For VHT however, NS-3 did not correctly predict the performance across all measured distances. Overall, the key observation from the measurements and predictions was that existing simulation models do not accurately portray the true performance of 802.11 standards when taking spatial multiplexing (spatial streams) into consideration.

The practical measurements showed that the theoretical increases in data rates, attributed to increases in channel width, held true. For HT, the data rate increased by 66Mbps for 20MHz compared to 40MHz. For VHT, the data rate increased by 90Mbps and 188 Mbps for 20MHz to 40MHz and 40MHz to 80MHz channels. Note that the data rate more than doubled in all cases. This effect was not expected. A possible cause for such an effect may have been cross-channel interference reductions as bandwidth increased, due to dynamic channel assignment taking effect at remote access points. It should also be noted that the decrease in data rate at 5 and 10 meters for VHT 20MHz was experienced across all 10 measurements at each distance. As the experiments were performed in an environment containing other wireless transmitters (APs, Devices etc.) this decrease may be attributed to interference from said transmitters at the time of the measurements being taken. This would explain why the effect was not measured across the 40MHz and 80MHz measurements. NS-3 and MATLAB were shown to produce realistic performance predictions, modelling both the data rate and performance over distance successfully for each channel width. As previously discussed, NS-3 was unable to successfully model transient environmental effects. Additionally, for both HT and VHT, the predicted performance degradation was again overly pessimistic for both simulations. Unlike previous experiments, this degradation did occur correctly between the 10 and 15 meter distances, however not to the degree predicted.

C. Guard Intervals

The practical measurements found for HT and VHT respectively a 10% and 9.5% increase in data rates when changing from 800ns guard intervals to 400ns. This closely matched the theoretical increase of 11.1% stated within the 802.11 specification.

Additionally, for VHT, the long guard interval was shown to exceed the performance of the short guard interval as the signal quality degraded. This matched theory as longer guard intervals are expected to improve performance due to additional time for reflected signals to disperse prior to transmission.

NS-3 predictions closely matched measurements for HT with a maximum variance of 20Mbps being observed. For VHT, NS-3 did not correctly model reduced data rates at 0 meters, nor the sharp decrease observed between 5m and 10m. However, for the 5m and 15m measurements, predictions for the long guard interval varied by only 5.7% and 1.6%.

MATLAB predictions were somewhat inaccurate. For HT, a sharp performance reduction was incorrectly predicted, occurring beyond 9 meters. However, the increase in data rate of 11.1% was correctly predicted. For VHT, whilst the model was overly pessimistic, it did

 ISBN: 978-1-68558-073-5

correctly predict a sharp performance reduction between 5m and 10m. It did not predict the long guard interval overtaking the performance of the short guard interval.

D. Coding Schemes

Practical measurements showed that the specification data rate increases for HT and VHT occurred. For HT, each increase in the ratio of data bits to error-checking bits theoretically results in the data rate increasing by 15Mbps (Between 2/3, 3/4 and 5/6). The true increase was measured to be 13Mbps moving from 2/3 to 3/4 coding and 11Mbps from 3/4 to 5/6 coding. VHT on the other hand measured only an 11Mbps increase between 3/4 and 5/6, compared to the stated theoretical increase of 30Mbps. Unlike HT, VHT showed a tendency for the achieved data rates to converge as the distance increases. This was expected as higher coding schemes are intended to offer higher data rates at the cost of reliable data transfer. A lower coding rate is expected to achieve greater data rates than higher coding rates as a reduction in signal quality is observed.

The predictions made by the NS-3 simulator were noted to be a closer representation of the true performance than MATLAB. For HT, MATLAB matched the theoretical increases of 15Mbps as the coding scheme increased. NS-3 predicted a 12Mbps and 15Mbps increase from 2/3 to 3/4 and 3/4 to 5/6 respectively. Additionally, at 15 meters, NS-3 correctly predicted a minor convergence of each coding scheme, as was observed to a lesser degree in the lab. Similarly, NS-3 most accurately predicted the VHT measurements both in terms of the data rate and the reduction in data rate between 5 and 10 meters. It did not however correctly predict the leveling off between 10 and 15 meters, nor did MATLAB.

IV. DISCUSSION

The performance of 802.11n and 802.11ac was found to be lacking when compared to the data rates stated in the specification. This performance was unable to be accurately modeled in a number of experiments by the two simulators investigated. Specifically, the following findings were made:

When simulating 802.11ac (VHT), both NS-3 and MATLAB were found, in general, to accurately model the increases in data rates. For non-zero distances it was found that the accuracy of the predictions made reduced as the distance between the transmitter and receiver increased due to the simulators' RF model(s) (channel model). Models could not be calibrated with known measurements.

When simulating 802.11n (HT) both simulators were found to more closely reflect the performance measured in reality, across all four distances. However, at 15 meters, the greatest variances were observed. MATLAB was shown to be the least accurate model of performance, with prediction accuracy falling greatly after 10 meters. NS-3 on the other hand more accurately predicts both the measured data rates and the reduction in data rates over distance, excluding predictions for multiple spatial streams.

Of the technical features evaluated in this report, the simulation predictions for spatial multiplexing (MIMO) were found to substantially differ from those measured in reality. The models used for NS-3 and MATLAB were found to more accurately reflect the measurements of HT and VHT respectively, with substantial prediction inaccuracies being made by NS-3 and MATLAB for VHT and HT respectively. Similar to the other experiments conducted, the factor responsible for such variations was the implementation of the RF model.

In high SNR conditions, advanced technical features aimed at increasing data rates fulfilled their function. However, these features did not necessarily do so to the degree stated in the 802.11 standard. As distances increased, and SNR fell, the performance gains decreased, in some cases to a point where the attainable data rate fell below that of features with lower stated data rates.

V. CONCLUSION

It has been shown that MATLAB and NS-3 successfully modeled the technical features of the 802.11 specification. For both 802.11n and 802.11ac, both simulators successfully modeled the increases in data rate provided by increasing the efficiency or complexity of each technical feature. Performance over distance is modeled through channel models. The inaccurate implementation of the channel models contributed greatly to the inaccuracy of the predictions made.

Better monitoring is recommended. E.g., Ettus Research X300 [19] for 160MHz channel bandwidth, full 802.11ac capabilities and support for low-level configuration using C++, Python or GNU Radio.

In future, 5G cellular contains innovations to increase data rates to up to 20Gbps per user with frequencies between 30 and 100GHz under consideration, full duplex, beamforming, MIMO and pico-cells [20]. Development simulators are the Third Generation Partnership Project's 3GPP [21] and the Novel Millimeter-Wave Channel Simulator (NYUSIM) simulator, built by researchers at New York University [22].

REFERENCES

[1] Y. Zeng, P. H. Pathak, and P. Mohapatra, "A first look at 802.11 ac in action: Energy efficiency and interference characterization," in Networking Conference, 2014 IFIP, 2014, pp. 1-9: IEEE.

[2] D. Newell, P. Davies, R. Wade, P. Decaux, and M. Shama, "Comparison of Theoretical and Practical Performances with 802.11 n and 802.11 ac Wireless Networking," in Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on, 2017, pp. 710-715: IEEE.

 ISBN: 978-1-68558-073-5

[3] M. Darbyshire. "Practical Performance, Interference Effects of Wireless Networks ", Bournemouth University, 2017.

[4] Z. Machrouh and A. Najid, "Performance Analysis of IEEE802. 11ac DCF Enhancement for VHT with Frame Aggregation," International Journal of Recent Contributions from Engineering, Science & IT (iJES), vol. 4, no. 3, pp. 17-21, 2016.

[5] W. Zhengzhong, Z. Xianxin, Y. Wenge, and L. Zilin, "System level simulation modeling of Bluetooth voice and its interference," in Signal Processing, 2004. Proceedings ICSP'04. 2004 7th International Conference on, 2015, vol. 1, pp. 29-32: IEEE.

[6] N. V. Doohan, D. K. Mishra, and S. Tokekar, "Reliability analysis for wireless sensor networks considering environmental parameters using MATLAB," in Computational Intelligence, Communication Systems and Networks (CICSyN), 2011 Third International Conference on, 2011, pp. 99-102: IEEE.

[7] "IEEE Standard for Information technology--Telecommunications and information exchange between systems--Local and metropolitan area networks--Specific requirements--Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications--Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz," IEEE Std 802.11ac(TM)-2013 (Amendment to IEEE Std 802.11-2012, as amended by IEEE Std 802.11ae-2012, IEEE Std 802.11aa-2012, and IEEE Std 802.11ad-2012), pp. 1-425, 2013.

[8] "IEEE Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012), pp. 1-3534, 2016.

[9] NSF, "Network Simulator 3. 3.27 ", ed. Alexandria, Virginia: National Science Foundation, 2018a.

[10] MathWorks, "MATLAB. 2017a ", ed. Natick, Massachusetts, 2017.

[11] J. Dugan, S. Elliott, B. A. Mah, J. Poskanzer, and K. Prabhu, "IPERF3. 3.1.3 ", ed. iperf.fr: iKoula, 2016.

[12] AirSpy, "SDR#. revision 1651 ", ed, 2018.

[13] Metageek, "inSSIDer. 3.1.2.1 " 2016.

[14] M. Ossmann, D. Spill, T. Streetman, E. Wooton, J. Graves, and M. Mclaurin, (2016, 25 February 2018). HackRF One. Available: https://greatscottgadgets.com/hackrf/

[15] MathWorks, "Packet (Frame) Processing", Massachusetts, 2018.

[16] M. Banchi, HT Wi-Fi Network [ns-3 script], 2018, Available: https://goo.gl/NuZKKq

[17] S. Deronne, VHT Wi-Fi Network [ns-3 script], 2018, Available: https://goo.gl/koR2KG

[18] M. Stoffers and G. Riley, "Comparing the ns-3 propagation models," in Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2012 IEEE 20th International Symposium on, 2012, pp. 61-67: IEEE.

[19] Ettus Research, (7 April 2018). USRP X300 Available: https://www.ettus.com/product/details/X300-KIT

[20] A. Nordrum and K. Clark, "Everything You Need to Know About 5G," IEEEE, 2017.

[21] 3GPP, "LTE; 5G; Study on channel model for frequency spectrum above 6 GHz ", Sophia Antipolis, France, 2017, Accessed on: 12 April 2018.

[22] S. Sun, G. R. MacCartney, and T. S. Rappaport, "A novel millimeter-wave channel simulator and applications for 5G wireless communications," in Communications (ICC), 2017 IEEE International Conference on, 2017, pp. 1-7: IEEE.

 ISBN: 978-1-68558-073-5