Leibniz Supercomputing Centre

Slurm on SuperMUC-NG at LRZ | 13.09.2024 | Dr. Alexander Block

Partner in digital transformation for science

~300
Colleagues

Since 1962
IT Services
for Science

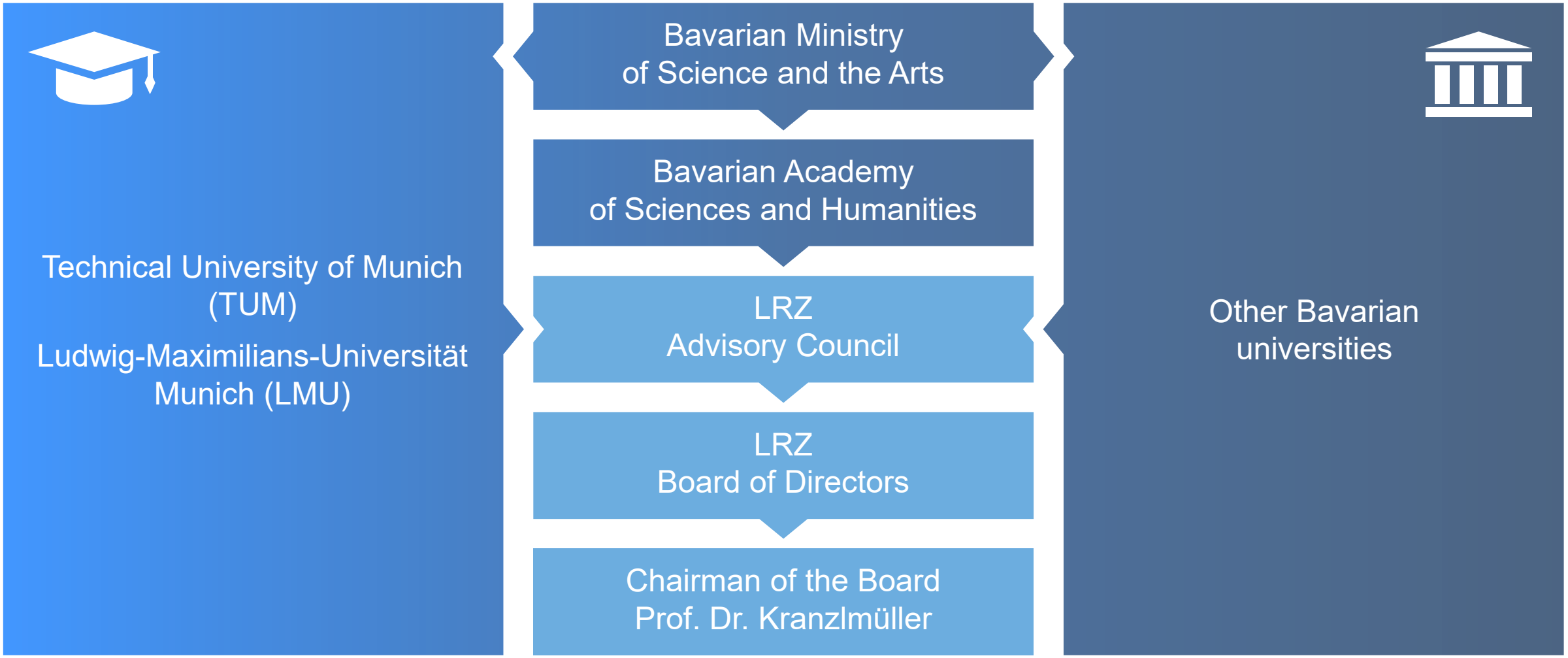
Computer Centre
for all Munich Universities

Regional Computer Centre
for all Bavarian Universities

National
Supercomputing Centre (GCS)

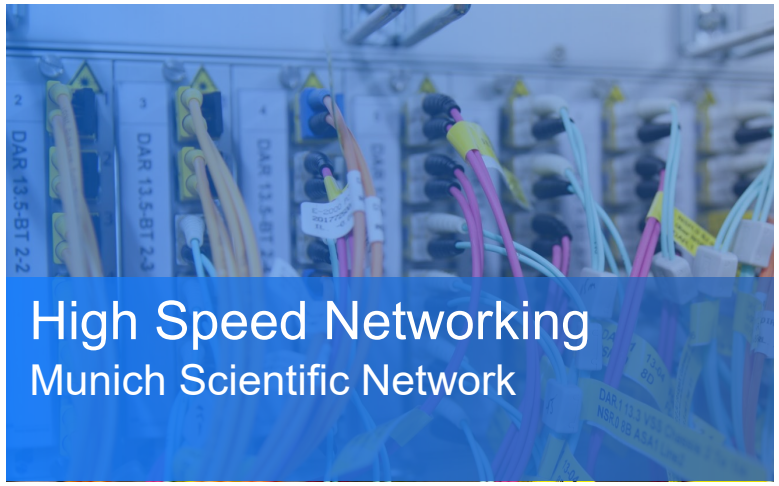
European
Supercomputing Centre

Our organisational structure



LRZ portfolio for science

IT services and technologies



High Speed Networking

Munich Scientific Network

Big Data

Bavarian State Library

Digital Archive

HPC

SuperMUC-NG, Linux Cluster,

CoolMUC-2/-3/-4

VR & Visualisation

V2C: CAVE, Powerwall

Future Computing

Technology exploration

Research & Development

LRZ as national supercomputing centre

We support science in Germany

GCS Gauss Centre for Supercomputing

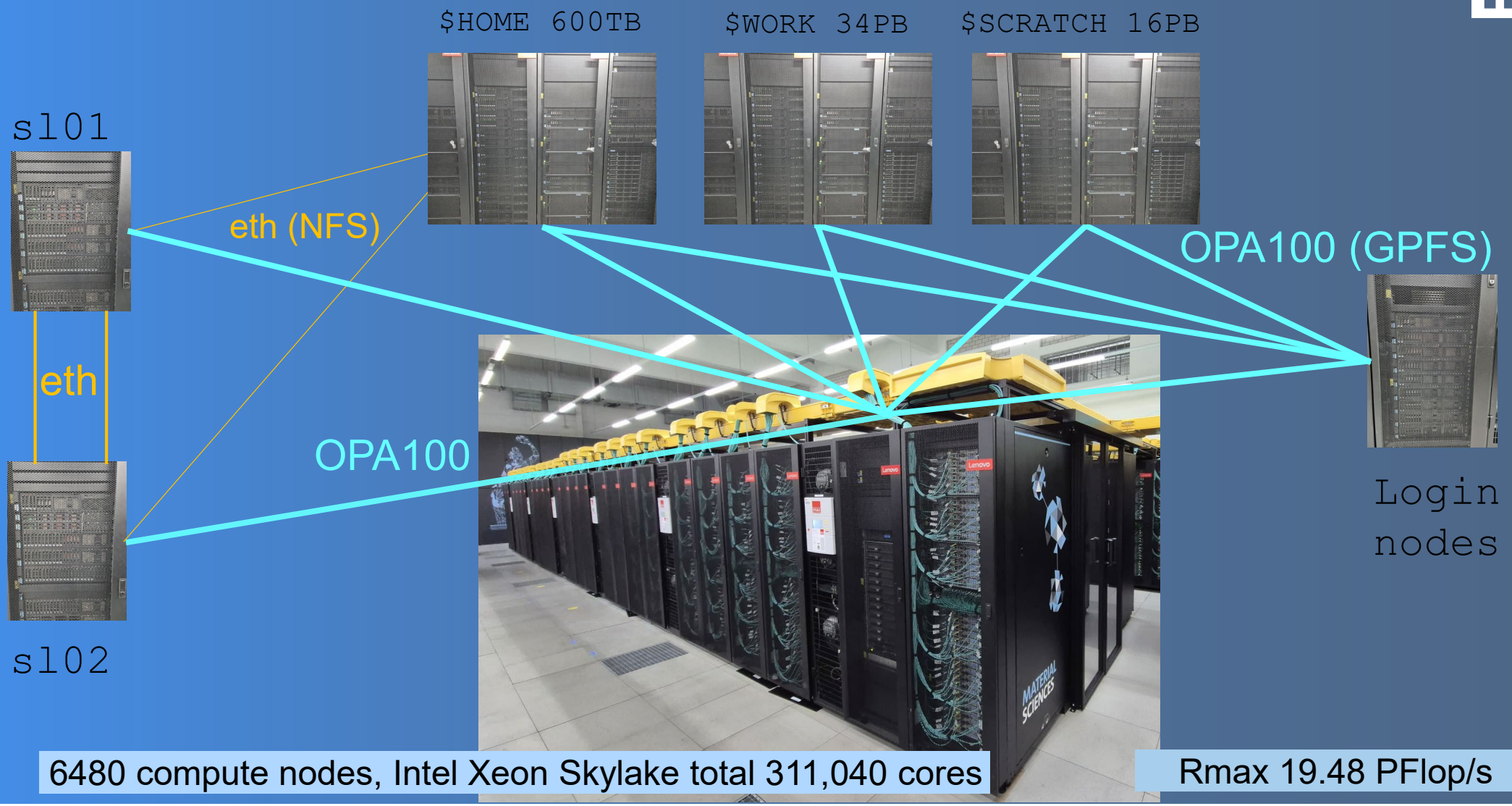
lrz

Tier-0
GCS

3
HPC centres

13/04/2007
founded

SuperMUC-NG (SNG)



6480 compute nodes, Intel Xeon Skylake total 311,040 cores

Rmax 19.48 PFlop/s

Slurm installation on SNG

- Download from SchedMD
- `rpmbuild`
- Build image (xcat)
- Install/boot image
- Diskless images for compute nodes

- Same procedure for ***munge***

➢ Updating Slurm takes more than `configure, make, install…`

- `/etc/slurm` mounted DSS shared FS ($SLURM_HOME)
- Logs also on shared DSS

Slurm configuration on SNG

slurm.conf

```
…
include /etc/slurm.specific.conf
…
```

slurm.specific.conf

```
…
SlurmdLogFile=/dss/sngslurm/log/i01/r01/slurmd.%n.log
…
SlurmctldLogFile=/var/log/slurmctld.log
…
```

- Node specific configuration for log files

slurmdbd.conf

```
…
DbdHost=sl01
DbdBackupHost=sl02
StorageHost=localhost
…
```

- HA set-up for slurmdbd

Slurm configuration on SNG

slurm.conf

```
…
PartitionName=test Nodes=i[01-04]r[01-11]c[01-06]s[01-12] AllowQOS=test MinNodes=1 MaxNodes=16 MaxTime=00:30:00 Default=YES state=up PriorityTier=2
                   PreemptMode=off

PartitionName=fat Nodes=f01r[01-02]c[01-06]s[01-12] AllowQOS=fat MinNodes=1 MaxNodes=128 MaxTime=48:00:00 PriorityJobFactor=0 state=up
                   PriorityTier=2 PreemptMode=off

PartitionName=micro Nodes=i01r[02-11]c[01-06]s[01-12],i[02-05]r[01-11]c[01-06]s[01-12] AllowQOS=micro MinNodes=1 MaxNodes=16 MaxTime=48:00:00
                   PriorityJobFactor=0 state=up PriorityTier=2 PreemptMode=off

PartitionName=general Nodes=i01r[07-11]c[01-06]s[01-12],i[02-08]r[01-11]c[01-06]s[01-12] AllowQOS=general MinNodes=17  MaxNodes=768  MaxTime=48:00:00
                   PriorityJobFactor=70 state=up PriorityTier=2 PreemptMode=off

PartitionName=large Nodes=i[03-08]r[01-11]c[01-06]s[01-12] AllowQOS=large MinNodes=769 MaxNodes=3168 MaxTime=24:00:00 PriorityJobFactor=100 state=up
                   PriorityTier=2 PreemptMode=off

PartitionName=tmp0 Nodes=f01r02c[01-06]s[01-12],i01r[02-11]c[01-06]s[01-12],i[02-08]r[01-11]c[01-06]s[01-12] AllowQOS=nolimit
                   AllowAccounts=pr28fa,pr27cu,pr27ca PriorityJobFactor=200 state=up PriorityTier=2 PreemptMode=off

PartitionName=preempt Nodes=f01r[01-02]c[01-06]s[01-12],i01r[02-11]c[01-06]s[01-12],i[02-08]r[01-11]c[01-06]s[01-12] AllowQOS=preempt MinNodes=1
                   MaxNodes=40   MaxTime=168:00:00 PriorityJobFactor=200 state=down PriorityTier=1 PreemptMode=requeue
…
```

- Overlapping partitions; different runtime, size, priority, QOS

Slurm configuration on SNG

slurm.conf

```
…
PriorityType=priority/multifactor
PriorityWeightAge=1000000
PriorityWeightJobSize=500000
PriorityWeightPartition=500000
PriorityMaxAge=14-0
…
```

- Higher priority for jobs with:
  - Longer waiting time
  - Higher node count
  - QOS/special partition

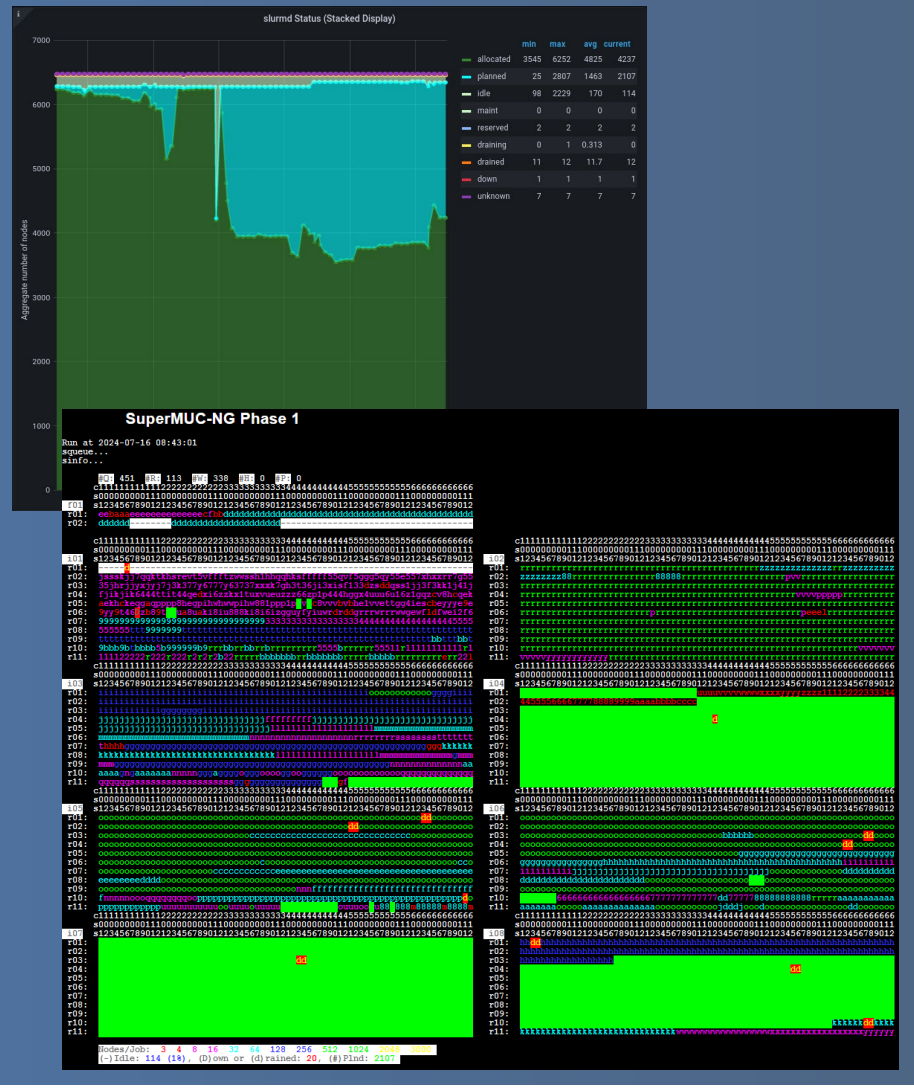
Slurm set-up on SNG

- User management
  - Custom Perl script
  - Update user/accounts twice a day from LDAP; SNG replica from LRZ master
  - Only users with compute attribute and accounts with positive budget

- Budgeting
  - Accounting in a dedicated DB, **`sacct`** once a day as input
  - Get info about current budget from DB
  - `preemption` no budget accounting

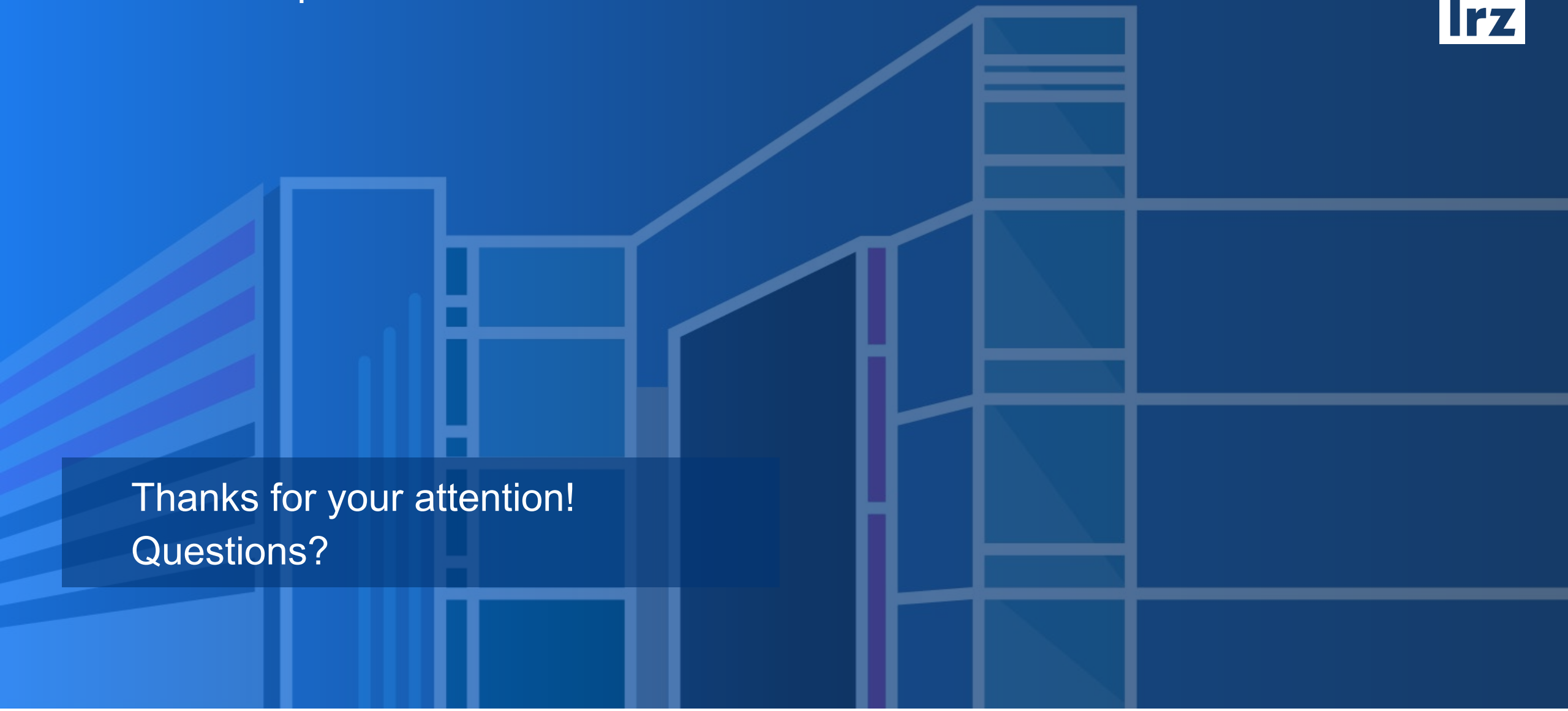
Slurm set-up on SNG

- Submit LUA
  - Setting default QOS
  - Forcing email address from LDAP user registration
- Prolog/Epilog
  - Tags for logging (SPLUNK)
  - Checks for left over processes, memory, caches
- Monitoring with Icinga2 and custom checks



Slurm set-up on SNG

- SPANK Plugin EAR
  - Energy Aware Runtime (EAR) is a system level tool for optimisation of energy consumption (https://www.eas4dc.com/ear)
  - Integrated in Slurm
    - `#SBATCH --ear=on/off`
  - Control energy consumption of the system
  - Set processor frequency
  - Power capping according to limit
- SNG power consumption ~2.5 MW



Thanks for your attention!
Questions?