Karlsruhe Institute of Technology

GridKa – operations and resources planning

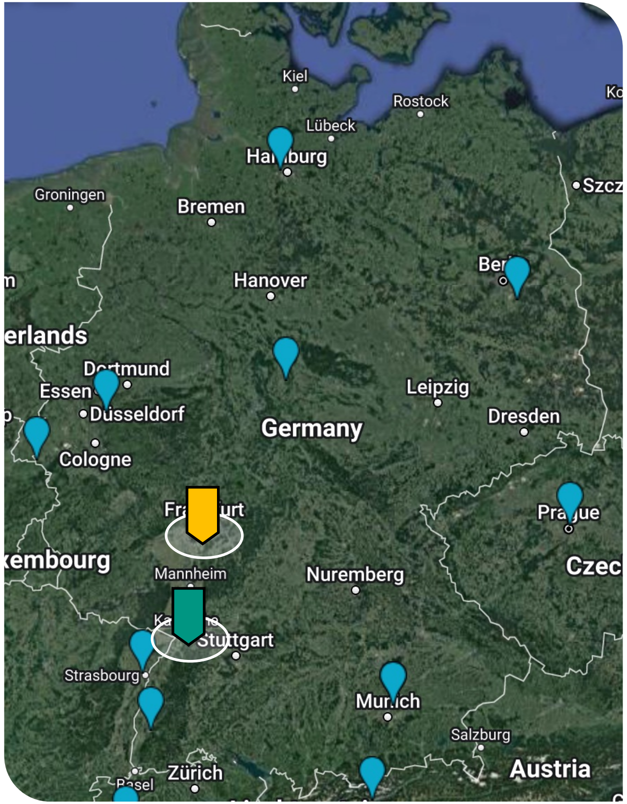
aka "KIT", "FZK", "FZK-LCG2"

ALICE Tier-1/Tier-2 Workshop 2022, Budapest

Max Fischer

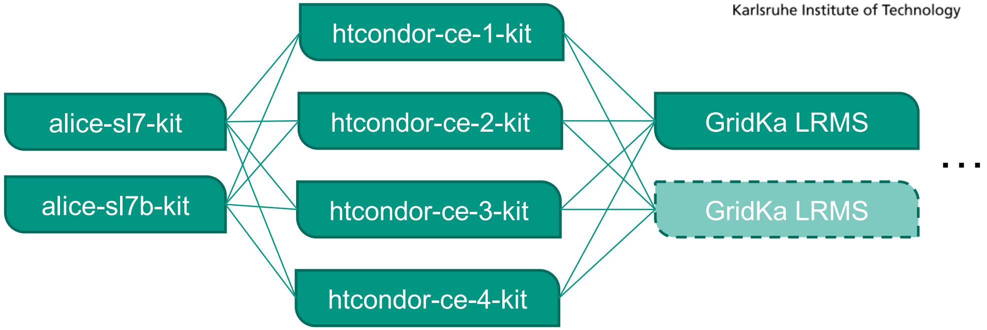
GridKa at a glance

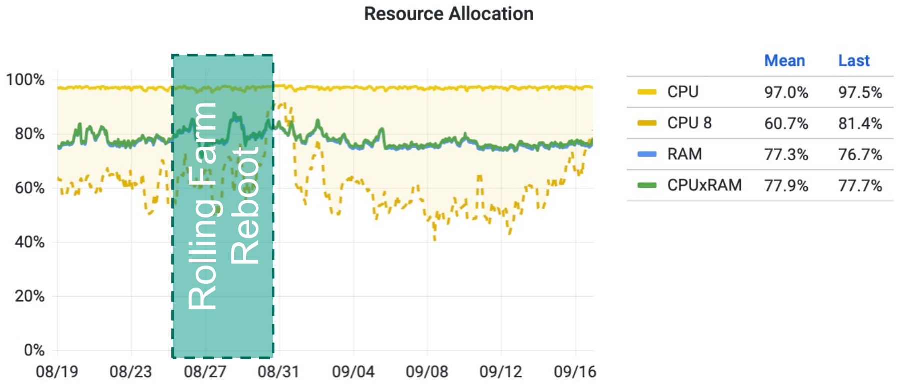
- Many German grid contributors
  - HGF centres: DESY, GSI, KIT
  - Several university compute centres
- GridKa: Multi-VO Tier 1 at KIT
  - Primarily LHC VOs and Belle2
  - Condor CE + LRMS
    - About 800k HS06
      ALICE: 124kHS ’22 / 143kHS ‘23
  - XRootD / dCache
    - About 50 PB Disk
      ALICE: 14PB ’22 / 16PB ‘23
    - About 100 PB Tape
      ALICE: 16PB ’22 / 21PB ‘23



Compute Middleware

- HTCondor-CE since ‘20
  - Previously ARC-CE v5
  - Very satisfied with scalability
    - (after contribs for accounting)
  - Token Auth transition ongoing
- HTC-LRMS setup rewritten in ’22
  - Adjust to new, larger WNs
  - Removed VO-specific partitions
  - Consistent +97% allocation for any workflow mix (MC vs SC, VOs, …)
- Preparing for HepScore22

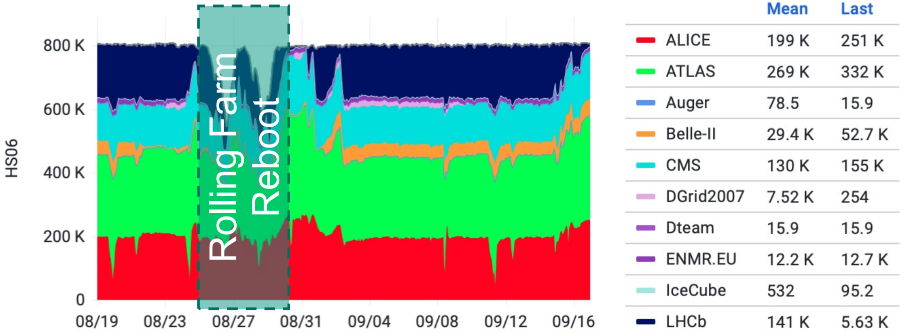




Compute Hardware

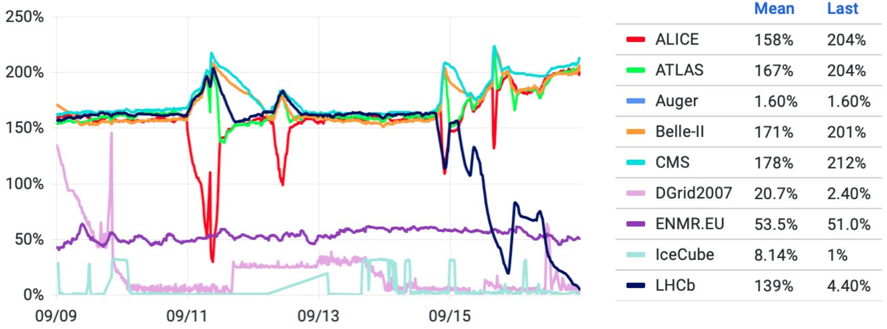
- Single procurement for ‘22 and ‘23
  - ~200 machines, each 192 Cores with ~3GB/Core at ~3KHS06
  - +previous: ~75 each 256 Cores ~2GB/Core at ~2.5kHS06
  - More than sufficient for ‘23 pledge
- Operational Considerations
  - Unclear: Future power costs
  - Long-term planning for new cooling
  - Older hardware (~200k HS06) already too inefficient to operate

Cumulative Usage



	Mean	Last
🟥 ALICE	199 K	251 K
🟩 ATLAS	269 K	332 K
🟦 Auger	78.5	15.9
🟧 Belle-II	29.4 K	52.7 K
CMS	130 K	155 K
DGrid2007	7.52 K	254
Dteam	15.9	15.9
🟪 ENMR.EU	12.2 K	12.7 K
IceCube	532	95.2
LHCb	141 K	5.63 K

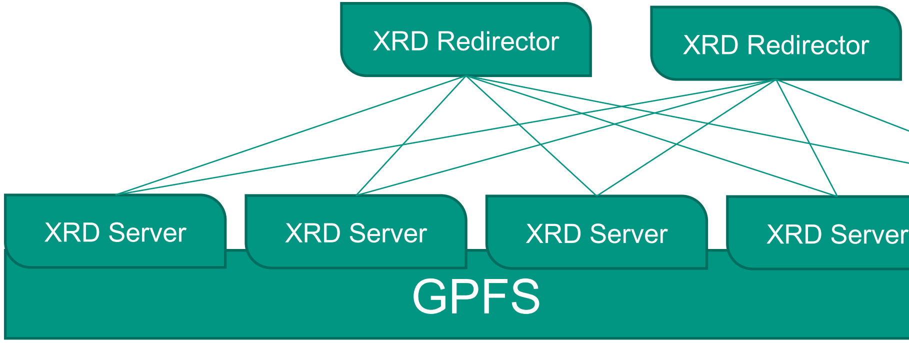
Relative Pledge Usage



	Mean	Last
🟥 ALICE	158%	204%
🟩 ATLAS	167%	204%
🟦 Auger	1.60%	1.60%
🟧 Belle-II	171%	201%
CMS	178%	212%
DGrid2007	20.7%	2.40%
🟪 ENMR.EU	53.5%	51.0%
IceCube	8.14%	1%
LHCb	139%	4.40%

Storage Middleware: ALICE::FZK::(SE|TAPE)

- One multi-PB GPFS per VO/SE
  - XRootD VM redirector, HW servers
  - Investigating HW redirectors for better metadata performance
  - XRootD still on v4 ☹
    - … see next slide
- Tape backend migration upcoming
  - TSM => HPSS, new tape libraries
  - Completed for CMS, LHCb, Belle2, currently ATLAS ongoing
  - Transparent via XRootD wrapper, manual migration of old data



XRD Redirector

XRD Redirector

XRD Server

XRD Server

XRD Server

XRD Server

GPFS

2022 Tape Challenge	ATLAS (TSM)	CMS (HPSS)	LHCb (HPSS)
Read/drive	125 MB/s	300MB/s	300MB/s
Write/drive	100 MB/s	300MB/s	300MB/s

Storage Hardware

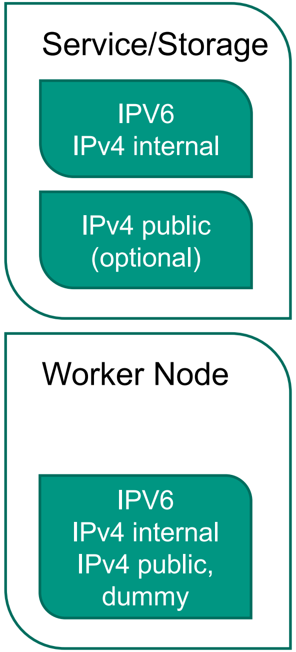
- HW of pledge ‘22 still not available
  - Delivery and performance problems
  - Final benchmarking, ETA 1 month
  - Will cover ‘23 pledge as well

- Same setup, new filesystem
  - Need to copy from old=>new HW, will be performed by GridKa
  - Simpler than ‘17/’18 migration, expected to be faster

- This illustration sadly left blank

Miscelaneous: IPv6 and Network

- Full IPv6 dual home/stack setup
  - Pushed by KIT internally as well
  - Many internal services IPv6 only
  - Hard to test thoroughly due to IPv4 fallbacks…
- Preparing public IPv4 on WNs
  - Deprecate old FW and NAT
  - New IPv4 cheaper than new FW
  - Better scalability and performance


Service/Storage

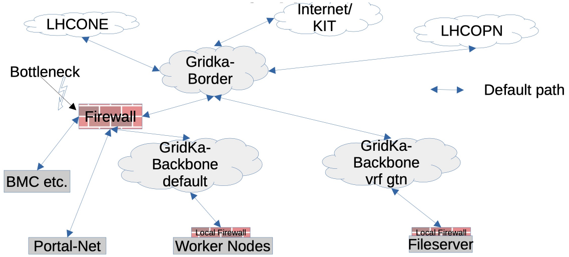
IPV6
IPv4 internal

IPv4 public
(optional)

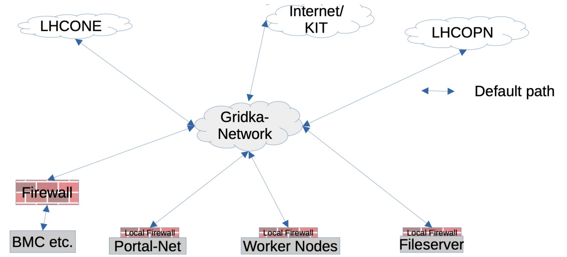
Worker Node

IPV6
IPv4 internal
IPv4 public,
dummy

Current Network (simplified)

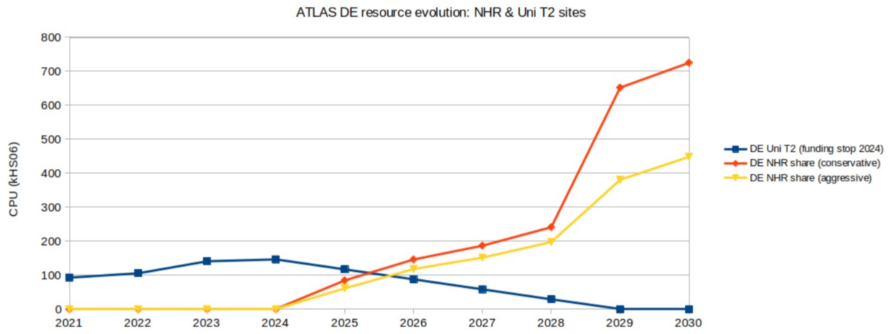


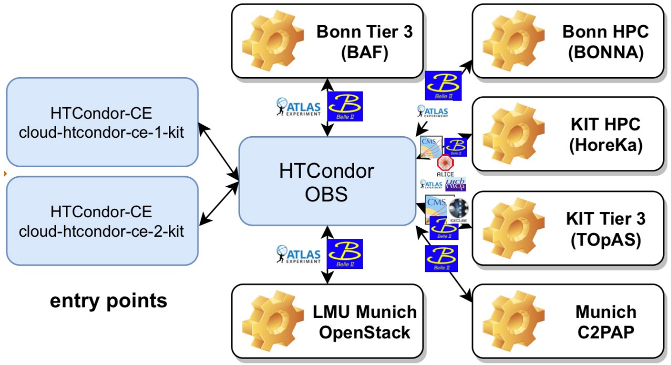
Target Network



Miscelaneous: WLCG landscape in Germany

- Shift in resource distribution
  - Phase out university hosted Tier 2
  - Concentration of storage at HGF centres, extend network+caches
  - Use of HPC compute, HEP groups and centres as stakeholders
- Several R&D infrastructures
  - GridKa provides separate CE for VO-agnostic overlay batch system
  - Mainly driven by CMS, ATLAS and some sites with co-located HPC




HTCondor-CE
cloud-htcondor-ce-1-kit

HTCondor-CE
cloud-htcondor-ce-2-kit

entry points

HTCondor
OBS

Bonn Tier 3
(BAF)

LMU Munich
OpenStack

Bonn HPC
(BONNA)

KIT HPC
(HoreKa)

KIT Tier 3
(TOpAS)

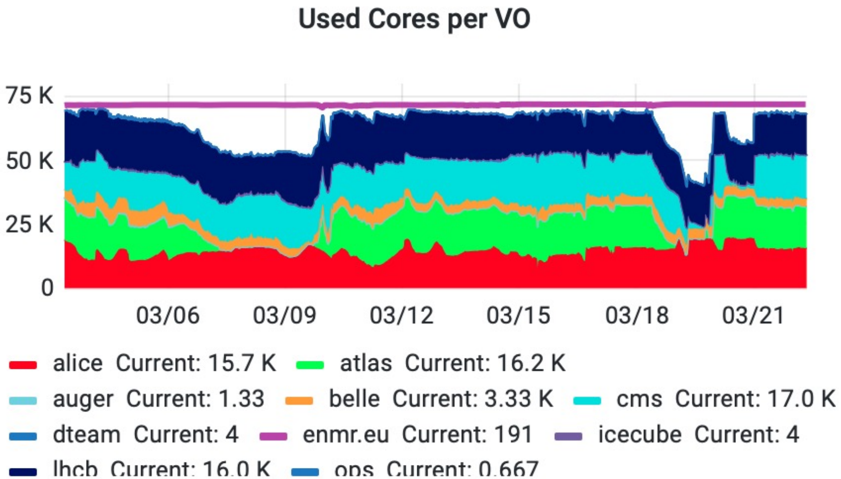
Munich
C2PAP

TLDR

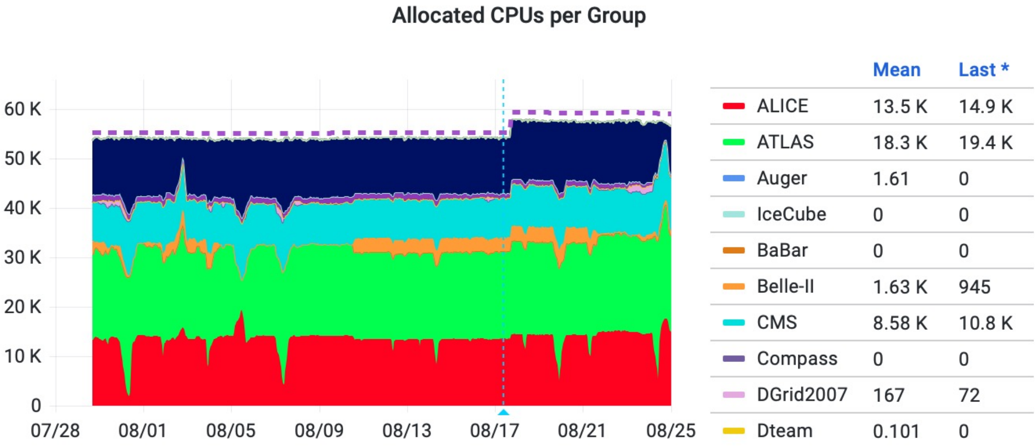
- Compute ready for ‘22+’23
  - Extra resources on best effort
- Delayed ‘22+’23 storage delivery
  - Starting preparations soon
  - Need migrations in backends
- Major changes on network
  - Done: IPv6 for VOs and services
  - Distributed FW and public IPv4

	‘22	‘23
HTCondor	👷	🙂
Worker Nodes	😳	🙂
XRootD	😴	👷
Disk	😱	🙂
Tape	😴	👷
Power	😀	🤷‍♂️
IPv6	😀	🥳

Old vs New Farm Scheduler Setup



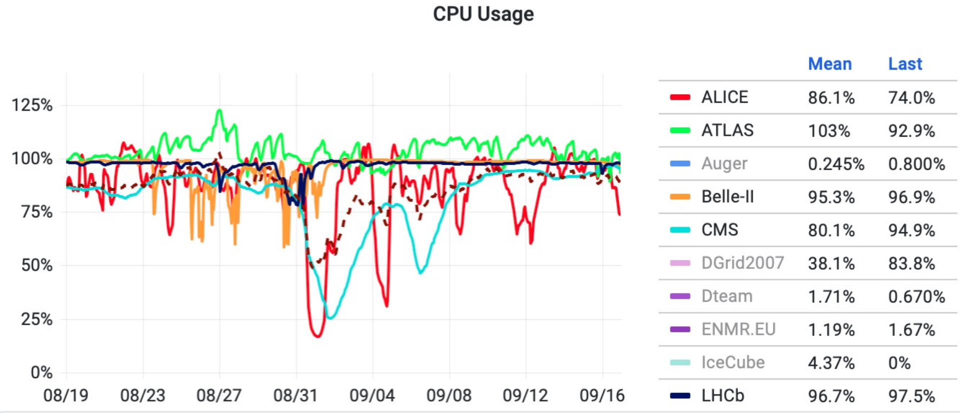
Used Cores per VO



Allocated CPUs per Group

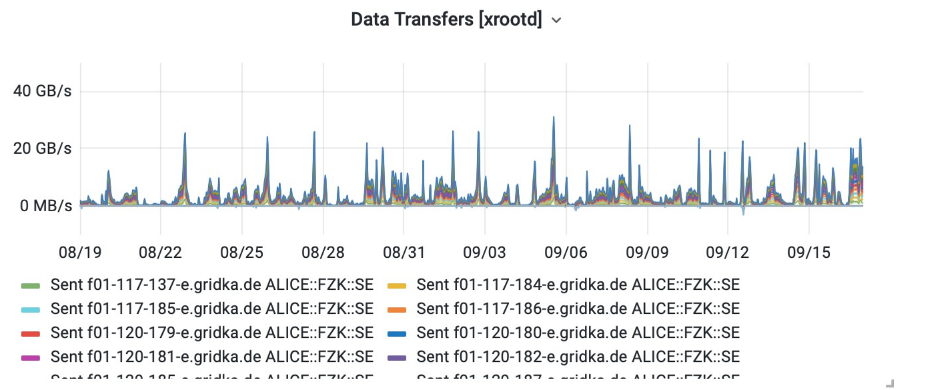
	Mean	Last *
🟥 ALICE	13.5 K	14.9 K
🟩 ATLAS	18.3 K	19.4 K
🟦 Auger	1.61	0
▬ IceCube	0	0
🟧 BaBar	0	0
🟧 Belle-II	1.63 K	945
▬ CMS	8.58 K	10.8 K
🟪 Compass	0	0
▬ DGrid2007	167	72
🟨 Dteam	0.101	0

resource usage



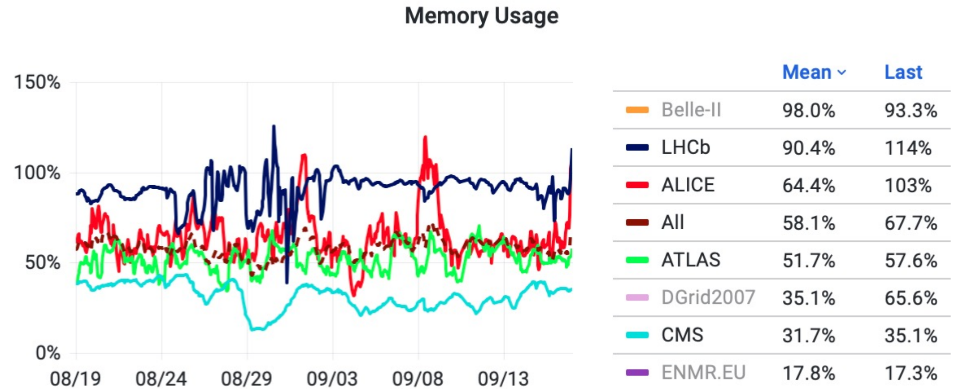
CPU Usage

	Mean	Last
▬ ALICE	86.1%	74.0%
▬ ATLAS	103%	92.9%
▬ Auger	0.245%	0.800%
▬ Belle-II	95.3%	96.9%
▬ CMS	80.1%	94.9%
▬ DGrid2007	38.1%	83.8%
▬ Dteam	1.71%	0.670%
▬ ENMR.EU	1.19%	1.67%
▬ IceCube	4.37%	0%
▬ LHCb	96.7%	97.5%



Data Transfers [xrootd] ⌄

- ▬ Sent f01-117-137-e.gridka.de ALICE::FZK::SE
- ▬ Sent f01-117-184-e.gridka.de ALICE::FZK::SE
- ▬ Sent f01-117-185-e.gridka.de ALICE::FZK::SE
- ▬ Sent f01-117-186-e.gridka.de ALICE::FZK::SE
- ▬ Sent f01-120-179-e.gridka.de ALICE::FZK::SE
- ▬ Sent f01-120-180-e.gridka.de ALICE::FZK::SE
- ▬ Sent f01-120-181-e.gridka.de ALICE::FZK::SE
- ▬ Sent f01-120-182-e.gridka.de ALICE::FZK::SE
- Sent f01-120-185-e.gridka.de ALICE::FZK::SE
- Sent f01-120-187-e.gridka.de ALICE::FZK::SE



Memory Usage

	Mean ⌄	Last
▬ Belle-II	98.0%	93.3%
▬ LHCb	90.4%	114%
▬ ALICE	64.4%	103%
▬ All	58.1%	67.7%
▬ ATLAS	51.7%	57.6%
▬ DGrid2007	35.1%	65.6%
▬ CMS	31.7%	35.1%
▬ ENMR.EU	17.8%	17.3%