LRZ-LMU (+MPPMU) Site Report
A+C Verbundtreffen

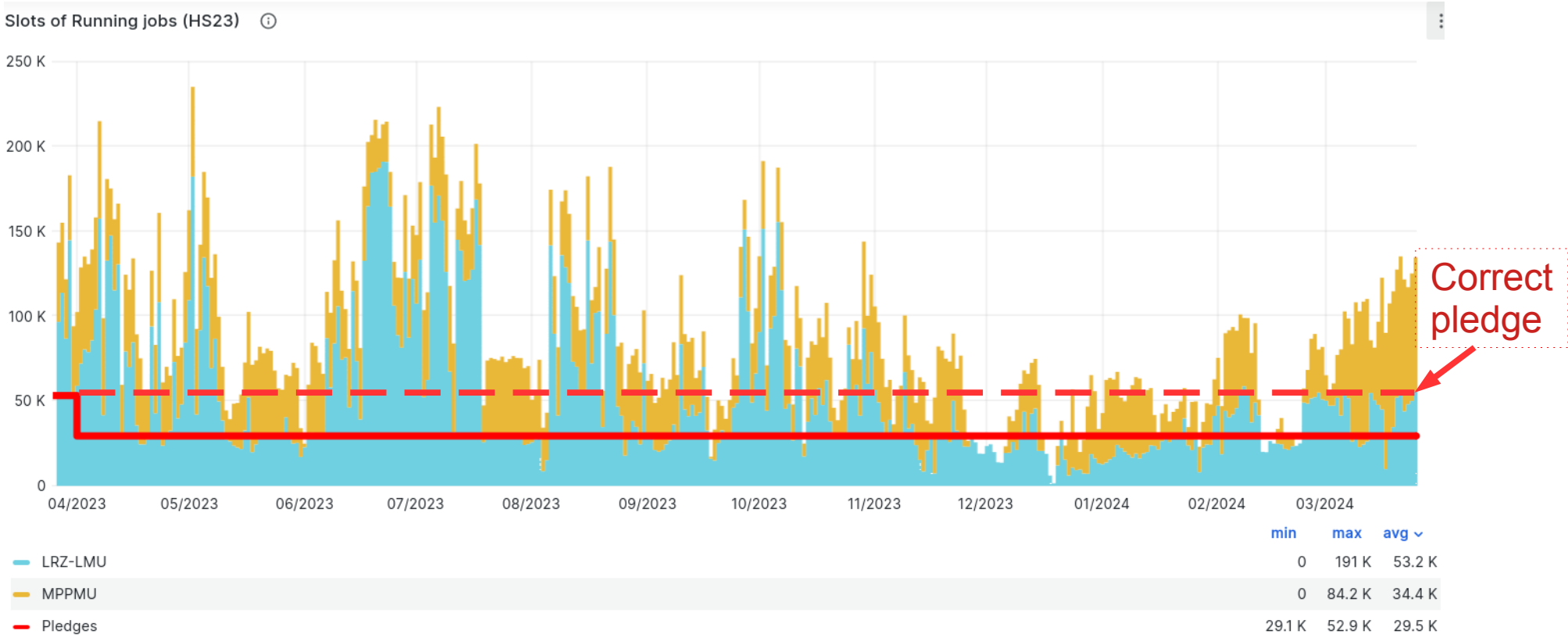
Mar 26, 2024
G. Duckeck
R. Walker

- Overview
- Production contribution
- Opportunistic add-ons
- Hardware status and procurement plans
- Sustainability considerations
- MPPMU status

Overview

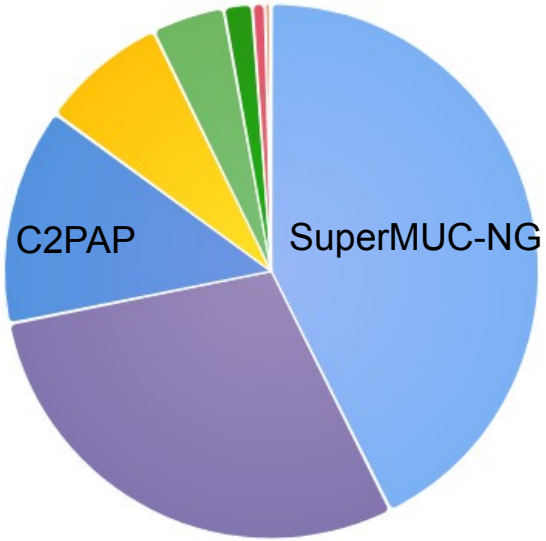
- Munich ATLAS T2 federation "MCAT" – 2 sites, 500 m apart in Garching, providing **WLCG pledges** since **2007**
  - LRZ-LMU:
    - Located at LRZ
    - CPU part of general Linux cluster ("attended housing")
      - Hardware setup, sys-admin, SLURM done by LRZ
      - CE operated by LMU team
    - Storage servers
      - Hardware setup done by LRZ
      - Sys-admin and dCache operated by LMU team
  - MPPMU
    - Located at MPCDF (former RZG, general HPC site for MPG)
    - CPU and storage operated by MPCDF staff
      - Partially paid by MPP
    - Started as ATLAS-only first
      - Now also service for small expts and theory groups of MPP
- Monthly "technical meetings"

LRZ-LMU and MPPMU jobs last Apr 23 – Mar 24



Opportunistic CPU resources for LRZ-LMU

Wall clock time. All jobs (HS23 seconds)

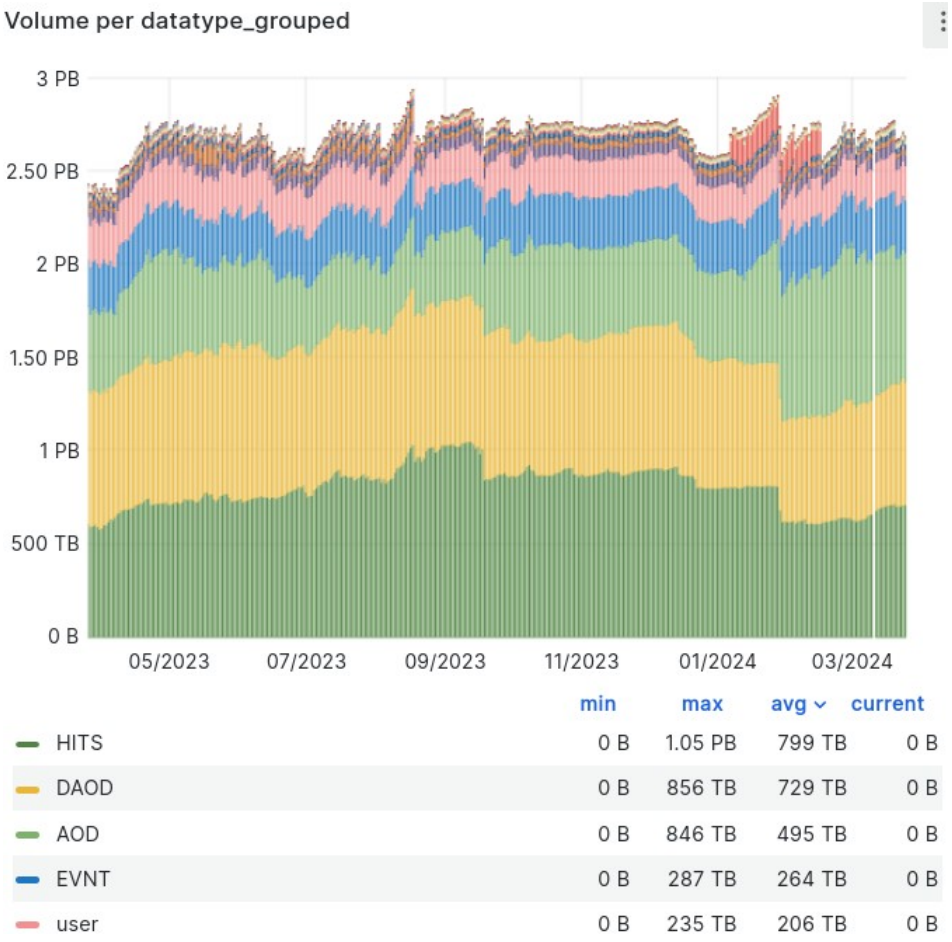


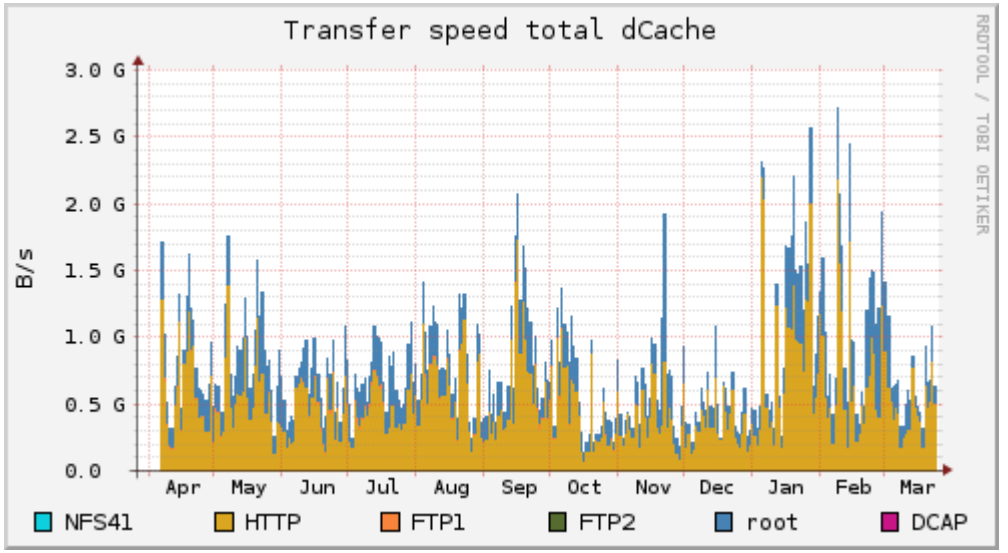
	Value	Percent
LRZ-LMU_MUC	722 Bil	43%
LRZ-LMU	495 Bil	29%
LRZ-LMU_C2PAP_FN	222 Bil	13%
LRZ-LMU_SIM	129 Bil	8%
LRZ-LMU_TEST	73.4 Bil	4%
ANALY_LRZ_VP	27.9 Bil	2%

About 60% of CPU from Opportunistic resources:
- SuperMUC-NG
- C2PAP
- LRZ-cloud (also via KIT C/T)

In addition HEPHY-UIBK (Innsbruck site, CPU only) uses LRZ-LMU storage for IO

LRZ-LMU storage usage





Hardware Status - LRZ-LMU

- CPU
  - Always purchased via LRZ “Rahmenvertrag”
    - Same config as used for LRZ HPC
    - SuSE SLES 15.1 enforced – causing problems sometimes
  - IBM/Lenovo (2014–2016) Xeon E5-2697v3
    - ~1200 cores – 24 kHS06
      - To be de-commisioned when new order in place
  - Megware (2020), Xeon 6230
    - ~700 cores – 17 kHS06
  - Ordered: Lenovo AMD EPYC 9654
    - ~1900 cores – 45 kHS06
    - Ordered last Nov, waiting for delivery – Apr?

Hardware Status - LRZ-LMU

- Disk – dCache
  - Debian 12 OS, brtfs on pools, dCache9.2
  - purchased via LMU “Rahmenvertrag”
  - Currently: HPE RAID6 servers
    - 18 nodes, 12x8TB → ~1400 TB from 2016 – out of maintenance
    - 19 nodes, 14x16 TB → ~3000 TB from 2021
  - Planned to order: Dell RAID6 server
    - ~10 nodes, 24x20 TB → ~4000 TB
    - Final iterations w/ vendor
- Disk – XCache
  - 2 disk-servers 80 TB each ((de-commisioned dCache pool nodes)
  - 1 SSD server with 20 TB
  - Integrated into ATLAS analysis via “virtual placement” service

Person power - LRZ-LMU

- Otmar Biebel – ATLAS group leader
- Guenter Duckeck
- Rod Walker – Fed A+C Comp (core-computing)
- Alex Lory – Fed A+C Comp (site and expt support – HammerCloud)
- Christoph Ames – Fed A+C Comp (site and expt support -- Rucio)
- Christoph Mitterer – (Storage and dCache)
- Nikolai Hartmann – (Xcache, Columnar analysis)
- David Koch – Fidium (Analysis Grand Challenge, OpenData)
- LRZ staff ...

Slide
C. d. Fratte

MPPMU site Report

- 7PB dCache storage mostly ATLAS but also BELLE, CRESST and others. dCache 7.2 but 9.2 ASAP

- 186 Computing servers shared, via Slurm among ATLAS, BELLE and local jobs

- Fixed our computing node SMP misconfiguration, as shown yesterday by David South, and evaluated our HEPScore23 carefully.

Slide
C. d. Fratte

MPPMU site Report

- dr. Meisam Tabriz just joined our group.

- We are in the process of buying new Computing and Storage servers

- The new computing resources will be assigned to the MPCDF Openstack Cloud. We already have a Slurm cluster deployment procedure using JADE

- the latter implies also we will also get rid of all the VMs and hypervisors moving all the main services into the Cloud as well as the computing

- the old cluster will be slowly decommissioned

- We have no estimation about power consumption and did not yet implemented a strategy concerning sustainability.

Sustainability considerations - LRZ-LMU

- LRZ HPC – CoolMuc
  - Direct warm-water cooling – PuE ~ 1.1 for compute nodes (~1.4 for storage)
  - Heat used for building
- Power contract LRZ:
  - 80% flat rate, 20% variable (Stromboerse) for LRZ overall
    - but not available for small consumers (ATLAS/WLCG ~20 kW vs SuperMUC-NG ~4 MW)
- CPU:
  - Xeon E5-2697v3 (2017) ~0.58 W/HS06
  - Xeon 6230 (2020): ~0.33 W/HS06
  - AMD EPYC 9654 (2024): ~0.10 W/HS23 (ordered)
  - Rod's CPU frequency regulator in principle in place: tested & operational on sub-set

Sustainability considerations – LRZ-LMU - 2

- Storage: current setup – single box servers:
  - HP Raid6 10(+2)x8 TB (2017): ~2.3 W/TB
  - HP Raid6 12(+2)x16 TB (2021): ~1.2 W/TB
  - *large fraction used by server CPU – rather go to bigger units for new purchase*
    Dell 20(+4)x20 TB : ~0.8 W/TB → tbc
  - *New purchase*
- Lifetime:
  - usually run CPU and Disk beyond 5 y maint – switch off problematic nodes and use as spare
- Substantial savings over time for LRZ-LMU:
  Capacity vs Power: factor 4.2 increase from 2016 → 2022

	2016	***2022***	***ratio***
CPU (HS06)	9433	27600	**2.9**
Disk (TB)	1200	2467	**2.1**
Power (kW)	42.6	25.2	**0.6**
Perf/Power			**4.2**

Slide
Rod Walker

In the year 2030, if we are still here ….

Capacity	2022(GW)	2030 (GW)	Factor
Offshore Wind	7.8	30	4
Onshore Wind	56	115	2
Solar	66	215	3

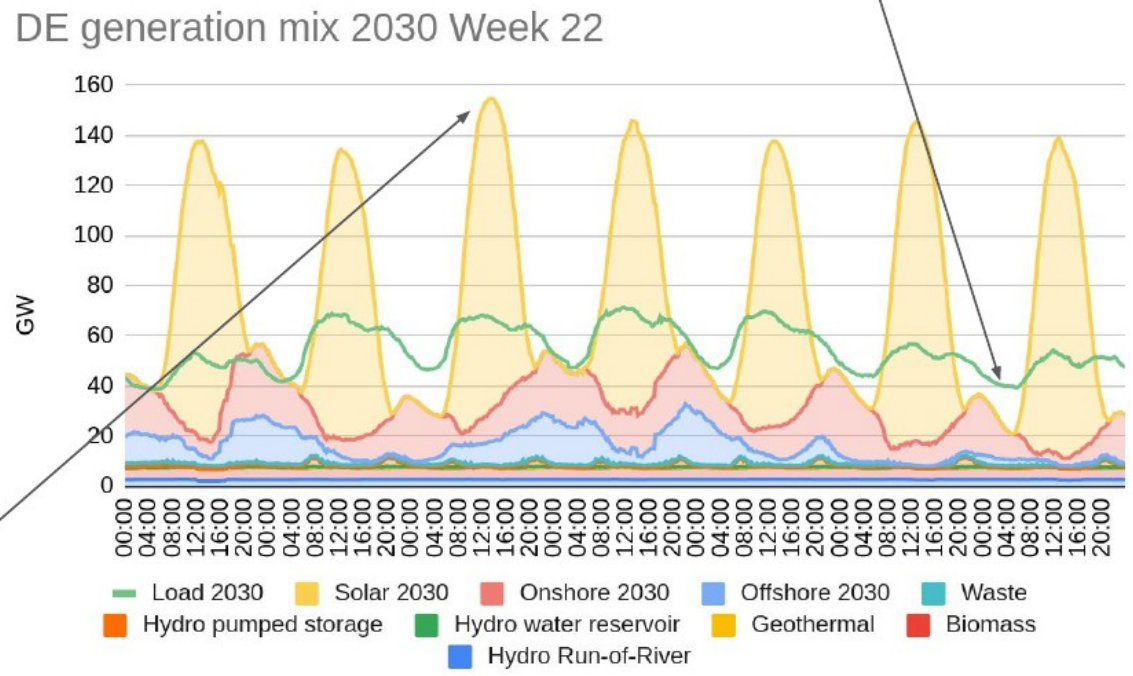
https://www.bmwk.de/Redaktion/DE/Dossier/erneuerbare-energien.html

Assume the same geographical distribution and weather, then these simply **scale up** the respective contributions.

Load will change too.
E-Auto, Heatpumps.
11% increase

Need 70GW on-demand

Still periods needing gas generation, unless we can reduce the load.



Slide
Rod Walker

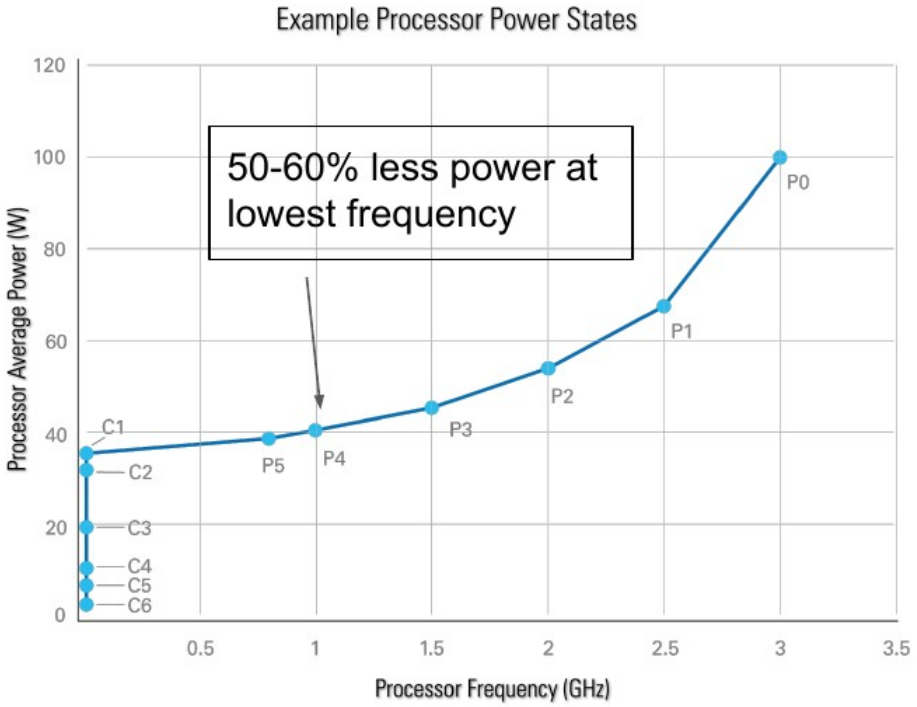
Can a Datacenter modulate power consumption?

- Mostly HTC where a few hours or days delay is irrelevant.
- Obvious saving is from turning off compute nodes
  - ok for longer predictable pauses
    - but lengthy draining of long jobs, without checkpointing
  - twice per day is unfeasible
- Can we reduce power without draining jobs
  - freeze processes to let CPU sleep
    - large drop in node power, but base usage still there for no work done
  - reduce CPU frequency to minimum
    - smaller drop in node power, ~50%, but still doing work
    - does it pay off, given base usage?
  - switch to battery
    - solar battery systems prices tumble: would add 3-6% to CPU server cost
      - battery/inverter lifetime costs with 6000 cycles gives ~ 10ct/kWh stored and returned. Price variations ~>10ct cover battery cost, today.

Slide
Rod Walker

CPU frequency modulation

Free, fast, repeatable, harmless to workloads
Set CPU governor to PowerSave



$P \sim f V^2$

dynamic voltage and frequency scaling (DVFS)

voltage reduces with frequency

Useful work ~ frequency, but power falls faster than frequency

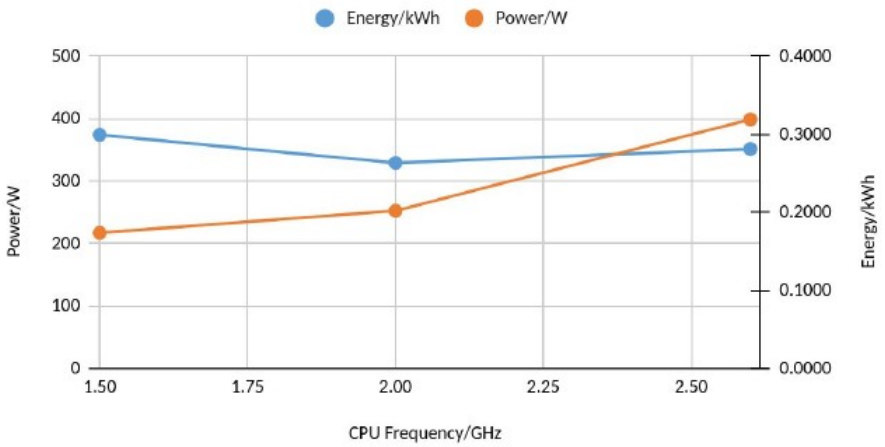
Could offset base/non-CPU node power consumption

Slide
Rod Walker

Real-world measurements: HEP work vs total node power

1) Same work at different frequencies:
1000evt GEANT4 simulation

dual-x86
E.Simili, Glasgow



- **HEP work per kWh not significantly less at lowest frequency**
  - Glasgow 6% & DESY 2%
- Middle frequency best for both!
  - fewer voltage steps?
  - highest frequency at lowest V

Frequency/GHz	HS06	Power/W	HS06/GHz	HS06/W	Ratio to high
1.5	1085	286	723	3.79	98%
2.15	1424	330	662	4.32	111%
2.85	2032	524	713	3.88	100%

2) AMD node HEPSpec vs f
T.Hartmann, DESY

Summary

- LRZ-LMU
  - Stable and efficient operation
    - Substantial extra CPU from opportunistic resources
  - Large and last purchase in progress
  - In good shape for next years
- MPPMU
  - Stable and efficient operation
  - Site provides important service also for other MPP groups/small experiments
  - To my knowledge operation expected to continue in coming years