Karlsruhe Institute of Technology

NATIONALES HOCHLEISTUNGS RECHNEN

GridKa Overview Report

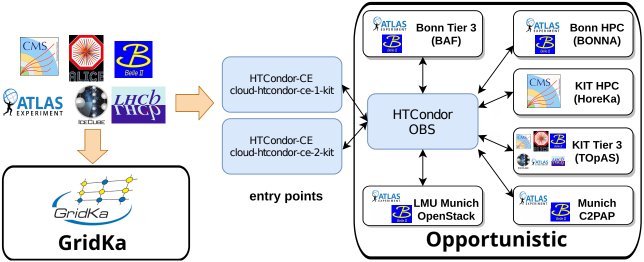
@ Annual Meeting of the ATLAS and CMS Computing Verbund

Robin Hofsaess on behalf of the GridKa CMS and R&D team | 26.03.2024

Outline

- GridKa Status and News
  - Pledges and Resources for CMS
  - Opportunistic Resources at GridKa
  - New NVMe cache for HPSS tape system
  - WLCG Data Challenge ’24
  - ARM workers delivered
  - Update to RHEL8
  - Update of Compute Elements
  - HappyFace4 development
  - Progress in GPU Usage for CMS

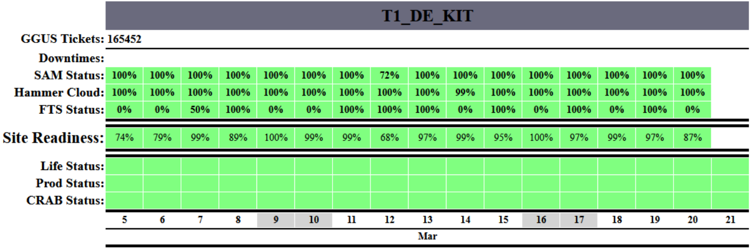
GridKa Overview



Opportunistic Resources

Successfully integrated with COBalD/TARDIS – developed at KIT

GridKa Status



T1_DE_KIT

GGUS Tickets: 165452

Downtimes:

	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
SAM Status:	100%	100%	100%	100%	100%	100%	100%	72%	100%	100%	100%	100%	100%	100%	100%	100%	
Hammer Cloud:	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%	100%	100%	100%	100%	
FTS Status:	0%	0%	50%	100%	0%	0%	100%	100%	100%	0%	100%	0%	100%	0%	100%	0%	
Site Readiness:	74%	79%	99%	89%	100%	99%	99%	68%	97%	99%	95%	100%	97%	99%	97%	87%	
Life Status:	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Prod Status:	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CRAB Status:	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Mar

Site Status

GridKa is one of the most reliable Tier 1 centers! (Readiness Report)

Pledges and Resources for CMS

Resource	Pledges 2024
CPUs	93k HEPscore23 (≈ 6900k Cores)
Disk	12.2 PB
Tape	38 PB

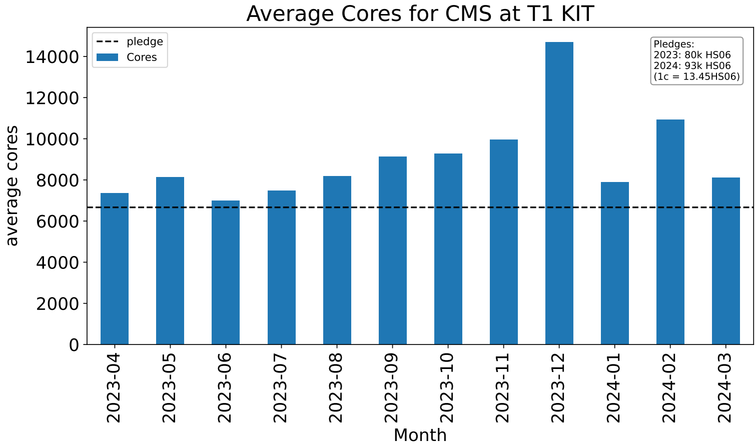
Tier 3: (opportunistic)

dCMS Disk	2.8 PB
CPUs	1000 cores
*GPUs	56

*Prototype integration of GPUs into the Grid

Summary

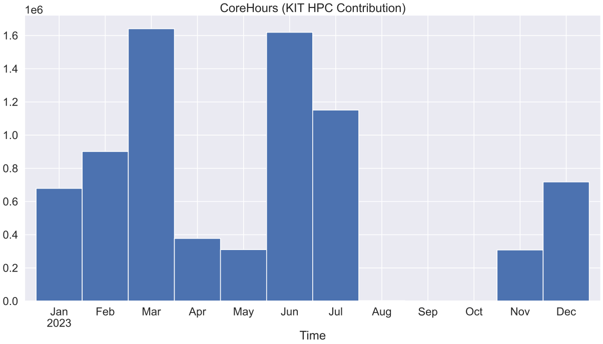
- The pledged tape and disk will be fully available in April
- All pledges will be fulfilled
- Additionally: We provide 56 GPUs with our Tier 3 (accessible for CMS production and via CRAB/CMS Connect for users)



Tier 1 Compute Resources

Always over pledge for all VOs

Integration of HoreKa



HoreKa HPC Resources

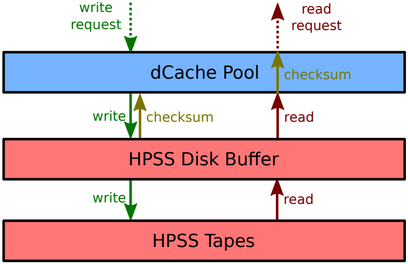
(Opportunistic) contribution up to an average T2!

NVMe Cache for HPSS Tape System

- New NVMe cache **in production** since Q4 ’23
  - 300TB fast NVMe
  - 4x100G network
  - To replace old HDD cache
- Currently, 95 PB on GridKa tape in total

Milestone:

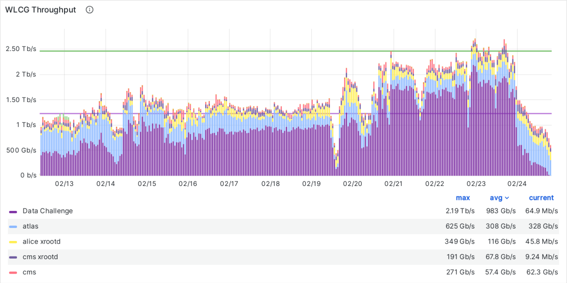
Stable operation with the desired **300 MB/s - 400 MB/s** read and write rate per tape drive



More details on our tape system in backup

WLCG Data Challenge ’24: Overview

- 12.02 to 24.02
- 1st week:
  1.2 Tb/s target ✓
- 2nd week:
  2.4 Tb/s target (✓)



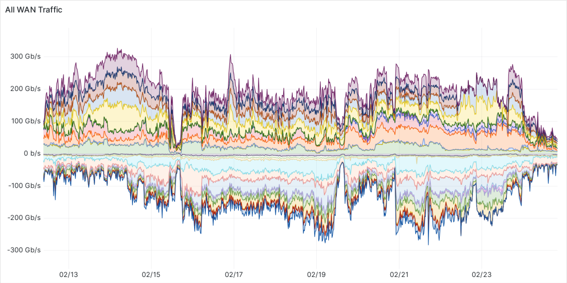
	max	avg	current
Data Challenge	2.19 Tb/s	983 Gb/s	64.9 Mb/s
atlas	625 Gb/s	308 Gb/s	328 Gb/s
alice xrootd	349 Gb/s	116 Gb/s	45.8 Mb/s
cms xrootd	191 Gb/s	67.8 Gb/s	9.24 Mb/s
cms	271 Gb/s	57.4 Gb/s	62.3 Gb/s

Result

In general, a success and valuable lessons learned for further optimization

WLCG Data Challenge ’24: GridKa Network Perspective

- LHCOPN: 300 Gb/s in each direction for DC 24
- LHCONE+Internet: 200 Gb/s in each direction



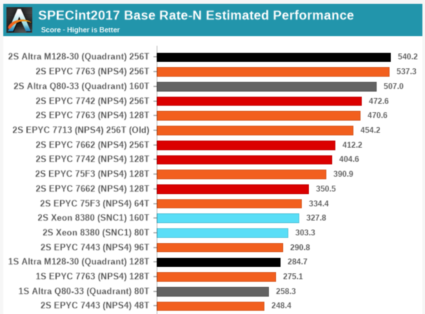
Conclusion

From the network perspective, GridKa was not at its limits

ARM Worker Nodes Delivered

- 15 nodes ordered with 2x AMPERE_Altra Max (2x128 cores)
- Test machines are promising (80 Cores 3,3 GHz)

AMD (64c)	ARM (80c)
2828 HS23 600W 4.71 HS/W	1606 HS23 380W **5.74** HS/W



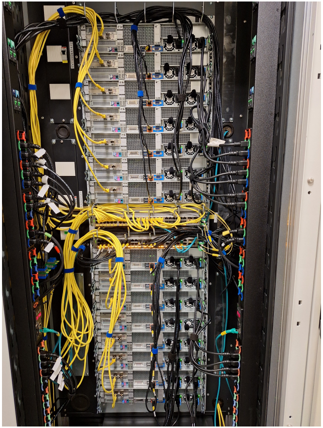
Source: anandtech.com

Provisioning: Q2 ’24 (expected)

Will accept jobs from our CEs

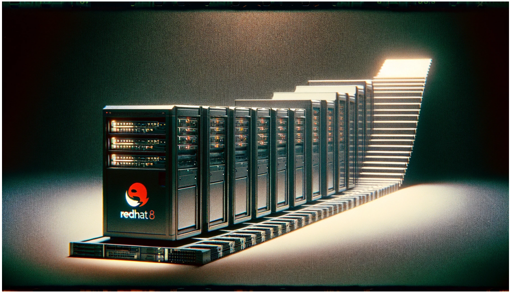
More info: CHEP poster on energy efficiency of ARM

ARM Worker Nodes Delivered

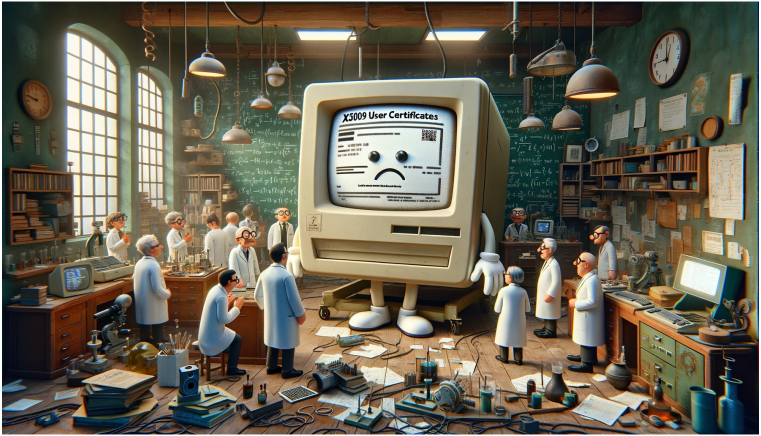


Update to RHEL8

- CentOS 7 will reach EOL end of June
- All machines will be updated to RHEL8 in the near future
  - Includes also the Tier 3 resources



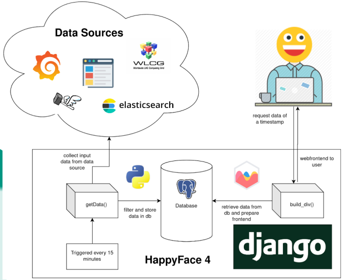
Update of CEs: Farewell to X509!



Within the next months, our **CEs** will be updated and X509 certificates will be deprecated.

HappyFace4

- What is HappyFace?
  - Meta monitoring tool for computing sites
  - Gives user a fast overview of the site status
  - GridKa production instance: happyface@ETP

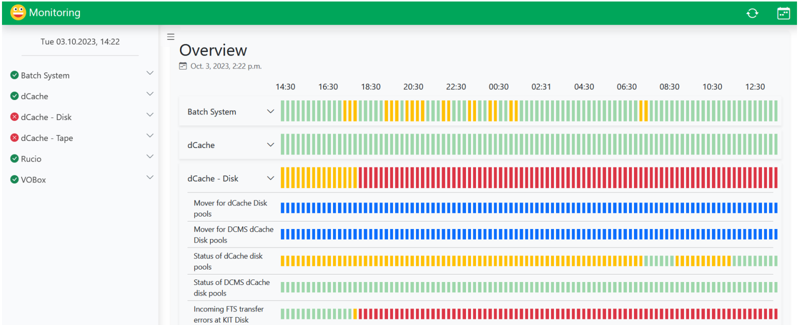

Data Sources
WLCG
Worldwide LHC Computing Grid
elasticsearch
request data of a timestamp
webfrontend to user
collect input data from data source
getData()
filter and store data in db
Database
retrieve data from db and prepare frontend
build_div()
Triggered every 15 minutes
HappyFace 4
django

Purpose

- Advanced (meta) monitoring for our Tier 1 center
- All necessary information for reliable operations gathered in one place

→ Extremely helpful for shifters to detect and report problems fast and in detail!

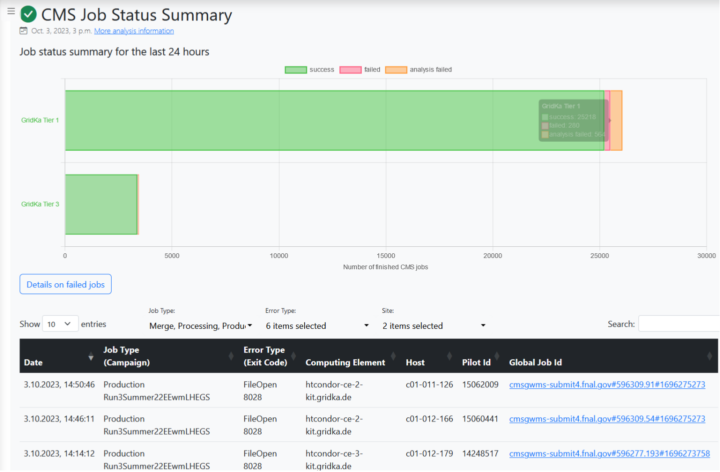
HappyFace4: Example



Don’t panic

- **Red** ≠ GridKa broken!
- **All** red and yellow issues are regularly followed up by our operations team to be understood :-)

HappyFace4: Details



CMS Job Status Summary

Oct. 3, 2023, 3 p.m. More analysis information

Job status summary for the last 24 hours

Details on failed jobs

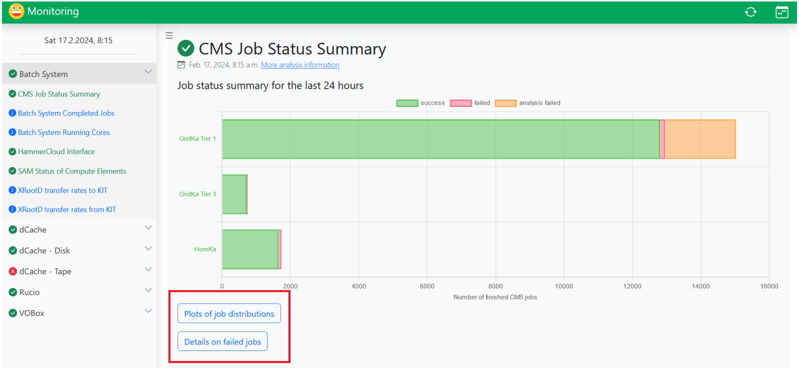
Job Type: Merge, Processing, Produ▾ | Error Type: 6 items selected ▾ | Site: 2 items selected ▾

Show 10 entries | Search:

Date	Job Type (Campaign)	Error Type (Exit Code)	Computing Element	Host	Pilot Id	Global Job Id
3.10.2023, 14:50:46	Production Run3Summer22EEwmLHEGS	FileOpen 8028	htcondor-ce-2-kit.gridka.de	c01-011-126	15062009	cmsgwms-submit4.fnal.gov#596309.91#1696275273
3.10.2023, 14:46:11	Production Run3Summer22EEwmLHEGS	FileOpen 8028	htcondor-ce-2-kit.gridka.de	c01-012-166	15060441	cmsgwms-submit4.fnal.gov#596309.54#1696275273
3.10.2023, 14:14:12	Production Run3Summer22EEwmLHEGS	FileOpen 8028	htcondor-ce-3-kit.gridka.de	c01-012-179	14248517	cmsgwms-submit4.fnal.gov#596277.193#1696273758

GridKa Status and News
○○○○○○○○○○○○○○●○○○○

HappyFace4: Recent Development


Monitoring
Sat 17.2.2024, 8:15
Batch System
CMS Job Status Summary
Batch System Completed Jobs
Batch System Running Cores
HammerCloud Interface
SAM Status of Compute Elements
XRootD transfer rates to KIT
XRootD transfer rates from KIT
dCache
dCache - Disk
dCache - Tape
Rucio
VOBox

CMS Job Status Summary

Feb. 17, 2024, 8:15 a.m. More analysis information

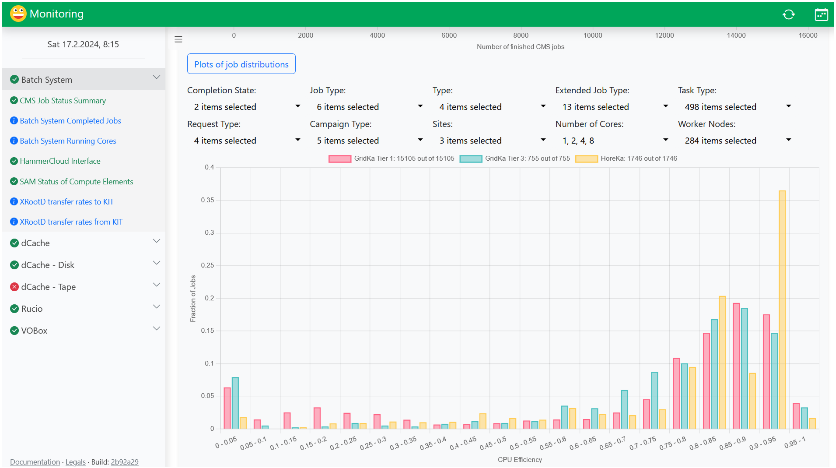
Job status summary for the last 24 hours

🟩 success 🟥 failed 🟧 analysis failed
GridKa Tier 1
GridKa Tier 3
HoreKa
0
2000
4000
6000
8000
10000
12000
14000
16000
Number of finished CMS jobs
Plots of job distributions
Details on failed jobs

New Features (credits to Artur Gottmann!)

CPU efficiency monitoring per job grouped by site, job type, cores, ...

HappyFace4: Recent Development


Monitoring
Sat 17.2.2024, 8:15
Batch System
CMS Job Status Summary
Batch System Completed Jobs
Batch System Running Cores
HammerCloud Interface
SAM Status of Compute Elements
XRootD transfer rates to KIT
XRootD transfer rates from KIT
dCache
dCache - Disk
dCache - Tape
Rucio
VOBox
Documentation · Legals · Build: 2b92a29
0
2000
4000
6000
8000
10000
12000
14000
16000
Number of finished CMS jobs
Plots of job distributions

Completion State: 2 items selected

Job Type: 6 items selected

Type: 4 items selected

Extended Job Type: 13 items selected

Task Type: 498 items selected

Request Type: 4 items selected

Campaign Type: 5 items selected

Sites: 3 items selected

Number of Cores: 1, 2, 4, 8

Worker Nodes: 284 items selected

Progress in GPU Usage for CMS

GridKa provides GPUs as a (opportunistic) prototype service via the Tier 3!

- **Current Development and Challenges:**
  - Efficient grid integration of GPU resources
  - Scheduling optimization of small scale GPU jobs
  - Contribution to the future of GPU in the HEPScore benchmark
  - Energy efficiency benchmarks of HEP GPU applications

- **Currently used by CMS DeepTau group**

Summary

GridKa Status

- **No problems** to report
- One of the most **reliable** Tier 1s
- All pledges will be **fulfilled**

Hardware

- New **NVME cache** is fully operational
- **ARM** machines delivered are about to be provisioned

Data Challenge 24

- GridKa participated successfully
- Within **Top-3** of sites
- Network has further potential

Opportunistic Resources

- Our opportunistic resources successfully provide up to **several 1000** additional cores

Upcoming Updates

- All machines will be upgraded to **RHEL8**
- With updating the Compute Elements, **X509 certificates** will be deprecated

Recent Development

- **HappyFace4**, our multi-facet observation tool for the Tier 1 operation, is constantly improved
- **GPU** integration and optimization

BACKUP

The GridKa CMS and R&D Team!

- Prof. Dr. Günter Quast
- Dr. Manuel Giffels
- Dr. Artur Gottmann
- Dr. Matthias Schnepf (BELLE, T3)
- Dr. Max Fischer (ALICE, T3)
- Dr. Maximilian Horzela
- Dr. Sebastian Brommer
- Robin Hofsaess
- Tim Voigtlaender
- Jonas Eppelt
- Lars Sowa
- Cedric Verstege
- Jost von den Driesch
- Christian Winter

Comparison of T1s

monit-grafana.cern.ch: cms-tier-1-utilization

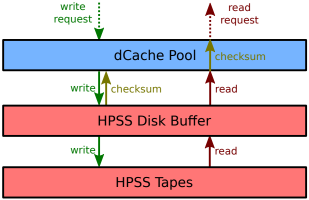
Tape Storage at GridKa

- Since March 2022
  - Larger tape capacity: 8 → 20 TB
  - Higher tape drive speed: 150 → 400 MB/s
  - 1/2 PB SSD+NVME buffer as part of the system
  - In full operation for CMS, Belle 2, and LHCb
- Data from the old system fully migrated for ATLAS, CMS, Belle 2, and LHCb
- Planning to finish migration for ALICE in summer 2024

top: tape cartridge & drive, bottom: tape library at KIT (current total capacity: 150 PB)

Schematic overview of GridKa tape system

- Write request:
  - Incoming file transfer at dCache disk pool
  - Written from dCache to HPSS disk buffer
  - Read back for checksum consistency test
  - Within HPSS, writing to tapes initiated afterwards in **file aggregates**
- Read request:
  - File read requests appear at dCache pool
  - Requests grouped by tape & aggregate
  - **Entire aggregates** read from tapes to HPSS disk buffer
  - Files read from HPSS disk to dCache pool



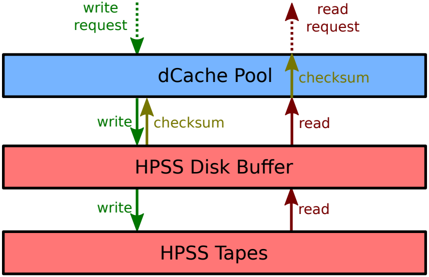
Files in the same directory collected
into aggregates of up to 300 GB

Important fraction of in-house written interface done by **ATLAS & CMS** representatives

Details on the NVMe Cache

- Current Setup:
  - AMD EPYC 9554P 64-Core
  - 300TB NVMe in XINNOR softraid (→ 500TB ssd cache in total)
  - 700k IOPS for r+w
- Optimized for low latency for many clients to
- Performance benchmark:
  - 10 or 5 times 10 Clients with each sequential reads (2/3) and writes (1/3) of a 5GB file
  - Throughput: around 50-70GB/s

  → constant 300 MB/s to 400 MB/s write speed per tape drive

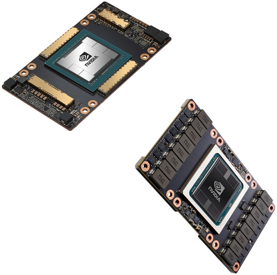


Final setup TBD (potential alternatives: GRAID or all-flash ararys)

Backup

GPUs at GridKa Made Available for CMS

- Several GPU’s deployed at GridKa TOpAS cluster and provided to entire CMS through the grid
  - 24 × Nvidia A100
  - 24 × Nvidia V100
  - 8 × Nvidia V100S
- GPU workflows sent by CMS were successfully completed
  - High Level Trigger Test Workflow
  - Release Validation Workflow



Conclusion

We are well prepared for heterogeneous computing era!