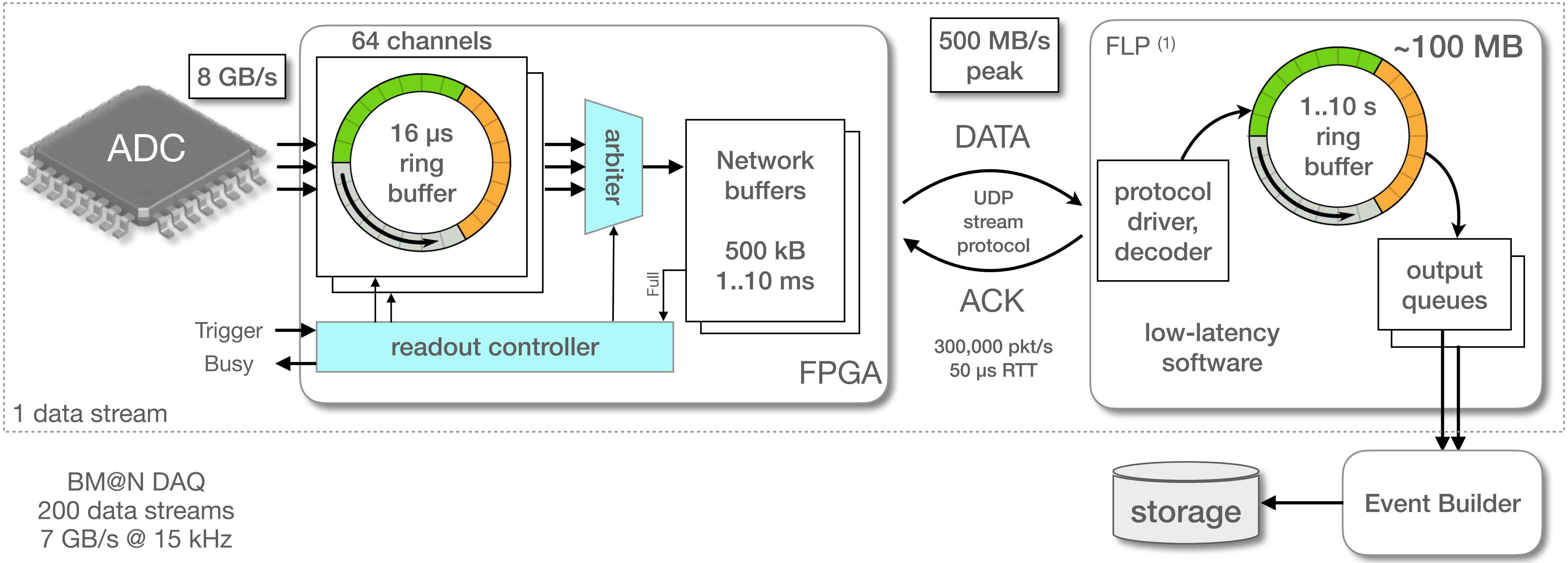
BM@N Data Acquisition IT Infrastructure

BM@N Experiment at the NICA Facility
10th Collaboration Meeting
St Petersburg, May 14 – 19, 2023

ILIA SLEPNEV, JINR

Data Acquisition

Data transfer from detector to storage system



BM@N DAQ
200 data streams
7 GB/s @ 15 kHz

Decouple fast microsecond-scale synchronous acquisition from slow second-scale software processing

Input queue full condition suspends data taking process causing missed physical events
Possible solutions

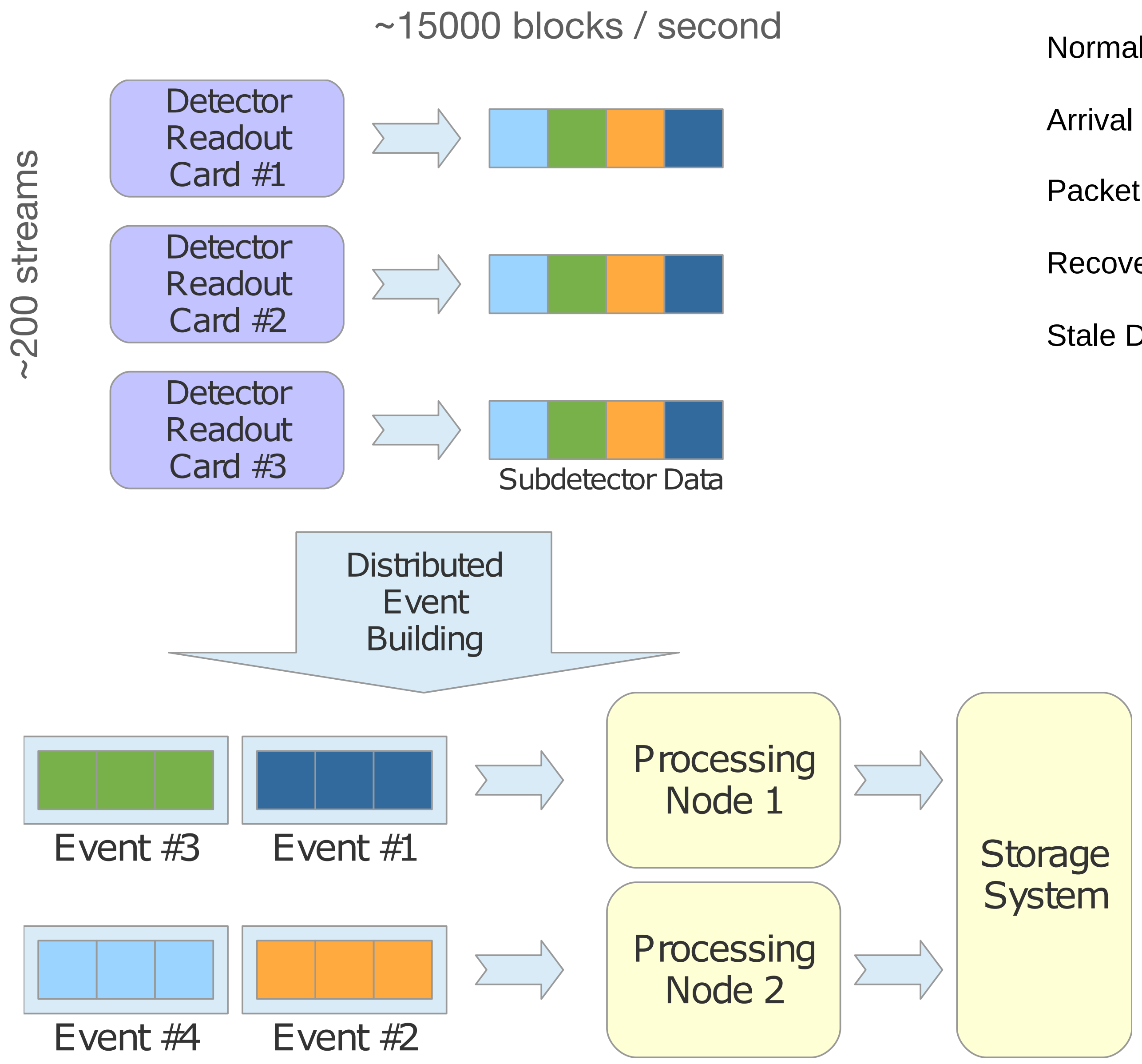
- discrete on-board DRAM memory chips
- custom PCIe data processing cards in FLP
- Ethernet based readout with commodity hardware

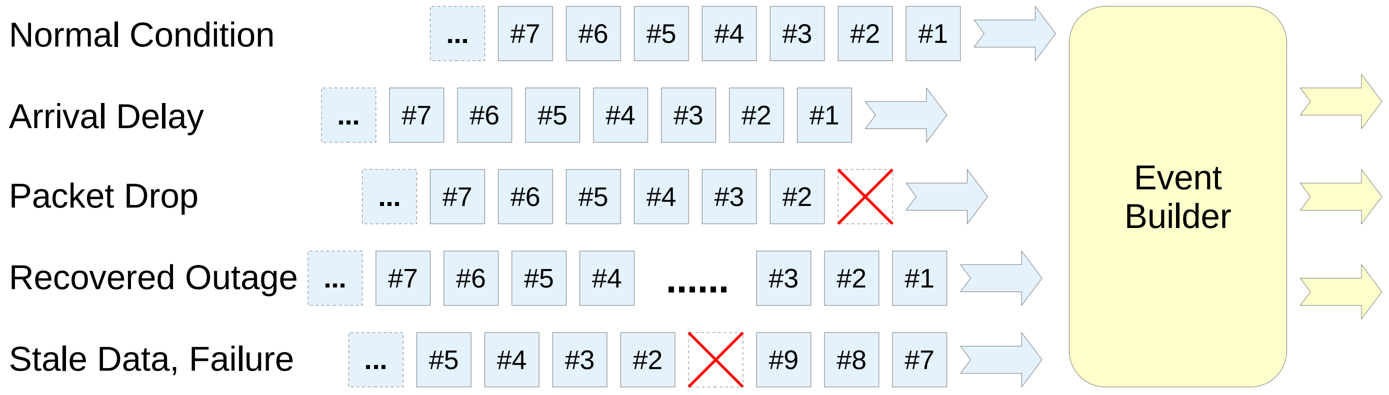
(1) FLP — First Level Processor (bare metal server)
Primary task: to receive data from readout electronics, buffer, validate, format and enqueue data blocks ready to be transferred to event building network.

FLP in BM@N is part of synchronous processing, it directly affects readout efficiency

Data Acquisition

Distributed Event Building




Normal Condition ... #7 #6 #5 #4 #3 #2 #1
Arrival Delay ... #7 #6 #5 #4 #3 #2 #1
Packet Drop ... #7 #6 #5 #4 #3 #2
Recovered Outage ... #7 #6 #5 #4 ...... #3 #2 #1
Stale Data, Failure ... #5 #4 #3 #2 #9 #8 #7
Event Builder

Event Building – process of sorting data fragments from subdetectors and assembling complete event data ready for physical analysis

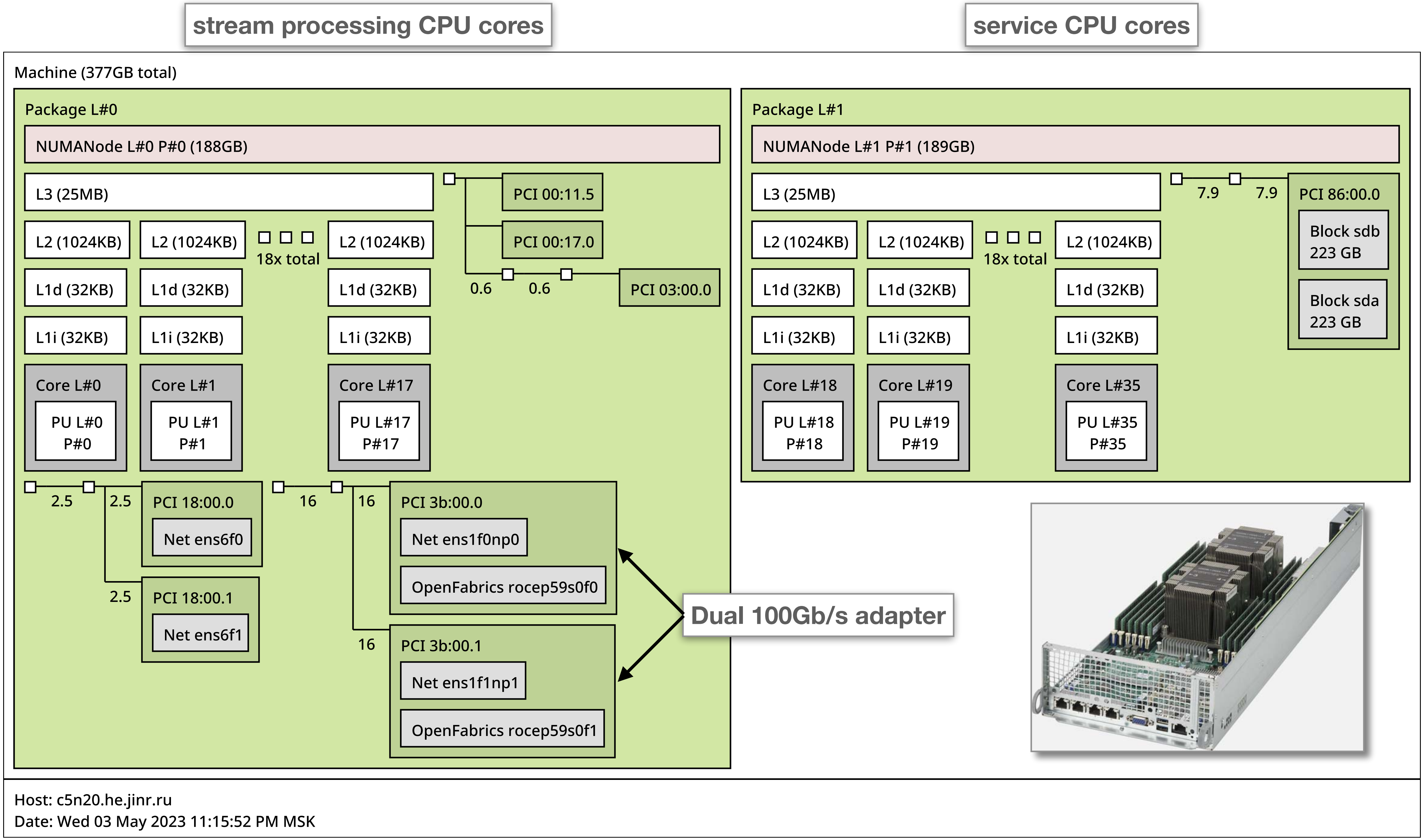
Reliability: handle single errors, data dropouts, corrupted data, timeouts, detector electronics restarts or servers restarts without process interruption

Event Building in BM@N is part of asynchronous processing, it does not affect readout efficiency under normal conditions

Data Acquisition

FLP Hardware Topology

Typical NUMA topology of dual-CPU server
used in BMN DAQ Data Center



stream processing CPU cores

service CPU cores

Machine (377GB total)

Package L#0

NUMANode L#0 P#0 (188GB)

L3 (25MB)

L2 (1024KB) | L2 (1024KB) | 18x total | L2 (1024KB)

L1d (32KB) | L1d (32KB) | L1d (32KB)

L1i (32KB) | L1i (32KB) | L1i (32KB)

Core L#0: PU L#0 P#0

Core L#1: PU L#1 P#1

Core L#17: PU L#17 P#17

PCI 00:11.5

PCI 00:17.0

0.6 0.6 PCI 03:00.0

2.5 2.5 PCI 18:00.0: Net ens6f0

2.5 PCI 18:00.1: Net ens6f1

16 16 PCI 3b:00.0: Net ens1f0np0, OpenFabrics rocep59s0f0

16 PCI 3b:00.1: Net ens1f1np1, OpenFabrics rocep59s0f1

Dual 100Gb/s adapter

Package L#1

NUMANode L#1 P#1 (189GB)

L3 (25MB)

L2 (1024KB) | L2 (1024KB) | 18x total | L2 (1024KB)

L1d (32KB) | L1d (32KB) | L1d (32KB)

L1i (32KB) | L1i (32KB) | L1i (32KB)

Core L#18: PU L#18 P#18

Core L#19: PU L#19 P#19

Core L#35: PU L#35 P#35

7.9 7.9 PCI 86:00.0: Block sdb 223 GB, Block sda 223 GB

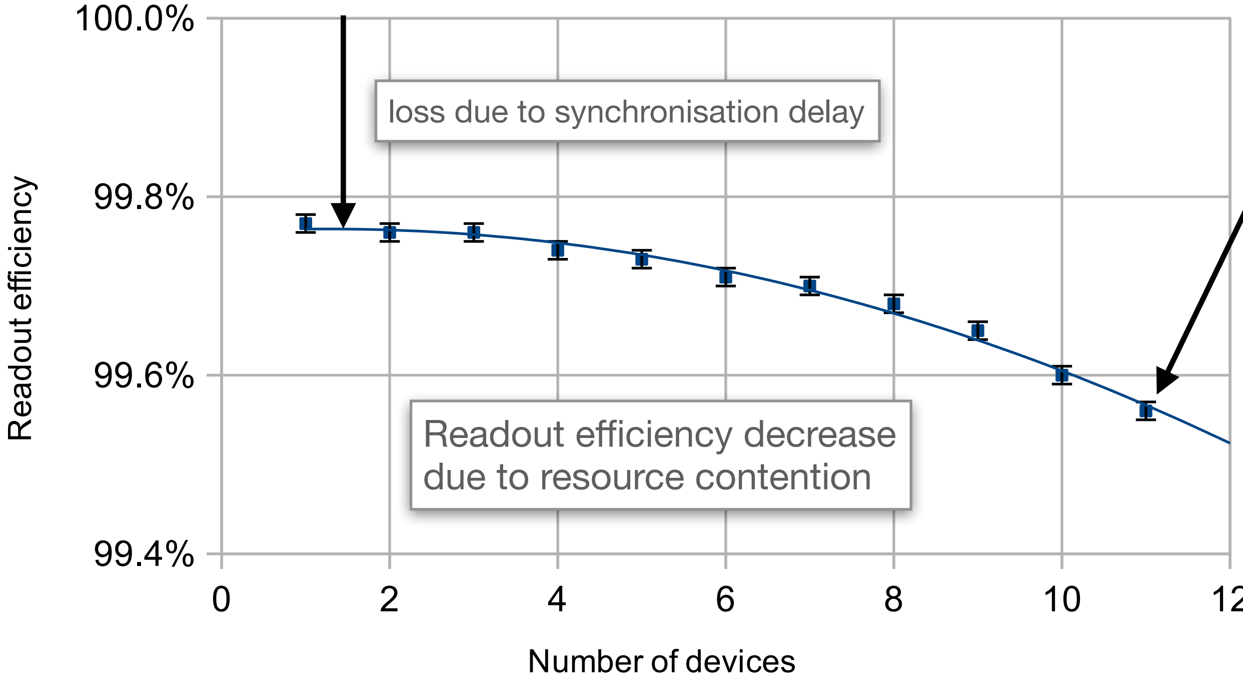
Host: c5n20.he.jinr.ru
Date: Wed 03 May 2023 11:15:52 PM MSK

Data Acquisition

OS Tuning for Real-Time

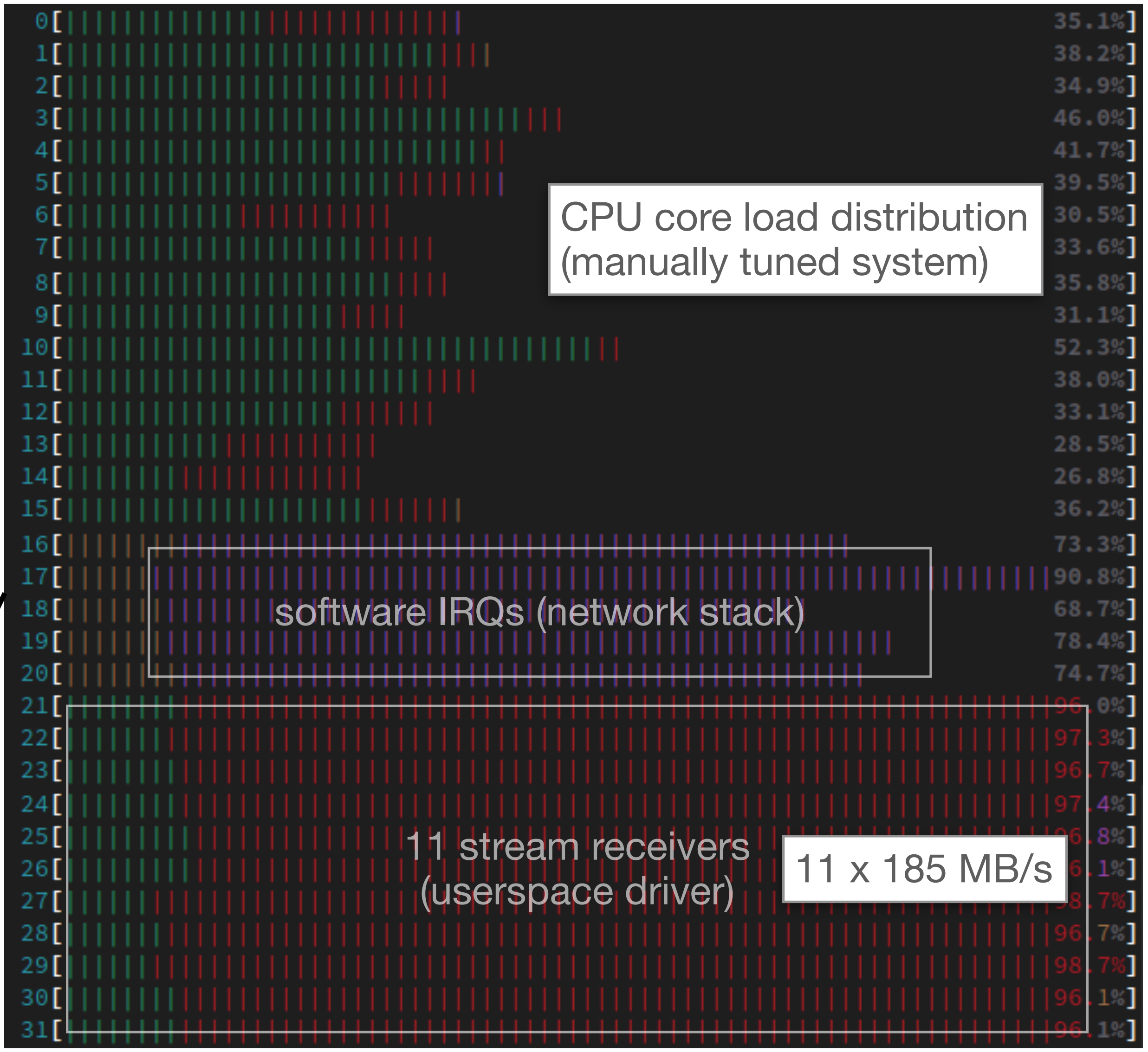
50 kHz random trigger
4.8 μs programmed dead time
…
10 kHz effective rate

Default system task scheduler is not optimal. Manual tuning is required.



loss due to synchronisation delay

Readout efficiency decrease due to resource contention



```
 0[ 35.1%]
 1[ 38.2%]
 2[ 34.9%]
 3[ 46.0%]
 4[ 41.7%]
 5[ 39.5%]
 6[ 30.5%]
 7[ 33.6%]
 8[ 35.8%]
 9[ 31.1%]
10[ 52.3%]
11[ 38.0%]
12[ 33.1%]
13[ 28.5%]
14[ 26.8%]
15[ 36.2%]
16[ 73.3%]
17[ 90.8%]
18[ 68.7%]
19[ 78.4%]
20[ 74.7%]
21[ 96.0%]
22[ 97.3%]
23[ 96.7%]
24[ 97.4%]
25[ 96.8%]
26[ 96.1%]
27[ 98.7%]
28[ 96.7%]
29[ 98.7%]
30[ 96.1%]
31[ 96.1%]
```

CPU core load distribution (manually tuned system)

software IRQs (network stack)

11 stream receivers (userspace driver)

11 x 185 MB/s

CPU cores assignment (11 streams test case)

0-15 — available to system scheduler
16-20 — software IRQ (network adapter jobs)
21-31 — stream receivers (userspace driver)

Platform: Supermicro X10DRT-PT, Dual Intel Xeon E5-2697A v4
Network: Mellanox ConnectX-5 100G
Readout modules: TQDC16VS-10G, f/w v2.9

Linux kernel 6.0.15
mitigations=off
intel_idle.max_cstate=1
idle=poll
skew_tick=1
isolcpus=16,17,…,31

Tuning
# tuned-adm profile network-latency

4 CPU cores per data stream

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux_for_real_time/9

IT Infrastructure

Compute Resource Virtualisation

Class	Task	Requirements	Where to run
Near real-time	FLP, data stream receiver	UDP protocol processing minimal latency, non-interruptible	Bare Hardware or LXC Manual tuning required
High Throughput	Event Building	High TCP/IP throughput, large buffers in RAM	Bare Hardware or LXC
Lightweight	Detector Control, Web services	Ease of Management	VM or container *KVM, LXC, Docker-in-KVM*

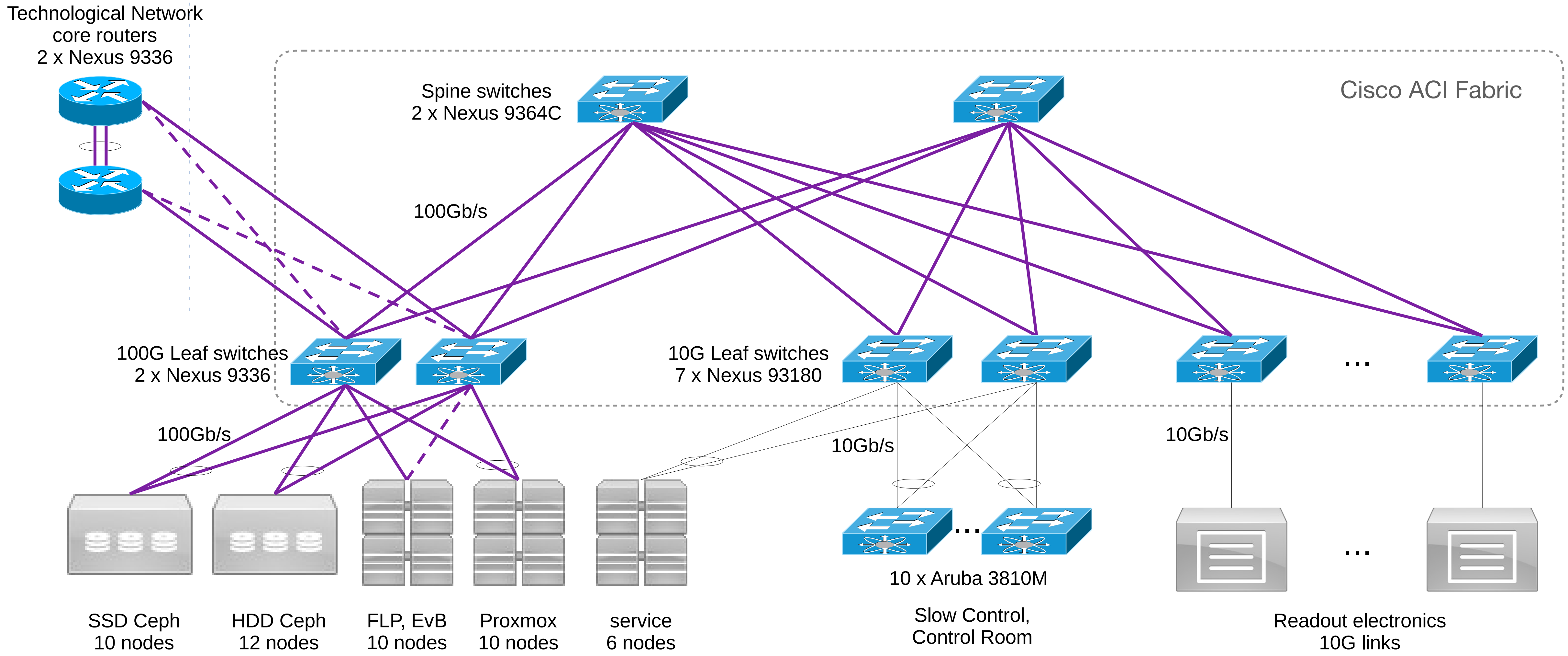
What should be tuned for real-time

- CPU power and frequency management
- Network adapter interrupt coalescing
- Disable firewall
- Process and soft-IRQ affinity, CPU core isolation

IT Infrastructure

BM@N DAQ Network

Network bandwidth
External: 200 Gb/s (400 planned)
Fabric: 200 Gb/s (extension possible)
Servers: 100 Gb/s (partially redundant)
Electronics: 1 and 10 Gb/s, no redundancy



Software Defined Storage: CephFS
HDD: 2.2 PB (EC-replicated)
SSD: 100 TB (triple replicated)

Compute resources
CPU: 720 cores total
RAM: 7500 GiB

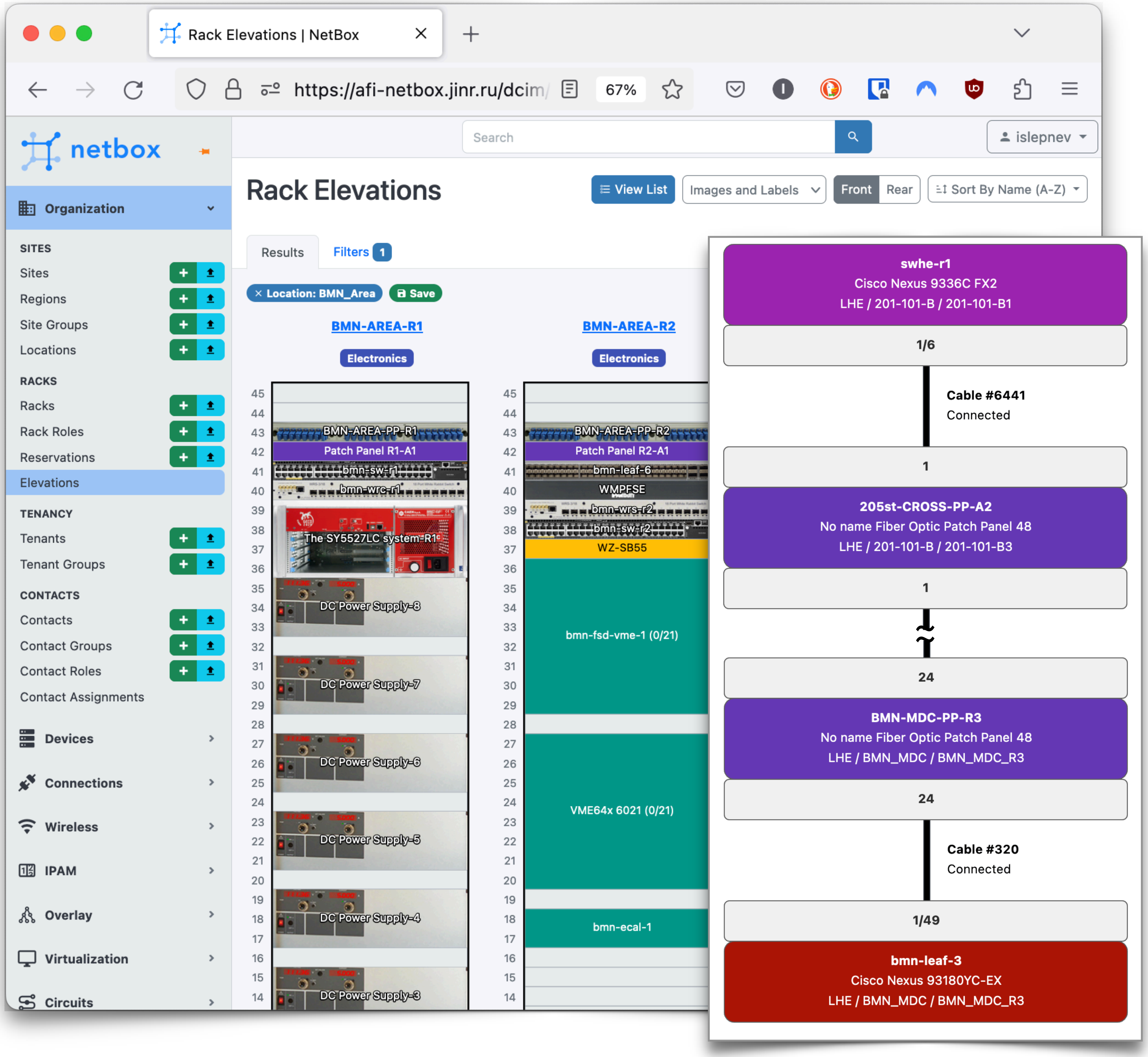
Network Redundancy and High Availability
- Network is base component of data taking process and is absolutely critical
- Thousands of devices communicate, loss of control or monitoring is critical
- Long data taking runs, no maintenance windows for critical upgrades and improvements
- Automated response to hardware and software failures, minimal operator intervention

Infrastructure Management

Data sources

NetBox, openDCIM

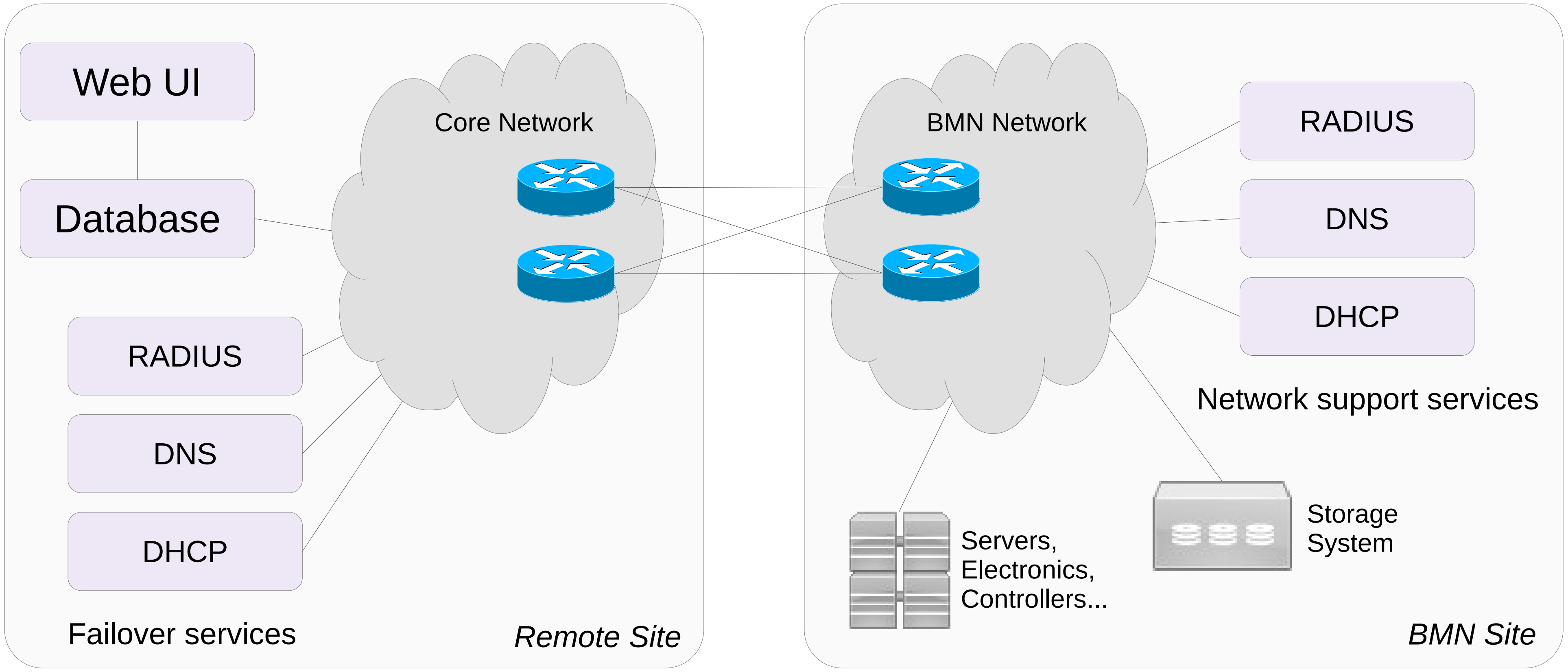
- Models hardware and networks
  - IP address space allocation
  - MAC, IP address, VLAN, User
  - Hardware items management
  - Cable and connections tracking
- Single source of truth. Not a monitoring
- Data source for automation tools
  - DHCP, DNS, RADIUS configs
- Documentation and reporting
- Existing and planned infrastructure



Infrastructure Management

High Availability of Network Services

- Database is source of configuration data, but not critical for operation
- Service configuration is updated periodically from database
  - Configuration is validated and service is restarted
  - Running service is announced to routers with IP Anycast address
- Failed service is excluded and traffic is routed to next available service automatically



Network support services

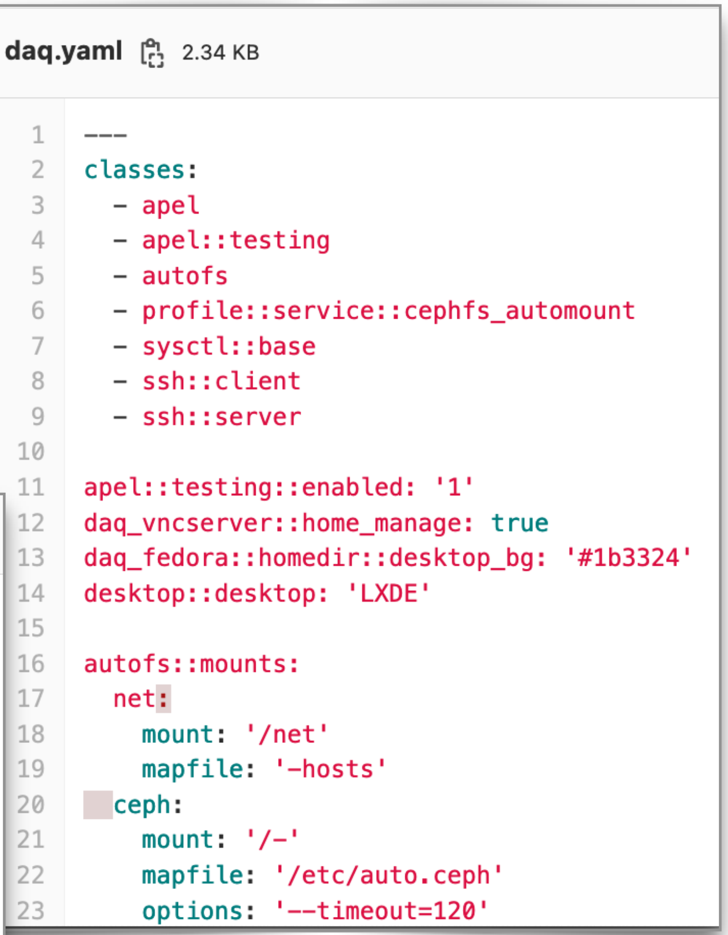
- DHCP: assignment of IP address by looking up database with hardware address
- RADIUS: dynamic assignment of virtual network, IEEE 802.1x
- DNS: hostname to IP address translation

Infrastructure Management

Infrastructure-as-Code

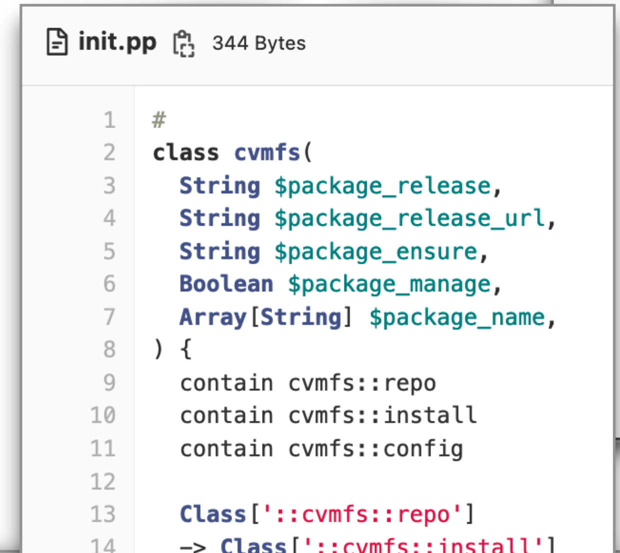
Tool	Method	Approach, our usage	Tasks
Puppet	Pull	functional (declarative)	configure services, settings
Ansible	Push	procedural (imperative)	updates, one-time tasks
—	—	manual admnistration	other complex tasks

- Machine-readable, version-controlled configuration files (YAML, Ruby)
- Puppet modules:
  - provision, configure, manage OS and application components
  - supported by community or our custom solution
- Hierarchical design: roles, profiles, classes are assigned to groups of computers. Dev and Prod environments.
- Documentation of IT Infrastructure configuration



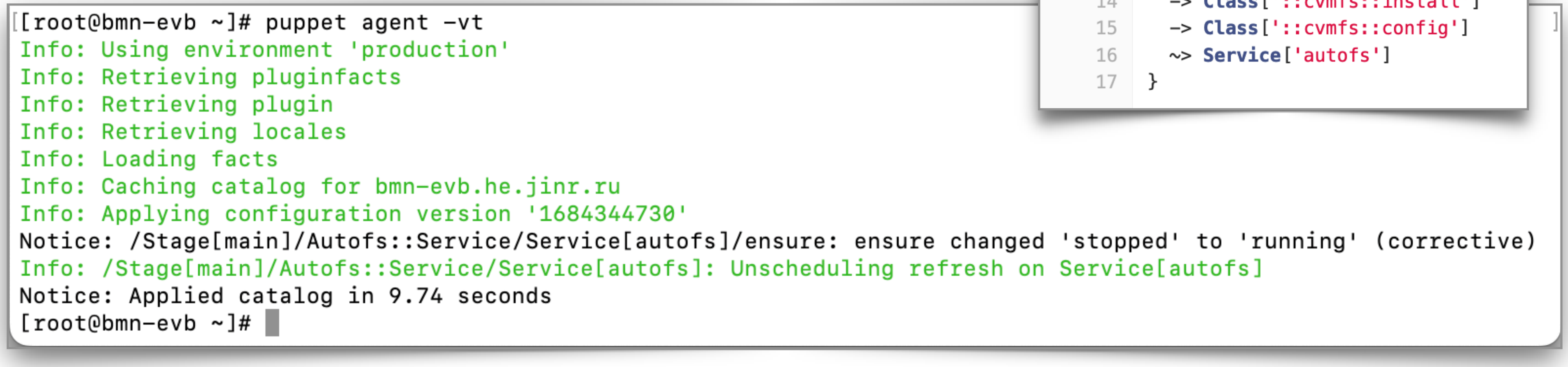
daq.yaml 2.34 KB

```yaml
---
classes:
  - apel
  - apel::testing
  - autofs
  - profile::service::cephfs_automount
  - sysctl::base
  - ssh::client
  - ssh::server

apel::testing::enabled: '1'
daq_vncserver::home_manage: true
daq_fedora::homedir::desktop_bg: '#1b3324'
desktop::desktop: 'LXDE'

autofs::mounts:
  net:
    mount: '/net'
    mapfile: '-hosts'
  ceph:
    mount: '/-'
    mapfile: '/etc/auto.ceph'
    options: '--timeout=120'
```



init.pp 344 Bytes

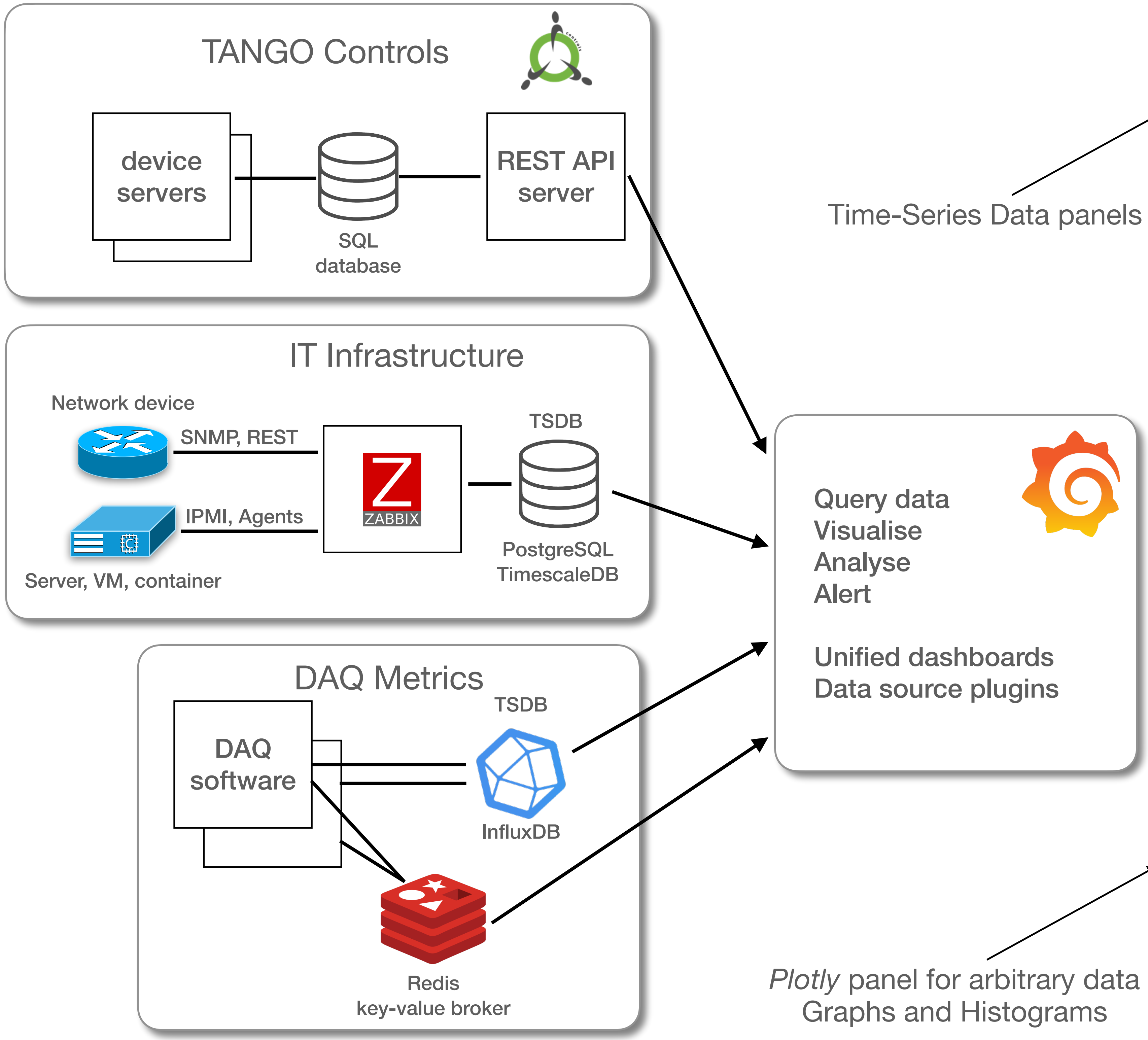
```puppet
#
class cvmfs(
  String $package_release,
  String $package_release_url,
  String $package_ensure,
  Boolean $package_manage,
  Array[String] $package_name,
) {
  contain cvmfs::repo
  contain cvmfs::install
  contain cvmfs::config

  Class['::cvmfs::repo']
  -> Class['::cvmfs::install']
  -> Class['::cvmfs::config']
  ~> Service['autofs']
}
```



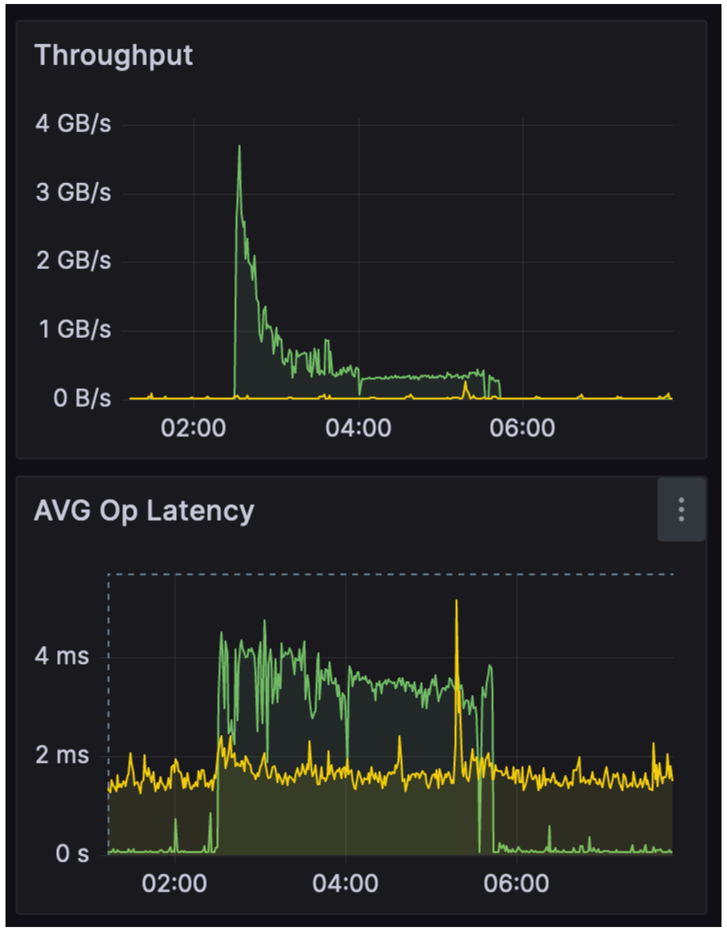
```
[root@bmn-evb ~]# puppet agent -vt
Info: Using environment 'production'
Info: Retrieving pluginfacts
Info: Retrieving plugin
Info: Retrieving locales
Info: Loading facts
Info: Caching catalog for bmn-evb.he.jinr.ru
Info: Applying configuration version '1684344730'
Notice: /Stage[main]/Autofs::Service/Service[autofs]/ensure: ensure changed 'stopped' to 'running' (corrective)
Info: /Stage[main]/Autofs::Service/Service[autofs]: Unscheduling refresh on Service[autofs]
Notice: Applied catalog in 9.74 seconds
[root@bmn-evb ~]#
```

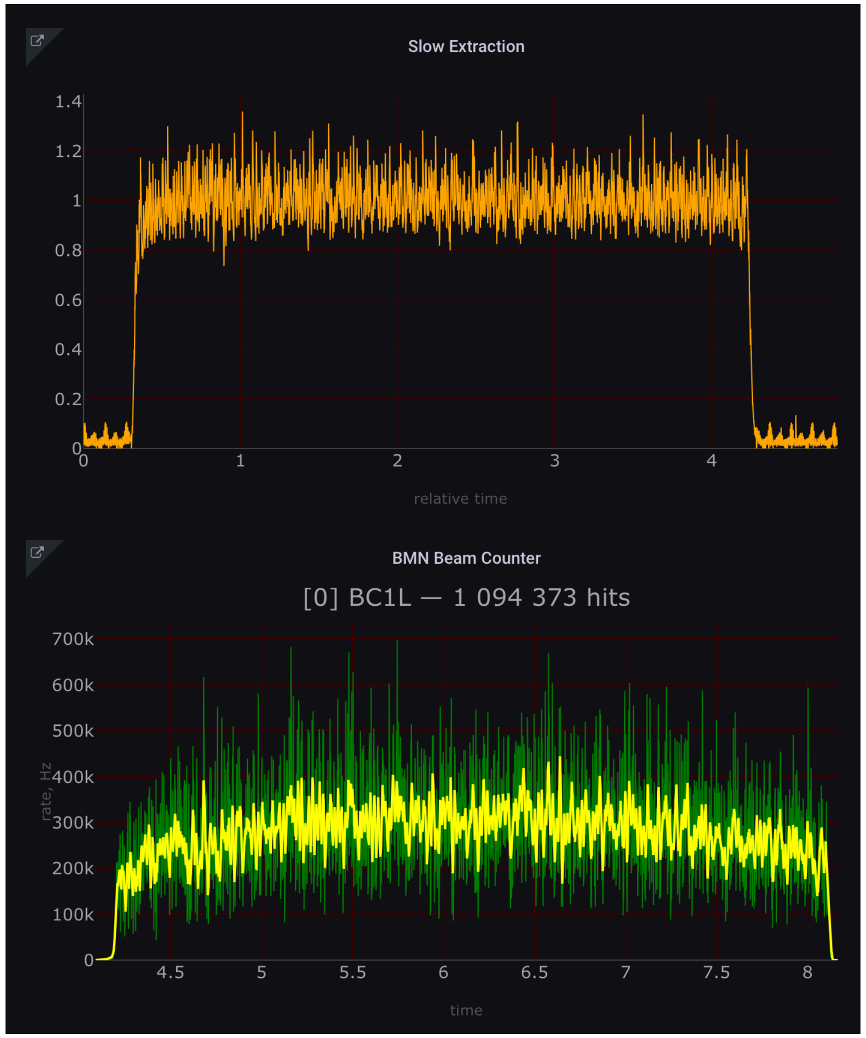
Observability: metrics

Grafana Unified Dashboards



Time-Series Data panels

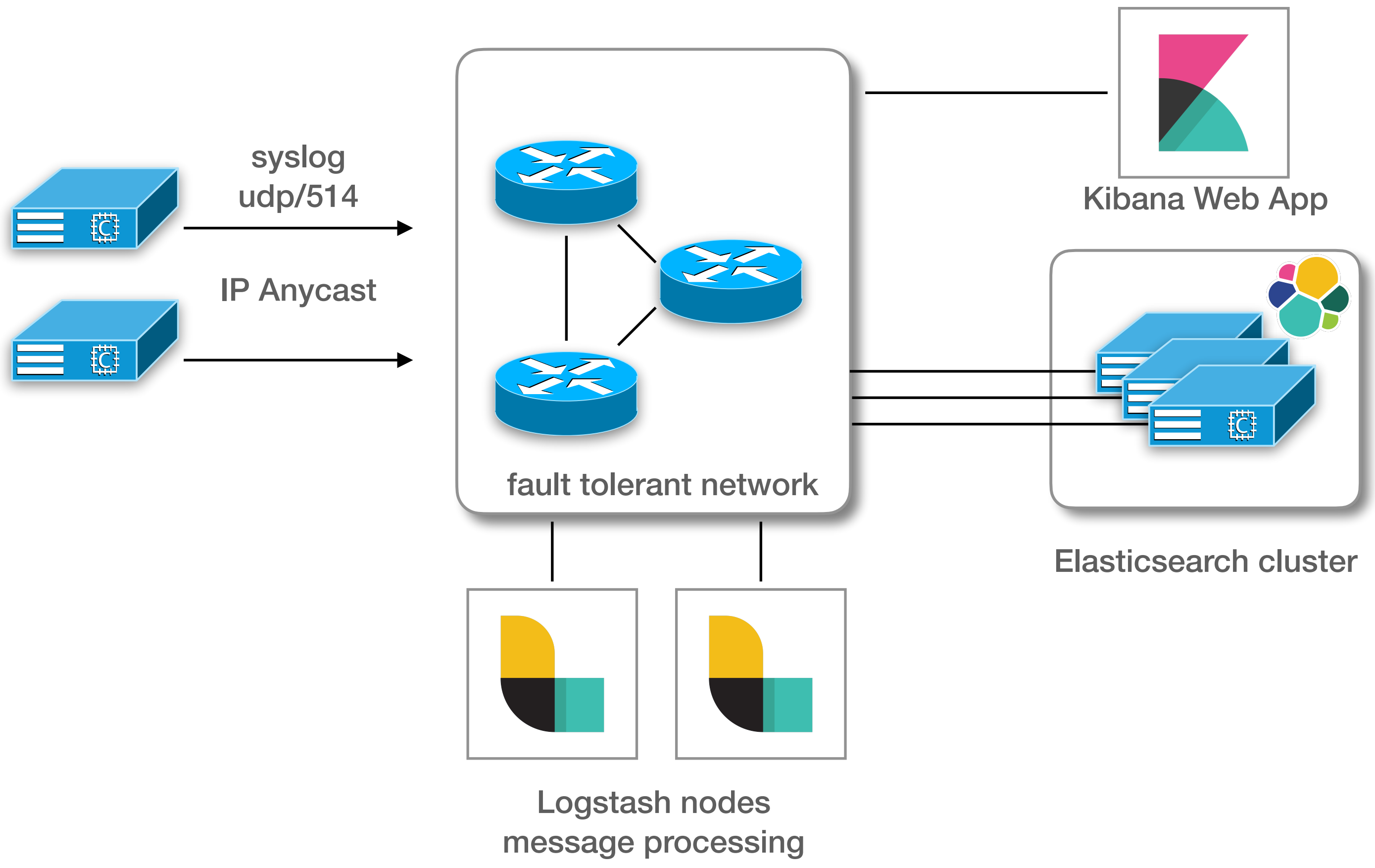




Plotly panel for arbitrary data
Graphs and Histograms

Observability: logs

Elasticsearch, Logstash, Kibana



syslog udp/514

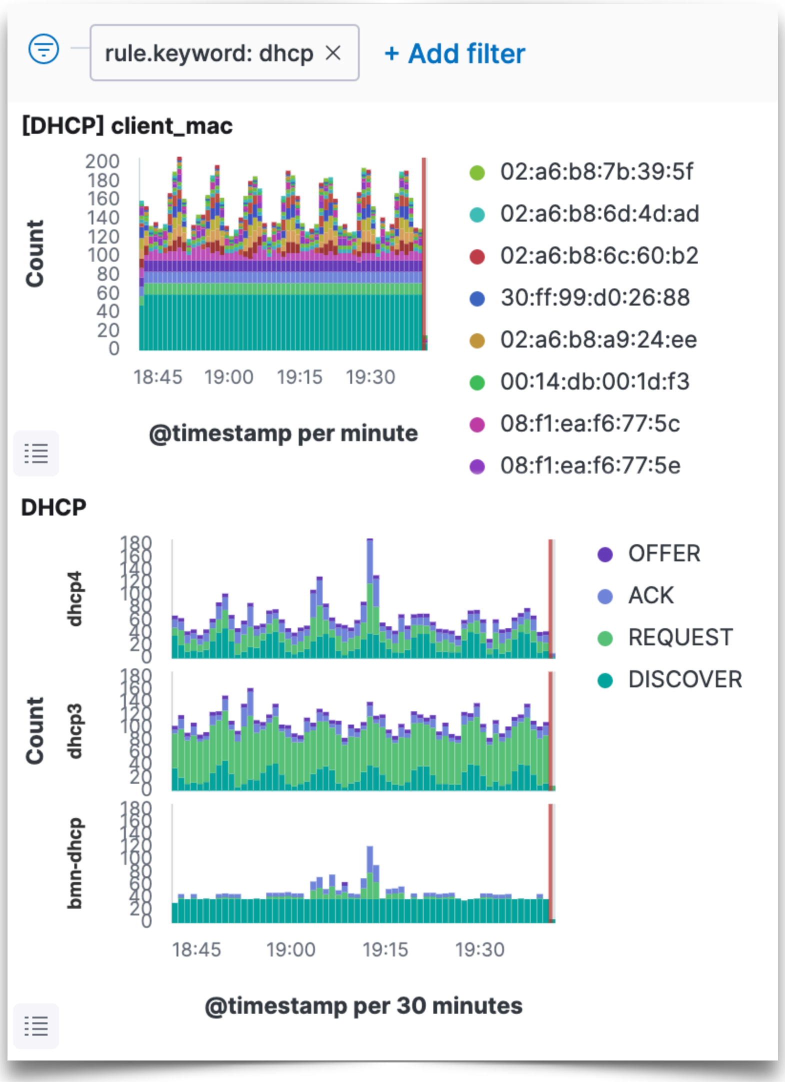
IP Anycast

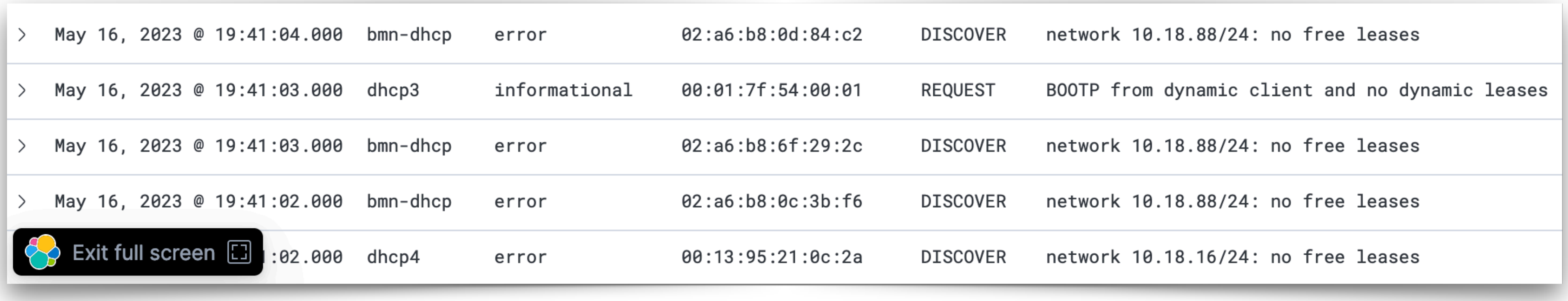
fault tolerant network

Kibana Web App

Elasticsearch cluster

Logstash nodes message processing





```
>  May 16, 2023 @ 19:41:04.000  bmn-dhcp  error          02:a6:b8:0d:84:c2  DISCOVER  network 10.18.88/24: no free leases
>  May 16, 2023 @ 19:41:03.000  dhcp3     informational  00:01:7f:54:00:01  REQUEST   BOOTP from dynamic client and no dynamic leases
>  May 16, 2023 @ 19:41:03.000  bmn-dhcp  error          02:a6:b8:6f:29:2c  DISCOVER  network 10.18.88/24: no free leases
>  May 16, 2023 @ 19:41:02.000  bmn-dhcp  error          02:a6:b8:0c:3b:f6  DISCOVER  network 10.18.88/24: no free leases
                       :02.000  dhcp4     error          00:13:95:21:0c:2a  DISCOVER  network 10.18.16/24: no free leases
```

#TODO

DAQ Infrastructure was adequate for past BM@N runs, no critical issues. However…

- Improve DAQ readout efficiency. Considering intermediate hardware buffer based on CRU16.
- New detectors, thus more data streams, additional FLP nodes are required.
- Trigger rate is expected to rise on next run, increasing total data throughput. More Event Builder nodes are required.
- Extend storage space for autonomous operation.
- BMN DAQ Data Center is at limit of cooling power. Considering to move storage system to building 14
- Routine maintenance for all information systems. OS and application upgrades.
- Last, not least. Finish integration with JINR SSO authentication.