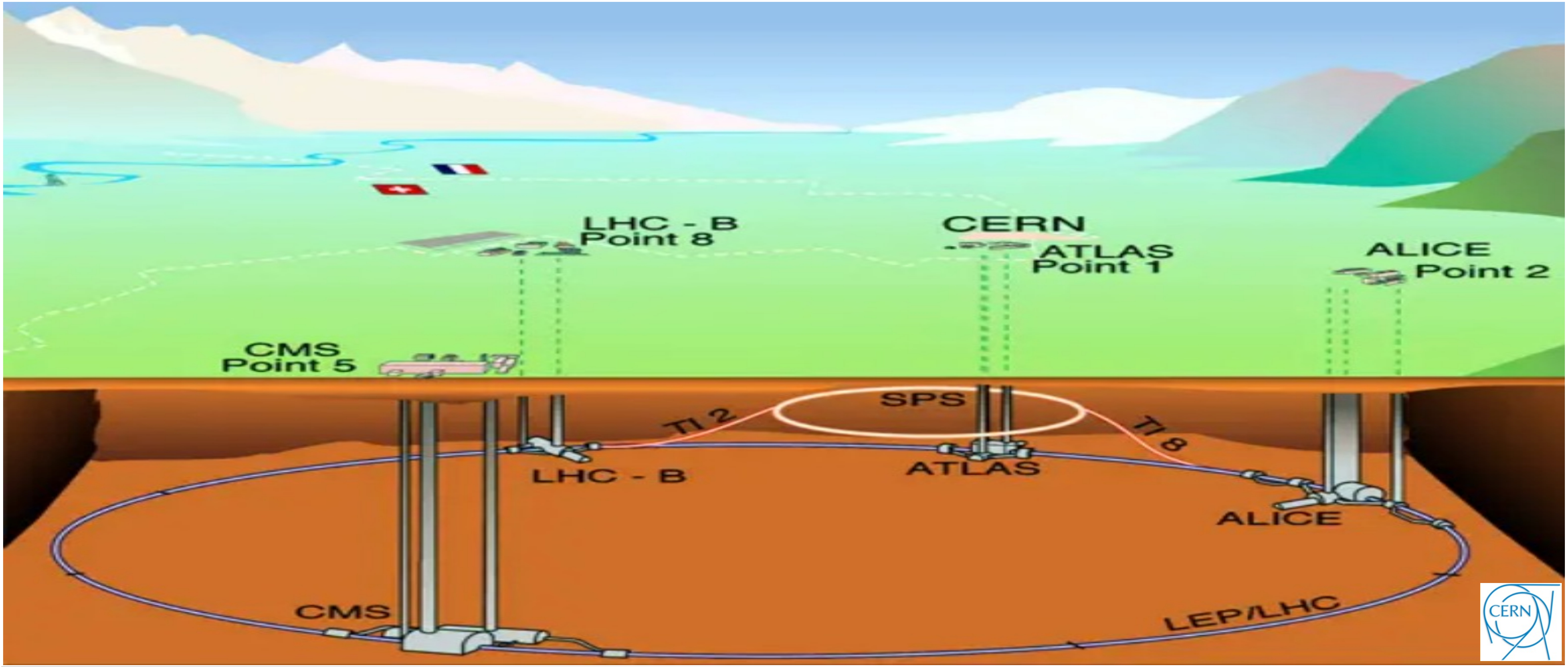
Moving towards IPv6-only in the German Tier-1 Data Center of the CERN Large Hadron Collider

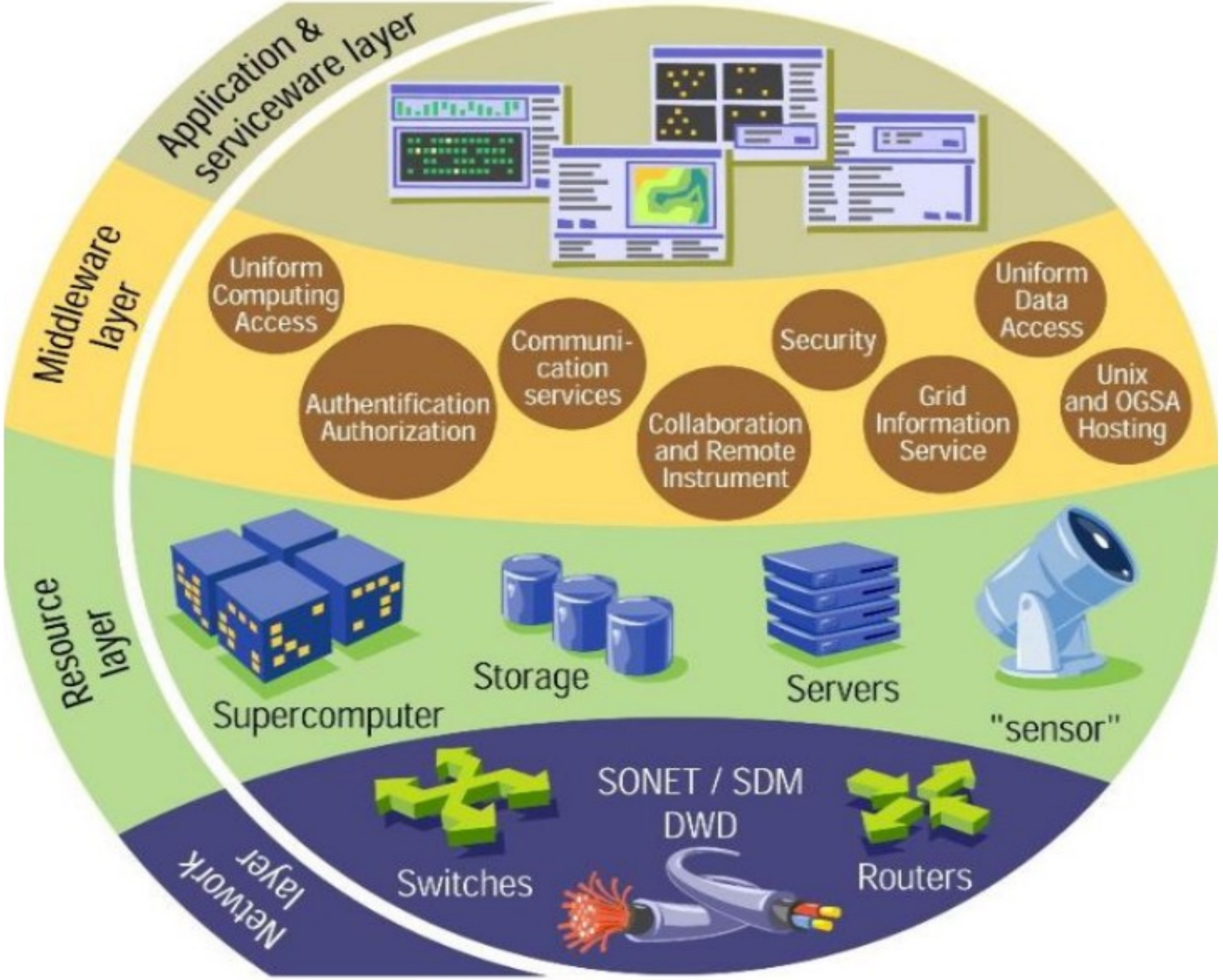
Bruno Hoeft, Matthias Schnepf, Max Fischer, Andreas Petzold

Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, {first-.familyname}@kit.edu

LHC accelerator and experiments



WLCG

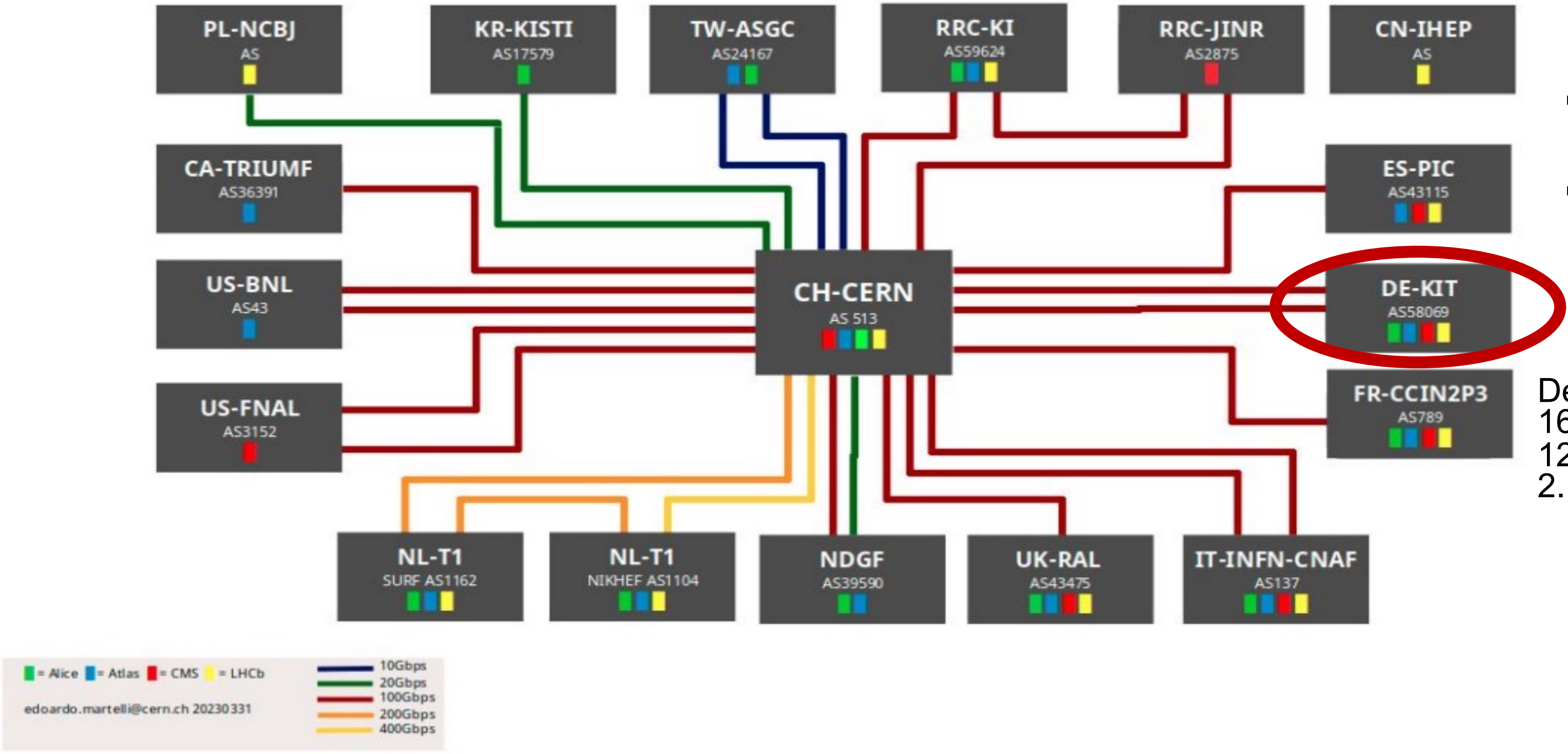


world wide distributed computing engine for the Large Hadron Collider emerging data Arround 2000:

the memberstates decided for remote and distributed installations. The foundation for the **W**orldwide **L**arge hadron collider **C**omputing **G**rid (WLCG) was layed

raw data calculation and tape storage centres

LHCOPN



PL-NCBJ AS 🟨

KR-KISTI AS17579 🟩

TW-ASGC AS24167 🟦🟩

RRC-KI AS59624 🟩🟦🟨

RRC-JINR AS2875 🟥

CN-IHEP AS 🟨

CA-TRIUMF AS36391 🟦

US-BNL AS43 🟦

US-FNAL AS3152 🟥

CH-CERN AS 513 🟥🟦🟩🟨

ES-PIC AS43115 🟦🟥🟨

DE-KIT AS58069 🟩🟦🟥🟨

FR-CCIN2P3 AS789 🟩🟦🟥🟨

NL-T1 SURF AS1162 🟩🟦🟨

NL-T1 NIKHEF AS1104 🟩🟦🟨

NDGF AS39590 🟩🟦

UK-RAL AS43475 🟩🟦🟥🟨

IT-INFN-CNAF AS137 🟩🟦🟥🟨

🟩 = Alice 🟦 = Atlas 🟥 = CMS 🟨 = LHCb

edoardo.martelli@cern.ch 20230331

10Gbps
20Gbps
100Gbps
200Gbps
400Gbps

Tier0 to Tier1s network

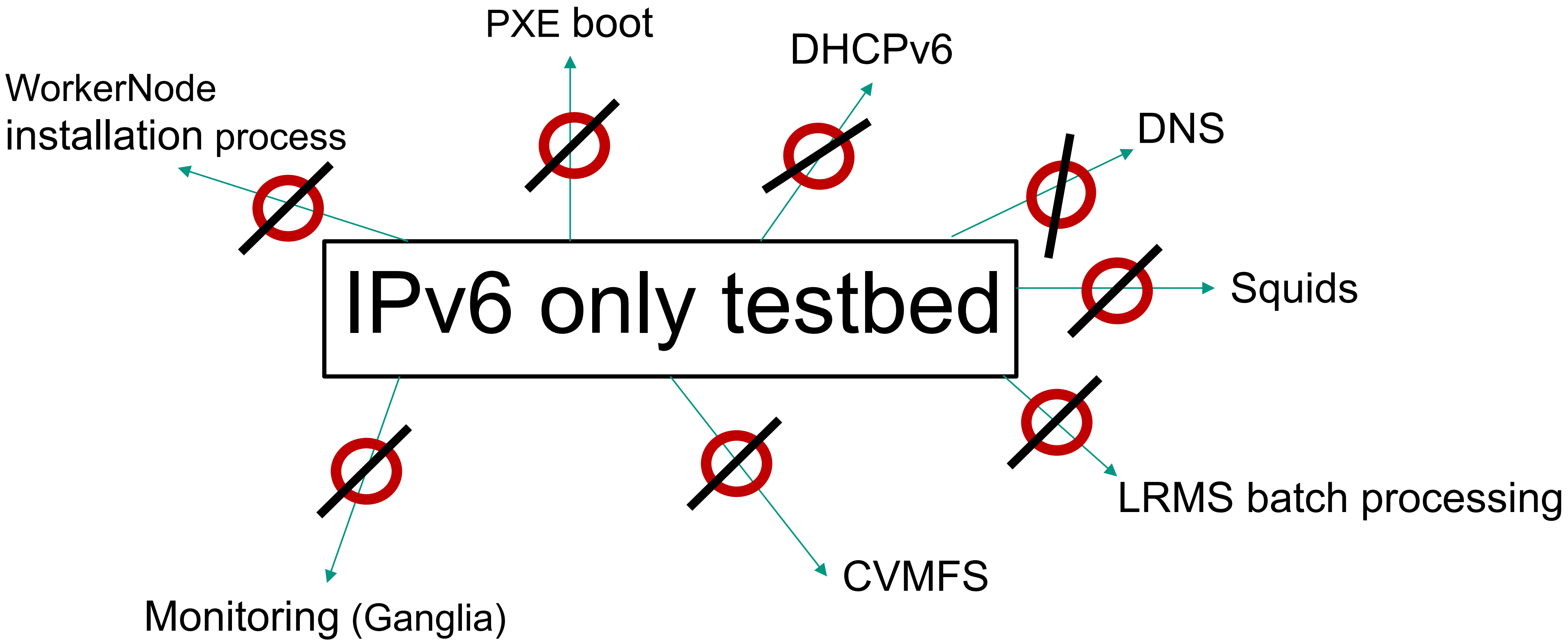
Dedicated 10/100/400Gbps links
16 Tier1s + 1 Tier0
12 countries in 3 continents
2.1 Tbps to the Tier0

GridKa

- worker node farm
  - 217 aktive hardware systems
  - 42500 compute cores
- online-storage
  - 99 PB effectiv storage capacity
  - 6824 HDDs
  - 100 Server
- nearline-storage
  - 85 PB saved on tapes
  - 135PB available capacity on tapes
- wan network
  - 2 x 100Gb/s direct to CERN (LHCOPN)
  - 2 x 100Gb/s to DFN (LHCONE overley)
  - 2 x 100Gb/s to Belwue

Building IPv6 Testbed

HEPiX- IPv6 working group asking for IPv6 only testbed



DE-KIT – workernode migration towards IPv6

Pro-active IPv6 Monitoring at DE-KIT

packet number decreased from monitoring in 2022 to 2023
- power budget depending workernodes were switched of (while still LHC MOU is full filled)

Detailed monitoring at DE-KIT (GridKa)

- Monitor all comunications between WorkerNodes and
    - administration
    - job submission
    - Storage
    - …

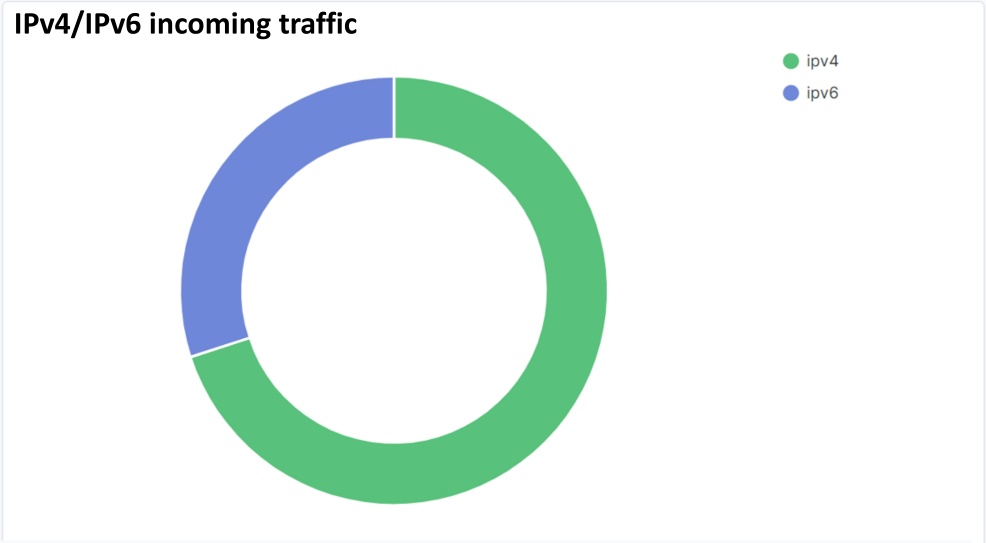
Monitoring of process intercomunication at DE-KIT (GridKa)

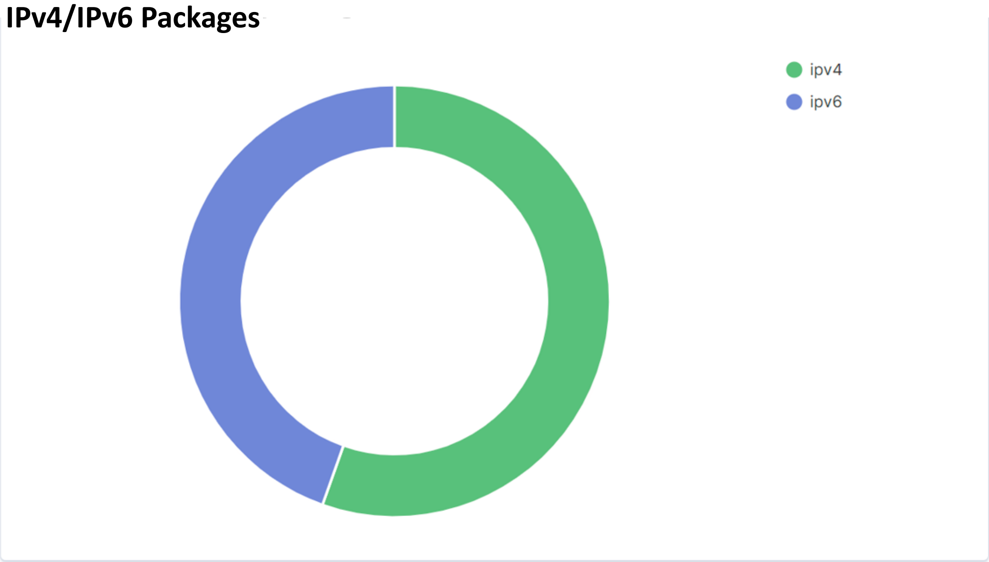
- with packetbeat collecting network data
- logstach pushing the data to opensearch (former elastic search) for storing the data
- kibana for visualizing (no opensearch – only easy search requests)
  - started with a small set of workernodes (storing the data „longterm“ → ~ 6 days)
  - while enlarging the set of workernodes graduately data keeping time had to be limited to less than one week only (for not exceeding the storage size of 0,5 Tbyte)

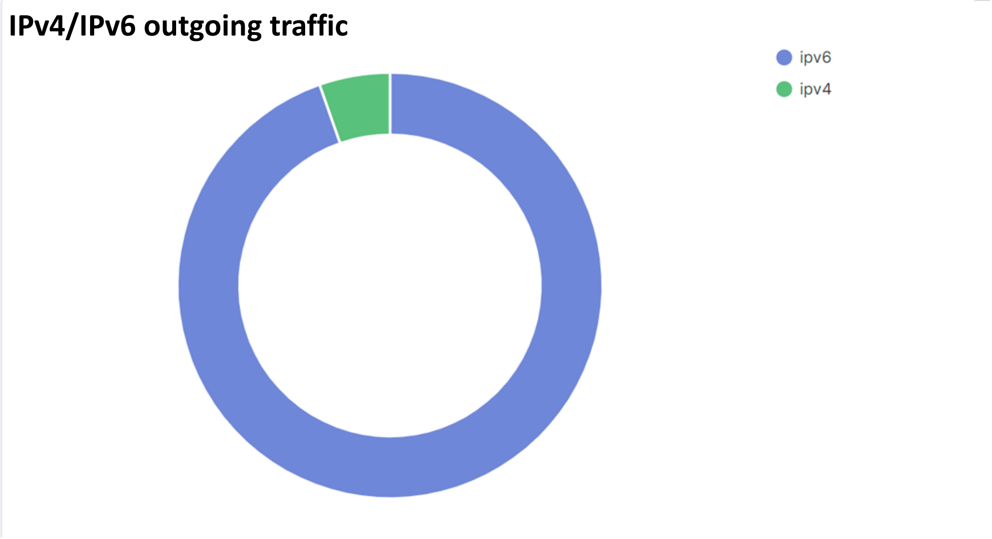
- Identify IPv4 protocol usage

Snapshot of a dashbord

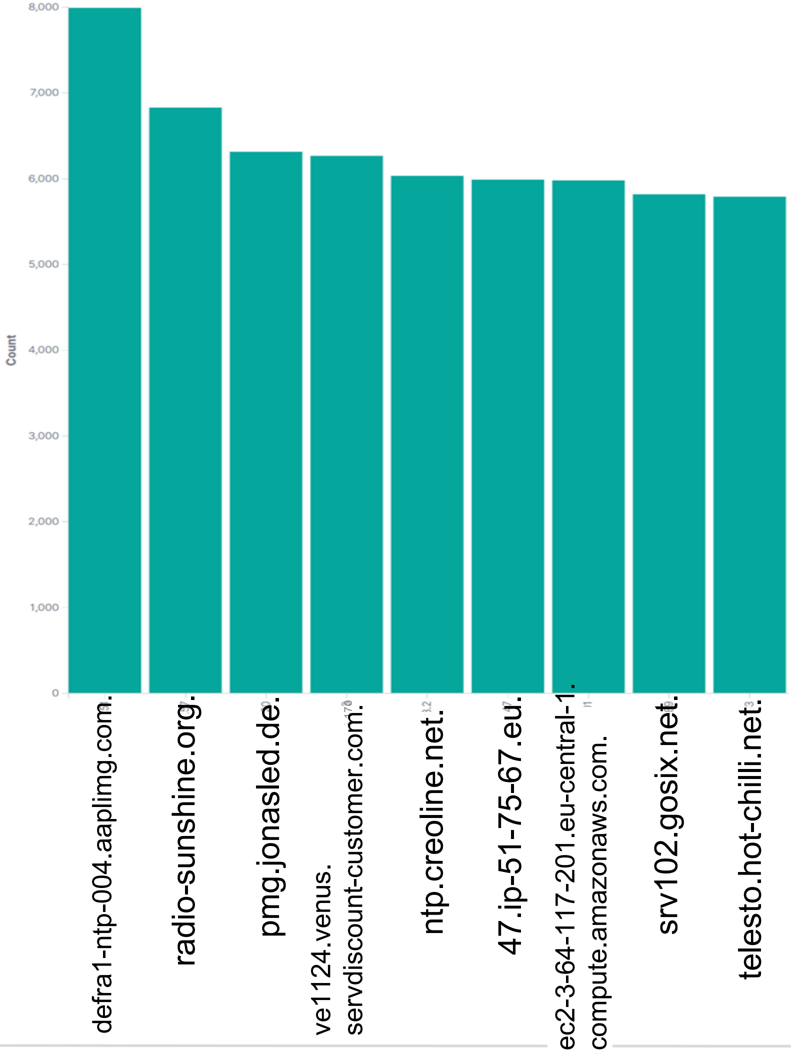
at 08.09.22
all worker nodes already dual-stack deployed

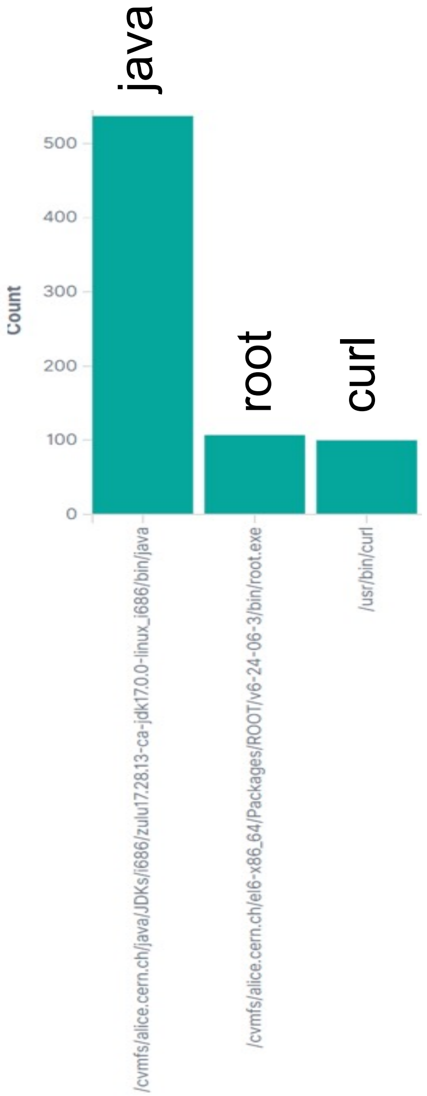

IPv4/IPv6 incoming traffic


IPv4/IPv6 Packages


IPv4/IPv6 outgoing traffic

NTP ?





- Many NTP / port 123 connections
  - During 24 hours approx. 210.000
  - NTP → IPv4 only (depending on dualstack enabling of rack-manager (40.000 internal))
  - Monitoring was first pointing especially 10.1.12 and 10.1.18 → checking later showed that much more racks running ntp check via private addr. (NAT)
  - 160.000 external communications → some of the destination server have quite dubious „names“
- process-tracking
  - The numbers of NTP communication process and matched process is not matching yet

SOLVED

- NTP.ORG

  → returns sometimes funny addresses

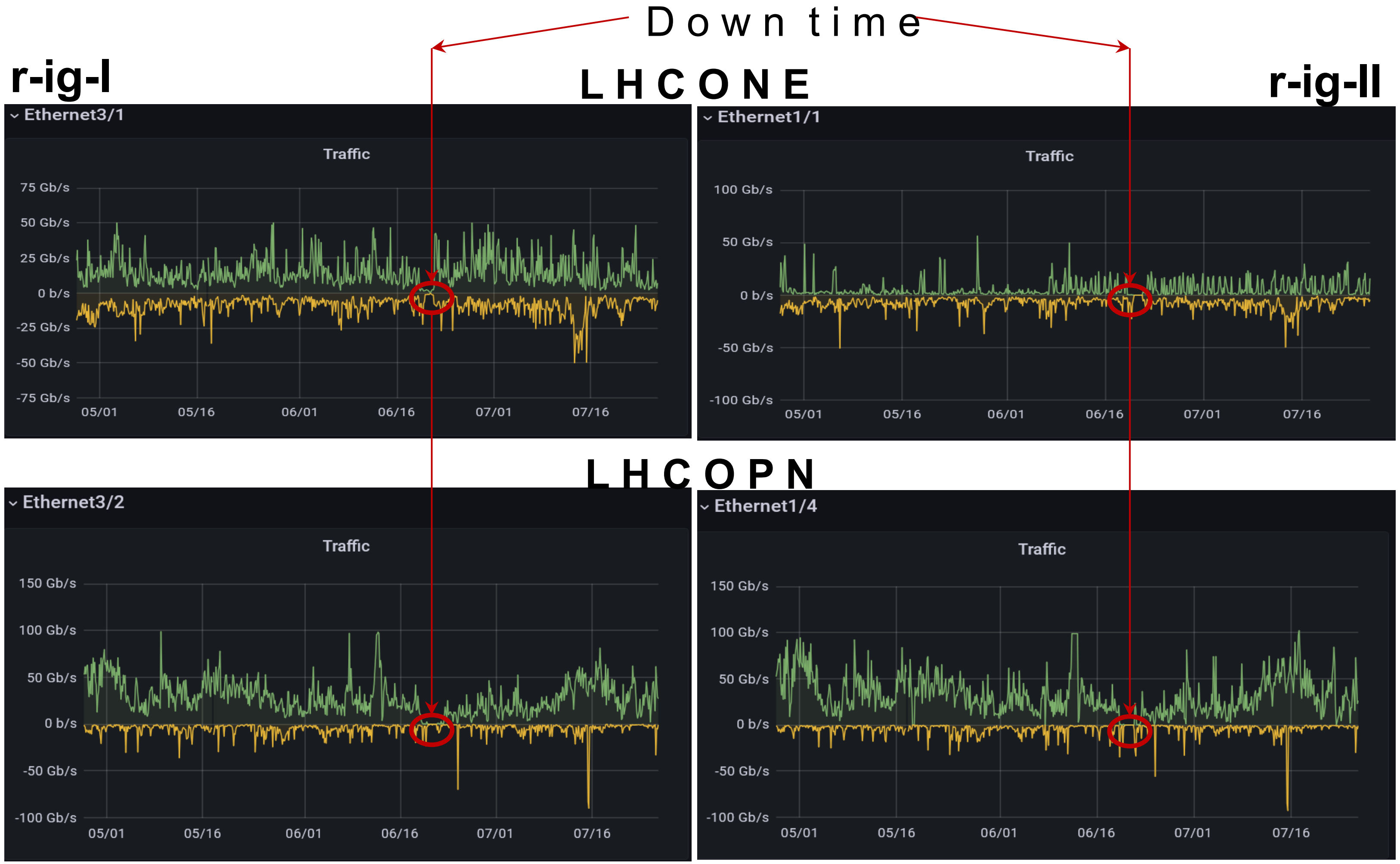
dCache upgrade to 7.2.15

Upgrade from dCache version 6.2.34 to 7.2.15

Two day downtime at June 20th and 21st 2022

- HTTP-TPC transfers now prefer IPv6 address, if both endpoints support it.
- fixed handling of Storage Resource Reporting (SRR) requests over IPv6
- Handle IPv6 address when running HTTP(s) Third Party Copy (TPC) with gridsite delegation
- Storage Resource Manager (SRM) : Fix IPV6 logging for SRM

WAN Interfaces



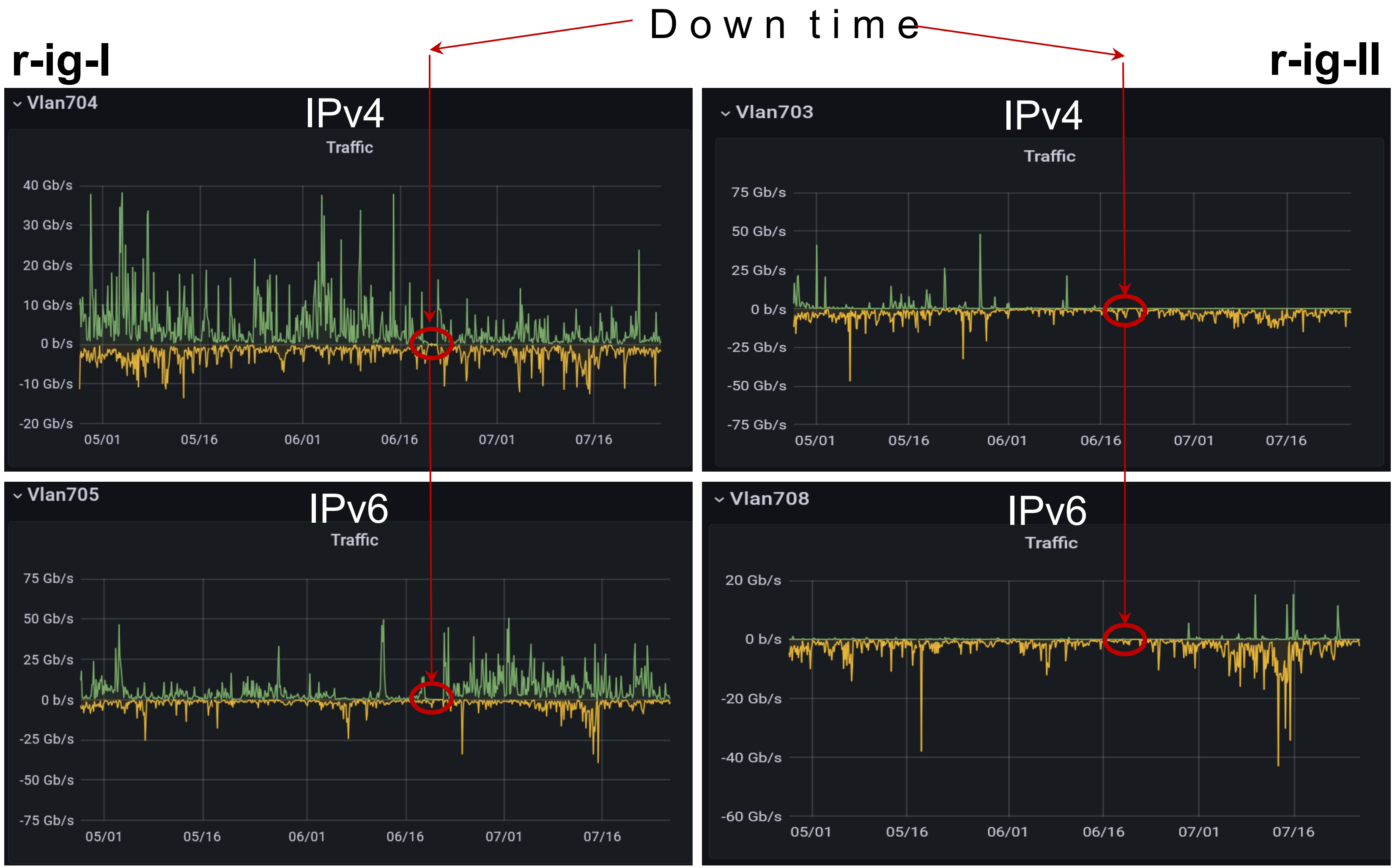
r-ig-I (DE-KIT Border Router):
left two Interfaces

- Ethernet 3/1 (Internet + LHCONE) +
- Ethernet 3/2 (LHCOPN)

r-ig-II (DE-KIT second Border Router):
right two Interfaces

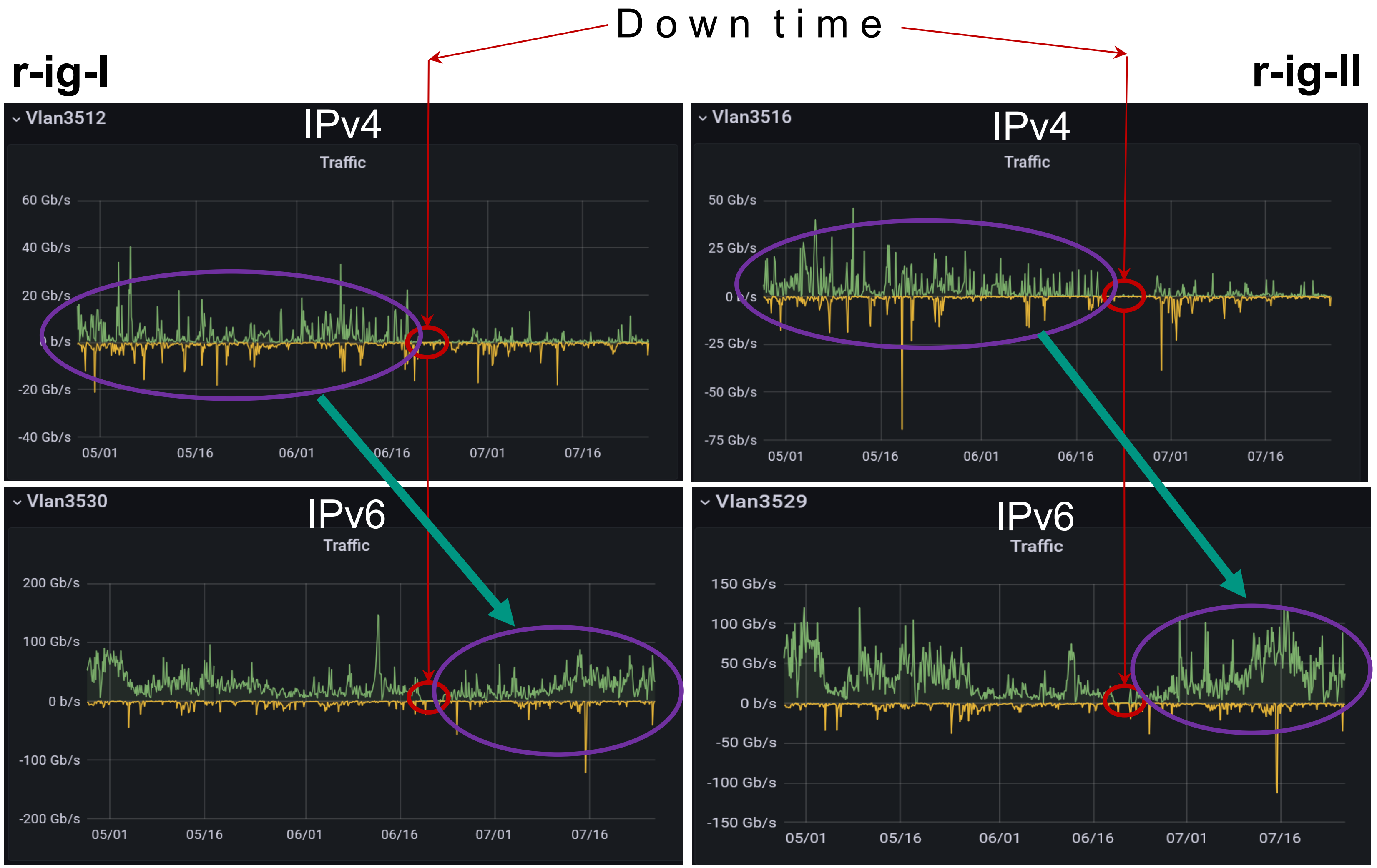
- Ethernet 1/1 (Internet + LHCONE) +
- Ethernet 1/4 (LHCOPN)

LHCONE IPv4 / IPv6
transfer pattern after downtime



Graph over 90 days
Traffic of LHCONE
moved partioly from the IPv4 vlans
after the downtime to the IPv6 Vlans

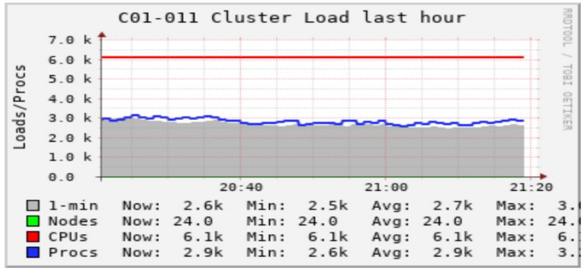
LHCOPN IPv4 / IPv6
transfer pattern after downtime



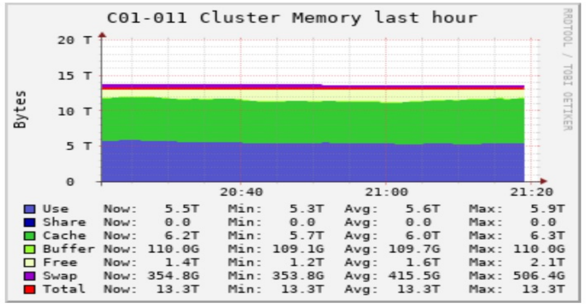
Graph over 90 days
Traffic of LHCOPN
moved from the IPv4 vlans
after the downtime to the IPv6 Vlans

Monitoring

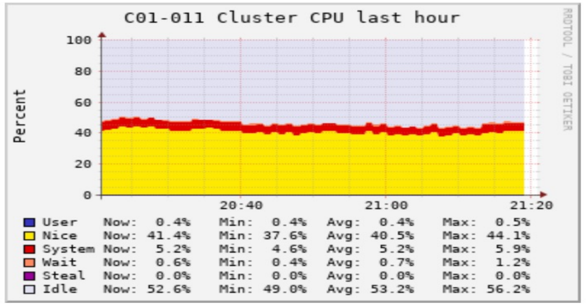
GANGLIA



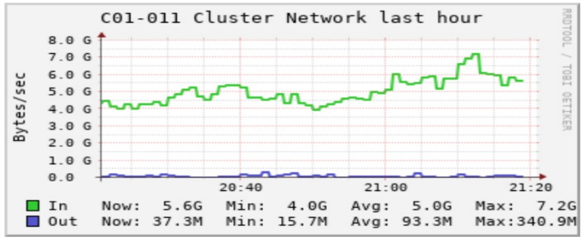
```
▨ 1-min  Now:  2.6k  Min:  2.5k  Avg:  2.7k  Max:  3.
🟩 Nodes  Now: 24.0   Min: 24.0   Avg: 24.0   Max: 24.
🟥 CPUs   Now:  6.1k  Min:  6.1k  Avg:  6.1k  Max:  6.
🟦 Procs  Now:  2.9k  Min:  2.6k  Avg:  2.9k  Max:  3.
```



```
🟦 Use    Now:   5.5T  Min:   5.3T  Avg:   5.6T  Max:   5.9T
🟦 Share  Now:   0.0   Min:   0.0   Avg:   0.0   Max:   0.0
🟩 Cache  Now:   6.2T  Min:   5.7T  Avg:   6.0T  Max:   6.3T
🟩 Buffer Now: 110.0G  Min: 109.1G  Avg: 109.7G  Max: 110.0G
□ Free   Now:   1.4T  Min:   1.2T  Avg:   1.6T  Max:   2.1T
🟪 Swap   Now: 354.8G  Min: 353.8G  Avg: 415.5G  Max: 506.4G
🟥 Total  Now:  13.3T  Min:  13.3T  Avg:  13.3T  Max:  13.3T
```

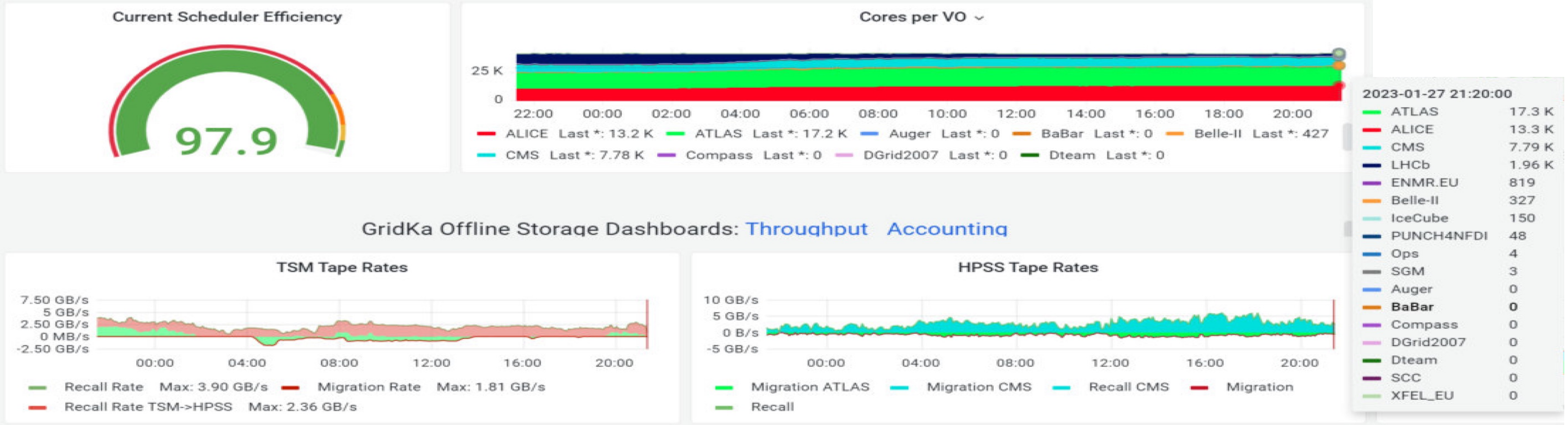


```
🟦 User   Now:  0.4%  Min:  0.4%  Avg:  0.4%  Max:  0.5%
🟨 Nice   Now: 41.4%  Min: 37.6%  Avg: 40.5%  Max: 44.1%
🟥 System Now:  5.2%  Min:  4.6%  Avg:  5.2%  Max:  5.9%
🟧 Wait   Now:  0.6%  Min:  0.4%  Avg:  0.7%  Max:  1.2%
🟪 Steal  Now:  0.0%  Min:  0.0%  Avg:  0.0%  Max:  0.0%
□ Idle   Now: 52.6%  Min: 49.0%  Avg: 53.2%  Max: 56.2%
```



```
🟩 In   Now:  5.6G  Min:  4.0G  Avg:  5.0G  Max:  7.2G
🟦 Out  Now: 37.3M  Min: 15.7M  Avg: 93.3M  Max:340.9M
```

- Migration of Ganglia to IPv6 will not persuit
- Ganglia will be replaced by opensearch, kibana and grafana



2023-01-27 21:20:00

ATLAS	17.3 K
ALICE	13.3 K
CMS	7.79 K
LHCb	1.96 K
ENMR.EU	819
Belle-II	327
IceCube	150
PUNCH4NFDI	48
Ops	4
SGM	3
Auger	0
BaBar	**0**
Compass	0
DGrid2007	0
Dteam	0
SCC	0
XFEL_EU	0

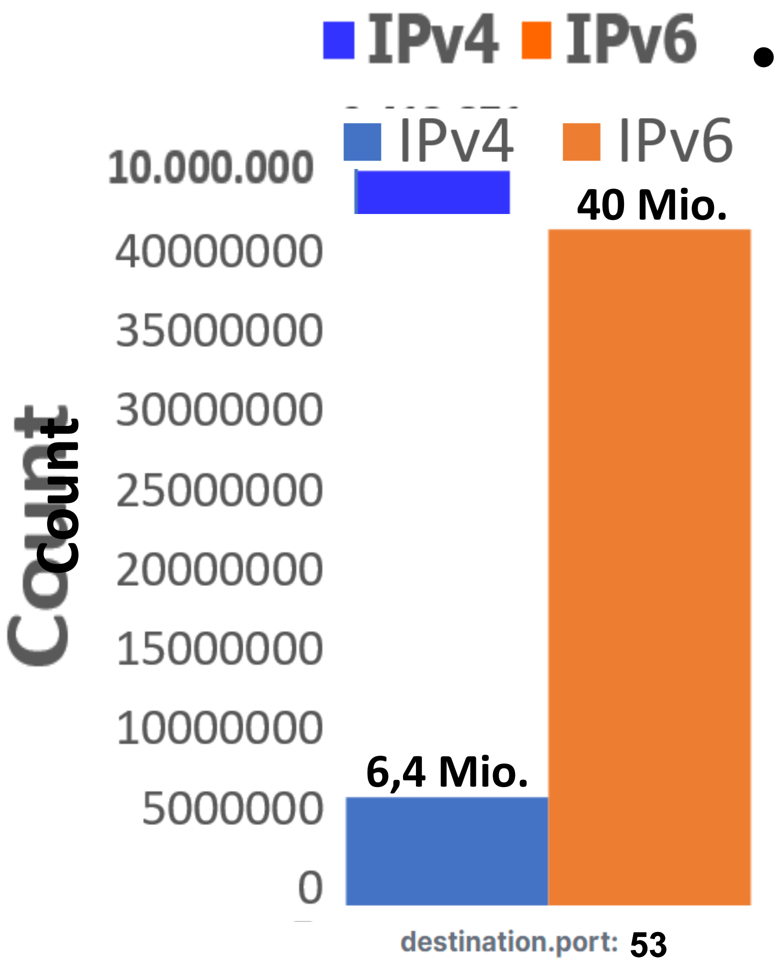
Logstach → is now IPv6

Logstach (port 5047) → dual-stack deeployed

statistic:

28-07-2022 → IPv4 385k – IPv6 1,41M
23-10-2022 → IPv4 476k – IPv6 1,39M
23-12-2022 → IPv4 227k – IPv6 450k
30-10-2023 → IPv4 906k – IPv6 864k

Closer look at DNS



- GridKa DNS:
  - IPv4 only count : 9,412,871 (24 hours)
  - DNS (Bind) Server and WN is already dual-stack
  - at WN resolve.conf first lines IPv4
    - Make sure IPv6 DNS server addresses listed and
    - place it before IPv4
    - every new deployed host: the first lines are IPv6 resolver addresses of the **resolve.conf** file **followed by the IPv4 addresses**

```
•   nameserver 2a00:139c:address
•   nameserver 2a00:139c:address
•   nameserver 10.privat-address
•   nameserver 10.privat-address
```

→ **Resolve.conf update: reprovisioning required**

Administrative Services

- at each rack is a Rack Manager deployed:
  - Starting in 2001 with private IPv4 only
  - Migration process initiated (but still in progress)
    → enable dual-stack (AAAA)
    - NTP
    - rsyslog (→ migration → still pending (port 514))
    - Monitoring (GmonD → Ganglia Client)
    - DHCP (→ migration to DHCPv6 pending)

WN – deployment process

- Redhat Satellite Server (foreman)
  - Used for management of most GridKa hosts:
    - Manages redhat Subscriptions
    - Controlls kickstart installations (DHCP / PXE)
    - Provides yum repos
    - Provides CA (certificate authority) and ENC (encryptor) functionalities for puppet
  - Uses modular architecture. Additional functionalities can be added via so called capsules
    - TFTP server (IPv6 ready - dual-stack)
    - Puppetmaster (IPv6 ready - dual-stack)
    - Pulp (software repository management (IPv6 ready - dual-stack))
    - DNS (IPv6 ready - dual-stack)
    - DHCP (currently DHCPv6 capsule not available)

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 → **not tested yet** (end of 2022)

SQUIDS migrated all to dual-stack

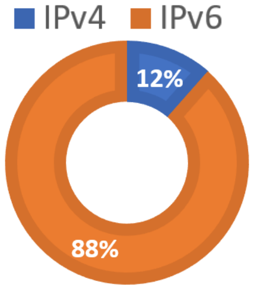
During the second half of 2022 all SQUIDS migrated to dual-stack deployment

CVMFS now
- manly 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:

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

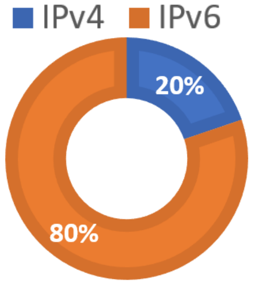
26-07-2022

IPv4 : 1,25 mio. IPv6: 9,6 mio.



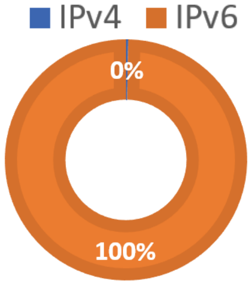
23-10-2022

IPv4 : 4,44 mio. IPv6: 18 mio.



25-10-2023

IPv4 : 64 k IPv6: 22 mio.



Batch-Processing -- LRMS (HT-Condor) all dual-stack

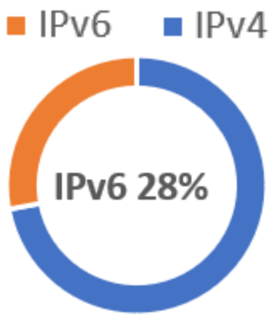
- LRMS (**Local Resource Management System**) HTCondor at GridKa (all dual-stack and set to **prefer** the protocoll **IPv6** (Port 9618/9)
  - 4080 – HTCondor (rooster-deamon) → migrated all towards IPv6 (HTCondor → startd)
  - percentage increased toward IPv6 at 28-06-2022→ IPv4: **895k** to IPv6: **255k**
  - 1,2% IPv4 28-07-2022 → IPv4: 27k, IPv6: 2,17 mio.
  - **11%** IPv4 02-01-2023 → IPv4: 287k, IPv6: 2,28 mio.
  - **18%** IPv4 31-10-2023 → IPv4: 2,68 mio., IPv6: 11,7 mio.

Less then 1% (0,0049%) of IPv4 is internal traffic

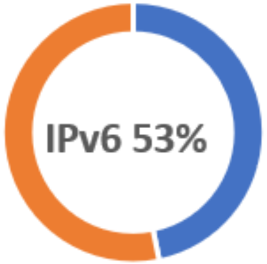
(communication with home → the LRMS demons uses protocol of Home-Institution)

A view statistics

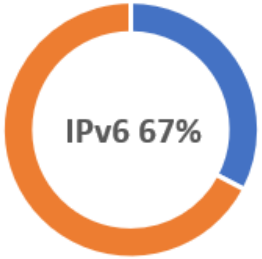
- 15-04-2022:
  - IPv4: → 80 mio.
  - IPv6: → 31 mio.



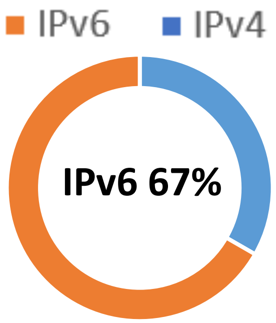
- 26-07-2022:
  - Ipv4 → 44 mio.
  - Ipv6 → 50 mio.



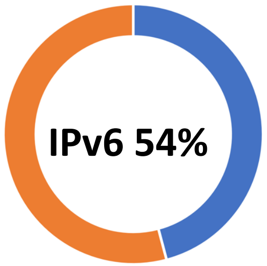
- 23-10-2022:
  - IPv4 → 69 mio.
  - IPv6 → 142 mio.



20-12-2022:
- IPv4: → 42 mio.
- IPv6: → 86 mio.



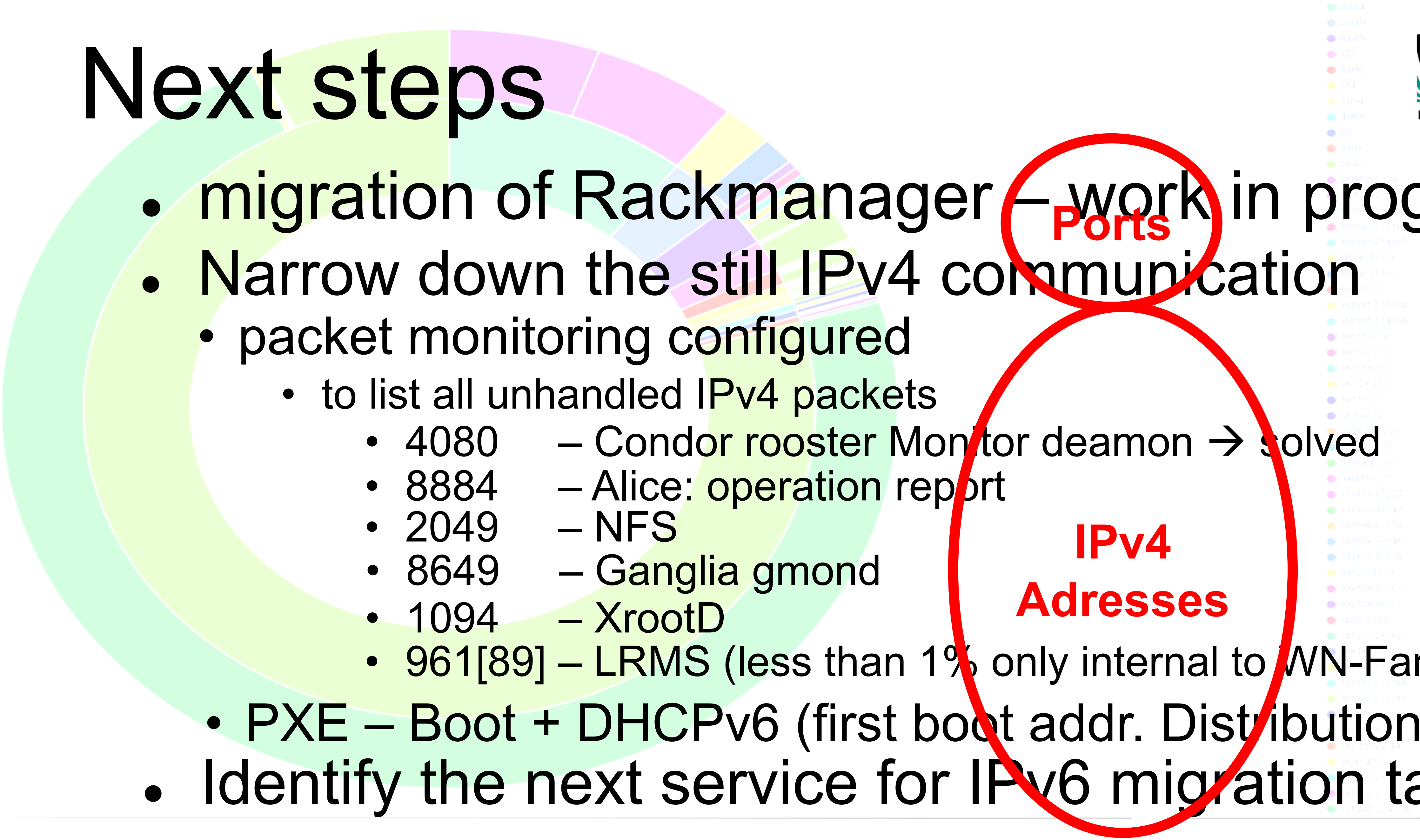
31-10-2023:
- IPv4: → 122 mio.
- IPv6: → 144 mio.



(packets in 24 hours)

# of **WorkerNodes** included in the statistic expanded

Next steps



- migration of Rackmanager – work in progress
- Narrow down the still IPv4 communication
  - packet monitoring configured
    - to list all unhandled IPv4 packets
      - 4080 – Condor rooster Monitor deamon → solved
      - 8884 – Alice: operation report
      - 2049 – NFS
      - 8649 – Ganglia gmond
      - 1094 – XrootD
      - 961[89] – LRMS (less than 1% only internal to WN-Farm)
  - PXE – Boot + DHCPv6 (first boot addr. Distribution)
- Identify the next service for IPv6 migration tasks

Ports

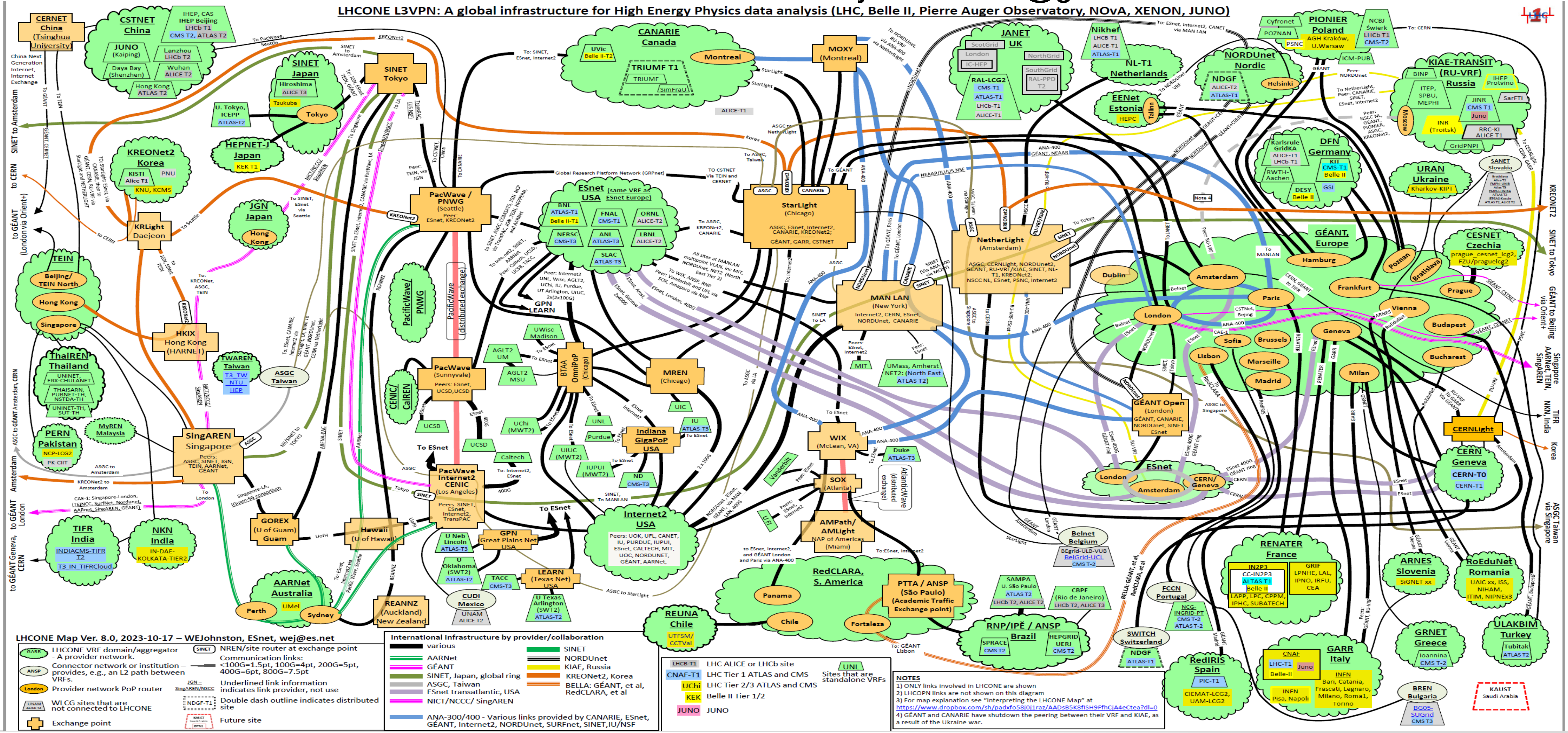
IPv4 Adresses

Thx for your attention

backup Slides

LHCONE

William E Johnston <johnstonwe@gmail.com>



LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)

LHCONE Map Ver. 8.0, 2023-10-17 – WEJohnston, ESnet, wej@es.net

- GARR — LHCONE VRF domain/aggregator - A provider network.
- ANSP — Connector network or institution – provides, e.g., an L2 path between VRFs.
- London — Provider network PoP router
- UNAM ALICE T2 — WLCG sites that are not connected to LHCONE
- Exchange point
- SINET — NREN/site router at exchange point
- Communication links: <100G=1.5pt, 100G=4pt, 200G=5pt, 400G=6pt, 800G=7.5pt
- JGN – SingAREN/NSCC — Underlined link information indicates link provider, not use
- NDGF-T1 — Double dash outline indicates distributed site
- KAUST Saudi Arabia — Future site

International infrastructure by provider/collaboration

- various
- AARNet
- GÉANT
- SINET, Japan, global ring
- ASGC, Taiwan
- ESnet transatlantic, USA
- NICT/NCCC/ SingAREN
- SINET
- NORDUnet
- KIAE, Russia
- KREONet2, Korea
- BELLA: GÉANT, et al, RedCLARA, et al
- ANA-300/400 - Various links provided by CANARIE, ESnet, GÉANT, Internet2, NORDUnet, SURFnet, SINET,IU/NSF

- LHCB-T1 — LHC ALICE or LHCb site
- CNAF-T1 — LHC Tier 1 ATLAS and CMS
- UChi — LHC Tier 2/3 ATLAS and CMS
- KEK — Belle II Tier 1/2
- JUNO — JUNO
- UNL — Sites that are standalone VRFs

NOTES

1) ONLY links involved in LHCONE are shown
2) LHCOPN links are not shown on this diagram
3) For map explanation see "Interpreting the LHCONE Map" at https://www.dropbox.com/sh/padxfo58j0j1raz/AADsB5K8fISH9FfhCjA4eCtea?dl=0
4) GÉANT and CANARIE have shutdown the peering between their VRF and KIAE, as a result of the Ukraine war.

Details of Alice VOBoxes:

- ALICE VOBoxes:
  - Client to VOBox prefers IPv4 (ALICE Monitoring (UDP))
  - => to check the possibility of IPv6 migration with ALICE (still ongoing)
    - dual-stack enabling works and
    - if Preference towards IPv6 is possible
    - ALICE is constrained by IPv6 unavailability on other sites
  - → advice of Alice : switch of IPv4 at VO-BOX (the none monitoring VO-BOX)
    - Timing still under discussion
  - Monitoring (port 8884 / IPv4 only) → 11 Mio. (/24 hours)
- XRootD:
  - via public IPv4 (ALICE)
  - All ALICE XRootD SE are dual-stack deployed
  - older version of XRootD → upgrade to current XRootD should improve, is still pending
  - → advice of Alice : get IPv6 ready – but wait for switching it on till complete Alice is IPv6 ready
- Dest port 1094 –Ipv4/ipv6 → XRootD (alice, belle2, atlas, cms)