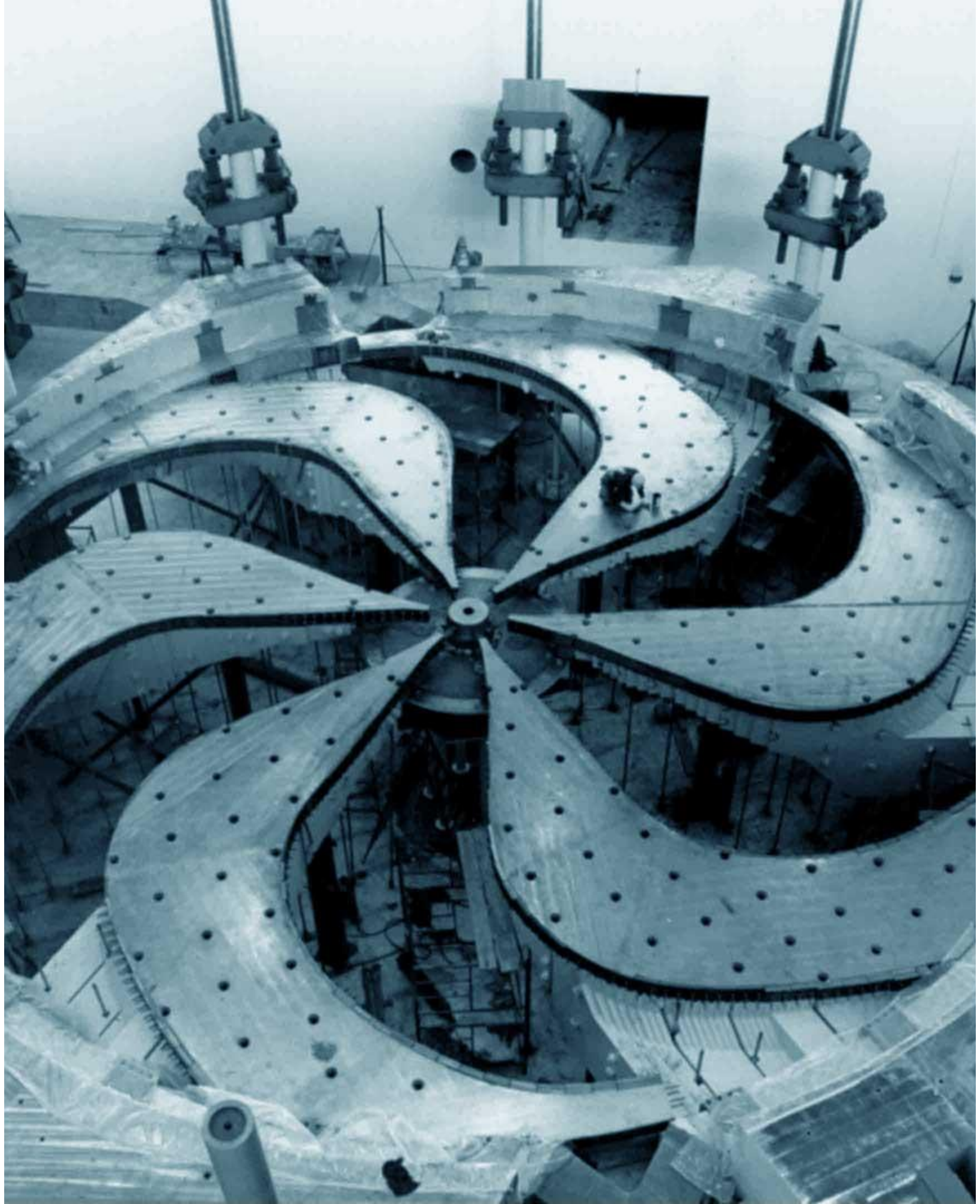
TRIUMF

Infrastructure and Analytics Summary (Tier–1)

Fernando Fernandez Galindo
TRIUMF Scientific Computing Department

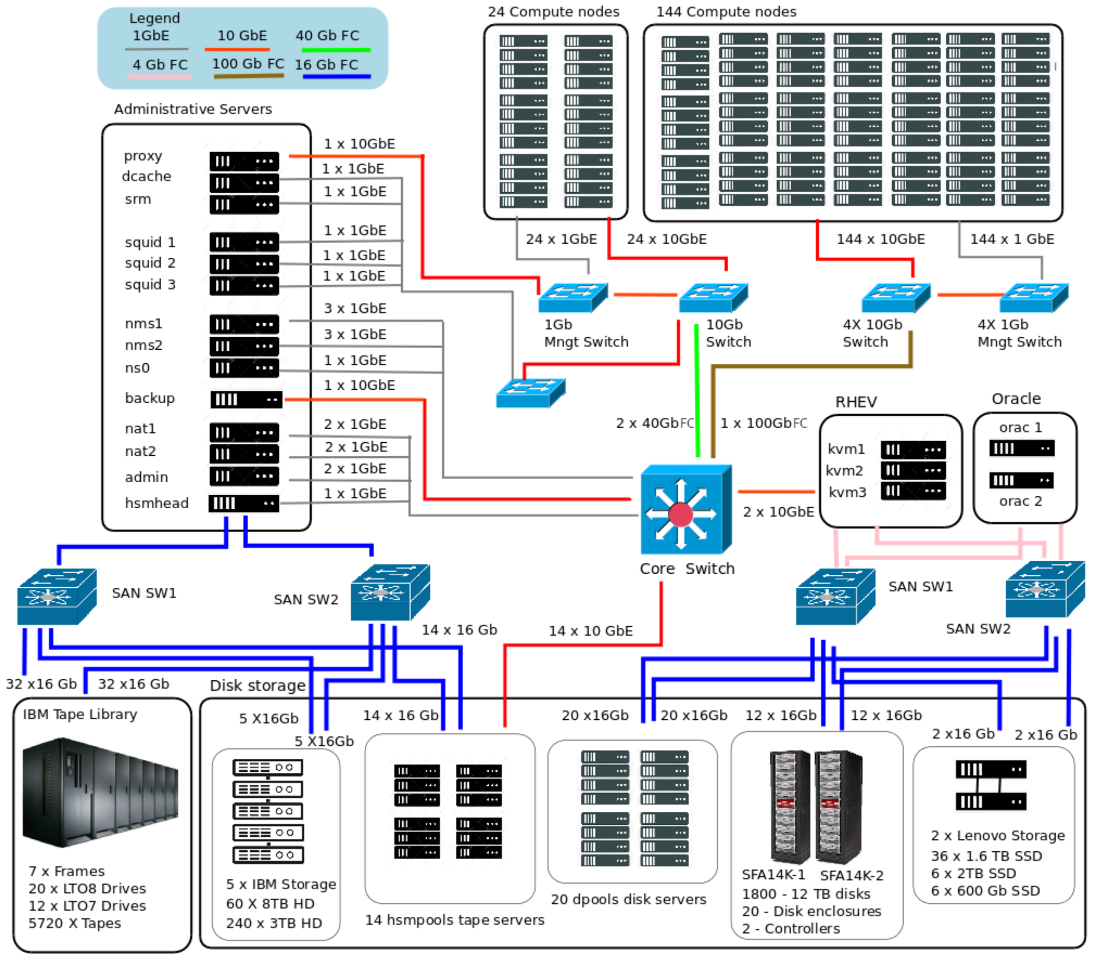
US ATLAS Computing Facilities Face-to-Face at SLAC
December 1st, 2022



Discovery, accelerated

INFRASTRUCTURE SUMMARY

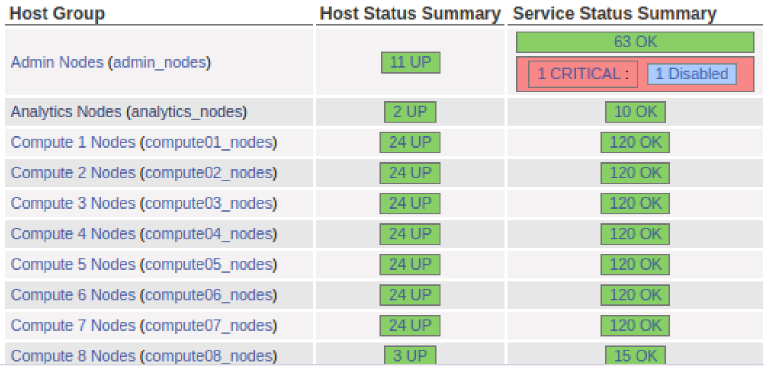
- Compute:
  - ARC CE 6.15.1
  - HTCondor 9.0.11
  - 1,600 cores at TRIUMF.
  - 7,820 at Simon Fraser University.
- Storage:
  - dCache (6.2.39 -> 7.2.26 soon)
    - 30 dpool serving 17PB of usable disk.
    - 13 hsmpool nodes 36PB of usable tape.
- OS:
  - Mostly running on SL 7.9
  - RedHat has given us 1,000 licenses of RHEL 9 to which we will be upgrading to in the near future.



ANALYTICS PROJECT MOTIVATION

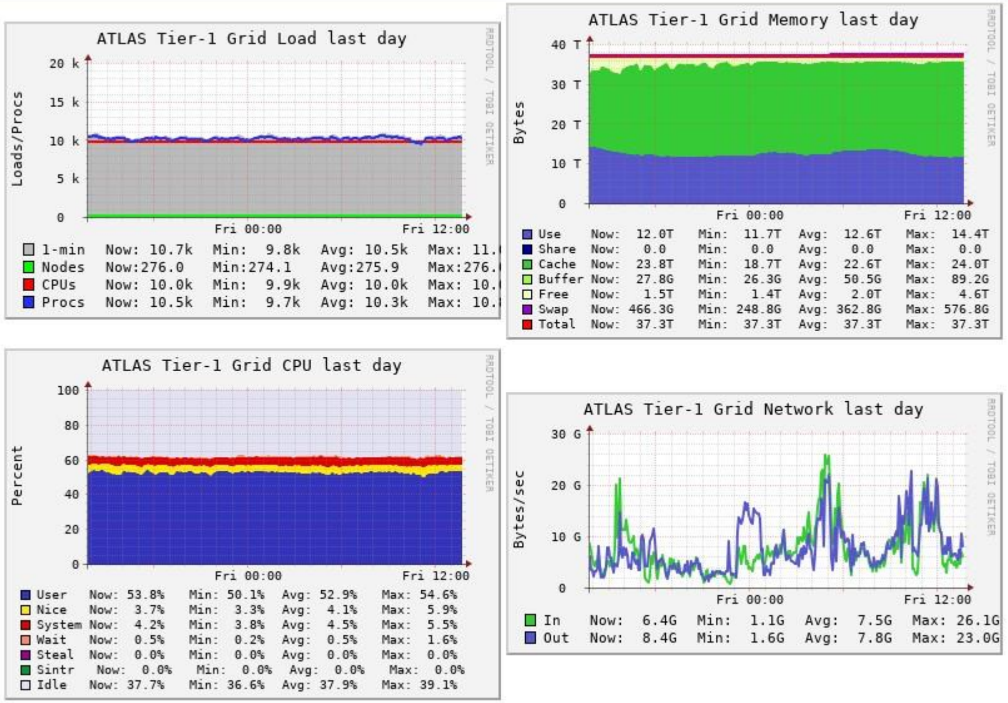
- With growing infrastructure an ever-increasing collection of heterogeneous monitoring data is produced.
- Our existing monitoring and alerting implementation is robust and stable but rather static and isolated.
- Early 2020 we started the analytics project on hardware that was deprecated.
- The project's main objectives are:
    - Create a framework where the different datasets can be brought in together for analysis and monitoring.
    - Experiment with tools and techniques like ML, to assist in finding correlations between different areas of our infrastructure, detect anomalies, inefficiencies and maybe even predict issues.

Nagios



Host Group	Host Status Summary	Service Status Summary
Admin Nodes (admin_nodes)	11 UP	63 OK 1 CRITICAL : 1 Disabled
Analytics Nodes (analytics_nodes)	2 UP	10 OK
Compute 1 Nodes (compute01_nodes)	24 UP	120 OK
Compute 2 Nodes (compute02_nodes)	24 UP	120 OK
Compute 3 Nodes (compute03_nodes)	24 UP	120 OK
Compute 4 Nodes (compute04_nodes)	24 UP	120 OK
Compute 5 Nodes (compute05_nodes)	24 UP	120 OK
Compute 6 Nodes (compute06_nodes)	24 UP	120 OK
Compute 7 Nodes (compute07_nodes)	24 UP	120 OK
Compute 8 Nodes (compute08_nodes)	3 UP	15 OK

Ganglia



NEW HARDWARE (FALL 2022)

Frontend (Elasticsearch and clients)

- 1x PowerEdge R650
- **CPU:**
  2x Xeon 6336Y – 24 cores with Scikit-learn extensions.
- **GPU:**
  Nvidia Telsa T4
- **Memory:**
  256GB
- **Network:**
  Nvidia Mellanox ConnectX-5
- **Storage:**
  2x 480GB SSD (OS)
  2x 3.84TB NVMe

Backend (Elasticsearch and Logstash)

- 2x PowerEdge R650
- **CPU:**
  Xeon 6326 – 16 Cores
- **Memory:**
  256GB
- **Network:**
  Nvidia Mellanox ConnectX-5
- **Storage:**
  2x 480GB SSD (OS)
  2x 3.84TB NVMe (Hot Data)
  2x 7.68TB SSD (Warm Data)
- KVM will be used to create 4 ES nodes: master, hot data, warm data, transform (with logstash).

SOFTWARE OVERVIEW

COLLECTION

- Custom Scripts
- Elastic beats
- Gmond
- Nagios
- SNMP traps
- Syslog
- Telegraph

ENRICHMENT

- ES pipelines
- Logstash

STORAGE

- MariaDB
- Elasticsearch
- influxDB
- RRD

VISUALIZATION

- Ganglia
- Grafana

ALERTING

- Email
- Grafana
- Nagios
- Pager (24/7)

*Purple denotes additions from the analytics project.

ELASTIC SUITE

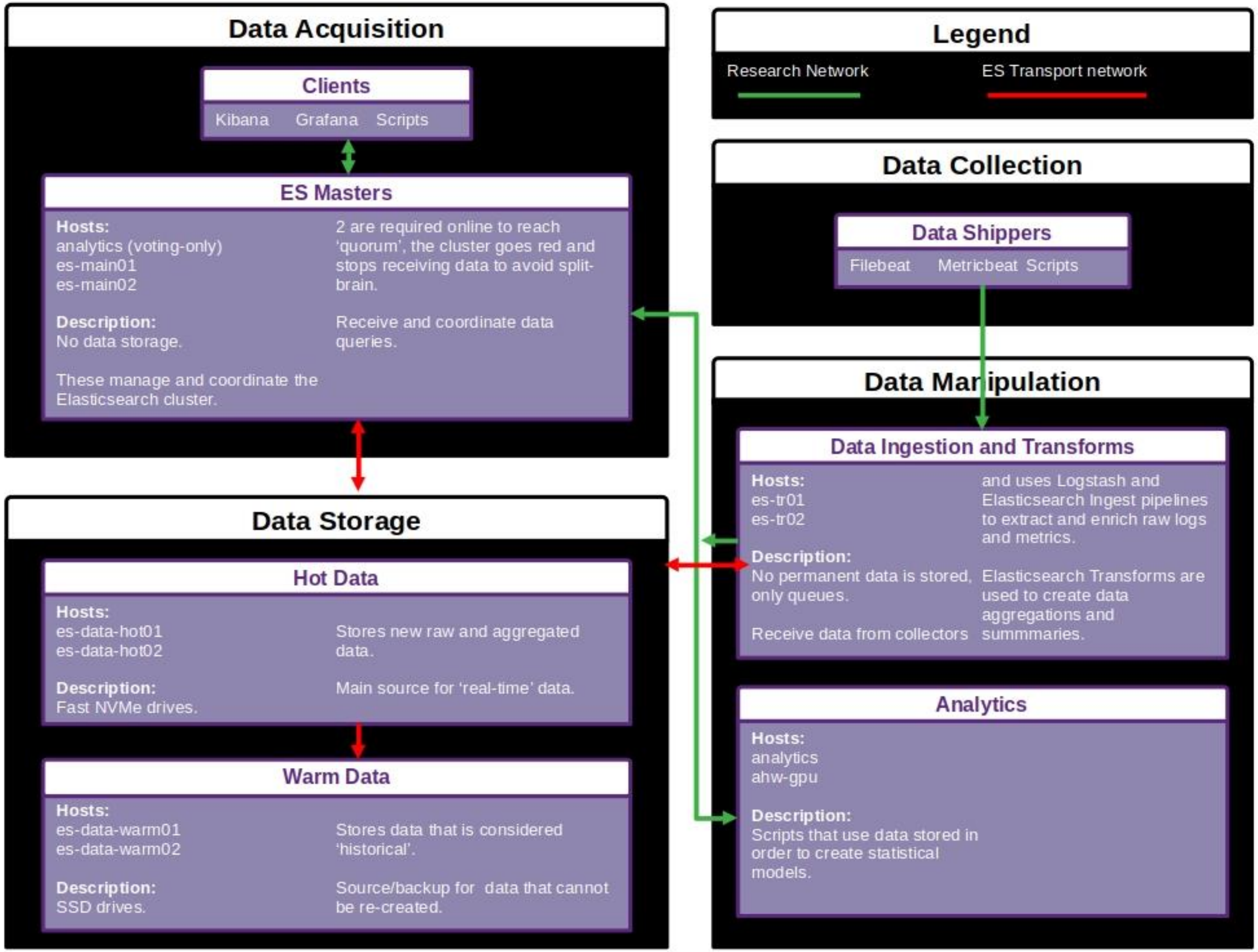
- It is the workhorse of the analytics platform.
- Originally interested in their built-in machine learning tools and while promising, licensing is expensive. The free 'basic' license provides all our needs so far.
- Currently using version 7.17.6, upgrading to 8.3.3 soon.
- **Elasticsearch**:
  - Flexible database, can hold heterogeneous data.
  - Easy to grow horizontally as demands increase.
  - Many tools to aggregate and transform data.
- **Logstash**:
  - Many filters to parse and enrich data.
  - Multiple instances and pipelines to balance the load.
  - Many input and output protocols.
  - Persistent and 'dead-letter' queues.
- **Beats**:
  - 'Smart' collectors that monitor log files (Filebeat), service metrics (Metricbeat) and network ports (Packetbeat).
  - Balance loads to multiple outputs and queues data if they are unavaliable.

elastic

GRAFANA

- Currently using version 9.2.1
- Our main visualization software for the following reasons:
    - It can use a large variety of different data sources.
    - Has many options for creating nice looking dashboards easily.
    - Powerful templating of panels.
    - It can further transform data on the fly.
    - It can generate alerts data query-based alerts.

ELASTICSEARCH PIPELINE



Legend

Research Network 🟩

ES Transport network 🟥

Data Acquisition

Clients

Kibana Grafana Scripts

ES Masters

Hosts:
analytics (voting-only)
es-main01
es-main02

2 are required online to reach 'quorum', the cluster goes red and stops receiving data to avoid split-brain.

Description:
No data storage.

Receive and coordinate data queries.

These manage and coordinate the Elasticsearch cluster.

Data Storage

Hot Data

Hosts:
es-data-hot01
es-data-hot02

Stores new raw and aggregated data.

Description:
Fast NVMe drives.

Main source for 'real-time' data.

Warm Data

Hosts:
es-data-warm01
es-data-warm02

Stores data that is considered 'historical'.

Description:
SSD drives.

Source/backup for data that cannot be re-created.

Data Collection

Data Shippers

Filebeat Metricbeat Scripts

Data Manipulation

Data Ingestion and Transforms

Hosts:
es-tr01
es-tr02

Description:
No permanent data is stored, only queues.

Receive data from collectors and uses Logstash and Elasticsearch Ingest pipelines to extract and enrich raw logs and metrics.

Elasticsearch Transforms are used to create data aggregations and summmaries.

Analytics

Hosts:
analytics
ahw-gpu

Description:
Scripts that use data stored in order to create statistical models.

DATASETS

Logs

Service	Size (GB)	Storage DB
dCache (billing, ftp|srm|webdav|xrootd access)	4,500	Elasticsearch
network (router)	2	Elasticsearch
SNMP (traps)	15	MariaDB
system (auth, iptables, kernel)	400	Elasticsearch

Metrics

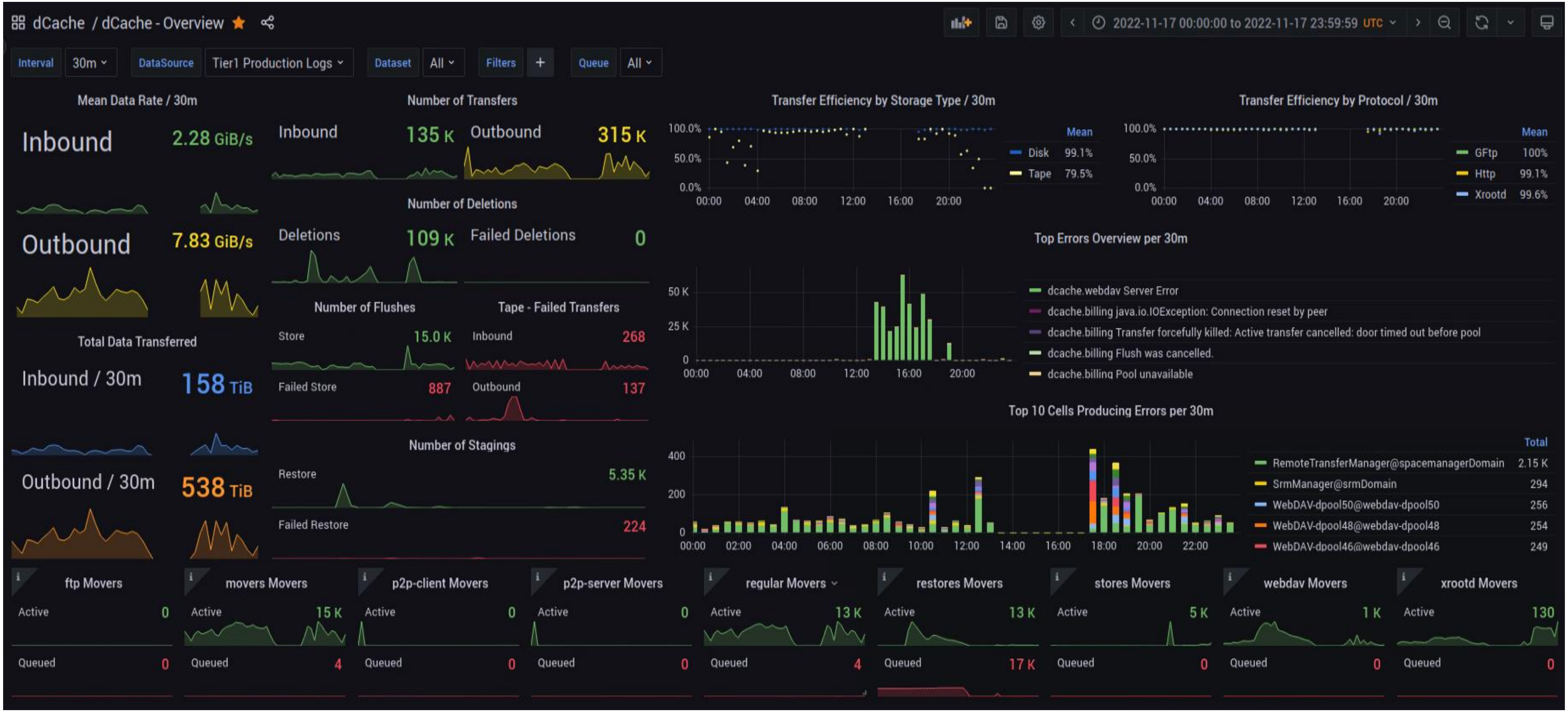
Service	Size (GB)	Storage
dCache (queues, movers)	25	Elasticsearch
dCache (netflows)	1,250	Elasticsearch
HTCondor (job history, status)	10	Elasticsearch
infrastructure (DDN and inlet temps, SSD TBW)	50	Elasticsearch
infrastructure (humidity, PDU, temps, etc)	2*	RRD
mySQL (status)	1	Elasticsearch
network (router sflow)	15	influxDB
postgreSQL (activity, bgwriter, database)	200	Elasticsearch
system (cpu, mem, net, etc)	2*	RRD
tape library stats (device, volume)	2	influxDB
tape library (consumption, performance, staging, etc)	2*	RRD

* 2GB for all datasets in RRD

DCACHE

- Filebeat monitors and ships the contents of dCache's logs.
  - Billing logs contain transaction information within the scope of dCache.
  - Access logs contain transaction information pertaining to the different door protocols (FTP, SRM, WebDAV, XRootD).
- Logstash parses these logs into fields to create Elasticsearch documents, and enrich them as necessary (DNS resolution, GeoIP, tags).
- Packetbeat monitors the door protocol ports to obtain network flows and TLS handshake response times.
- A custom script parses dCache's pool queue table and sends it to Elasticsearch.

DCACHE OVERVIEW DASHBOARD



dCache / dCache - Overview

2022-11-17 00:00:00 to 2022-11-17 23:59:59 UTC

Interval: 30m | DataSource: Tier1 Production Logs | Dataset: All | Filters: + | Queue: All

Mean Data Rate / 30m

Inbound 2.28 GiB/s

Outbound 7.83 GiB/s

Total Data Transferred

Inbound / 30m 158 TiB

Outbound / 30m 538 TiB

Number of Transfers

Inbound 135 K

Outbound 315 K

Number of Deletions

Deletions 109 K

Failed Deletions 0

Number of Flushes

Store 15.0 K

Failed Store 887

Tape - Failed Transfers

Inbound 268

Outbound 137

Number of Stagings

Restore 5.35 K

Failed Restore 224

Transfer Efficiency by Storage Type / 30m

	Mean
Disk	99.1%
Tape	79.5%

Transfer Efficiency by Protocol / 30m

	Mean
GFtp	100%
Http	99.1%
Xrootd	99.6%

Top Errors Overview per 30m

- dcache.webdav Server Error
- dcache.billing java.io.IOException: Connection reset by peer
- dcache.billing Transfer forcefully killed: Active transfer cancelled: door timed out before pool
- dcache.billing Flush was cancelled.
- dcache.billing Pool unavailable

Top 10 Cells Producing Errors per 30m

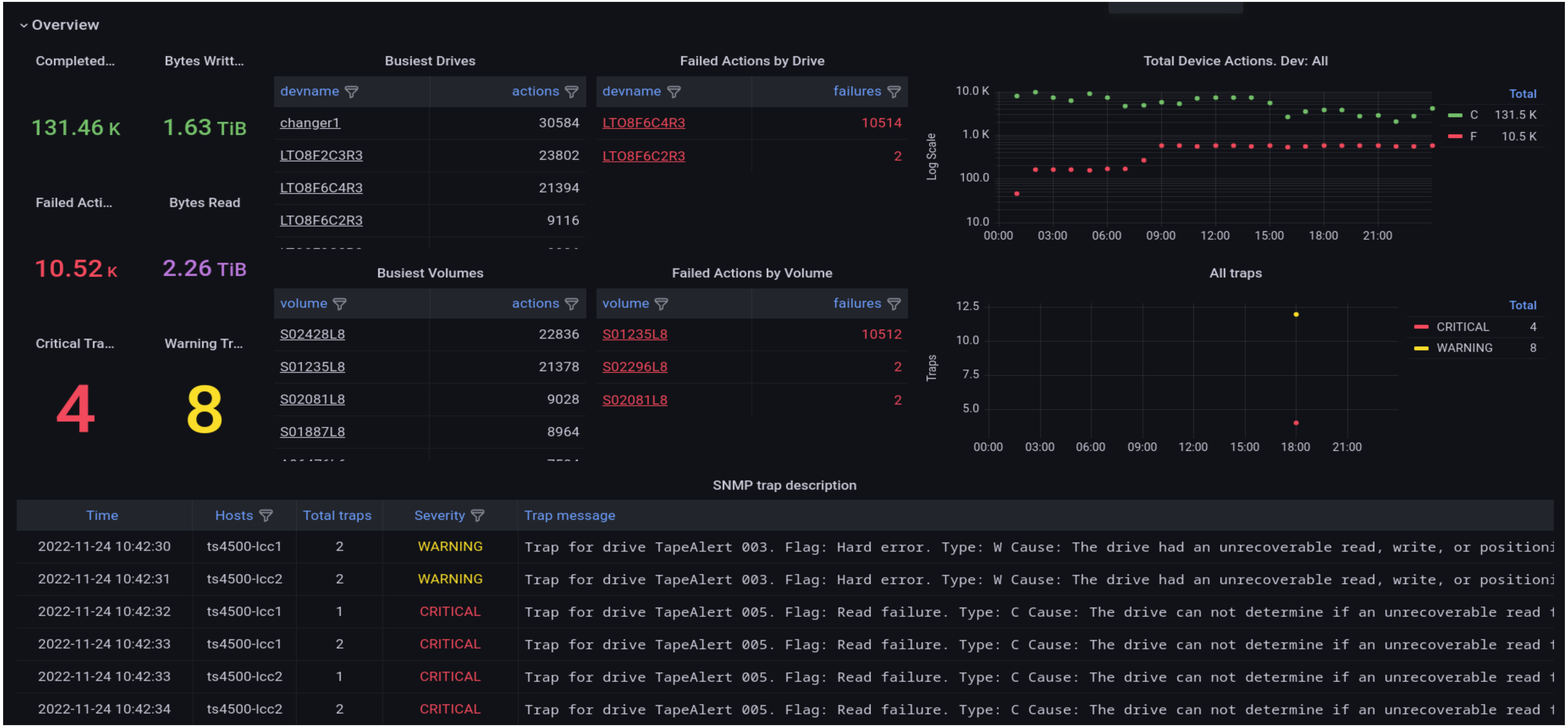
	Total
RemoteTransferManager@spacemanagerDomain	2.15 K
SrmManager@srmDomain	294
WebDAV-dpool50@webdav-dpool50	256
WebDAV-dpool48@webdav-dpool48	254
WebDAV-dpool46@webdav-dpool46	249

	Active	Queued
ftp Movers	0	0
movers Movers	15 K	4
p2p-client Movers	0	0
p2p-server Movers	0	0
regular Movers	13 K	4
restores Movers	13 K	17 K
stores Movers	5 K	0
webdav Movers	1 K	0
xrootd Movers	130	0

TAPE LIBRARY (HSM)

- The library produces SNMP traps when there are failures which are stored in MariaDB instance from where Grafana queries the information directly.
- Another MariaDB instance that records the tape library's devices and volumes activities.
- A custom script extracts the data and sends it to influxDB where we can manipulate it for later visualization.
- Here we chose influxDB to test and see how it compares against Elasticsearch for metrics data. Main benefit so far is the smaller storage footprint. Still investigating if there are other benefits.

TAPE LIBRARY DASHBOARD



⌄ Overview

Completed...
131.46 K

Bytes Writt...
1.63 TiB

Failed Acti...
10.52 K

Bytes Read
2.26 TiB

Critical Tra...
4

Warning Tr...
8

Busiest Drives

devname	actions
changer1	30584
LTO8F2C3R3	23802
LTO8F6C4R3	21394
LTO8F6C2R3	9116

Failed Actions by Drive

devname	failures
LTO8F6C4R3	10514
LTO8F6C2R3	2

Busiest Volumes

volume	actions
S02428L8	22836
S01235L8	21378
S02081L8	9028
S01887L8	8964

Failed Actions by Volume

volume	failures
S01235L8	10512
S02296L8	2
S02081L8	2

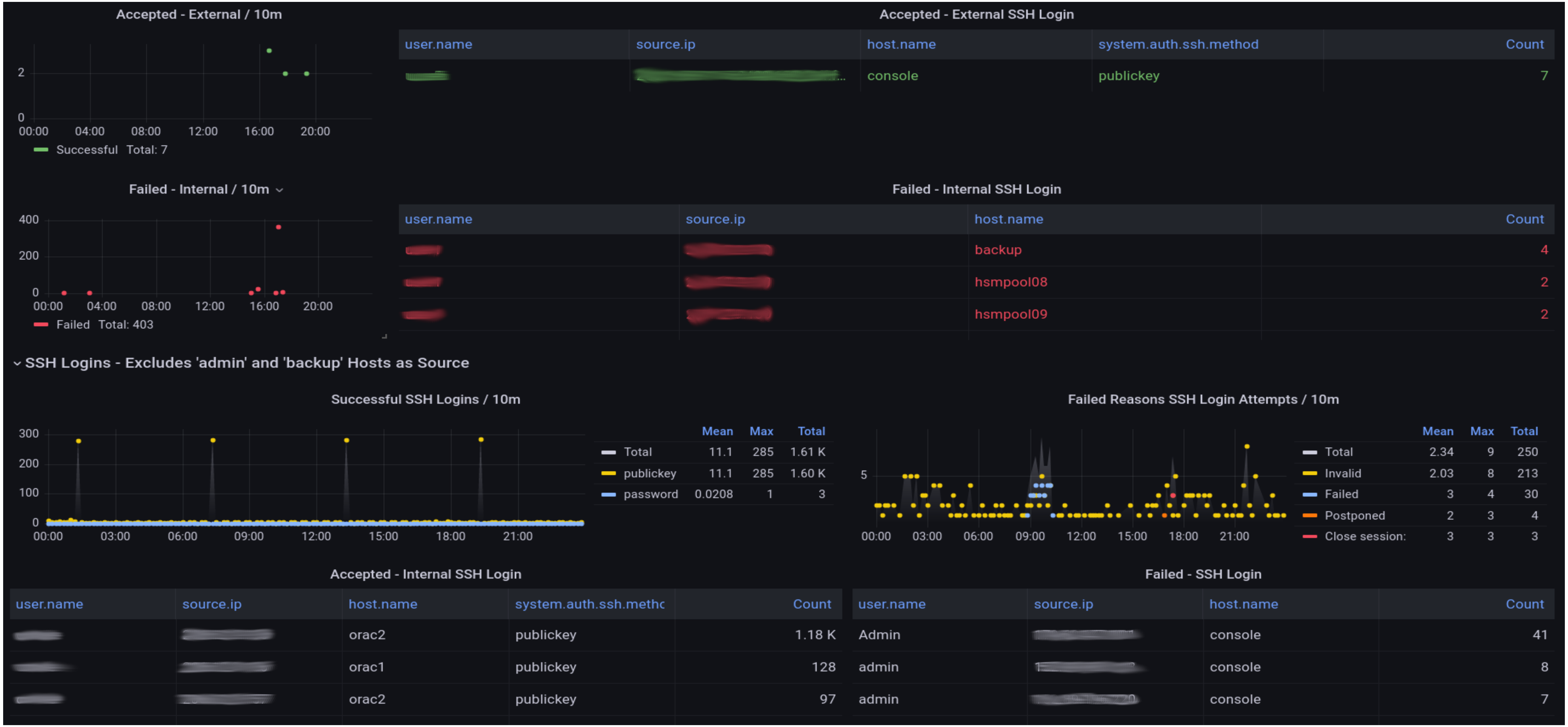
SNMP trap description

Time	Hosts	Total traps	Severity	Trap message
2022-11-24 10:42:30	ts4500-lcc1	2	WARNING	Trap for drive TapeAlert 003. Flag: Hard error. Type: W Cause: The drive had an unrecoverable read, write, or positioni
2022-11-24 10:42:31	ts4500-lcc2	2	WARNING	Trap for drive TapeAlert 003. Flag: Hard error. Type: W Cause: The drive had an unrecoverable read, write, or positioni
2022-11-24 10:42:32	ts4500-lcc1	1	CRITICAL	Trap for drive TapeAlert 005. Flag: Read failure. Type: C Cause: The drive can not determine if an unrecoverable read i
2022-11-24 10:42:33	ts4500-lcc1	2	CRITICAL	Trap for drive TapeAlert 005. Flag: Read failure. Type: C Cause: The drive can not determine if an unrecoverable read i
2022-11-24 10:42:33	ts4500-lcc2	1	CRITICAL	Trap for drive TapeAlert 005. Flag: Read failure. Type: C Cause: The drive can not determine if an unrecoverable read i
2022-11-24 10:42:34	ts4500-lcc2	2	CRITICAL	Trap for drive TapeAlert 005. Flag: Read failure. Type: C Cause: The drive can not determine if an unrecoverable read i

SYSTEM SYSLOG

- All our hosts' kernel, auth and iptables logs are sent centralized via Syslog to one location and file.
- Filebeat monitors and ships the data.
- Logstash separates the three datasets, parsing and enriching as necessary.
- Our goal is to monitor and detect hardware issues, unauthorized logins and network traffic rejections.
- This is one of the datasets we would like to apply machine learning anomaly detection.

LOGINS OVERVIEW DASHBOARD



Accepted - External SSH Login

user.name	source.ip	host.name	system.auth.ssh.method	Count
[redacted]	[redacted]	console	publickey	7

Failed - Internal SSH Login

user.name	source.ip	host.name	Count
[redacted]	[redacted]	backup	4
[redacted]	[redacted]	hsmpool08	2
[redacted]	[redacted]	hsmpool09	2

Successful SSH Logins / 10m

	Mean	Max	Total
Total	11.1	285	1.61 K
publickey	11.1	285	1.60 K
password	0.0208	1	3

Failed Reasons SSH Login Attempts / 10m

	Mean	Max	Total
Total	2.34	9	250
Invalid	2.03	8	213
Failed	3	4	30
Postponed	2	3	4
Close session:	3	3	3

Accepted - Internal SSH Login

user.name	source.ip	host.name	system.auth.ssh.methc	Count
[redacted]	[redacted]	orac2	publickey	1.18 K
[redacted]	[redacted]	orac1	publickey	128
[redacted]	[redacted]	orac2	publickey	97

Failed - SSH Login

user.name	source.ip	host.name	Count
Admin	[redacted]	console	41
admin	[redacted]	console	8
admin	[redacted]	console	7

CURRENT AND FUTURE WORK

- Migrating Elasticsearch to the new hardware.
- Cleanup of existing datasets and re-processing in some instances.
- Creating 'events' database for overlay on Grafana and classification.
- Creating of time aggregated datasets (e.g. 1hour bins) to both reduce storage usage and normalization.
- Creating Grafana alerts from existing dashboards.
- “Tokenizing” logs. (e.g. 1.1.1.1 -> <IPADDRESS>).
- Creating tools for testing and implementing machine learning tools like anomaly detection, classification, correlation, prediction.
- Identification of datasets that can be brought together to create "vectors" for correlation analysis.
- Investigate the use of GPUs for this type of work.

TRIUMF

Thank you
Merci

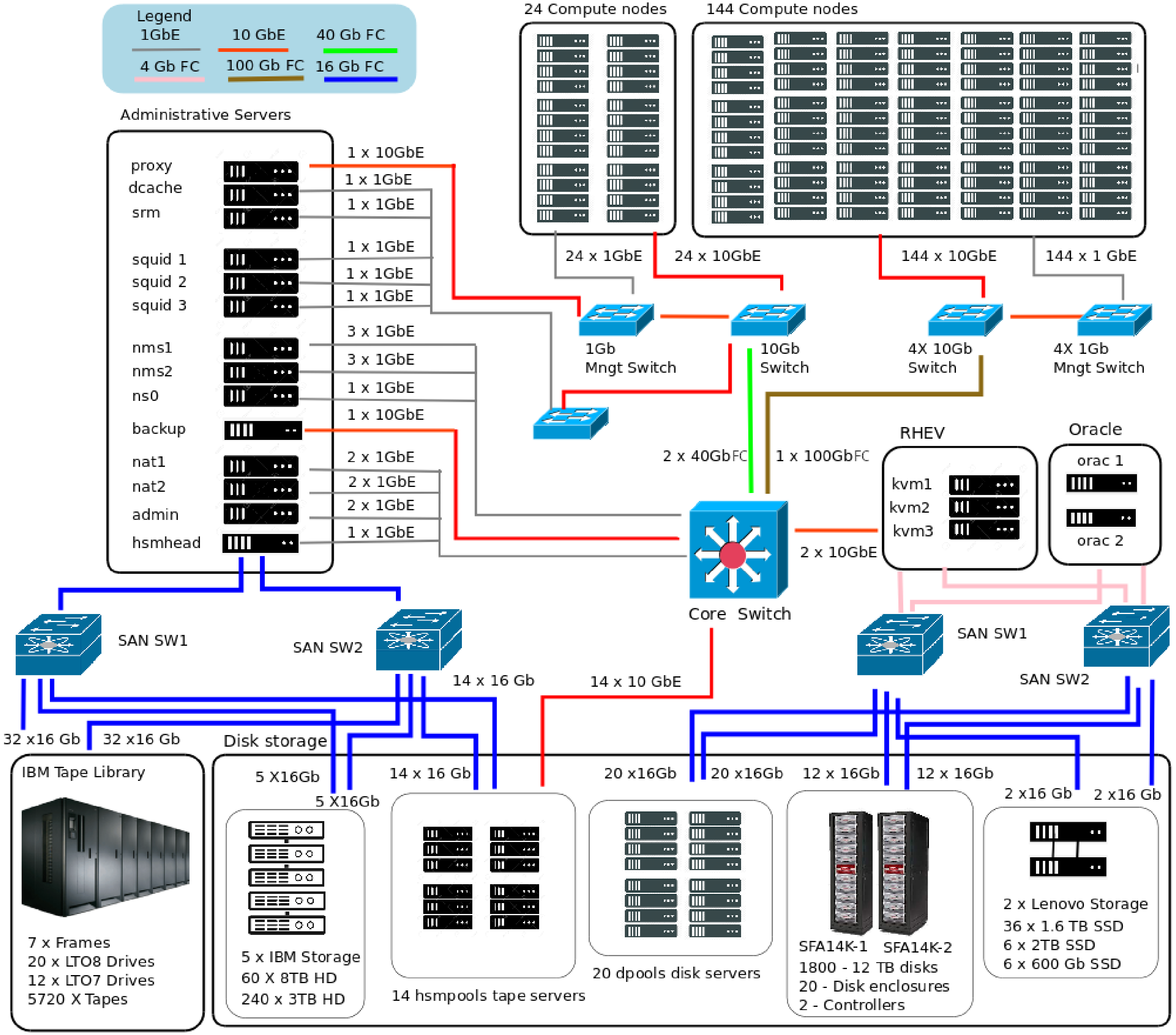
www.triumf.ca

Follow us **@TRIUMFLab**

Discovery,
accelerated

ADDITIONAL MATERIAL

TIER-1 INFRASTRUCTURE DIAGRAM



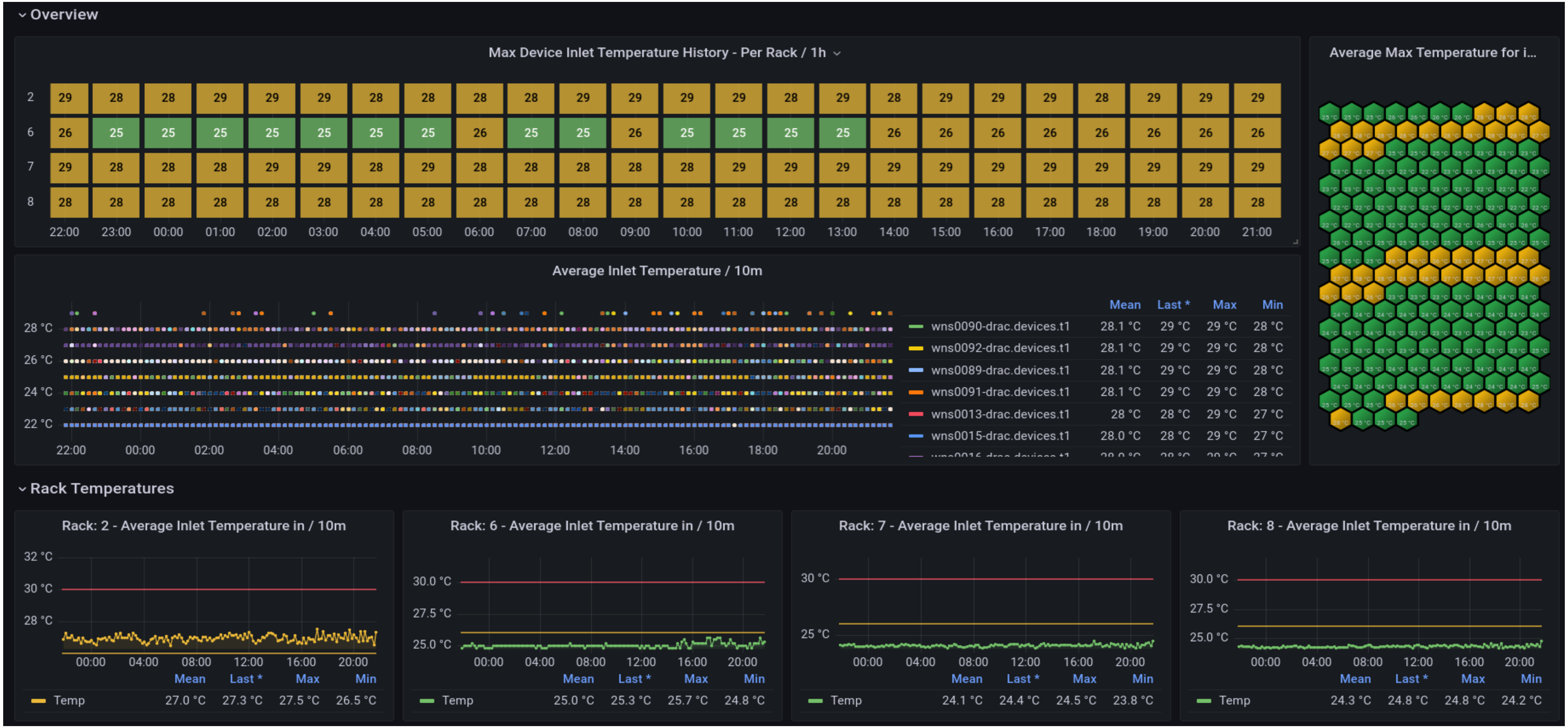
CORE ROUTER SFLOW

- Telegraf receives sflow data from our Juniper core router. Only a percentage sample of all data is captured due to its large magnitude.
- Data is stored on influxDB.
- One idea is to implement Snort/Suricata as an intrusion detection system.
- Logs would be sent to Elasticsearch

WORKER NODES INLET TEMPS

- A custom script queries all worker nodes' iDrac interfaces to obtain current temperature.
- It writes all information to a logfiles.
- Filebeat monitors and ships the data.
- Logstash parses these logs into fields to create Elasticsearch documents, enriching it with infrastructure data.

WORKER NODES INLET TEMPS



Overview

Max Device Inlet Temperature History - Per Rack / 1h

Rack	22:00	23:00	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00
2	29	28	28	29	29	29	28	28	28	28	29	29	29	29	28	29	28	29	29	29	28	29	29	29
6	26	25	25	25	25	25	25	25	26	25	25	26	25	25	25	25	26	26	26	26	26	26	26	26
7	29	28	28	28	29	29	28	29	28	28	28	28	28	29	29	29	29	29	29	29	29	29	29	29
8	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28

Average Max Temperature for i...

Average Inlet Temperature / 10m

	Mean	Last *	Max	Min
wns0090-drac.devices.t1	28.1 °C	29 °C	29 °C	28 °C
wns0092-drac.devices.t1	28.1 °C	29 °C	29 °C	28 °C
wns0089-drac.devices.t1	28.1 °C	29 °C	29 °C	28 °C
wns0091-drac.devices.t1	28.1 °C	29 °C	29 °C	28 °C
wns0013-drac.devices.t1	28 °C	28 °C	29 °C	27 °C
wns0015-drac.devices.t1	28.0 °C	28 °C	29 °C	27 °C

Rack Temperatures

Rack: 2 - Average Inlet Temperature in / 10m

	Mean	Last *	Max	Min
Temp	27.0 °C	27.3 °C	27.5 °C	26.5 °C

Rack: 6 - Average Inlet Temperature in / 10m

	Mean	Last *	Max	Min
Temp	25.0 °C	25.3 °C	25.7 °C	24.8 °C

Rack: 7 - Average Inlet Temperature in / 10m

	Mean	Last *	Max	Min
Temp	24.1 °C	24.4 °C	24.5 °C	23.8 °C

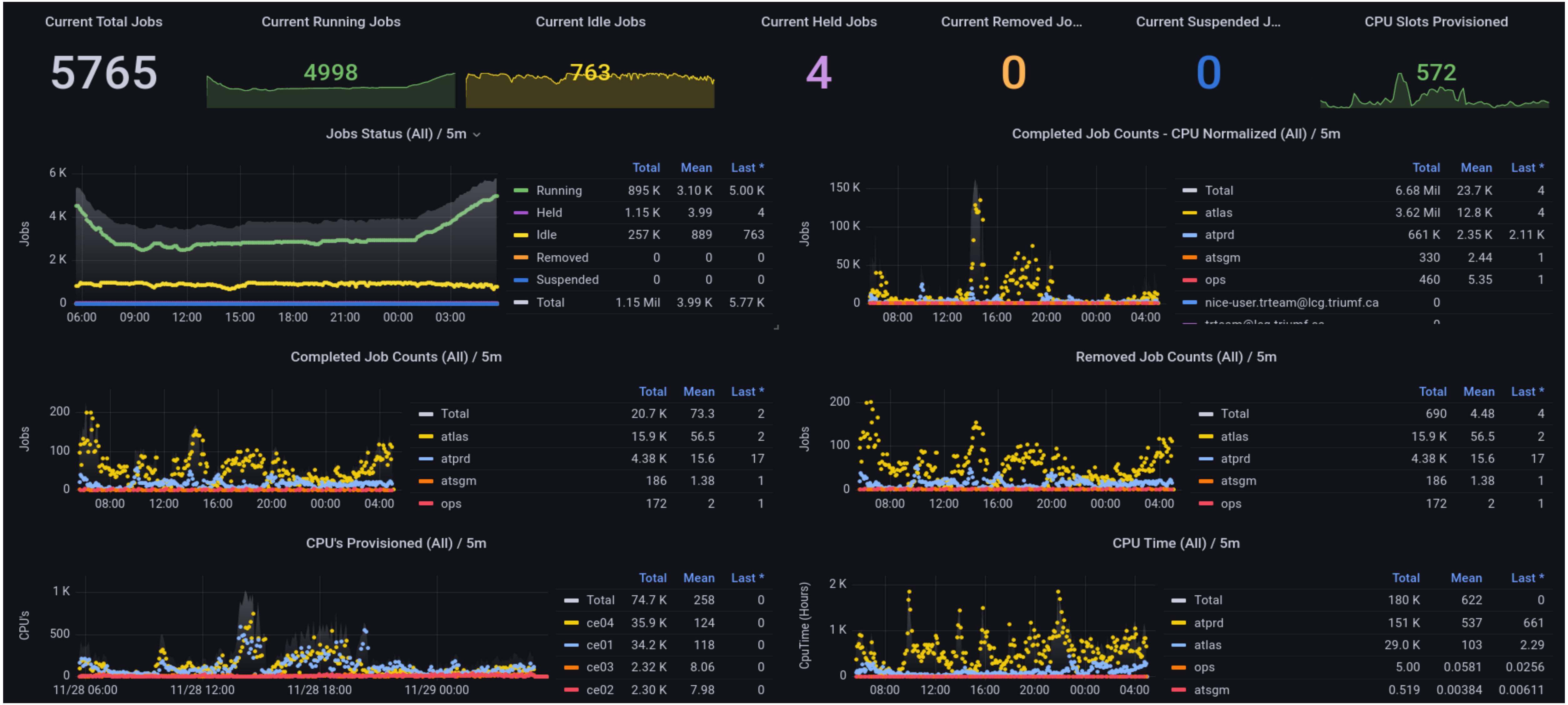
Rack: 8 - Average Inlet Temperature in / 10m

	Mean	Last *	Max	Min
Temp	24.3 °C	24.8 °C	24.8 °C	24.2 °C

HTCONDOR

- Two custom scripts query the HTCondor:
  - Every 15 minutes to obtain current jobs status.
  - Every 1 hour to obtain job history.
- Both write all information to two different logfiles.
- Filebeat monitors and ships the data.
- Logstash parses these logs into fields to create Elasticsearch documents.

HTCONDOR JOBS' STATUS



Current Total Jobs	Current Running Jobs	Current Idle Jobs	Current Held Jobs	Current Removed Jo...	Current Suspended J...	CPU Slots Provisioned
5765	4998	763	4	0	0	572

Jobs Status (All) / 5m

	Total	Mean	Last *
Running	895 K	3.10 K	5.00 K
Held	1.15 K	3.99	4
Idle	257 K	889	763
Removed	0	0	0
Suspended	0	0	0
Total	1.15 Mil	3.99 K	5.77 K

Completed Job Counts - CPU Normalized (All) / 5m

	Total	Mean	Last *
Total	6.68 Mil	23.7 K	4
atlas	3.62 Mil	12.8 K	4
atprd	661 K	2.35 K	2.11 K
atsgm	330	2.44	1
ops	460	5.35	1
nice-user.trteam@lcg.triumf.ca	0		
trteam@lcg.triumf.ca	0		

Completed Job Counts (All) / 5m

	Total	Mean	Last *
Total	20.7 K	73.3	2
atlas	15.9 K	56.5	2
atprd	4.38 K	15.6	17
atsgm	186	1.38	1
ops	172	2	1

Removed Job Counts (All) / 5m

	Total	Mean	Last *
Total	690	4.48	4
atlas	15.9 K	56.5	2
atprd	4.38 K	15.6	17
atsgm	186	1.38	1
ops	172	2	1

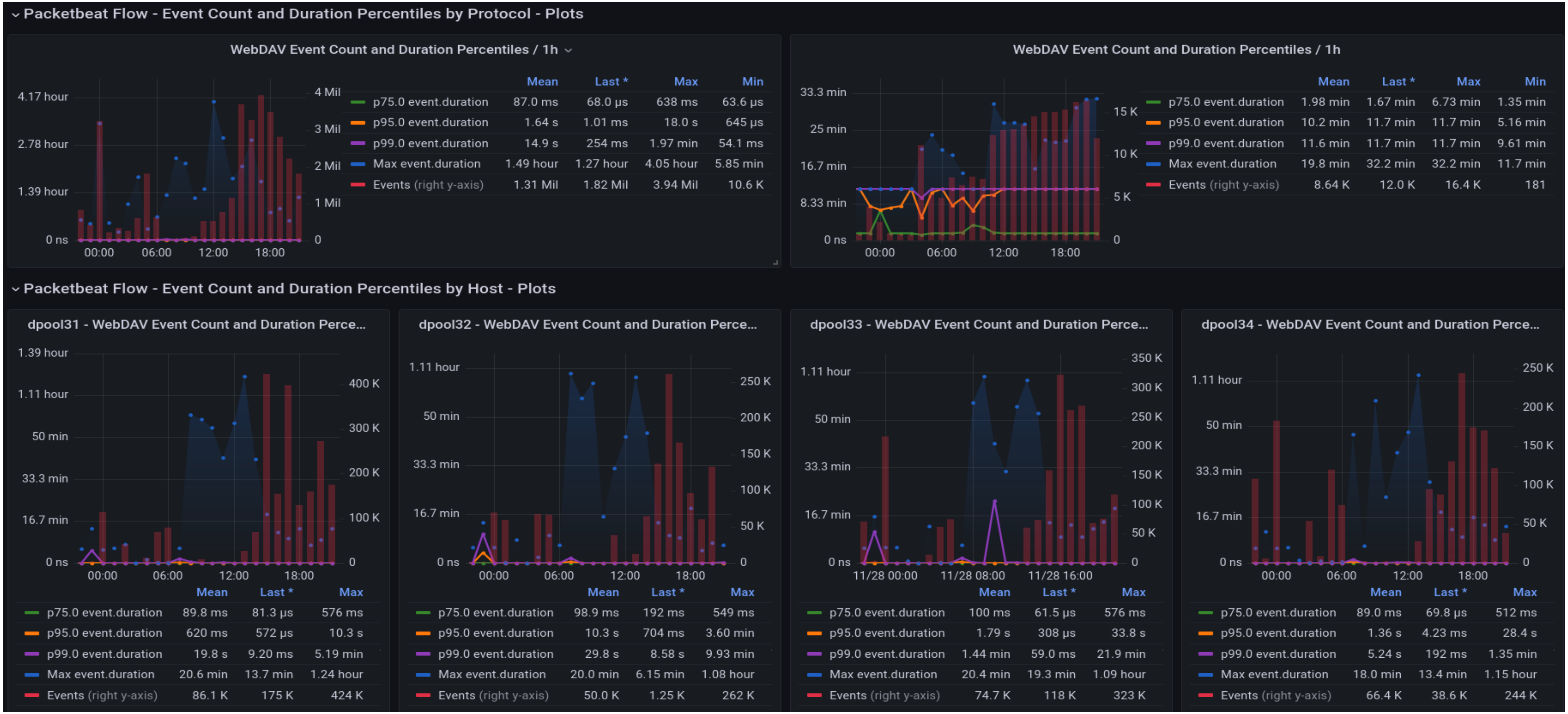
CPU's Provisioned (All) / 5m

	Total	Mean	Last *
Total	74.7 K	258	0
ce04	35.9 K	124	0
ce01	34.2 K	118	0
ce03	2.32 K	8.06	0
ce02	2.30 K	7.98	0

CPU Time (All) / 5m

	Total	Mean	Last *
Total	180 K	622	0
atprd	151 K	537	661
atlas	29.0 K	103	2.29
ops	5.00	0.0581	0.0256
atsgm	0.519	0.00384	0.00611

DCACHE PROTOCOLS (PACKETBEAT)



Packetbeat Flow - Event Count and Duration Percentiles by Protocol - Plots

WebDAV Event Count and Duration Percentiles / 1h

	Mean	Last *	Max	Min
p75.0 event.duration	87.0 ms	68.0 µs	638 ms	63.6 µs
p95.0 event.duration	1.64 s	1.01 ms	18.0 s	645 µs
p99.0 event.duration	14.9 s	254 ms	1.97 min	54.1 ms
Max event.duration	1.49 hour	1.27 hour	4.05 hour	5.85 min
Events (right y-axis)	1.31 Mil	1.82 Mil	3.94 Mil	10.6 K

WebDAV Event Count and Duration Percentiles / 1h

	Mean	Last *	Max	Min
p75.0 event.duration	1.98 min	1.67 min	6.73 min	1.35 min
p95.0 event.duration	10.2 min	11.7 min	11.7 min	5.16 min
p99.0 event.duration	11.6 min	11.7 min	11.7 min	9.61 min
Max event.duration	19.8 min	32.2 min	32.2 min	11.7 min
Events (right y-axis)	8.64 K	12.0 K	16.4 K	181

Packetbeat Flow - Event Count and Duration Percentiles by Host - Plots

dpool31 - WebDAV Event Count and Duration Perce...

	Mean	Last *	Max
p75.0 event.duration	89.8 ms	81.3 µs	576 ms
p95.0 event.duration	620 ms	572 µs	10.3 s
p99.0 event.duration	19.8 s	9.20 ms	5.19 min
Max event.duration	20.6 min	13.7 min	1.24 hour
Events (right y-axis)	86.1 K	175 K	424 K

dpool32 - WebDAV Event Count and Duration Perce...

	Mean	Last *	Max
p75.0 event.duration	98.9 ms	192 ms	549 ms
p95.0 event.duration	10.3 s	704 ms	3.60 min
p99.0 event.duration	29.8 s	8.58 s	9.93 min
Max event.duration	20.0 min	6.15 min	1.08 hour
Events (right y-axis)	50.0 K	1.25 K	262 K

dpool33 - WebDAV Event Count and Duration Perce...

	Mean	Last *	Max
p75.0 event.duration	100 ms	61.5 µs	576 ms
p95.0 event.duration	1.79 s	308 µs	33.8 s
p99.0 event.duration	1.44 min	59.0 ms	21.9 min
Max event.duration	20.4 min	19.3 min	1.09 hour
Events (right y-axis)	74.7 K	118 K	323 K

dpool34 - WebDAV Event Count and Duration Perce...

	Mean	Last *	Max
p75.0 event.duration	89.0 ms	69.8 µs	512 ms
p95.0 event.duration	1.36 s	4.23 ms	28.4 s
p99.0 event.duration	5.24 s	192 ms	1.35 min
Max event.duration	18.0 min	13.4 min	1.15 hour
Events (right y-axis)	66.4 K	38.6 K	244 K