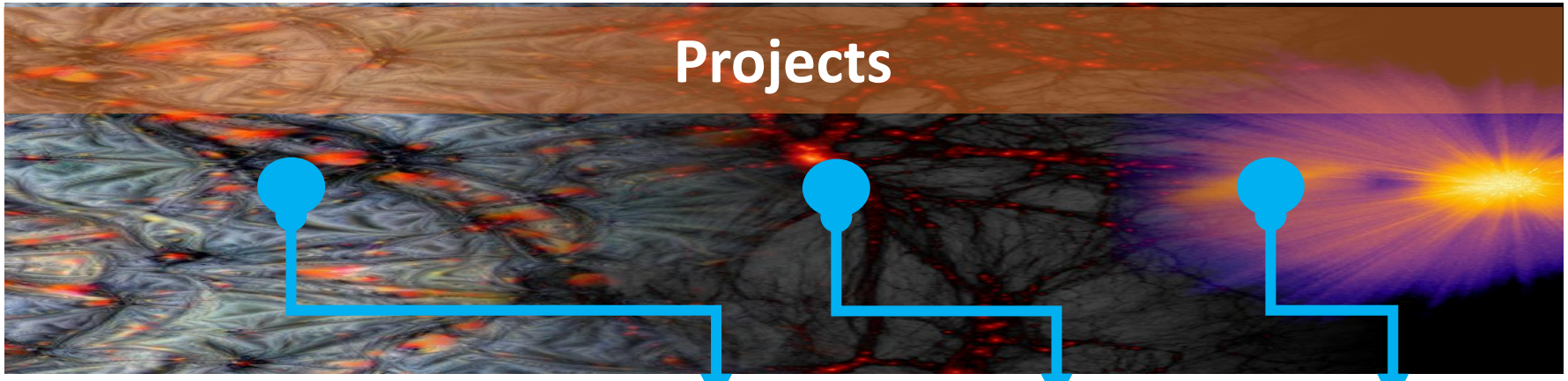
HPC at AIP

Dr A.Khalatyan

AIP,2025

Research interests



HPC
BigDATA
ML
MPI
OpenMPI
Visualisation

HESTIA:
High-resolution Environmental Simulations of The Immediate Area

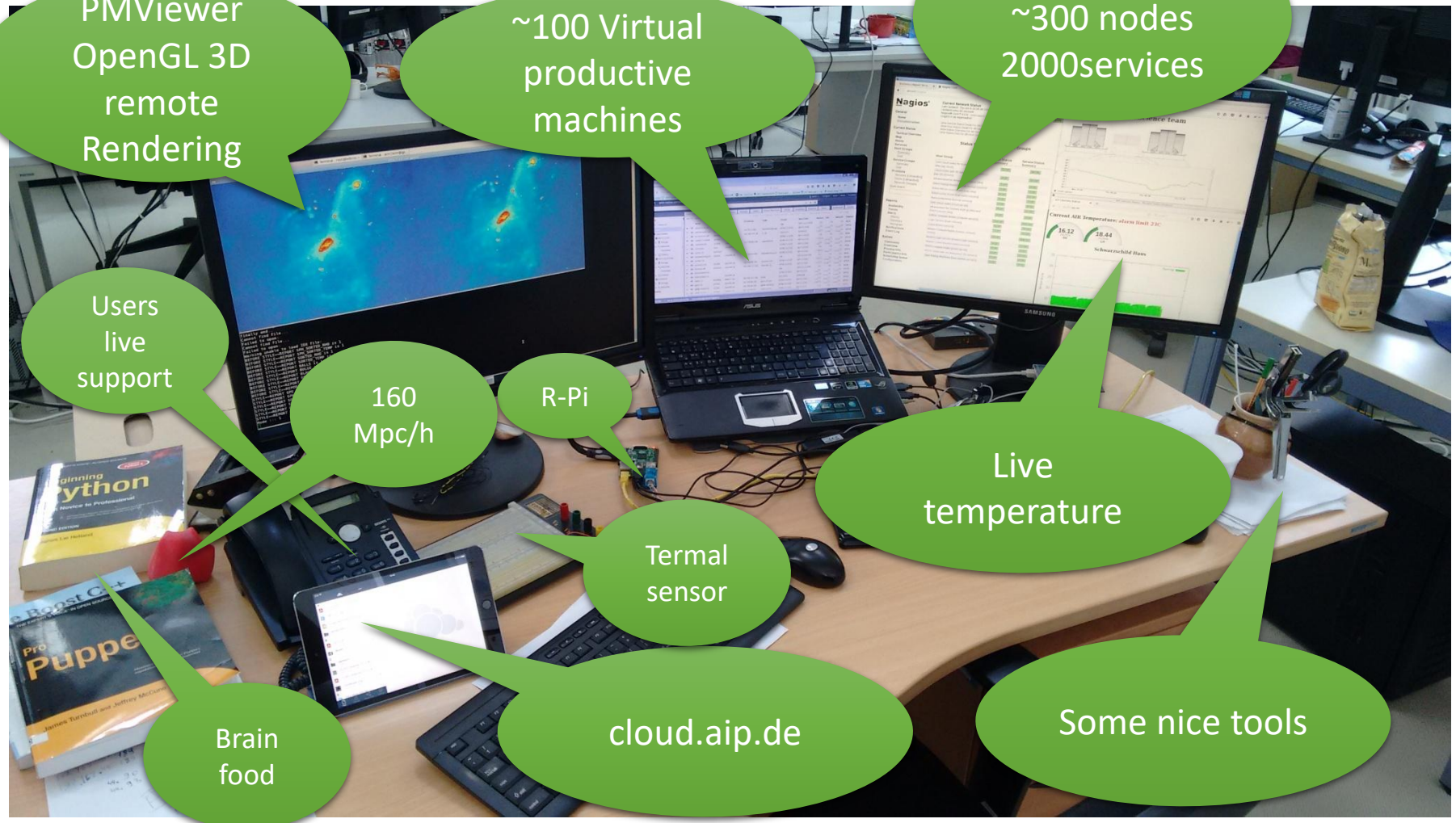
StarHorse:
Photo-astrometric distances, extinctions, and astrophysical parameters for Gaia stars brighter than G = 18

colab.aip.de cloud.aip.de vr.aip.de

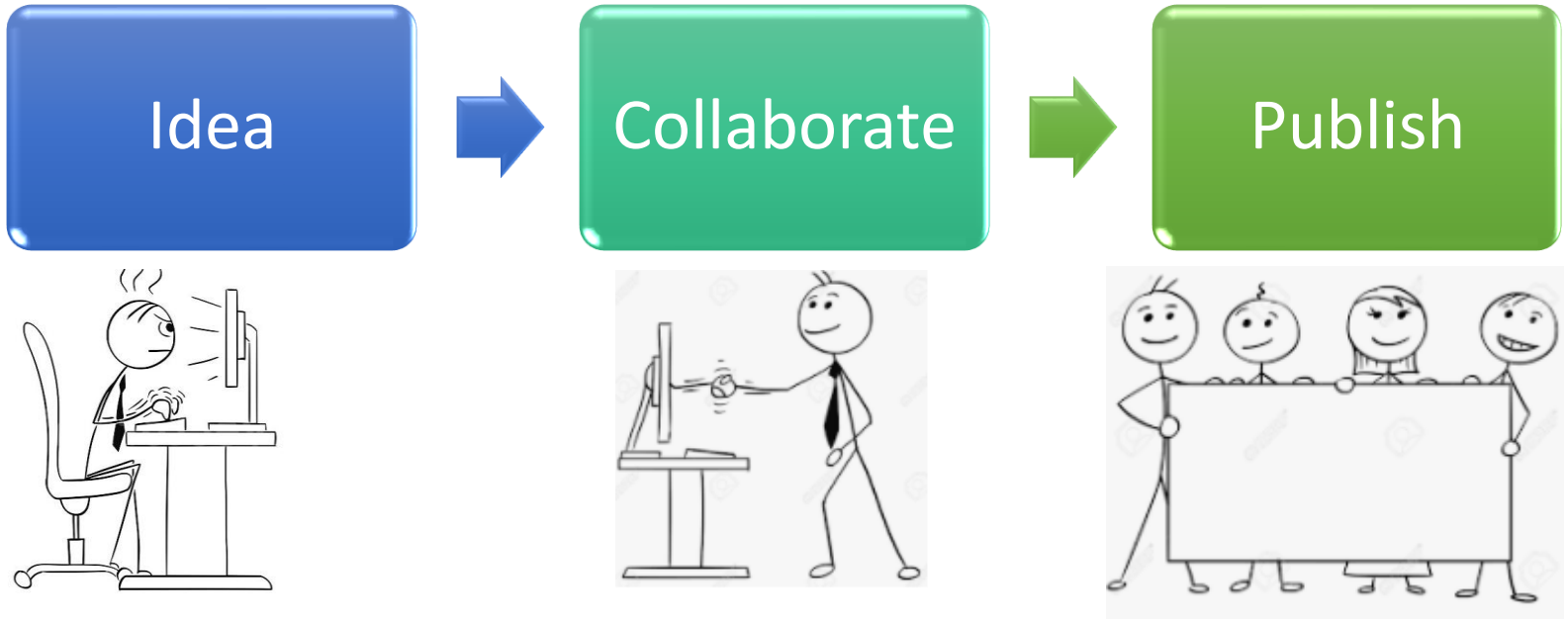
Nationale Forschungs-Daten Infrastruktur NFDI

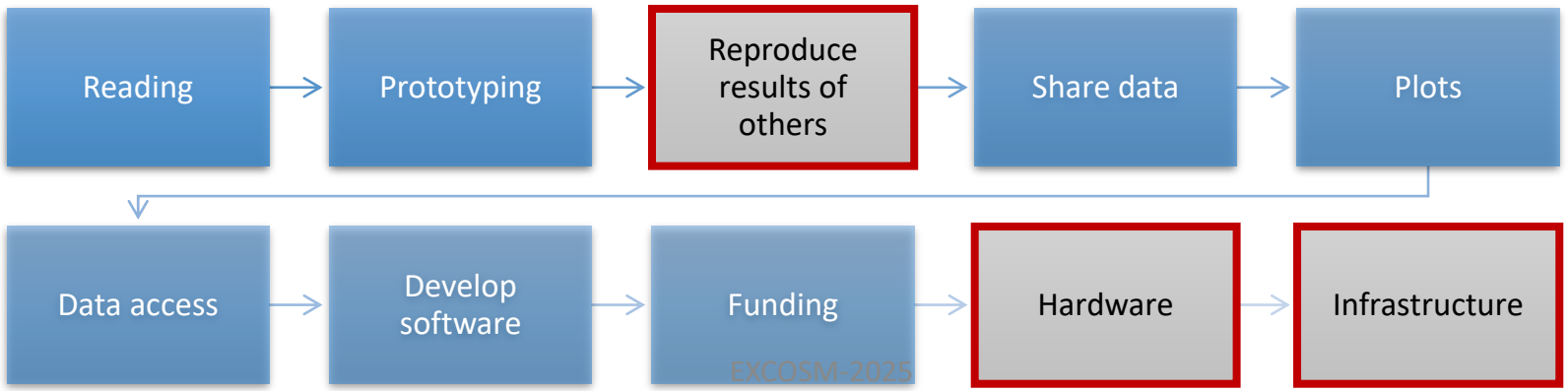
PUNCH 4NFDI

Daily workflow

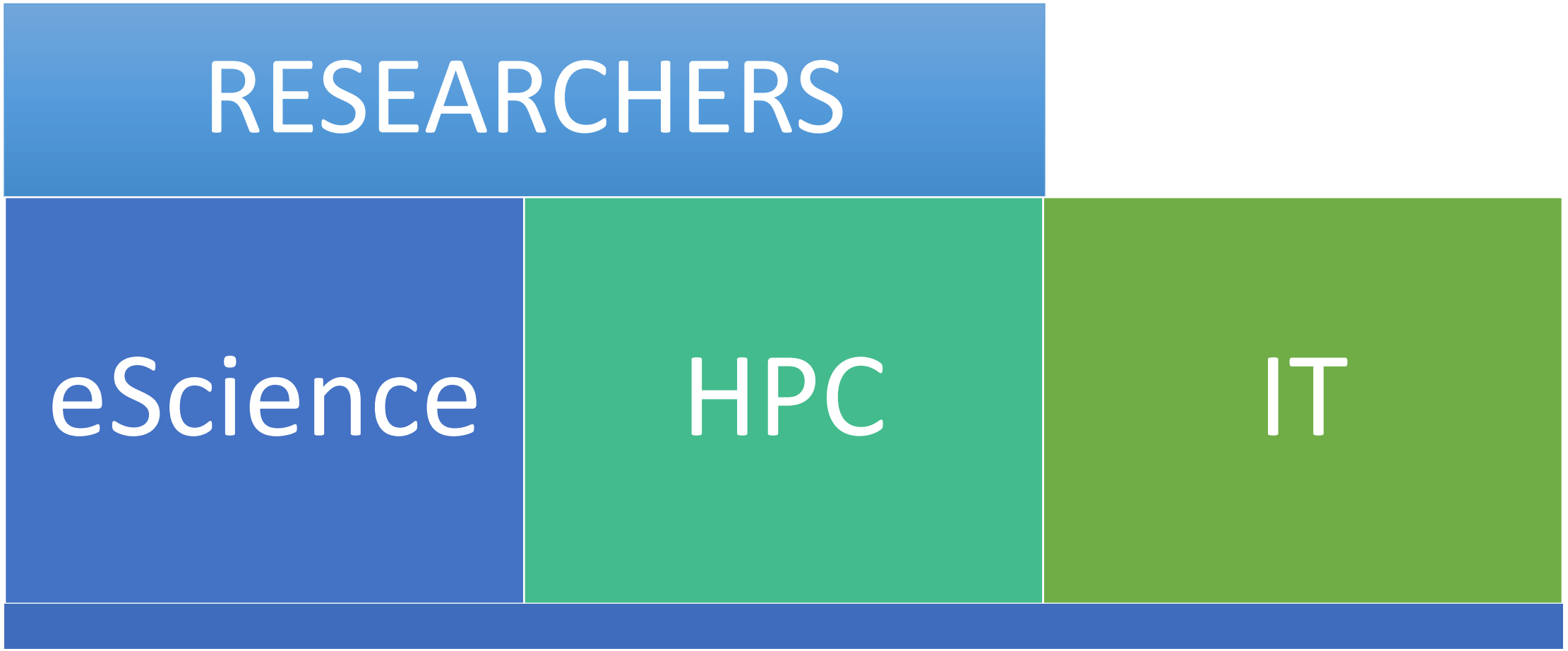


Scientific life (top to down)





eScience+HPC+IT



Infrastructure (down to top)

Funding: **EFRE**

Hardware

Compute nodes

Storage

Software

Users

Rocky Linux

oVirt

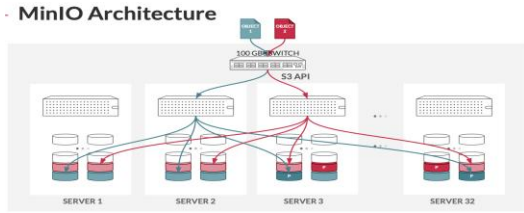
docker

RANCHER

HPC Cloud friendly

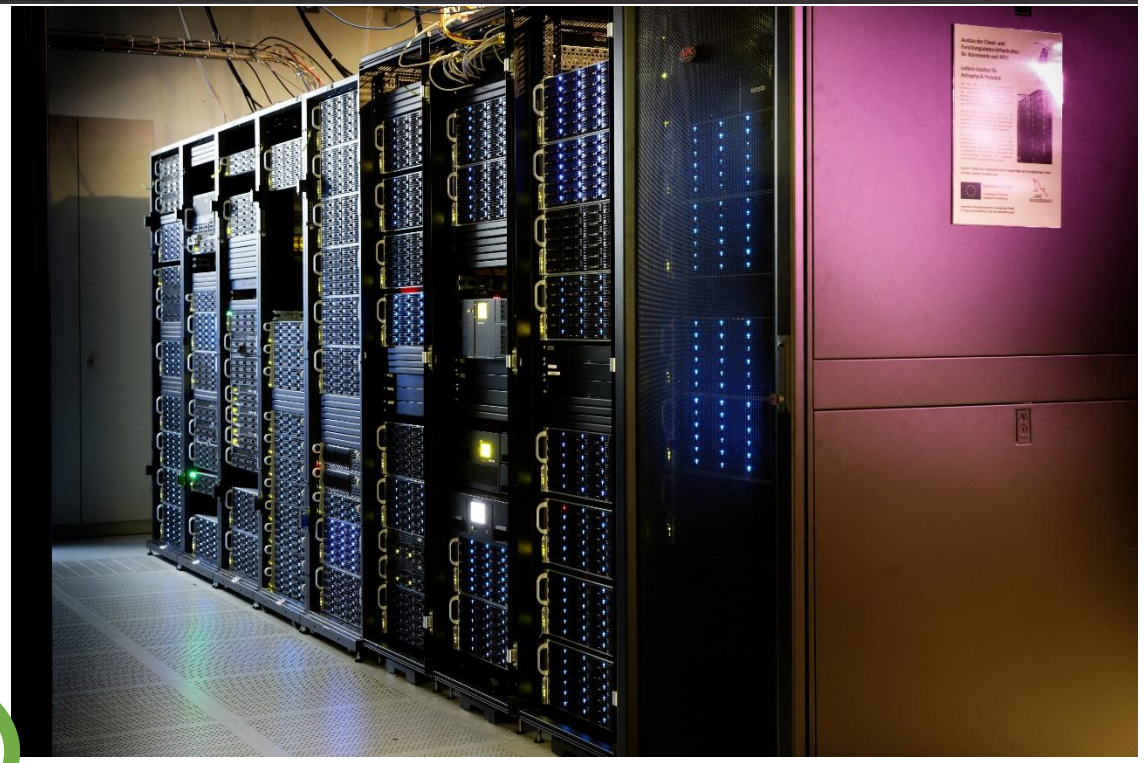
lustre™

Cloud friendly? Recycle your hardware



Newton 21

- EFRE funding ~**1.2M** Euro: *“Ausbau der Cloud- und Forschungsdaten-Infrastruktur für Astronomie und NFDI“*
- We started hardware order in Covid-19 time (end of Nov 2020) with many delivery delays on different components



01.04.2022
HW/SW –DONE!

Dieses Projekt wird unterstützt durch Fördermittel der Europäischen Union und des Landes Brandenburg.

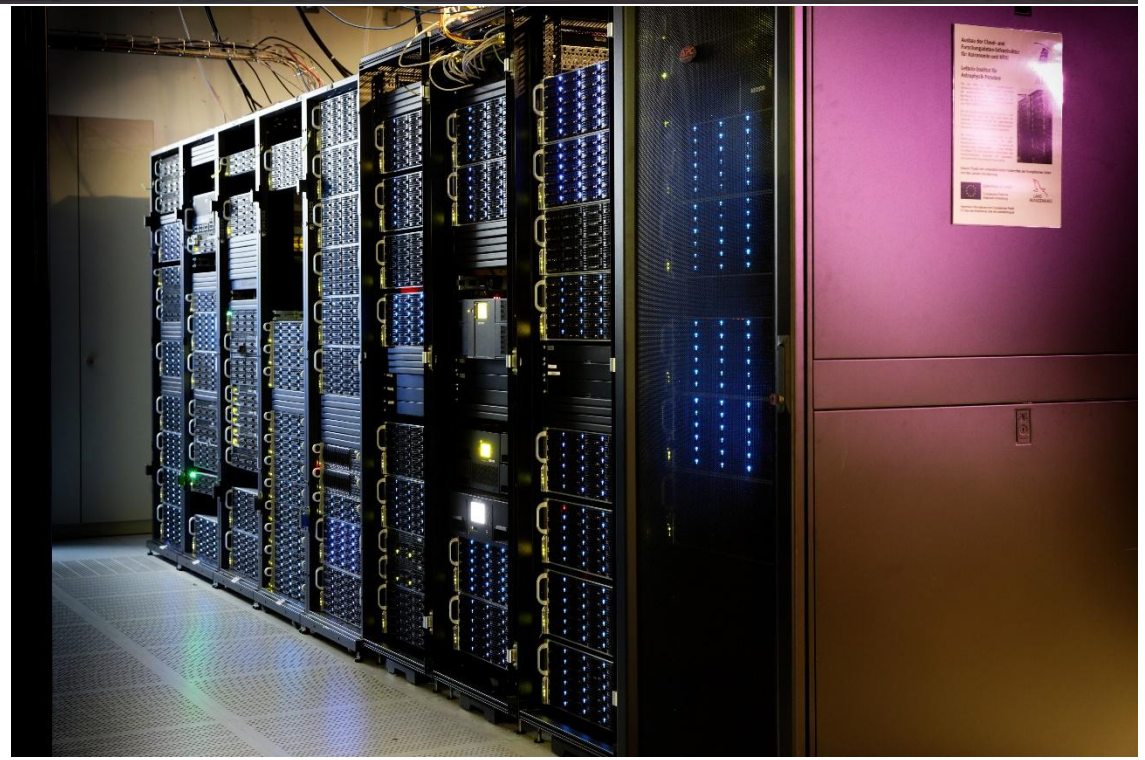
EUROPÄISCHE UNION
Europäischer Fonds für Regionale Entwicklung

LAND BRANDENBURG

Allgemeine Informationen zum Europäischen Fonds für regionale Entwicklung unter efre.brandenburg.de

Newton 21

Infiniband: 120GBit/s
Ethernet: 10GBit



2020-2021: EFRE funding ~**1.2M** Euro:
"Ausbau der Cloud- und Forschungsdaten-Infrastruktur für Astronomie und NFDI"

Dieses Projekt wird unterstützt durch Fördermittel der Europäischen Union und des Landes Brandenburg.

EUROPÄISCHE UNION
Europäischer Fonds für Regionale Entwicklung

LAND BRANDENBURG

Allgemeine Informationen zum Europäischen Fonds für regionale Entwicklung unter efre.brandenburg.de

Newton 21: Hardware

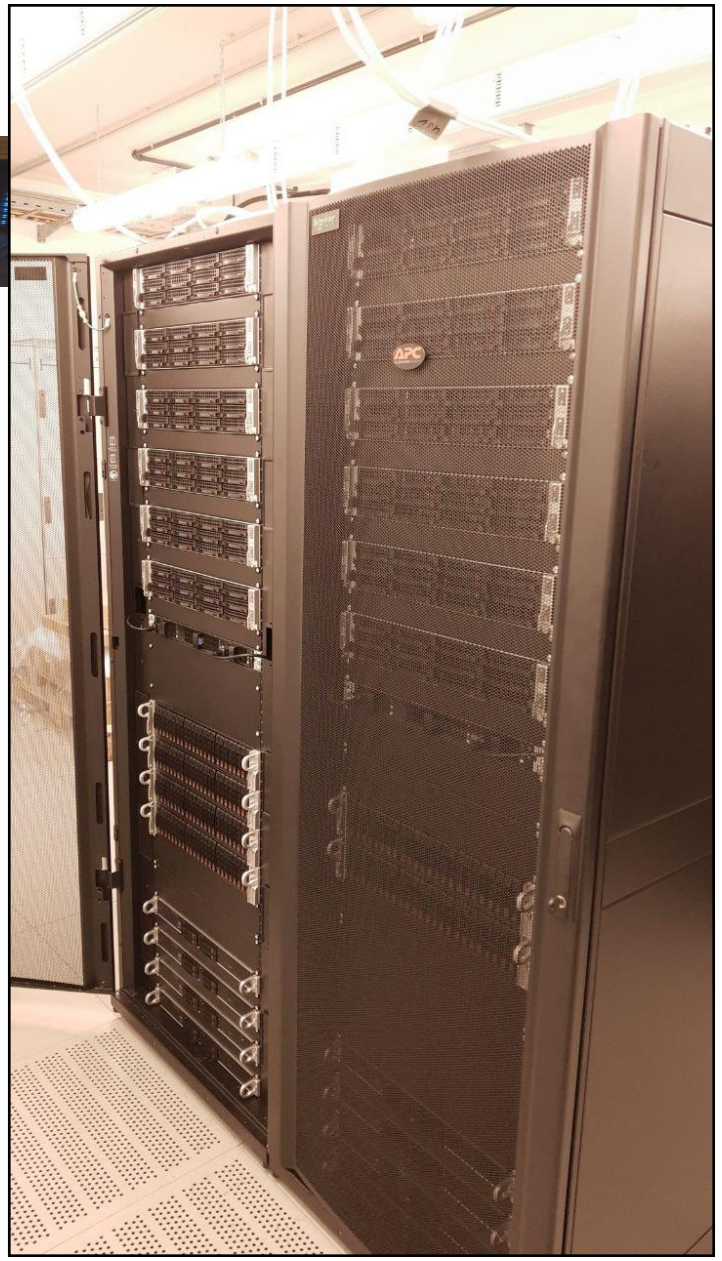
- Login nodes:
  - 2 each 2x Intel Xeon Gold **6226**, 12cores **8GB**/Core,196GB RAM
- Regular nodes:
  - 48 each 2x Intel Xeon Gold **6252**, 24cores **8GB**/Core,384GB RAM
- Himem nodes:
  - 8 each 2x Intel Xeon Gold **6252**, 24cores **16GB**/Core,768GB RAM
- GPU nodes(regular node+):
  - 4x Nvidia GPU RTX8000 **48GB** vRAM
  - 4x Nvidia GPU A100 **40GB** vRAM
- Storage:
  - 2x**1.5PT** Lustrefs

Summary:

Cores: 3120

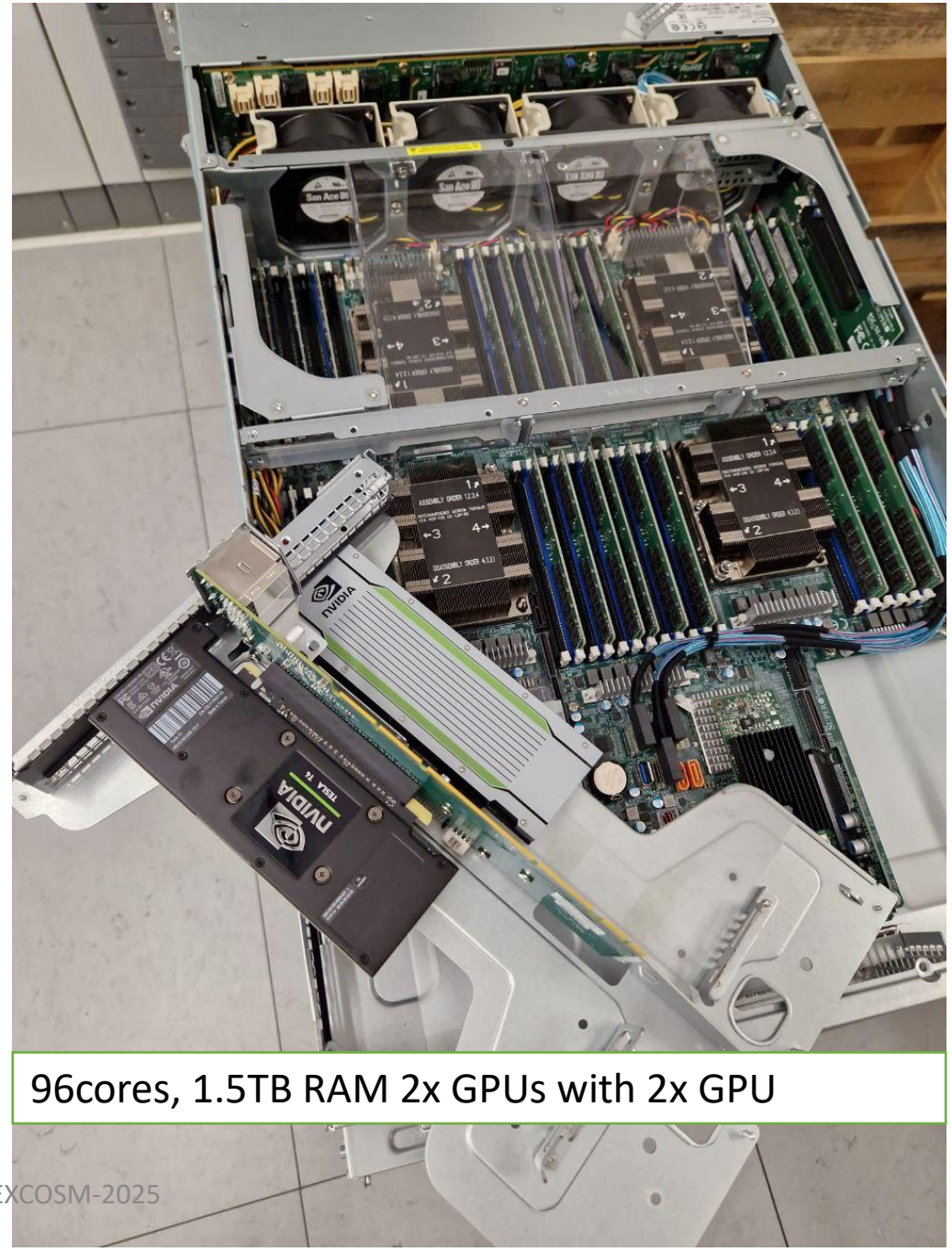
RAM: 28TB

Disk: 2.5PT



HW upgrade

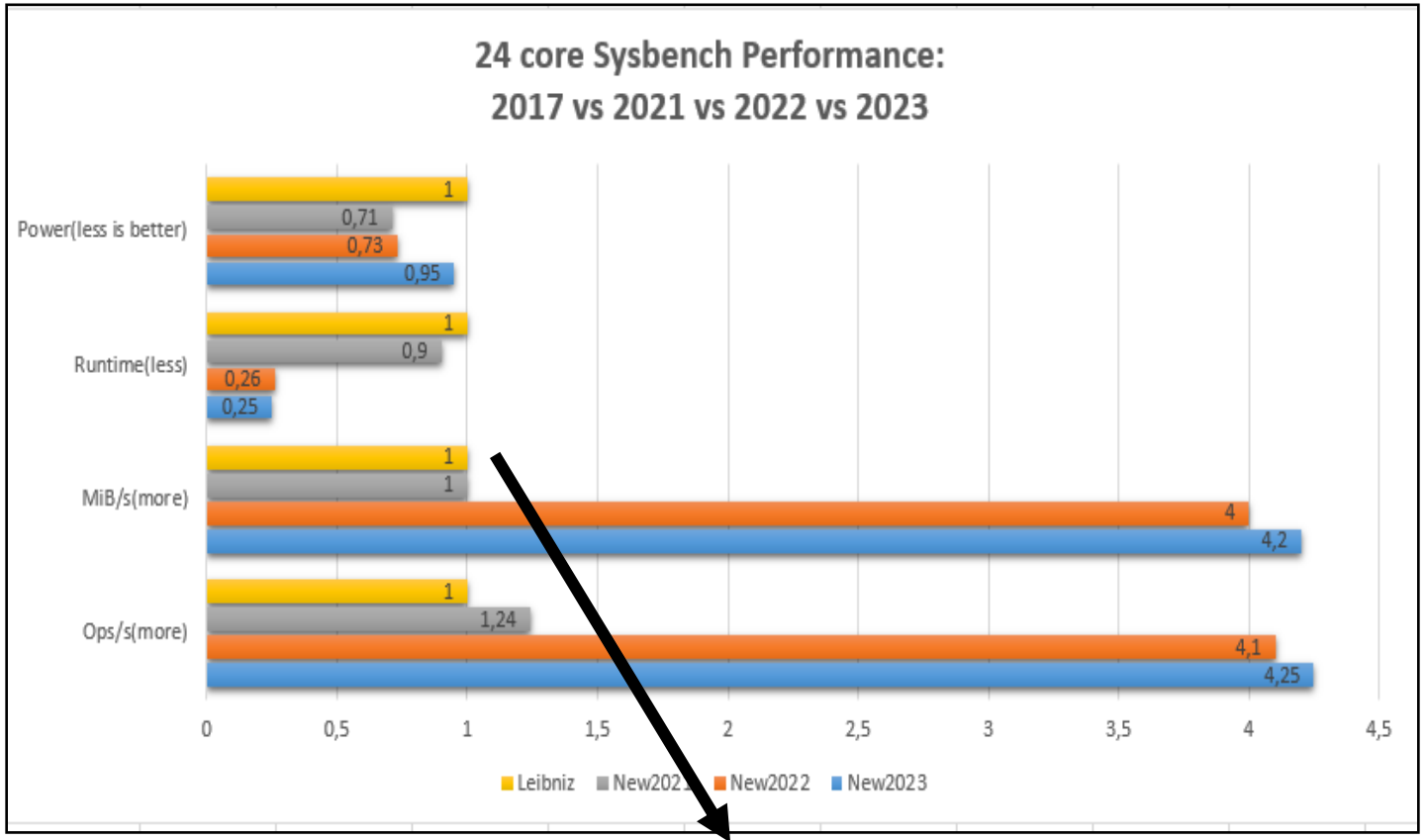
The surgery went ok ☺



96cores, 1.5TB RAM 2x GPUs with 2x GPU

HPC at AIP:
What we've achieved and what's on the horizon…

- Old Leibniz is turned off
- We have installed the New **20** nodes:
  - thanks to the Oliver's and Rainer's successful projects
- The benchmarks are started to test the stability of the cluster
- HW:
  - CPU: Intel Xeon Platinum **8452Y**
  - Cores/node: **72**
  - RAM/node: **512**GB, DDR5
- Release dates:
  - **2024 Jan-early Feb**
- Job queue availability:
  - to ALL AIP co-workers
- The optimal number of queues:
  - A subject of ongoing discussion and exploration.

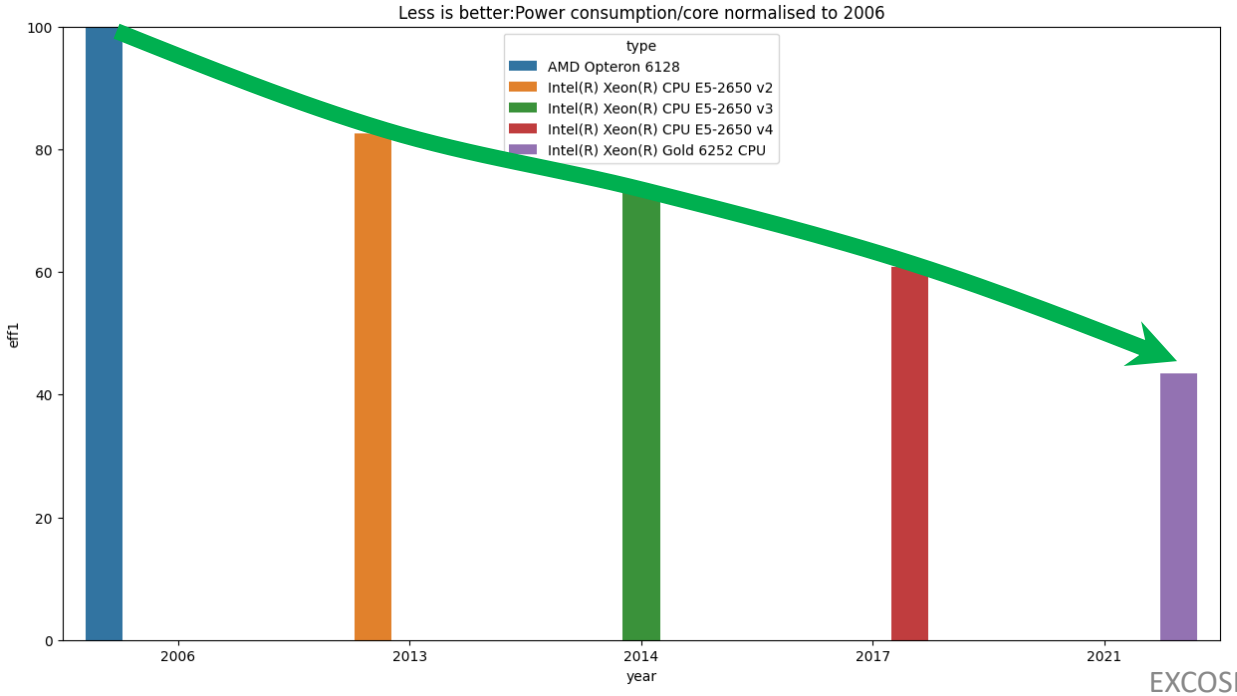


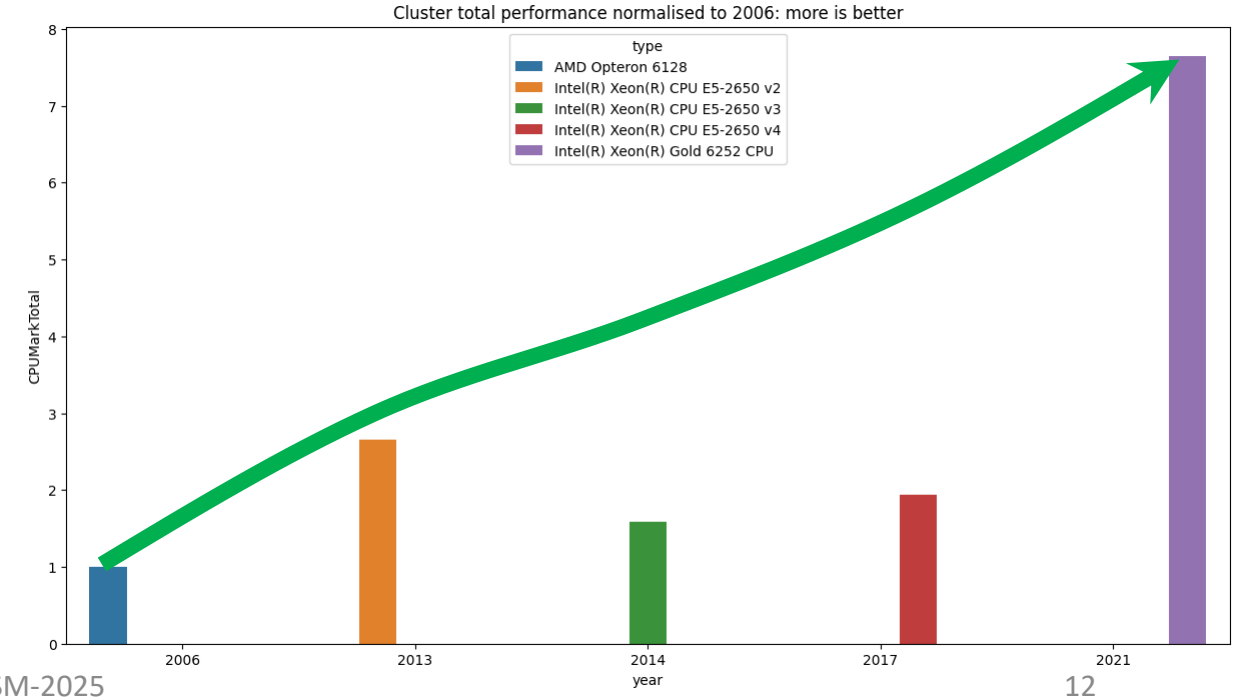
A huge Jump due to the Fabrication Process: 14nm vs 10nm:

Green computing strategy

We try to keeping the power consumption same as possible over the years, but increasing the clusters performance.
We should renew clusters/hardware <5-7y to be “green”

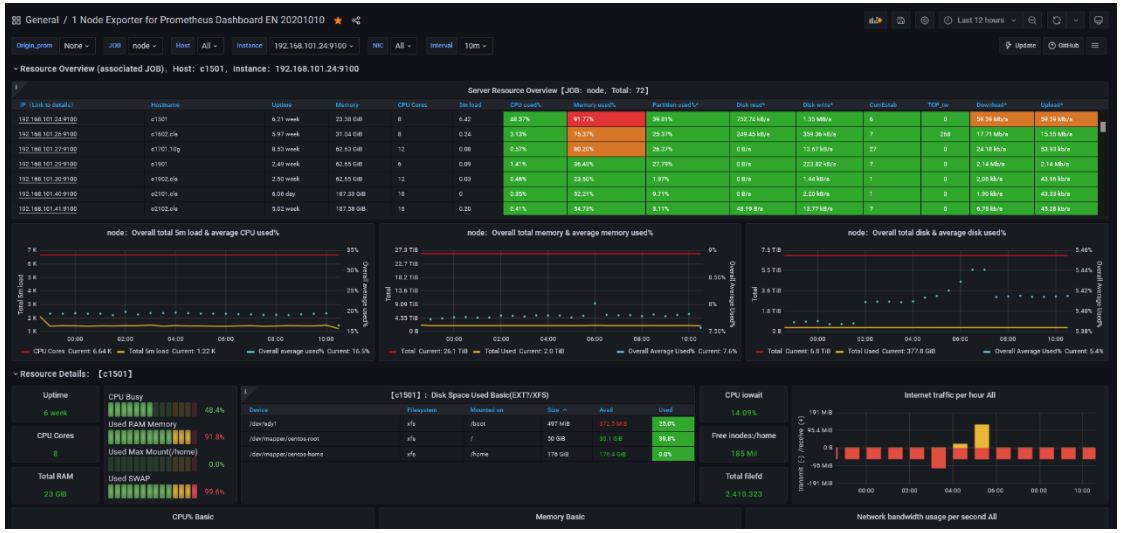
	year	type	socket	cores	threads	TDPW	CPUMark
0	2006	AMD Opteron 6128	2	8	8	115	2826
1	2013	Intel(R) Xeon(R) CPU E5-2650 v2	2	8	16	95	9989
2	2014	Intel(R) Xeon(R) CPU E5-2650 v3	2	10	20	105	11931
3	2017	Intel(R) Xeon(R) CPU E5-2650 v4	2	12	24	105	13489
4	2021	Intel(R) Xeon(R) Gold 6252 CPU	2	24	48	150	32410

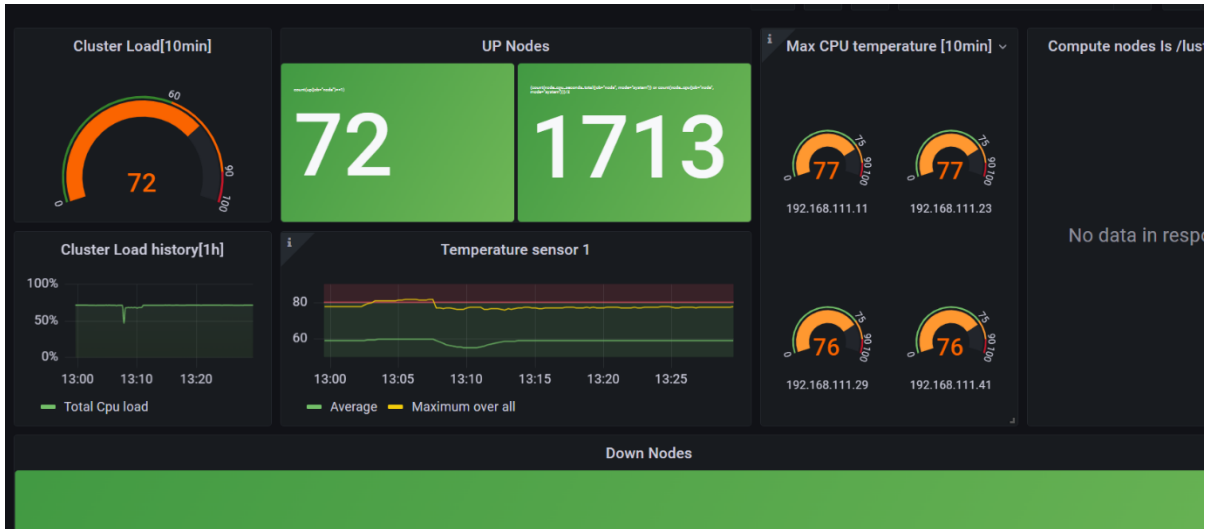




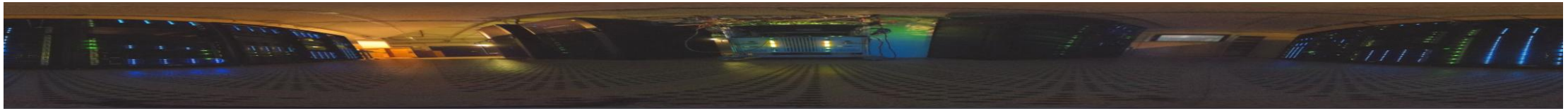
Monitoring is important!

Cluster Monitoring (Grafana)





Newton 21: Now is stopped(04.2022)



<70

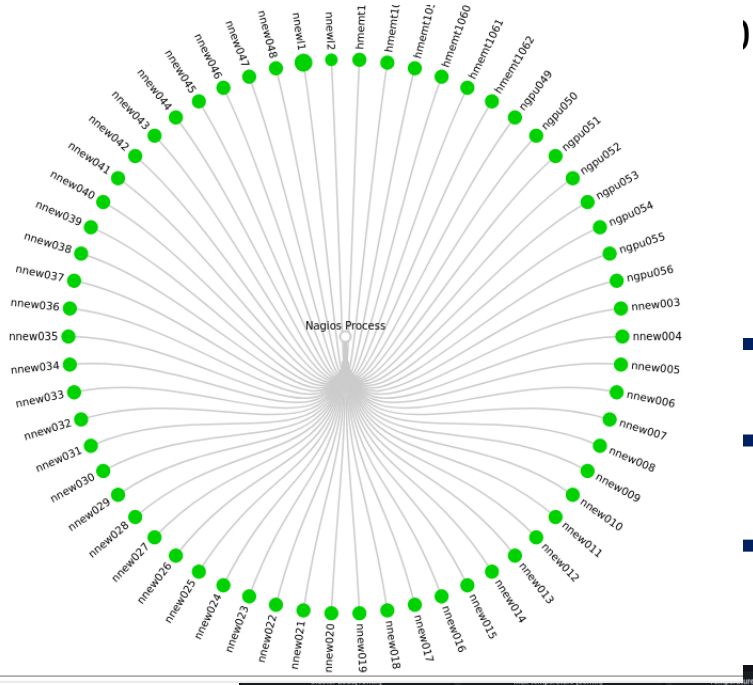
Dieses Projekt wird unterstützt durch Fördermittel der Europäischen Union und des Landes Brandenburg.

EUROPÄISCHE UNION
Europäischer Fonds für Regionale Entwicklung

LAND BRANDENBURG

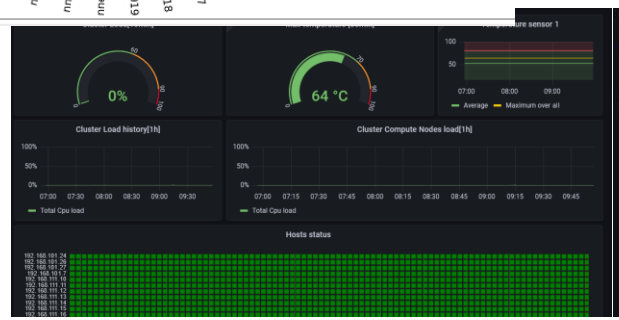
Allgemeine Informationen zum Europäischen Fonds für regionale Entwicklung unter efre.brandenburg.de

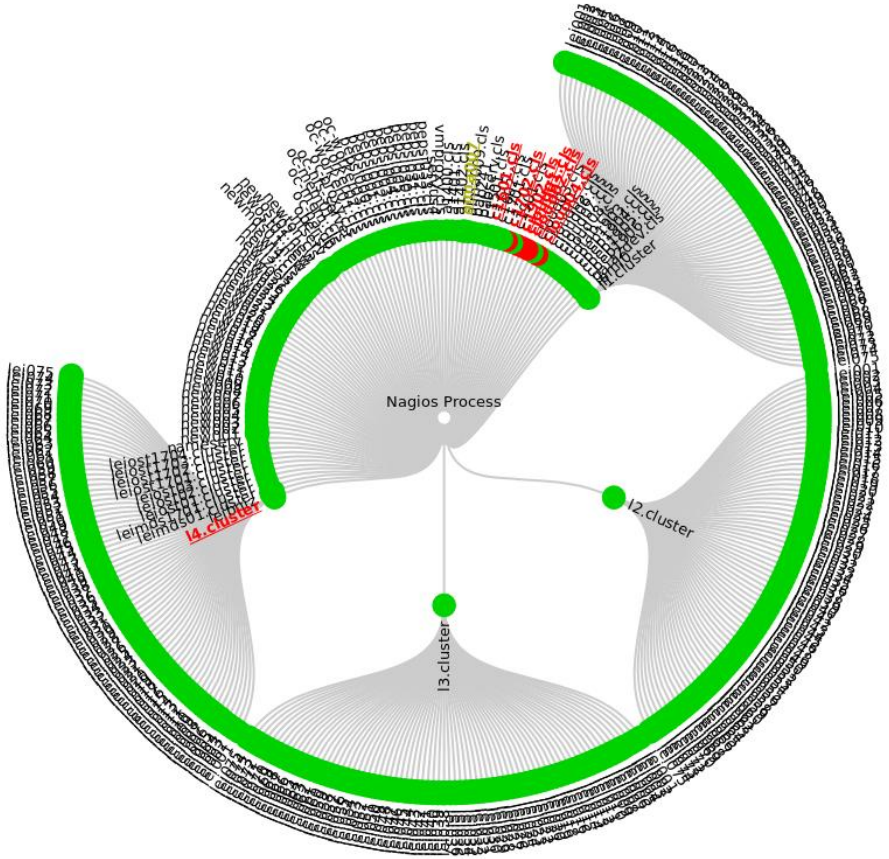
Monitoring stack



http://1

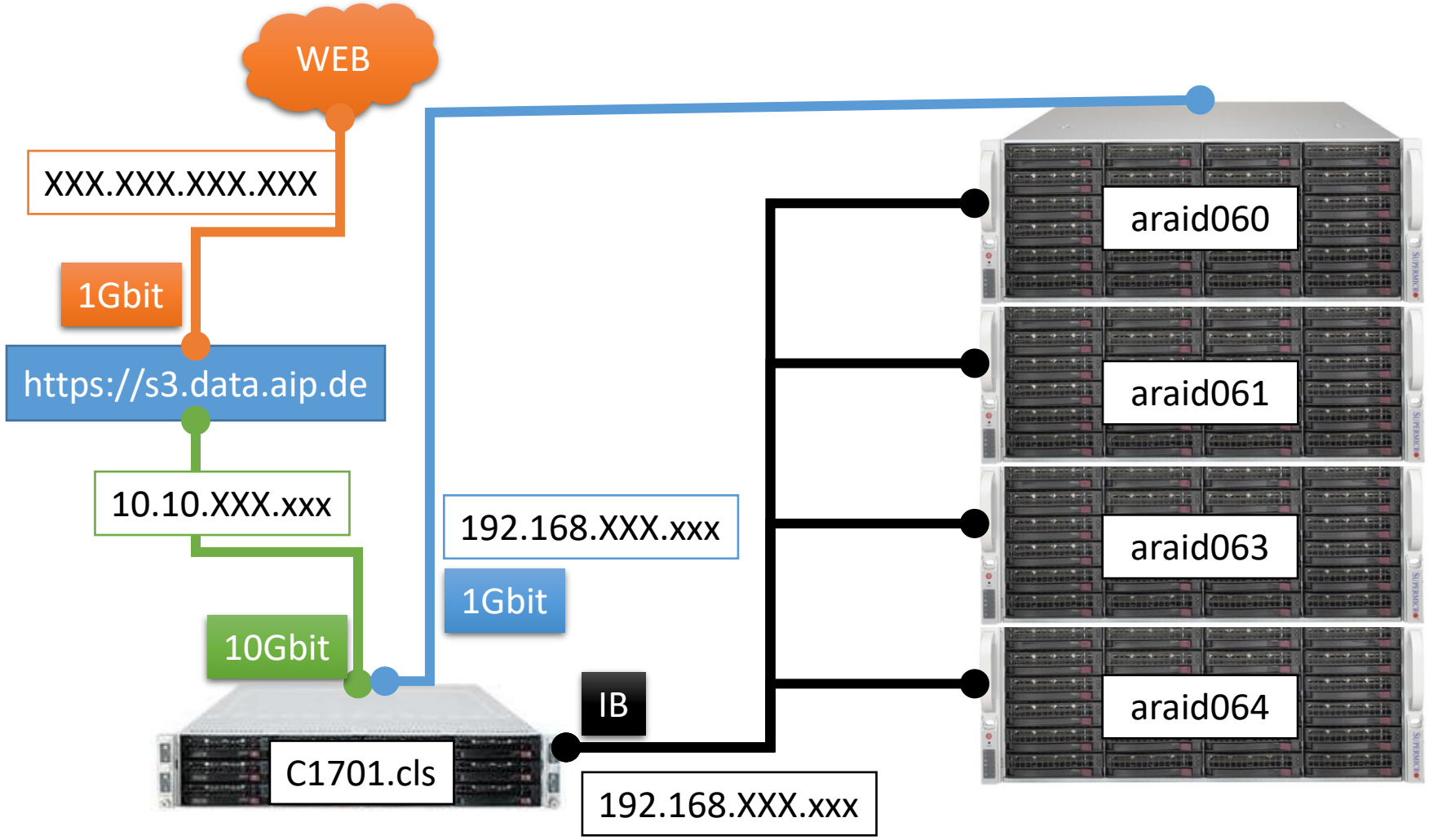
http://1





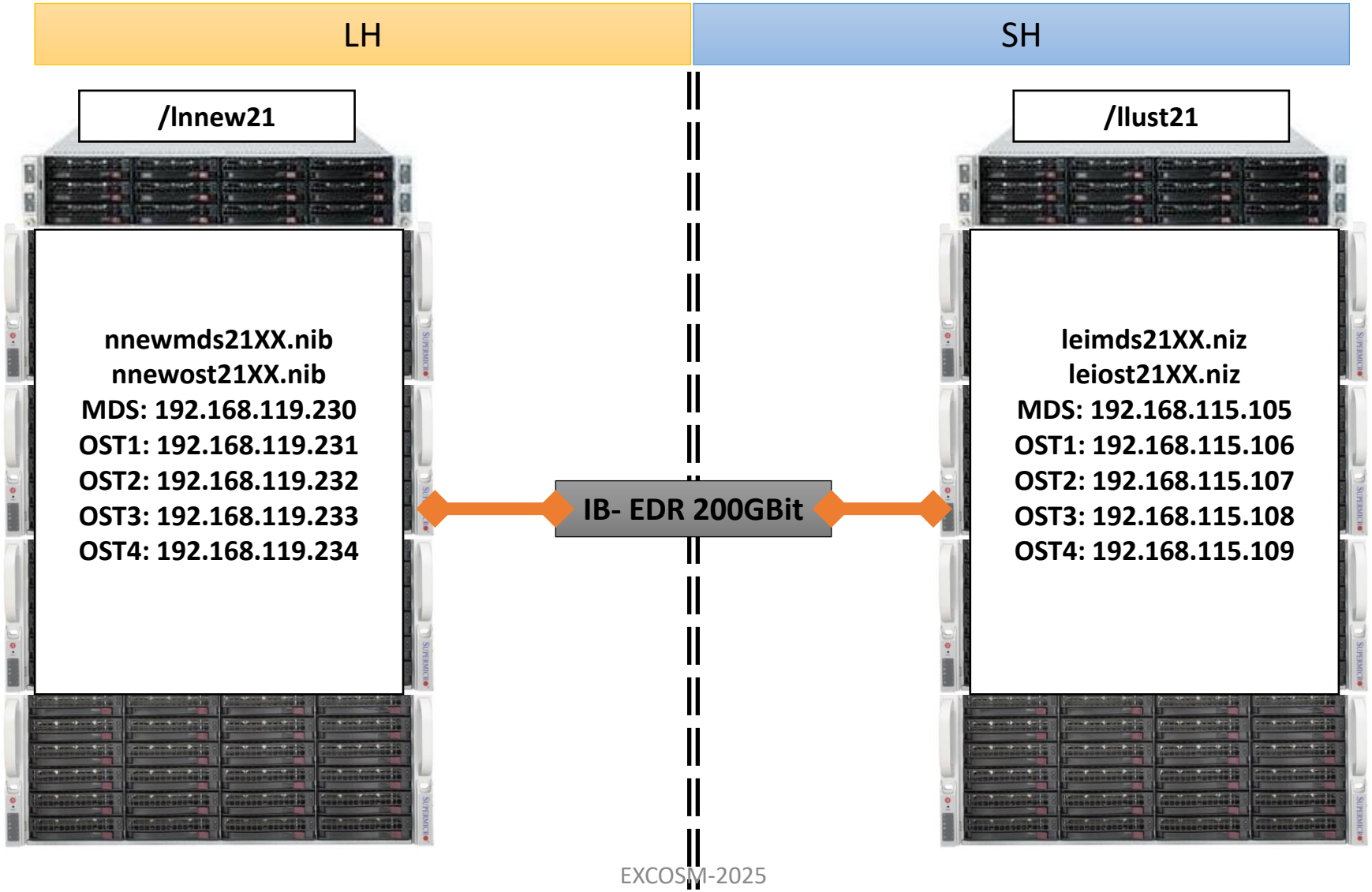
TODO

S3 storage at AIP: MinIO network

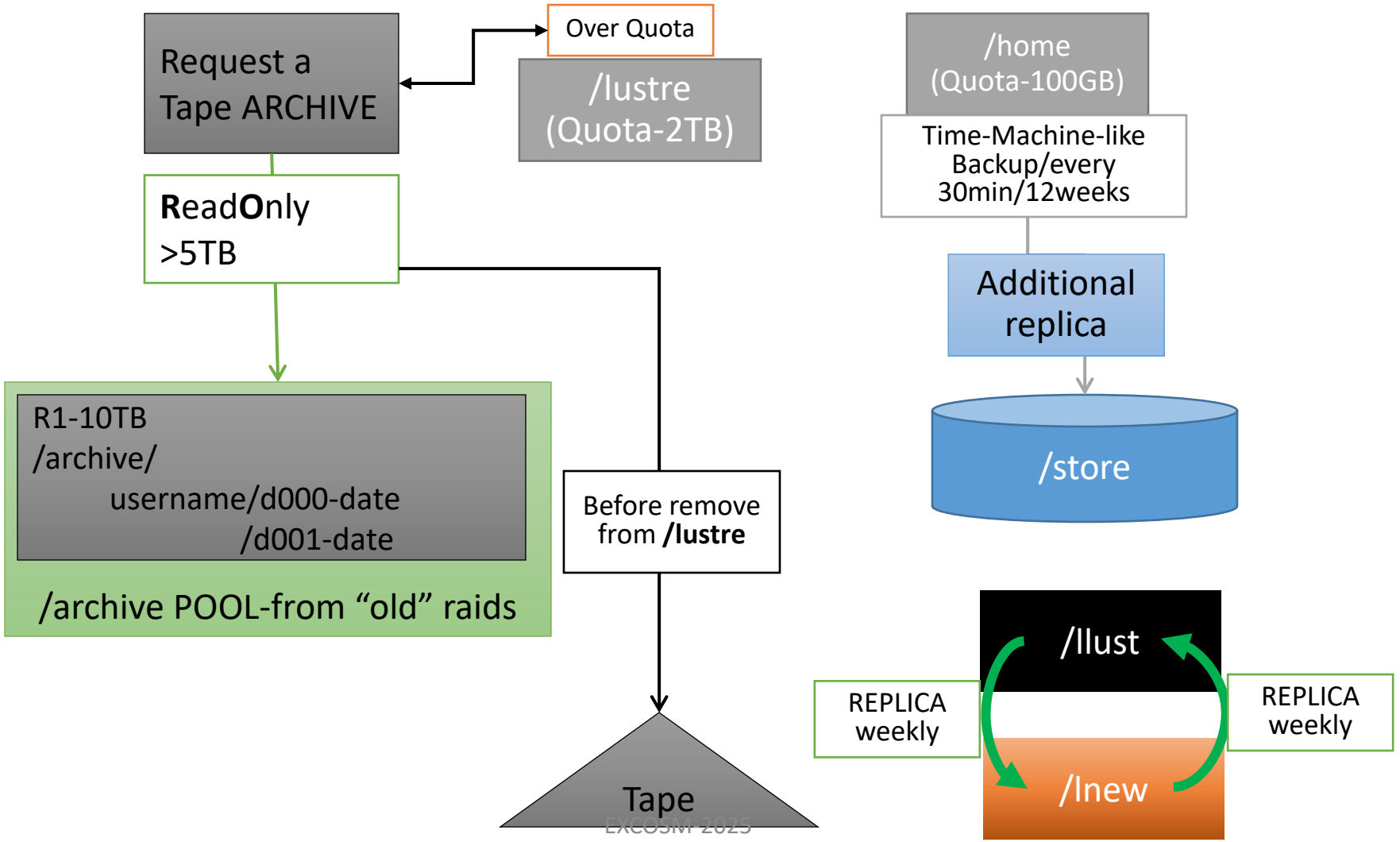


Data management

LustreFS


LH
SH
/lnnew21
/llust21
nnewmds21XX.nib
nnewost21XX.nib
MDS: 192.168.119.230
OST1: 192.168.119.231
OST2: 192.168.119.232
OST3: 192.168.119.233
OST4: 192.168.119.234
IB- EDR 200GBit
leimds21XX.niz
leiost21XX.niz
MDS: 192.168.115.105
OST1: 192.168.115.106
OST2: 192.168.115.107
OST3: 192.168.115.108
OST4: 192.168.115.109

Data management on HPC Clusters



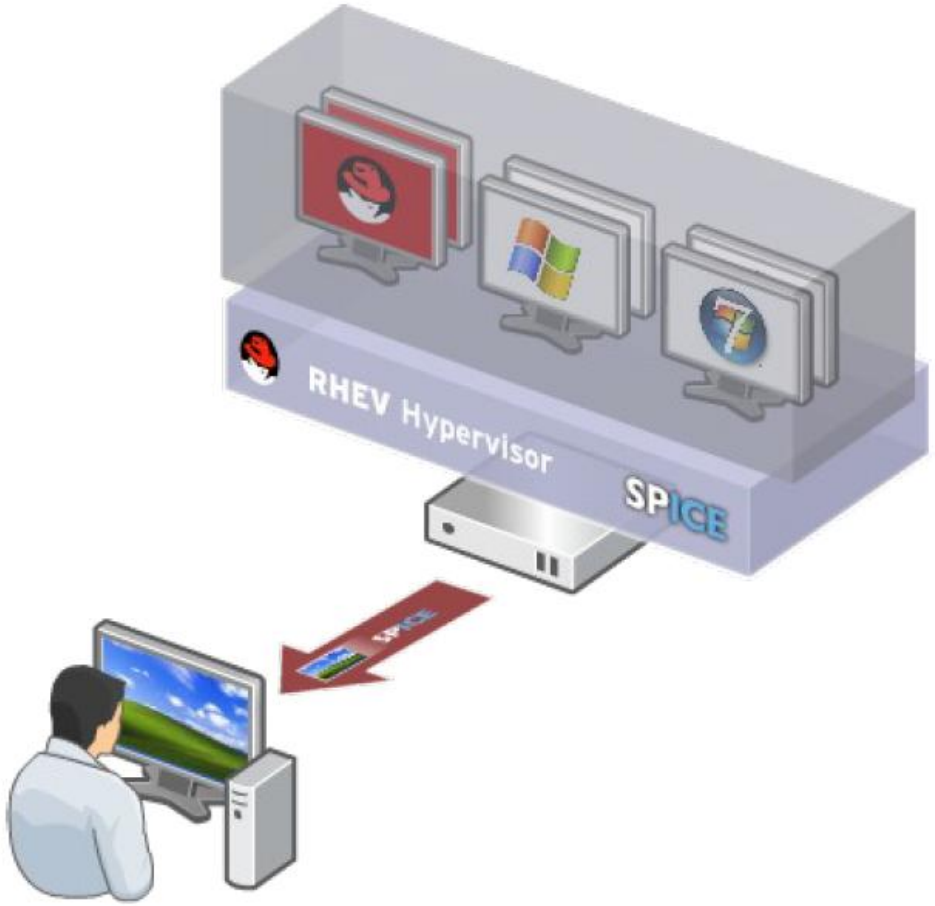
Users in HPC

Parallel programming models

- single-instruction multiple-data (SIMD)
  - OpenMP
  - Pthreads
  - auto-parallelization
- multiple-instruction single-data (MISD)
  - MPI
- Serial job sharding
  - **xargs–PN** aka shared
  - GNU **parallel –PN** aka OpenMP+MPI
  - slurm job arrays

HPC admins: Users software

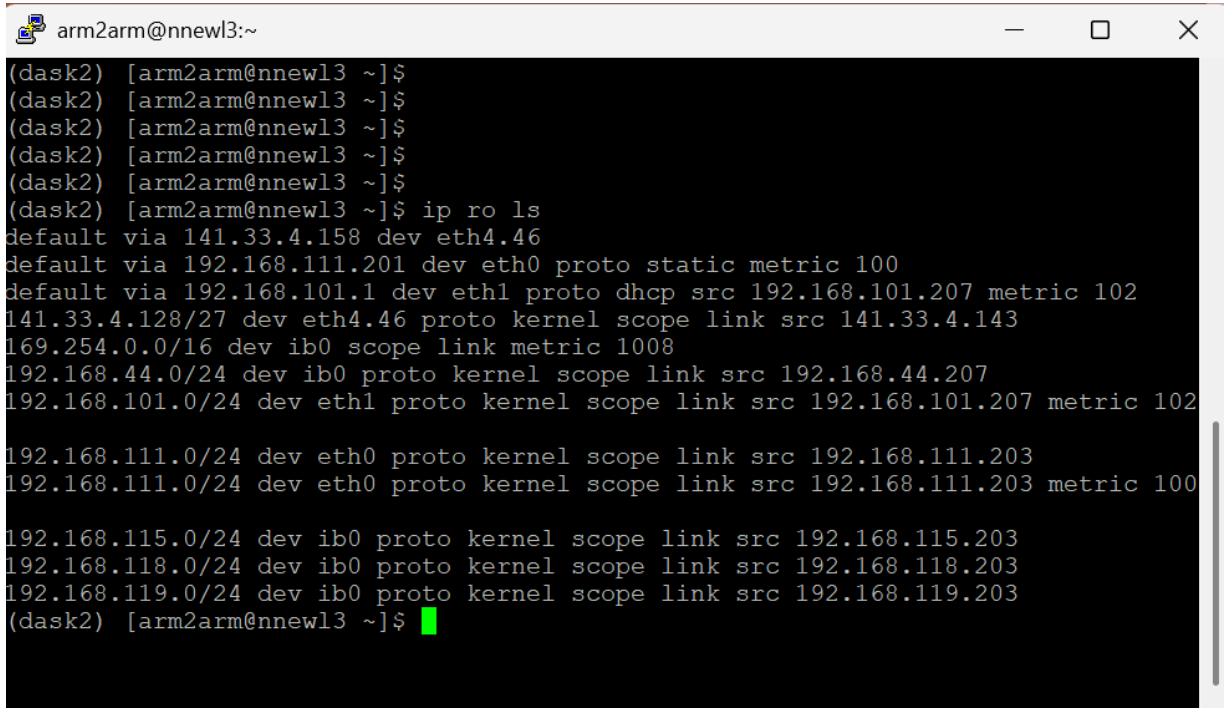
Users: What do they see?



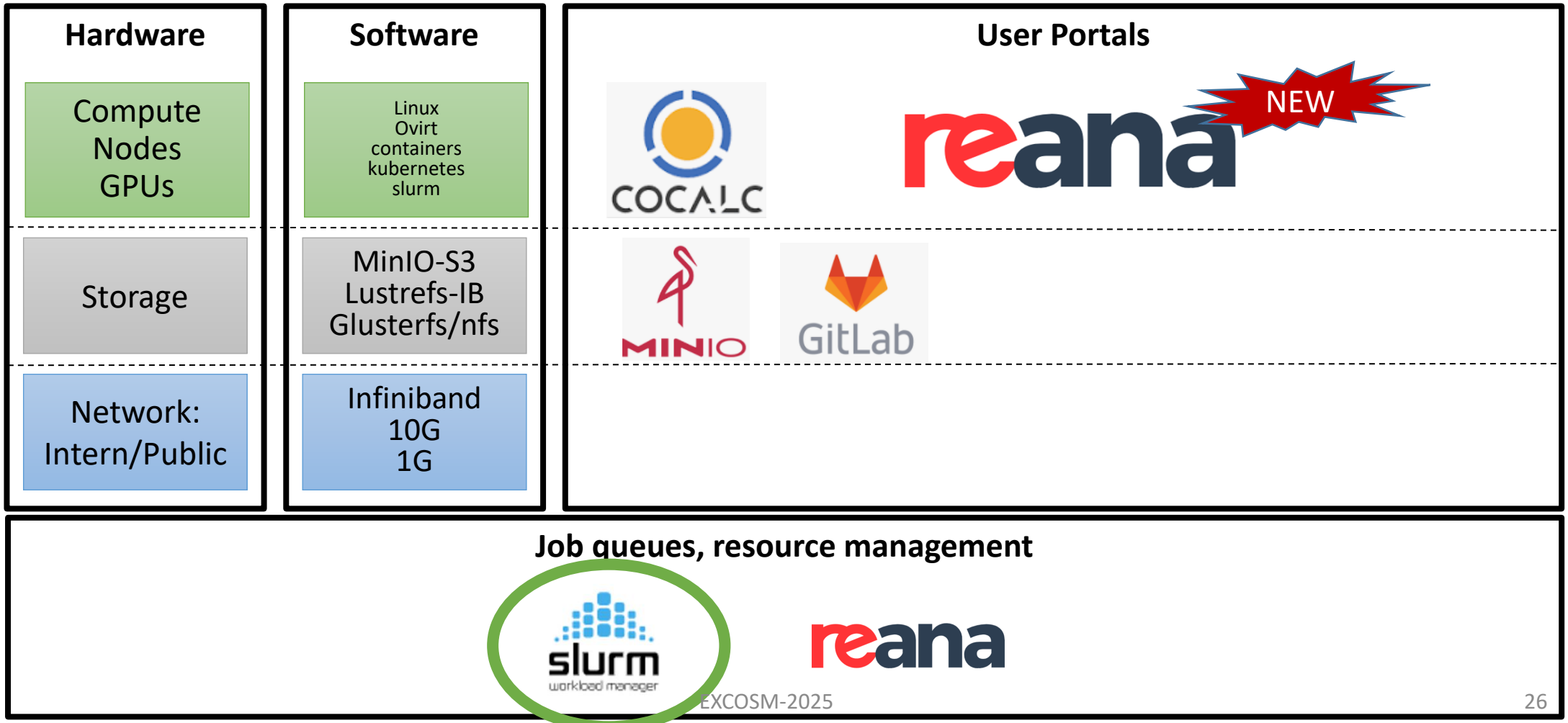
Getting Started

If you have an AIP-account, but no **cluster** account
mailto: **cluster-adm** @ **aip.de** with your name and username

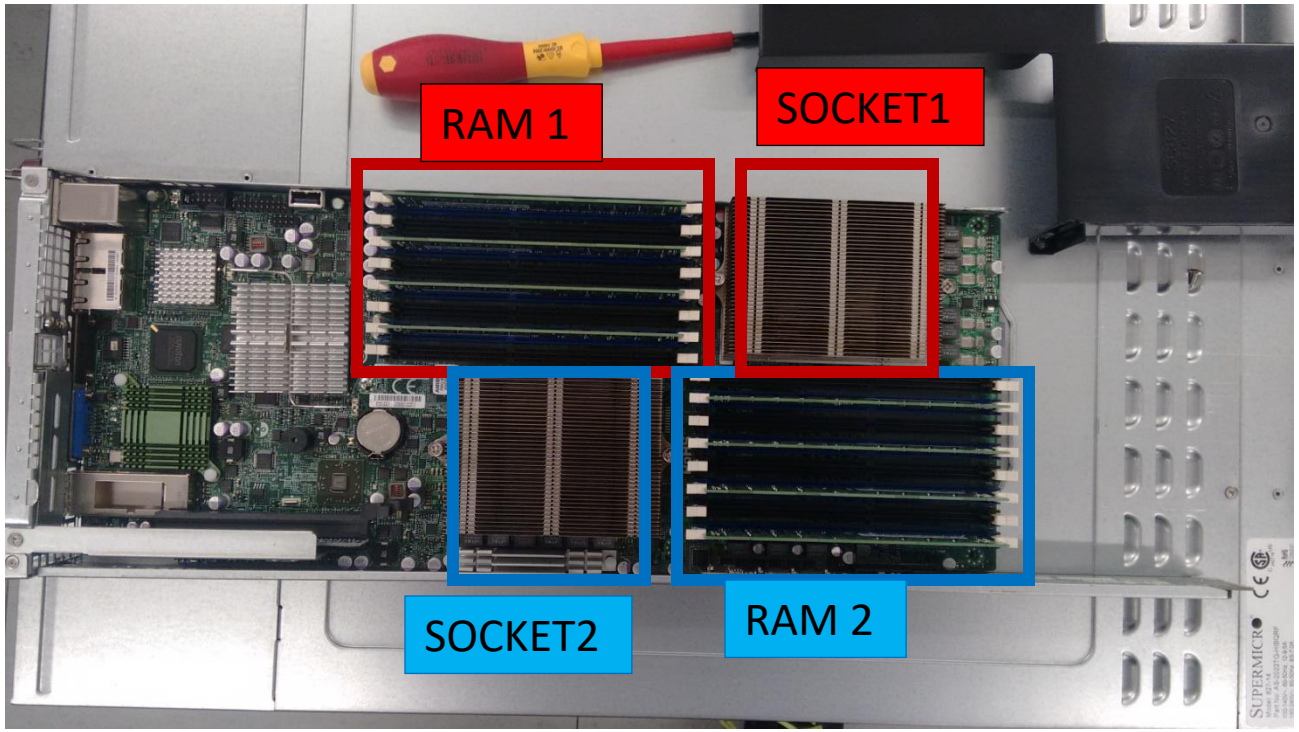
Access: **ssh XXX.XXX.XXX.XXX**



SAAS,IAAS and PAAS

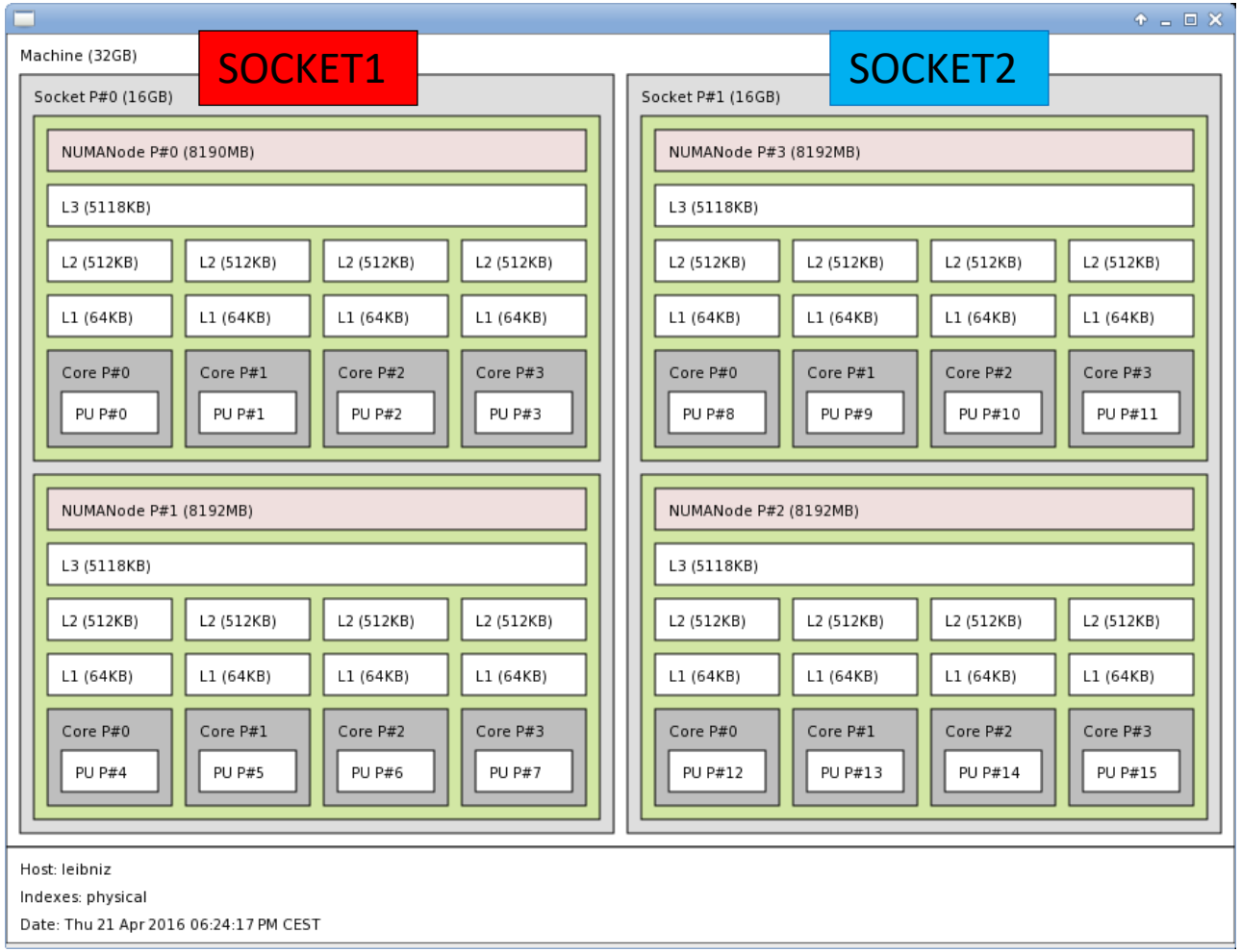


what do you get as a **node**?

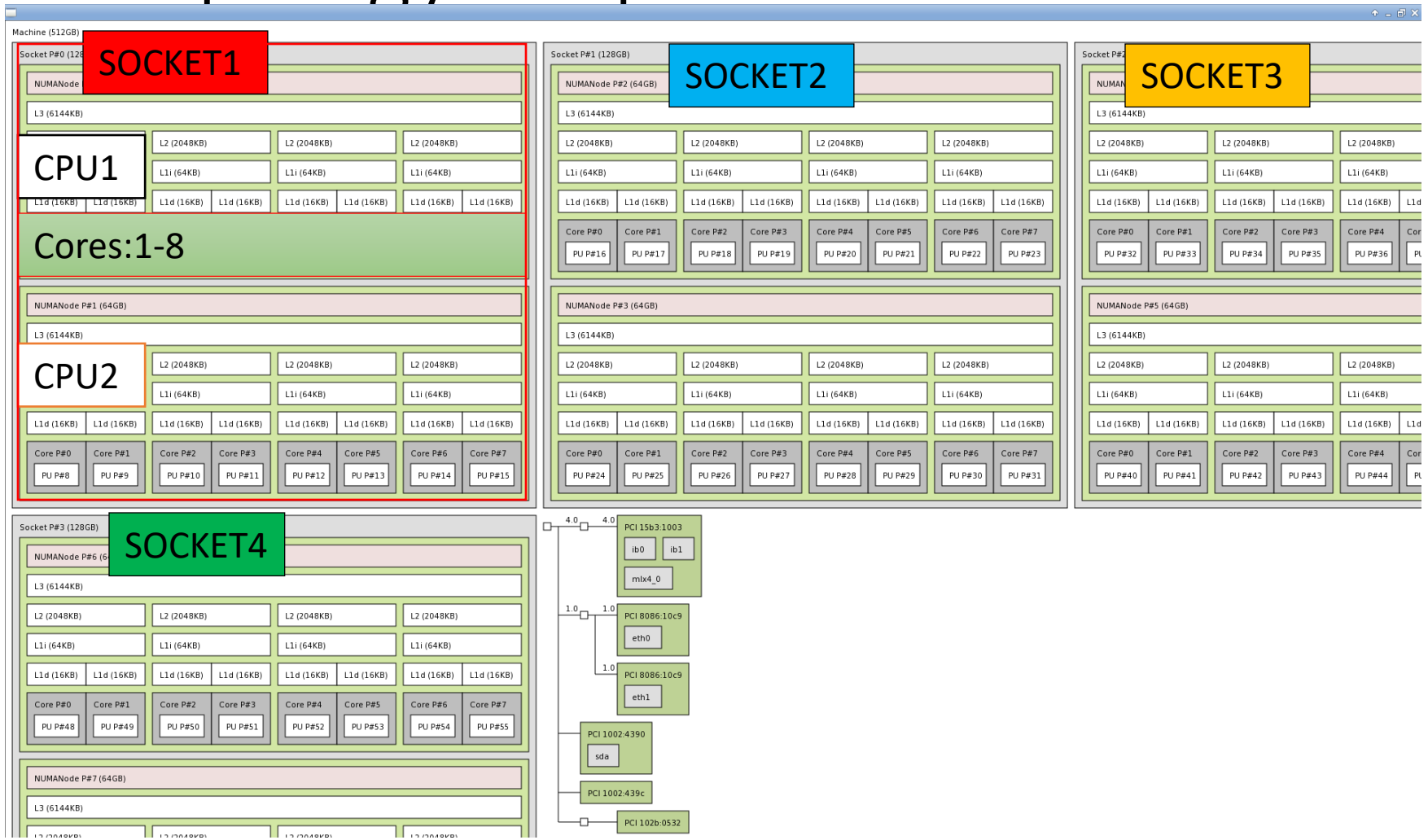


2x Intel Xeon Gold 6252, 24 cores – 2 numa nodes

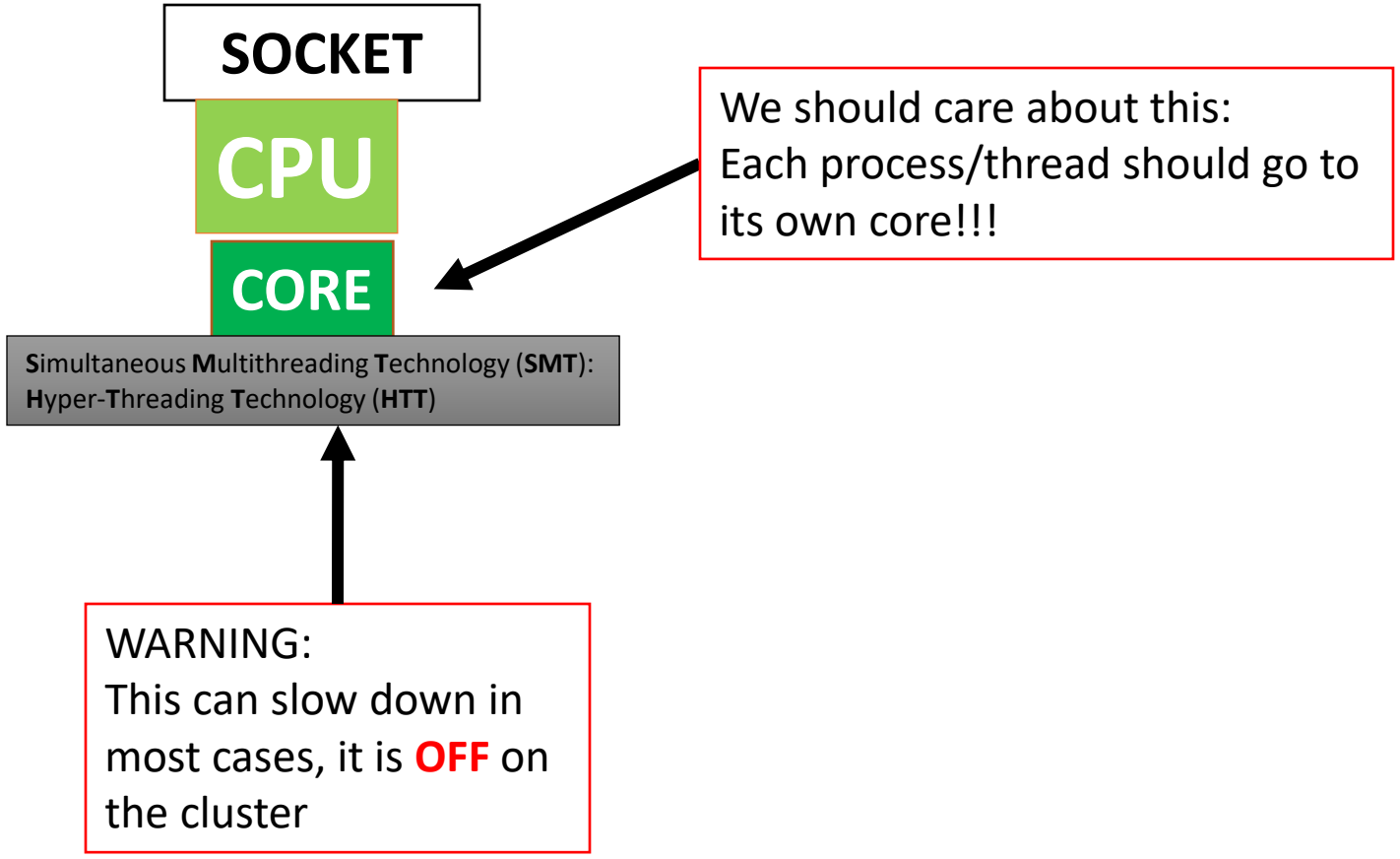
The HW topology lstopo: AMDLei



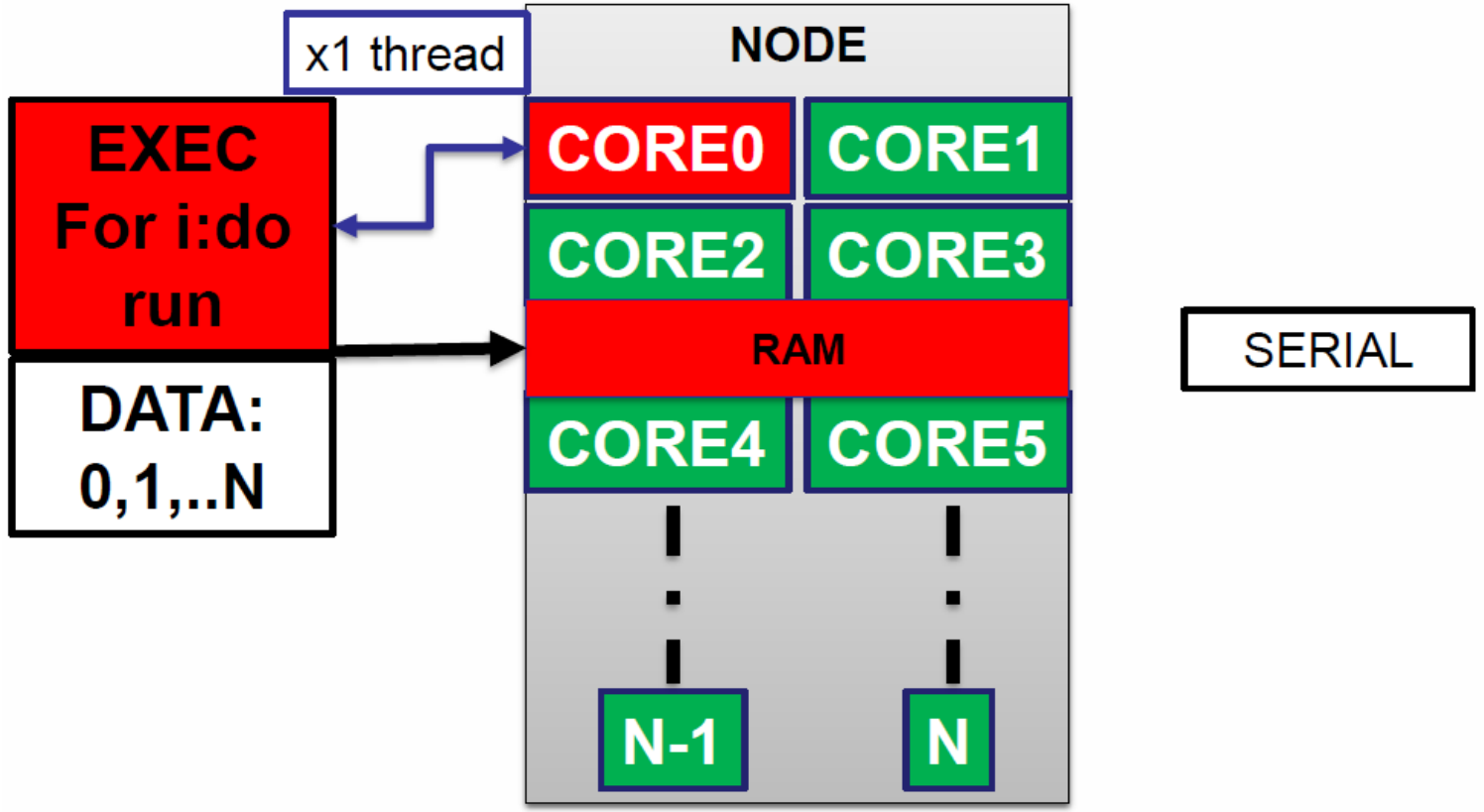
The HW topology lstopo:himem



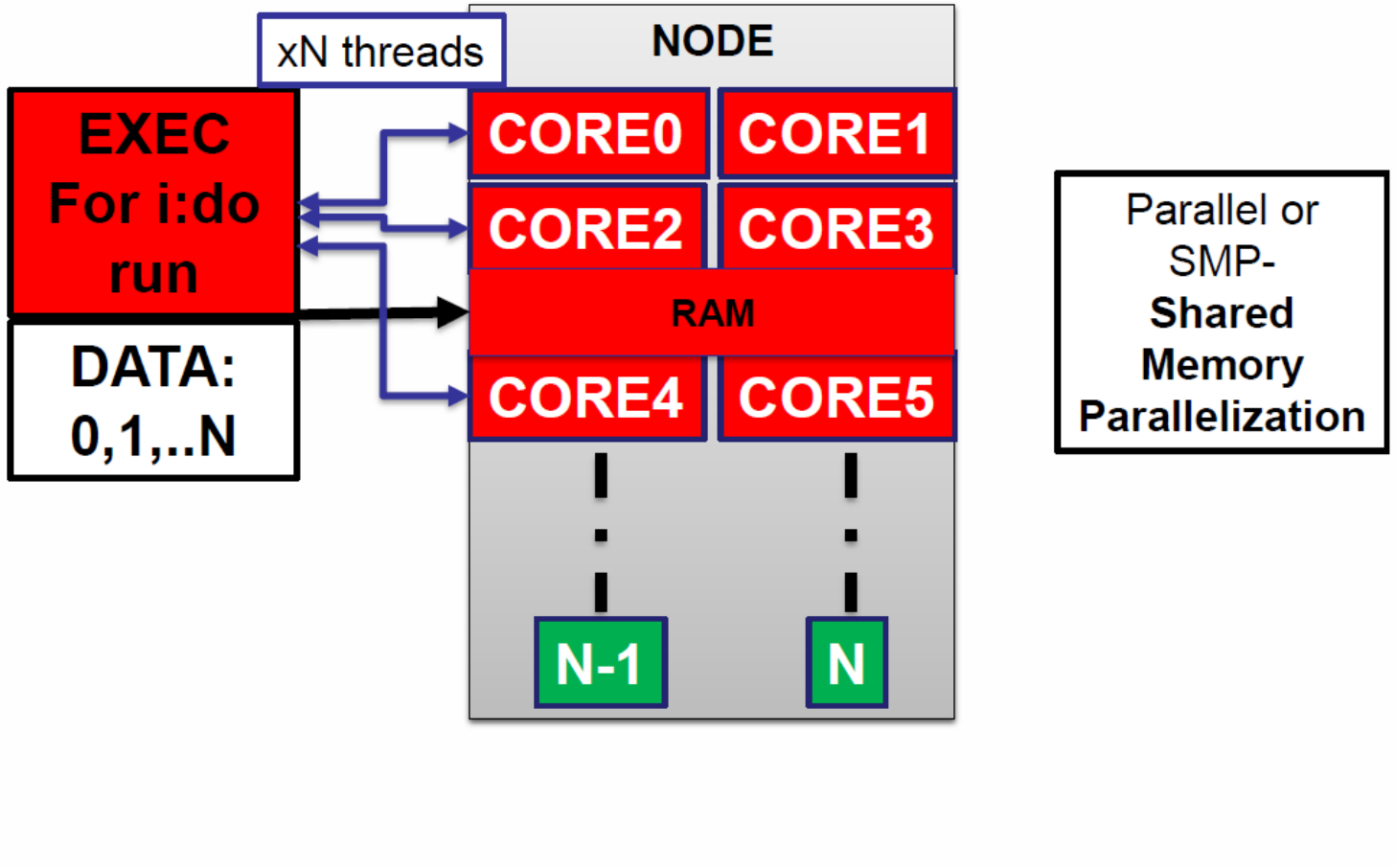
The HW topology lstopo



Shared Memory Devices



Shared Memory Devices: OpenMP/thr

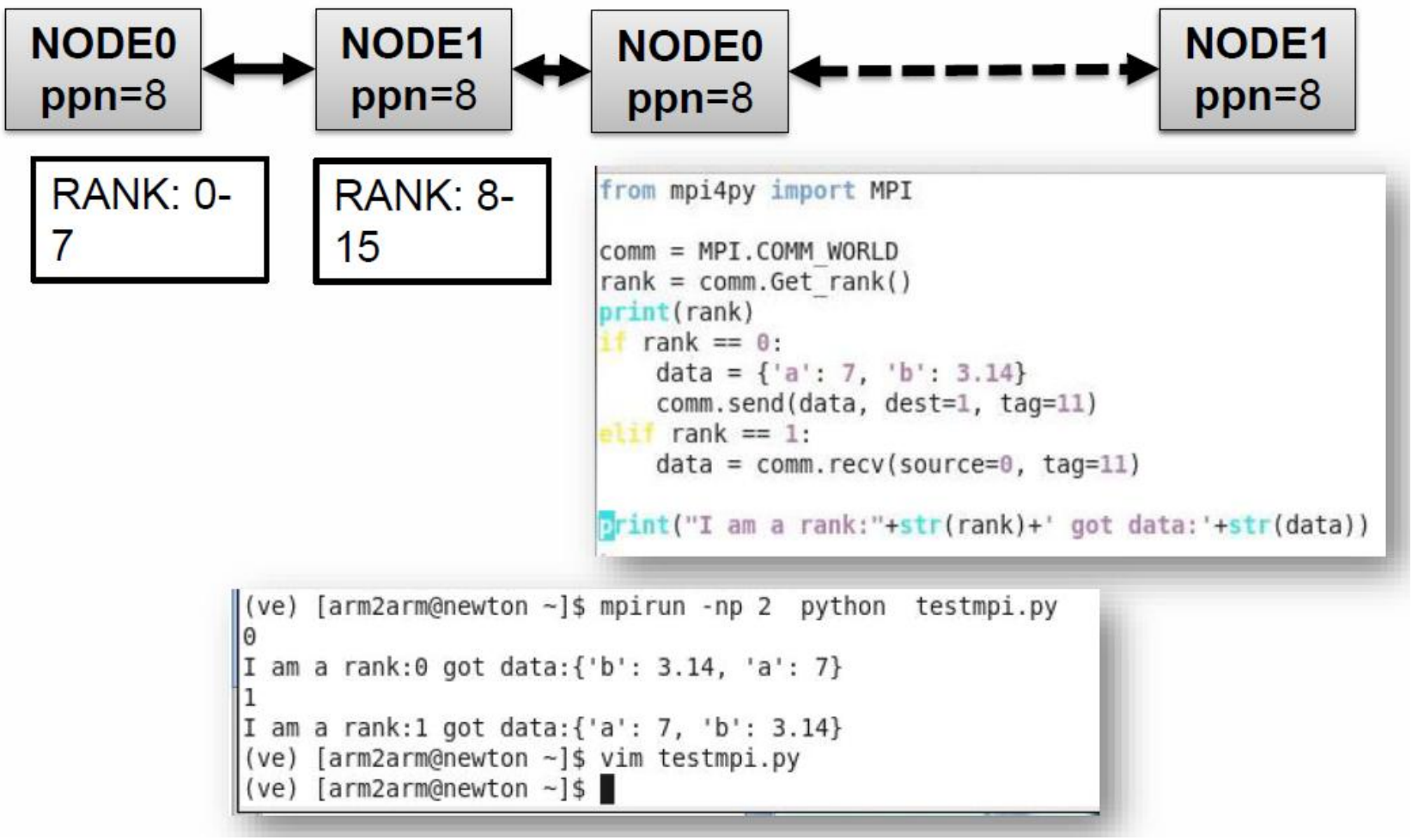


Shared Memory Devices: OpenMP/threads

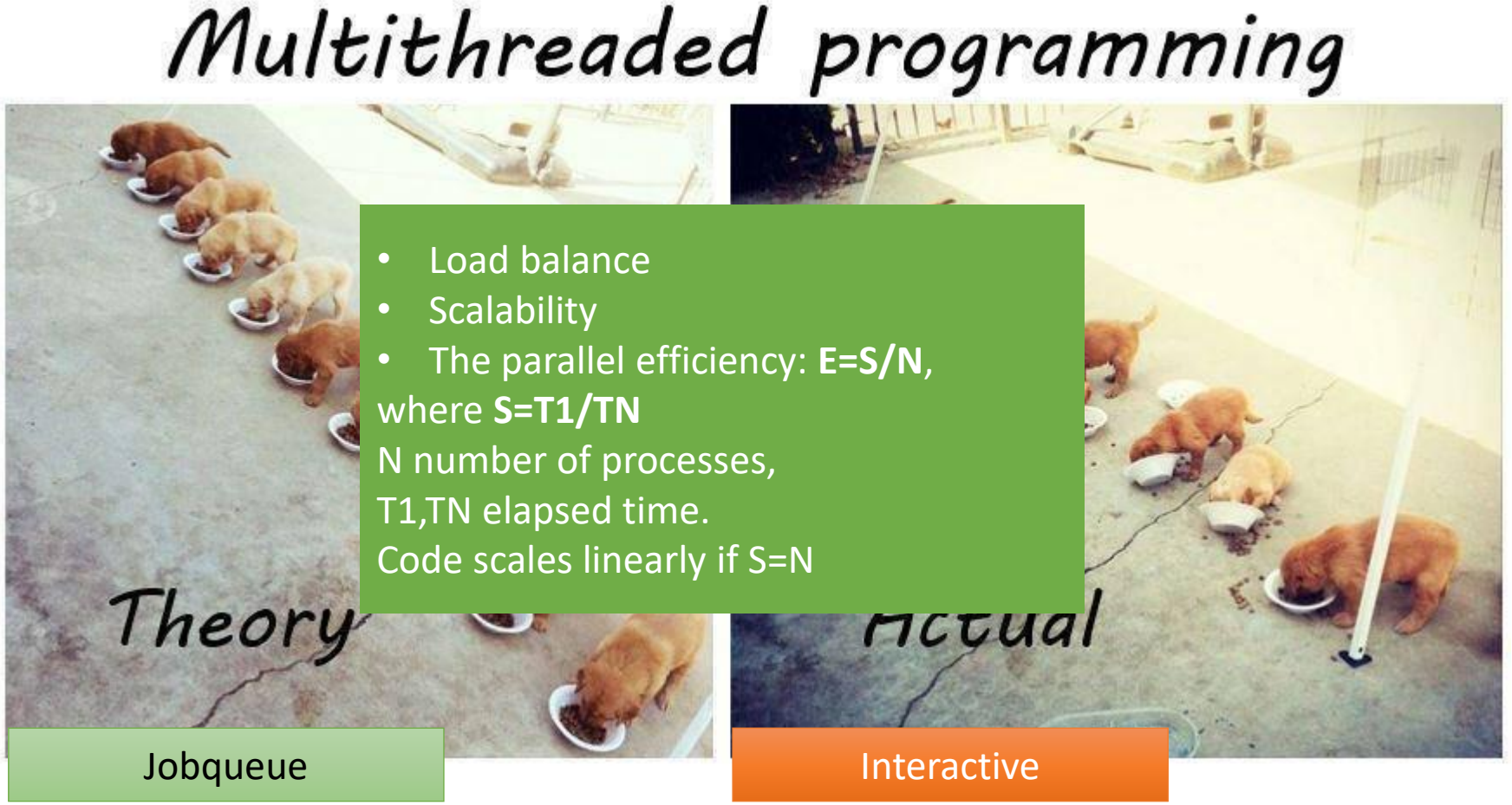
```
from joblib import Parallel, delayed
import multiprocessing

# Assuming N inputs with long calculations
N=10
inputs = range(N)
def VeryLongCalculation(i):
    return i * i

# Process each input in parallel
num_cores = multiprocessing.cpu_count()
results = Parallel(n_jobs=num_cores)(delayed(VeryLongCalculation)(i)
for i in inputs)
results
```

```
[0, 1, 4, 9, 16, 25, 36, 49, 64, 81]
```

The distributed memory machines: MPI



Reminder…



- Load balance
- Scalability
- The parallel efficiency: **E=S/N**,

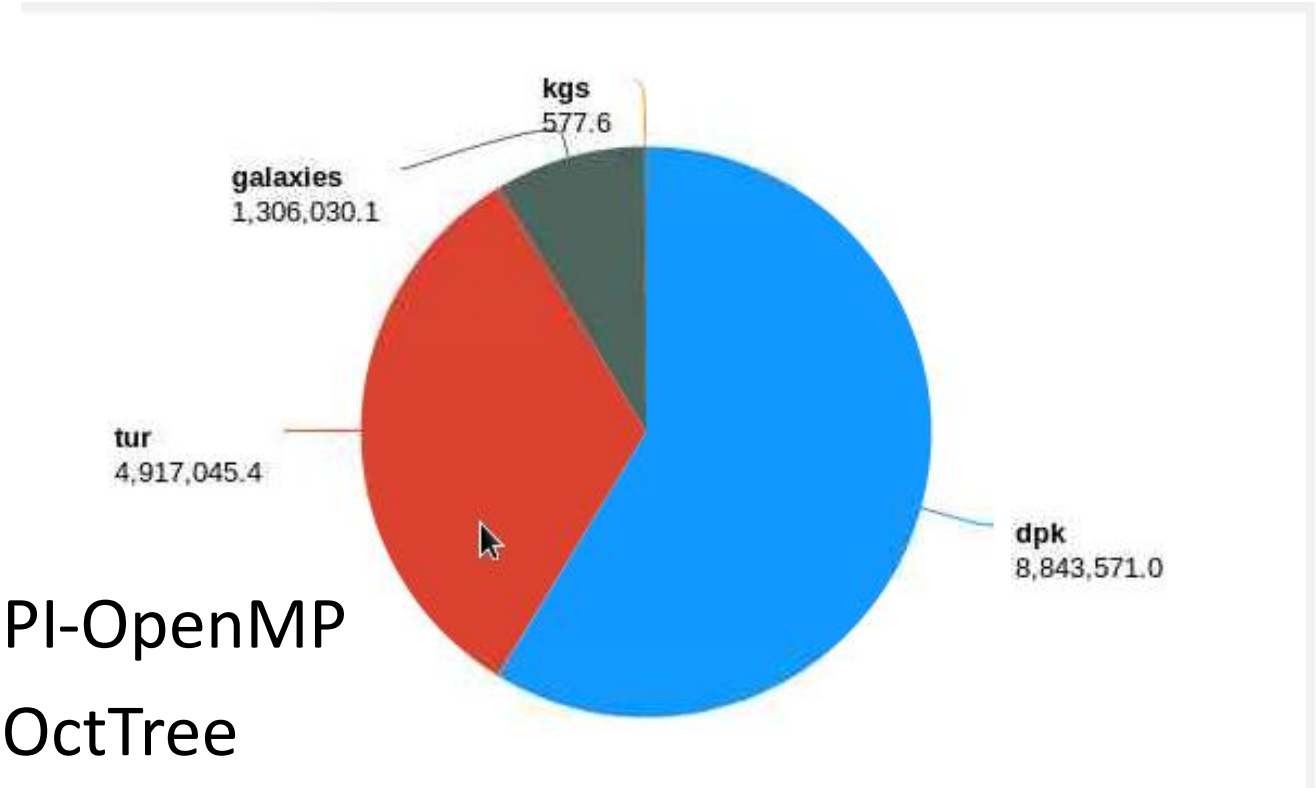
where **S=T1/TN**
N number of processes,
T1,TN elapsed time.
Code scales linearly if S=N

Jobqueue

Interactive

Who is using most of the CPU time?

- Cosmology:
  - MHD+Gravity+Gasdynamics
  - Starformation, Cosmic Rays,BH…
- Magneto-hydrodynamics: MHD
- Data processing from telescopes



- Adaptive unstructured mesh, MPI-OpenMP
- Magneto-hydrodynamics: AMR-OctTree
- Data processing from telescopes: python, c, java, other

Software stack

(dask2) [arm2arm@nnewl3 ~]$ **module avail**

```
---------------------------------------------------- /opt/ohpc/pub/moduledeps/gnu9-openmpi4 -----------------------------------------------------
  boost/1.76.0    hypre/2.18.1    mumps/5.2.1    opencoarrays/2.9.2    phdf5/1.10.8    py3-mpi4py/3.0.3    scalapack/2.1.0    superlu_dist/6.4.0
  fftw/3.3.8      mfem/4.3        netcdf/4.7.4   petsc/3.16.1          ptscotch/6.0.6  py3-scipy/1.5.1     slepc/3.16.0       trilinos/13.2.0

------------------------------------------------------------ /opt/ohpc/pub/moduledeps/gnu9 -------------------------------------------------------
  gsl/2.7     (L)    impi/2021.4.0    mpich/3.4.2-ofi    openblas/0.3.7       py3-numpy/1.19.5
  hdf5/1.10.8 (L)    metis/5.1.0      mvapich2/2.3.6     openmpi4/4.1.1 (L)   superlu/5.2.1

---------------------------------------------------------- /opt/ohpc/pub/modulefiles ----------------------------------------------------------
  EasyBuild/4.8.2      gnu9/9.4.0     (L)    intel/2022.0.1    intel/2023.2.1    (D)    ohpc          (L)    prun/2.2          (L)
  autotools     (L)    hwloc/2.7.0           intel/2022.0.2    intel/2024.0.0           openmpi-x86_64       singularity/3.7.1
  cmake/3.24.2         hwloc/2.7.2    (L,D)  intel/2022.1.0    intel/2024.0.1           os                   ucx/1.15.0        (L)
  gnu12/12.3.0         intel/2021.4.0        intel/2022.2.0    libfabric/1.19.0 (L)     papi/6.0.0
```

Software stack

Selecting right tools for the right tasks is hard.

	Energy
(c) C	1.00
(c) Rust	1.03
(c) C++	1.34
(c) Ada	1.70
(v) Java	1.98
(c) Pascal	2.14
(c) Chapel	2.18
(v) Lisp	2.27
(c) Ocaml	2.40
(c) Fortran	2.52
(c) Swift	2.79
(c) Haskell	3.10
(v) C#	3.14
(c) Go	3.23
(i) Dart	3.83
(v) F#	4.13
(i) JavaScript	4.45
(v) Racket	7.91
(i) TypeScript	21.50
(i) Hack	24.02
(i) PHP	29.30
(v) Erlang	42.23
(i) Lua	45.98
(i) Jruby	46.54
(i) Ruby	69.91
(i) Python	75.88
(i) Perl	79.58

Interactive usage of tensorflow on GPU

Reserve GPU:

- **srun -p gpu --pty bash**

On the gpu node:

- **module load singularity**
- **singularity pull –docker-login docker://gitlab.aip.de:5005/akhalatyan/gpu-on-newton:main**
- **singularity run --bind /lustre/arm2arm:/lustre/arm2arm --nv gpu-on-newton_main.sif**

Runnig jobs

job.mpi

```
[arm2arm@nnewl1 ~]$ cat job.mpi
#!/bin/bash

#SBATCH -p debug         # partition name
#SBATCH -J test-hybrid               # Job name
#SBATCH -o job.%j.out         # Name of stdout output file (%j expands to jobId)
#SBATCH -e job.%j.err


### compute nodes
#SBATCH --nodes=1
###  MPI ranks
#SBATCH --ntasks=4
###  MPI ranks per node
#SBATCH --ntasks-per-node=4
###  tasks per MPI rank(eg OMP tasks)
#SBATCH --cpus-per-task=3
#SBATCH -t 01:30:00           # Run time (hh:mm:ss) - 1.5 hours

# Launch OMP+MPI-based executable

export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK


# to compile
# mpif90 test-hybrid.f90  -o test-hybrid.x -fopenmp

module load gnu9
module load openmpi4
# Run the code
prun test-hybrid.x
```

- to submit in the jobqueue:

```
sbatch   job.mpi
```

- check the status

```
[arm2arm@nnewl1 ~]$ squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
                22    normal     test  arm2arm  R       0:02      1 nnew003
```

- show job full status

```
[arm2arm@nnewl1 ~]$ scontrol show job
JobId=23 JobName=test
 UserId=arm2arm(1266) GroupId=dpk(1230) MCS_label=N/A
 Priority=4294901758 Nice=0 Account=(null) QOS=(null)
 JobState=RUNNING Reason=None Dependency=(null)
 Requeue=1 Restarts=0 BatchFlag=1 Reboot=0 ExitCode=0:0
 RunTime=00:00:02 TimeLimit=01:30:00 TimeMin=N/A
 SubmitTime=2021-11-19T00:35:18 EligibleTime=2021-11-19T00:35:18
 AccrueTime=2021-11-19T00:35:18
 StartTime=2021-11-19T00:35:18 EndTime=2021-11-19T02:05:18 Deadline=N/A
 SuspendTime=None SecsPreSuspend=0 LastSchedEval=2021-11-19T00:35:18
 Partition=normal AllocNode:Sid=nnewl1:1883529
 ReqNodeList=(null) ExcNodeList=(null)
 NodeList=nnew003
 BatchHost=nnew003
 NumNodes=1 NumCPUs=96 NumTasks=4 CPUs/Task=3 ReqB:S:C:T=0:0:*:*
 TRES=cpu=96,node=1,billing=96
 Socks/Node=* NtasksPerN:B:S:C=4:0:*:* CoreSpec=*
 MinCPUsNode=12 MinMemoryNode=0 MinTmpDiskNode=0
 Features=(null) DelayBoot=00:00:00
 OverSubscribe=NO Contiguous=0 Licenses=(null) Network=(null)
 Command=/home/arm2arm/job.mpi
 WorkDir=/home/arm2arm
 StdErr=/home/arm2arm/job.23.out
 StdIn=/dev/null
 StdOut=/home/arm2arm/job.23.out
 Power=
 NtasksPerTRES:0
```

Questions?

~~Use CLOUD everywhere!!!~~

Free Software Foundation Europe: fsfe.org