Towards an Industry Data Gateway: An Integrated Platform for the Analysis of Wind Turbine Data

Alvaro Aguilera*, Richard Grunzke*, Ulf Markwardt
Center for Information Services and
High Performance Computing (ZIH)
Technische Universität Dresden
01062 Dresden, Germany
*{alvaro.aguilera, richard.grunzke}@tu-dresden.de

Dirk Habich
Database Systems Group
Technische Universität Dresden
01062 Dresden, Germany

Dirk Schollbach
Bosch Rexroth Monitoring Systems GmbH
Else-Sander-Str. 8
01099 Dresden, Germany

Jochen Garcke
Fraunhofer SCAI
Schloss Birlinghoven
53754 Sankt Augustin, Germany and
Institut für Numerische Simulation
Universität Bonn
53115 Bonn, Germany

***Abstract*—The increasing amount of data produced in many scientific and engineering domains creates as many new challenges for an efficient data analysis, as possibilities for its application. In this paper, we present one of the use-cases of the project VAVID, namely the condition monitoring of sensor information from wind turbines, and how a data gateway can help to increase the usability and security of the proposed system. Starting by briefly introducing the project, the paper presents the problem of handling and processing large amount of sensor data using existing tools in the context of wind turbines. It goes on to describe the innovative approach used in VAVID to meet this challenge, covering the main goals, numerical methods used for analysis, the storage concept, and architectural design. It concludes by offering a rational for the use of a data gateway as the main entry point to the system and how this is being implemented in VAVID.**

I. Introduction

The partners in the project VAVID [1] are developing methods to tackle the enormous volumes of data that accumulate at engineering departments. Examples of such data include simulation results and the sensor data received from machines and installations. We aim to develop improved techniques for data compression as well as new methods of data analysis, data management and interactive data visualization. This saves on the costs of data storage and creates the transparency needed by engineers to optimize both production and products.

Before fabrication begins, it is essential to computationally analyze the product's characteristics in a way that mirrors reality as faithfully as possible. The computations and high performance computing (HPC) systems required for this task are generating an ever growing mountain of data, in particular we study data from numerical simulations of car crashes and from wind turbine design. An exponential rise in data volumes is also being seen due to the acquisition of sensor data during the operation of machines and plant. These measurement data allow engineers to draw important conclusions on how well control systems are working and how they can further optimize production. The real life data we are investigating are measurement data taken from wind turbine monitoring systems.

Just archiving the huge masses of data pose great challenges to technology companies. Moreover, important information carried by data is frequently not recognized because the company does not have the necessary data extraction methods at its disposal. By performing joint and comparative analysis of data from different industries, we aim to develop methodologies for efficient data analysis across application domains. These methods and techniques are going into the creation of a high performance data management system that will allow centralized data storage as well as efficient data access and retrieval.

One of the problems arising from the use of complex HPC infrastructure to process industrial data is that most users are unfamiliar with working on such platforms. Relying on command line interfaces and dealing with batch systems is often found to be tedious and sometimes even challenging. VAVID works around this issue by: (1) using a workflow engine to take care of the execution and monitoring of different algorithmic compositions created by the users to analyze the data, and (2) offering a data gateway as the central entry point to the system, with which the users can create, launch, and evaluate such compositions in a user-friendly manner.

In the following sections we present the current stage of our ongoing effort.

II. Related Work

Science and data gateways are build to efficiently integrate complex underlying infrastructures to enable novel research methods and make these available in a user-friendly way. The data gateway described here will integrate the UNICORE middleware [2] that manages computing and data resources. It is mature, flexible, continuously being developed, and used in major research infrastructures such as PRACE [3], XSEDE [4], and the Human Brain Project [5]. UNICORE consists of the target system, middleware, and client layers. The first integrates underlying HPC systems, the second contains services that manage jobs and data, and the third offers various client interfaces. The libraries of the command line client are used in gUSE (grid User Support Environment) [6] to provide access to HPC systems via UNICORE. gUSE is a high-level middleware that provides web services for workflow management, execution and sharing, and integrates with cloud and distributed computing infrastructures. WS-PGRADE [6] provides graphical user interfaces to gUSE for workflow creation, modification, execution, monitoring, and related data management. Together, gUSE and WS-PGRADE form a science gateway framework that runs within the open source Liferay portal framework [7]. By virtue of providing web portals, Liferay only requires a web browser and internet connection, and is widely used in industry and research projects. The distributed file system XtreemFS [8] is object-based and can be integrated with cloud and distributed computing environments with replication features and various access methods being available.

The MoSGrid (Molecular Simulation Grid) [9] science gateway was built by integrating and extending the above technologies. It enables complex molecular simulations via workflows and HPC resources for the three major chemical application domains (molecular dynamics, quantum chemistry and docking) via easy-to-use graphical user interfaces. By logging in once, the user gains access to all underlying systems, avoiding the need to re-authenticate [10]. Furthermore, workflows and data is annotated with metadata based on MSML (Molecular Simulation Markup Language) [11] and thus, improving reproducibility and usability. The VAVID gateway described in this manuscript will be built upon the MoSGrid experiences and architecture. It will be extended to facilitate the wind energy turbine use case.

III. Condition Monitoring of Wind Turbines

The adequate simulation of wind turbines plays an important role when designing and certifying new installations, as well as when determining the optimal location where to place them. Moreover, simulations are also applied to fine-tune the installations and explore the effect of possible upgrades. In order to evaluate a wind turbine, a series of thousands of transient simulations are conducted using data from hundreds of sensors. The gigabytes of data produced in this way model the behavior of the installation for a time window in the order of minutes, using only one of hundreds of possible configurations. Given the amount of data produced and the number of configurations that have to be explored, the use of simulation for the design and optimization of wind turbines is very complicated. Existing tools only allow punctual comparisons of time series or parameters from a small number of configurations. This forces the manufacturers to rely on Finite Element Modeling (FEM) with a real prototype. The downsides of finite element modeling are the cumbersome analysis of cause and effects, as well as the fact that many manufacturers never make the models available.

Condition monitoring systems (CMS) rely on empiric models derived from sensor data. Applied to wind turbines, they have the goal of determining the health of the components, ensuring safety (e. g. by detecting ice build-ups on rotor blades), and reducing the wear of the mechanical parts. Diagnosis and fine tuning are made available thanks to the condition information of hundreds of sensors collected by the turbine control station. Precise control strategies can be derived from this information in order to reduce wear, hence prolonging the healthy state of the installation. The condition monitoring of rotating machine parts has a long tradition, whereas the monitoring of rotor blades with oscillation analysis has been studied far less deeply up until now. Therefore, VAVID will concentrate its effort on the analysis of rotor blade oscillation data enriched with operational data of the wind turbine. The operational data includes for example the meteorological conditions and the operational modes of the wind turbine. The analysis of the rotor blade oscillation data has to be based on robust behavior models of the rotor blade. This means that these models need to be generally applicable for different wind turbine types, and also valid for a broad range of operational modes and meteorological conditions. Since in condition monitoring the models themselves are deduced from historical data, it is necessary to examine the data from a vast amount of wind turbines over a broad range of operational conditions in order to validate these behavior models. Various kinds of analysis methods are used for this task, especially signal analysis and statistical methods for regression, clustering, and optimization. Since the approach is empirical, the results are fraught with statistical uncertainty. VAVID should guide the users to evaluate those uncertainties. For example, the system should produce a warning when no cross validation data is selectable, e. g. when the user initially selected to process all available data. Otherwise, data for cross validation shall be suggested or the user should be guided to make different but comparable model hypotheses. Since parameters and the selection of data varies for every analysis, it is necessary to store these parameters and selections of data, thus linking the analysis' results to the algorithms that produced them and the data they were applied to. This will enable VAVID to make the results comparable, even with metrics not existing during early experiments, which can then be reproduced. It should also be possible to run a parametrized workflow on different data for cross validation. These envisioned features will guide the user to find the best analysis methods for their particular goal and evaluate the robustness of a behavior model.

In the context of the VAVID project, the sensor information coming from up to 600 wind turbines is to be analyzed, each of them producing a data stream of about 100 MiB per hour. This data is typically highly fragmented, consisting of millions of single measurements structured in individual files and databases. The sensor data is to be kept at least as long as the wind turbine remains operational, which is in the order of 20 years. VAVID will provide a data gateway for massive data analysis with a user interface designed for the day to day use by engineers enabling them to execute empirical analysis of the data and gain results with an educated trial-and-error approach comparable to physical experiments.

IV. Efficient Data Management

Data management is a core service for every business and scientific application, especially in the big data era. The data life cycle comprises different phases starting from understanding external data sources and integrating data into a common database schema, when databases are used. Here, the life cycle continues with an exploitation phase by answering queries, for example, against a potentially very large database and possibly closes with archiving activities to data with respect to legal requirements and cost efficiency. While understanding the data and creating a common database schema is a challenging task from a modeling perspective, efficiently and flexibly storing and processing large datasets is the core requirement form a system architectural point of view. Within the VAVID project, high performance computing concepts and technologies are applied within a novel in-memory database technology to address key challenges for large-scale data management from an application perspective.

Currently, database systems are considered as software components providing a comprehensive SQL language interface and exploiting common services provided by the operating system and the hardware layer. However, current applications as considered in VAVID demand different data management solutions. To move towards for large scale data analytics, an alternative approach with regard to modular algorithm design inspired by MapReduce [12] was proposed. The approach using algorithms from data mining, in particular from the data clustering domain was illustrated. Fundamentally, data clustering is a highly applicable analysis method that is used to reduce the amount of data or to gain understanding and acquire novel, previously unknown knowledge. The proposed modular building blocks approach as a unified construction kit for clustering algorithms corresponds as database query language in the same way. Based on this, all necessary operations like distance measurement, filtering and association are modeled as mathematical functions on matrices. To complete the set of building blocks, conditions and loops were introduced to represent the control-flow of a clustering algorithm.

In the VAVID project, this research on hybrid database query languages will be extended. The SQL database language was originally intended for the application programmer. However, after more than 20 years of extending the language, SQL can only be generated by a software component and is no longer suitable for users like knowledge workers or data scientists, interactively working with the data. The original idea of declarative query languages consists in telling the system what to retrieve and not how to retrieve the required information and is still relevant. Additionally, procedural elements are extremely worthwhile and should be part of a next generation data programming language. First, ideas like JAQL, direct access to database elements via R, or the MapReduce programming paradigm relying on 2nd order functions to provide an automated parallel execution of application logic are some examples of novel database languages. The data clustering approach [12] generally extends the MapReduce paradigm and we think this is the right way to go.

V. Numerical Algorithms

A main requirement for the analysis of the wind turbine data is to be able to compare the measurements from the different situations, e. g. change of wind turbine type, of operational mode or in the weather conditions. Roughly speaking, three different kind of algorithms will be involved.

The first step involves the computation of suitable features derived from the raw time dependent data [13], starting with the typical representation of the arising data in the frequency domain, which involves the fast Fourier transform. In the course of the project a large range of algorithms will be studied, including Hilbert transform or Cepstrum, both based on Fourier transforms, or wavelet analysis. Another aspect in the study of such time dependent data are signal decompositions, to allow the extraction of unwanted effects or noise, for these exists a range of established so-called filters, which for example allow the decomposition into low and high frequency signals. But also new approaches from recent years will be investigated to allow the data analyst the utilization of a range of algorithms for the development and identification of suitable features for the task at hand.

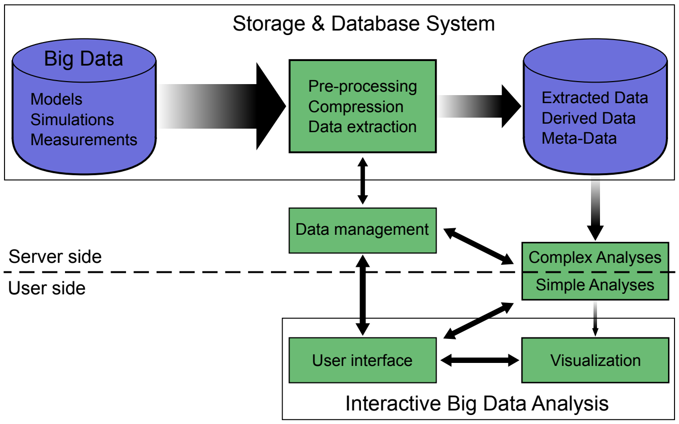

Fig. 1. Overview of the data life cycle in VAVID.

In the second step (dimensionality reduction) we employ methods to obtain a low dimensional embedding of the data which still describes most of their characteristics. The standard dimensionality reduction method is principal component analysis (PCA), which gives a low-dimensional representation of data which are assumed to lie in a linear subspace. But if the data resides in a non-linear structure, PCA gives an inefficient representation with too many dimensions. Therefore, and in particular in recent years, algorithms for non-linear dimensionality reduction, or manifold learning, have been introduced, see e. g. [14]. The goal consists in obtaining a low-dimensional representation of the data which respects the intrinsic non-linear geometry. To allow the latter, different distance measures will need to be investigated, which then provide an efficient comparison and interactive navigation of bundles of time series data. These first two steps need to connect in a suitable fashion. E. g. the signals need to be "clean" enough for the embedding, which requires the extraction of unwanted parts of the spectrum. In the other direction, the identification of suitable distance measures for this kind of data will give insight into the choice of features.

Based on the derived features and exploiting dimensionality reduction, classification and regression algorithms will be used. The aim is to obtain characteristic values, which then can be provided to the controller for decision making. A linear regression will likely not be enough, so non-linear classification or regression methods like support vector machines, logistic regression, random forests, or neural networks will be studied [15]. In particular for the latter, HPC computing capabilities can nowadays be exploited, for instance, using GPU setups with so called deep learning algorithms. Since in the application domain the aspect of dataset shift likely occurs, we will also investigate our approaches for covariate shift [16] or transfer learning [17]. Also relevant are machine learning algorithms for anomaly detection.

VI. A Data Gateway Solution

The VAVID data gateway is responsible for providing the end users with a friendly way of controlling the processes and interactions taking place on the different components of the VAVID architecture. It's aimed, thus, at increasing the overall acceptance of the system. The big picture of the data life cycle in VAVID is illustrated in Figure 1. On the left side, data coming from the wind turbines is collected in a central repository for analysis and long term storage. In order to efficiently analyze the data, a smaller subset of the arriving data is generated using pre-processing algorithms for data anonymization, compression, and pattern extraction. This subset of data is kept and managed in the in-memory database presented in Section IV and serves as input for the different numerical analyses described in Section V. Both the pre-processing and the numerical analysis are multi-step processes involving many tools with different levels of parallelization, performance, and system requirements. These building-blocks are to be implemented in modular form following a general convention. In this way, all modules will be available to the user when creating workflow compositions using the data gateway. The resulting workflows are controlled by a workflow orchestration & mapping system, which in turn is accessed by the users through the data gateway. The advantage of this approach is that the users do not have to spend time creating and maintaining orchestration and mapping scripts, or dealing with the batch system of HPC clusters. Additionally, the data gateway possesses extensible visualization capabilities that users can employ to inspect the data and the results of the analyses.

The technical realization of this concept is illustrated in Figure 2. At the lowest level, a cluster file system is responsible for storing the bulk of data coming from the geographically distributed wind turbines. In order to keep up with the data inflow, as well as to provide enough bandwidth for an effective data analysis on HPC clusters, a parallel file system like Lustre [18] or GPFS [19] is the most sensible choice. In VAVID, Lustre is the natural choice since it's the parallel file system used on the available HPC systems at the ZIH. Even though parallel file systems offer unmatched bandwidths, they often lack in areas such as fault tolerance, replication, data management, and client support. To improve the data management functionality in these areas, the VAVID data gateway will access the parallel file system indirectly through a dCache [19] layer that allows end users a more comfortable interaction with the underlying file system without hindering the HPC components from directly accessing the parallel file system to maximize performance. The HPC cluster *Taurus* [20] located at the ZIH will be the primary HPC system used to power the different data analyses during development and evaluation. However, given the site-replication capabilities of dCache, other remote HPC clusters could be seamlessly integrated into the data gateway.

Processing wind turbine data is a multi-step process involving data compression, pre-processing, dimensionality reduction as well as different analyses based on numerical methods that generate characteristic values. These values are then evaluated by scientists and engineers in order to optimize the setup of the turbines and improve their models. In order to succeed, VAVID will provide its users with an effective way of creating, executing and monitoring such workflows. Building on the experience of the past project MoSGrid, the responsibility of managing the workflows will be shared amongst two state of the art technologies: UNICORE and gUSE. The former will be in charge of the workflow orchestration and mapping to the HPC hardware, and the latter will be used for the creation, parametrization and high-level management of the workflows. Given that the wind turbines continuously generate large amounts of files that have to be processed, it is planned to automatize the transmission and launch of the pre-processing workflow. This automation will be achieved by creating adequate interfaces for the transmission of the data, coupled with the data-oriented processing capabilities of UNICORE [21].

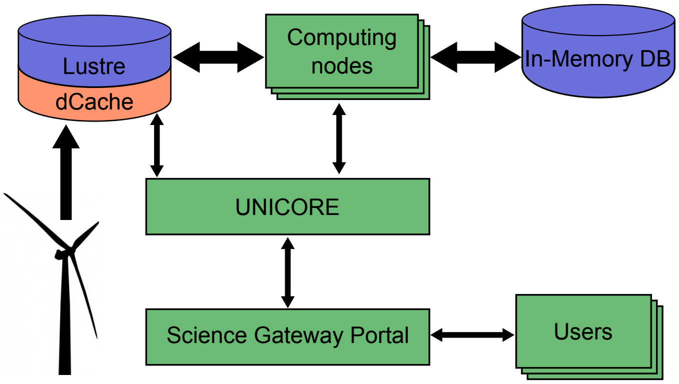

Fig. 2. Projected architecture of the VAVID system.

Data gateways offer different ways of workflow composition that range from text-based descriptions to various graphical paradigms [22]. In VAVID, workflows are composed graphically using the web interface of the data gateway. As part of the project, the Java-based workflow editor available in gUSE/WS-PGRADE has been substituted by a modern one that relies solely on HTML and Javascript to function. This was achieved thanks to a collaboration between the ZIH, the University of Edinburgh, and the MTA SZTAKI. Furthermore, this new editor will replace the former one in upcoming versions of gUSE to profit the community.

A vital aspect for VAVID is securing the access to the measurement and simulation data. The requirement of strict confidentiality is a result of the industrial context in which this data is produced. Ensuring the seamless integration of all the components present in VAVID in a secure way while keeping the user acceptance high is a complex and challenging task. With the help of digital certificates, VAVID will tackle both issues at once. Certificate-based encryption of network transmissions is widely supported by most of the components used in VAVID, so certificates will be used to increase the security this way. On the other hand, certificate-based single sign-on protocols also enjoy a high level of support in most of the VAVID components and they help the users by removing the burden of having to separately authenticate with each of the different components in the system. The overall challenge is to enable a high usability especially from the users point of view. This goal will be significantly facilitated by freeing users from having to apply for, configure and manage user certificates. The identity management solution Unity [23] will be deployed and integrated with the architecture to enable certificate-free usage of the VAVID data gateway.

VII. Conclusion and Outlook

In this paper, we have introduced the novel VAVID architecture and the aim of providing a platform for big data analysis in different industrial domains. The focus is on the use case of the condition monitoring of wind energy turbines, for which a motivation was provided explaining the current situation as well as the need for a platform like VAVID. Furthermore, we presented our concept for the in-memory processing of hot data, and described the generalities of the numerical methods used for data analysis. Finally, we described the central concept of a novel big data gateway serving as the main entry point to the system in order to improve both its usability and overall security.

Amongst the things deserving further studies are the performance and scalability of the numerical methods and storage subsystem, the interoperability and interfaces of the different components, production workflows and their aggregated performance, as well as the improvement cycle of the data gateway in order to maximize user acceptance.

Acknowledgment

The authors would like to thank the German Federal Ministry of Education and Research (BMBF) for the opportunity to do research in the VAVID project under grant 01IS14005. Furthermore, financial support by the German Research Foundation (DFG) for the MASi project is gratefully acknowledged. The research leading to these results has partially been supported by the LSDMA project of the Helmholtz Association of German Research Centres.

References

[1] Project's website, "Vergleichende Analyse von ingenieurrelevanten Mess- und Simulationsdaten (VAVID)," 2015. [Online]. Available: http://vavid.de

[2] A. Streit, P. Bala, A. Beck-Ratzka, K. Benedyczak, S. Bergmann, R. Breu, J. M. Daivandy, B. Demuth, A. Eifer, A. Giesler *et al.*, "UNICORE 6 - Recent and future advancements," *Annals of Telecommunications - Annales des Télécommunications*, vol. 65, no. 11-12, pp. 757–762, 2010.

[3] PRACE, "PRACE research infrastructure," 2015. [Online]. Available: http://www.prace-ri.eu

[4] XSEDE, "Extreme science and engineering discovery environment," 2015. [Online]. Available: https://www.xsede.org

[5] HBP, "The human brain project," 2015. [Online]. Available: https://www.humanbrainproject.eu

[6] P. Kacsuk, Z. Farkas, M. Kozlovszky, G. Hermann, A. Balasko, K. Karoczkai, and I. Marton, "WS-PGRADE/gUSE generic DCI gateway framework for a large variety of user communities," *Journal of Grid Computing*, vol. 10, no. 4, pp. 601–630, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10723-012-9240-5

[7] Liferay, "Enterprise open source portal and collaboration software," 2015. [Online]. Available: http://www.liferay.com/

[8] F. Hupfeld, T. i. Cortes, B. Kolbeck, J. Stender, E. Focht, M. Hess, J. Malo, J. Marti, and E. Cesario, "The XtreemFS architecture - a case for object-based file systems in grids," *Concurrency and Computation: Practice and Experience*, vol. 20, no. 17, pp. 2049–2060, 2008.

[9] J. Krüger, R. Grunzke, S. Gesing, S. Breuers, A. Brinkmann, L. de la Garza, O. Kohlbacher, M. Kruse, W. E. Nagel, L. Packschies, R. Müller-Pfefferkorn, P. Schäfer, C. Schärfe, T. Steinke, T. Schlemmer, K. D. Warzecha, A. Zink, and S. Herres-Pawlis, "The MoSGrid science gateway – a complete solution for molecular simulations," *Journal of Chemical Theory and Computation*, vol. 10(6), p. 2232–2245, 2014. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ct500159h

[10] S. Gesing, R. Grunzke, J. Krüger, G. Birkenheuer, M. Wewior, P. Schäfer, B. Schuller, J. Schuster, S. Herres-Pawlis, S. Breuers, Á. Balaskó, M. Kozlovszky, A. S. Fabri, L. Packschies, P. Kacsuk, D. Blunk, T. Steinke, A. Brinkmann, G. Fels, R. Müller-Pfefferkorn, R. Jäkel, and O. Kohlbacher, "A single sign-on infrastructure for science gateways on a use case for structural bioinformatics," *Journal of Grid Computing*, vol. 10, no. 4, pp. 769–790, 2012. [Online]. Available: http://link.springer.com/article/10.1007%2Fs10723-012-9247-y

[11] R. Grunzke, S. Breuers, S. Gesing, S. Herres-Pawlis, M. Kruse, D. Blunk, L. de la Garza, L. Packschies, P. Schäfer, C. Schärfe, T. Schlemmer, T. Steinke, B. Schuller, R. Müller-Pfefferkorn, R. Jäkel, W. E. Nagel, M. Atkinson, and J. Krüger, "Standards-based metadata management for molecular simulations," *Concurrency and Computation: Practice and Experience*, vol. 26(10), p. 1744–1759, 2014. [Online]. Available: http://dx.doi.org/10.1002/cpe.3116

[12] M. Hahmann, D. Habich, and W. Lehner, "Modular data clustering - algorithm design beyond mapreduce," in *Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference (EDBT/ICDT 2014), Athens, Greece, March 28, 2014.*, 2014, pp. 50–59.

[13] K.-D. Kammeyer and K. Kroschel, *Digitale Signalverarbeitung*. Springer Vieweg, 2012.

[14] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer, 2007.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Second Edition*. Springer, 2001.

[16] T. Vanck and J. Garcke, "Using hyperbolic cross approximation to measure and compensate covariate shift," in *ACML 2013, Canberra*, C. S. Ong and T. B. Ho, Eds., 2013, pp. 435–450.

[17] J. Garcke and T. Vanck, "Importance weighted inductive transfer learning for regression," in *ECMLPKDD 2014, Nancy*, ser. Lecture Notes in Computer Science, T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, Eds., vol. 8724. Springer, 2014, pp. 466–481.

[18] Lustre, "The Lustre parallel file system," 2015. [Online]. Available: http://lustre.org

[19] F. B. Schmuck and R. L. Haskin, "GPFS: A shared-disk file system for large computing clusters," in *FAST*, vol. 2, 2002, p. 19.

[20] Center for Information Services and High Performance Computing of the TU Dresden, "HPC cluster Taurus," 2015. [Online]. Available: https://doc.zih.tu-dresden.de/hpc-wiki/bin/view/Compendium/SystemTaurus

[21] B. Schuller, R. Grunzke, and A. Giesler, "Data oriented processing in unicore," in *UNICORE Summit 2013 Proceedings*, ser. IAS Series, vol. 21, 2013, pp. 1–6.

[22] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-science: An overview of workflow system features and capabilities," 2008.

[23] Unity, "Unity - cloud identity and federation management," 2014. [Online]. Available: http://unity-idm.eu