Leibniz Supercomputing Centre

of the Bavarian Academy of Sciences and Humanities

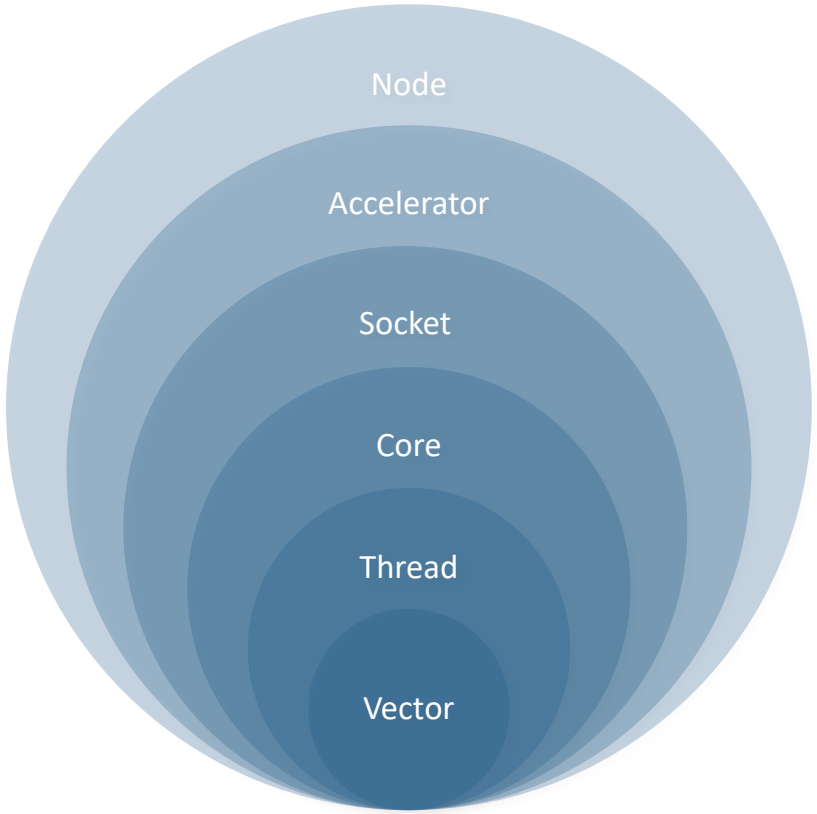
AI Training Series: Orientation Session

October, 9th/10th 2023 | J. Albert-von der Gönna, F. Dufour, M. Ohlerich

Panorama of systems @ LRZ

SuperMUC-NG Phase 1

Levels of Parallelism



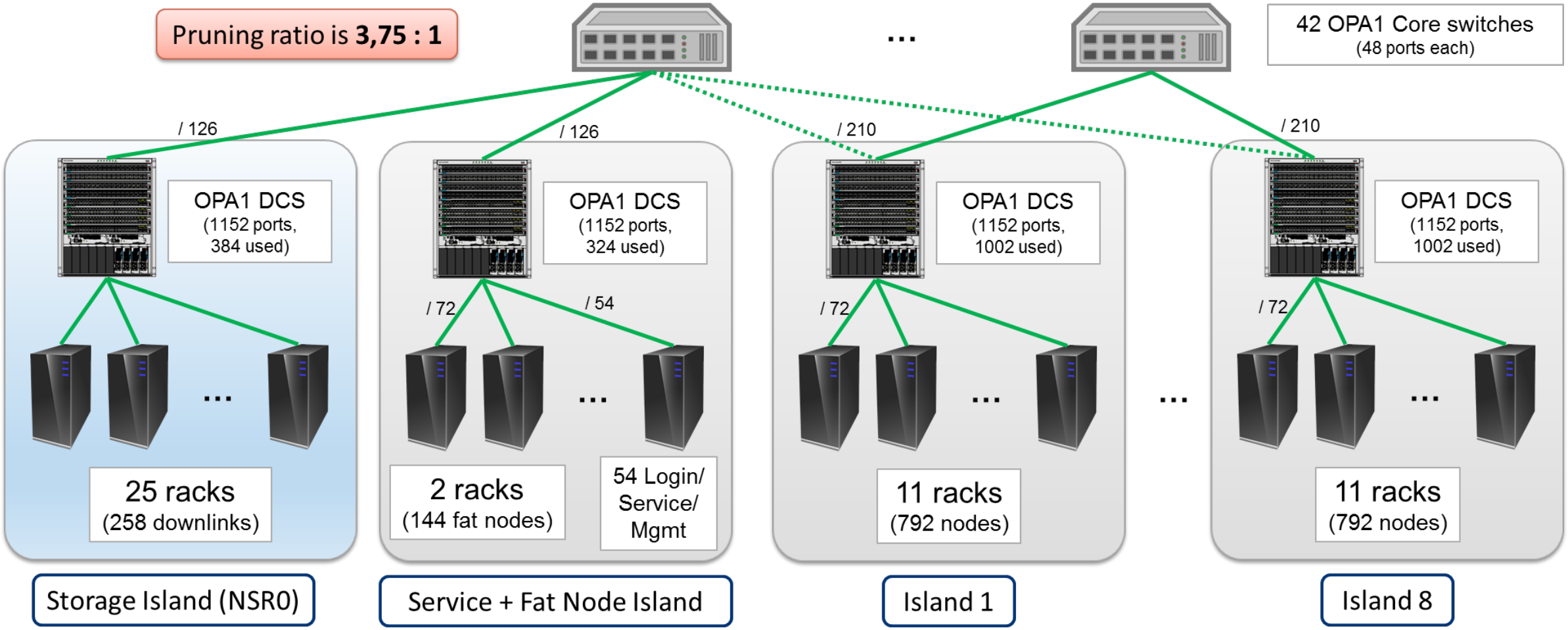
- **Node Level** (*e.g.,* SuperMUC-NG Phase 1 has 6480 nodes)
- ***Accelerator Level*** *(e.g., a Nvidia DGX A100 has 8 GPUs)*
- **Socket Level** (*e.g.,* Linux Cluster Teramem has 4 sockets [with 24 cores each])
- **Core Level** (*e.g.,* Linux Cluster CoolMUC-3 nodes have 64 cores [on a single socket])
- ***Thread Level*** *(e.g., Linux Cluster CoolMUC-2 nodes allow 2 threads per core)*
- **Vector Level** (*e.g.,* AVX-512 has 32 512-bit vector registers)

SuperMUC-NG Phase 1 theoretical peak performance:

6480 Nodes x **2** Sockets x **24** Cores x **32** Vectors x **2,7** GHz

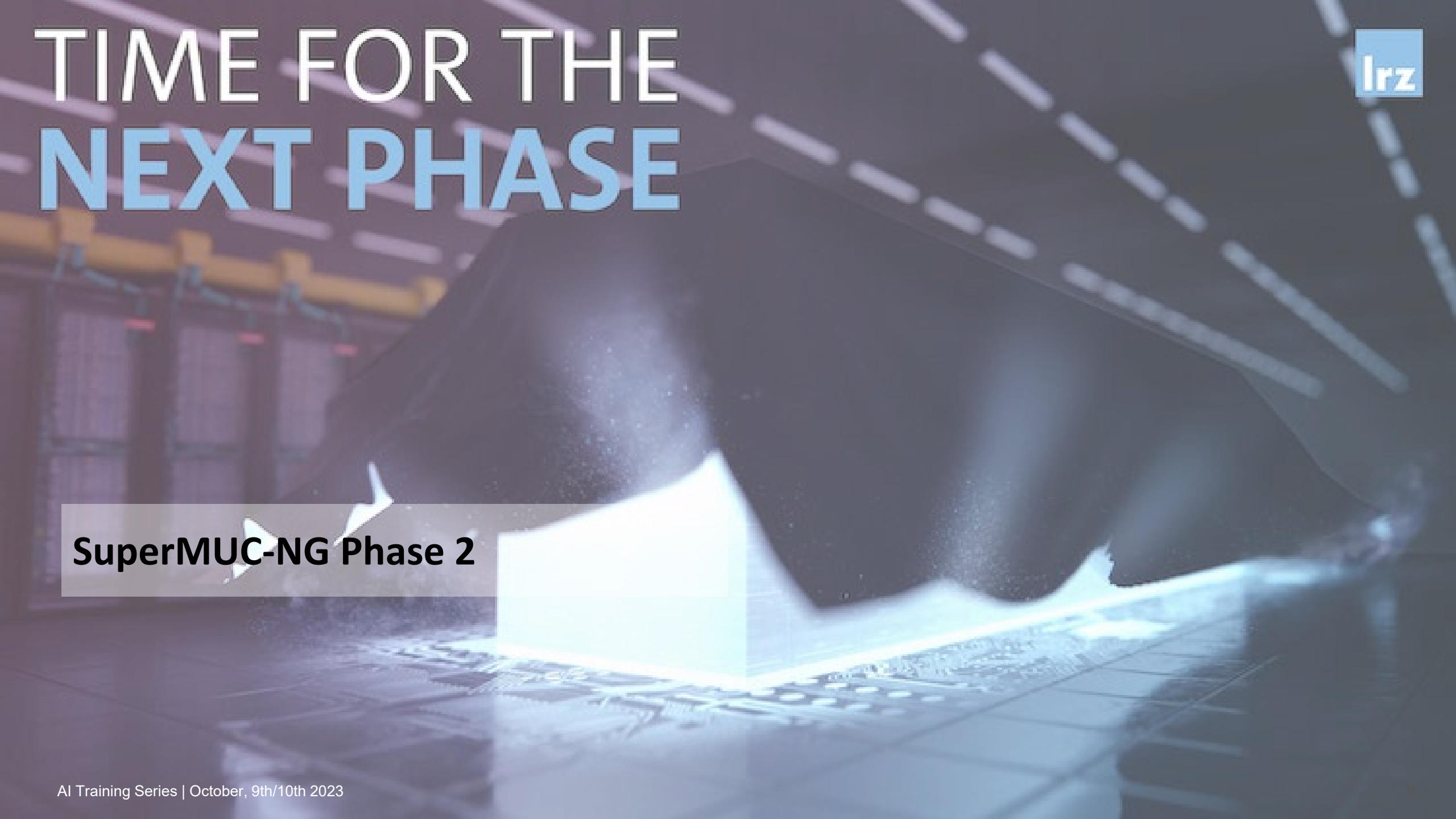
= 26 873 856 000 000 000 Flop/s

SuperMUC-NG Phase 1: High Level System Architecture



SuperMUC-NG Phase 1: Hardware Overview

Compute Nodes	Thin Nodes	Fat Nodes	Total (Thin + Fat)
Processor Type	Intel Skylake Xeon Platinum 8174	Intel Skylake Xeon Platinum 8174	Intel Skylake Xeon Platinum 8174
Cores per Node	48	48	48
Memory per Node [GByte]	96	768	N/A
Number of Nodes	6,336	144	6,480
Number of Cores	304,128	6,912	311,040
Peak Performance @ nominal [PFlop/s]	26.3	0.6	26.9
Linpack [PFlop/s]	–	–	19.476
Memory [TByte]	608	111	719
Number of Islands	8	1	9
Nodes per Island	792	144	N/A
Filesystems			
High Performance Parallel Filesystem	50 PiB @ 500 GB/s		
Data Science Storage	20 PiB @ 70 GB/s		
Home Filesystem	256 TiB		
Infrastructure			
Cooling	Direct warm water cooling		
Waste Heat Reuse	For producing cold water with adsorption coolers		
Software			
Operating System	Suse Linux Enterprise Server (SLES)		
Batch Scheduling System	SLURM		
High Performance Parallel Filesystem	IBM Spectrum Scale (GPFS)		
Programming Environment	Intel Parallel Studio XE, GNU compilers		
Message Passing	Intel MPI, (OpenMPI)		



SuperMUC-NG Phase 2

SuperMUC-NG Phase 2:

- The system will be equipped with next-generation **Intel Xeon Scalable processors** (codenamed Sapphire Rapids) **and "Ponte Vecchio", Intel's upcoming GPU** based on the Xe-HPC micro-architecture **for high performance computing and AI**.

- The **storage system** will feature distributed asynchronous object storage (**DAOS**).

- **240** Intel **compute nodes**

- Its **capacity is 1 petabyte of data storage**, but more importantly, this technology enables fast throughput of large data volumes. This system architecture is **particularly well-suited** then **for highly scalable, compute- and data-intensive workloads and artificial intelligence applications**.

While the DAOS storage system is expected to arrive in Garching in the fall of 2021, the compute system will follow in the spring of 2022… 😐

https://www.lrz.de/presse/ereignisse/2021-05-04-SuperMUC-NG-Phase-2_ENG/

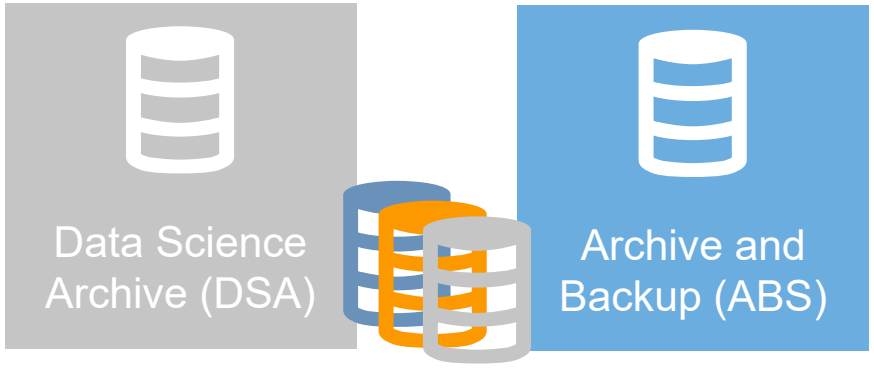
SuperMUC-NG: Access

There are three options for research groups with a Principal Investigator (PI) affiliated with a German University or German Research Institutions:

1. GCS test project: rolling call, fast review (short abstract), < 300.000 core-h
2. GCS regular project: rolling call, technical & scientific review, < 45m core-h, 2 years
3. GCS large scale project: biannual, technical & scientific review, > 45m core-h, 1 year

For further details, see
https://doku.lrz.de/application-for-a-project-on-supermuc-ng-11482823.html

HPC & BDAI Systems for Bavarian Universities



Data Science Storage (DSS)

LRZ Linux Cluster

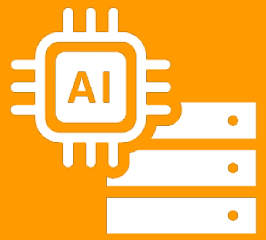
CoolMUC-2 Teramem-2 CoolMUC-3

lxlogin[1-4].lrz.de

lxlogin8.lrz.de

LRZ AI Systems

- Multiple “Big Data” CPU nodes
- HPE P100 node
- Multiple V100 nodes
- DGX-1 P100, DGX-1 V100
- Multiple DGX A100



login.ai.lrz.de
(ssh:// and https://)

LRZ Compute Cloud

LRZ Compute Cloud
(w/ some GPUs)

https://cc.lrz.de

HPC & BDAI Systems for Bavarian Universities: Account Management

- In order to use LRZ services provided to Bavarian universities, a "**LRZ Kennung**" (user ID belonging to an LRZ project**) with appropriate permission**s is needed.
  While student/staff accounts from LMU and TUM are (to some extent) managed by LRZ, they are restricted to certain services (e.g., E-Mail, Cloud Storage, LRZ Sync+Share) and can not be used to obtain access to most (high performance) systems
  (see https://doku.lrz.de/x/64N6Aw for an overview).
- Department/institute heads and/or professors/PIs can request **new LRZ projects** and appoint a **master user** (or more) for the project. The master user(s) can manage IDs and permissions within these dedicated LRZ projects.

Linux Cluster

Linux Cluster: Hardware Overview

Name	CPU	Cores/Node	RAM/Node (GB)	Nodes (total)	Cores (total)
CoolMUC-2	Intel Xeon E5-2690 v3 ("Haswell")	28	64	812	22736
CoolMUC-3	Intel Xeon Phi (“Knights Landing”)	64	96	148	9472
Teramem(-2)	Intel Xeon Platinum 8360HL (“Cooper Lake”)	96	6144	1	96
CoolMUC-4	tba	tba	tba	tba	tba

https://doku.lrz.de/display/PUBLIC/Compact+Guide+to+first-time+Linux+Cluster+Access+Process

Linux Cluster: Access in Case LRZ Project Exists

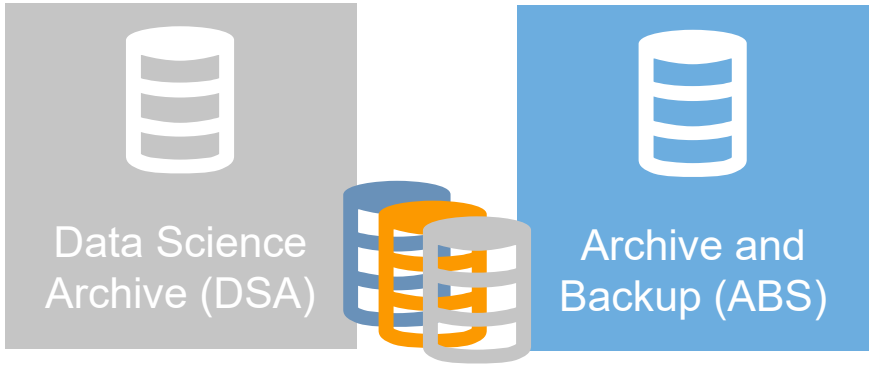
- The master user has to check if the LRZ project is already eligible for Linux Cluster usage.
  - If not, the master user must contact the LRZ contact person for the project (advisor).
    The LRZ advisor will then explain the next steps to the master user.
  - If the LRZ project is eligible for Linux Cluster usage, the master user can create a new personal LRZ user ID (or import an existing one) and enable Linux Cluster access rights for this ID through the LRZ Identity Management (IDM) Portal.
    The master user will need your nationality. Please provide this information. It is a necessary requirement for the export control regulations affecting all HPC/HPDA/HPAI services at LRZ.
- After you get the new user ID from your master user, please use the password reset function of the LRZ IDM Portal using your new ID and your contact e-mail address:
  https://idmportal.lrz.de/pwreset

https://doku.lrz.de/display/PUBLIC/Compact+Guide+to+first-time+Linux+Cluster+Access+Process

Linux Cluster: Access in Case No LRZ Project Exists

- Your chair/research group has to apply for a new LRZ project.
- Use PDF application form "Antrag auf ein LRZ-Projekt“ to be found here: https://doku.lrz.de/x/CgCiAQ (only available in German, unfortunately)
  - Pay attention to "Gewünschte LRZ-Serviceklassen" in the application form. For Linux Cluster access you need to
    - select “High Performance Computing” and
    - fill in the phrase “Linux Cluster” at “andere Dienste:”
  - Send the filled in and signed application form to the responsible LRZ contact person (advisor). The original document is needed (you may send a scanned copy to speed up the process, but this does not replace sending the physical letter via snail mail…).
- The master user of the newly requested LRZ project will get instructions from the LRZ advisor:
  - Fill out the Service Request Template for "Linux Cluster Project Activation" and submit it to LRZ Servicedesk (https://servicedesk.lrz.de/en/ql/createsr/12)
  - Afterwards, the Linux Cluster access for the new LRZ project is typically approved
  - Now, your new master user can create new LRZ user IDs (or import existing ones) and provide them with access to the Linux Cluster (see previous slide)

HPC & BDAI Systems for Bavarian Universities



Data Science Storage (DSS)

LRZ Linux Cluster

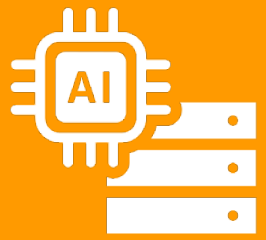
CoolMUC-2 Teramem-2 CoolMUC-3

lxlogin[1-4].lrz.de

lxlogin8.lrz.de

LRZ AI Systems

- Multiple “Big Data” CPU nodes
- HPE P100 node
- Multiple V100 nodes
- DGX-1 P100, DGX-1 V100
- Multiple DGX A100



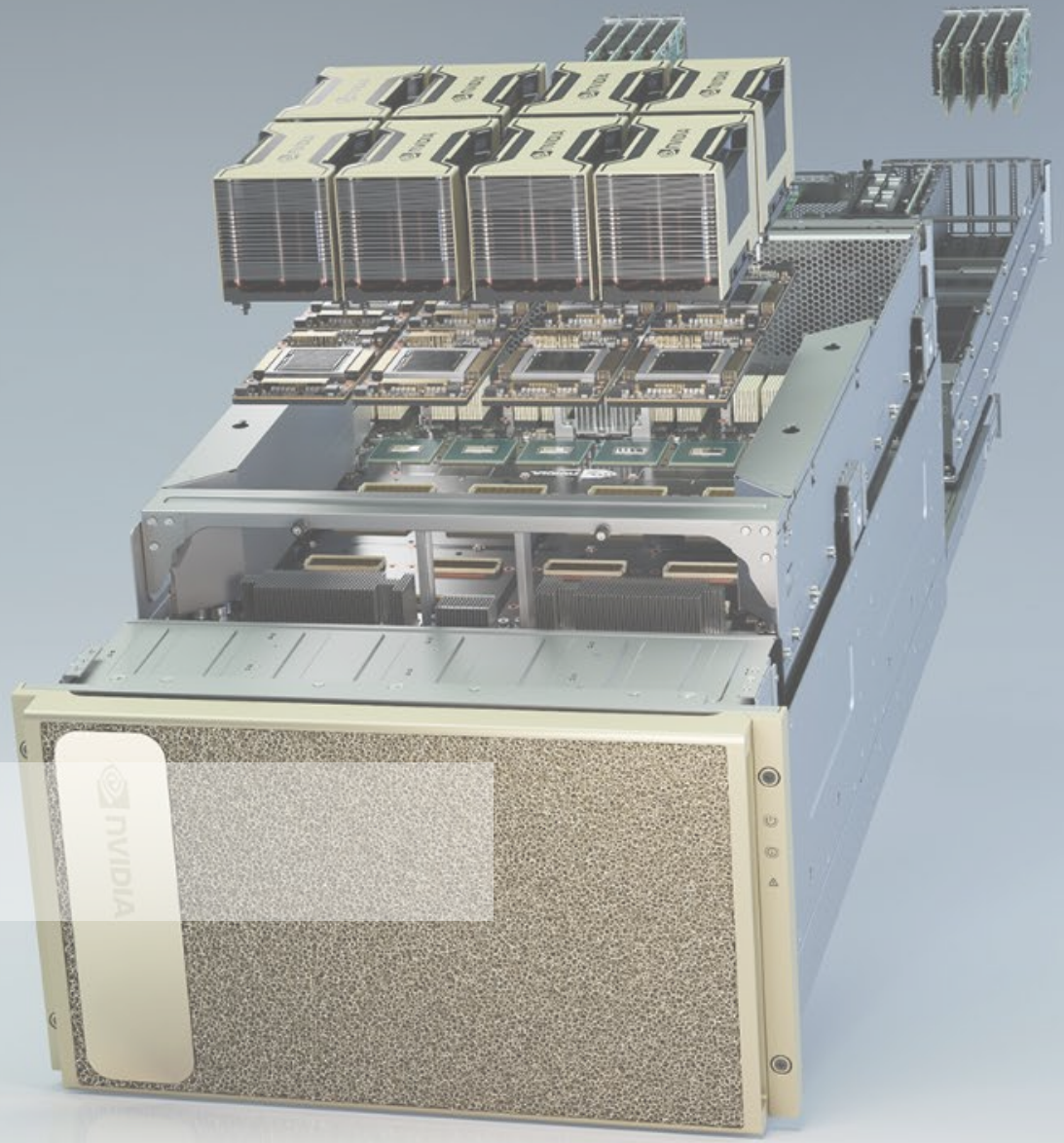
login.ai.lrz.de
(ssh:// and https://)

LRZ Compute Cloud

LRZ Compute Cloud
(w/ some GPUs)

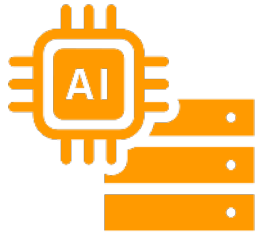
https://cc.lrz.de

LRZ AI Systems



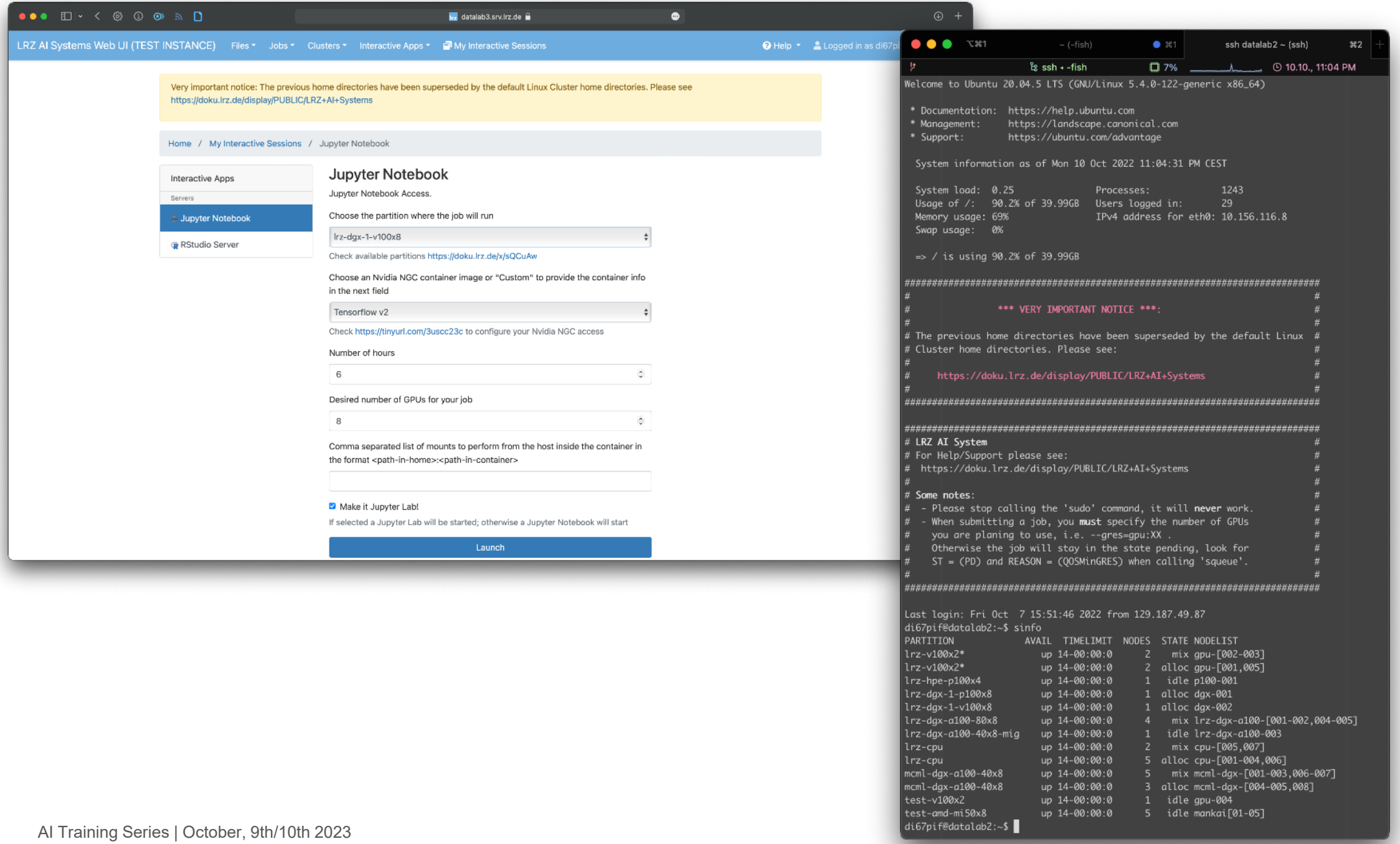
(BD)AI Systems: Hardware Overview

Type	Nodes	CPUs (Node)	Memory (Node)	GPUs (Node)	Memory (GPU)
CPU Nodes	12	up to 20	up to 800GB	-	-
HPE P100 Node	1	64	256 GB	4x P100	16 GB
V100 Nodes	4	40	368 GB	2x V100	16 GB
DGX-1 P100	1	80	512 GB	8x P100	16 GB
DGX-1 V100	1	80	512 GB	8x V100	16 GB
DGX A100/40	1	256	1 TB	8x A100	40 GB
DGX A100/80	4	256	2 TB	8x A100	80 GB



(BD)AI Systems: Hardware Overview

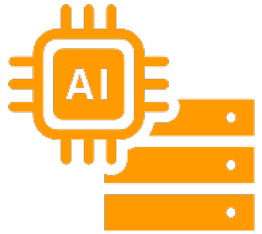
- 3 xNVIDIA A100 nodes rack-mounted at the Argonne National Lab
- 143kg / node
- 8 GPUs / node
- 400 W
- (not actually made of gold)


datalab3.srv.lrz.de
LRZ AI Systems Web UI (TEST INSTANCE)
Files
Jobs
Clusters
Interactive Apps
My Interactive Sessions
Help
Logged in as di67pi
Very important notice: The previous home directories have been superseded by the default Linux Cluster home directories. Please see
https://doku.lrz.de/display/PUBLIC/LRZ+AI+Systems
Home / My Interactive Sessions / Jupyter Notebook
Interactive Apps
Servers
Jupyter Notebook
RStudio Server
Jupyter Notebook
Jupyter Notebook Access.
Choose the partition where the job will run
lrz-dgx-1-v100x8
Check available partitions https://doku.lrz.de/x/sQCuAw
Choose an Nvidia NGC container image or "Custom" to provide the container info in the next field
Tensorflow v2
Check https://tinyurl.com/3uscc23c to configure your Nvidia NGC access
Number of hours
6
Desired number of GPUs for your job
8
Comma separated list of mounts to perform from the host inside the container in the format <path-in-home>:<path-in-container>
Make it Jupyter Lab!
If selected a Jupyter Lab will be started; otherwise a Jupyter Notebook will start
Launch
~ (-fish)
ssh datalab2 ~ (ssh)
ssh • -fish
7%
10.10., 11:04 PM
Welcome to Ubuntu 20.04.5 LTS (GNU/Linux 5.4.0-122-generic x86_64)
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage
System information as of Mon 10 Oct 2022 11:04:31 PM CEST
System load: 0.25 Processes: 1243
Usage of /: 90.2% of 39.99GB Users logged in: 29
Memory usage: 69% IPv4 address for eth0: 10.156.116.8
Swap usage: 0%
=> / is using 90.2% of 39.99GB
*** VERY IMPORTANT NOTICE ***:
The previous home directories have been superseded by the default Linux Cluster home directories. Please see:
https://doku.lrz.de/display/PUBLIC/LRZ+AI+Systems
LRZ AI System
For Help/Support please see:
https://doku.lrz.de/display/PUBLIC/LRZ+AI+Systems
Some notes:
- Please stop calling the 'sudo' command, it will never work.
- When submitting a job, you must specify the number of GPUs you are planing to use, i.e. --gres=gpu:XX .
Otherwise the job will stay in the state pending, look for ST = (PD) and REASON = (QOSMinGRES) when calling 'squeue'.
Last login: Fri Oct 7 15:51:46 2022 from 129.187.49.87
di67pif@datalab2:~$ sinfo
PARTITION AVAIL TIMELIMIT NODES STATE NODELIST
lrz-v100x2* up 14-00:00:0 2 mix gpu-[002-003]
lrz-v100x2* up 14-00:00:0 2 alloc gpu-[001,005]
lrz-hpe-p100x4 up 14-00:00:0 1 idle p100-001
lrz-dgx-1-p100x8 up 14-00:00:0 1 alloc dgx-001
lrz-dgx-1-v100x8 up 14-00:00:0 1 alloc dgx-002
lrz-dgx-a100-80x8 up 14-00:00:0 4 mix lrz-dgx-a100-[001-002,004-005]
lrz-dgx-a100-40x8-mig up 14-00:00:0 1 idle lrz-dgx-a100-003
lrz-cpu up 14-00:00:0 2 mix cpu-[005,007]
lrz-cpu up 14-00:00:0 5 alloc cpu-[001-004,006]
mcml-dgx-a100-40x8 up 14-00:00:0 5 mix mcml-dgx-[001-003,006-007]
mcml-dgx-a100-40x8 up 14-00:00:0 3 alloc mcml-dgx-[004-005,008]
test-v100x2 up 14-00:00:0 1 idle gpu-004
test-amd-mi50x8 up 14-00:00:0 5 idle mankai[01-05]
di67pif@datalab2:~$

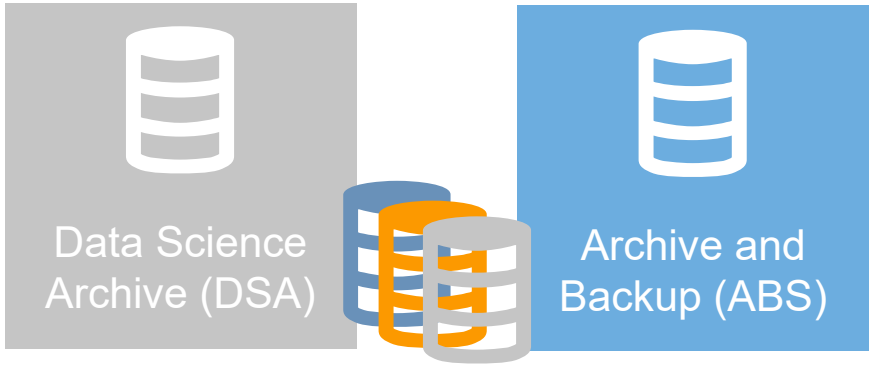
https://doku.lrz.de/display/PUBLIC/3.+Access+and+Getting+Started

(BD)AI Systems: Access

- Account management of the LRZ AI Systems is associated with the LRZ Linux Cluster:
  - If you don’t have an account for the Linux Cluster yet, you need to set that up first (see previous slides)
  - If you already have an account for the Linux Cluster, you will additionally have to request access to the LRZ AI Systems for this account
- Submit a dedicated service request to LRZ Servicedesk:
  https://servicedesk.lrz.de/en/ql/create/23
  Select "Service Request" from the drop-down list and subsequently "LRZ AI Systems - Request for Access".



HPC & BDAI Systems for Bavarian Universities



Data Science Storage (DSS)

LRZ Linux Cluster

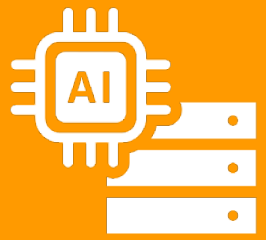
CoolMUC-2 Teramem-2 CoolMUC-3

lxlogin[1-4].lrz.de

lxlogin8.lrz.de

LRZ AI Systems

- Multiple “Big Data” CPU nodes
- HPE P100 node
- Multiple V100 nodes
- DGX-1 P100, DGX-1 V100
- Multiple DGX A100



login.ai.lrz.de
(ssh:// and https://)

LRZ Compute Cloud

LRZ Compute Cloud
(w/ some GPUs)

https://cc.lrz.de

LRZ Compute Cloud

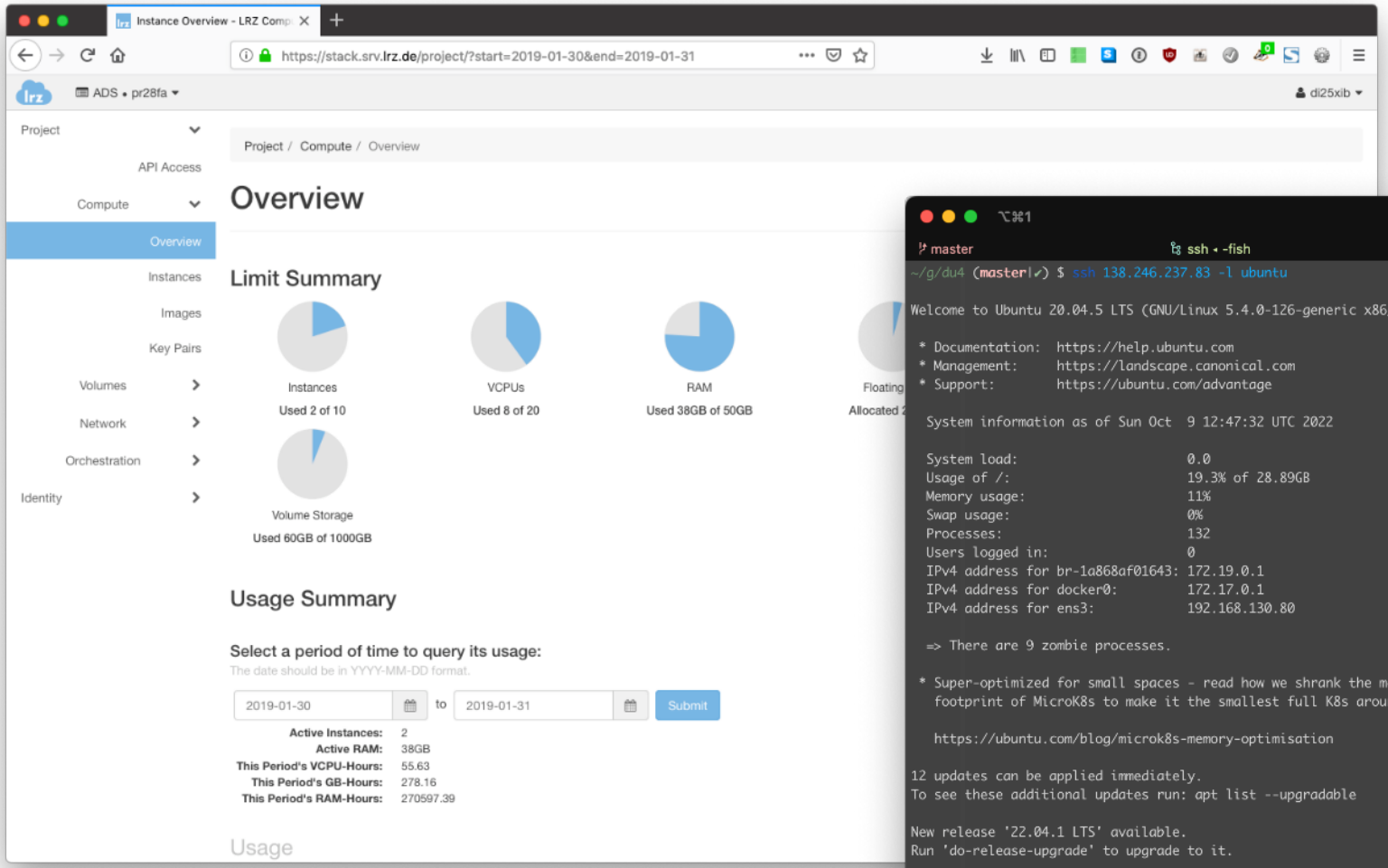
LRZ Compute cloud: Hardware Overview

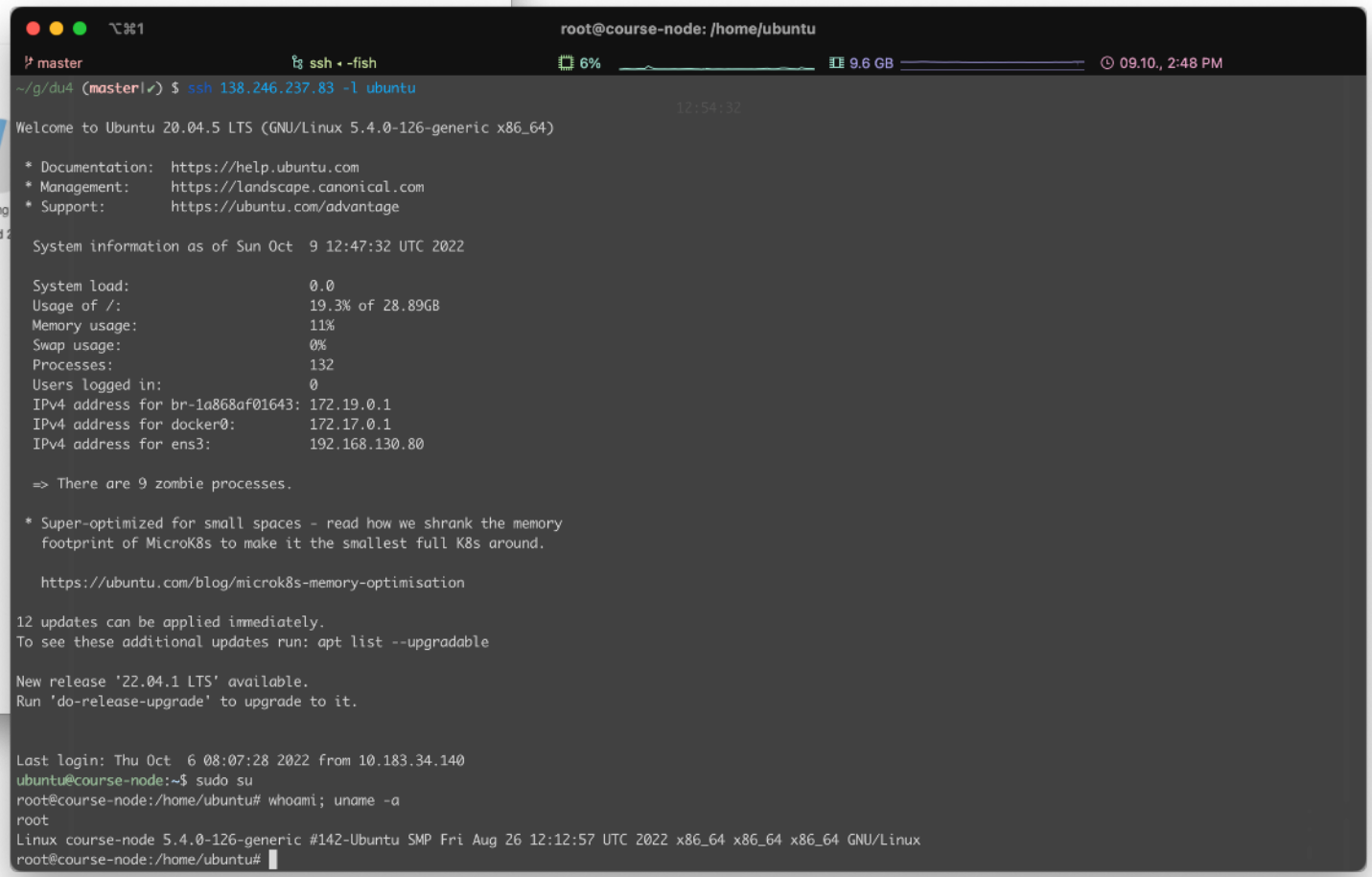
Compute	200 Nodes 192 GB to 1024 GB RAM Intel® Xeon® ~2.40 GHz
	25 x 2 GPUs Nodes 2x Nvidia Tesla V100 16 GB/node 768GB RAM/node
Storage	15 nodes 2 PB Raw Storage
Networking	100G Intel OmniPath
Software	OpenStack & CEPH

Access to more than 10 vCPUs and/or other restricted resources can be requested by contacting the cloud support team: https://servicedesk.lrz.de/ql/create/105

40000 vCPU capacity with overcommitment
2000 users and 1500 active VMs

Compute Cloud: Hardware Overview





https://doku.lrz.de/display/PUBLIC/FAQ#FAQ-HowtogetaccesstotheComputeCloud?

Compute Cloud: Access

- You need to have an **ID in a cloud-enabled LRZ project**.
  - If your user ID belongs to a project that is maintained by LRZ but not cloud-enabled yet, you might ask the project's master user(s) to contact LRZ and ask to activate this project for the compute cloud.
- The project's **master user can enable your account** for the LRZ Compute Cloud.
- It might take **some minutes to synchronize** your new permissions with the cloud system. You can not log in until your permissions have been synchronized with the cloud system. After your account's permissions have been synchronized you will receive emails providing further information on how to use the Cloud.

LRZ Data Storage

https://doku.lrz.de/display/PUBLIC/Data+Science+Storage

Data Storage: Overview

- The LRZ HPC/HPDA/HPAI Infrastructure is backed by the Data Science Storage (DSS)
  - Solution for potentially vast amounts of data
  - Long-term storage (Linux Cluster DSS) and/or high-bandwidth, low latency (AI Systems DSS)
  - Directly connected to the LRZ high-performance computing ecosystem
  - Flexible data sharing among LRZ users
  - Web interface for world-wide access and transfer
  - Data sharing with external users (invite per e-mail, access per web interface)
- Additionally, we also provide a new type of Data Archive, based on the DSS Solution stack, called Data Science Archive (DSA) (this basically relates to DSS like AWS Glacier relates to AWS S3).
- Disk space and access is managed (as DSS projects and containers) by data curators.
  This can be LRZ personnel (e.g., $HOME directories) or PIs/master users/dedicated data curators (e.g., project storage).

https://doku.lrz.de/display/PUBLIC/File+Systems+and+IO+on+Linux-Cluster

Data Storage: Linux Cluster & AI Systems

- $HOME (DSS-backed home directory, managed by LRZ)
  - 100GB per user
  - Path: /dss/dsshome1/##/<user>
  - Automatic tape backup and file system snapshots
    (see “/dss/dsshome1/.snapshots/” directory)
  - All your important files/anything you invested a lot of work into should be here
  - BUT: Not suitable for heavy and/or high-frequency I/O operations, i.e. most machine learning applications. Use the AI Systems DSS instead.

https://doku.lrz.de/display/PUBLIC/2.+Storage+on+the+LRZ+AI+Systems

Data Storage: AI Systems

- AI Systems DSS
  - Up to 4 TB per project **upon request**, typically shared among project members
  - Path(s): $ dssusrinfo all
  - Configuration (e.g., exports, quota) to be managed by data curator
  - Use this for e.g., high bandwidth, low latency I/O
  - Can also be accessed from the Linux Cluster

https://doku.lrz.de/display/PUBLIC/File+Systems+and+IO+on+Linux-Cluster

Data Storage: Linux Cluster

- Linux Cluster DSS project storage
  - Up to 10 TB per project **upon request**, shared among project members
  - Path(s): $ dssusrinfo all
  - Configuration (e.g., exports, backup, quota) to be managed by data curator
  - Use this for e.g., large raw data (and consider backup options)
  - Can also be accessed from the AI Systems

https://doku.lrz.de/display/PUBLIC/File+Systems+and+IO+on+Linux-Cluster

Data Storage: Linux Cluster

- Legacy $SCRATCH (scratch file system, “temporary file system”, not available on CoolMUC-4 and Teramem)
  - 1.4 PB, shared among all users
  - Access: /gpfs/scratch/<group>/<user>
- New $SCRATCH_DSS (not available on CoolMUC-2 compute nodes)
  - 3.1 PB, shared among all users
  - Access: /dss/lxclscratch/##/<user>
- No backup (!) and sliding window file deletion, i.e. old files will eventually be deleted (!!) – a data retention time of approx. 30 days may be assumed, but is not guaranteed
- This is the place for e.g., very large, temporary files or intermediate results, directly feeding into additional analyses
- Data integrity is not guaranteed. Do not save any important data exclusively on these file systems! Seriously, don’t do it!

https://doku.lrz.de/display/PUBLIC/DSS+documentation+for+users

Data Storage: Compute Cloud

- The storage backend of the Compute Cloud is used to host the virtual disks belonging to the VMs in the cloud. It is not meant to store large data sets. No backups are created.
- DSS containers can be made available for VMs running in the LRZ Compute Cloud without the need to copy data into the VM.
  - The data curator of the data project, to which the relevant container belongs, needs to export the container to the IP address used by your VM via NFS.
  - You should only export DSS containers to IPs that are statically assigned to and trusted by you. NFS exports follow a "host based trust" semantic, which means the DSS NFS server will trust any IP/system to which a DSS container is exported. There is no additional user authentication between NFS server and client enforced.