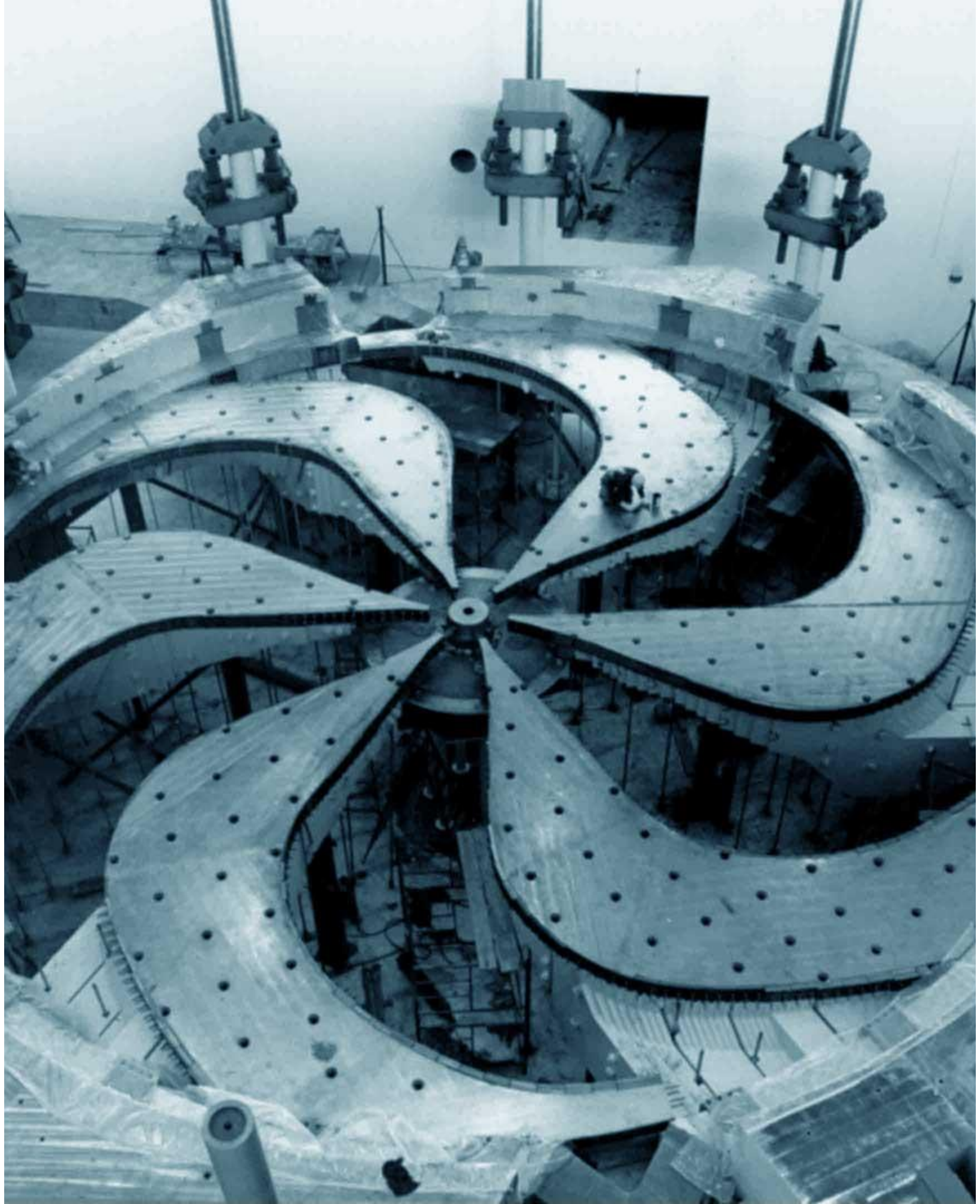
# Canadian ATLAS Tier-1 Analytics Infrastructure

Fernando Fernandez Galindo

TRIUMF Scientific Computing Department
ATLAS Tier-1 Group

HEPiX Autumn 2023 Workshop

October 16-20, 2023



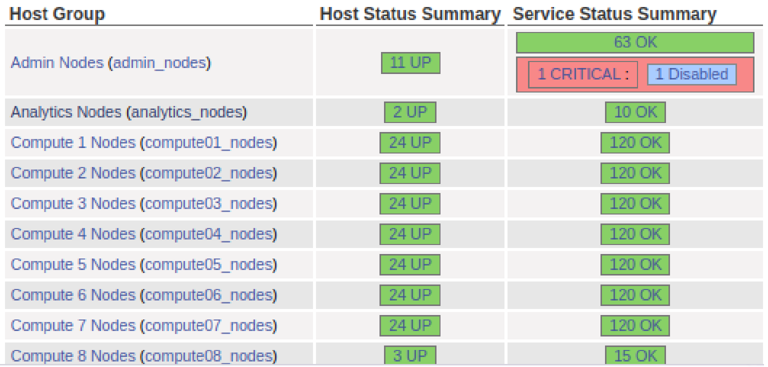


Discovery, accelerated

# ANALYTICS PROJECT MOTIVATION

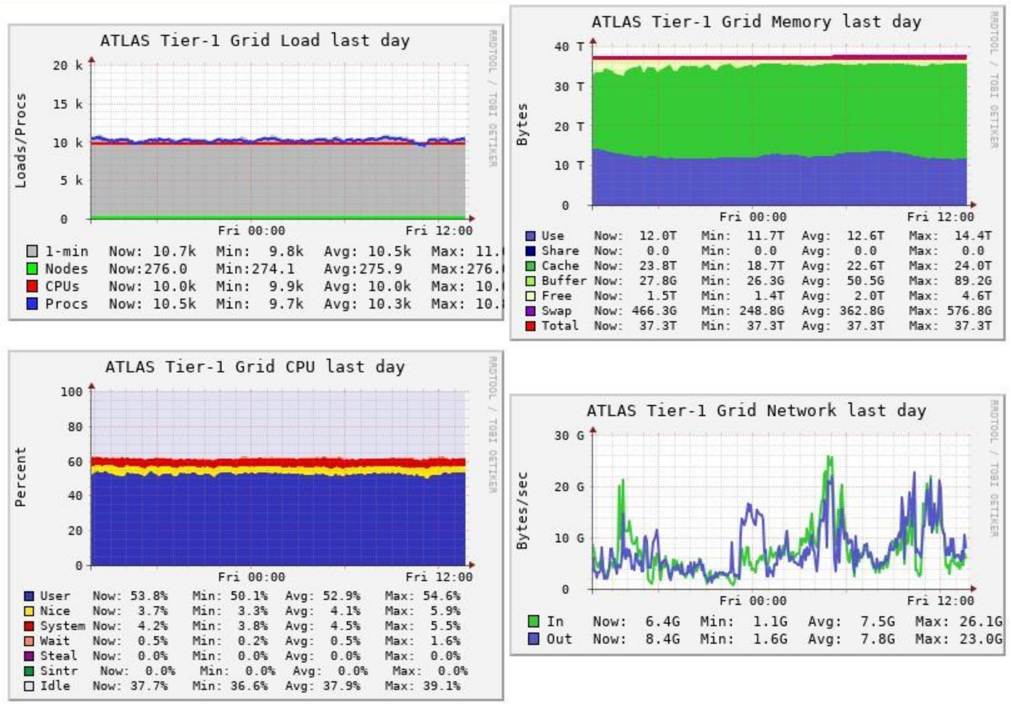
- With growing infrastructure an ever-increasing collection of heterogeneous monitoring data is produced.
- Our existing monitoring and alerting implementation is robust and stable but rather static and isolated.
- Early 2020 we started the analytics project on hardware that was deprecated.
- The project's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

**Nagios**



| Host Group | Host Status Summary | Service Status Summary |
|---|---|---|
| Admin Nodes (admin_nodes) | 11 UP | 63 OK<br>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**




Fernando Fernandez Galindo
TRIUMF Scientific Computing Department
HEPiX Autumn 2023 Workshop
October 16-20, 2023

# 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)

All our software is deployed using Ansible.

*Purple denotes additions from the analytics project.

# DATASETS

## Logs

| Service | Size (GB) | Storage DB |
|---|---|---|
| dCache (billing, ftp\|srm\|webdav\|xrootd access) | 9,850 | Elasticsearch |
| network (router) | 30 | Elasticsearch |
| SNMP (traps) | 15 | MariaDB |
| system (auth, iptables, kernel) | 1,000 | Elasticsearch |

## ELK Ingestion Rate

| Service | events/sec |
|---|---|
| Mean | 600 |
| Max | 12,000 |

## Metrics

| Service | Size (GB) | Storage |
|---|---|---|
| dCache (queues, movers) | 215 | Elasticsearch |
| dCache (netflows) | 110 | Elasticsearch |
| HTCondor (job history, status) | 32 | Elasticsearch |
| infrastructure (DDN and inlet temps, SSD TBW) | 160 | Elasticsearch |
| infrastructure (humidity, PDU, temps, etc) | 2* | RRD |
| mySQL (status) | 4 | Elasticsearch |
| network (router sflow) | 15 | influxDB |
| postgreSQL (activity, bgwriter, database) | 920 | Elasticsearch |
| system (cpu, mem, net, etc) | 2* | RRD |
| tape library stats (device, volume) | 2 | influxDB |
| tape library (consumption, staging, etc) | 2* | RRD |

* 2GB for all datasets in RRD

# GRAFANA

- Currently using version 10.0.3
- 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.

# ELASTIC SUITE

elastic

- It is the workhorse of the analytics platform.
- We obtained a trial for the full license to investigate their ML tools but due to technical limitations, we didn't get much out of it. Pricing was very expensive to keep the license.
- We use the free 'basic' license.
- Currently using version 8.3.3.
- **Elasticsearch**:
  - Flexible database, can hold heterogeneous data.
  - Easy to grow horizontally.
  - Many tools to manage 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 unavailable.

# ELK INFRASTRUCTURE

elastic

- Data collection consists of Filebeat, Metricbeat and Packetbeat instances installed on each host of interest.
- Data is sent to Logstash for processing before being stored in Elasticsearch.
- The Elasticsearch cluster consists of 9 instances running on VM's.
- Kibana shares a VM and is mostly used for administration of the cluster.

- There are 5 types of Elasticsearch instances:
  - 1x Voting-only Master Node (also running Kibana and main endpoint)
  - 2x Master Nodes.
  - 2x "Hot" Data Nodes (nvme storage)
  - 2x "Warm" Data Nodes (ssd storage)
  - 2x Ingest and Transform Nodes (also running Logstash)

# 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).

# ELK INFRASTRUCTURE

## ahw-fe01

### analytics

CPU: 24

GPU: Nvidia Tesla T4

Network: 2x Mellanox VFS (10Gb)

Memory: 220GB

Storage: QCOW2 (from SSD):
- Data 100GB

NVMe 3.84T x 2

Roles: master, Kibana

elastic

## ahw-kvm01

### es-data-hot01

CPU: 10

Memory: 92GB

Network: 2x Mellanox VFS (10Gb)

Roles: data-content, data-hot

Storage: NVMe 3.84T x 2

### es-data-warm01

CPU: 10

Memory: 80GB

Network: 2x Mellanox VFS (10Gb)

Roles: data-warm

Storage: SSD 7.68TB x 2

### es-main01

CPU: 4

Memory: 8GB

Network: 2x Mellanox VFS (10Gb)

Roles: master

### es-tr01

CPU: 8

Memory: 52GB

Network: 2x Mellanox VFS (10Gb)

Roles: ingest, remote-cluster-client, transform, Logstash

## ahw-kvm02

### es-data-hot02

CPU: 10

Memory: 92GB

Network: 2x Mellanox VFS (10Gb)

Roles: data-content, data-hot

Storage: NVMe 3.84T x 2

### es-data-warm02

CPU: 10

Memory: 80GB

Network: 2x Mellanox VFS (10Gb)

Roles: data-warm

Storage: SSD 7.68TB x 2

### es-main02

CPU: 4

Memory: 8GB

Network: 2x Mellanox VFS (10Gb)

Roles: master

### es-tr02

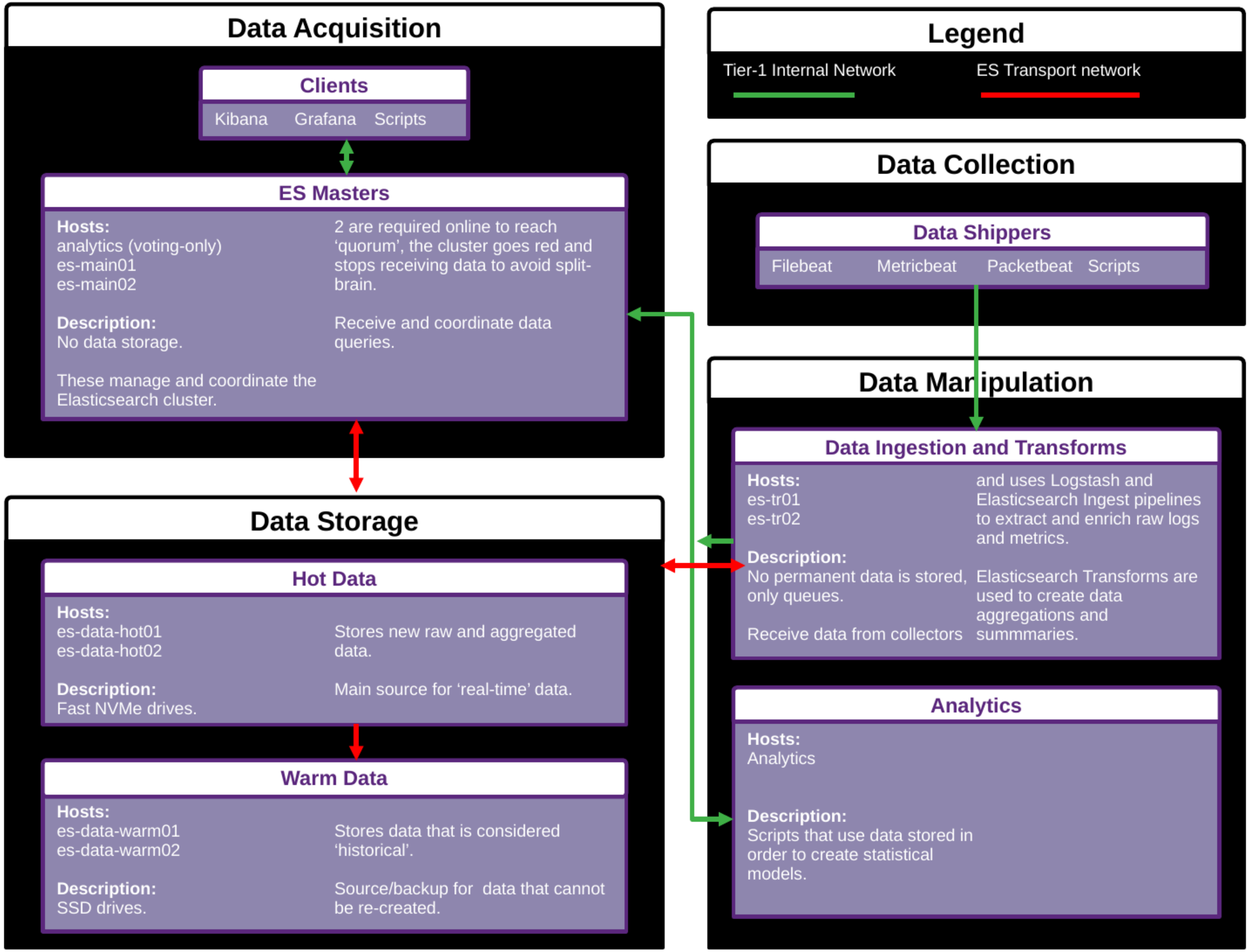
CPU: 8

Memory: 52GB

Network: 2x Mellanox VFS (10Gb)

Roles: ingest, remote-cluster-client, transform, Logstash

# ELK PIPELINE

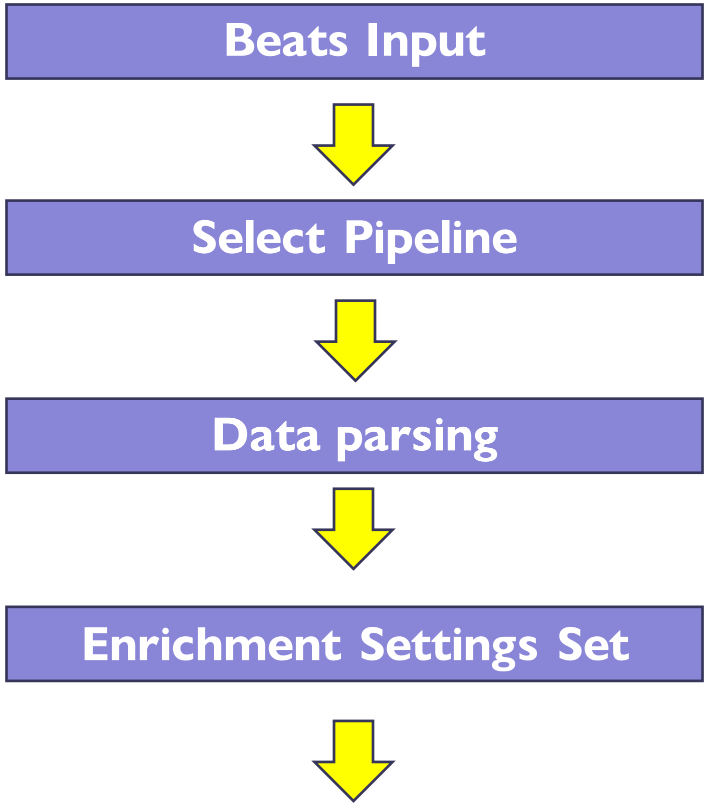


# LOGSTASH PIPELINE - PARSING

1. Logstash receives data from the beats
2. A pipeline name is injected in the configuration specific to each of the different logs.

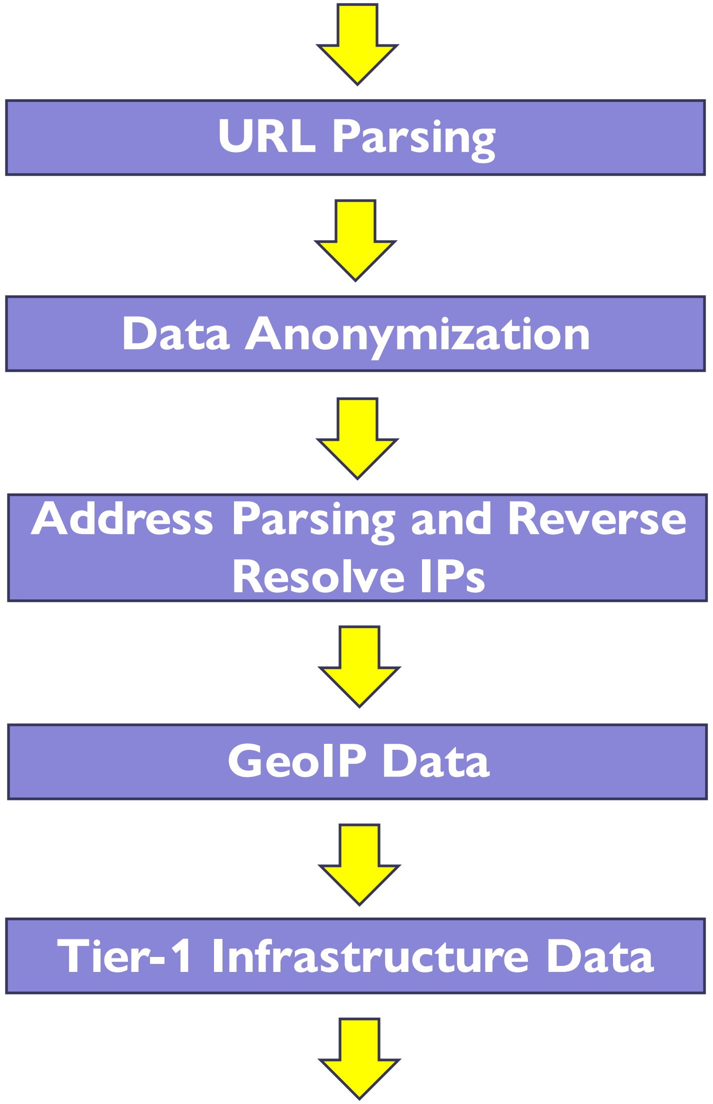
   Example:

   - tier1-logs-dcache-billing-log-v4
3. Data is parsed into fields for each dataset.
4. Enrichment settings are applied, so that the pipeline knows if each of the next steps should be applied or not (shown in the next slide).



# LOGSTASH PIPELINE - ENRICHMENT

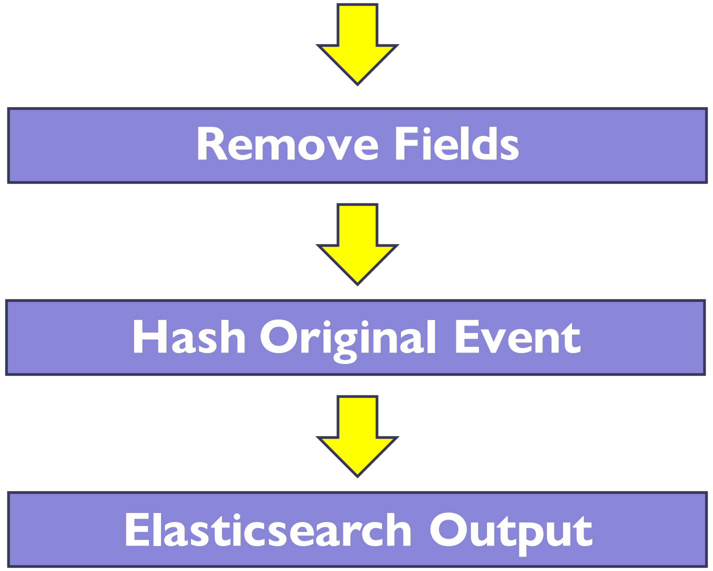
1. URLs parsed into fields for protocol, hostname, path, query.
2. Anonymize data (specific to each pipeline).
3. Parse addresses and reverse resolve IP's.
4. Resolve DNS names.
5. Add GeoIP data.
6. Add Tier-1 Infrastructure location information (datacenter, rack, row, etc).




Fernando Fernandez Galindo
TRIUMF Scientific Computing Department

# LOGSTASH PIPELINE – FINAL STAGE

1. Remove unwanted fields like metadata from the Beats or just fields we don't really need.
2. Create hash of the original message to use for deduplication and as document ID.
3. Output to Elasticsearch.



# TOKENIZING - BEFORE

- We are drowned by the details.
- Unique error message count: **3,423**

Tier-1 - Logs - dChache - Billing - Error Messages

| Top 1000 values of | Top 1000 values of error.message.keyword | Count of records |
|---|---|---|
| 10006 | Flush was cancelled. | 329,292 |
| 10006 | Stage was cancelled. | 130,937 |
| 10006 | Request to [>PoolManager@dCacheDomain] timed out. | 103,987 |
| 10006 | Failed to select pool: Request to [>PoolManager@dCacheDomain] timed out. | 31,541 |
| 10006 | No connection from client after 300 seconds. Giving up. | 13,517 |
| 10006 | No redirect from mover on pool PoolName=sfa14kx_1_lun8 PoolAddress=sfa14kx_1_lun8@sfa14kx_1_lun8 after 3 min | 10,787 |
| 10006 | No redirect from mover on pool PoolName=sfa14kx_1_lun68 PoolAddress=sfa14kx_1_lun68@sfa14kx_1_lun68 after 3 min | 10,437 |
| 10006 | No redirect from mover on pool PoolName=hsmnostage PoolAddress=hsmnostage@hsmpool14_nostage after 3 min | 3,291 |
| 10006 | No redirect from mover on pool PoolName=hsmpool19_sfa14_2_lun48_2 PoolAddress=hsmpool19_sfa14_2_lun48_2@hsmpool19_sfa14_ | 2,770 |
| 10006 | No redirect from mover on pool PoolName=hsmpool19_sfa14_2_lun48 PoolAddress=hsmpool19_sfa14_2_lun48@hsmpool19_sfa14_2_lur | 2,715 |
| 10006 | No redirect from mover on pool PoolName=hsmpool18_sfa14_1_lun48_2 PoolAddress=hsmpool18_sfa14_1_lun48_2@hsmpool18_sfa14_ | 2,690 |
| 10006 | No redirect from mover on pool PoolName=sfa14kx_2_lun86 PoolAddress=sfa14kx_2_lun86@sfa14kx_2_lun86 after 3 min | 2,404 |
| 10006 | No redirect from mover on pool PoolName=hsmpool20_sfa14_1_lun50_2 PoolAddress=hsmpool20_sfa14_1_lun50_2@hsmpool20_sfa14_ | 2,213 |
| 10006 | No redirect from mover on pool PoolName=sfa14kx_2_lun43 PoolAddress=sfa14kx_2_lun43@sfa14kx_2_lun43 after 3 min | 2,050 |

# TOKENIZING

- By "tokenizing" we mean normalizing log messages by applying a placeholder or "token" instead of highly variable data.
- For example, IPs and Hostnames are replaced with "IP" or "HOSTNAME" or in dCache cell names are replaced with "CELL_NAME".
- Therefore, we get the base error message which can then be aggregated, used in statistics, clean up dashboards and allow for more visibility on other error messages.
- The next slide shows and example of dCache billing data from August and September 2023.
- **Notes:**
  - Requires regular expressions specific to each dataset as trying to do general substitution failed.
  - This is work in progress, only 'completed' for dCache Billing logs and tested in preproduction (but with production datasets).

# TOKENIZING - AFTER

- Better visibility of all errors.
- Unique error message count: **124**

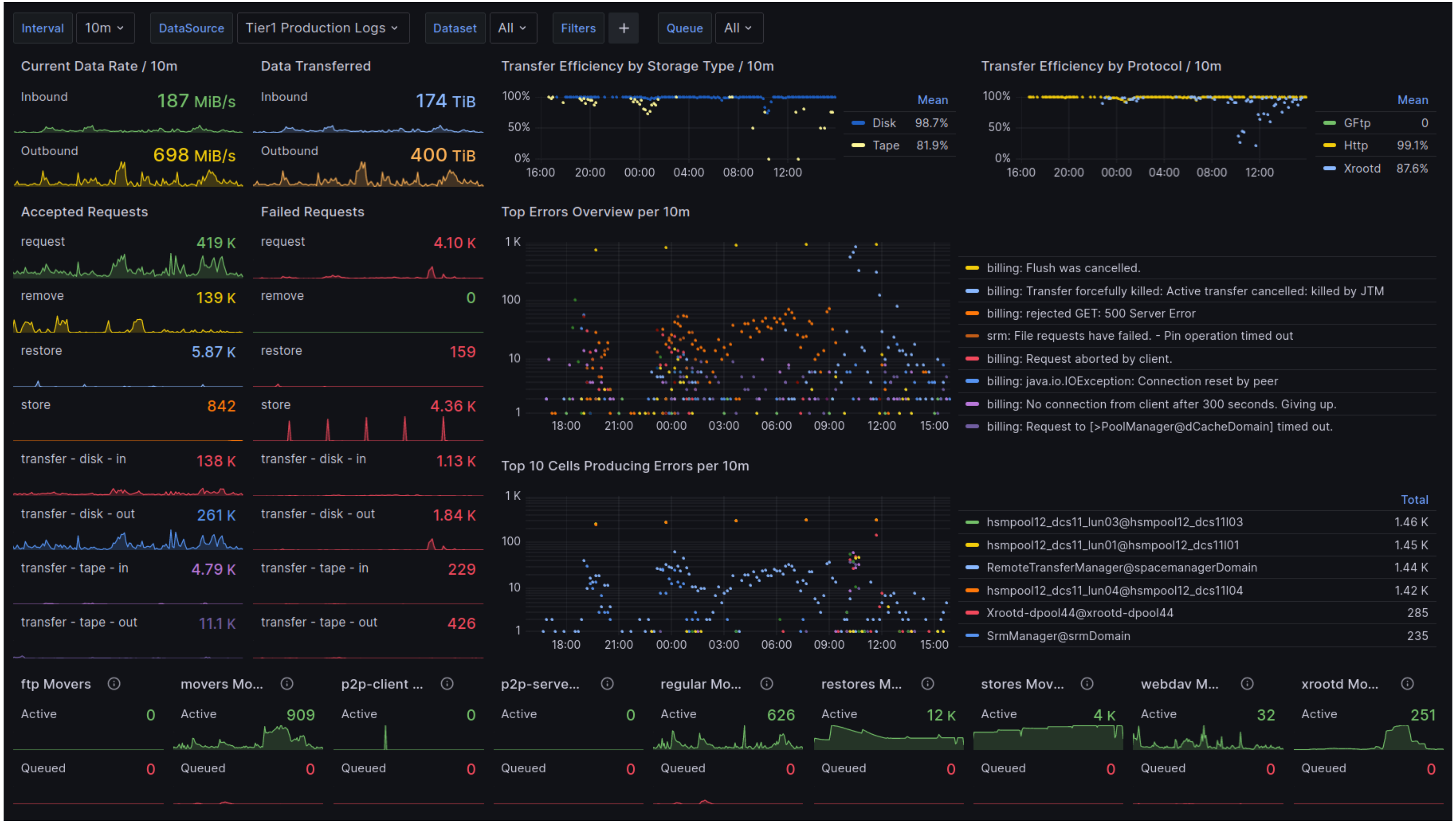
| Top 1000 values of | Top 1000 values of error.message.keyword | Count of records |
|---|---|---|
| 10006 | Flush was cancelled. | 260,899 |
| 10006 | Stage was cancelled. | 104,869 |
| 10006 | Request to [>CELL_ADDRESS] timed out. | 75,804 |
| 10006 | No redirect from mover on pool PoolName=CELL_NAME PoolAddress=CELL_ADDRESS after 3 min | 49,270 |
| 10006 | Failed to select pool: Request to [>CELL_ADDRESS] timed out. | 21,632 |
| 10006 | No connection from client after 300 seconds. Giving up. | 9,768 |
| 10006 | Transfer killed by door due to failure for mover PoolName=CELL_NAME PoolAddress=CELL_ADDRESS: Request to [>CELL_ADDRESS] timec | 189 |
| 10006 | Request to [>CELL_ADDRESS:CELL_ADDRESS] timed out. | 181 |
| 10006 | Transfer killed by door due to failure for mover PoolName=CELL_NAME PoolAddress=CELL_ADDRESS: (0) Job not found : JOB_ID | 55 |
| 10006 | Failed to deliver message <NUM:NUM> to [>CELL_ADDRESS]: CELL_ADDRESS is busy (its estimated response time of NUM ms is longer tł | 31 |
| 10006 | Failed in state -1: Request to [>CELL_ADDRESS] timed out. [10006] | 17 |
| 10006 | Failed to select pool: Stage was cancelled. | 3 |
| 10006 | Staging timed out | 2 |
| 10006 | Failed to select pool: Staging timed out | 1 |

# 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 and is aggregated in 1m in and 1h datasets.
- A custom script parses dCache's webadmin pool queue table and sends it to Elasticsearch.

Fernando Fernandez Galindo
TRIUMF Scientific Computing Department

# DCACHE OVERVIEW DASHBOARD




Fernando Fernandez Galindo
TRIUMF Scientific Computing Department

# TAPE LIBRARY (HSM)

- The library produces SNMP traps when there are failures which are stored in a MariaDB instance from where Grafana queries the information directly.
- Another MariaDB instance that records the tape library's devices and volumes actions.
- A custom script extracts the data and sends it to influxDB where we can manipulate it for later visualization.

# TAPE LIBRARY DASHBOARD

Overview

| Completed... | Bytes Writt... |
|---|---|
| 131.46 K | 1.63 TiB |

| Failed Acti... | Bytes Read |
|---|---|
| 10.52 K | 2.26 TiB |

| Critical Tra... | Warning Tr... |
|---|---|
| 4 | 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 |

Total Device Actions. Dev: All


All traps

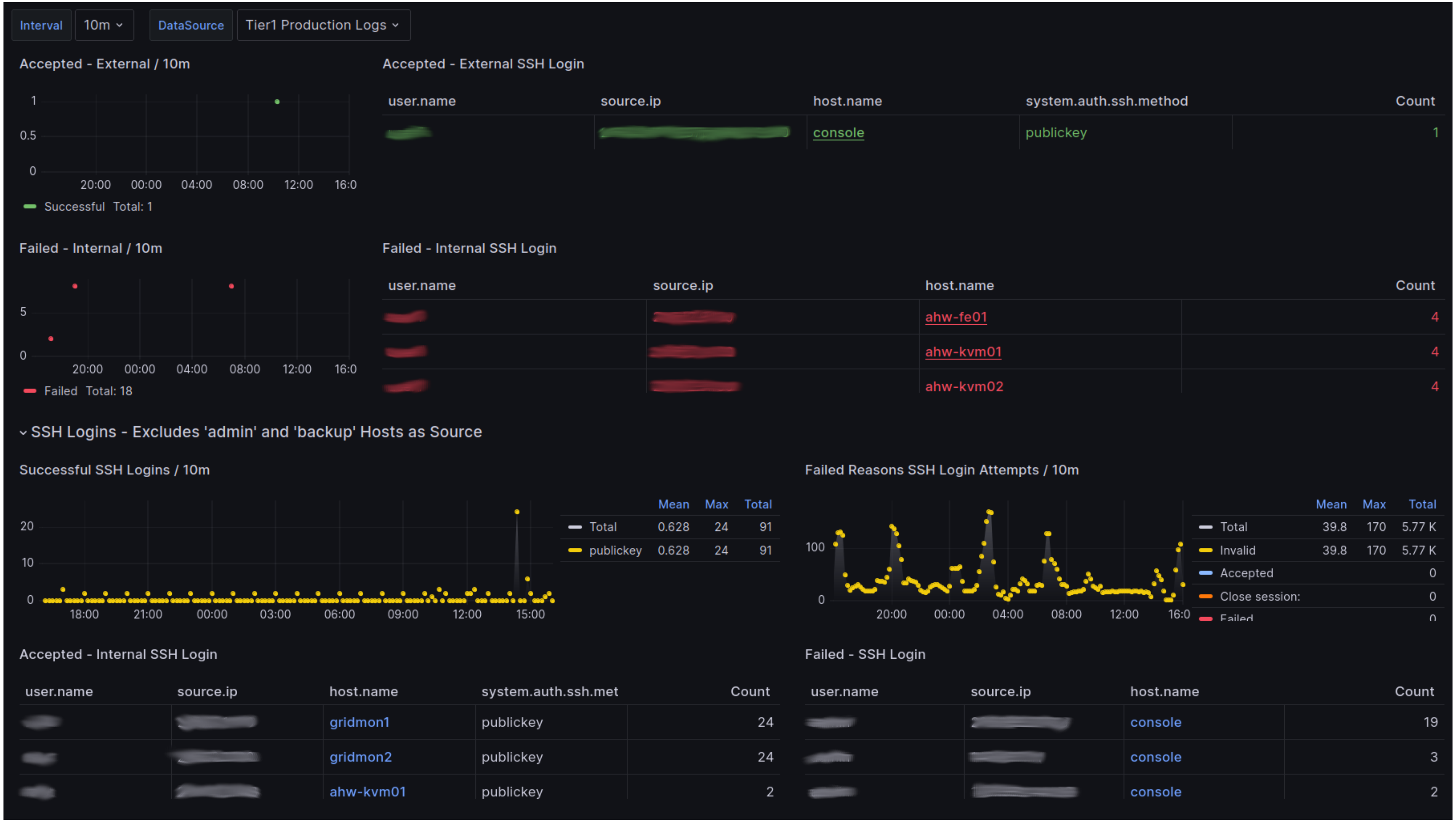

**SNMP trap description**

| 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 (auth, iptables, syslog), 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



# LOGINS HOST DASHBOARD

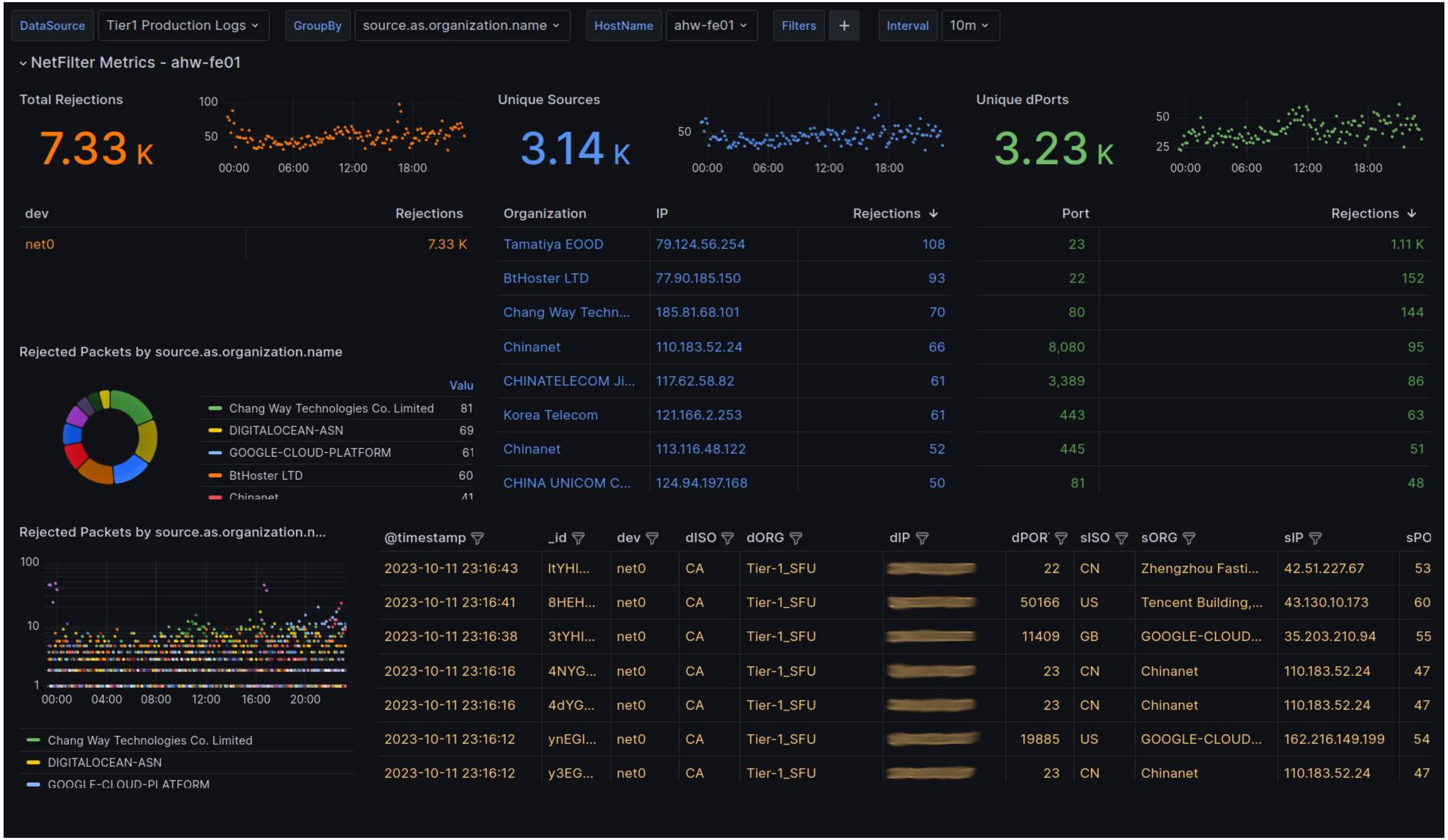

Fernando Fernandez Galindo
TRIUMF Scientific Computing Department
HEPiX Autumn 2023 Workshop
October 16-20, 2023

# IPTABLES OVERVIEW DASHBOARD



Fernando Fernandez Galindo
TRIUMF Scientific Computing Department

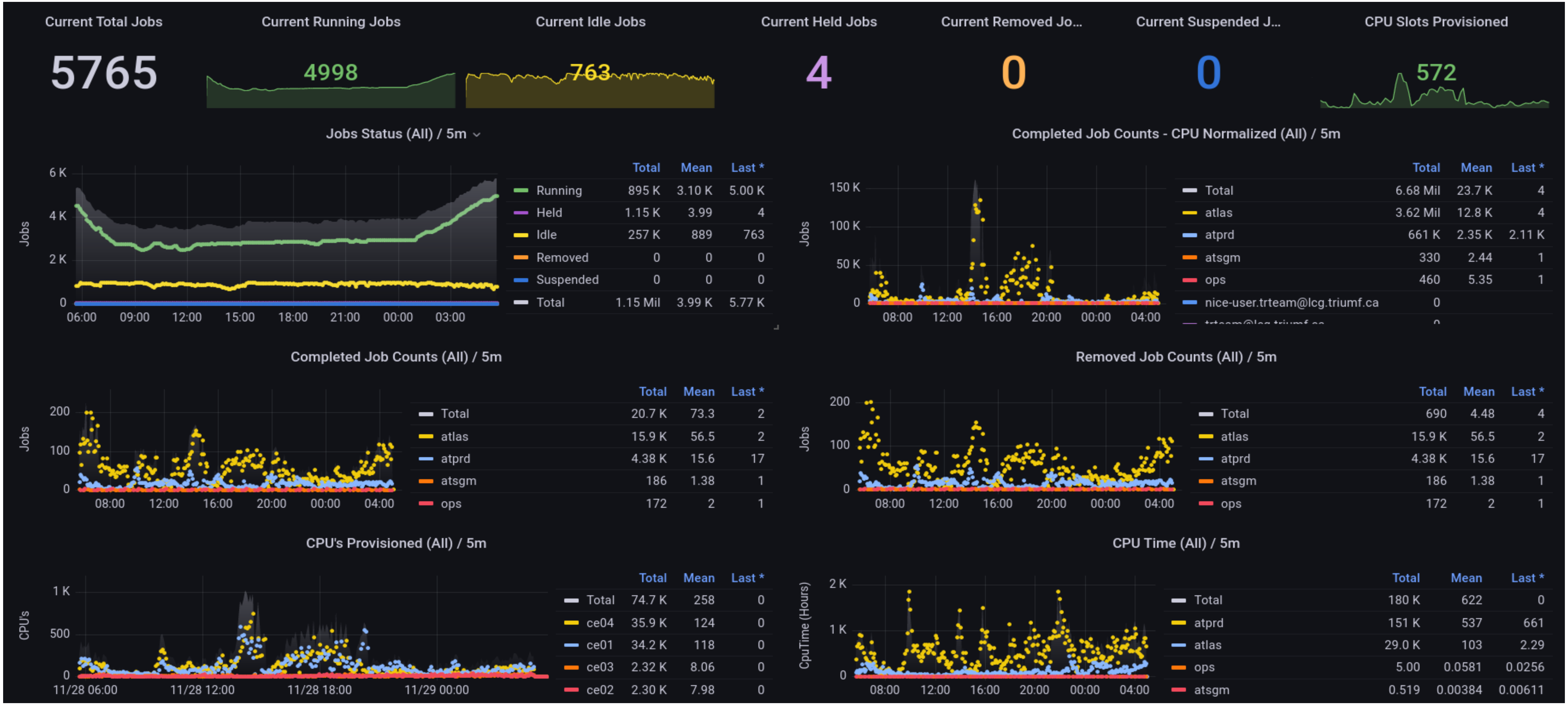
# IPTABLES HOST DASHBOARD



# HTCONDOR

- Two custom python 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.

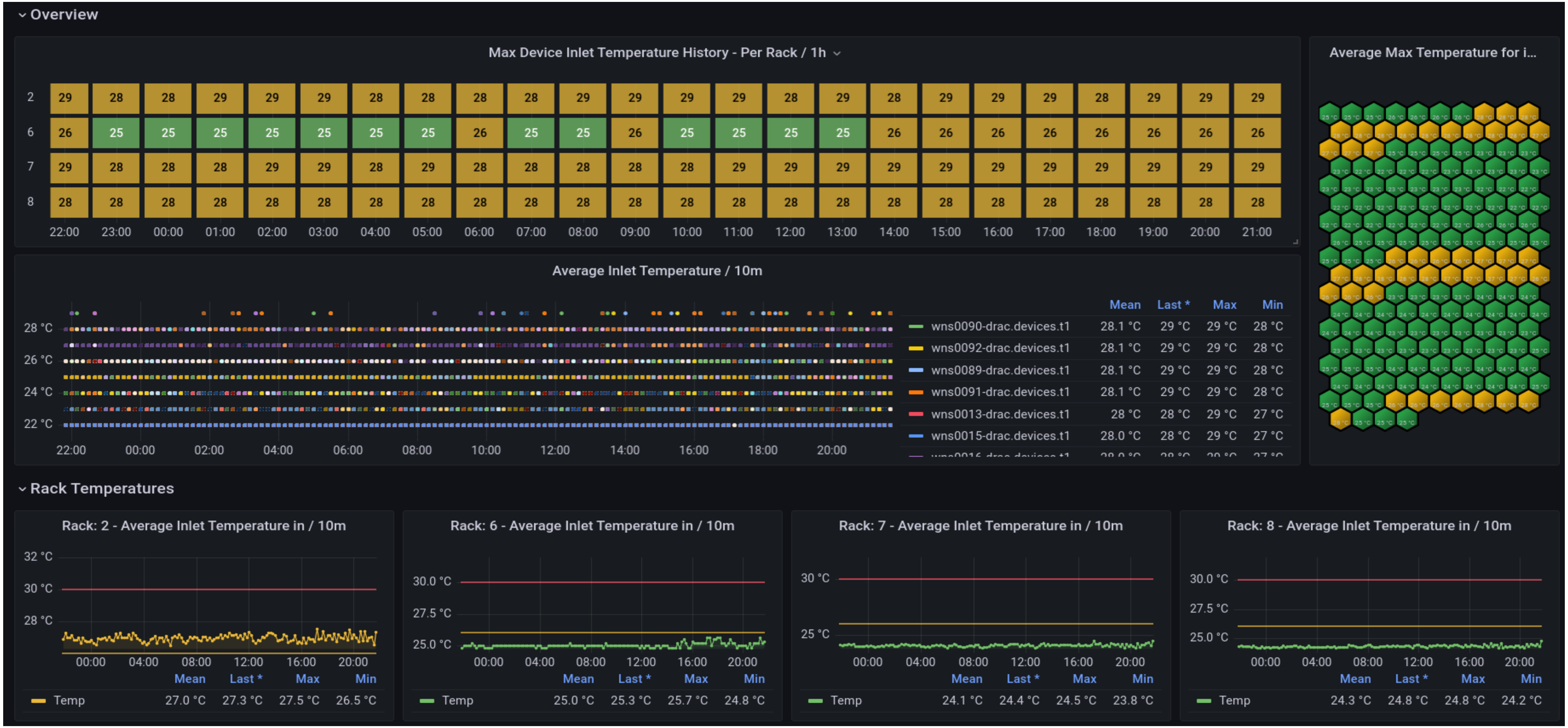
# HTCONDOR JOBS' STATUS



# 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





Fernando Fernandez Galindo
TRIUMF Scientific Computing Department
HEPiX Autumn 2023 Workshop
October 16-20, 2023

# CURRENT AND FUTURE WORK

- 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” more datasets.
- 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

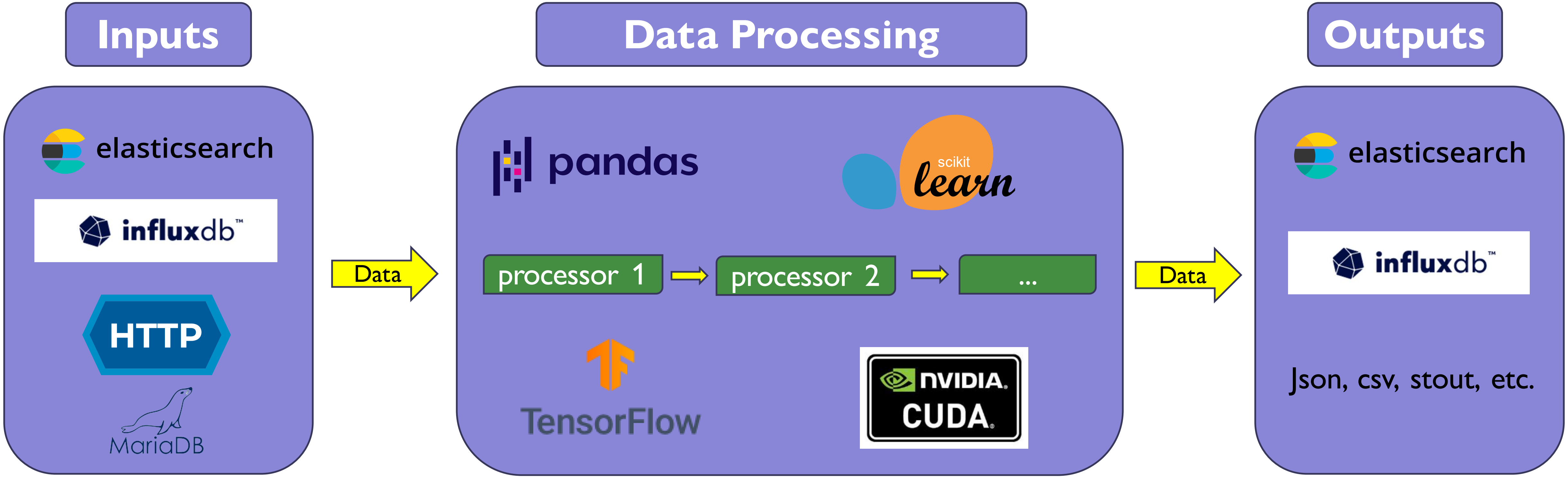
Follow us **@TRIUMFLab**

Discovery,
accelerated

# ADDITIONAL MATERIAL

# AFW PYTHON PACKAGE (WIP)

Python package/module for obtaining/sending data to/from given databases using custom "Request" objects and processing it with a desired tool via "processors".



# AFW CONFIG FILE EXAMPLE

Request objects configured via JSON files. Can pass configuration arguments and database queries. Example:

```
{
  "nodes": [
    {
      "src": {
        "name": "src",
        "type": "elasticsearch",
        "url": "http://localhost:9200"
      }
    },
    {
      "dst": { … }
    }
  ],
```

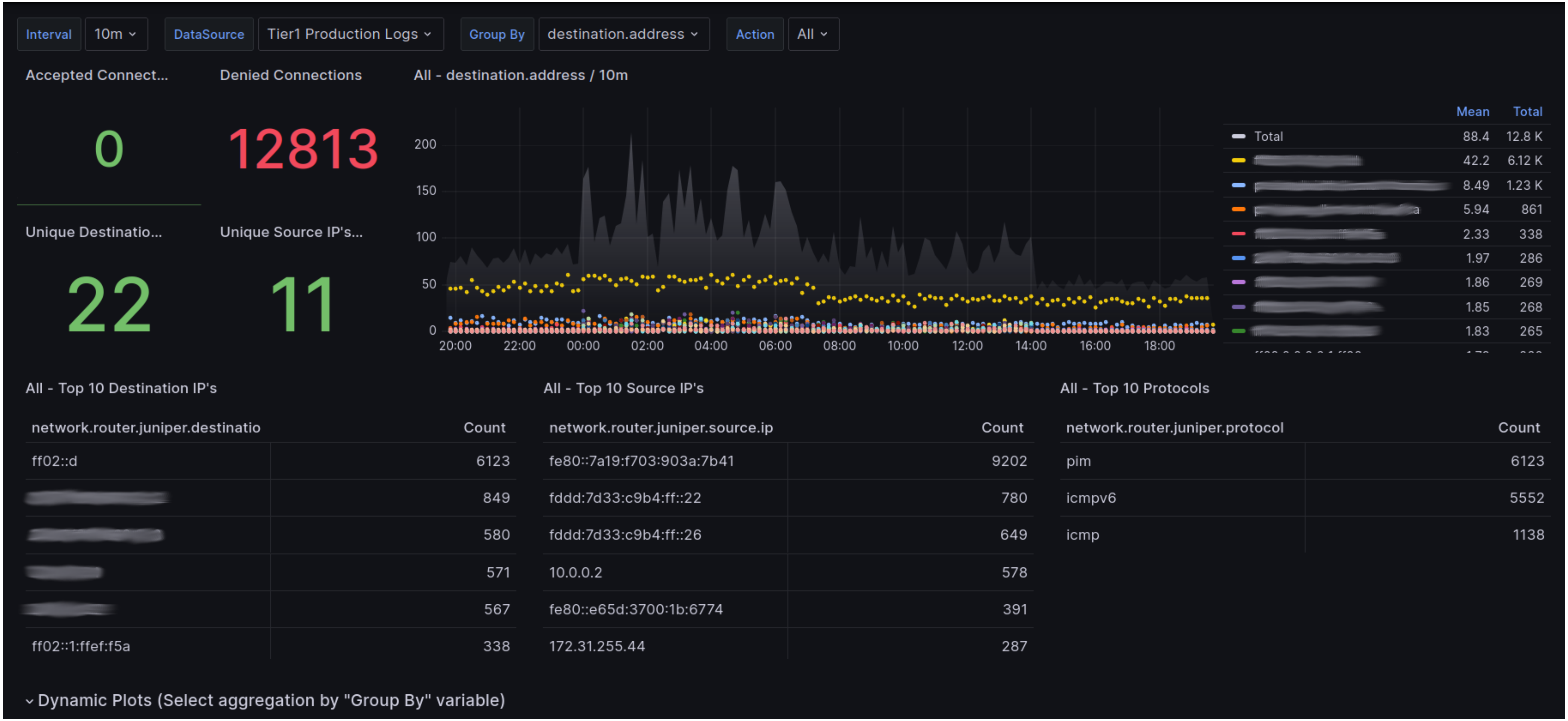
```
"requests": [
  {
    "name": "src",
    "node": "src",
    "args": {…},
    "body": {
      "query": { … },
    }
  },
  {
    "name": "dst",
      …
  }
]
}
```

# 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

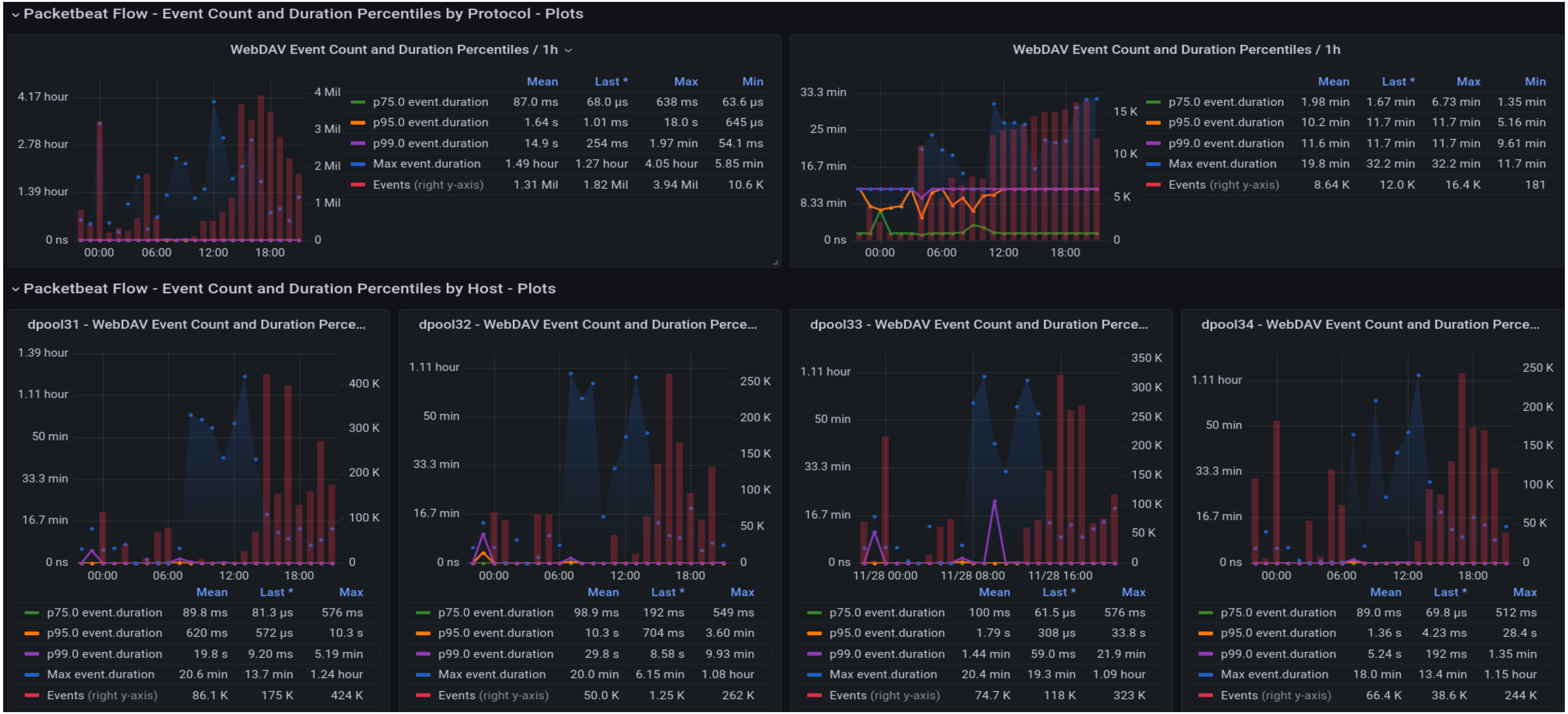
Fernando Fernandez Galindo
TRIUMF Scientific Computing Department

# CORE ROUTER SFLOW




Fernando Fernandez Galindo
TRIUMF Scientific Computing Department

# DCACHE PROTOCOLS (PACKETBEAT)





Fernando Fernandez Galindo
TRIUMF Scientific Computing Department
HEPiX Autumn 2023 Workshop
October 16-20, 2023

# SYSTEM SYSLOG HDD ERRORS

Interval | auto | DataSource | Tier1 Production Logs | HostName | All | Filters | +

**WNs With SSD Error Messages**

| host.name | process.module | Count |
|---|---|---|
| wn334 | sd | 32 |
| wn334 | blk_update_request | 8 |

**Events per Host / 1d**


**WNs Block Kernel Error Messages**

| host.name | process.module | device.name | message.keyword | Count |
|---|---|---|---|---|
| wn334 | blk_update_request | sda | critical medium error | 8 |

**WNs SD Kernel Error Messages**

| host.name | device.name | message.keyword | Count |
|---|---|---|---|
| wn334 | sda | Sense Key : Medium Error [current] | 8 |
| wn334 | sda | Add. Sense: Unrecovered read error | 8 |
| wn334 | sda | FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=0s | 5 |
| wn334 | sda | FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=1s | 2 |
| wn334 | sda | CDB: Read(10) 28 00 0f 1d d3 d8 00 00 08 00 | 2 |
| wn334 | sda | FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=2s | 1 |
| wn334 | sda | CDB: Read(10) 28 00 0f 1d d5 40 00 00 80 00 | 1 |