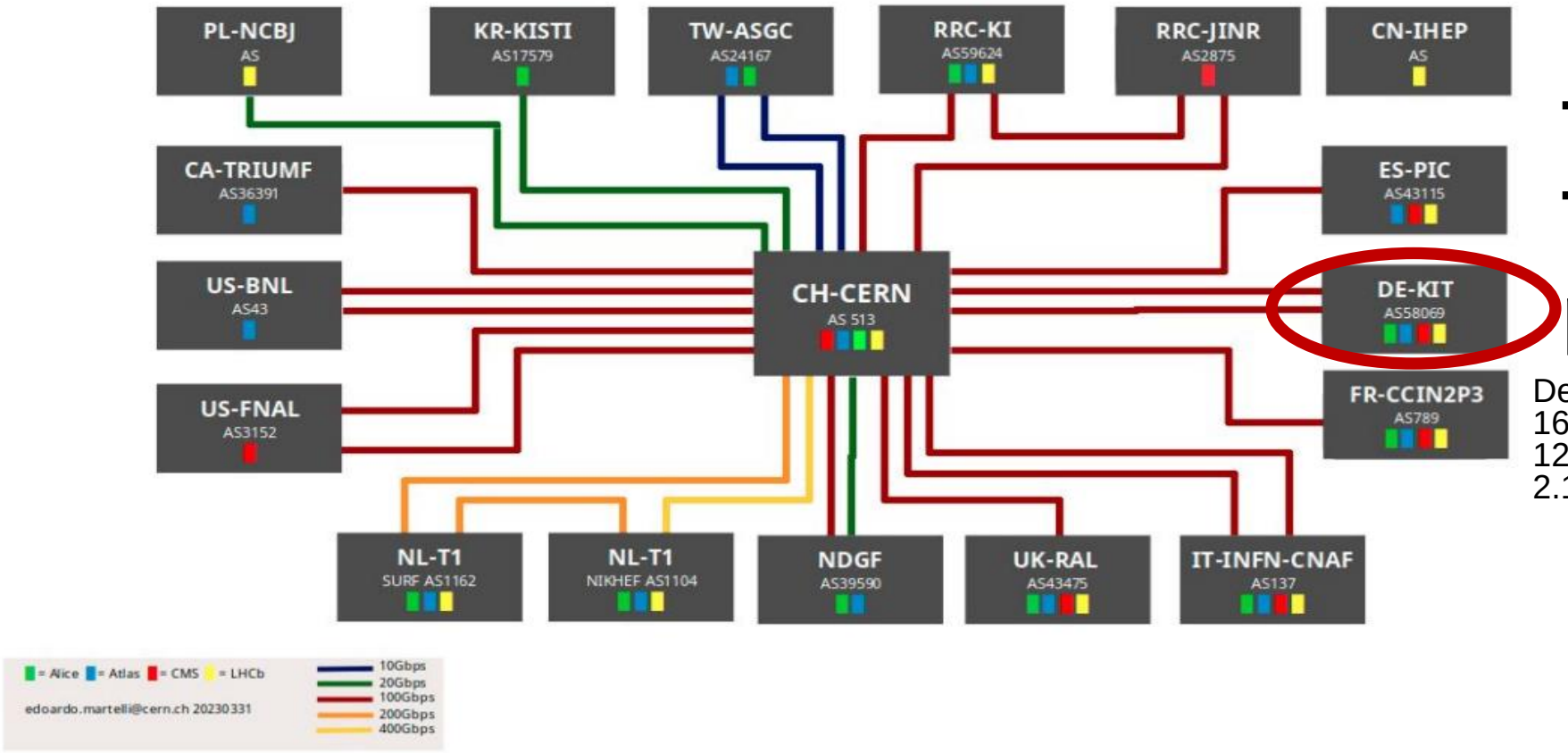
Karlsruhe Institute of Technology

IPv4 to 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

raw data calculation and tape storage centres

LHCOPN



Tier0 to Tier1s network

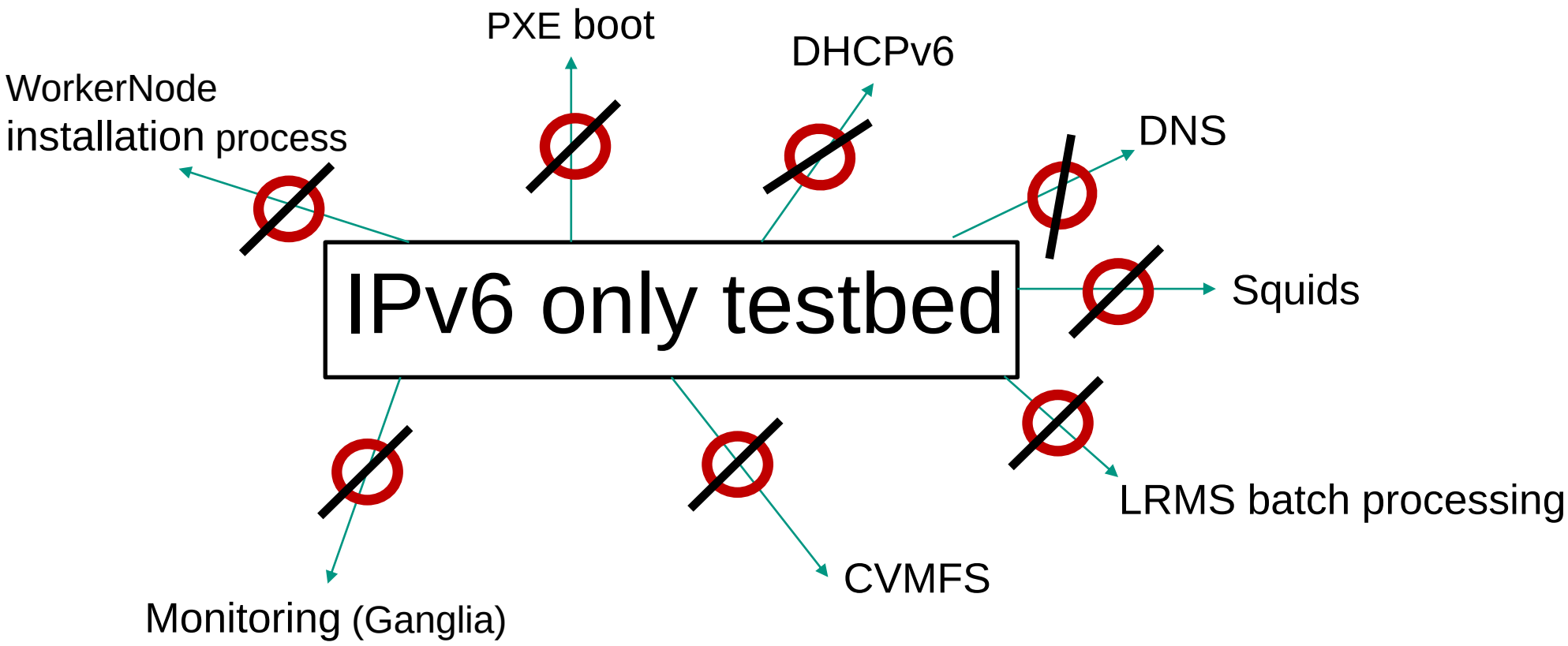
Dedicated 10/100/400Gbps links
16 Tier1s + 1 Tier0
12 countries in 3 continents
2.1 Tbps to the Tier0

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 Textbed

HEPiX- IPv6 working group asking for IPv6 only testbed



DE-KIT – workernode migration towards IPv6

Karlsruhe Institute of Technology

Pro-active IPv6 Monitoring at DE-KIT

detailed monitoring at DE-KIT (GridKa)

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

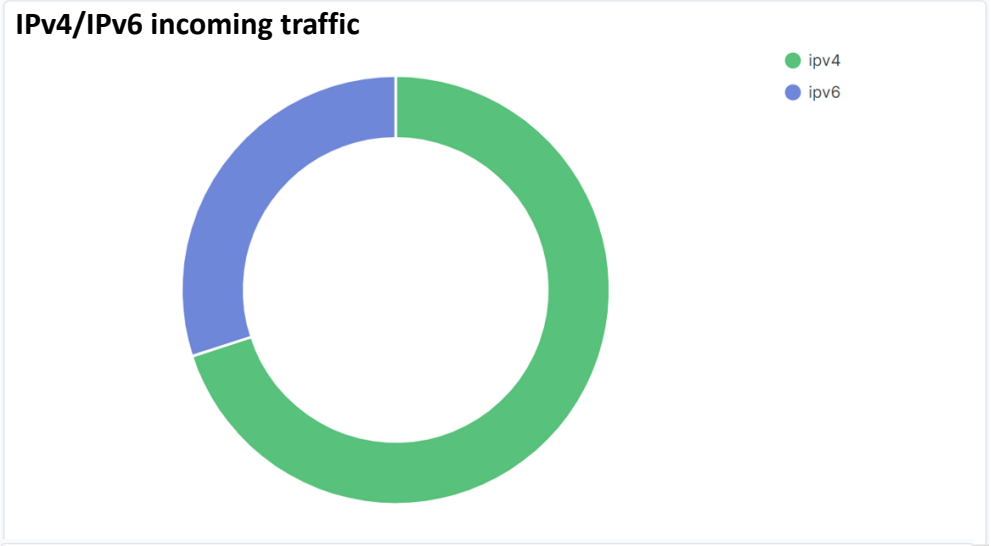
monitoring of process intercomunication at DE-KIT (GridKa)

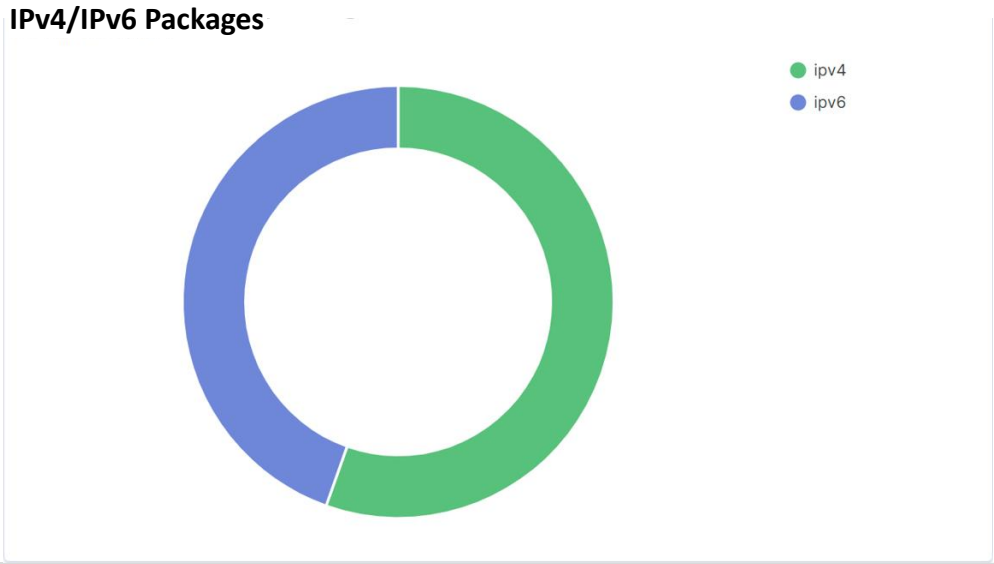
- with packetbeat collecting network data
- logstash pushing the data to opensearch (former elastic search) for storing the data
- kibana for visualizing
  (no opensearch – only easy search requests)
  - started with a small set of workernodes (storing the data „longterm“ → ~ 6 days)
  - while enlarging 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)

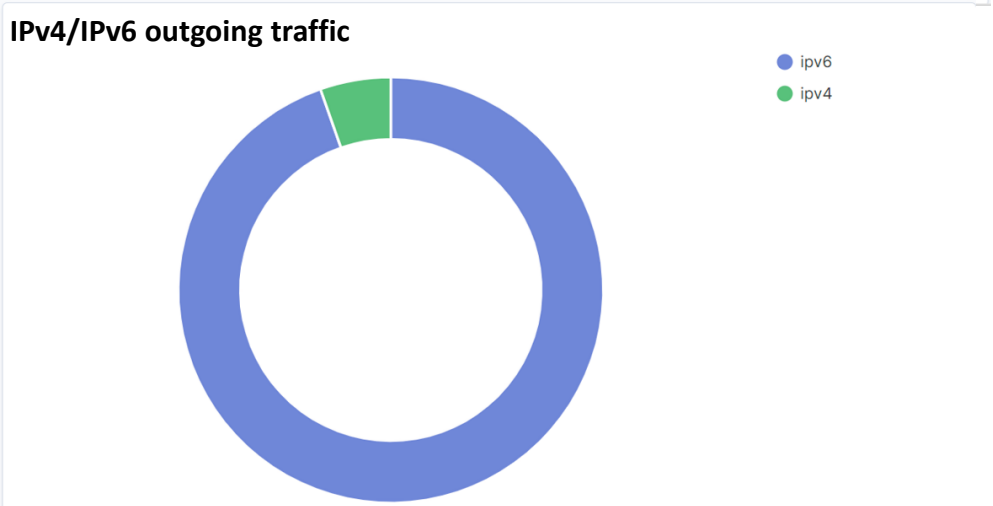
- Identify IPv4 protocol usage

snapshot of a dashbord

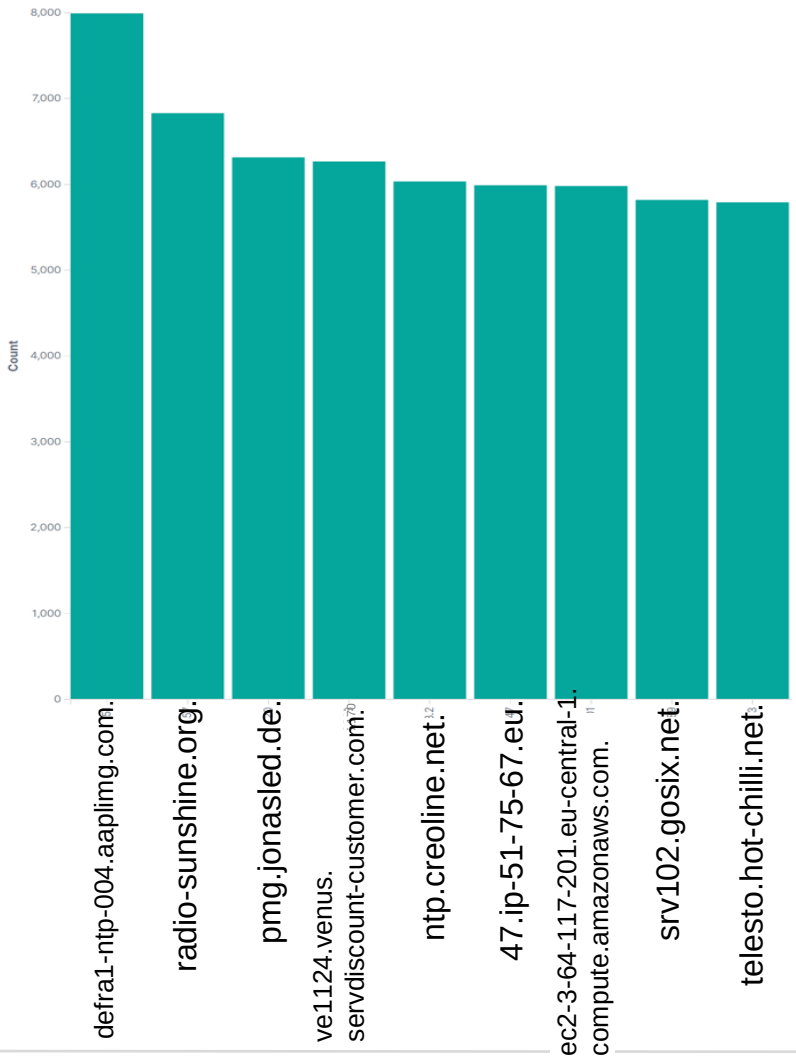
at 08.09.22
all worker nodes already dual-stack deployed

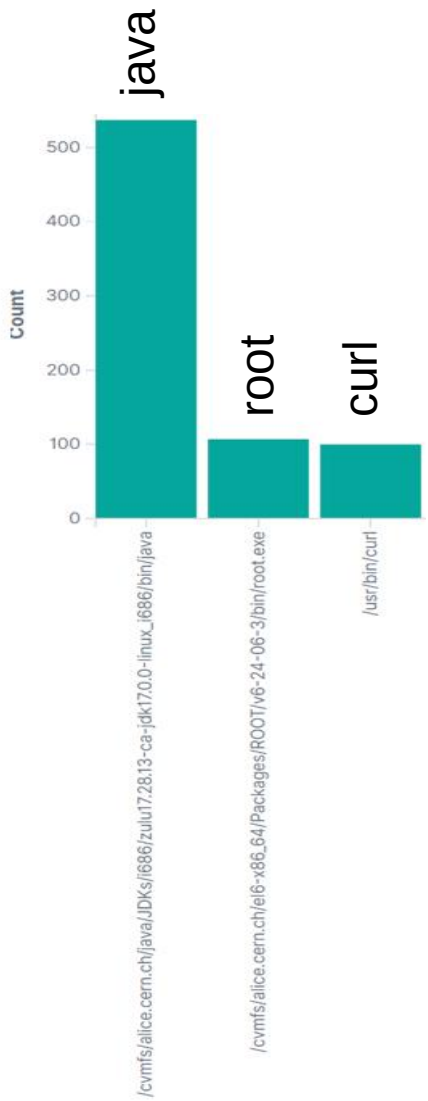






NTP ?





- 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

S O L V E D

- NTP.ORG
  → returns sometimes funny addresses

Logstash → is now IPv6

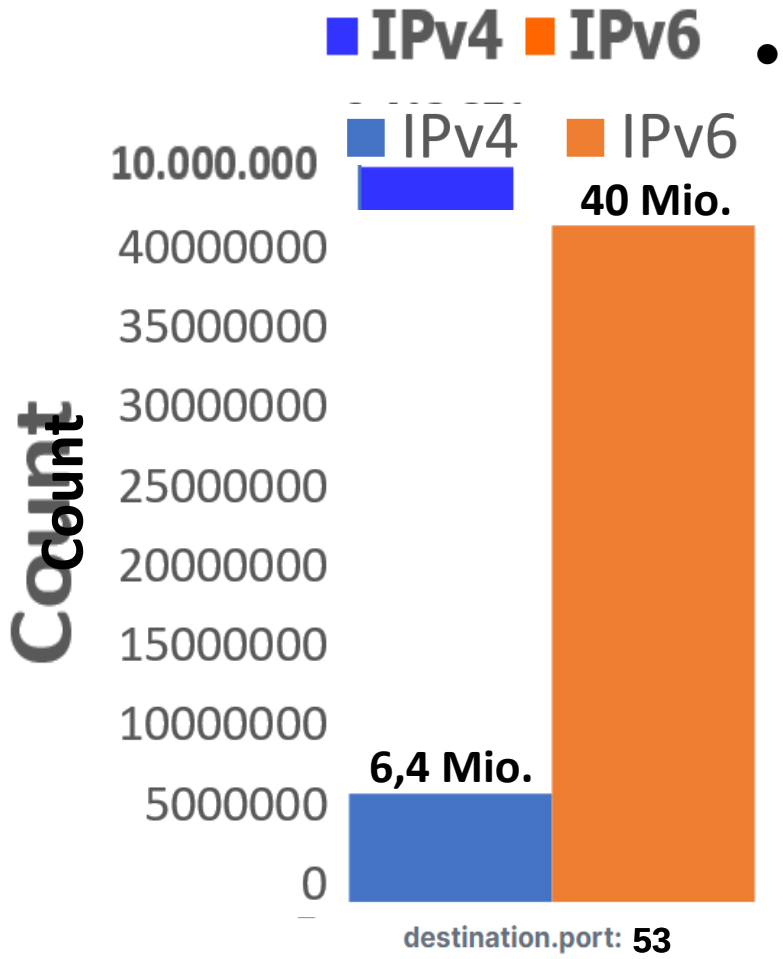
Karlsruhe Institute of Technology

Logstash (port 5047) → dual-stack deeployed

statistic:

28-07-2022 → IPv4 385k – IPv6 1,41 M
23-10-2022 → IPv4 476k – IPv6 1,39 M
23-12-2022 → IPv4 227k – IPv6 450k
30-10-2023 → IPv4 906k – IPv6 864k
04-04-2024 → IPv4 1,0 M – IPv6 1,0 M

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**

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

SQUIDS migrated all to dual-stack

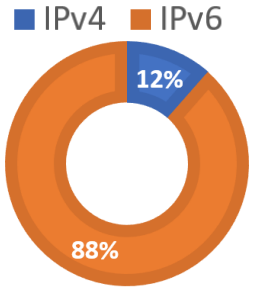
During the second half of 2022 all SQUIDS migrated to dual-stack deployment

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>
```

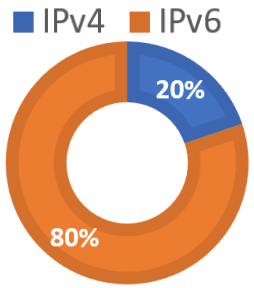
26-07-2022

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



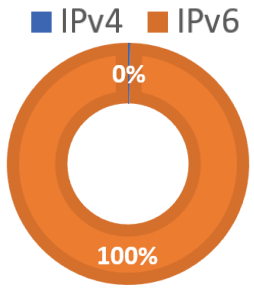
23-10-2022

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



25-10-2023

IPv4 : 64 k IPv6: 22 mio.



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

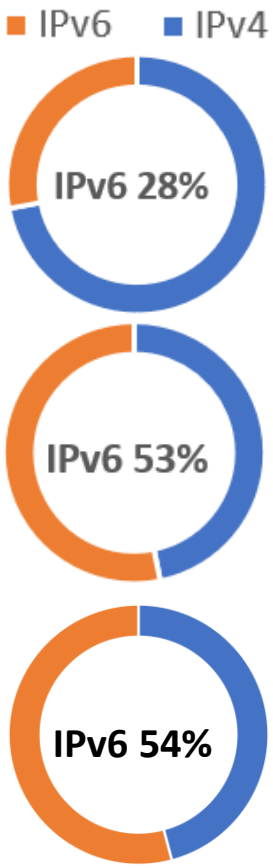
- LRMS (**Local Resource Management System**) HTCondor at GridKa (all dual-stack and set to **prefer** the protocoll **IPv6** (Port 9618/9)
  - percentage increased toward IPv6 at 28-06-2022→ IPv4: **895k** to IPv6: **255k**
  - 1,2% IPv4 28-07-2022 → IPv4: 27k, IPv6: 2,17 mio.
  - **11%** IPv4 02-01-2023 → IPv4: 287k, IPv6: 2,28 mio.
  - **4,3%** IPv4 04-40-2023 → IPv4: 767k, IPv6: 17,8 mio.

Less then 50% of IPv4 is internal traffic

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

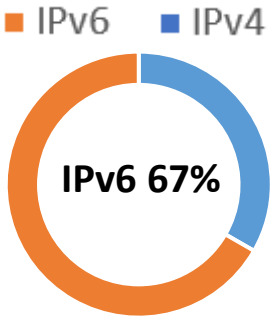
A view statistics

- 15-04-2022:
  - IPv4: → 80 mio.
  - IPv6: → 31 mio.
- 26-07-2022:
  - Ipv4 → 44 mio.
  - Ipv6 → 50 mio.
- 23-10-2023:
  - IPv4 → 122 mio.
  - IPv6 → 144 mio.



04-04-2024:

- IPv4: → 99,4 mio.
- IPv6: → 210,6 mio.



(packets in 24 hours)

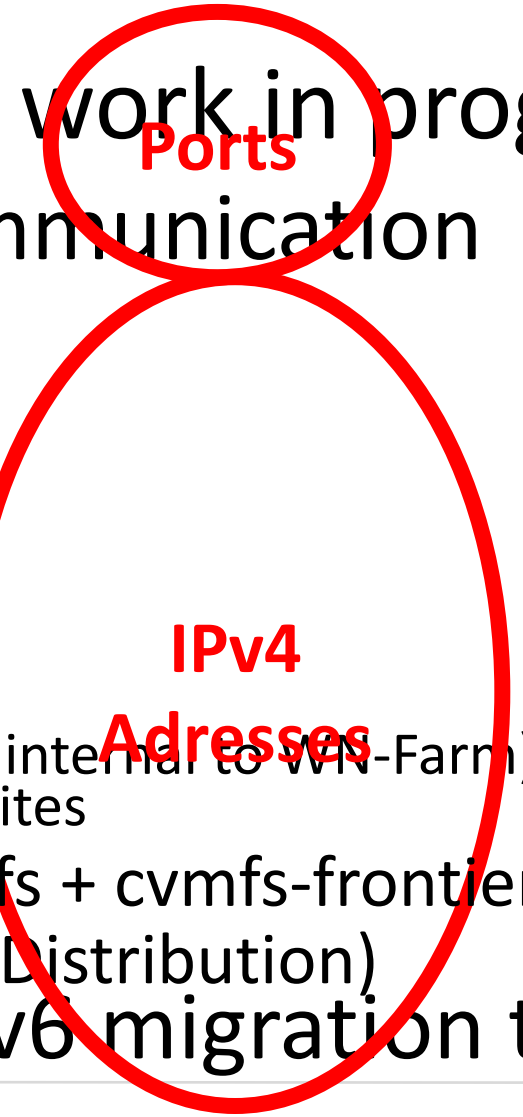
# of WorkerNodes included in the statistic expanded

IPv6 @ LHCB

- DE-KIT WN farm dual-stack deployed
- representative Alexey Zhelezov
  and Raja Nandakumar assured that LHCB is IPv6-only ready
- WN subset for LHCB IPv6 only deployed
  - undeploying IPv4
- Jobs running without ony failure

Conclusion

- migration of Rackmanager – work in progress
- Narrow down the still IPv4 communication
  - packet monitoring configured
    - to list all unhandled IPv4 packets
      - 8884 – Alice: operation report
      - 2049 – NFS
      - 1094 – XrootD
      - 961[89] – LRMS (less than 50% only internal to WN-Farm) be aware → dependency on other sites
  - IPv6 linux internals (resolve.conf, cvmfs + cvmfs-frontier, …)
  - PXE – Boot + DHCPv6 (first boot addr. Distribution)
- Identify the next service for IPv6 migration tasks



Karlsruhe Institute of Technology

Karlsruhe Institute of Technology

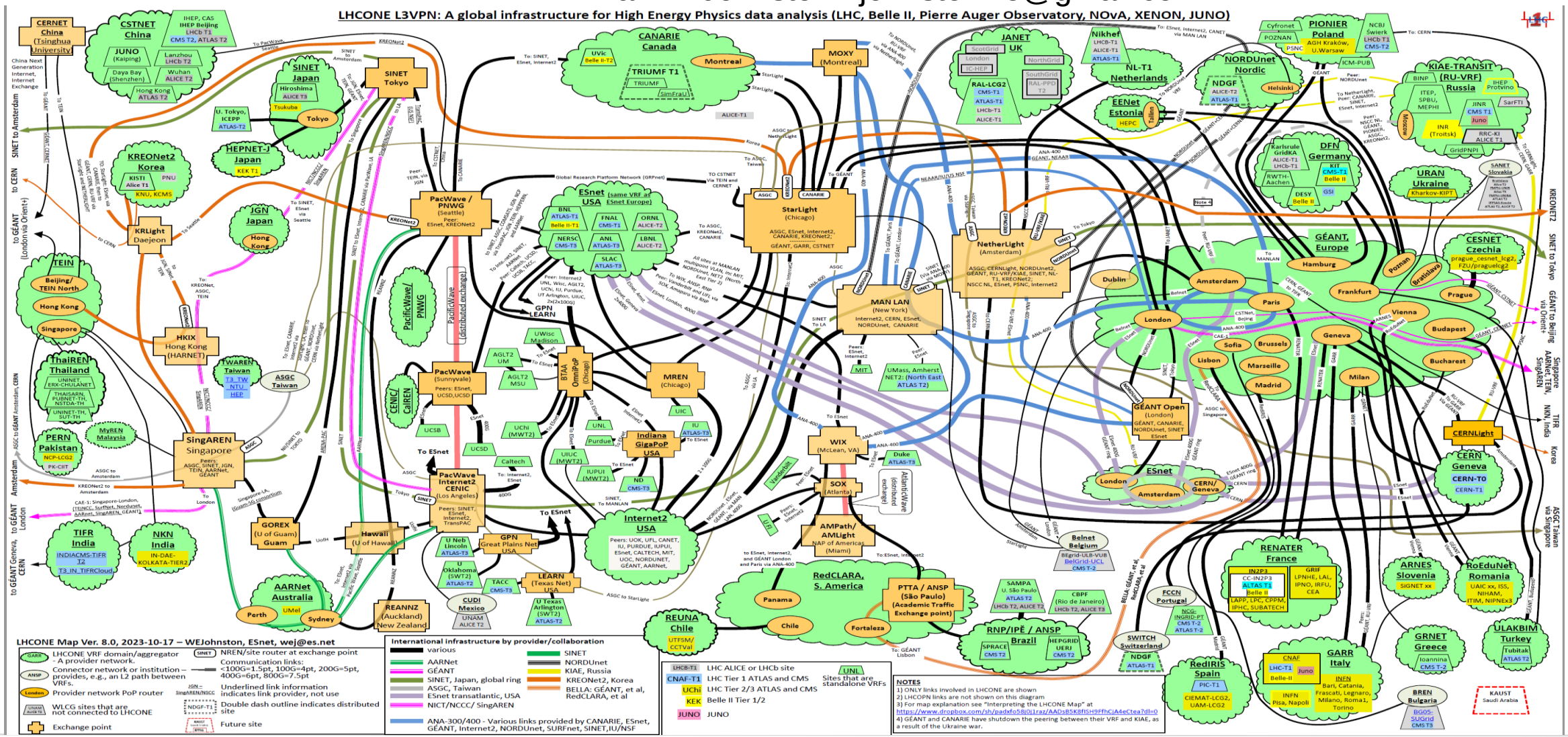
Thx for your attention

backup Slides

LHCONE

William E Johnston <johnstonwe@gmail.com>

Karlsruhe Institute of Technology



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)

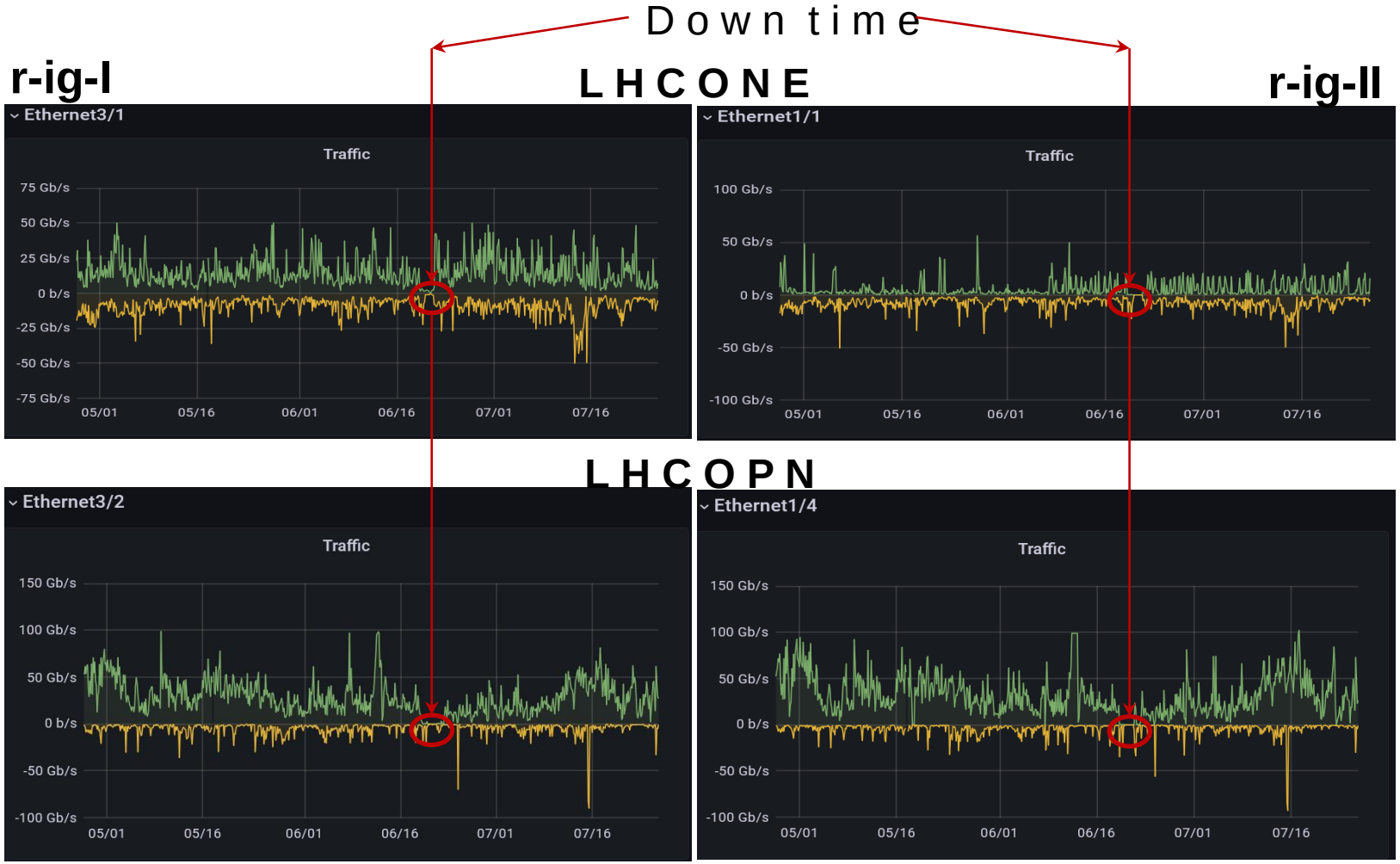
dCache upgrade to 7.2.15

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-l (DE-KIT Border Router):
left two Interfaces
- Ethernet 3/1 (Internet + LHCONE) +
- Ethernet 3/2 (LHCOPN)

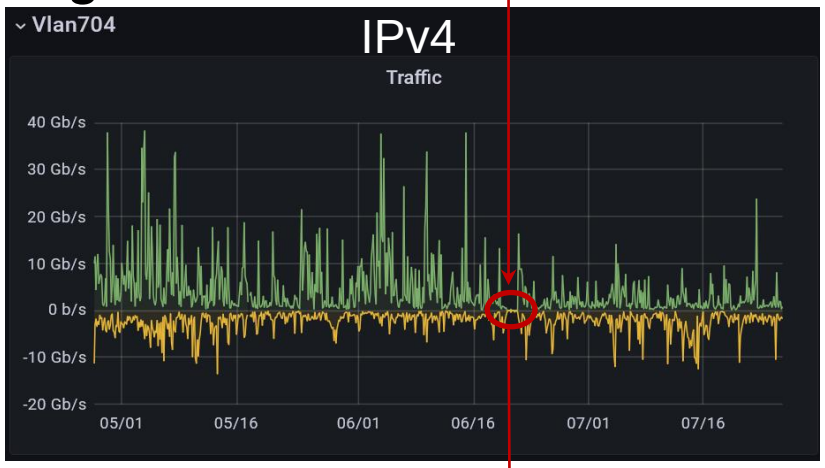
r-ig-ll (DE-KIT second Border Router):
right two Interfaces
- Ethernet 1/1 (Internet + LHCONE) +
- Ethernet 1/4 (LHCOPN)

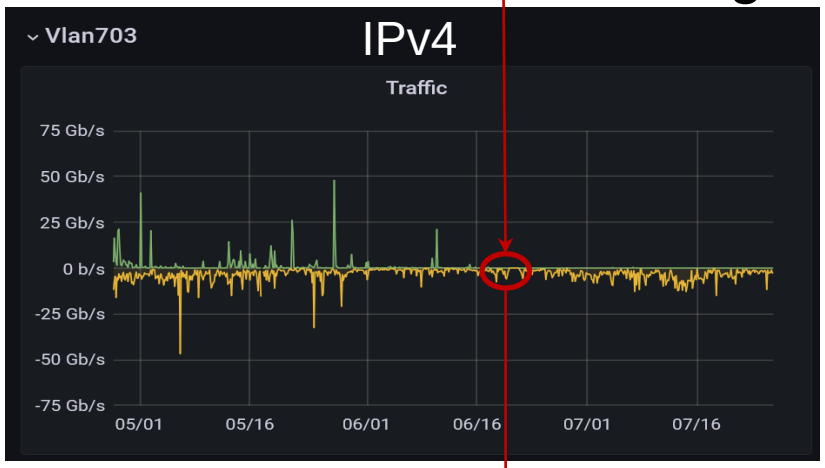
LHCONE IPv4 / IPv6
transfer pattern after downtime

Karlsruhe Institute of Technology

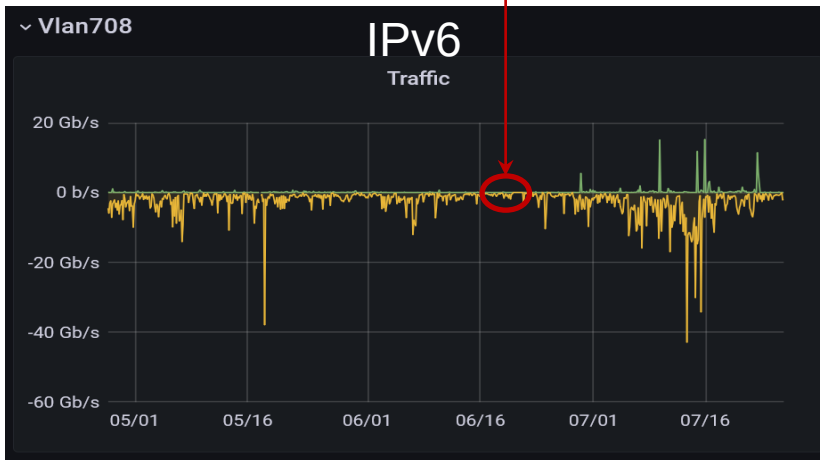
Down time

r-ig-I





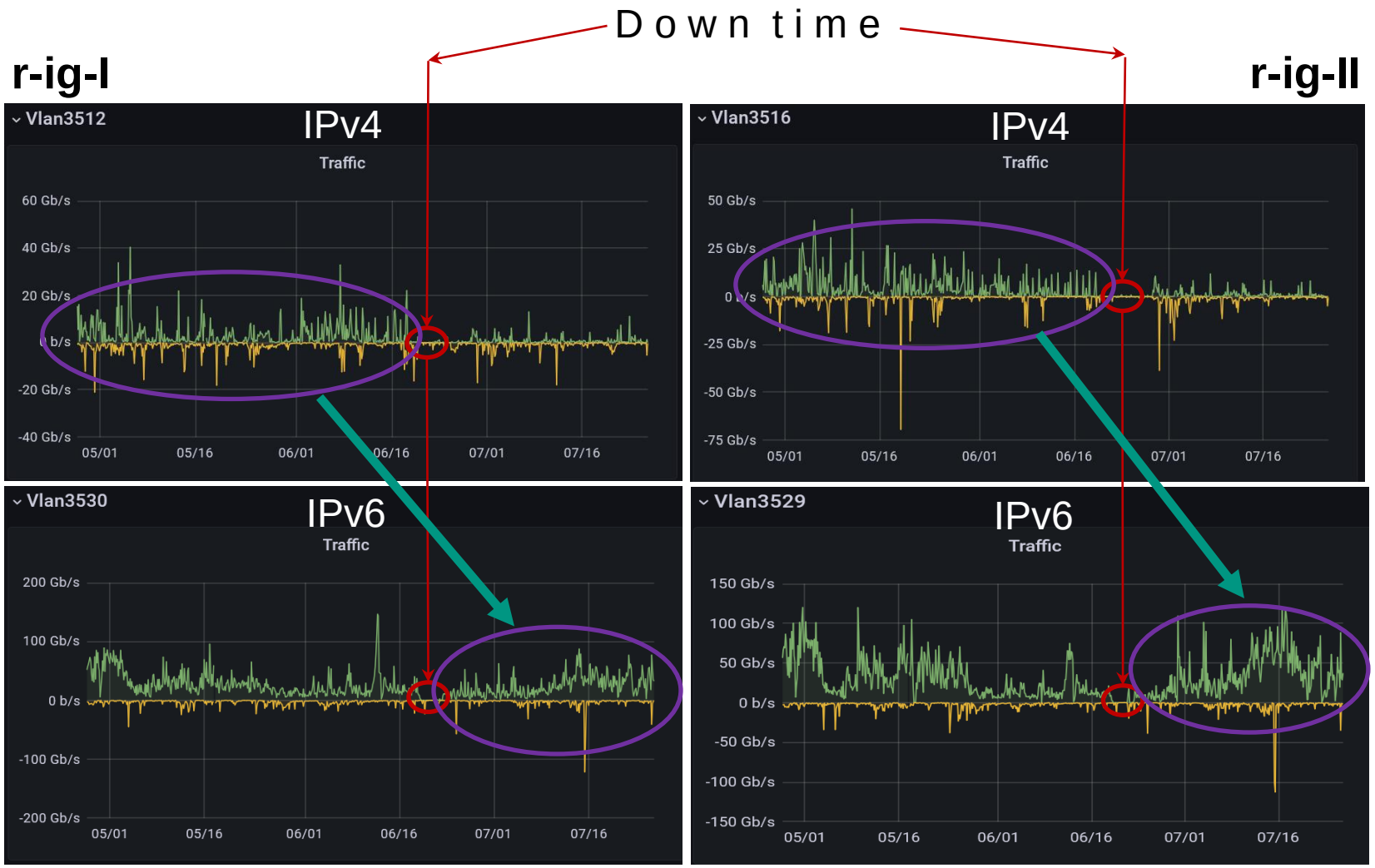
r-ig-II



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

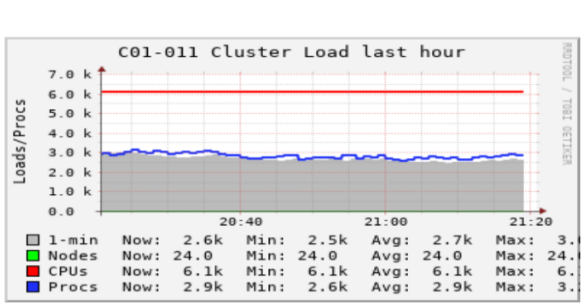
Karlsruhe Institute of Technology

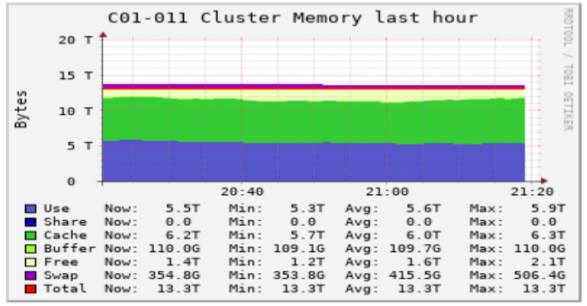


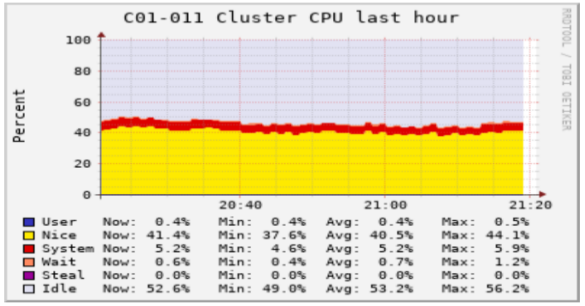
Graph over 90 days
Traffic of LHCOPN
moved from the IPv4 vlans
after the downtime to the IPv6 Vlans

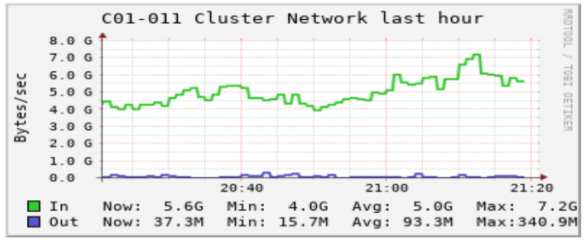
Monitoring

G A N G L I A

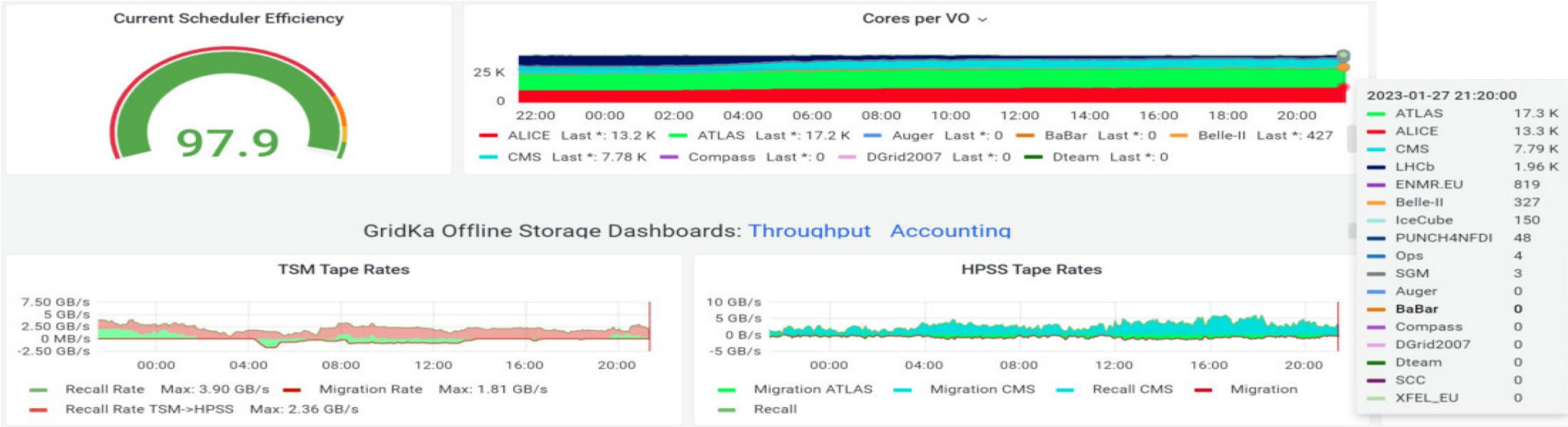








- Migration of Ganglia to IPv6 will not persuit
- Ganglia will be replaced by opensearch, kibana and grafana



administatrative 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)