Business Model with Alternative Scenarios
Deliverable D4.1

For each Deliverable, a single file in PDF (max 50MB) can be uploaded

This project has received funding from the European Union's
Horizon 2020 Research and Innovation Programme
under Grant Agreement No 675191

About this document

Work package in charge: *WP4 Exploitability*
Actual delivery date for this deliverable: *24 February 2017*
Dissemination level: *PU*

Lead author:
Jakob Lüttgau, DKRZ

Other contributing authors:
Julian Kunkel, Jakob Lüttgau (DKRZ)
Jens Jensen (STFC)
Bryan Lawrence (STFC and the University of Reading)

Contacts: esiwace@dkrz.de
Visit us on: www.esiwace.eu
Follow us on Twitter: @esiwace

Index

1. Abstract /publishable summary

This report summarizes the work on requirements and business models for the storage infrastructure within weather and climate data centres (although much of the work has wider applicability). The report concentrates on identifying and evaluating the interplay of important cost factors, along with an introduction to relevant (storage and data movement) hardware and software technology, terminology and performance metrics.

The report begins with a description of how climate and weather applications make use of HPC systems, the arising challenges and data requirements, and some trends that are likely to impact future data centre designs. Related work follows that introduces some cost developments, cost modelling and technological developments. The body of the work is an integrated graph-based approach to modelling costs, resilience and performance for storage systems. Storage models are evaluated in several scenarios each introducing some architectural changes to currently deployed high-performance systems and discusses the cost and performance implications. These discussions are made on a high-level of abstraction as no model is able to predict the non-linear behaviour when scaling out big systems accurately. Conclusions identify the potential benefits that more refined models might offer and outline future work.

2. Conclusion & Results

The report looked at the evolution of data centres and various approaches concerned with data access that have been used in scientific contexts. We covered the trends in both climate and weather research as well as the development of storage technologies and, in particular, discussed the performance divergence of compute vs. storage. In this context, we also reviewed cloud technologies that have the potential to change the scientific computing landscape. The report also looked at the data centre perspective and vendor perceptions on the cost developments of disks, NAND and tape.

We described high-level considerations for costs, performance and resilience considerations. A hierarchical graph-based approach is proposed allowing us to specify component and system characteristics and costs and visualize them. This also allows to add or remove complexity and detail as becomes necessary and open the possibility for automation. In the model, we also look at the core components that are usually found in data centres. Starting with the smallest components such as hard drives, that provide little tuning opportunities, the report moves on to discuss how derive the emergent performance and cost of small subsystems like individual compute and I/O nodes as well as of larger subsystems like the network or a parallel file system.

The impact of cloud computing is also considered and some configurations costed and found to be significantly more expensive than on-site systems for typical heavily used weather and climate data centres.

In the evaluation, we provided the characteristics for costs and performance of currently deployed systems. Starting from these systems as baseline, we explored changes to the system and the impact on cost, power consumption and performance applying coarse grained models. The scenarios discussed do not aim to quantify the costs accurately but instead provide a qualitative perspective on the implications of various modifications of the system. In that sense, they serve as blueprints for subsequent scenarios. For example, we discussed reducing the storage budget in favour of compute resources – with the goal of doing so without reduction in scientific productivity, and conclude that this requires more intelligent scheduling and staging mechanisms. Researchers would also likely need to drastically change their applications as well as their workflows.

The report shows that the cost developments for the technologies are an important but unknown factor and that it is therefore advisable to push for more flexible infrastructures to make the integration of new technologies easier. One such technology, NVRAM, might shape future data centres radically, but the cost-benefits of this technology is difficult to quantify as the cost-prognostics of this technology does not exist. Similarly, it is clear that cloud-like services from within data centres will become increasingly important, and will depend on flexible infrastructures. Centralizing some resources such as (pooled) memory, introduces some opportunity, but the benefits are difficult to judge right now as we do not yet have all the relevant data. Not withstanding these caveats, the abstract models shed light on the available design space.

While the high-level models can show certain cases for which a certain technology is useful, they cannot quantify the benefit for individual workloads as the dimension of time and the workload characteristics is abstracted. An approach for more detailed models is showcased: Using discrete event simulation, it is possible to account for the workload behaviour of a system. In particular, we introduced a simulation for hierarchical storage systems that integrate tape libraries and online storage. The case study demonstrates the overhead associated with fine grained models but also shows that it is possible to approximate the observed behaviour in the actual system monitoring. It is then shown that by varying the configuration, we can make forecasts for impacts on the quality of service of an alternative system configuration.

3. Project objectives

This deliverable contributes directly and indirectly to the achievement of all the macro-objectives and specific goals indicated in section 1.1 of the Description of the Action:

Macro-objectives	**Contribution of this deliverable?**
Improve the efficiency and productivity of numerical weather and climate simulation on high-performance computing platforms	Yes
Support the end-to-end workflow of global Earth system modelling for weather and climate simulation in high performance computing environments	Yes
The European weather and climate science community will drive the governance structure that defines the services to be provided by ESiWACE	No
Foster the interaction between industry and the weather and climate community on the exploitation of high-end computing systems, application codes and services.	Yes

Increase competitiveness and growth of the European HPC industry	Yes

Specific goals in the workplan	Contribution of this deliverable?
Provide **services** to the user community that will impact beyond the lifetime of the project.	No
Improve **scalability** and shorten the time-to-solution for climate and operational weather forecasts at increased resolution and complexity to be run on future extreme-scale HPC systems.	Yes
Foster **usability** of the available tools, software, computing and data handling infrastructures.	Yes
Pursue **exploitability** of climate and weather model results.	Yes
Establish governance of common software management to avoid unnecessary and redundant development and to deliver the best available solutions to the user community.	No
Provide **open access** to research results and **open source** software at international level.	Yes
Exploit **synergies** with other relevant activities and projects and also with the global weather and climate community	No

4. Detailed report on the deliverable

The work done covers

1. review and summary of the state of the art and related work;
2. gathering of information of data centre characteristics;
3. development of the high-level models;
4. prototyping example scenarios and evaluating the results;
5. providing a model and simulation of a hierarchical storage system; and
6. documentation in the deliverable.

DKRZ was involved in all activities. STFC was involved in 1, 2, 5, 6.

5. References *(Bibliography)*

[Aga17] Agam Shah. Intel ships first Optane memory modules for testing, January 2017.
[Ben15] Ben Stopford. Log Structured Merge Trees, February 2015.
[BJZ14] Lorenzo Blasi, Jens Jensen, and Wolfgang Ziegler. Expressing quality of service and protection using federation-level service level agreement. In Proc.Euro-Par2013, pages 146--156. Springer, 2014.
[CB02] Thomas Connolly and Carolyn Begg. Database Systems, chapter19. Addison Wesley, 3rd ed. edition, 2002.
[DFH13] GDecad, RFontana, and SHetzler. The Impact of Areal Density and Millions of Square Inches (MSI) of Produced Memory on Petabyte Shipments for TAPE, NAND Flash, and HDD Storage Class, 2013.
[Eva10] Evangelos Eleftheriou. Trends in Storage Technologies, 2010.
[FDH13] RobertE. Fontana, GaryM. Decad, and S.R. Hetzler. The impact of areal density and millions of square inches (MSI) of produced memory on petabyte shipments of TAPE, NAND flash, and HDD storage

class memories. In 2013 IEEE 29th Symposium on Mass Storage Systems and Technologies (MSST), pages 1--8. IEEE, 2013.
[Gir16] Maria Girone. Experiences using commercial clouds in CMS. Computing in High Energy Physics (CHEP), October 2016.
[GWM+14] Preeti Gupta, Avani Wildani, EthanL. Miller, Daniel Rosenthal, IanF. Adams, Christina Strong, and Andy Hospodor. An economic perspective of disk vs. flash media in archival storage. In 2014 IEEE 22nd International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems, pages 249--254. IEEE, 2014.
[HFD+09] James Hughes, Dave Fisher, Kent Dehart, Benny Wilbanks, and Jason Alt. HPSS RAIT Architecture. White paper of the HPSS collaboration, www. hpss-collaboration. org/documents/HPSS RAIT Architecture. pdf, 2009.
[KKL14] Julian Kunkel, Michael Kuhn, and Thomas Ludwig. Exascale Storage Systems An Analytical Study of Expenses. In Supercomputing Frontiers and Innovations, 2014.
[KL12] JulianM. Kunkel and Thomas Ludwig. IOPmModeling the I/O Path with a Functional Representation of Parallel File System and Hardware Architecture. In 2012 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, pages 554--561. IEEE, 2012.
[KMKL10] Julian Kunkel, Olga Mordvinova, Michael Kuhn, and Thomas Ludwig. Collecting Energy Consumption of Scientific Data. Computer Science - Research and Development, pages 1--9, 2010.
[Ko15] PankajDeep Kaur and others. A survey on Big Data storage strategies. In Green Computing and Internet of Things (ICGCIoT), 2015 International Conference On, pages 280--284. IEEE, 2015.
[Kov15] Kove. About Xpress Disk (Xpd). Kove Corporation, 2015.
[LPG+11] Jay Lofstead, Milo Polte, Garth Gibson, Scott Klasky, Karsten Schwan, Ron Oldfield, Matthew Wolf, and Qing Liu. Six degrees of scientific data: Reading patterns for extreme scale science IO. In Proceedings of the 20th International Symposium on High Performance Distributed Computing, pages 49--60. ACM, 2011.
[MA15] Marco Mancini and Giovanni Aloisio. How advanced cloud technologies can impact and change HPC environments for simulation. In High Performance Computing & Simulation (HPCS), 2015 International Conference On, pages 667--668. IEEE, 2015.
[MAP05] I.Mandrichenko, W.Allcock, and T.Perelmutov. Gridftp v2 protocol description, May 2005.
[Mel12] Mellanox. Building a Scalable Storage with InfiniBand, 2012.
[MG11] Peter Mell and Tim Grance. 800-145: The NIST definition of cloud computing. 2011.
[OCGO96] Patrick O'Neil, Edward Cheng, Dieter Gawlick, and Elizabeth O'Neil. The log-structured merge-tree (LSM-tree). Acta Informatica, 33(4):351--385, 1996.
[OMBE11] JonathanT. Overpeck, GeraldA. Meehl, Sandrine Bony, and DavidR. Easterling. Climate data challenges in the 21st century. science, 331(6018):700--702, 2011.
[OSP14] Catherine Olschanowsky, Susmit Shannigrahi, and Christos Papadopoulos. Supporting climate research using named data networking. In 2014 IEEE 20th International Workshop on Local & Metropolitan Area Networks (LANMAN), pages 1--6. IEEE, 2014.
[PWB07] Eduardo Pinheiro, Wolf-Dietrich Weber, and LuizAndr'e Barroso. Failure trends in a large disk drive population. Proc.5th USENIX Conf. on File and Storage Technologies, Feb 2007.
[SZ17] J.D. Silver and C.S. Zender. The compression--error trade-off for large gridded data sets. Geoscientific Model Development, 10(1):413--423, 2017.
[WZK+13] Simon Waddington, Jun Zhang, Gareth Knight, Jens Jensen, Roger Downing, and Cheney Ketley. Cloud repositories for research data -- addressing the needs of researchers. J.Cloud Computing, 2013 2:13, June 2013.
[YWX+13] Chao Yin, Jianzong Wang, Changsheng Xie, Jiguang Wan, Changlin Long, and Wenjuan Bi. Robot: An efficient model for big data storage systems based on erasure coding. In Big Data, 2013 IEEE International Conference On, pages 163--168. IEEE, 2013.

6. Dissemination and uptake

6.1 Dissemination

This formal deliverable is the beginning of dissemination. The project participants expect to further refine and publish further results in the academic literature.

Peer reviewed articles

None as yet.

Publications in preparation OR submitted

Additional work beyond that listed below is planned.

In preparation OR submitted?	Title	All authors	Title of the periodical or the series	Is/Will ***open access*** be provided to this publication?
In preparation	Understanding Costs of Tape Libraries for Hierarchical Storage Management Systems	Jakob Lüttgau, Julian Kunkel	High Performance Computing, Proceedings	Yes, green open access

6.2 Uptake by the targeted audience

As indicated in the Description of the Action, the audience for this deliverable is:

X	The general public (PU)
	The project partners, including the Commission services (PP)
	A group specified by the consortium, including the Commission services (RE)
	This reports is confidential, only for members of the consortium, including the Commission services (CO)

We recognise that the "General Public" will get little out of this work, however, by being in the public domain, we expect that we will be able to discuss the work with a range of other important audiences, including vendors, and the wider weather and climate (and storage) community. It is these last two categories: vendors and those responsible for providing data centres for the weather and climate community who are the true intended audience.

7. The delivery is delayed: No

8. Changes made and/or difficulties encountered, if any

The original intention had been to make a tool available which encoded the work described in the deliverable. That has not yet been possible, but it will form part of future work. However, the intellectual objectives of the deliverable have been met, and while they have not yet been applied to as many uses cases (including within partners) as anticipated, that too will form part of future work when the necessary input data may be available.

9. Efforts for this deliverable

Person-months spent on this deliverable:

Beneficiary	**Person-months**	**Period covered**	**Names of scientists involved, including third parties (if appropriate)** **and their gender (f/m)**
DKRZ	8	M1-18	Jakob Lüttgau (m), Julian Kunkel (m)
ECMWF			
CNRS-IPSL			
MPG			
CERFACS			
BSC			
STFC	3	M1-18	Bryan Lawrence (m), Jens Jensen (m), Brian Davies (m)
MET O			
UREAD			Bryan Lawrence (m)
SMHI			
ICHEC			
CMCC			
DWD			
SEAGATE			
BULL			
ALLINEA			
Total	**9**		

10. Sustainability

10.1. Lessons learnt: from the experiences of the work to date (positive and negative)

The modelling bears a lot of potential to understand the cost-benefit factors and can serve as a discussion point between vendors and data centres but also between data centres and users. These models need the costs and performance characteristics as input and assume linear dependency between variables to simplify the situation

(sometimes too much). But even so, determining the base characteristics such as costs for storage may be non-trivial when procurements cover multiple components (storage + compute part of a supercomputer like for DKRZ). Fine-grained models bear the difficulty to obtain the respective data needed for the simulation. In our case, obtaining the workload traces for tape access is non-trivial but shows that we have to increase the monitoring effort on the sites.

10.2 Links built with other deliverables, WPs, and synergies created with other projects

Based on these results, we initiated a collaboration between Argonne National Laboratory and the company Kove to discuss cost-benefit of several storage related architectures.

11. Dissemination activities

Type of dissemination and communication activities	**Number**	**Total funding amount**	**Type of audience reached In the context of all dissemination & communication activities** ('multiple choices' is possible)	**Estimated number of persons reached**
Participation to a conference	1	2.500 €	Scientific Community Vendors	Scientific Community
Poster "Modelling and Simulation of Tape Libraries for Hierarchical Storage Systems" (J. Lüttgau, J. Kunkel) at Supercomputing, 2016.				
Total funding amount		2.500 €		

Intellectual property rights resulting from this deliverable

Copyright only!