Tapping Data Spaces in HPC: Enabling, Exploring and Benchmarking Database & Object Store Usage

M. Hayek[a], S. Hachinger[a], H. Gurbanov[a], V. Pauw[a], M. Golasowski[b], J. Martinovič[b], P. Karagoz[c], I.H. Toroslu[c]

[a]Leibniz Supercomputing Centre (LRZ), Bavarian Academy of Sciences & Humanities, Garching near Munich, Germany; [b]IT4Innovations National Supercomputing Center (IT4I), VŠB – Technical University of Ostrava, Czech Republic; [c]Department of Computer Engineering, Middle East Technical University (METU), Ankara, Turkey

OUR RESEARCH

Context. HPC relies on optimised, parallel file input/output (I/O). Enterprise customers envisage to use HPC-scale systems – but they are familiar with Cloud Computing and methodically rely on **database management systems (DBMS) or object stores (OBS)**. The **EXA4MIND** Extreme Data analytics project thus explores **combinations of versatile data stores, supercomputing, and data ecosystems** with four pilot applications [1].

Aims and Conjectures. Via data-transfer benchmarks, we explore direct DBMS and OBS usage from HPC clusters. We investigate the following ideas: (I) there should be a **speed hierarchy** "DBMS I/O < OBS I/O < FILE I/O"; (II) **DBMS I/O will probably be limited by the DBMS-server or DBMS-client application** (processing or data handling), (III) **OBS I/O will be limited by the network or by OBS-server/-client-dependent overheads,** depending on the setting, transfer direction, and file sizes. Building upon others' earlier work [2], we aim to understand **when/how direct DBMS/OBS use is feasible in supercomputing** or when **data staging or caching is needed – for which we provide tools.**

Methods. We **measure data-transfer speeds** using **LRZ OpenStack Compute Cloud** [3] Virtual Machines (VMs) with 100 Gbit/s (shared) connectivity, and **LRZ CoolMUC-4** [3] / **IT4I Karolina** [4] **High-Performance-Computing (HPC)** nodes in different combinations. The server and client DBMS/OBS systems used are PostgreSQL (representing a common DBMS [5]), MINIO and S3Proxy (S3-compliant OBS servers [6,7]) in current versions – **for which EXA4MIND provides installation recipes.**

ENVIRONMENTS USED FOR BENCHMARKING

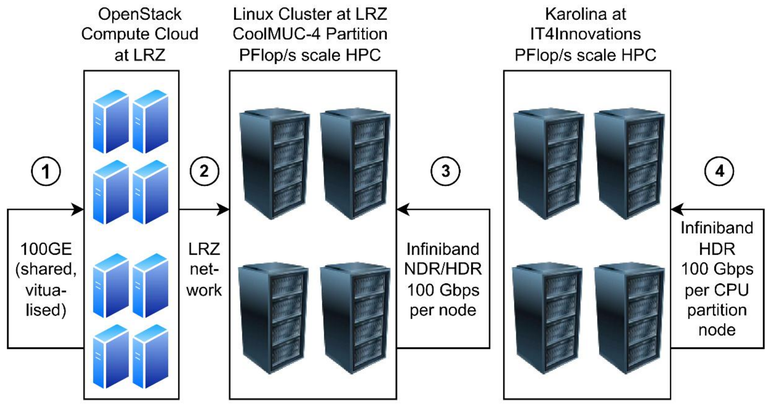

Figure 1. Environments / Data Paths in our Benchmarks for DBMS & OBS Access.

We measure read (download) speeds from DBMS and OBS within the LRZ Compute Cloud (Figure 1, Data Path 1), from a DBMS/OBS server on the Cloud to the CoolMUC-4 HPC system (Path 2), within CoolMUC-4 (Path 3) and within IT4I's Karolina HPC system. Write (upload) speeds are tested on selected occasions.

IS DBMS USAGE ON A HPC SYSTEM POSSIBLE?

We explored whether DBMS may be directly used from HPC applications, with DBMS servers running on the LRZ Cloud (Fig. 2, Data Path 2) or on CoolMUC-4 (Path 3). Table 1 shows that direct usage is only feasible for limited (e.g. metadata) I/O. Write rates in particular are obviously limited by DBMS processing, confirming conjecture (II). Reads can reach 0.8 GB/s within HPC systems.

Target system / Direction	*PostgreSQL DBMS* on LRZ Compute Cloud	*PostgreSQL DBMS* on CoolMUC-4	CoolMUC-4 Scratch *cp* (for comparison)
Rate from system to CoolMUC-4 Scratch	54.6 MB/s	792 MB/s	2610 MB/s
Rate from CoolMUC-4 Scratch to system	28,3 MB/s	24,0 MB/s	2610 MB/s

Table 1. Data rates obtained communicating with an LRZ-Cloud or internal PostgreSQL server from the LRZ CoolMUC-4 HPC system. Colours indicate low (red) to high (green) rates. Rates for a file copy (cp) from the Scratch file system to itself are shown for comparison (right col.). A "SELECT *" type query is used for reading, and a "COPY" for writing.

OBS USAGE ON A HPC SYSTEM

Figure 2 shows our exploration of OBS usage on LRZ and IT4I HPC clusters (cf. Figure 1, Data Paths 2, 3 and 4). With 10 GB files and multiple streams, network bandwidth can be largely utilised (cf. conjecture (III) in introduction).

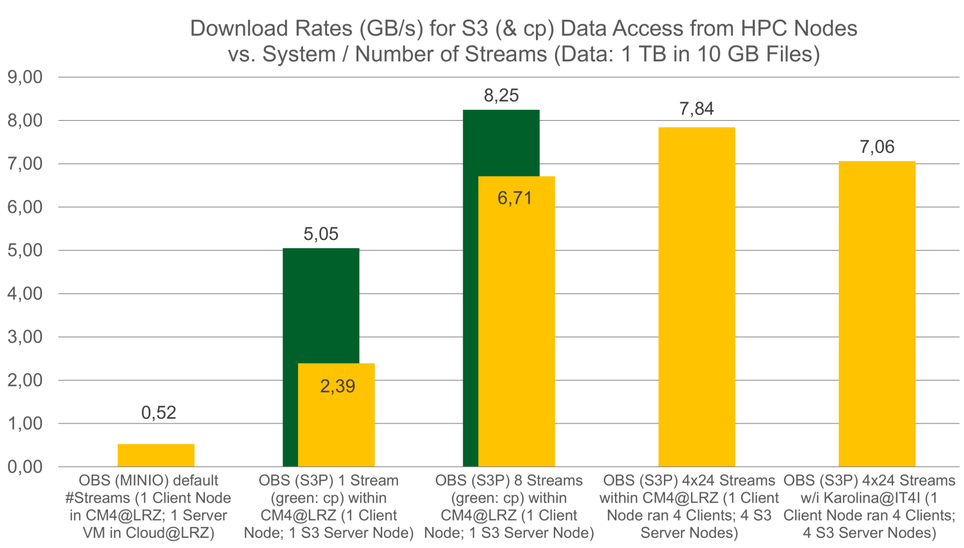

Figure 2. Rates (yellow bars) retrieving data from S3Proxy OBS (MINIO where noted) on nodes of LRZ's CoolMUC-4 and IT4I's Karolina (rightmost bar). We use s5cmd as a client and a thread number as indicated (number of streams); rates obtained with one or eight cp processes/ streams (green bars) copying the same data from SCRATCH to /dev/null are given to roughly indicate max. read bandwidth. In case of 4x24 streams, four server nodes are used and four client instances (24 threads each) on one node, to test whether client-node network bandwidth can be utilized.

UNDERSTANDING THE LIMITATIONS OF OBS

OBS middleware (e.g. MINIO, S3Proxy) allows us to explore OBS usage in a supercomputing environment. To understand the limitations of OBS systems, we are benchmarking up- and downloads within the LRZ Compute Cloud (Fig. 1, Data Path 1). First results (Figure 3) show that uploads are slower (depending on the server product); performance is worse for small files (in particular < 100MB), hinting at middleware- or network-related latencies. We do not observe overly high CPU loads in our tests (neither on Cloud VMs nor previously on HPC).

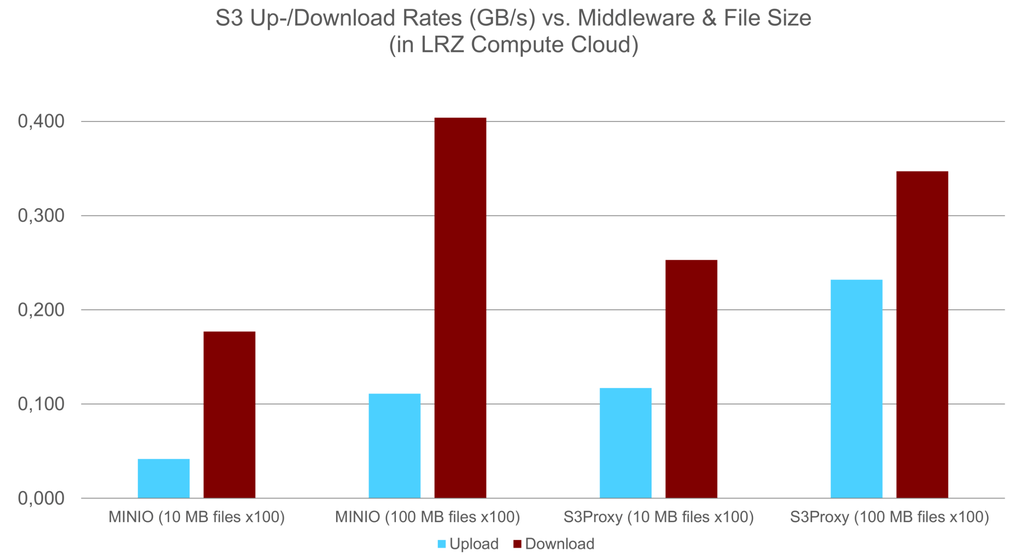

Figure 3. Exploring OBS performance dependency on software (S3Proxy/MINIO), transfer direction (up-/download) and file size (10/100 MB) – with server/client VMs on LRZ Compute Cloud (client: boto3).

MAIN TAKEAWAY FROM BENCHMARKS

Typical HPC applications may have an **I/O of 0.01-1 PB in a day.** If I/O shall occur only 1% of the time – **necessary "burst" output rates are of 10-1000 GB/s. DBMS do not reach such rates,** but can e.g. be utilised to read and write **metadata** in an image-annotation job. **The OBS explored by EXA4MIND reach promising rates** – if file sizes and thread numbers are high enough. Work on parallel I/O to a OBS server group (cf. [2]) may be worth re-exploring.

OVERCOMING DATA ANALYTICS OBSTACLES WITH EXA4MIND

The EXA4MIND project goes much beyond providing benchmarks and installation recipes for DMBS and OBS for extreme data analytics in supercomputing environments. It provides a flexibly-deployable **Advanced Query and Indexing System** framework where **analytics workflows** can be **scripted with Apache Airflow and Dask** (see [8]), with **libraries for efficient data staging and caching matching this environment.** Figure 4 shows the concept of the **EXA4MIND/AQIS Caching Library,** which stages data between a slower OBS (or DBMS) and a "caching space" on a fast HPC File System. The cache status is kept in a lightweight internal database. The interface allows for accessing the cached objects and requesting the caching of additional data. The library is available through the **Open Source repository of the EXA4MIND Platform** (https://opencode.it4i.eu/exa4mind/platform), whose modules can be flexibly deployed in a mixed (HPC/Cloud/...) supercomputing environment.

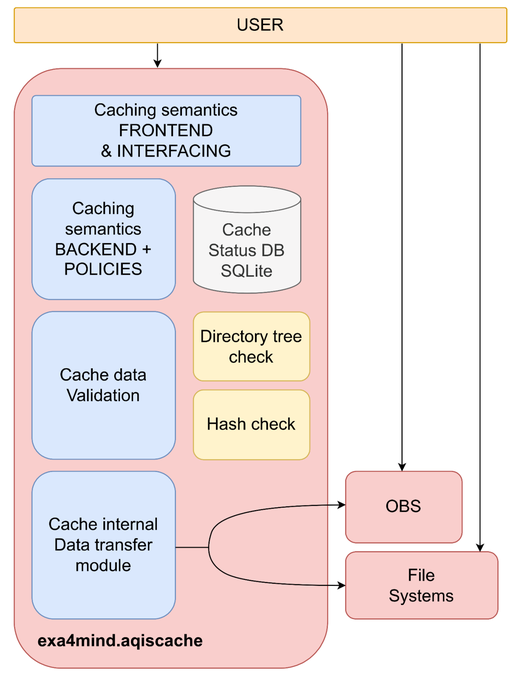

Figure 4. EXA4MIND/AQIS Caching Library Concept.

REFERENCES & FURTHER INFORMATION

[1] https://exa4mind.eu; [2] Gadban, F.; Kunkel, J. Analyzing the Performance of the S3 Object Storage API for HPC Workloads. Appl. Sci. 2021, 11, 8540. https://doi.org/10.3390/app11188540; [3] https://doku.lrz.de/; [4] https://docs.it4i.cz; [5] Stonebraker, M.; Rowe, L.A. The design of POSTGRES. SIGMOD Rec. 1986, 15, 340–355. https://doi.org/10.1145/16856.16888 [5] https://www.postgresql.org; [6] https://min.io; [7] https://github.com/gaul/s3proxy; [8] EXA4MIND Consortium. D3.1 Architecture of AQIS. Zenodo. 2024. https://doi.org/10.5281/zenodo.15695601

M. Hayek | EXA4MIND on YouTube | Project Site | E4M Kickstarter Initiative | Open Source Repository

ACKNOWLEDGEMENT

This research received the support of the **EXA4MIND** project, funded by the European Union's Horizon Europe Research and Innovation Programme, under Grant Agreement N° 101092944. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.
Ideation of this work was strongly catalysed by meetings using the project grant **ICBxBCI** for Czech-Bavarian Collaboration and Researcher Mobility of the Bavarian State Chancellery (Bayerische Staatskanzlei). The authors gratefully acknowledge the computational and data resources provided by **IT4Innovations** and by the **Leibniz Supercomputing Centre**. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the **e-INFRA CZ** (ID:90254).

Funded by the European Union