Karlsruhe Institute of Technology

Karlsruhe Institute of Technology

HEPiX

Closer to IPv6-only on WLCG

Closer to IPv6-only on WLCG

Bruno Hoeft[1], David Kelsey[3], Edoardo Martelli[4], Carmen Misa Moreira[4]
ISGC 2024, Taipei, 27 March 2024

1 - Karlsruhe Institute of Technology (KIT), Germany
3 - UKRI Science and Technology Facilities Council (STFC UK), United Kingdom
4 - Conseil Européen pour la Recherche Nucléaire (CERN), Switzerland

On behalf of all co-authors in the HEPiX IPv6 working group

Active in HEPiX IPv6 Working Group – last 12 months

- M Babik (CERN), M Bly (RAL), N Buraglio (ESnet), T Chown (Jisc), D Christidis (U Texas/ATLAS), J Chudoba (FZU Prague), P Demar (FNAL), J Flix (PIC), C Grigoras (CERN/ALICE), B Hoeft (KIT), H Ito (BNL), D P Kelsey (RAL), E Martelli (CERN), S McKee (U Michigan), C Misa Moreira (CERN), R Nandakumar (RAL/LHCb), K Ohrenberg (DESY), F Prelz (INFN), D Rand (Imperial), A Sciabà (CERN/CMS), C Walker (Jisc)

- many more in the past, and others join from time to time
- *and thanks also to WLCG operations, WLCG sites, LHC experiments, networking teams, monitoring groups, storage developers…*

Outline

- The HEPiX IPv6 working group
- WLCG Data Challenge 2024 (DC24)
- Special connection between DE-KIT and CERN for DC24
- IPv6/IPv4 traffic analysis of the Tier-0 to Tier-1 LHCOPN
- DC24 results
- Tier-1/2 Dual-Stack WorkerNodes GGus ticket campaign
- Migration towards IPv6 at DE-KIT
- Summary

HEPiX-IPv6 Working group

working group
-- created in April 2011

Phase 1 to consider whether and how IPv6 should be deployed in HEP (especially for WLCG). Readiness and gap analysis of

- HEP applications,
- middleware,
- security issues,
- system management and monitoring tools and
- end to end network monitoring

Phase 2 → dual-stack deployment:

- dual-stack at fileserver (at Tier-1 – realized end of 2020 / at Tier-2 – realized at end of 2023)
- a GGUS-Ticket campaign at end of 2023 → dual-stack WorkerNodes incl. middleware (@Tier-1/2)

Phase 3 : move towards single protocol
→ decommission of IPv4 and migrate to IPv6 only

WLCG Data Challenge 2024 (DC24)

DC24

- LHC-HighLuminosity (HL) run – starting 2029
- 25% load pre-run at DC24
- Organizing off all scheduled activities
  - CERN (Tier-0) to Tier-1 and Tier-1 to Tier-2 traffic flow
  - of the four LHC experiments (Alice, Atlas, CMS, LHCB)
  - including Belle-2 and Dune
- Minimal Scenario in 2029
  4810 Gbps
- Flexible Scenario in 2029
  9620 Gbps

T1	LHC Network Needs (Gbps) Flexible Scenario in 2029	Data Challenge target 2024 (Gbps)
CA-TRIUMF	400	100
DE-KIT	1200	300
ES-PIC	400	100
FR-CCIN2P3	1140	290
IT-INFN-CNAF	1380	350
KR-KISTI-GSDC	100	30
NDGF	280	70
NL-T1	360	90
NRC-KI-T1	240	60
UK-T1-RAL	1220	310
RU-JINR-T1	400	100
US-T1-BNL	900	230
US-FNAL-CMS	1600	400
(atlantic link)	2500	630
Sum	9620	2430

DC24 February 12 - 23

Proposal to distribute different exercises (independently of the experiment exercises) over the challenge days, e.g.

Day 1-3: T0 export
Day 4-5: Reprocessing like traffic
Day 6,7 (weekend): Keep things running…
Day 8-9: MC like traffic
Day 10-11: Increase to flexible scenario + experimental traffic flow (NOTED)
Day 12: Repeat things e.g. with adjusted setting
(Day 13-14 (tentative weekend): Hope that nothing completely broken)

URL: https://indico.cern.ch/event/1307338/contributions/5657133/attachments/2749065/4784409/DC24%20Welcome%20Logistics%20Introduction.pdf

Experiment plans at DC24

special Exeriments plans

- CMS → 250Gbps T0 to T1 (to DE-KIT 60Gbps) +…
- Atlas → 270Gbps T0 to T1 (to DE-KIT 61Gbps) + …

  detailed table with expected transfer rates of each site

HEPiX-IPv6 → IPv6 only Networking

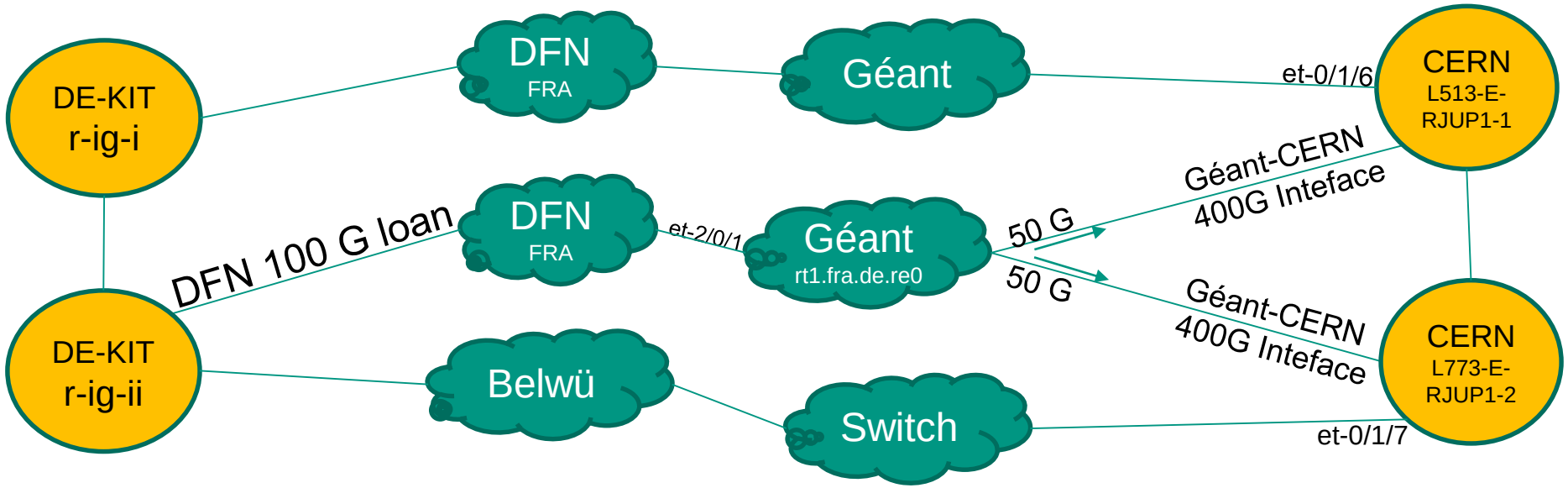
IPv4 → finding of leftovers at LHCOPN

URL: Inspect remaining IPv4 dataflows LHCOPN between T0 (CERN) and T1 (DE-KIT)

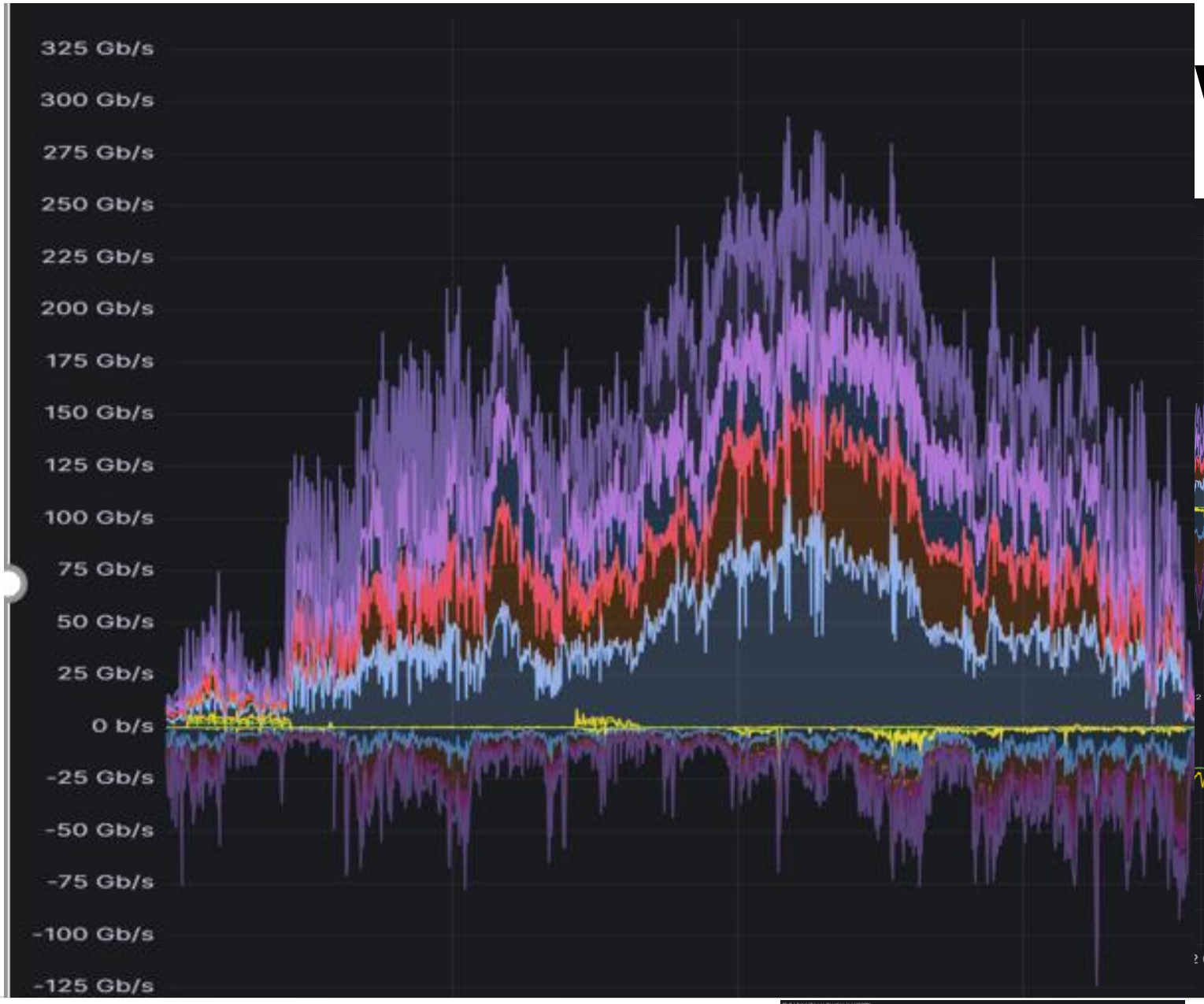
https://indico.cern.ch/event/1307338/contributions/5663197/attachments/2749306/4785227/DC24-IPv6-project-1.pdf

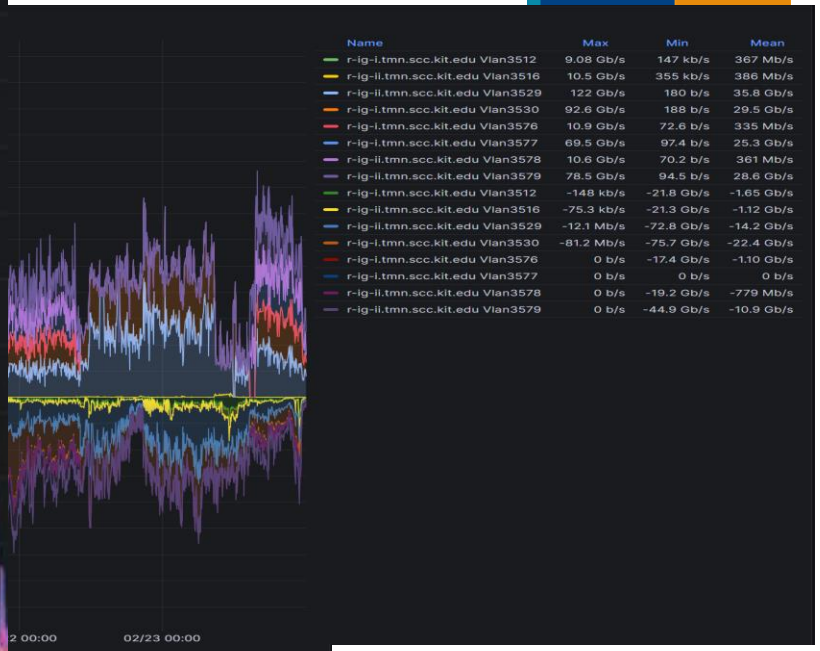
DE-KIT (3+100G) LHCOPN only for DC24

LHCOPN – Connecting DE-KIT to CERN → 2 x 100 G
Additional 100 G LHCOPN DE-KIT to CERN
DFN loan of 100 G DE-KIT to DFN PoP Frankfurt incl. link to Géant Router in Frankfurt
Géant Frankfurt to Géant Geneva
Splitting the 1 x 100 G to 2 x 50 G as overlay to the 2x400 G CERN/Géant interfaces



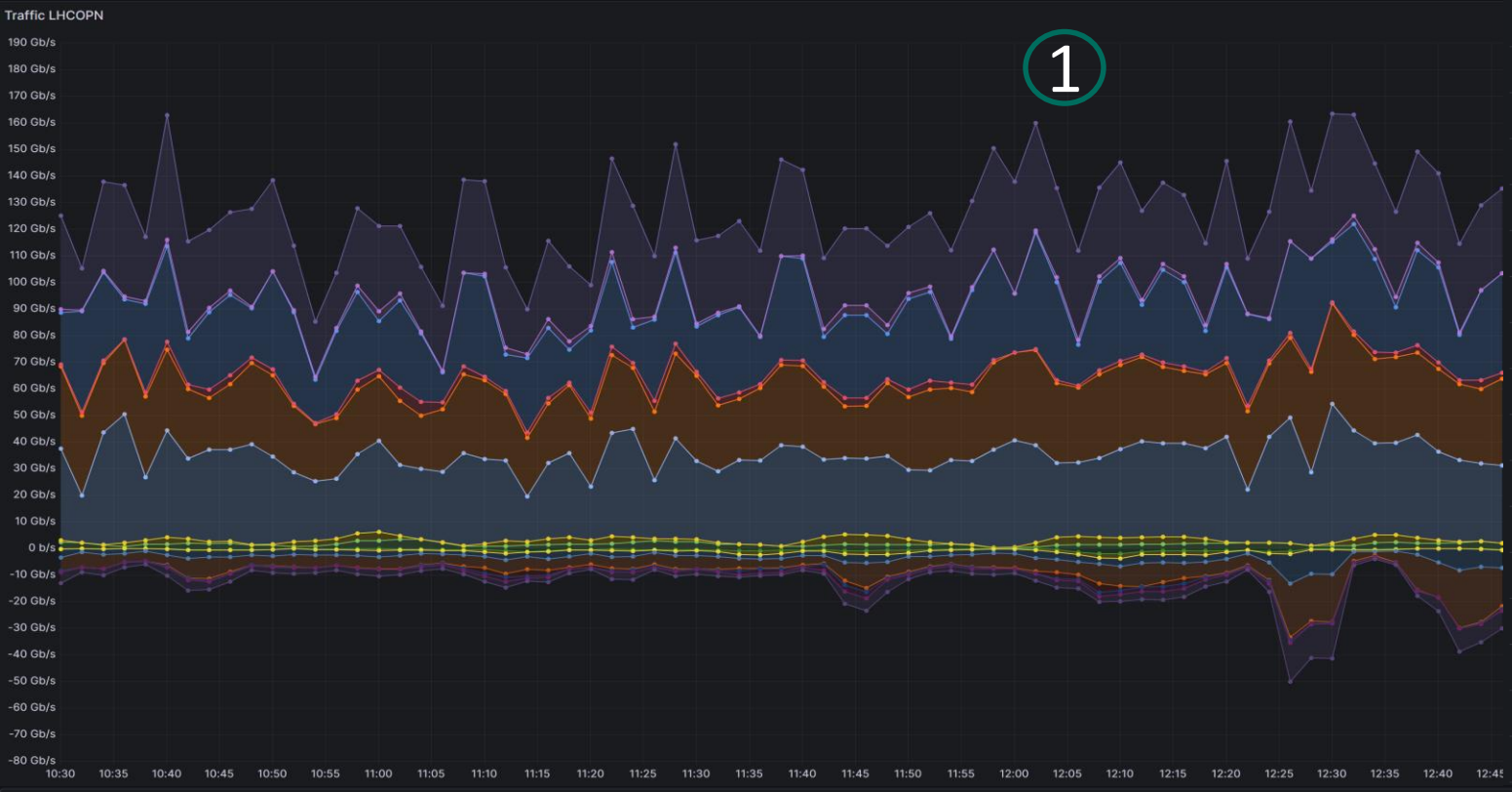
weeks





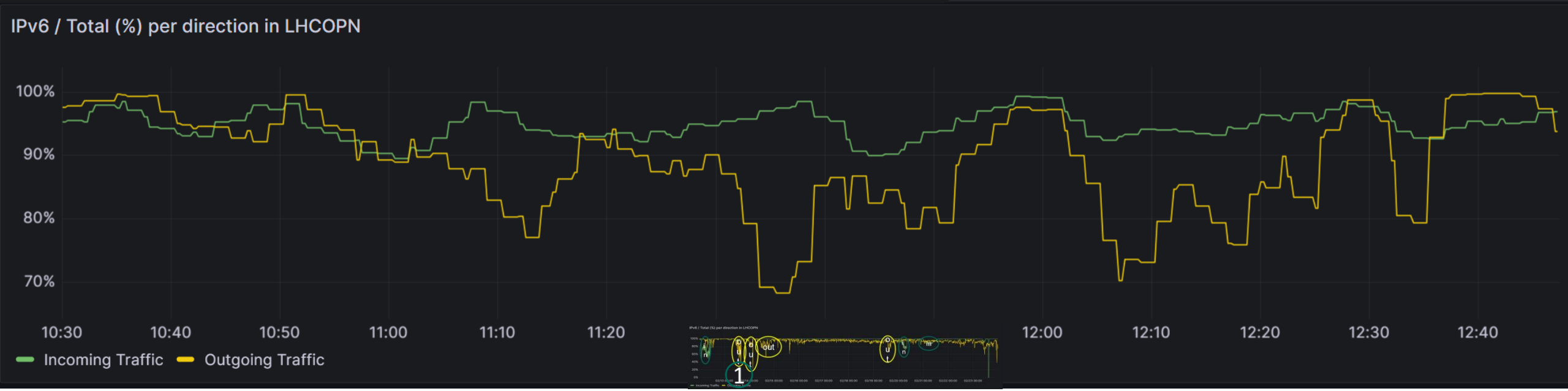
Name	Max	Min	Mean
r-ig-i.tmn.scc.kit.edu Vlan3512	9.08 Gb/s	147 kb/s	367 Mb/s
r-ig-ii.tmn.scc.kit.edu Vlan3516	10.5 Gb/s	355 kb/s	386 Mb/s
r-ig-ii.tmn.scc.kit.edu Vlan3529	122 Gb/s	180 b/s	35.8 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3530	92.6 Gb/s	188 b/s	29.5 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3576	10.9 Gb/s	72.6 b/s	335 Mb/s
r-ig-i.tmn.scc.kit.edu Vlan3577	69.5 Gb/s	97.4 b/s	25.3 Gb/s
r-ig-ii.tmn.scc.kit.edu Vlan3578	10.6 Gb/s	70.2 b/s	361 Mb/s
r-ig-ii.tmn.scc.kit.edu Vlan3579	78.5 Gb/s	94.5 b/s	28.6 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3512	-148 kb/s	-21.8 Gb/s	-1.65 Gb/s
r-ig-ii.tmn.scc.kit.edu Vlan3516	-75.3 kb/s	-21.3 Gb/s	-1.12 Gb/s
r-ig-ii.tmn.scc.kit.edu Vlan3529	-12.1 Mb/s	-72.8 Gb/s	-14.2 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3530	-81.2 Mb/s	-75.7 Gb/s	-22.4 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3576	0 b/s	-17.4 Gb/s	-1.10 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3577	0 b/s	0 b/s	0 b/s
r-ig-ii.tmn.scc.kit.edu Vlan3578	0 b/s	-19.2 Gb/s	-779 Mb/s
r-ig-ii.tmn.scc.kit.edu Vlan3579	0 b/s	-44.9 Gb/s	-10.9 Gb/s

DC24 graph over two weeks



Name	Max	Min	Mean
r-ig-i.tmn.scc.kit.edu Vlan3512	3.58 Gb/s	426 Mb/s	1.72 Gb/s
r-ig-ii.tmn.scc.kit.edu Vlan3516	3.68 Gb/s	7.33 Mb/s	1.42 Gb/s
r-ig-ii.tmn.scc.kit.edu Vlan3529	52.5 Gb/s	17.1 Gb/s	32.2 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3530	37.8 Gb/s	19.4 Gb/s	28.0 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3576	5.23 Gb/s	35.0 Mb/s	1.95 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3577	43.8 Gb/s	11.2 Gb/s	28.4 Gb/s
r-ig-ii.tmn.scc.kit.edu Vlan3578	3.85 Gb/s	6.46 Mb/s	1.58 Gb/s
r-ig-ii.tmn.scc.kit.edu Vlan3579	47.3 Gb/s	15.3 Gb/s	31.6 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3512	-5.09 Mb/s	-1.92 Gb/s	-412 Mb/s
r-ig-ii.tmn.scc.kit.edu Vlan3516	-4.01 Mb/s	-1.87 Gb/s	-569 Mb/s
r-ig-ii.tmn.scc.kit.edu Vlan3529	-585 Mb/s	-11.1 Gb/s	-2.66 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3530	-1.41 Gb/s	-21.8 Gb/s	-6.35 Gb/s
r-ig-i.tmn.scc.kit.edu Vlan3576	-1.88 Mb/s	-3.19 Gb/s	-669 Mb/s
r-ig-i.tmn.scc.kit.edu Vlan3577	0 b/s	0 b/s	0 b/s
r-ig-ii.tmn.scc.kit.edu Vlan3578	-4.92 Mb/s	-2.41 Gb/s	-613 Mb/s
r-ig-ii.tmn.scc.kit.edu Vlan3579	-518 Mb/s	-14.8 Gb/s	-2.89 Gb/s

IPv4
IPv4
IPv4
IPv4
IPv4
IPv4



DC24 first day ①

- shown graph are network monitoring graphs with dedicated SVI interface for IPv4 and IPv6
- further analysis with netflow (nfsen)
  - source/destination host and port

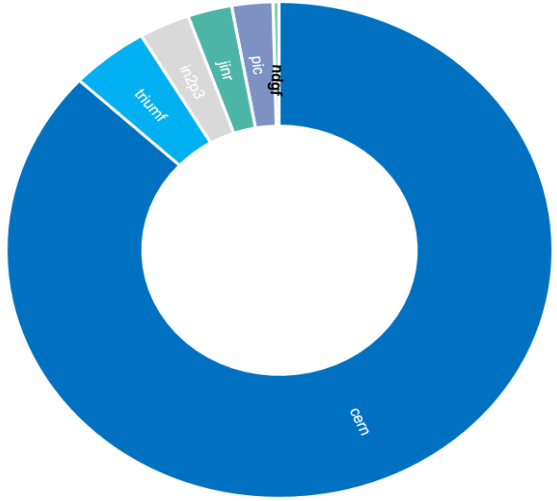
Outgoing traffic during first DC24 day

IPv4 traffic was mainly to CERN – 3,0 TeraByte (data transfer to aliendb (experiment Alice)) and file server (failed transfers are falling back from IPv6 to IPv4 (e.g. Alice))

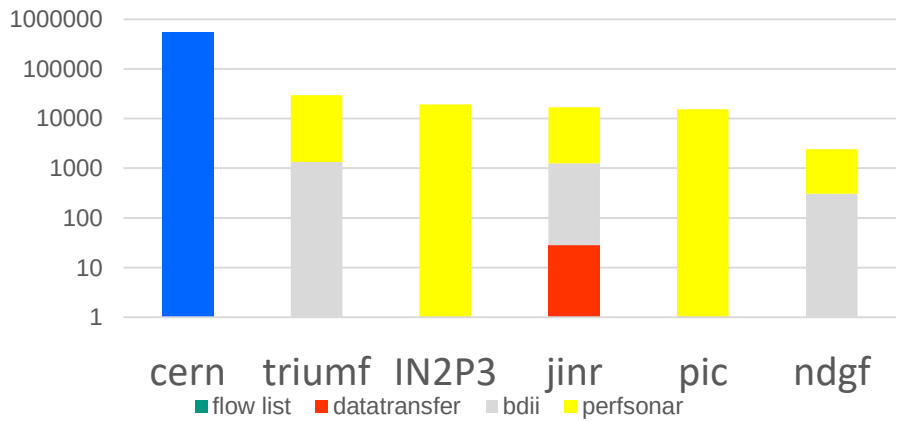
IPv4 flows to other sites are mainly data exchanged between perfsonar server

→ except a little dataflow to jinr and bdii flows

→ bdii at DE-KIT IPv4 only (bdii shall be decommisioned)

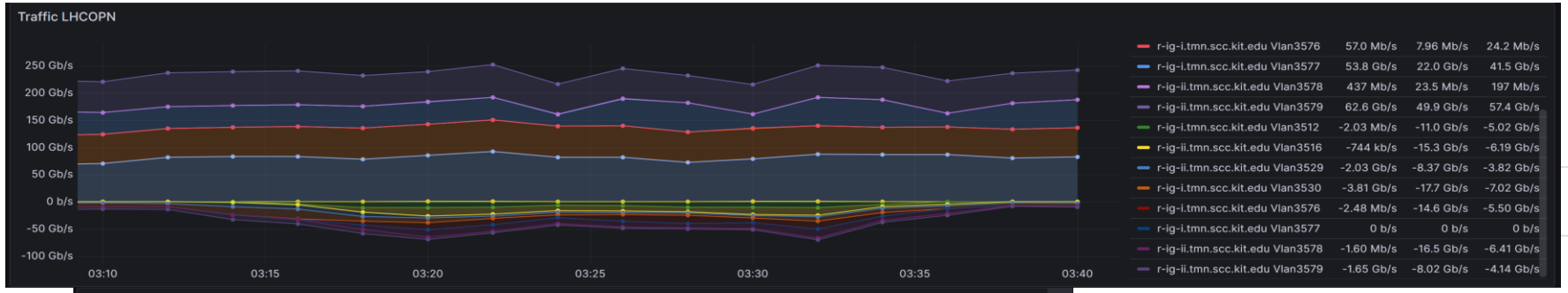


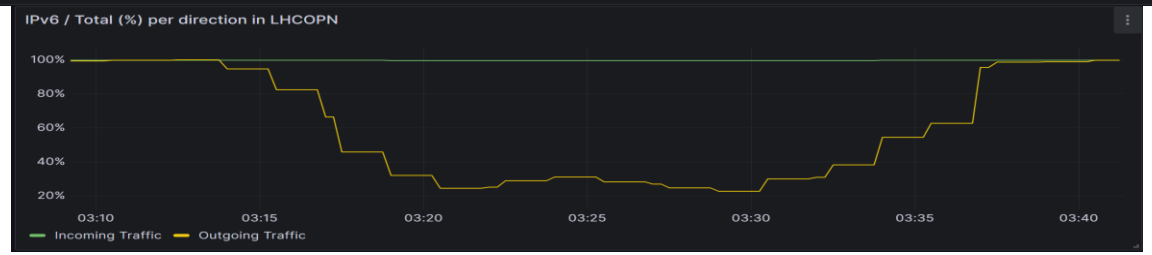
	datatransfer	perfsonar	bdii	
cern	556875			3,0 TeraByte -- aliendb
triumf		28113	1341	neglectible volume
in2p3		19158		neglectible volume
jinr	27	15510	1221	neglectible volume
pic		15534		neglectible volume
ndgf		2121	308	neglectible volume

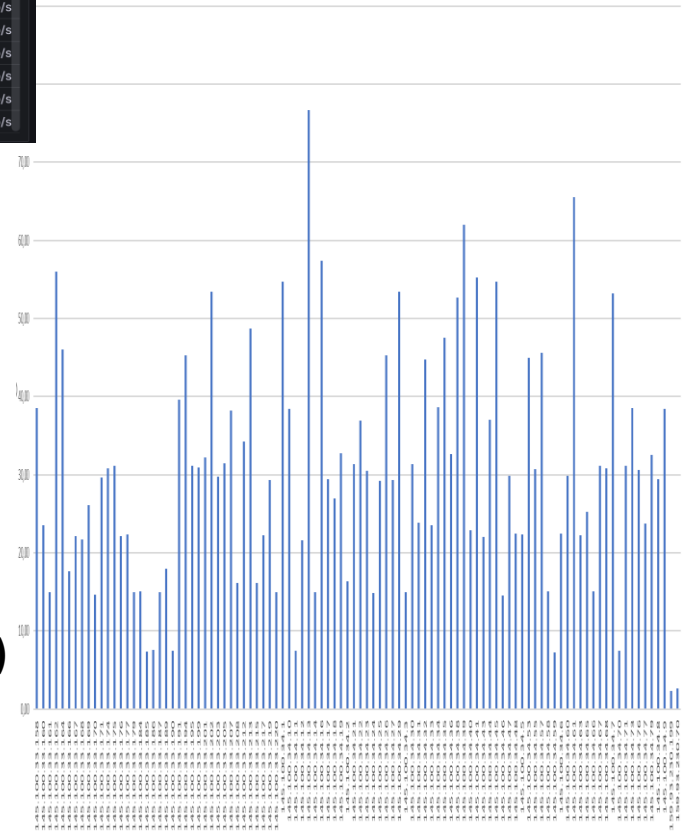


server at CERN :
atlasfrontier09.cern.ch.
front-lcg0[1234].cern.ch.
p06636710n27239.cern.ch.
p06636710e56718.cern.ch.
p06636710x02074.cern.ch.
p06636710g94481.cern.ch.
p06636710d42351.cern.ch.
p06636710t37894.cern.ch.
p06636710r35645.cern.ch.
p06636710e11348.cern.ch.
aliendb[45].cern.ch.
aliendb[78].cern.ch.
aliendb10.cern.ch.
pcalimonitor.cern.ch.
pcaliendb06a.cern.ch.
st-048-bb35f9c9.cern.ch.
st-096-hh1510b1.cern.ch.
st-192-e1bf9e17.cern.ch.
st-192-c6995aa0.cern.ch.
vocms0334.cern.ch.

3,0TeraByte Data ② DE-KIT to NLT1







Summary: total flows: 9634, total bytes: 3.0 TeraByte
→ At NLT1 98 different hosts involved

Reason:

NLT1 turned to accept IPv6 connections, but initiate IPv4 (to reduce amount of IPv6 connections)

- **instability reason with IPv6 connections**
  - **neighbour discovery (nd) not reliable**
  - **memory overflow with many concurrent IPv6 connections**

③

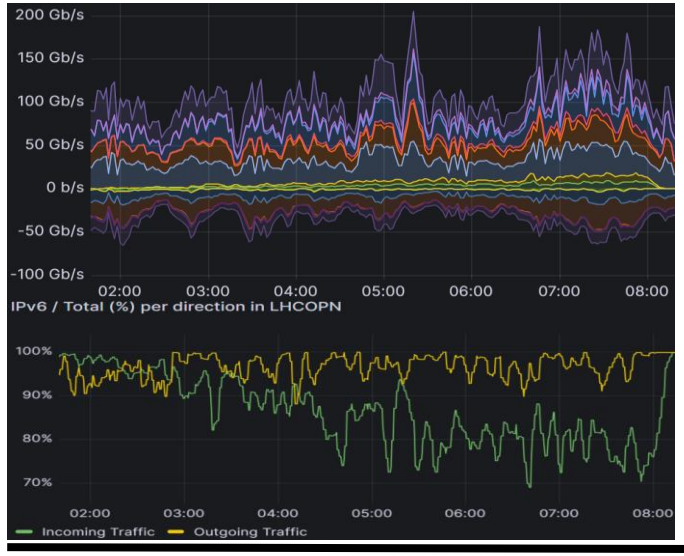
2,1TeraByte Data DE-KIT to NLT1

Summary: total flows: 536, total bytes: 2.1 TeraByte
2 file server at DE-KIT transfer data to 98 file server at NLT1, all hosts at DE-KIT and NLT1 dual-stack deployed

* 64GByte
transfered to CERN, kiae, BNL
and a small leftover to other sites (manly perfsonar)

- **BNL hosts (IPv4 only) (Experiment Belle2):**
  - 192.33.128.21 bldiracvm02.sdcc.bnl.gov.
  - 192.33.128.23 bldiracvm04.sdcc.bnl.gov.
  - 192.33.128.25 bldiracvm05.sdcc.bnl.gov.
  - 192.33.128.41 blrucio01-vrrp.sdcc.bnl.gov.
  - 192.33.128.42 blrucio02-vrrp.sdcc.bnl.gov.
  - 192.33.128.44 blcond03.sdcc.bnl.gov.
  - 192.33.128.45 blcond04.sdcc.bnl.gov.
  - 192.33.128.49 blcond06.sdcc.bnl.gov.
  - 192.33.128.98 blrucio03.sdcc.bnl.gov.
  - 192.33.128.99 blrucio04.sdcc.bnl.gov.

XRootD file transfer from CERN



4

```
2024-02-20 06:02:38.012  16.000 TCP  128.142.57.111 40594  192.108.47.89  1094  2.7 M  4.1 G  2.1 G  1499  1
2024-01-31 09:33:31.833  11.653 TCP  128.142.63.105 43670  192.108.46.89  1094  2.8 M  4.2 G  2.9 G  1498  1
```

Summary: total flows: 597053, total bytes: 33.0 TeraByte

1625 Server at CERN

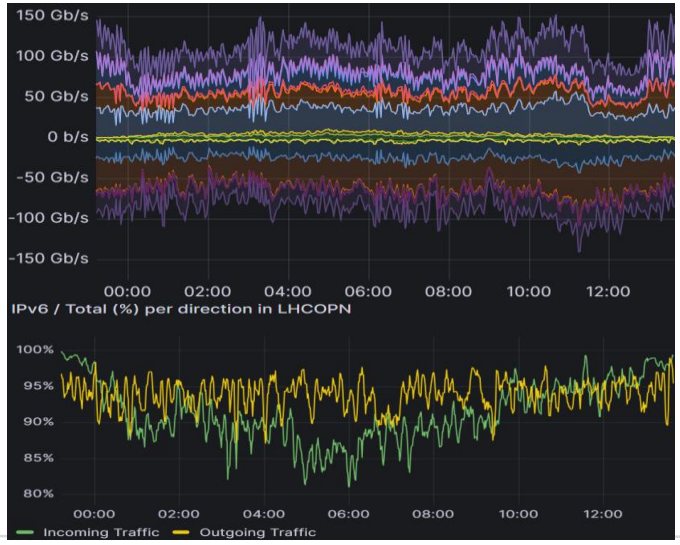
Only 16 Server at KIT

Squid service Port 3401

- cvmfs-sq4.gridka.de. dual-stack
- cvmfs-sq1.gridka.de. dual-stack
- cvmfs-sq3.gridka.de. dual-stack
- cvmfs-sq5.gridka.de. dual-stack
- cvmfs-sq6.gridka.de. dual-stack
- cvmfs-sq2.gridka.de. dual-stack
- frontier-sq1.gridka.de. dual-stack
- fw-nat-inside-outside.gridka.de.

XRootD Port 1094

- f01-032-114-e.gridka.de.
- f01-124-110-e.gridka.de. dual-stack
- f01-124-159-e.gridka.de. dual-stack
- f01-124-160-e.gridka.de. dual-stack
- f01-124-161-e.gridka.de. dual-stack
- f01-125-159-e.gridka.de. dual-stack
- f01-125-160-e.gridka.de. dual-stack
- f01-125-161-e.gridka.de. dual-stack



5

```
2024-02-20 23:26:09.012  0.500 TCP  128.142.249.74 38908  192.108.68.144 1094   12   2266  36256  188  1
2024-02-21 02:39:33.262  0.250 TCP  128.142.240.76 55700  192.108.46.89  1094   10   706   22592  70   1
```

Summary: total flows: 1460049, total bytes: 43.1 TeraByte

2426 Server at CERN

Only 25 Server at KIT

Squid service Port 3401

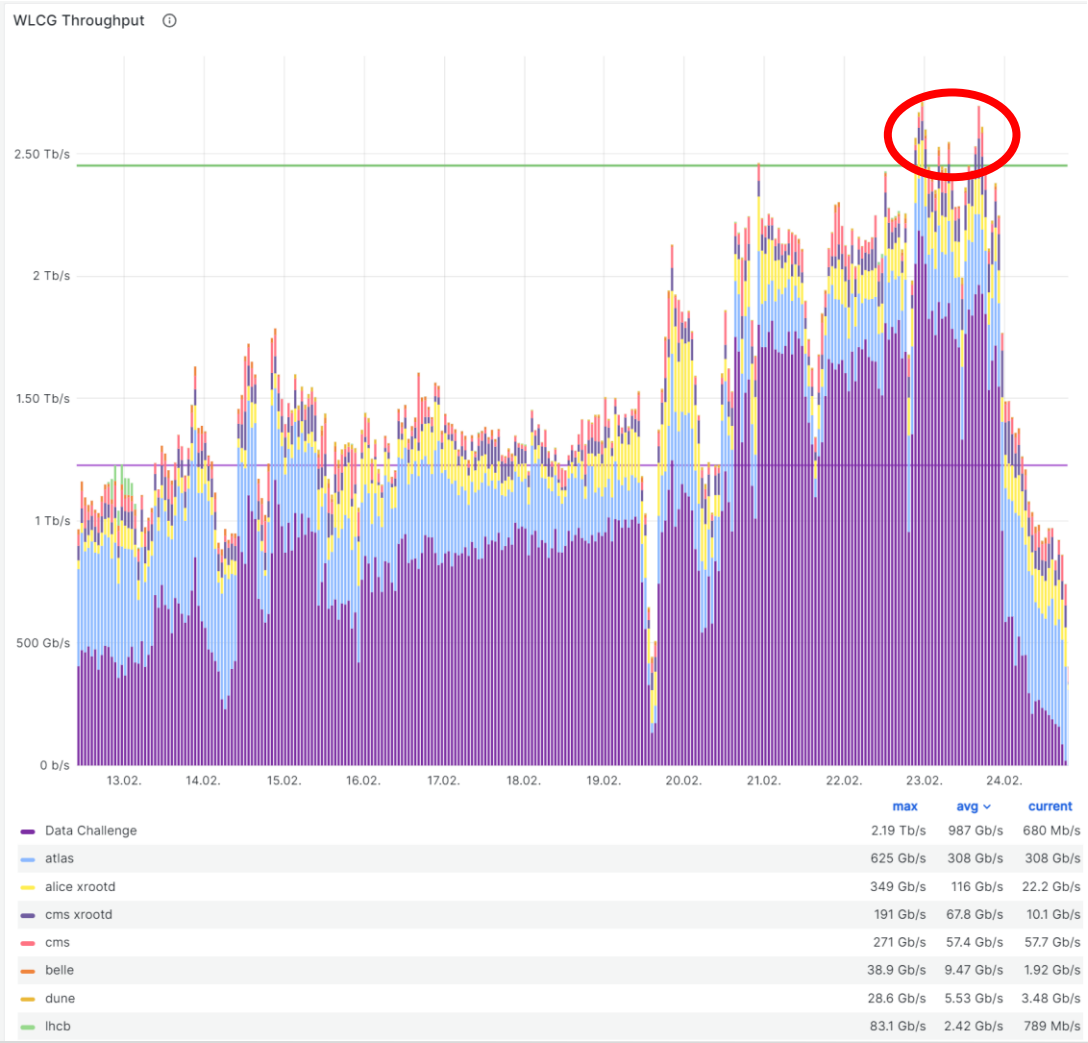
cvmfs-sq4.gridka.de.
cvmfs-sq1.gridka.de.
cvmfs-sq3.gridka.de.
cvmfs-sq5.gridka.de.
cvmfs-sq6.gridka.de.
cvmfs-sq2.gridka.de.
frontier-sq1.gridka.de.
fw-nat-inside-outside.gridka.de.

XRootD Port 1094

f01-124-109-e.gridka.de.
f01-124-110-e.gridka.de.
f01-124-112-e.gridka.de.
f01-124-155-e.gridka.de.
f01-124-159-e.gridka.de.
f01-124-160-e.gridka.de.
f01-124-161-e.gridka.de.
f01-125-109-e.gridka.de.
f01-125-110-e.gridka.de.
f01-125-155-e.gridka.de.
f01-125-159-e.gridka.de.
f01-125-160-e.gridka.de.
f01-125-161-e.gridka.de.
f01-117-137-e.gridka.de.
f01-152-140-e.gridka.de.
f01-152-191-e.gridka.de.
f01-152-192-e.gridka.de.

token

Some DC24 results



- Aggreagated WLCG Traffic over 48 hours above 2.5Tbit/sec
  - not realised, but the traffic rate itself where reached (25% of LHC-HL)
- Source/outbound performance of DE-KIT was good (1.033)
- Destination/inbound performance was a little less (0.897), there were sites identified with more open points to solve as DE-KIT, DC24 brought issues to light
  - the week bevor DC24 DE-KIT could achive a sustained inbound rate of 300 Gbps,
    but during DC24 the following issu was hitting us:
    - problems with HTTP WebDAV third-party transfers
- „file deletion process“ initiated by FTS only was to slow – multiple deletion proccesses were required
- Change from x.509 to token was not as smooth as envisioned
  → livetime of tokens too short (running out before / or during execution)

Tier-1/2 Dual-Stack WorkerNodes GGus ticket campaign

Tier-1/2 WN and CE @ Dual-Stack

Motivations

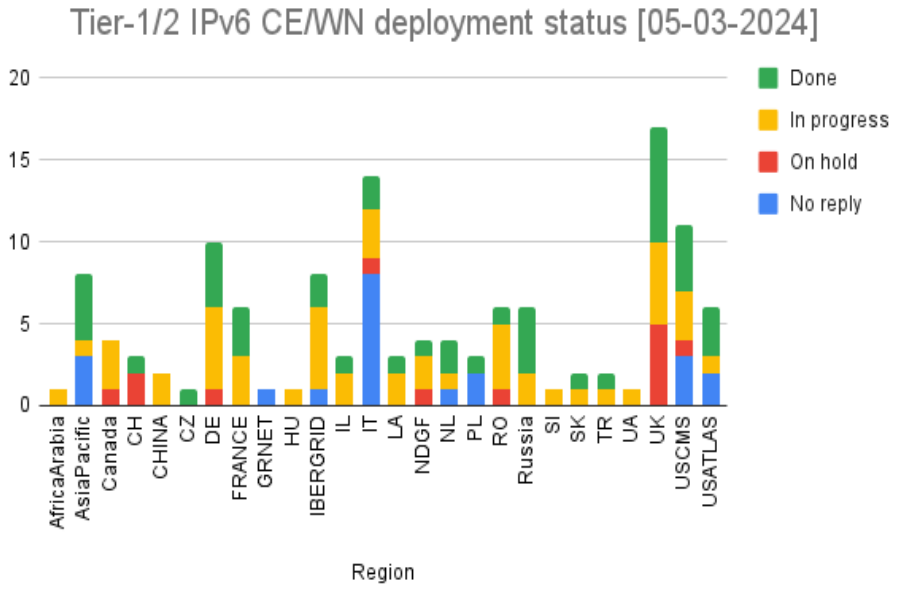
- scarcity of IPv4 addresses
- full IPv6 support
- by the experiment middleware stacks,
- readiness for adoption of packet marking in network traffic
- reduction of residual IPv4 traffic
- fulfillment of government mandates by some countries

Goal

- Storage services successful deployed
- Next step is deploying IPv6 fully enabled infrastructure at entire WLCG
- WLCG management board and LHC experiments approved
  - deployment plan for IPv6 requires that Tier-1 and Tier-2 sites
  - deploy dual-stack connectivity (IPv4 + IPv6)
  - on their computing elements (CE) and worker nodes (WN)
  - by 30 June 2024

→ For more details see https://indico.cern.ch/event/1225423/

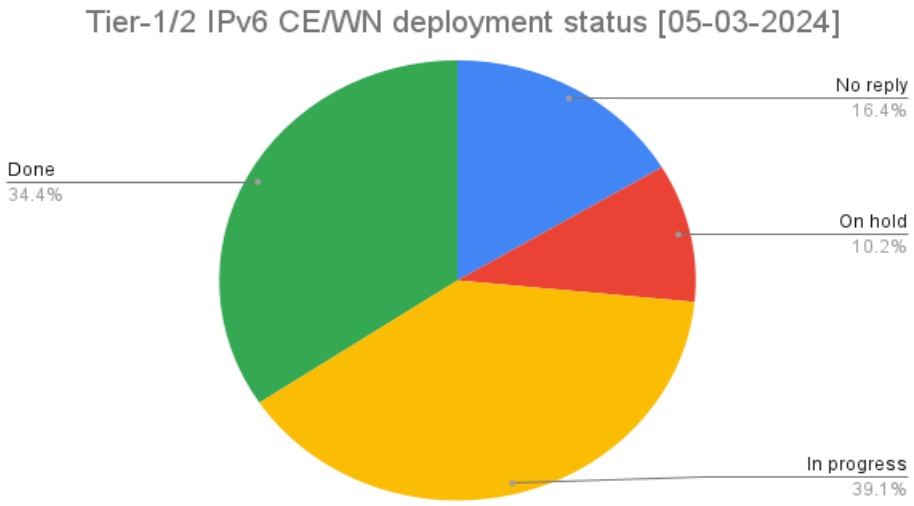
GGus Ticket campaign

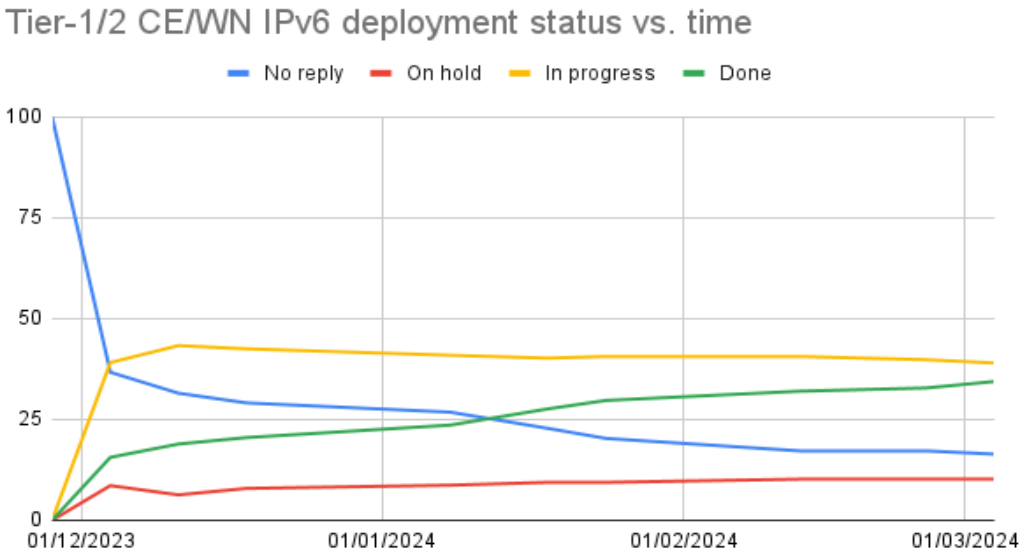


IPv6 deployment steps

Switching off IPv4 is not requested nor recommended at this stage: any step in this direction should first be discussed with the LHC experiments you support and WLCG. In general terms we should have to wait until all traffic happens via IPv6.

The practical objective is to allow worker nodes to contact IPv6-only central services, and central job submission services (e.g. pilot factories) having only IPv6 connectivity to submit jobs to computing elements. There is no requirement to allow incoming connections in the WNs, as it has always been the case.





Migration towards IPv6 at DE-KIT

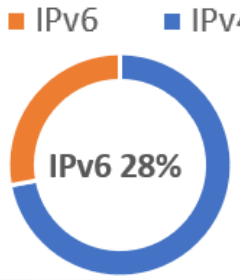
DE-KIT IPv6 WorkerNodes migration

detailed monitoring is still in place for narrowring down IPv4 communication

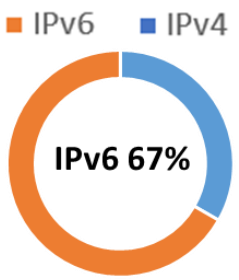
- to list all unhandled IPv4 packets
  - 8884 – Alice: operation report
  - 2049 – NFS
  - 1094 – XrootD
  - 961[89] – LRMS (20% only internal to WN-Farm)

A view statistics

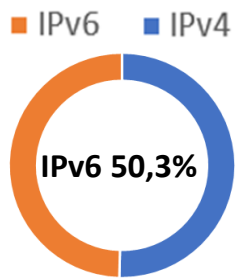
20220415:
IPv4: → 80 Mio
IPv6: → 31 Mio



20221220:
- IPv4: → 42 Mio
- IPv6: → 86 Mio



20240313:
- IPv4: → 184,8 Mio
- IPv6: → 187,7 Mio



Details of Squid

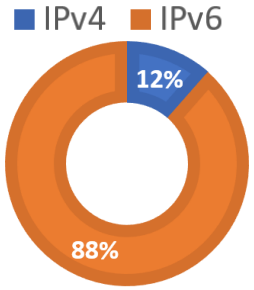
- SQUIDS (Proxyserver and Web-Cache):
  - some SQUIDS still IPv4 only (**migration to dualstack in proccess**)
  - significant part of connections via public IPv4
  - => to check: if CVMFS can prefer IPv6?
    (CVMFS → CernVM-File-System)
    - CVMFS sending via http request to squid
    - CVMFS has DN configuriert that needs to be resolved
      → default chooses IPv4 address
    - **Solution** => cvmfs_ipfamily_prefer=6

SQUIDS migrated all to dual-stack

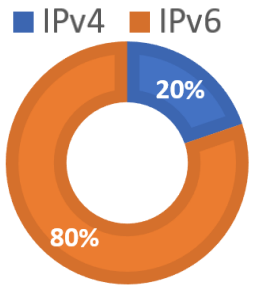
later all SQUIDS migrated to dual-stack deployment

CVMFS now
- mainly IPv6 but:
- on WorkerNodes uses IPv6 (with deployed flag: CVMFS_IPFAMILY_PREFER=6)
- CVMFS frontier uses still IPv4 even while both systems dual-stack
- but switching of IPv4 → froniters will operate over IPv6
- the CMS CVMFS frontiers offers in site-local-config.xml the Option:

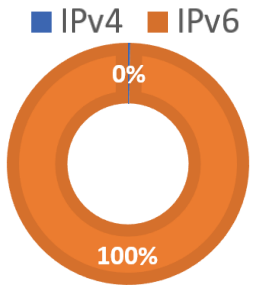
26-07-2022
IPv4 : 1,25 mio. IPv6: 9,6 mio.



23-10-2022
IPv4 : 4,44 mio. IPv6: 18 mio.



25-10-2023
IPv4 : 64 k IPv6: 22 mio.




```
<frontier-connect>
...
        <prefer ipfamily="6"/>
...
</frontier-connect>
```

Summary

WN + CE -- next step towards IPv6

- dual-stack deplyment
- ongoing and activley monitored (by HEPiX-IPv6 Working Group (Andrea Sciaba))

DC24

- analysing all the figures collected during DC24 still ongoing
- detailed (/in depth) result understanding
- Enabing LHC-HL requested performance (data throughput)
- still pending LHCOPN IPv4 traffic
  - ongoing analysis and following up required

DE-KIT IPv6 migration not done yet

- new WorkerNodes with ARM processor will be deployed → analyse protocol behaviour

new experiments → with IPv6 only,

- no migration „nightmare“

Thanks for your attention

Questions