SPONSORED BY THE

Federal Ministry of Education and Research

IDIUM

Physics Institute III A

RWTH AACHEN UNIVERSITY

Status of Integration of NHR at CMS

Alexander Jung for the Grid Teams from KIT and RWTH

Annual Meeting

27th of March

Future German CMS Tier-2 Concept

- German ATLAS and CMS Tier-2 concept undergoes major adaption
- Foreseen scenario:
  - BMBF funded Tier-2 hardware at universities (age < 5 years) still part of WLCG pledge
  - Starting 2025: gradual shift from university to federated resources
    - For storage: Data Lakes at Hamburg and Karlsruhe (Helmholtz sites)
    - For CPUs: NHR centres at Aachen and Karlsruhe (and DESY continues to provide 2/3 of the total pledge)
  - Hybrid operation during fadeout

Overview CLAIX @ RWTH

- CLAIX-2018 (NHR-Tier-2 + NHR-Tier-3) currently in used production
- Reaches EOL soon

Peak performance, hosts and GPUs	**2.3 + 0.7 PFlops** • 400 + 80 MCoreHours
MPI	**1032 + 216 nodes** • 2-socket Intel Skylake processors • Platinum 8160, 2x24 cores, 192 GB, 2.1 GHz • Additional dialog and service nodes
GPUs	**48 + 6 nodes** • 2-socket Intel Skylake processors • Platinum 8160, 2x24 cores, 192 GB, 2.1 GHz • 2 NVIDIA Tesla V100, 16 GB HBM2
Fabric	**Intel OmniPath network (OPA)** • 1:2 blocking, 16X PCIgen3
Storage	• 10 PB Lustre storage • BEEOND on SSDs (480 GB)

Overview CLAIX @ RWTH

- CLAIX-2023 (NHR-Tier-2 + NHR-Tier-3) in pilot phase (in production next week)

Peak performance	• 2.6 + 1.4 PFlops (CPU) • 4.4 + 0.7 PFlops (GPU)
Available resources	• 346 + 185 MCoreHours (CPU) • 27 + 4 MCoreHours (GPU) [1 GPU-h = 24 CPU-h]
HPC segment	**412 + 220 nodes** • 2- socket Intel Sapphire Rapids • Xeon 8468, 2x48 cores, 2.1 GHz • 470 nodes with 256GB • 260 with 512GB • 2 nodes with 1024GB
ML segment	**32 + 5 nodes** • 2-socket Intel Sapphire Rapids • Xeon 8468, 2x48 cores, 2.1 GHz, 256GB • 4 NVIDIA H100, 96GB HBM2e per node
Interactive segment	• Additional nodes with smaller GPUs (e.g. for JupyterHub usage)
Fabric	• Infiniband NDR network (OPA) • 2:1 blocking
Storage	• 26PiB Lustre Storage • BEEOND on SSDs (1.4TB per node)

Overview CLAIX @ RWTH

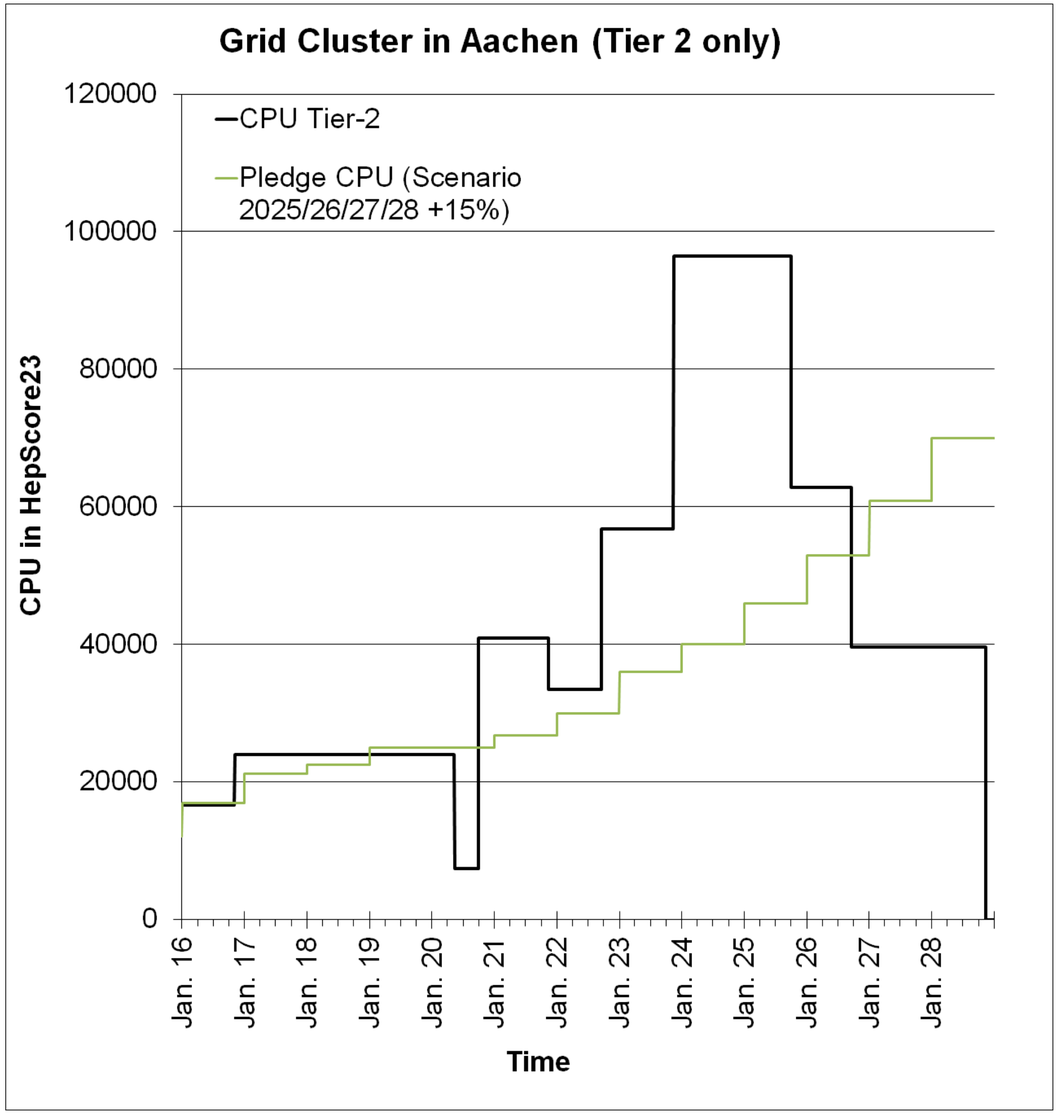
- CLAIX-2025 planned to replace CLAIX-2018 and start running in first quarter of 2026
- Projected resources for NHR: about the same order of magnitude as for CLAIX-2023
- Resources of both CLAIX-2023 and CLAIX-2025 available for us

Integration of CLAIX @ RWTH

- Running Rocky Linux 8, no problems for CMS noticeable
- At the moment only small grant of $\mathcal{O}(1)$ worker node to ensure that CMS jobs and COBalD/TARDIS are running properly, preparation for NHR
- Transparent for CMS users and operation
- Storage access to „local“ dCache and worldwide CMS storage by WAN
- Operational since more than 2 years already, already successfully "Xmas" stress tested with $\mathcal{O}(10{,}000)$ cores for about two weeks

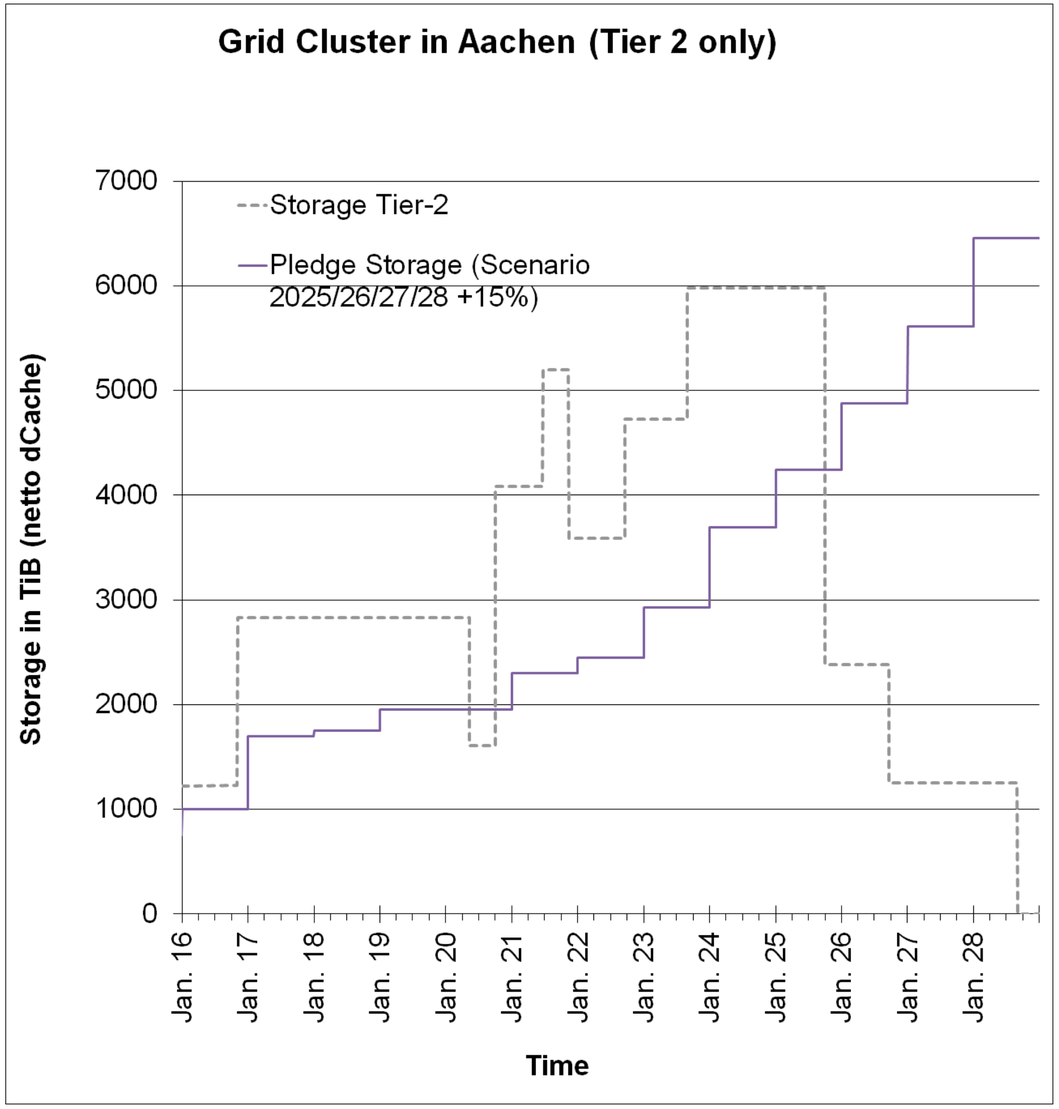
Aachen WLCG-Tier-2

- Local resources fading out in next 5 years
- Replacement and new pledges provided by NHR
- Assumed: +15%/a scenario, might be more



Aachen WLCG-Tier-2 and Access from CLAIX

- Access via dcap, XrootD (ro) and SRM (rw)
- WebDAV (rw) to local WLCG-Tier-2 and WLCG-Tier-3
- CVMFS uses Frontier Squid from local WLCG-Tier-2 and WLCG-Tier-3
- Assumed: +15%/a scenario, might be more

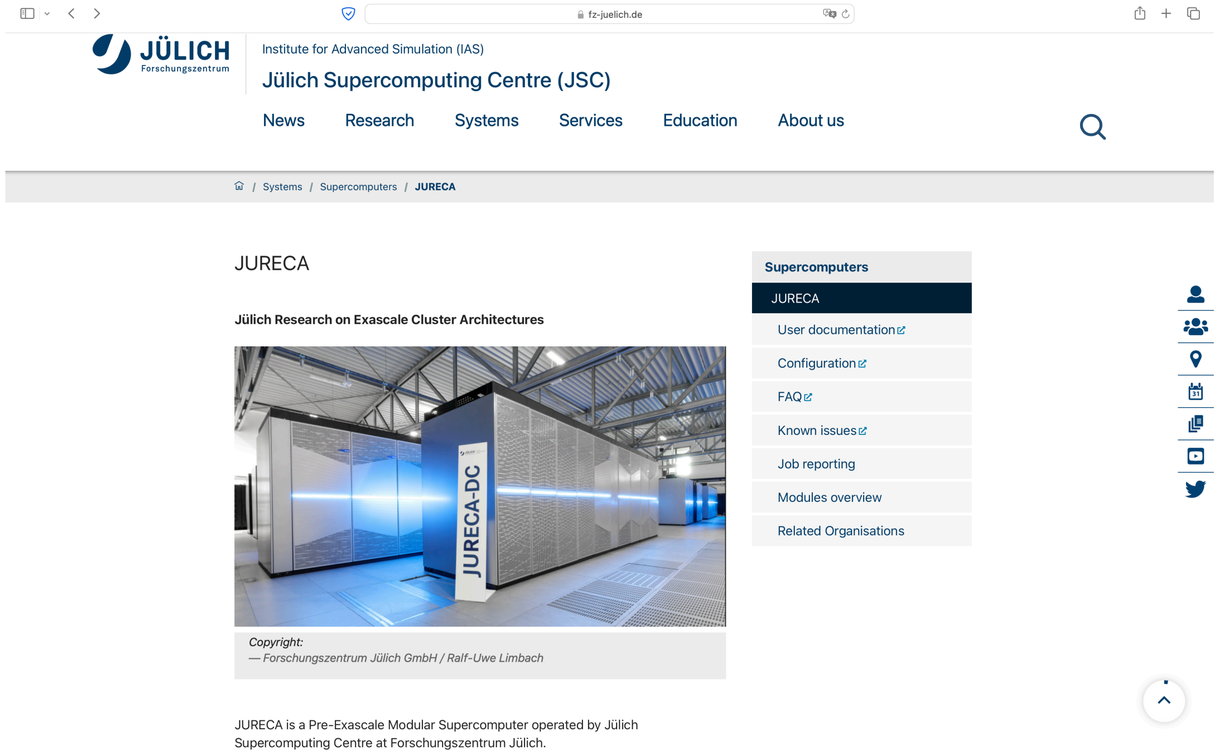


Integration of CLAIX @ RWTH

- No IPv6 available, but planned for the future
- Firewall policy compatible with our use case
- Recently mandatory activation of 2FA
  - Automated command execution only possible with static commands and command keys
  - Currently used setup will need modifications
- Urgent security risk triggered deactivation of username spaces in the past
  - COBalD/TARDIS setup not running without them
  - We had to wait for reactivation by HPC team
- CLAIX team has no access to CMS operation directly → local CMS IT experts needed

Test of JURECA @ JSC

- JSC ≠ NHR site, different project (within FIDIUM)
- Restricted firewall → non trivial setup
- As non HEP site: needed to make CVMFS available
- Similar bandwidth bottleneck as seen for HoreKa (see later slides)
- Additionally: strict firewall blocking all traffic to/from worker node from/to outside
- Ongoing connection issues between glidein and HTCondor pool
  - Reason unclear
  - No support from local IT staff
- Project on hold

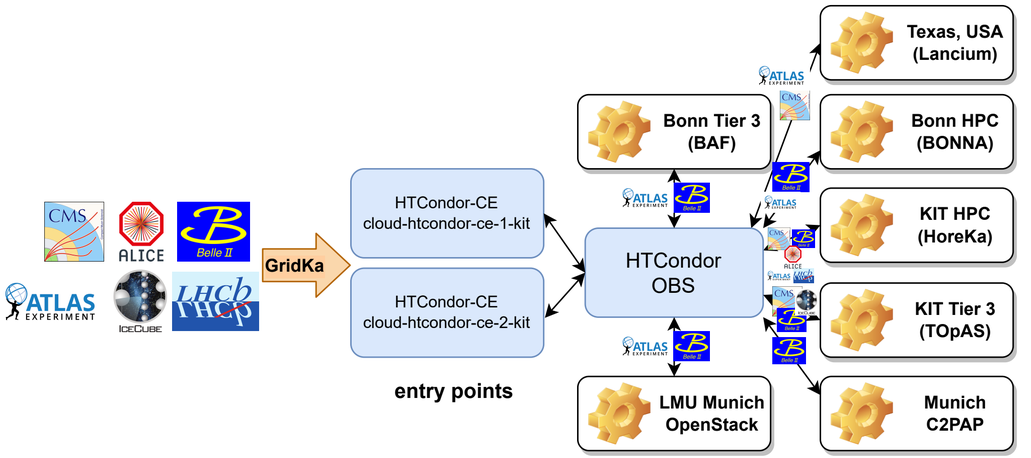


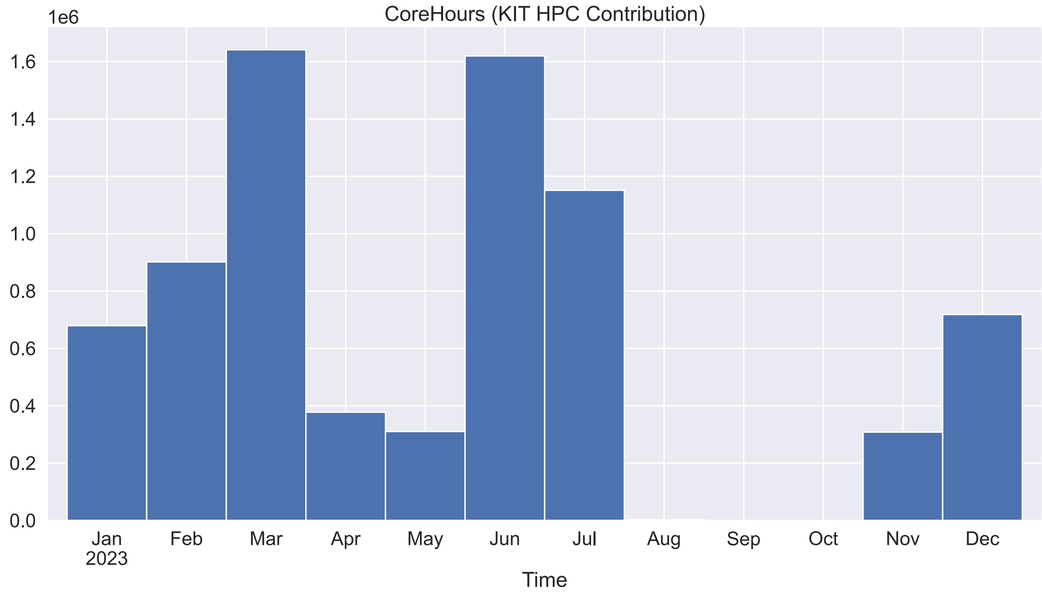
Integration of HoreKa @ KIT

- HoreKa is currently integrated as opportunistic resource into KIT's Tier-1 via COBalD/TARDIS
  - HoreKa is a midsize HPC within NHR centre
  - $60\,000$ Intel Xeon „Ice Lake“, 220 TB RAM
  - 1 Gbit/s LAN per worker node
  - 2 parallel file systems (> 15 PB)
    - Fair share: 250 TB disk

Integration of HoreKa @ KIT

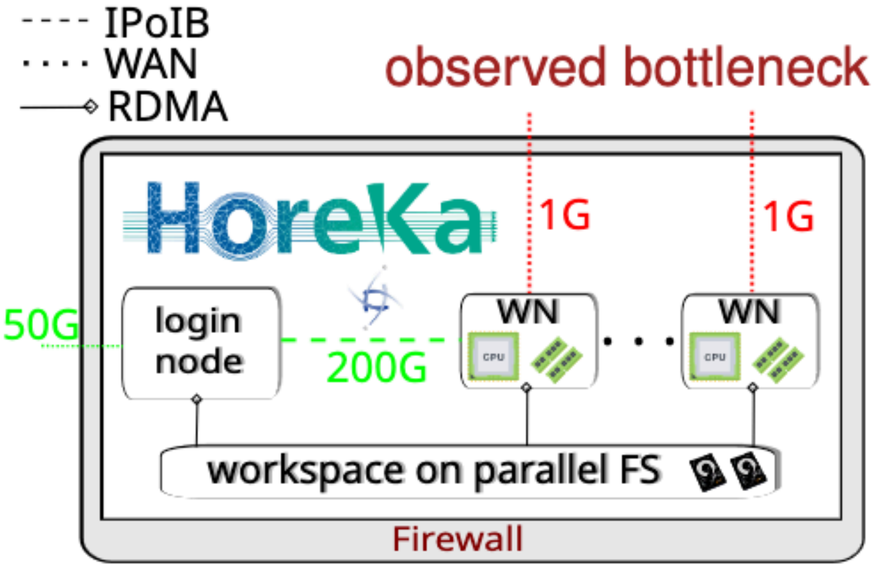
- Part of the FIDIUM Opportunistic Compute Cloud operated at GridKa
  - Dynamic, transparent and on-demand integration via COBalD/TARDIS (developed by KIT)
  - Provide community-overarching unified entry points to a variety of resources (HPCs, Clouds, …)
  - Demonstrated production scale operation during scale test together with HoreKa (KIT HPC cluster)
  - HoreKa provided in 2023 about 7.7 MCoreHours to the CMS experiment
- Similar setup deployed at CLAIX HPC (RWTH) and ongoing deployment at Emmy (Göttingen)





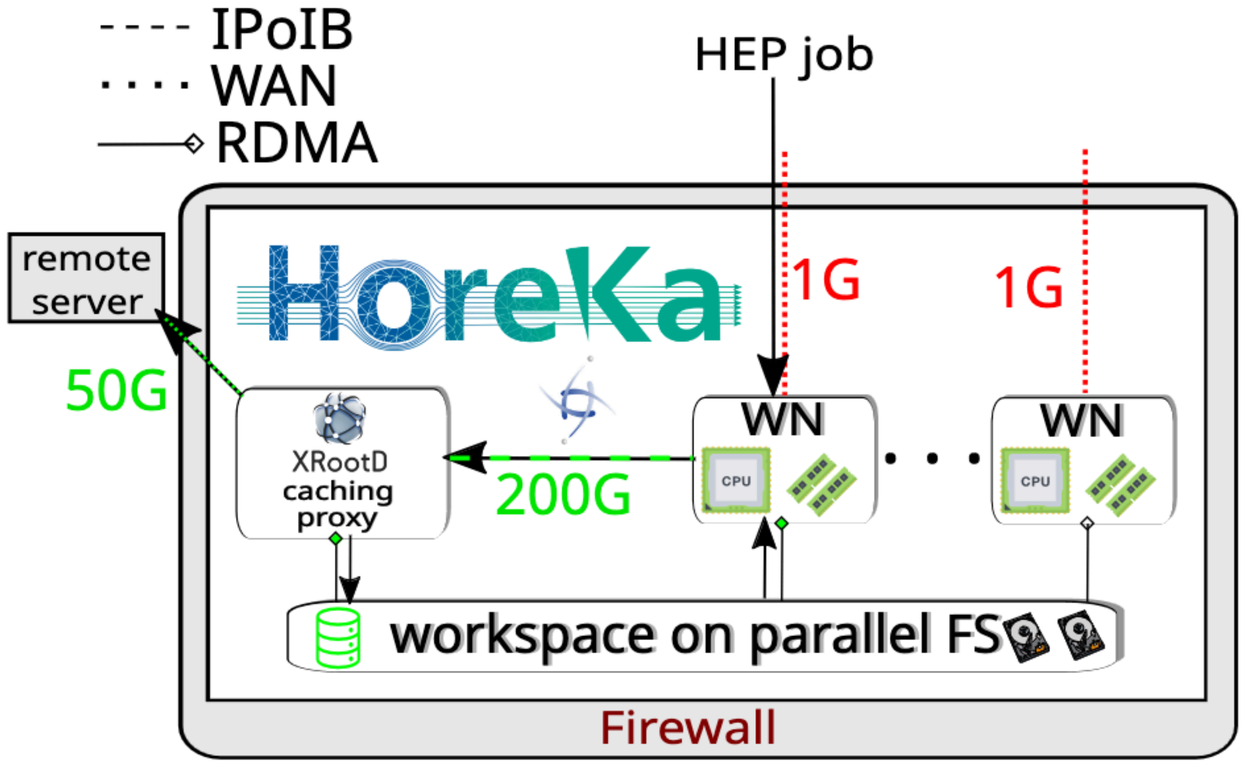
Prototype: XrootD Buffer @ HoreKa

- The external 1 Gbit/s bandwidth is a bottleneck for data intensive jobs
- XrootD based approach to mitigate the effects of the slow WAN connection
  - Caching proxy on login node that prefetches data with 50 Gbit/s (like buffering)
  - Using share on parallel FS as (volatile) cache
- Advantages
  - Increasing bandwidth for external transfers
  - Cache hits accelerate transfers extremely
- Running Proof of Concept on HoreKa



Prototype: XrootD Buffer @ HoreKa

- Idea: run XrootD caching proxy as buffer on the login node
- Benefit from internal IP over IB connection (200 Gbit/s)
- Benefit from higher external bandwidth of the login node (50 Gbit/s)
- Directly access fully cashed files on the parallel file system (RDMA)
- Dedicated data transfer node within the LHCONE network planned for the future (200 Gbit/s external bandwidth)



CMS Jobs Triggered RHEL 8.6 Kernel Bug

- KIT had to shutdown the HoreKa integration on 1st of March 2024
- CMS jobs somehow triggered a kernel bug in RHEL 8.6
- Nodes were stuck and needed a reboot
- Will potentially only be solved during site maintenance in April 2024
- NHR integrations need dedicated HEP personal to take care

RHEL 8.8/8.6(EUS): hung_task_timeout_secs at migration_entry_wait_on_locked

SOLUTION VERIFIED - Updated March 6 2024 at 8:51 AM - English

Issue

- After upgrading to RHEL 8.8, few commands are going to hung state and system load average showing very high.
- Logs shows several hung_task_timeout_secs with `migration_entry_wait_on_locked()` in backtrace.

Raw

```
INFO: task task1:1618 blocked for more than 120 seconds.
      Not tainted 4.18.0-477.10.1.el8_8.x86_64 #1
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
task:task1     state:D stack:    0 pid: 1618 ppid:     1 flags:0x00000080
Call Trace:
 __schedule+0x2d1/0x870
 schedule+0x55/0xf0
 io_schedule+0x12/0x40
 migration_entry_wait_on_locked+0x1ea/0x290
 do_swap_page+0x5b0/0x710
 __handle_mm_fault+0x453/0x6c0
 handle_mm_fault+0xca/0x2a0
 __do_page_fault+0x1f0/0x450
 do_page_fault+0x37/0x130
 page_fault+0x1e/0x30
```

- After upgrading to RHEL 8.6 EUS `kernel-4.18.0-372.91.1.el8_6`, few commands are going to hung state and system load average showing very high.
- Logs shows several hung_task_timeout_secs with `migration_entry_wait_on_locked()` in backtrace.

IPv6 and CMS Pilot Factories

- Discovered CMS pilots that are stuck roughly 1 h into their initialisation phase
  - In the Tier-1 it took only seconds
  - Only $1/3$ of the pilots were affected
- It turns out that the Fermilab factory did not respond on IPv6
- Different behaviour between LHCONE and non LHCONE networks
- NHR integrations need dedicated HEP staff to take car

```
[root@c01-013-166 ~]# time curl -4 http://cmssi-factory02.fnal.gov:8319/monitor/ 1> /dev/null
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  7003  100  7003    0     0  33586      0 --:--:-- --:--:-- --:--:-- 33668

real    0m0.214s
user    0m0.004s
sys     0m0.004s
[root@c01-013-166 ~]# time curl -6 http://cmssi-factory02.fnal.gov:8319/monitor/ 1> /dev/null
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0curl: (7)
Failed connect to cmssi-factory02.fnal.gov:8319; Connection refused

real    0m0.124s
user    0m0.003s
sys     0m0.006s
```

on the HPC worker I got:

```
[scc-sdm-hep-0001@hkn0825 ~]$ time curl -4 http://cmssi-factory02.fnal.gov:8319/monitor/ 1> /dev/
null
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  7003  100  7003    0     0  33668      0 --:--:-- --:--:-- --:--:-- 33668

real    0m0,215s
user    0m0,004s
sys     0m0,005s
[scc-sdm-hep-0001@hkn0825 ~]$ time curl -6 http://cmssi-factory02.fnal.gov:8319/monitor/ 1> /dev/
null
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:--  0:02:11 --:--:--     0curl: (7)
Failed to connect to cmssi-factory02.fnal.gov port 8319: Die Wartezeit für die Verbindung
ist abgelaufen

real    2m12,275s
user    0m0,008s
sys     0m0,015s
```

Future Access Patterns

- ATM: 2 mainly bilateral setups
  - HoreKa ↔ Grid Karlsruhe
  - CLAIX ↔ Grid Aachen
- Extension necessary to include DESY Tier-2 and Helmholtz storage (DESY and KIT)
- Full concept should be developed in the forthcoming months
- Afterwards: practice tests of storage access patterns within FIDIUM
- Could NHR CMS interoperate with NHR ATLAS?

Summary

- NHR resources tested in CMS context at CLAIX and HoreKa
- Pilot setup using COBalD/TARDIS (developed by KIT) very successfully
- First stress tests done
- Technical problems can occur at any time → HEP experts needed on site

Thanks to Manuel Giffels (KIT), Robin Hofsaess (KIT) and Tim Cramer (RWTH) for helping with the content for this talk.