Message-Passing Interface for Java Applications: Practical Aspects of Leveraging High Performance Computing to Speed and Scale Up the Semantic Web

Alexey Cheptsov and Bastian Koller
High Performance Computing Center Stuttgart (HLRS)
University of Stuttgart
70550 Stuttgart, Germany
Email: {cheptsov,koller}@hlrs.de

***Abstract*—The age of Big Data introduces a variety of challenges in how to store, access, process, and stream massive amounts of structured and unstructured data effectively. Among those domains that are impacted by the Big Data problem at most, the Semantic Web holds a leading position. By current estimates, the volume of Semantic Web data is exceeding the order of magnitude of billions of triples. Using High Performance Computing infrastructures is essential in dealing with these massive data volumes. Unfortunately, the most Semantic Web applications are developed in Java language, which makes them incompatible with the traditional high performance computing software solutions, which are tailored for compiled codes developed in C and Fortran languages. The known attempts to port existing parallelization frameworks, such as the Message-Passing Interface, to the Java platform have proved either a poor efficiency in terms of performance and scalability, or a limited usability due to a considerable configuration and installation overhead. We present an efficient porting of Java bindings based on Open MPI - one of the most popular Message-Passing Interface implementations for the traditional (C, C++, and Fortran) supercomputing applications.**

***Keywords*-High Performance Computing, Big Data, Semantic Web, Performance, Scalability, Message-Passing Interface, Open MPI.**

I. Introduction

The data volumes collected by the Semantic Web have already reached the order of magnitude of billions of triples and is expected to further grow in the future, which positions this Web extension to dominate the data-centric computing in the oncoming decade. Processing (e.g., inferring) such volume of data, such as generated in the social networks like Facebook or Twitter, or collected in domain-oriented knowledge bases like pharmacological data integration platform OpenPHACTS, poses a lot of challenges in terms of reaching the high performance and scalability by the software applications. As discussed in our previous publication [1], while there is a number of existing highly-scalable software solutions for storing data, such as Jena [2], the scalable data processing constitutes the major challenge for data-centric applications. This work is discussing application of the techniques elaborated in the previous paper to the Big Data application domain. In the literature, it is often referred as "Big Data" a set of issues related to scaling existing processing techniques to large amounts of data, for which standard computing platforms have proved inefficient [3]. Among those data-centric communities that address the Big Data, the Semantic Web enjoys a prominent position. Semantic Data are massively produced and published at the speed that makes traditional processing techniques (such as reasoning) inefficient when applied to the real-scale data. It is worth mentioning that the typical Semantic Web application workflows are highly parallel in their nature (see Figure 1) and are well-suited to run in high performance computing environments.

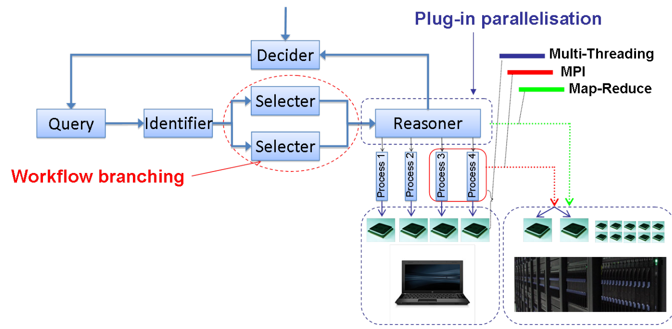


Figure 1. Parallelization patterns in a Reasoning application's workflow.

The data scaling problem in the Semantic Web is considered in two its main aspects - horizontal and vertical scale. Horizontal scaling means dealing with heterogeneous, and often unstructured data acquired from heterogeneous sources. The famous Linked Open Data cloud diagram [4] consists of hundreds of diverse data sources, ranging from geo-spatial cartographic sources like *Open Street Map*, to governmental data, opened to the publicity, like *data.gov*. Vertical scaling implies scaling up the size of similarly structured data. Along the open government data spawns over 851,000 data sets across 153 catalogues from more than 30 countries, as estimated in [5] at the beginning of 2012. Processing data in such an amount is not straightforward and challenging for any of the currently existing frameworks and infrastructures. Whereas there are some known algorithms dealing with the horizontal scaling complexity, such as

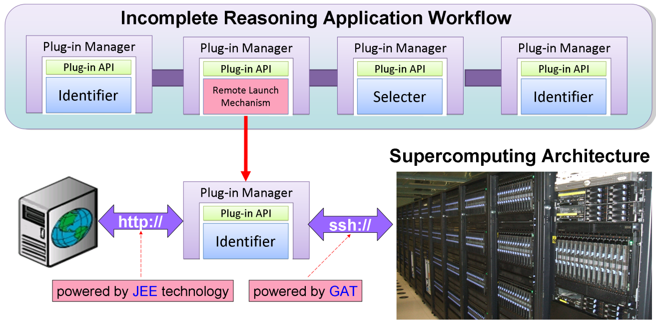


Figure 2. Execution of a reasoning application's workflow on a high performance computing system.

identification of the information subsets related to a specific problem, i.e., subsetting, the vertical scaling remains the major challenge for all existing algorithms.

Another essential property of the Big Data is complexity. Semantic applications must deal with rich ontological models describing complex domain knowledge, and at the same time highly dynamic data representing recent or relevant information, as produced by streaming or search-enabled data sources. A considerable part of the web data is produced as a result of automatic reasoning over streaming information from sensors, social networks, and other sources, which are highly unstructured, inconsistent, noisy and incomplete.

The availability of such an amount of complex data makes it attractive for Semantic Web applications to exploit High Performance Computing (HPC) infrastructures to effectively process the Big Data. There have been several pilot research projects aiming to enable the potential of supercomputing infrastructures to the Semantic Web application development. One of the prominent examples of such projects is the Large Knowledge Collider (LarKC), which is a software platform for large-scale incomplete reasoning. In particular, LarKC provides interfaces for loading off the computation-intensive part of a reasoning application's workflow to a supercomputing infrastructure (see Figure 2).

Both commodity and more dedicated HPC architectures, such as the Cray XMT [6], have been held in focus of the data-intensive Web applications. The XMT dedicated system, however, has proved successful only for a limited number of tasks so far, which is mainly due to the complexity of exploiting the offered software frameworks (mainly non-standard pragma-based C extensions).

Unfortunately, most Semantic Web applications are written in the Java programming language, whereas current frameworks that make the most out of HPC infrastructures, such as the Message Passing Interface (MPI), only target C or Fortran applications. MPI is a process-based parallelization strategy, which is a de-facto standard in the area of parallel computing for C, C++, and Fortran applications. Known alternative parallelization frameworks to MPI that conform with Java, such as Hadoop[7] or Ibis [8], prove to be scalable though but are not even nearly as efficient or well-developed as numerous open-source implementations of MPI, such as MPICH or Open MPI[9].

The implementation in Java has prevented MPI to be adopted by Semantic Web applications. However, given the vast data size addressed by the modern Web applications, and given the emergence of the new communities interested in adopting MPI, it seems natural to explore the benefits of MPI for Java applications on the HPC platforms as well. Introducing MPI to Java poses several challenges. First, the API set should be compliant with the MPI standard [9], but not downgrade the flexibility of the native Java language constructions. Second, the hardware support should be offered in a way that overcomes the limitation of the Java run-time environment, but meet such important requirements as thread-safety. Third, MPI support should be seamlessly integrated in the parallel application's execution environment. All of these three issues of functionality, adaptivity, and usability must complexly be addressed to make the use of MPI in Java applications practical and useful.

We look how to resolve the above-mentioned issues in a way that leverages the advances of the existing MPI frameworks. We present and evaluate our solution for introducing Java support in Open MPI [10], which is one of the most popular open source MPI-2 standard's implementations nowadays. Our approach is based on the integration of Java MPI bindings developed for mpiJava [11] directly in the native C realization of Open MPI, thus minimizing the bindings overhead and leveraging the Open MPI's run-time and development environment to ensure the high scalability of the Java parallel application. We also give examples of successful pilot scenarios implemented with our solution and discuss future work in terms of the development, implementation, and standardization activities.

II. Related Work

There are only a few alternatives to MPI in introducing the large-scale parallelism to Java applications. The most promising among those alternatives in terms of the performance and usability are solutions offered by IBIS/JavaGAT and MapReduce/Hadoop.

IBIS [12] is a middleware stack used for running Java applications in distributed and heterogeneous computing environments. IBIS leverages the peer-to-peer communication technology by means of the proprietary Java RMI (Remote Memory Invocation) implementation, based on GAT (Grid Application Toolkit) [13]. The Java realization of GAT (JavaGAT) is a middleware stack that allows the Java application to instatiate its classes remotely on the network-connected resource, i.e., a remote Java Virtual Machine. Along with the traditional access protocols. e.g., telnet or ssh, the advanced access protocols, such as ssh-pbs for clusters with PBS(cluster Portable Batch System)-like job scheduling or gsissh for grid infrastructures are supported. IBIS implements a mechanism of multiple fork-joins to detect and decompose the application's workload and execute its parts concurrently on distributed machines. While [8] indicates some successful Java applications implemented with IBIS/JavaGAT and shows a good performance, there is no clear evidence about the scalability of this solution for more complex communication patterns, involving nested loops or multiple split-joins. Whereas IBIS is a very effective solution for the distributed computing environments, e.g., Grid or Cloud, it is definitively not the best approach to be utilized on the tightly-coupled productional clusters.

The MapReduce framework [14] and its most prominent implementation in Java, Hadoop, has got a tremendous popularity in modern data-intensive application scenarios. MapReduce is a programming model for data-centric applications exploiting large-scale parallelism, originally introduced by Google in its search engine. In MapReduce, the application's workflow is divided into three main stages (see Figure 3): map, process, and reduce. In the map stage, the input data set is split into independent chunks and each of the chunks is assigned to independent tasks, which are then processed in a completely parallel manner (process stage). In the reduce stage, the output produced by every map task is collected, combined and the consolidated final output is then produced. The Hadoop framework is a service-based implementation of MapReduce for Java. Hadoop considers a parallel system as a set of master and slave nodes, deploying on them services for scheduling tasks as jobs (Job Tracker), monitoring the jobs (Task Tracker), managing the input and output data (Data Node), re-executing the failed tasks, etc. This is done in a way that ensures a very high service reliability and fault tolerance properties of the parallel execution. In Hadoop, both the input and the output of the job are stored in a special distributed file-system. In order to improve the reliability, the file system also provides an automatic replication procedure, which however introduces an additional overhead to the inter-node communication. Due to this overhead, Hadoop provides much poorer performance than MPI, however offering better QoS characteristics related to the reliability and fault-tolerance. Since MPI and MapReduce paradigms have been designed to serve different purposes, it is hardly possible to comprehensively compare them. However they would obviously benefit from a cross-fertilization; e.g., MPI could serve a high-performance communication layer to Hadoop, which might help improve the performance by omitting the disk I/O usage for distributing the map and gathering the reduce tasks across the compute nodes.

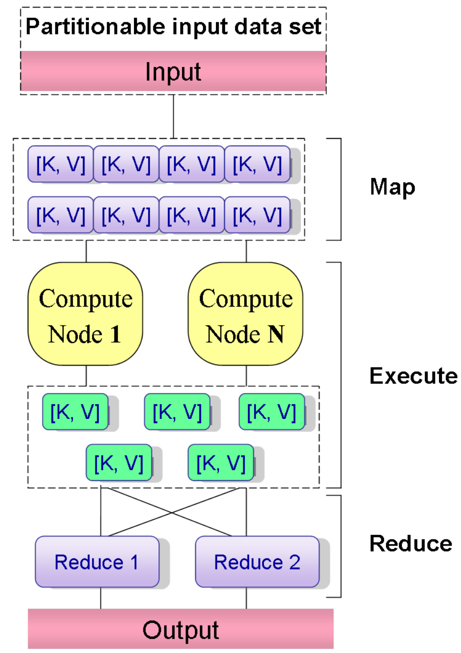


Figure 3. MapReduce processing schema.

III. Data-Centric Parallelization and MPI

By "data-centric parallelization" we mean a set of techniques for: (i) identification of non-overlapping application's dataflow regions and corresponding to them instructions; (ii) partitioning the data into subsets; and (iii) parallel processing of those subsets on the resources of the high performance computing system. For Semantic Web applications utilizing the data in such well-established formats as RDF [15], parallelization relies mainly on partitioning (decomposing) the RDF data set on the level of statements (triples), see Figure 4a. The ontology data (also often referred as *tbox*) usually remains unpartitioned as its size is relatively small as compared with the actual data (*abox*), so that it is just replicated among all the compute nodes.

The Message-Passing Interface (MPI) is a process-based standard for parallel applications implementation. MPI processes are independent execution units that contain their own state information, use their own address spaces, and only interact with each other via interprocess communication mechanisms defined by MPI. Each MPI process can be executed on a dedicated compute node of the high performance architecture, i.e., without competing with the other processes in accessing the hardware, such as CPU and RAM, thus improving the application performance and achieving the algorithm speed-up. In case of the shared file system, such as Lustre [16], which is the most utilized file system standard of the modern HPC infrastructures, the MPI processes can effectively access the same file section in parallel without any considerable disk I/O bandwidth degradation. With regard to the data decomposition strategy presented in Figure 4a, each MPI process is responsible for processing the data partition assigned to it proportionally to the total number of the MPI processes (see Figure 4b). The position of any MPI process within the group of processes involved in the execution is identified by an integer *R* (rank) between 0 and *N*-1, where *N* is a total number of the launched MPI processes. The rank *R* is a unique integer identifier assigned incrementally and sequentially by the MPI run-time environment to every process. Both the MPI process's rank and the total number of the MPI processes can be acquired from within the application by using MPI standard functions, such as presented in Listing 1.

```
import java.io.*;
import mpi.*;

class Hello {
  public static void main(String[] args) throws
      MPIException
  {
    int my_pe, npes; // rank and overall number of MPI
        processes
    int N;       // size of the RDF data set (number of
        triples)

    MPI.Init(args); // intialization of the MPI RTE

    my_pe = MPI.COMM_WORLD.Rank();
    npes  = MPI.COMM_WORLD.Size();

    System.out.println("Hello␣from␣MPI␣process" + my_pe +
        "␣out␣of␣" + npes);
    System.out.println("I'm␣processing␣the␣RDF␣triples␣
        from␣" + my_pe/npes + "␣to␣" + (my_pe+1)/npes);

    MPI.Finalize(); // finalization of the MPI RTE
  }
}
```

Listing 1. Acquiring rank and total number of processes in a simple MPI application.

The typical data processing workflow with MPI can be depicted as shown in Figure 5. The MPI jobs are executed by means of the *mpirun* command, which is an important part of any MPI implementation. *mpirun* controls several aspect of parallel program execution, in particular launches MPI processes under the job scheduling manager software like OpenPBS [17]. The number of MPI processes to be started is provided with the *-np* parameter to *mpirun*. Normally, the number of MPI processes corresponds to the number of the compute nodes, reserved for the execution of parallel job. Once the MPI process is started, it can request its rank as well as the total number of the MPI processes associated with the same job. Based on the rank and total processes number, each MPI process can calculate the corresponding subset of the input data and process it. The data partitioning problem remains beyond the scope of this work; particularly for RDF, there is a number of well-established approaches discussed in several previous publications, e.g., horizontal [18], vertical [19], and workload driven [20] partitioning.

Since a single MPI process owns its own memory space and thus can not access the data of the other processes directly, the MPI standard foresees special communication functions, which are necessary, e.g., for exchanging the data subdomain's boundary values or consolidating the final output from the partial results produced by each of the processes. The MPI processes communicate with each other

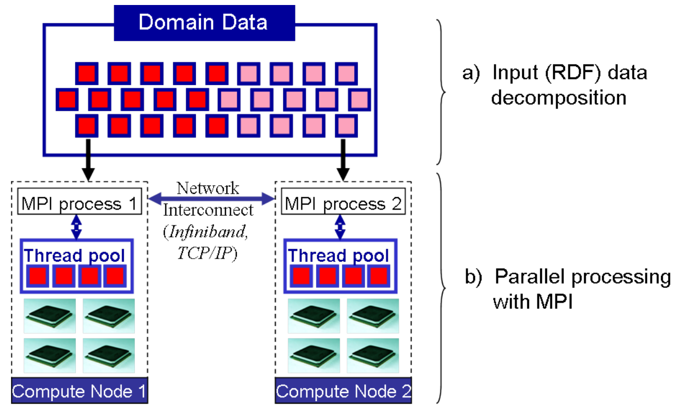


Figure 4. Data decomposition and parallel execution with MPI.

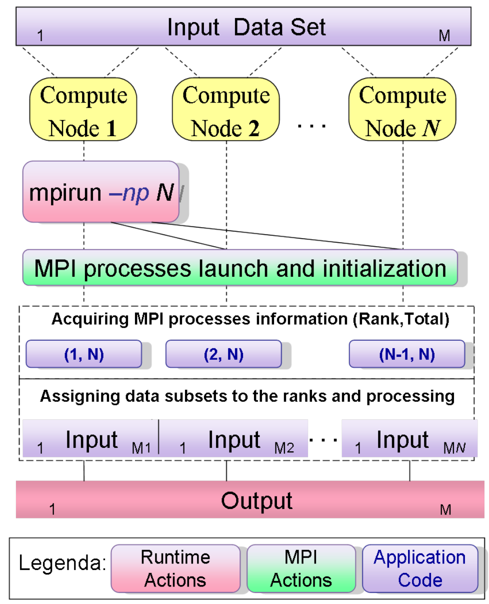


Figure 5. Typical MPI data-centric application's execution workflow.

by sending messages, which can be done either in "point-to-point"(between two processes) or collective way (involving a group of or all processes).

More details about the MPI communication can also be found in our previous publication [21].

IV. Open MPI Java Bindings

This section discusses implementation details of Java bindings for the Open MPI library.

A. *MPI bindings for Java*

Although the official MPI standard's bindings are limited to C and Fortran languages, there has been a number of standardization efforts made towards introducing the MPI bindings for Java. The most complete API set, however, has been proposed by mpiJava [22] developers.

There are only a few approaches to implement MPI bindings for Java. These approaches can be classified in two following categories:

- Pure Java implementations, e.g., based on RMI (Remote Method Invocation) [23], which allows Java objects residing in different virtual machines to communicate with each other, or lower-level Java sockets API.
- Wrapped implementations using the native methods implemented in C languages, which are presumably more efficient in terms of performance than the code managed by the Java run-time environment.

In practice, none of the above-mentioned approaches satisfies the contradictory requirements of the Web users on application portability and efficiency. Whereas the pure Java implementations, such as MPJ Express [24] or MPJ/Ibis [8], do not benefit from the high speed interconnects, e.g., InfiniBand [25], and thus introduce communication bottlenecks and do not demonstrate acceptable performance on the majority of today's production HPC systems [26], a wrapped implementation, such as mpiJava [27], requires a native C library, which can cause additional integration and interoperability issues with the underlying MPI implementation.

In looking for a trade-off between the performance and the usability, and also in view of the complexity of providing Java support for high speed cluster interconnects, the most promising solution seems to be to implement the Java bindings directly in a native MPI implementation in C.

B. Native C Implementation

Despite a great variety of the native MPI implementations, there are only a few of them that address the requirements of Java parallel applications on process control, resource management, latency awareness and management, and fault tolerance. Among the known sustainable open-source implementations, we identified Open MPI[28] and MPICH2[29] as the most suitable to our goals to implement the Java MPI bindings. Both Open MPI and MPICH2 are open-source, production quality, and widely portable implementations of the MPI standard (up to its latest 2.0 version). Although both libraries claim to provide a modular and easy-to-extend framework, the software stack of Open MPI seems to better suit the goal of introducing a new language's bindings, which our research aims to. The architecture of Open MPI [10] is highly flexible and defines a dedicated layer used to introduce bindings, which are currently provided for C, F77, F90 and some other languages (see also Figure 7). Extending the OMPI-Layer of Open MPI with the Java language support seems to be a very promising approach to the the discussed integration of Java bindings, taking benefits of all the layers composing Open MPI's architecture.

C. Design and Implementation in Open MPI

We have based our Java MPI bindings on the *mpiJava* code, originally developed in HPJava[30] project and currently maintained by the High Performance Computing Center Stuttgart[31]. mpiJava provides a set of Java Native Interface (JNI) wrappers to the native MPI v.1.1 communication methods, as shown in Figure 6. JNI enables the programs running inside a Java run-time environment to invoke native C code and thus use platform-specific features and libraries [32], e.g., the InfiniBand software stack. The application-level API is constituted by a set of Java classes, designed in conformance to the MPI v.1.1 and the specification in [22]. The Java methods internally invoke the MPI-C functions using the JNI stubs. The realization details for mpiJava can be obtained from [11][33].

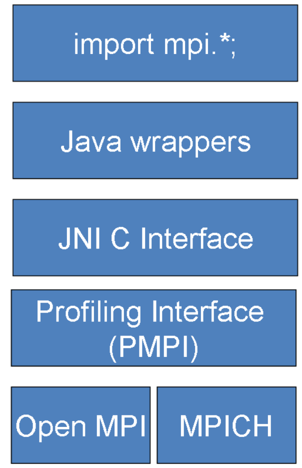


Figure 6. mpiJava architecture.

Open MPI is a high performance, production quality, MPI-2 standard compliant implementation. Open MPI consists of three combined abstraction layers that provide a full featured MPI implementation: (i) OPAL (Open Portable Access Layer) that abstracts from the peculiarities of a specific system away to provide a consistent interface adding portability; (ii) ORTE (Open Run-Time Environment) that provides a uniform parallel run-time interface regardless of system capabilities; and (iii) OMPI (Open MPI) that provides the application with the expected MPI standard interface. Figure 7 shows the enhanced Open MPI architecture, enabled with the Java bindings support.

The major integration tasks that we performed were as follows:

- extend the Open MPI architecture to support Java bindings,
- extend the previously available mpiJava bindings to MPI-2 (and possibly upcoming MPI-3) standard,
- improve the native Open MPI configuration, build, and execution system to seamlessly support the Java bindings,
- redesign the Java interfaces that use JNI in order to better conform to the native realization,
- optimize the JNI code to minimize its invocation over-

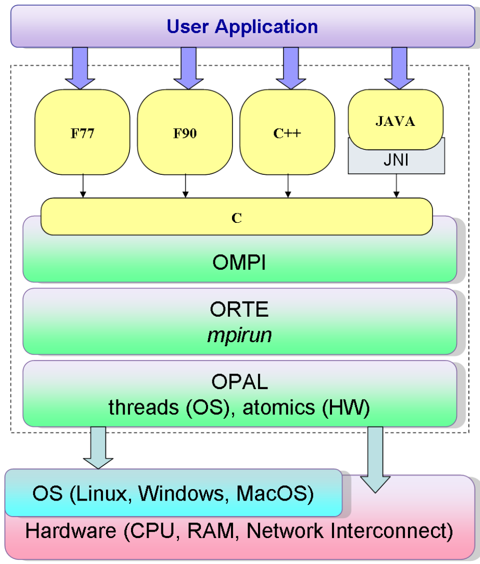


Figure 7. Open MPI architecture.

head,

- and create test applications for performance benchmarking.

Both Java classes and JNI code for calling the native methods were integrated into Open MPI. However, the biggest integration effort was required at the OMPI (Java classes, JNI code) and the ORTE (run-time specific options) levels. The implementation of the Java class collection followed the same strategy as for the C++ class collection, for which the opaque C objects are encapsulated into suitable class hierarchies and most of the library functions are defined as class member methods. Along with the classes implementing the MPI functionality (MPI package), the collection includes the classes for error handling (Errhandler, MPIException), datatypes (Datatype), communicators (Comm), etc. More information about the implementation of both Java classes and JNI-C stubs can be found in previous publications [11][26].

D. *Performance*

In order to evaluate the performance of our implementation, we prepared a set of Java benchmarks based on those well-recognized in the MPI community, such as NetPIPE [34] or NAS [35]. Based on those benchmarks, we compared the performance of our implementation based on Open MPI and the other popular implementation (MPJ Express) that follows a "native Java" approach. Moreover, in order to evaluate the JNI overhead, we reproduced the benchmarks also in C and ran them with the native Open MPI. Therefore, the following three configurations were evaluated:

- **ompiC** - native C implementation of Open MPI (the actual trunk version), built with the GNU compiler (v.4.6.1),

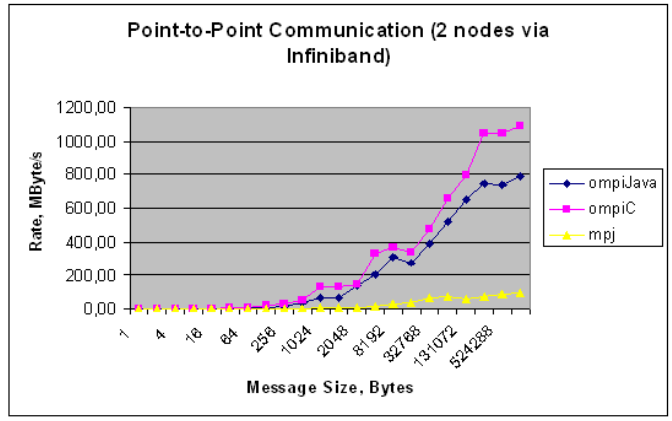


Figure 8. Message rate for the point-to-point communication.

- **ompiJava** - our implementation of Java bindings on top of *ompiC*, running with Java JDK (v.1.6.0), and
- **mpj** - the newest version of MPJ Express (v.0.38), a Java native implementation, running with the same JDK.

We examined two types of communication: point-to-point (between two nodes) and collective (between a group of nodes), varying the size of the transmitted messages. We did intentionally not rely on the previously reported benchmarks[36] in order to eliminate the measurement deviations that might be caused by running tests in a different hardware or software environment. Moreover, in order to ensure a fair comparison between all these three implementations, we ran each test on the absolutely same set of compute nodes.

The point-to-point benchmark implements a "ping-pong" based communication between two single nodes; each node exchanges the messages of growing sizes with the other node by means of blocking Send and Receive operations. As expected, our *ompiJava* implementation was not as efficient as the underlying *ompiC*, due to the JNI function calls overhead, but showed much better performance than the native Java based *mpj* (Figure 8). Regardless of the message size, *ompiJava* achieves around eight times higher throughput than *mpj* (see Figure 9).

The collective communication benchmark implements a single blocking message gather from all the involved nodes. Figure 10 shows the results collected for $P = 2^k$ (where k=2-7) nodes, with a varying size of the gathered messages. The maximal size of the aggregated data was 8 GByte on 128 nodes. Figure 11 demonstrates the comparison of collective gather performance for all tested implementations on the maximal number of the available compute nodes (128). Whereas the InfiniBand-aware *ompiJava* and *ompiC* scaled quite well, the native Java based *mpj* has shown very poor performance; for the worst case (on 128 nodes) a slow-down up to 30 times compared with *ompiJava* was observed.

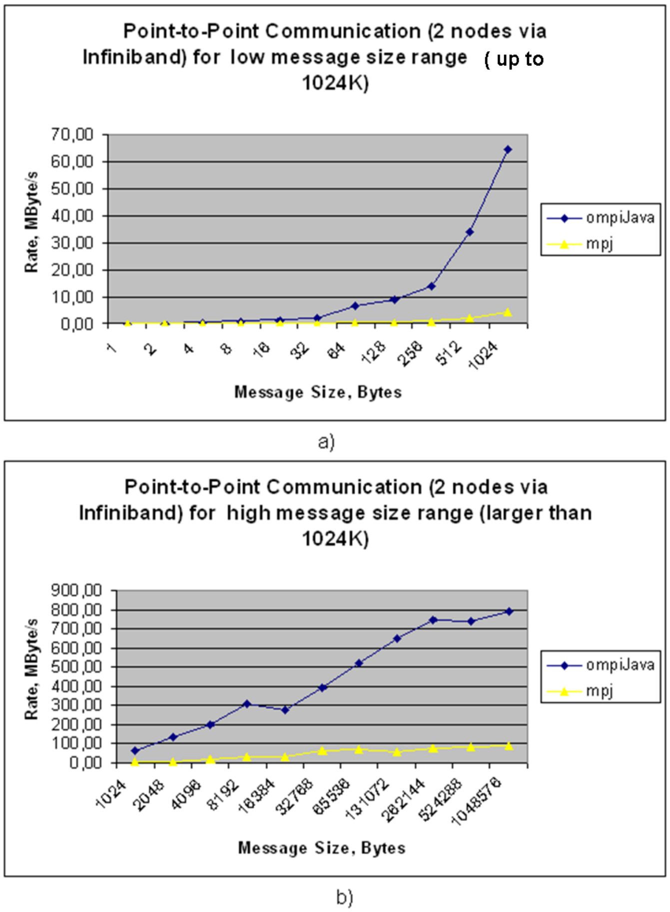


Figure 9. Comparison of the message rate for ompiJava and mpj for a) low and b) high message size range.

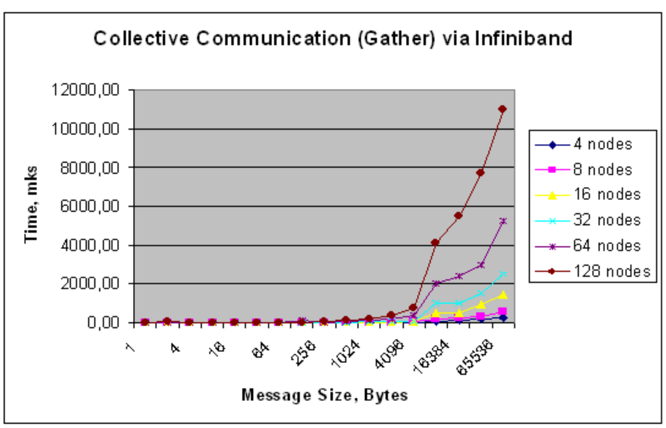


Figure 10. Collective gather communication performance of ompiJava.

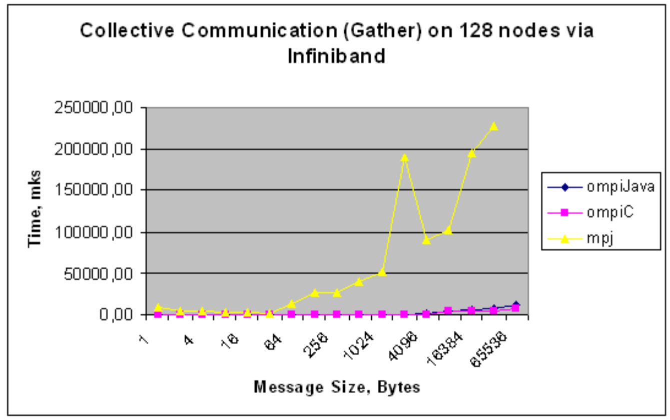


Figure 11. Collective gather communication performance on 128 nodes.

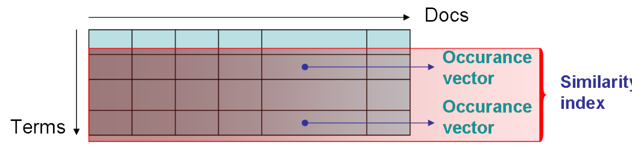


Figure 12. Similarity index computation in a document collection.

V. MPI IMPLEMENTATION OF RANDOM INDEXING

Random indexing [37] is a word-based co-occurrence statistics technique used in resource discovery to improve the performance of text categorization. Random indexing offers new opportunities for a number of large-scale Web applications performing the search and reasoning on the Web scale [38]. We used Random Indexing to determine the similarity index (based on the words' co-occurance statistic) between the terms in a closed document collection, such as Wikipedia or Linked Life Data (see Figure 12).

The main challenges of the Random Indexing algorithms lay in the following:

- Huge and high-dimensional vector space. A typical random indexing search algorithm performs traversal over all the entries of the vector space. This means, that the size of the vector space to the large extent determines the search performance. The modern data stores, such as Linked Life Data or Open PHACTS consolidate many billions of statements and result in vector spaces of a very large dimensionality. Performing Random indexing over such large data sets is computationally very costly, with regard to both execution time and memory consumption. The latter poses a hard constraint to the use of random indexing packages on the serial mass computers. So far, only relatively small parts of the Semantic Web data have been indexed and analyzed.
- High call frequency. Both indexing and search over the vector space is highly dynamic, i.e., the entire indexing

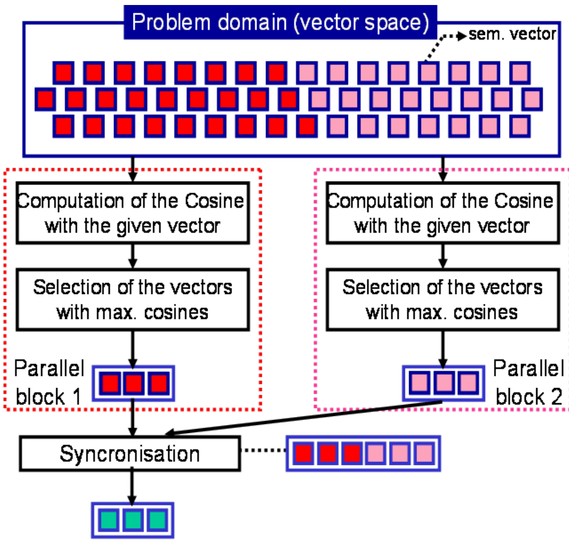
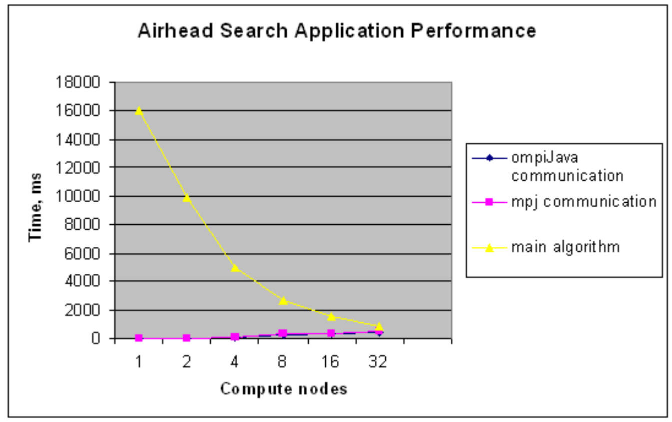
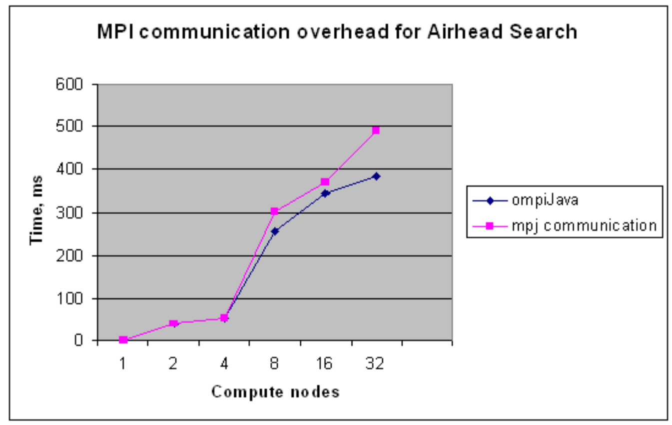


Figure 13. MPI-based parallel implementation of Airhead Search.



a)



b)

Figure 14. Airhead performance with ompiJava and mpj.

process repeats from scratch every time new data is encountered.

In our previous work [39], we have already reported on the efforts done on parallelizing the search operation of Airhead - an open source Java implementation of Random Indexing algorithm. Our MPI implementation of the Airhead search is based on a domain decomposition of the analyzed vector space and involves both point-to-point and collective gather and broadcast MPI communication (see the schema in Figure 13). In our current work, we evaluated the MPI version of Airhead with both *ompijava* and *mpj* implementations.

We performed the evaluation for the largest of the available data sets reported in [39] (namely, Wiki2), which comprises 1 Million of high density documents and occupies 16 GByte disk storage space. The overall execution time (wall clock) was measured. Figure 14a shows that both *ompijava* and *mpj* scale well until the problem size is large enough to saturate the capacities of a single node. Nevertheless, our implementation was around 10% more efficient over *mpj* (Figure 14b).

VI. Performance Analysis And Optimization Tools

Development of parallel communication patterns with MPI is quite a nontrivial task, in particular for large-scale use cases, which consist of hundreds and even thousands of parallel processes. The synchronization among the MPI processes of the parallel application can be a key performance concern. Among the typical problems the following appear most frequently:

- non-optimal balancing of the MPI processes load (i.e., wrong data decomposition),
- misconfiguration of the communication pattern preventing the applications scalability to the growing number of compute nodes,
- incorrect usage of the MPI communication functions (e.g., when point-to-point communication are used instead of the collective ones, which lowers the performance and also prevents the scalability).

One of the advantages of the C-based Java binding implementation as compared with the "native-Java" approach is the possibility to use numerous performance optimization tools available for the traditional HPC applications. This is leveraged by the special profiling interface provided by the MPI standard - PMPI (see Figure 6). Using PMPI, performance analysis tools can inject the measurement code directly in the parallel application's object file and capture and aggregate statistics about the application execution at run-time. Among the parameters measured with PMPI are duration of a single MPI communication, total number of communications, processes that are involved in the communication, etc. The profiling code is dynamically linked with the MPI library and thus does not require any changes in either the application code or the MPI library. The captured events are stored in trace files using a special format, such as OTF - the Open Trace Format, which can then be

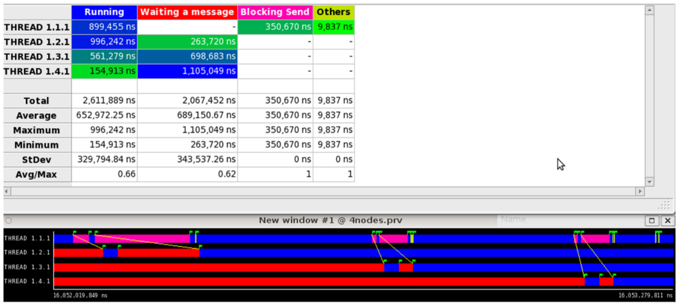


Figure 15. MPI Global Broadcast Communication visualization for four MPI processes with Paraver.

analyzed in order to retrieve and visualize the application's communication profile.

In our pilot investigations, we evaluated the ability of the *Extrae* [40] profiling library, developed by the Barcelona Supercomputing Center, to collect event traces of the MPI-parallelized Airhead Search application. For this purpose, we linked *Extrae* with our Java-enabled version of Open MPI and run the instrumented version of Airhead on the cluster. The traces collected as result of the execution were visualized with the Paraver [41] tool (see Figure 15), similar to any other MPI application in C or Fortran.

VII. Future Work

Our future work will concentrate on promoting both MPI standard and our ompiJava implementation to Semantic Web applications as well as improving the current realization of Java bindings in Open MPI.

With regard to promotion activities, we will be introducing our data-centric and MPI-based parallelization approach to further challenging data-intensive applications, such as Reasoning [42]. Regarding this application, there are highly successful MPI imlementations in C, e.g., the parallel RDFS graph closure materialization presented in [43], which are indicatively much more preferable over all the existing Java solutions in terms of performance. Our implementation will allow the developed MPI communication patterns to be integrated in existing Java-based codes, such as Jena [2] or Pellet [44], and thus drastically improve the competitiveness of the Semantic Web application based on such tools.

The development activities will mainly focus on extending the Java bindings to the full support of the MPI-3 specification. We will also aim at adding Java language-specific bindings into the MPI standard, as a reflection of the Semantic Web value in supercomputing.

The integration activities will concentrate on adapting the performance analysis tools to the specific of Java applications. Unfortunately, the existing performance analysis tools, such as Extrae discussed in the previous section, does not provide a deep insight in the intrinsic characteristics of the Java Virtual Machine, which however might be as important for the application performance optimization as the communication profile tailoring. For this purpose, the traditional performance analysis tools for the Java applications, such as ones provided by the Eclipse framework, must be extended with the communication profiling capabilities. Several EU-projects, such as JUNIPER, are already working in this direction.

VIII. Conclusion

High Performance Computing is a relatively new trend for the Semantic Web, which however has gained a tremendous popularity thanks to the recent advances in developing data-intensive applications.

The Message Passing Interface provides a very promising approach for developing parallel data-centric applications. Unlike its prominent alternatives, the MPI functionality is delivered on the library-level, and thus does not require any considerable development efforts to parallelize an existing serial application. Apart from a very flexible parallelization strategy, which foresees a number of parallelization options, either on the code, data, or both levels, but also delivers a very efficient communication mechanism, which takes full advantages of the modern supercomputing communication networks. Using MPI, the Semantic Web applications can enjoy the full backing of the high performance computing architectures. We would like to point out, that the current work is in no case an attempt to undermine the value of data-centric parallel implementations (like Hadoop), nor it is a replacement for any current data processing infrastructures. However many of the current parallel data processing systems can benefit from adopting MPI and ompiJava offers a set of good tools for this.

We introduced a new implementation of Java bindings for MPI that is integrated in one of the most popular open source MPI-2 libraries - Open MPI. The integration allowed us to deliver a unique software environment for flexible development and execution of parallel MPI applications, integrating the Open MPI framework's capabilities, such as portability and usability, with those of mpiJava, such as an extensive set of Java-based API for MPI communication. We evaluated our implementation for Random Indexing, which is one of the most challenging Semantic Web applications in terms of the computation demands currently. The evaluation has confirmed our initial considerations about the high efficiency of MPI for parallelizing Java applications. In the following, we are going to investigate further capabilities of MPI for improving the performance of data-centric applications, in particular by means of MPI-IO (MPI extension to support efficient file input-output). We will also concentrate on promoting the MPI-based parallelization strategy to the other challenging and performance-demanding applications, such as Reasoning. We believe that our implementation of Java bindings of MPI will attract Semantic Web development community to increase the scale of both its serial and parallel applications. The successful pilot application implementations done based on MPI, such as materialization of the finite RDFS closure presented in [43], offer a very promising outlook regarding the future perspectives of MPI in the Semantic Web domain.

ACKNOWLEDGMENT

Authors would like to thank the Open MPI consortium for the support with porting mpiJava bindings, to the EU-ICT HPC-Europa project for granting access to the computing facilities, as well as the EU-ICT JUNIPER project for the support with the Java platform and parallelization.

REFERENCES

[1] A. Cheptsov, "Enabling high performance computing for semantic web applications by means of open mpi java bindings," in *Proc. the Sixth International Conference on Advances in Semantic Processing (SEMAPRO 2012) Conference*, Barcelona, Spain, 2012.

[2] P. McCarthy. Introduction to jena. IBM developerWorks. [Online]. Available: http://www.ibm.com/developerworks/xml/library/j-jena [retrieved: January, 2013]

[3] (2011) Two kinds of big data. R. Gonzalez. [Online]. Available: http://semanticweb.com/two-kinds-of-big-dat_b21925

[4] Lod cloud diagram. [Online]. Available: http://richard.cyganiak.de/2007/10/lod/ [retrieved: January, 2013]

[5] R. Gonzalez. (2012) Closing in on a million open government data sets. [Online]. Available: http://semanticweb.com/closinginona-millionopen-governmentdatasets_b29994 [retrieved: January, 2013]

[6] E. Goodman, D. J. Haglin, C. Scherrer, D. Chavarria, J. Mogill, and J. Feo, "Hashing strategies for the cray xmt," in *Proc. 24th IEEE Int. Parallel and Distributed Processing Symp.*, 2010.

[7] Hadoop mapreduce framework homepage. [Online]. Available: http://hadoop.apache.org/mapreduce [retrieved: January, 2013]

[8] M. Bornemann, R. van Nieuwpoort, and T. Kielmann, "Mpj/ibis: A flexible and efficient message passing platform for java," *Concurrency and Computation: Practice and Experience*, vol. 17, pp. 217–224, 2005.

[9] (1995) Mpi: A message-passing interface standard. Message Passing Interface Forum. [Online]. Available: http://www.mcs.anl.gov/research/projects/mpi/mpi-standard/mpi-report-1.1/mpi-report.htm [retrieved: January, 2013]

[10] E. G. et al., "Open MPI: Goals, concept, and design of a next generation MPI implementation," in *Proc., 11th European PVM/MPI Users' Group Meeting*, Budapest, Hungary, September 2004, pp. 97–104.

[11] M. Baker, B. Carpenter, G. Fox, S. Ko, and S. Lim, "mpiJava: An object-oriented java interface to mpi," in *Proc. International Workshop on Java for Parallel and Distributed Computing IPPS/SPDP*, San Juan, Puerto Rico, 1999.

[12] R. van Nieuwpoort, J. Maassen, G. Wrzesinska, R. Hofman, C. Jacobs, T. Kielmann, and H. Bal, "Ibis: a flexible and efficient java based grid programming environment," *Concurrency and Computation: Practice and Experience*, vol. 17, pp. 1079–1107, June 2005.

[13] R. van Nieuwpoort, T. Kielmann, and H. Bal, "User-friendly and reliable grid computing based on imperfect middleware," in *Proc. ACM/IEEE Conference on Supercomputing (SC'07)*, November 2007.

[14] J. Dean and S. Ghemawat, "Mapreduce- simplified data processing on large clusters," in *Proc. OSDI04: 6th Symposium on Operating Systems Design and Implementation*, 2004.

[15] (2004, February) Resource description framework (RDF). RDF Working Group. [Online]. Available: http://www.w3.org/RDF/ [retrieved: January, 2013]

[16] "Lustre file system - high-performance storage architecture and scalable cluster file system," White Paper, SunMicrosystems, Inc., December 2007.

[17] Portable batch systems. [Online]. Available: http://en.wikipedia.org/wiki/Portable_Batch_System [retrieved: January, 2013]

[18] A. Dimovski, G. Velinov, and D. Sahpaski, "Horizontal partitioning by predicate abstraction and its application to data warehouse design," in *ADBIS*, 2010, pp. 164–175.

[19] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach, "Scalable semantic web data management using vertical partitioning," in *Proc. The 33rd international conference on Very large data bases (VLDB'07))*.

[20] C. Curino, E. P. C. Jones, S. Madden, and H. Balakrishnan, "Workload-aware database monitoring and consolidation," in *SIGMOD Conference*, 2011, pp. 313–324.

[21] A. Cheptsov, M. Assel, B. Koller, R. Kbert, and G. Gallizo, "Enabling high performance computing for java applications using the message-passing interface," in *Proc. The Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering (PARENG'2011)*.

[22] B. Carpenter, G. Fox, S.-H. Ko, and S. Lim, "mpiJava 1.2: Api specification," Northeast Parallel Architecture Center. Paper 66, 1999. [Online]. Available: http://surface.syr.edu/npac/66 [retrieved: January, 2013]

[23] T. Kielmann, P. Hatcher, L. Boug, and H. Bal, "Enabling java for high-performance computing: Exploiting distributed shared memory and remote method invocation," *Communications of the ACM*, 2001.

[24] M. Baker, B. Carpenter, and A. Shafi, "MPJ Express: Towards thread safe java hpc," in *Proc. IEEE International Conference on Cluster Computing (Cluster'2006)*, September 2006.

[25] R. K. Gupta and S. D. Senturia, “Pull-in time dynamics as a measure of absolute pressure,” in *Proc. IEEE International Workshop on Microelectromechanical Systems (MEMS’97)*, Nagoya, Japan, Jan. 1997, pp. 290–294.

[26] G. Judd, M. Clement, Q. Snell, and V. Getov, “Design issues for efficient implementation of mpi in java,” in *Proc. the 1999 ACM Java Grande Conference*, 1999, pp. 58–65.

[27] B. Carpenter, V. Getov, G. Judd, A. Skjellum, and G. Fox, “MPJ: Mpi-like message passing for java,” *Concurrency and Computation - Practice and Experience*, vol. 12(11), pp. 1019–1038, 2000.

[28] Open mpi homepaga. [Online]. Available: http://www.open-mpi.org [retrieved: January, 2013]

[29] Mpich2 project website. [Online]. Available: http://www.mcs.anl.gov/research/projects/mpich2/ [retrieved: January, 2013]

[30] Hpjava project website. [Online]. Available: http://www.hpjava.org [retrieved: January, 2013]

[31] mpijava website. [Online]. Available: http://sourceforge.net/projects/mpijava/ [retrieved: January, 2013]

[32] S. Liang, Ed., *Java Native Interface: Programmer’s Guide and Reference*. Addison-Wesley, 1999.

[33] M. Vodel, M. Sauppe, and W. Hardt, “Parallel high-performance applications with mpi2java - a capable java interface for mpi 2.0 libraries,” in *Proc. The 16th Asia-Pacific Conference on Communications (APCC)*, Nagoya, Japan, 2010, pp. 509–513.

[34] Net pipe parallel benchmark website. [Online]. Available: http://www.scl.ameslab.gov/netpipe/ [retrieved: January, 2013]

[35] Nas parallel benchmark website. [Online]. Available: http://sourceforge.net/projects/mpijava/ [retrieved: January, 2013]

[36] Mpj express benchmarking results. [Online]. Available: http://mpj-express.org/performance.html [retrieved: January, 2013]

[37] M. Sahlgren, “An introduction to random indexing,” in *Proc. Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE)’2005*, 2005, pp. 1–9.

[38] D. Jurgens, “The S-Space package: An open source package for word space models,” in *Proc. the ACL 2010 System Demonstrations*, 2010, pp. 30–35.

[39] M. Assel, A. Cheptsov, B. Czink, D. Damljanovic, and J. Quesada, “Mpi realization of high performance search for querying large rdf graphs using statistical semantics,” in *Proc. The 1st Workshop on High-Performance Computing for the Semantic Web*, Heraklion, Greece, May 2011.

[40] Extrae performance trace generation library homepage. [Online]. Available: http://www.bsc.es/computer-sciences/extrae [retrieved: January, 2013]

[41] Extrae performance trace generation library homepage. [Online]. Available: http://www.bsc.es/computer-sciences/performance-tools/paraver [retrieved: January, 2013]

[42] D. Fensel and F. van Harmelen, “Unifying reasoning and search to web scale,” *IEEE Internet Computing*, vol. 11(2), pp. 95–96, 2007.

[43] J. Weaver and J. A. Hendler, “Parallel materialization of the finite rdfs closure for hundreds of millions of triples,” in *Proc. International Semantic Web Conference (ISWC) 2009*, A. B. et al., Ed., 2009.

[44] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. Pellet: a practical owl-dl reasoner. Journal of Web Semantics. [Online]. Available: http://www.mindswap.org/papers/PelletJWS.pdf [retrieved: January, 2013]