Data Science Storage for SuperMUC

1. Overview

GCS has funded the investment costs for two LRZ Data Science Storage systems via the SuperMUC-NG and InHPC-DE projects. Each of the system provides a capacity of 10PB and approximately 2.7 billion I-nodes.

LRZ’s Data Science Storage (DSS) is a novel approach at LRZ to solve the demands and requirements of data intensive science. Therefore, DSS implements a data centric management approach, which gives users the ability to:

- Store vast amounts of data for as long as funding is secured and the data is important to them or the science community
- Access this data from the whole LRZ computing ecosystem (SuperMUC, LinuxCluster, Compute Cloud, VMWare, Remote Visualization, Housed Customer Compute Systems)
- Share this data between arbitrary users of the LRZ computing ecosystem
- Access/Transfer this data world wide via a high performance, WAN optimized transfer protocol, using a simple Graphical User Interface in the Web
- Share this data with arbitrary users around the globe, like people are already used to from services like LRZ Sync+Share, Dropbox or Google Drive

All DSS systems are managed by a cloud-like Self-Service portal. The development of this portal as well as the operations and backups of the two systems are performed by LRZ’s own resources.

2. Disclaimer

GCS has funded the SuperMUC-NG and InHPC DSS systems until end of 2024. While LRZ of course aims to raise funding for a successor system, which will replace the current systems once they’re phased out, we currently cannot give any guarantees for retaining data on SuperMUC-NG and InHPC-DE DSS beyond end of 2024. However, if we fail to raise funds for a successor system, we will announce this well in advance, to give users enough time to move their data somewhere else.

3. Eligible Research Projects

Storage space on SuperMUC-NG and InHPC-DE DSS is available only upon explicit request by the research project. All approved SuperMUC-NG projects are eligible to file a motion for storage space on the SuperMUC-NG and InHPC-DE DSS systems.

4. Maximum Storage Life

As stated in chapter 2, currently funding is secured until the end of 2024 and we cannot give any guarantees on storage life beyond this date. However beyond the implications that funding may have on storage life, the following rules apply:

For **Closed Data**, the maximum storage life is determined by the lifetime of the compute project on SuperMUC-NG. We keep data on SuperMUC-NG and InHPC-DE DSS for a grace period of one additional year after the compute project ends. After this grace period, data is automatically deleted. However if the PI or one of his proxies or successors get's granted another compute project on SuperMUC-NG within this grace period, data can be formally handed over to the successor project and with this, storage life can be extended. A formal process for this hand-over has to be developed by LRZ.

Exceptions are possible and have to be approved by the steering comitee

For **Open Data**, meaning data on SuperMUC-NG DSS and InHPC-DE DSS, that a SuperMUC-NG project makes freely available to the science community, the maximum storage life is not organisational limited. This means that we try to retain the data on a best effort basis for as long as the data is valuable for the science community, a contact person who takes responsibility for the data exists and reconfirms prolongation of the data storage on a yearly basis and funding by GCS or the government respectively is secured. If the contact person does not reconfirm prolongation to the requested yearly dates, data is automatically deleted after a grace period of 1 month. Additionally, LRZ reserves the right to run audits in reasonable intervals to verify that the data on SuperMUC-NG and InHPC-DE DSS still provides reasonable scientific value. The projects have to agree to actively take part in these audits and provide feedback in a timely manner.

5. Application Process

The application process for storage on SuperMUC-NG and InHPC-DE DSS is determined by the amount of storage needed. Basically we distinguish between 3 cases:

1. The amount of requested storage is smaller than 0,2TB per granted million core hours of the compute project that requests DSS storage (**Small Allocation**)
2. The amount of requested storage exceeds the limit of number one but is still less than or equal to 1PB in total (**Medium Allocation**)
3. The amount of requested storage exceeds 1PB in total (**Large Allocation**)

5.1. Application Process for Small Allocations
For projects that require just a Small Allocation on SuperMUC-NG and InHPC-DE DSS, there is no formal approval process, as these requests are pre-approved. The application is done by filling out the SuperMUC-NG_DSS_Data_Management_Plan_Small.xlsx and submitting it via the LRZ Service desk to the Data Science Storage Technical Team.

5.2. Application Process for Medium Allocations

For projects that require a Medium Allocation on SuperMUC-NG and InHPC-DE DSS, an LRZ-internal approval is required. The application is done by filling out the SuperMUC-NG_DSS_Data_Management_Plan.xlsx and submitting it via the LRZ Service desk or already together with the SuperMUC-NG compute project application. The Data Management Plan is then reviewed by the Data Science Storage Governance Team which decides of the allocation is granted or not. For applications rejected by the LRZ-internal committee there is the possibility for the applicant to request an escalation if his or her request to the SuperMUC-NG steering committee.

5.3. Application Process for Large Allocations

For projects that require a Large Allocation on SuperMUC-NG and InHPC-DE DSS, an approval from the SuperMUC-NG steering committee is required. The application is done by filling out the SuperMUC-NG_DSS_Data_Management_Plan.xlsx and submitting it via the LRZ Service desk or already together with the SuperMUC-NG compute project application. The Data Science Storage Governance Team reviews the DMP and passes the request along together with its remarks to the SuperMUC-NG steering committee which decides on the application.

5.4. Members of the Data Science Storage Governance Team

The following people are members of the LRZ Data Science Storage Governance Team

- Herbert Huber
- Werner Baur
- Matthias Brehm
- Nicolay Hammer
- Stephan Hachinger

6. Cost

SuperMUC-NG and InHPC-DE Data Science Storage services are provided free of charge.

7. Forms

- SuperMUC-NG_DSS_Data_Management_Plan_Small.xlsx
- SuperMUC-NG_DSS_Data_Management_Plan.xlsx

---

Approved by SuperMUC Steering Committee 2019-02-22