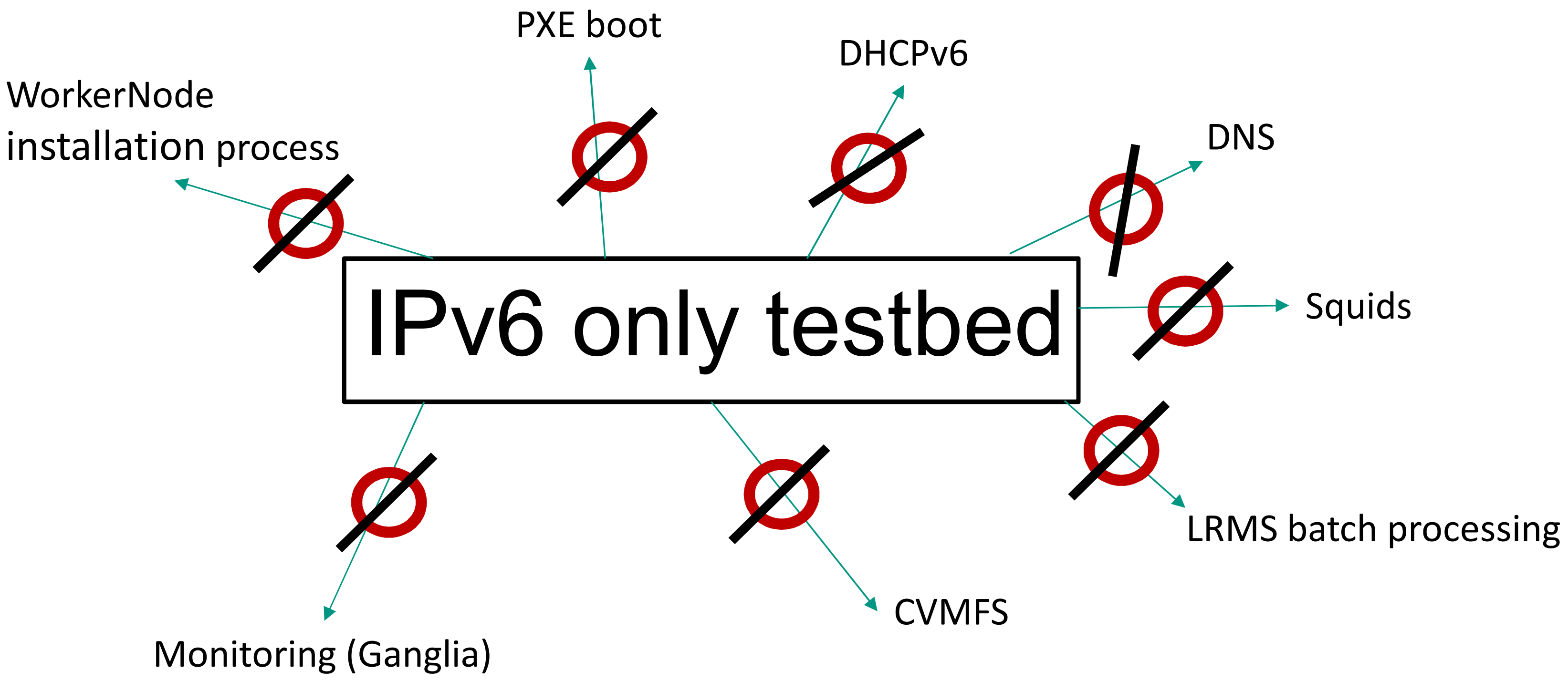
IPv4 to IPv6 Worker Node migration in WLCG

Bruno Hoeft, Matthias Schnepf, Max Fischer, Andreas Petzold
Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen
{first.familyname}@kit.edu

building IPv6 testbed

HEPiX- IPv6 working group asking for IPv6 only testbed



DE-KIT – workernode farm migration towards IPv6

for identifying migration tasks a
– Pro-active Monitoring at DE-KIT – is deployed

monitor all communications between WorkerNodes (WN)
and
- administration
- job submission
- storage
- …

monitoring of process inter-comunication at DE-KIT (GridKa)

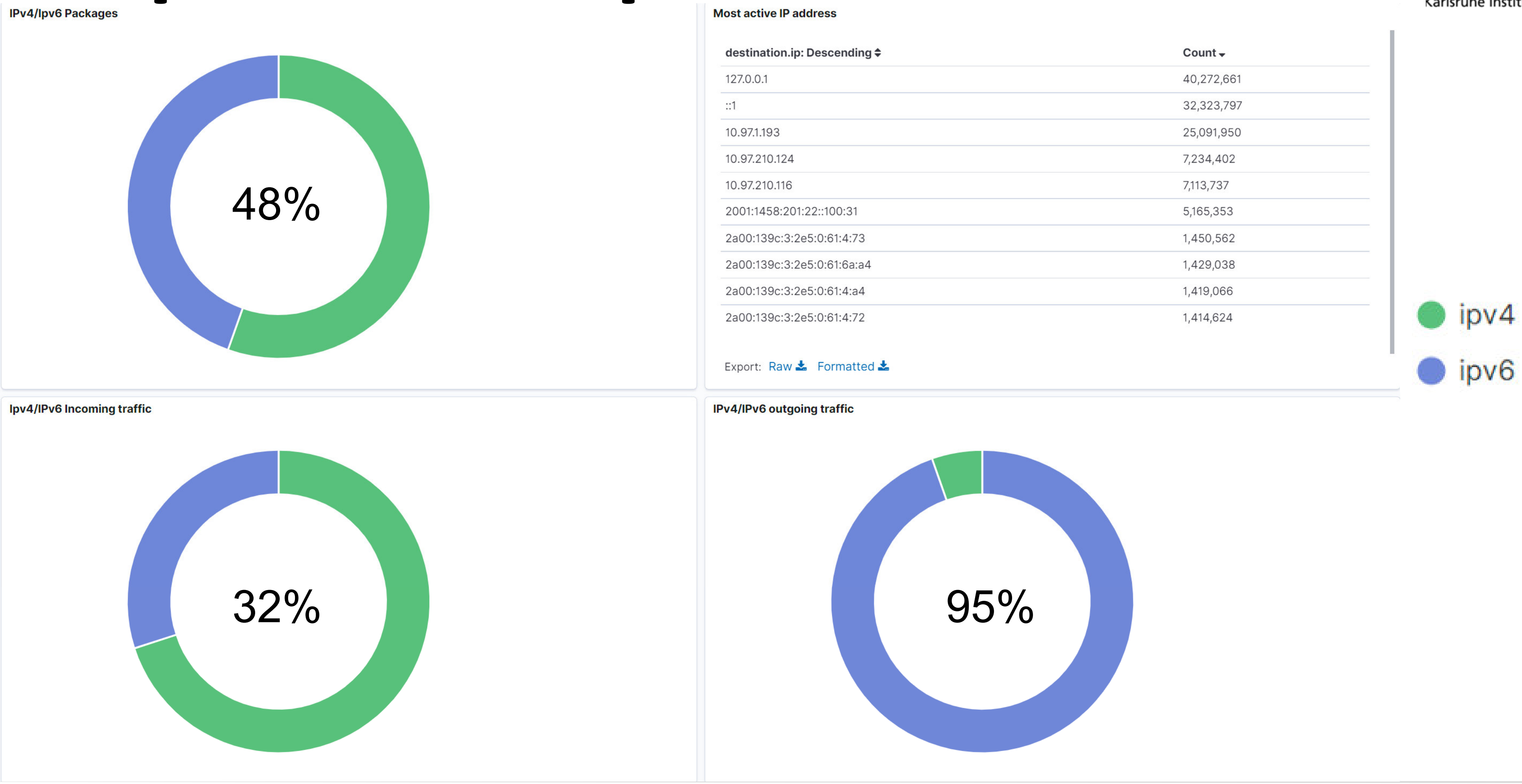
- packetbeat is collecting the network data
- logstach is pushing the data to opensearch (former elastic search) for storing the data
- kibana for visualizing

  → the monitoring started with a small set of workernodes (storing the data „longterm“ → ~ 6 weeks)

  → while enlarging the set of workernodes gradually 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 at Sept. 2022

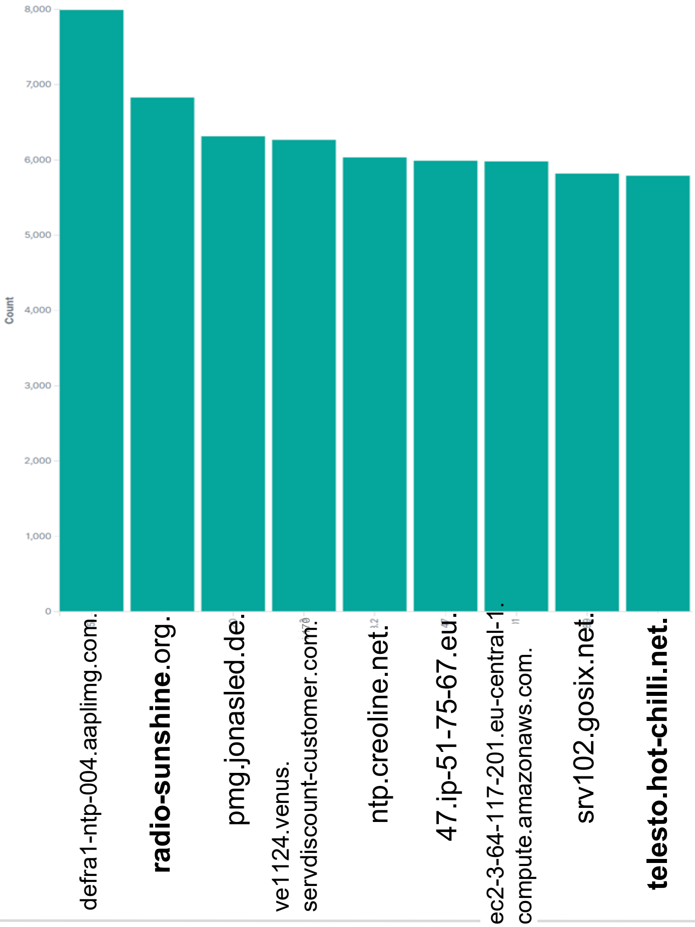


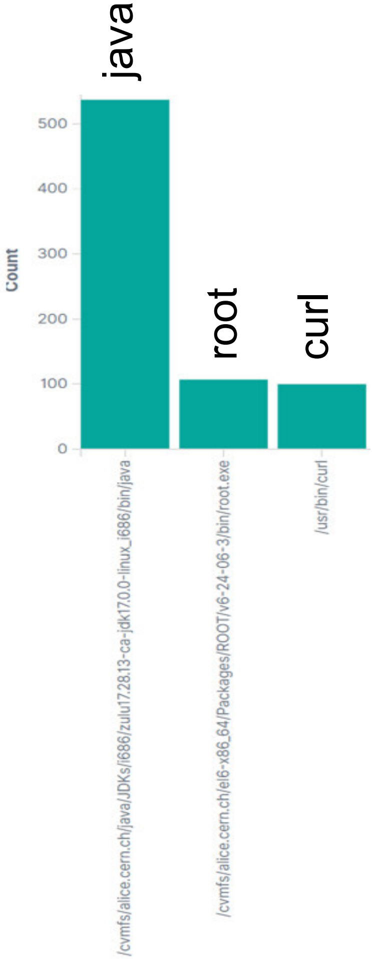
Most active IP address

destination.ip: Descending ⬍	Count ▾
127.0.0.1	40,272,661
::1	32,323,797
10.97.1.193	25,091,950
10.97.210.124	7,234,402
10.97.210.116	7,113,737
2001:1458:201:22::100:31	5,165,353
2a00:139c:3:2e5:0:61:4:73	1,450,562
2a00:139c:3:2e5:0:61:6a:a4	1,429,038
2a00:139c:3:2e5:0:61:4:a4	1,419,066
2a00:139c:3:2e5:0:61:4:72	1,414,624

Export: Raw ⬇ Formatted ⬇

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 to certain subnets 10.1.12.0/24 and 10.1.18.0/24
    → futher investigation 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

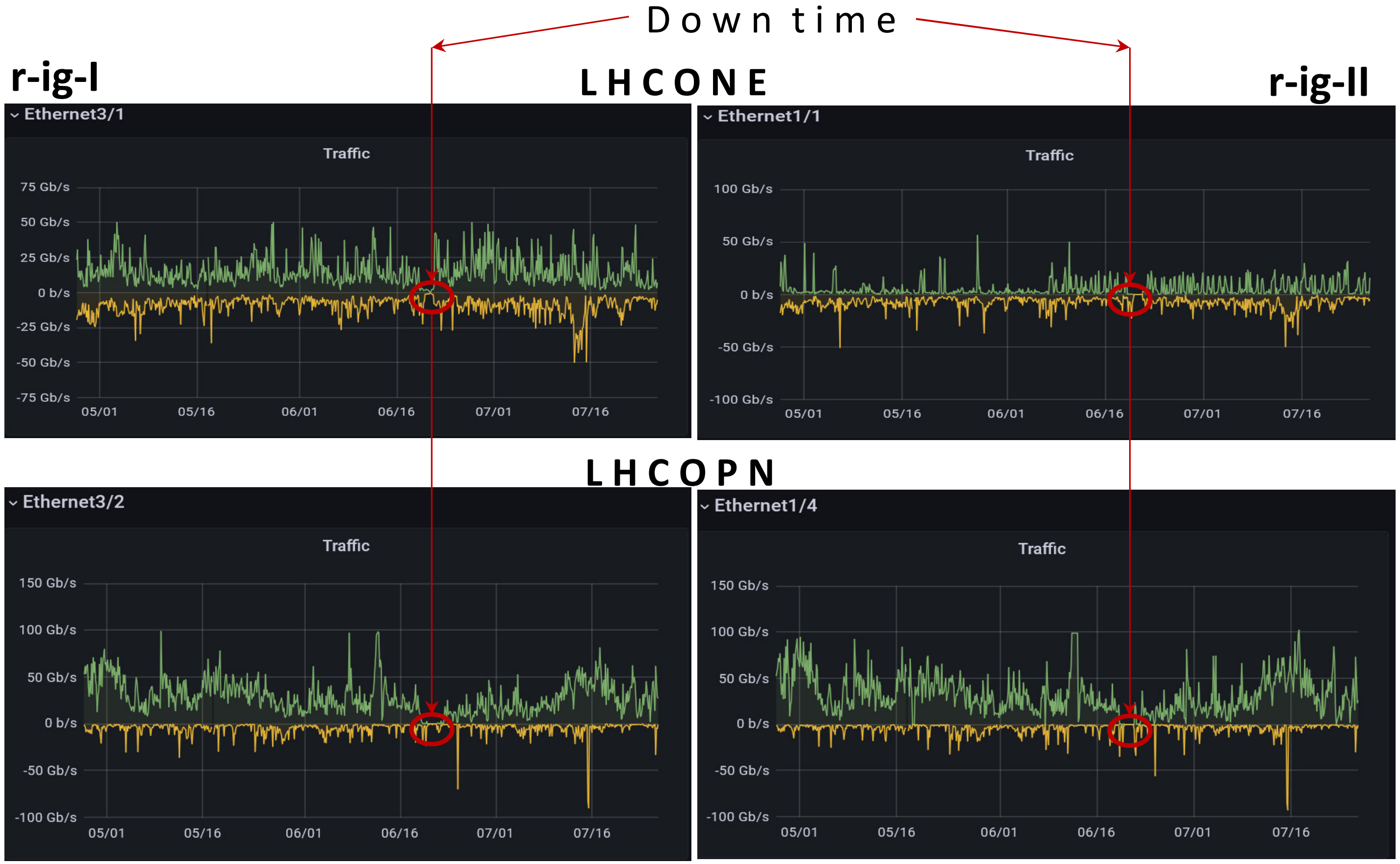
dCache upgrade to 7.2.15

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



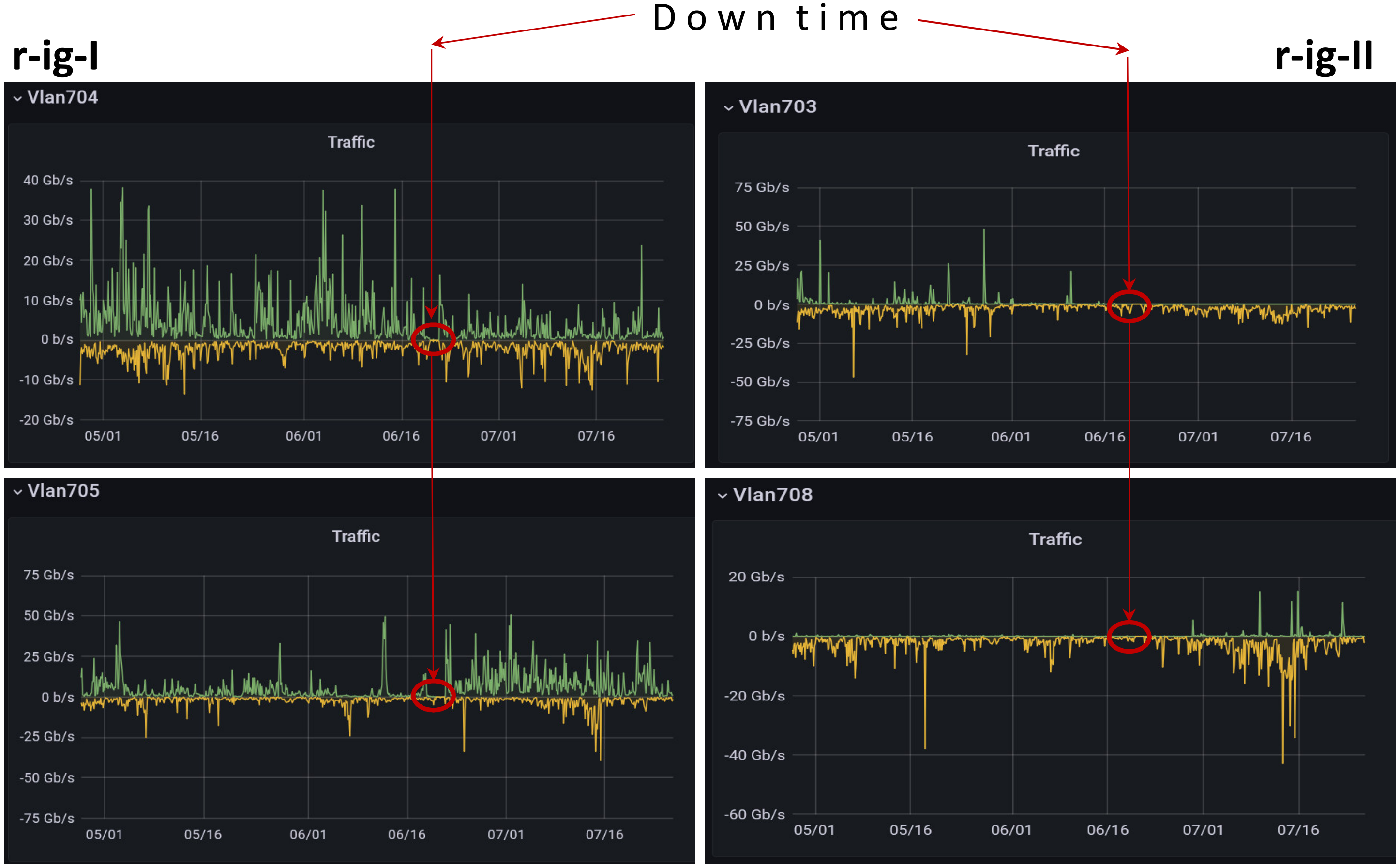
r-ig-I (DE-KIT border router):
left two Interfaces

- 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



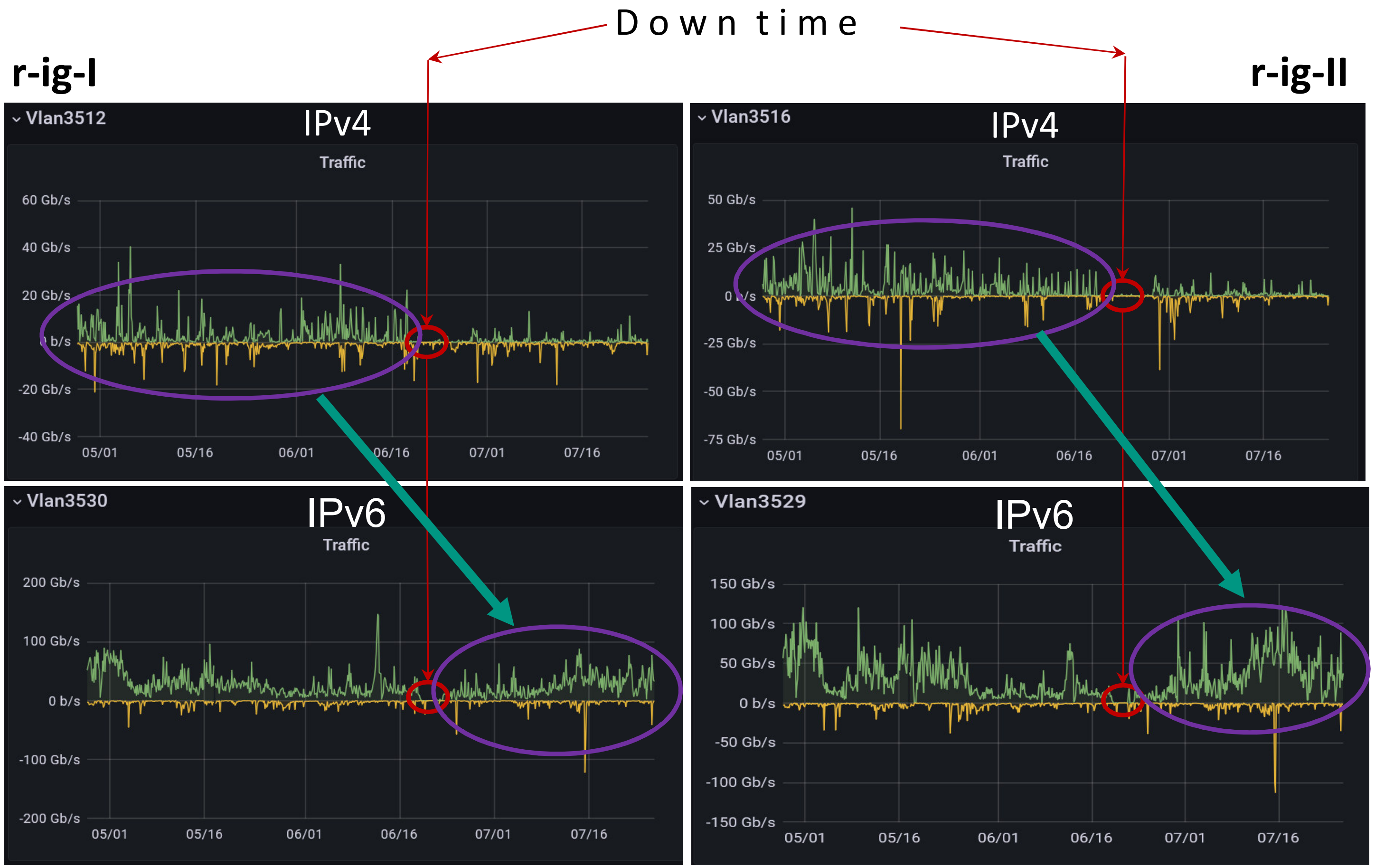
Down time

r-ig-I

r-ig-II

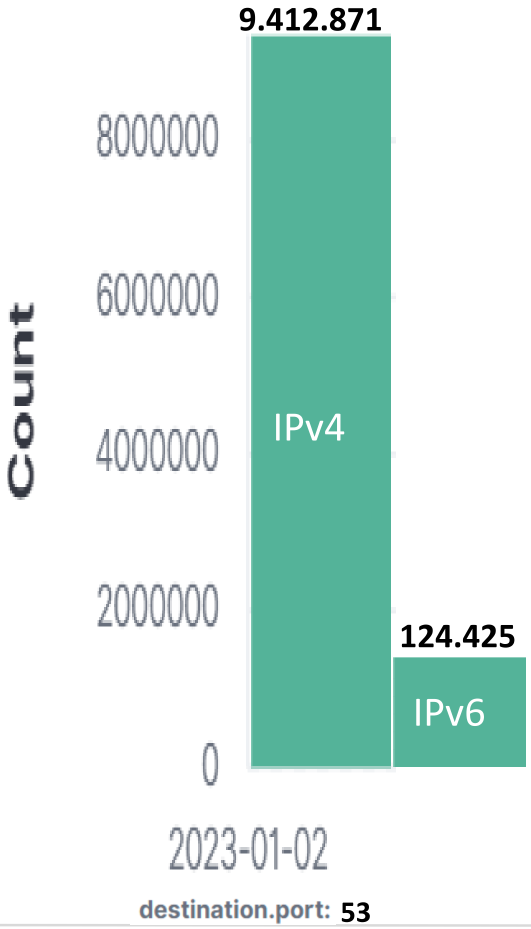
graph over 90 days
traffic of LHCONE
moved partioly from the IPv4 vlans
after the downtime to the IPv6 Vlans

LHCOPN IPv4 / IPv6 transfer pattern after downtime



graph over 90 days
traffic of LHCOPN
moved from the IPv4 vlans
after the downtime to the IPv6 Vlans

closer look at DNS



- GridKa DNS(Port 53):
  - IPv4 only count : 9,412,871 (24 hours)
  - DNS (Bind) server and WN are 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 10.privat-address
        ...

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

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**

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

- statistic:

• July :	IPv4 : 1,25 mio. IPv6: 9,6 mio. (tcp port 3128, 3401)
• **October :**	**IPv4 : 4,44 mio. IPv6: 18 mio.** (tcp port 3128, 3401)
• December :	IPv4 : 1,47 mio. IPv6 : 2,3 mio**.** (tcp port 3128, 3401)

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
  - Ratio increased toward IPv6 at 20220628→ IPv4: **895k** to IPv6: **255k**
  - Ratio today 20220728 → IPv4: 27k, IPv6: 2,17 mio. (per 24 hour)
  - **Ratio today 20221023 → IPv4: 10k, IPv6: 3,38 mio.** (per 24 hour)
  - Ratio today 20230102 → IPv4: 287k, IPv6: 2,28 mio. (per 24 hour)

Less then 20% of IPv4 is internal traffic

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

Logstash → is now IPv6

Logstash → dual-stack deployed

Ratio 78% IPv6 20220728 → IPv4 385k – IPv6 1,41M
Ratio 74% IPv6 20221023 → IPv4 476k – IPv6 1,39M
Ratio 66% IPv6 today 20221223 → IPv4 227k – IPv6 450k

migration still in progress

administatrative services

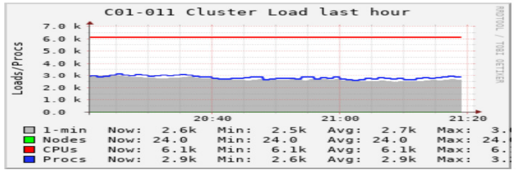
- at each rack is a Rack Manager deployed:
  - starting in 2001 with private IPv4 only
  - migration process initiated (but still in progress)
    → enable dual-stack (AAAA)
    - NTP
    - rsyslog (→ migration → still pending (port 514))
    - monitoring (GmonD → Ganglia client)
    - DHCP (→ migration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

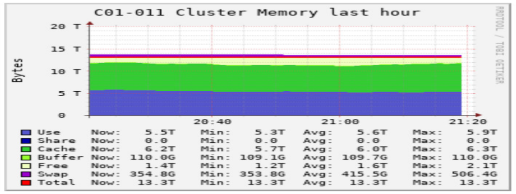
WN – deployment process

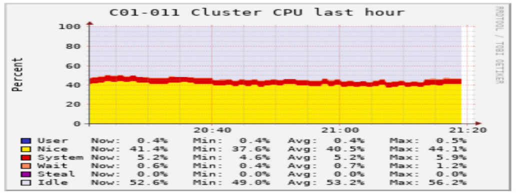
- Redhat Satellite Server (foreman)
  - used for management of most GridKa hosts:
    - manages redhat subscriptions
    - controlls kickstart installations (DHCP / PXE)
    - provides yum repos
    - provides CA (certificate authority) and ENC (encryptor) functionalities for puppet
  - uses modular 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)

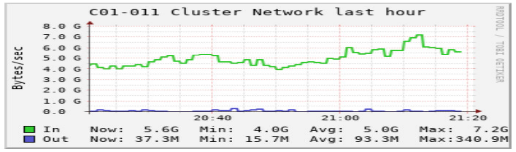
Monitoring

G A N G L I A

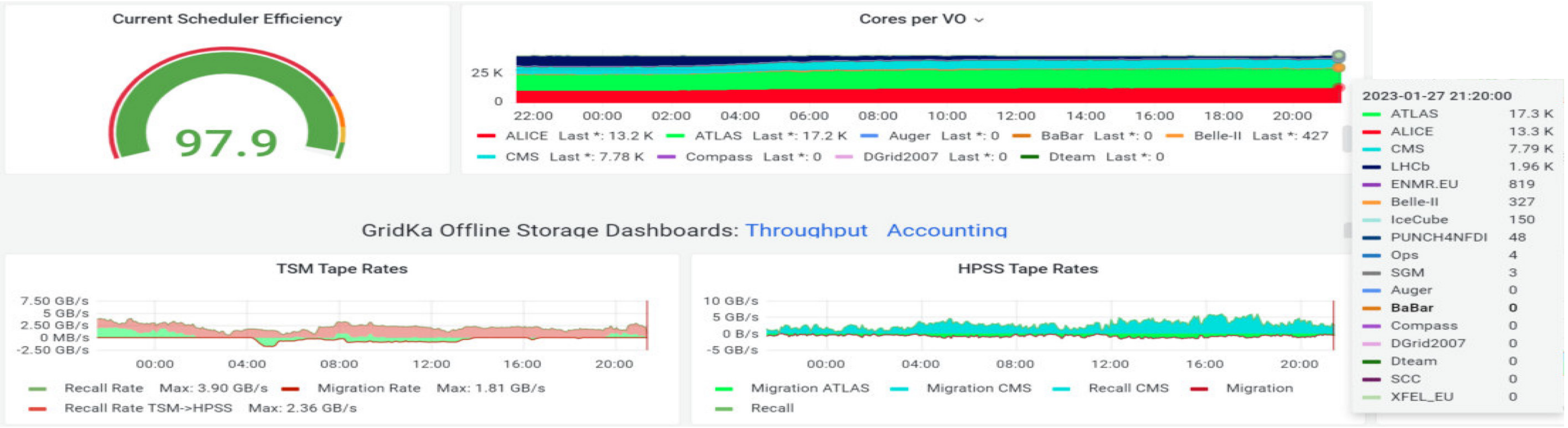








- Ganglia will not migrate to IPv6
- Ganglia will be replaced by opensearch, kibana and grafana

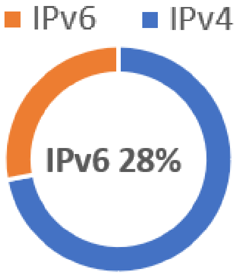


2023-01-27 21:20:00	
ATLAS	17.3 K
ALICE	13.3 K
CMS	7.79 K
LHCb	1.96 K
ENMR.EU	819
Belle-II	327
IceCube	150
PUNCH4NFDI	48
Ops	4
SGM	3
Auger	0
BaBar	**0**
Compass	0
DGrid2007	0
Dteam	0
SCC	0
XFEL_EU	0

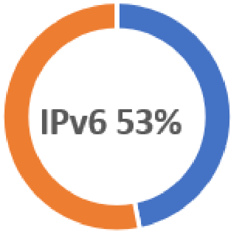
a few statistics

Karlsruhe Institute of Technology

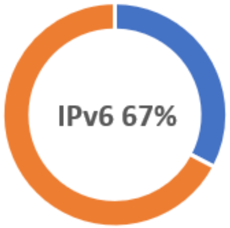
- 20220415:
  - IPv4: → 80 Mio
  - IPv6: → 31 Mio



- 20220726:
  - Ipv4 → 44 Mio
  - Ipv6 → 50 Mio

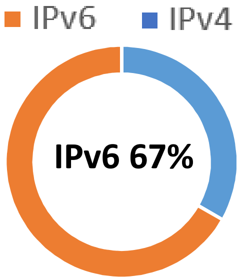


- 20221023:
  - IPv4 → 69 Mio
  - IPv6 → 142 Mio



20221220:

- IPv4: → 42 Mio
- IPv6: → 86 Mio



(packets in 24 hours)

of **Worker**Nodes included in the statistic expanded

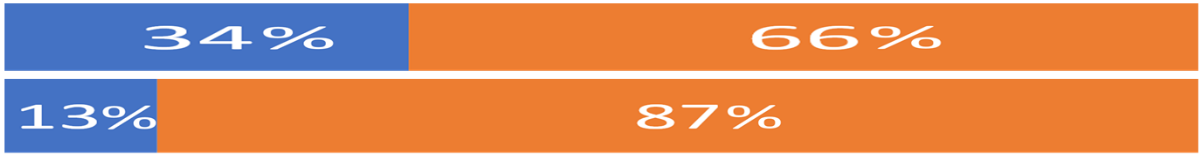
details of ALICE VOBoxes:

- 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)

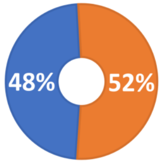
Japan KEK Belle-2

sites with Dual Stack Storage: 34%

sites with Dual Stack WN: 13%



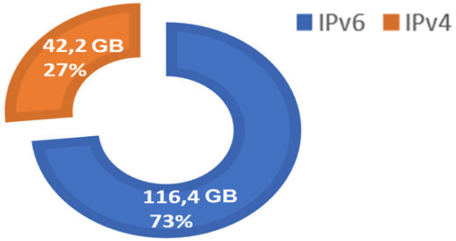
Sites within LHCONE: 48%



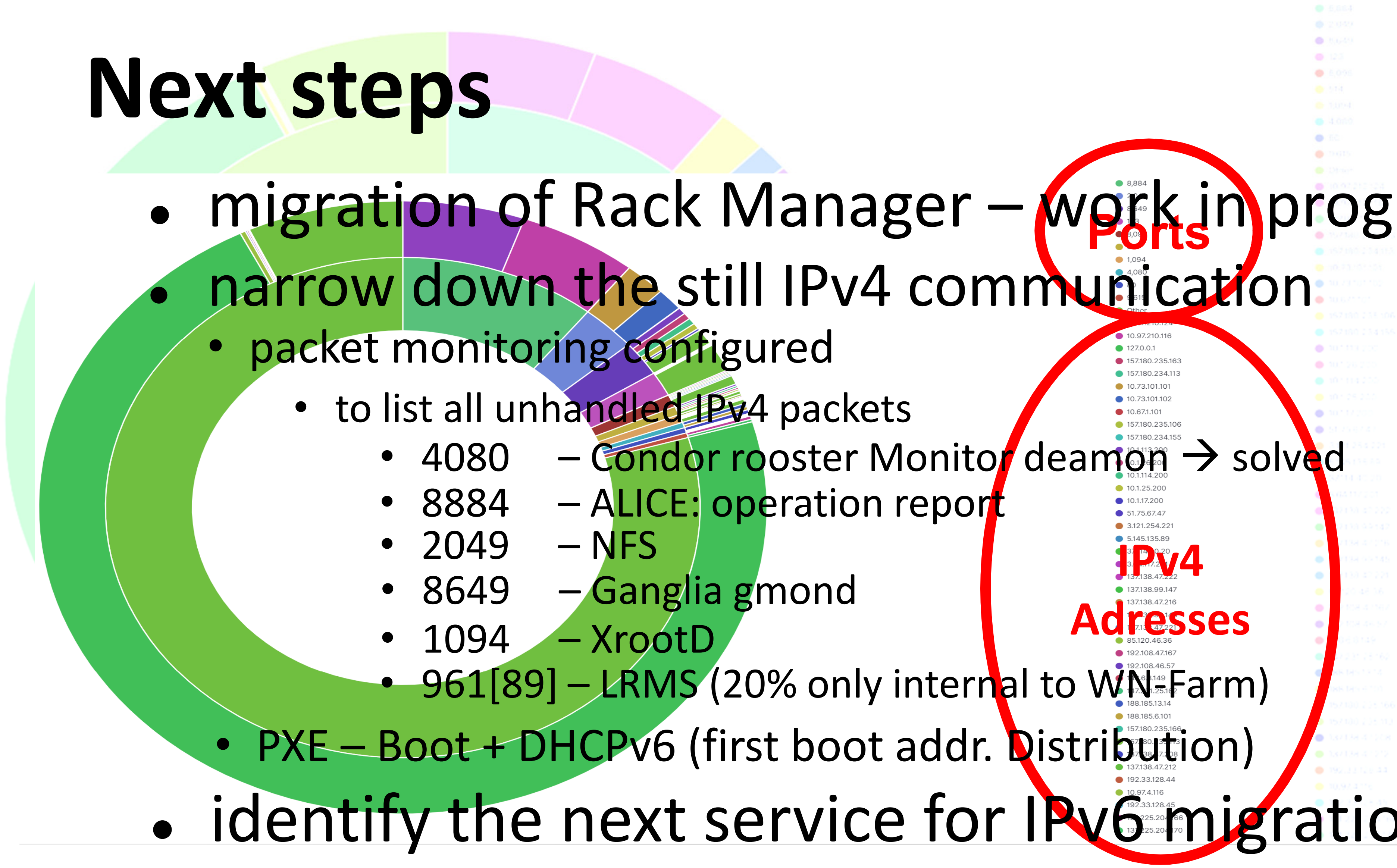
Sites at General IP: 52%

detector DB status data (non-/operational) → Ipv4 only

Snapshot (24 hours) End of Jan. 2023



Next steps



- migration of Rack Manager – 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 (20% only internal to WN-Farm)
  - PXE – Boot + DHCPv6 (first boot addr. Distribution)
- identify the next service for IPv6 migration tasks

Thx for your attention