





Data Analysis for Improving High-Performance Computing Operations and Research

An Eucor Seed Money Project

Florina M. Ciorba*  Gerhard Schneider†  Dirk von Suchodoletz‡ 
Aurélien Cavelan*  Thierry Sengstag§ Sabine Gless¶
Nicolas Lachiche‡ Ahmed Samet‡

*Department of Mathematics and Computer Science, University of Basel, Switzerland

†eScience, University of Freiburg, Freiburg, Germany

§Scientific Computing Center (sciCORE), University of Basel, Switzerland

¶Faculty of Law, University of Basel, Switzerland

‡ICube Laboratory, University of Strasbourg, France

This work addresses the challenges associated with analysis of data generated by high-performance computing (HPC) systems under data protection and privacy requirements. The HPC systems are the workhorse of simulation science, enabling unique insights across many disciplines (climate modeling, life sciences, weather forecast, etc.). System monitoring and analysis of monitoring data are highly significant for the efficient operation and research in performance optimization of HPC systems. Such systems generate various and large volumes of data as they operate, constituting a case of Big Data that challenges key data protection and privacy principles. This paper describes the *Data Analysis for Improving High Performance Computing Operations and Research* (DA-HPC-OR) project funded through the Eucor – The European Campus EVTZ via the Seed Money program¹. The main goal in this project is the analysis of data collected since July 2016 on the HPC system (NEMO) at the University of Freiburg in order to improve their

¹www.eucor-uni.org

research and operations activities. Data collected on the sciCORE cluster in Basel will be used to validate the knowledge extracted from NEMO. This knowledge will be used to improve the monitoring, operational, and research activities of the three HPC systems (Freiburg, Basel, and Strasbourg). Data protection requires legal monitoring of the relevant Swiss, German, and EU legislation. Compliance with such laws will be ensured via data de-identification and anonymization prior to analysis. We leverage the HPC, legal, and data analysis expertise of the consortium to develop solutions that can be transferred to other Eucor members at no additional legislative inquiries or overheads.

1 Introduction

Each of the four pillars (experiments, theory, simulation, and data) of the scientific method produce and consume large amounts of data. Breakthrough science will occur at the interface between empirical, analytical, computational, and data-based observation. Parallel computing systems are the workhorse of the third pillar: simulation science. These systems are highly complex ecosystems, with multiple layers, ranging from the hardware to the application layer (cf. Figure 1). Monitoring solutions exist at every single layer.

Access to the monitoring data varies between the different communities, which also have different interests in the data. For instance, computational scientists in general have access to the monitoring data from the application layer and in certain cases also from the application environment layer. Their interests may include the running time of their applications but also understanding the application performance by profiling or tracing. Computer scientists may have access to monitoring data from application environment and cluster software layers. They are typically interested in performance data from both software and hardware components and subsystems. System administrators typically have access to data monitored at the hardware, system software, and cluster software layers. The monitoring interests cover both operational aspects of the system (e. g., availability, fair usage), as well as research-oriented aspects of the system (e. g., scheduling of batch jobs, resource utilization, fault-tolerance).

Regardless of the community interest, the access to and the sharing of monitoring data of HPC systems for analysis purposes requires (i) non-disclosure agreements between the data owners and/or (ii) de-identification and anonymization of the

sensitive information within the data. The former requirement prevents any public release of data, thus hindering the reproducibility of the insights and of the research results derived from the data. The latter requirement imposes compliance with the data protection and privacy regulations in force at the location where the data is produced and collected.

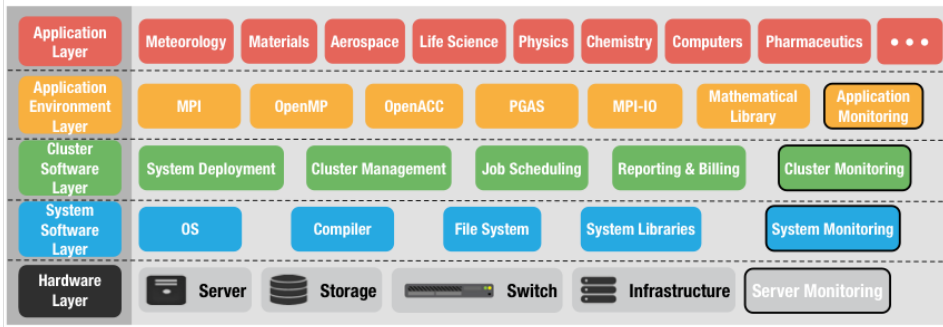


Figure 1: A typical HPC ecosystem with a layer-based monitoring approach

1.1 Goals and Expected Results

The focus of this project is on satisfying the second requirement (de-identification and anonymization that complies with data protection and privacy regulations) for the data produced and collected on the systems of the HPC centers at the member institutions: NEMO at University of Freiburg (NEMO-UniFR), sciCORE at University of Basel (sciCORE-UniBas), and HPC at University of Strasbourg (HPC-UniStra). This focus raises important legal and regulatory questions regarding the data protection and privacy laws in force in Germany, Switzerland, and France that this project needs to consider and comply with. Answering these questions requires legal expertise and a thorough analysis of the applicable individual, national, and European legislation.

1.1.1 Goal

The goal of this project is to analyze the (de-identified and anonymized) data collected at one of the three HPC centers of the consortium (NEMO-UniFR) to improve their research and operations activities, as well as offer monitoring, operational, and research insights for improving the activities at the other two HPC centers

(sciCORE-UniBas and HPC-UniStra). The rationale for concentrating on NEMO-UniFR is due to the monitoring and data integration activities running at UniFR since August 2016.

1.1.2 Approach

The approach towards achieving this goal involves: Monitoring of software and hardware components; Use of de-identification and anonymization methods; and Analysis of the data processed in the previous step. Monitoring data is collected (in a first step), under various types, formats, and sizes. Therefore, meaningful integration of the various types and formats represents a significant challenge. This challenge can be addressed by ensuring that the HPC monitoring data follows the FAIR (findable, accessible, interoperable, and reusable) data principles (Wilkinson et al., 2016) already in the data collection stage.

To comply with the data protection and privacy laws of the three partner countries, the monitoring data needs to be de-identified and anonymized (in a second step). The de-identification and anonymization also need to preserve the data usefulness. This can be addressed by extending the FAIR principles with another U (usefulness) principle, resulting in FAIRU.

Via a meaningful analysis (using well established data mining methods) of the FAIRU HPC data (in a third step) actionable insights can be generated that will result in improvements both for HPC operation as well as for the research performed in the context of performance optimization of HPC applications and systems.

The novelty of this project is that no effort towards data analysis for HPC has explored the legal challenges crosscutting any transferrable solution or knowledge.

1.1.3 Methods

The methods employed for achieving the above goals include: monitoring, FAIRU data principles, de-identification and anonymization, legal data controlling, data aggregation and mining, and insight extraction. Of these, the consortium has expertise in the following methods: monitoring, de-identification and anonymization, legal data controlling, and data mining. The consortium will develop monitoring solutions for all member institutions, apply the FAIR and useful data principles, perform legal data controlling, data aggregation and mining, and extract insights.

Expected outcome The expected outcome of this project is in the form of solutions for improving the HPC operations and research at the member institutions, that comply with the diverse applicable data protection and privacy legislations.

1.1.4 Significance

The significance of this work is in improving the research and operations activities of three HPC centers within Eucor. The solutions proposed herein will be transferable to improve the HPC operations and research at other Eucor member institutions, at the benefit of no (or minimal) additional legislative inquiries and data management overhead.

2 Current State and Challenges

Monitoring of HPC systems and applications generates various and large volumes of data, constituting a case of Big Data (FDPIC, 2017b) that challenges key privacy and data protection principles, as highlighted in (Koops et al., 2014). To help reduce the risks associated with the use and analysis of Big Data, a resolution for Big Data has been proposed (ICDPPC, 2015). Any effort towards Big Data analytics in a data protection- and privacy-aware manner needs to carefully examine and abide by the laws applicable in the country where the data is produced. In the present, in Switzerland, this corresponds to the Swiss Data Protection (SDP) law (FDPIC, 2017a). In Germany, the Federal Data Protection Act² (FDPA) applies at the moment, while Baden-Württemberg has its own Data Protection (BWDP) law (LfDI Baden-Württemberg, 2017b). The law 78-17 of 6 January 1978 on information technology, data files and civil liberties (ITDFCIL) (CNIL, 2014) presently applies in France. Starting in May of 2018, France and Germany, as member states of the European Union, will enforce the EU guidelines for data protection (EUGDP) (LfDI Baden-Württemberg, 2017a).

This will render legal monitoring between France and Germany (and throughout the EU) easier than before. However, legal monitoring between Switzerland and any EU country will remain difficult. This difficulty is also captured by a very recent census of privacy and data protection authorities (ICDPPC, 2017). Nonetheless,

²Federal Data Protection Act: http://byds.juris.de/byds/012_1.4_BDSG_1990_Inhaltsuebersicht.html, (visited on 27.09.2017)

data protection and privacy cannot be hardcoded (Koops et al., 2014) in HPC systems, runtime systems, or in programming languages. Therefore, ensuring that data complies with all legal provisions within sectoral, state, national, and European legislation, that contain data protection requirements is a non-trivial, yet critical task for this project.

A recent overview of the state of the art monitoring solutions in HPC systems can be found in (Jha et al., 2017; Layton, 2017; Brown et al., 2017). Monitoring is also used proactively for system maintenance (Röder, 2016) as well as for application optimization. Faults have become a major threat for the execution of HPC applications at scale (Geist, 2016; Snir et al., 2014; Cappello et al., 2009). Resilience has been identified as one of the top ten Exascale challenges (Dongarra et al., 2011). In this context, the analysis of system logs (Martino et al., 2014; Tiwari et al., 2015; Sridharan et al., 2015) paves the way for understanding the distribution of faults and their impact on the system and its performance. Research in this direction is urgently needed to improve current fault-detection methods, fault models, and other resilience techniques (Benoit et al., 2016). Existing work on (manual) data analysis or (automatic) data mining for HPC includes fault prediction (Gainaru et al., 2012) and coping with recall and precision of soft error detectors (Bautista-Gomez et al., 2016), respectively. However, there is room for improvement. Specifically, the use of data mining to HPC fits into Industry 4.0 where there is a strong interest on improving the performance of processes via log mining. This is described in a recent manifesto on process mining (Aalst et al., 2011).


3 Conclusion and Outlook


This seed money project will significantly raise the visibility of Eucor via the use of its label on each of the members' website, publications and via their network channels. The DA-HPC-OR project already tightens the collaboration between scientists or the three neighboring countries beyond this project.


Corresponding Author


Florina M. Ciorba: florina.ciorba@unibas.ch
Department of Mathematics and Computer Science,
University of Basel, Switzerland


ORCID

Florina M. Ciorba  <https://orcid.org/0000-0002-2773-4499>

Gerhard Schneider  <https://orcid.org/0000-0002-3214-002X>

Dirk von Suchodoletz  <https://orcid.org/0000-0002-4382-5104>

Aurélien Cavelan  <https://orcid.org/0000-0002-1784-0730>

License  <https://creativecommons.org/licenses/by-sa/4.0>

References

- Aalst, W. van der et al. (2011). *Process Mining Manifesto*. URL: http://www.win.tue.nl/ieetfpm/lib/exe/fetch.php?media=shared:process_mining_manifesto-small.pdf (visited on 27.09.2017).
- Bautista-Gomez, L. et al. (2016). »Coping with recall and precision of soft error detectors«. In: *Journal of Parallel and Distributed Computing*, 2016, 98, 8 - 24 98, pp. 8–24.
- Benoit, A., A. Cavelan, Y. Robert and H. Sun (2016). »Optimal Resilience Patterns to Cope with Fail-Stop and Silent Errors«. In: *Proceedings of the 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. Chicago, USA.
- Brown, R. and O. Graß (2017). *Application Monitoring with openITCOCKPIT*. URL: <http://www.admin-magazine.com/Archive/2017/37/Application-Monitoring-with-openITCOCKPIT> (visited on 27.09.2017).
- Cappello, F. et al. (2009). »Toward Exascale Resilience«. In: *Int. Journal of High Performance Computing Applications* 23, pp. 374–388.
- CNIL (2014). *Act n°78-17 of 6 January 1978 on information technology, data files and civil liberties*. URL: <https://www.cnil.fr/sites/default/files/typo/document/Act78-17VA.pdf> (visited on 27.09.2017).
- Dongarra, J. and et al. (2011). »The International Exascale Software Project Roadmap«. In: *Int. J. High Perform. Comput. Appl.* 25, pp. 3–60.
- FDPIC (2017a). *Data Protection*. URL: <https://www.edoeb.admin.ch/datenschutz/index.html?lang=en> (visited on 27.09.2017).
- (2017b). *Explanatory notes on Big Data*. URL: <https://www.edoeb.admin.ch/datenschutz/00683/01169/index.html?lang=en> (visited on 27.09.2017).
- Gainaru, A., F. Cappello, M. Snir and W. Kramer (2012). »Fault Prediction Under the Microscope: A Closer Look into HPC Systems«. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. Salt Lake City.

- Geist, A. (2016). *How To Kill A Supercomputer: Dirty Power, Cosmic Rays, and Bad Solder*. URL: <https://spectrum.ieee.org/computing/hardware/how-to-kill-a-supercomputer-dirty-power-cosmic-rays-and-bad-solder>.
- ICDPPC (2015). *Resolution Big Data*. URL: <https://icdppc.org/wp-content/uploads/2015/02/Resolution-Big-Data.pdf> (visited on 27.09.2017).
- (2017). *Counting on Commissioners: High level results of the ICDPPC Census 2017*. URL: <https://icdppc.org/wp-content/uploads/2017/09/ICDPPC-Census-Report-1.pdf> (visited on 27.09.2017).
- Jha, S. et al. (2017). »Holistic Measurement Driven System Assessment«. In: *Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA)*. Honolulu HI, USA.
- Koops, B.-J. and R. Leenes (2014). »Privacy regulation cannot be hardcoded. A critical comment on the ‘privacy by design’ provision in data-protection law«. In: *International Review of Law, Computers & Technology* 28 (2), pp. 159–171.
- Layton, J. (2017). *Resource Monitoring For Remote Applications*. URL: <http://www.admin-magazine.com/HPC/Articles/REMORA> (visited on 27.09.2017).
- LfDI Baden-Württemberg (2017a). *EU guidelines on data protection*. URL: <https://www.baden-wuerttemberg.datenschutz.de/eu-richtlinien-zum-datenschutz/> (visited on 27.09.2017).
- (2017b). *Laws / Regulations*. URL: <https://www.baden-wuerttemberg.datenschutz.de/gesetzeverordnungen/> (visited on 27.09.2017).
- Martino, C. D. et al. (2014). »Lessons Learned from the Analysis of System Failures at Petascale: The Case of Blue Waters«. In: *Proceedings of the 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. DSN '14*, pp. 610–621.
- Röder, D. (2016). *Proactive Monitoring*. URL: <http://www.admin-magazine.com/Articles/Proactive-Monitoring> (visited on 27.09.2017).
- Snir, M. and et al. (2014). »Addressing Failures in Exascale Computing«. In: *Int. J. High Perform. Comput. Appl.* 28, pp. 129–173.
- Sridharan, V. et al. (2015). »Memory errors in modern systems: The good, the bad, and the ugly«. In: *ACM SIGPLAN Notices*. Vol. 50, pp. 297–310.
- Tiwari, D., S. Gupta, G. Gallarno, J. Rogers and D. Maxwell (2015). »Reliability Lessons Learned From GPU Experience With The Titan Supercomputer at Oak Ridge Leadership Computing Facility«. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. Austin, Texas.
- Wilkinson, M. D. and et al (2016). »The FAIR Guiding Principles for scientific data management and stewardship«. In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.18.