Hewlett Packard
Enterprise

Monitoring HPE Cray HPC Systems

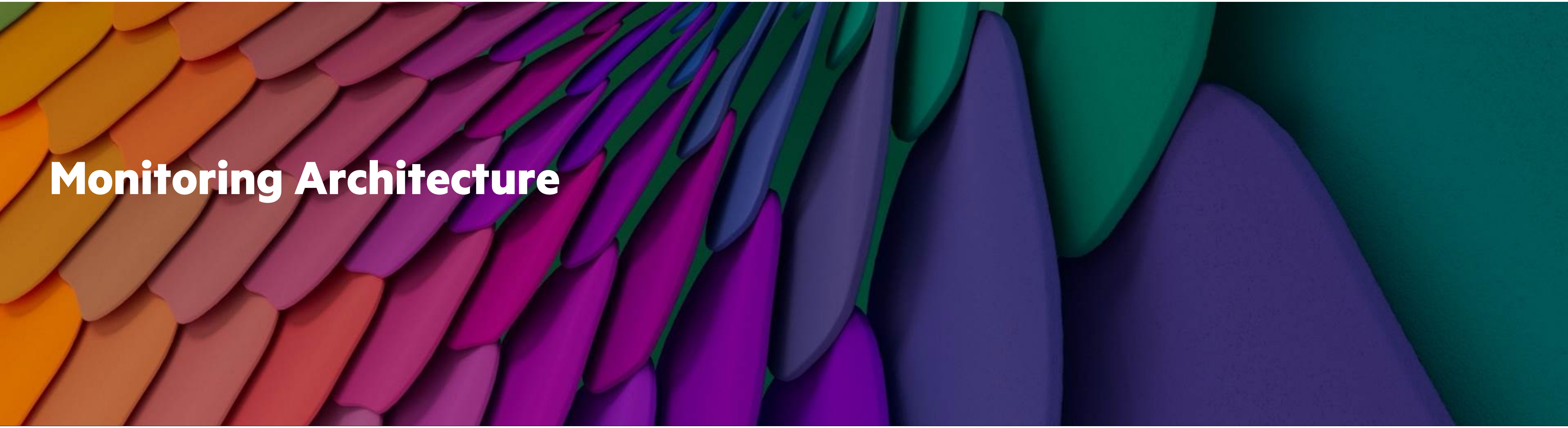
Harold Longley, Sue Miller,
Raghul Vasudevan, Pete Guyan, HPE

May 5, 2025

Agenda

- Monitoring Architecture
- HPCM (HPE Performance Cluster Manager)
  - Monitoring Configuration
  - Kafka and the Consumers
  - Producers
  - Alerting, SIM and rackmap
  - AIOps
- CSM (Cray System Management)
  - Monitoring Configuration
  - AIOps Configuration
  - Alerting Configuration
  - System Management Health Monitoring
  - SMA Monitoring
  - Logs
- System Exploration
- Resources

Monitoring Architecture



HPC Hardware and Software Monitoring

HPE offers fine-grained centralized monitoring and management of your system to keep it performing at its best

CPUs

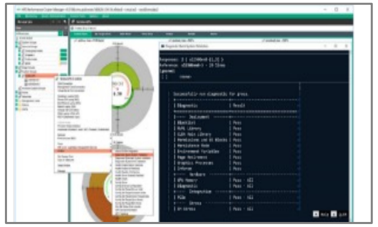
GPUs

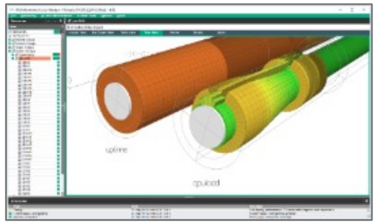
Memory

Networking

Power & cooling

Software





- CLI support to manage the monitoring framework – Setup, teardown
- View metrics and events via GUI, CLI, Grafana and OpenSearch Dashboards
- Support for wide range of components with HPC systems
- Customize system telemetry and alerts to best suit your needs
- AIOps for anomaly detection of system and selected datacenter metrics

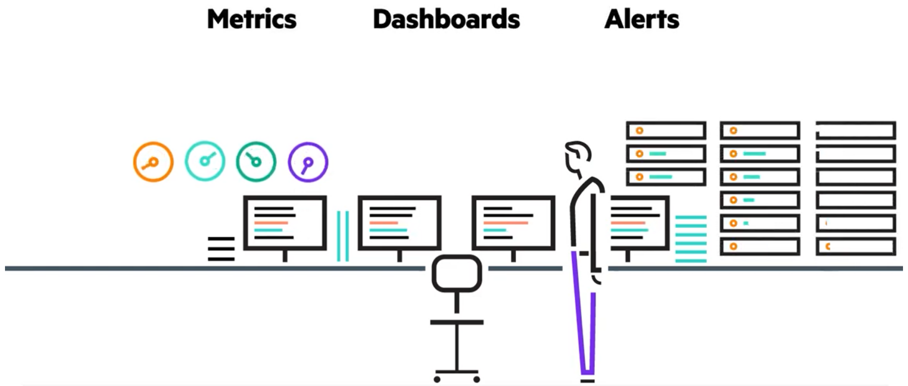
Unified Alerting Framework

Help you generate proactive alerts for monitoring the system

- Centralized alerting for various components and subsystems
- Faster incident reporting
- Customaizable alert rules
- Interactive visualization of alerts – GUI, CLI
- Action trigger policy for automative actions – plugin called first-responder
- Proactive notifications – Integration of email, slack, etc

HPE HPC Systems – monitoring

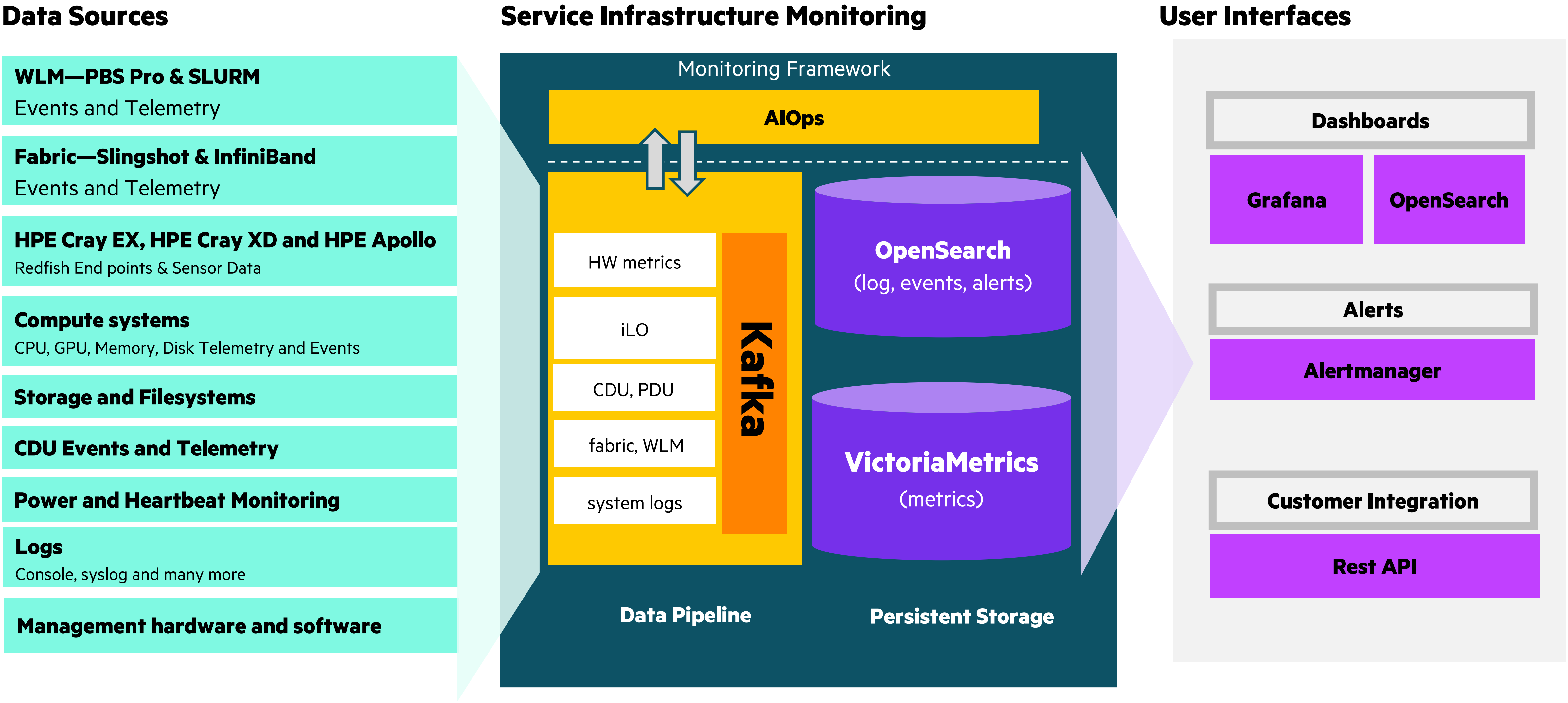
Customize metrics, dashboards, and alerts based on system monitoring needs



- workload managers
- high speed fabric
- compute nodes with CPUs and GPUs
- power and cooling systems
- services infrastructure components

- Performs active monitoring of infrastructure health
- Highly scalable and extensible solution able to support largest Exascale systems
- Provides visual dashboards and command line interfaces

Monitoring Framework



Data Sources

WLM—PBS Pro & SLURM
Events and Telemetry

Fabric—Slingshot & InfiniBand
Events and Telemetry

HPE Cray EX, HPE Cray XD and HPE Apollo
Redfish End points & Sensor Data

Compute systems
CPU, GPU, Memory, Disk Telemetry and Events

Storage and Filesystems

CDU Events and Telemetry

Power and Heartbeat Monitoring

Logs
Console, syslog and many more

Management hardware and software

Service Infrastructure Monitoring

Monitoring Framework

AIOps

HW metrics

iLO

CDU, PDU

fabric, WLM

system logs

Kafka

OpenSearch
(log, events, alerts)

VictoriaMetrics
(metrics)

Data Pipeline

Persistent Storage

User Interfaces

Dashboards

Grafana

OpenSearch

Alerts

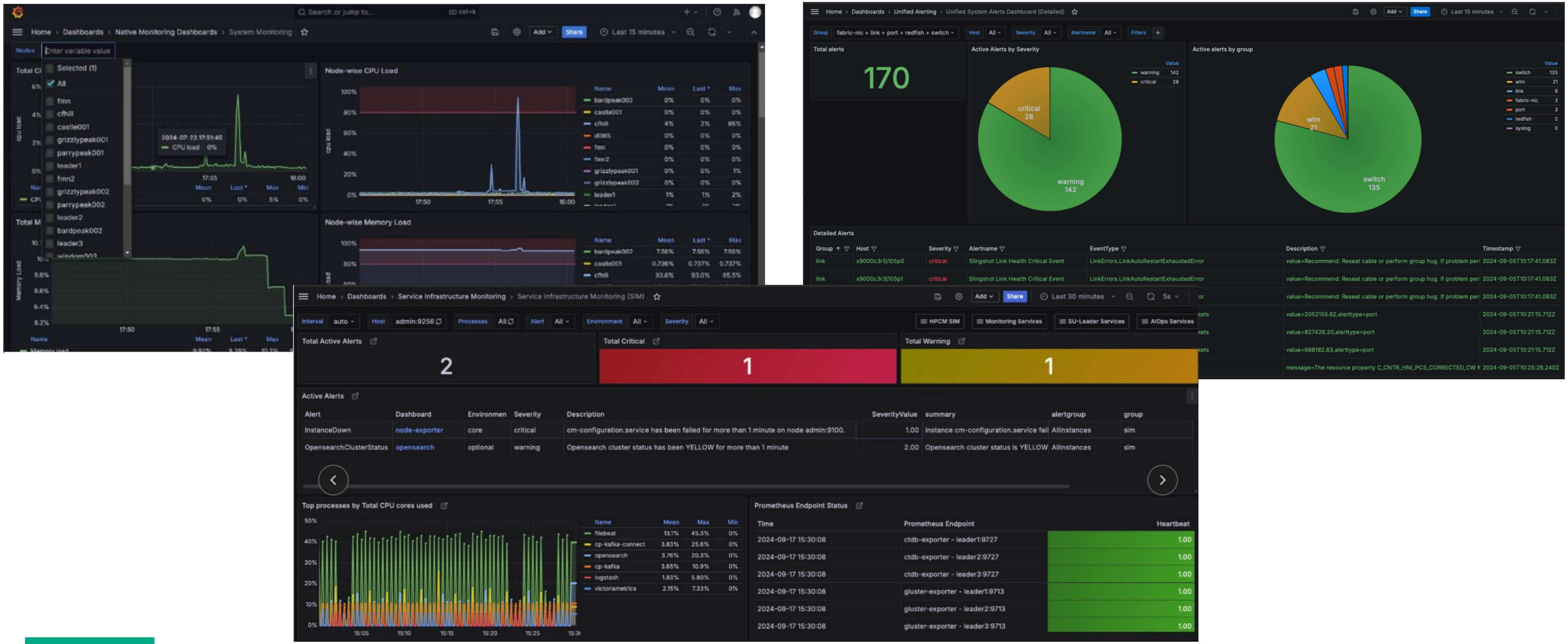
Alertmanager

Customer Integration

Rest API

Monitoring Dashboards

- Visualize health status of nodes, system, fabric, workload managers, GPU, CDU, services infrastructure



Detailed Alerts

Group	Host	Severity	Alertname	EventType	Description	Timestamp
link	x9000c3r5j105p0	critical	Slingshot Link Health Critical Event	LinkErrors.LinkAutoRestartExhaustedError	value=Recommend: Reseat cable or perform group hug. If problem per:	2024-09-05T10:17:41.083Z
link	x9000c3r3j105p1	critical	Slingshot Link Health Critical Event	LinkErrors.LinkAutoRestartExhaustedError	value=Recommend: Reseat cable or perform group hug. If problem per:	2024-09-05T10:17:41.083Z
				or	value=Recommend: Reseat cable or perform group hug. If problem per:	2024-09-05T10:17:41.083Z
				kets	value=2052155.62,alerttype=port	2024-09-05T10:21:15.712Z
				kets	value=827426.20,alerttype=port	2024-09-05T10:21:15.712Z
				kets	value=688182.63,alerttype=port	2024-09-05T10:21:15.712Z
					message=The resource property C_CNTR_HNI_PCS_CORRECTED_CW N	2024-09-05T10:25:26.240Z

Active Alerts

Alert	Dashboard	Environmen	Severity	Description	SeverityValue	summary	alertgroup	group
InstanceDown	node-exporter	core	critical	cm-configuration.service has been failed for more than 1 minute on node admin:9100.	1.00	Instance cm-configuration.service fail	AllInstances	sim
OpensearchClusterStatus	opensearch	optional	warning	Opensearch cluster status has been YELLOW for more than 1 minute	2.00	Opensearch cluster status is YELLOW	AllInstances	sim

Name	Mean	Max	Min
filebeat	13.1%	45.3%	0%
cp-kafka-connect	3.83%	25.6%	0%
opensearch	3.76%	20.3%	0%
cp-kafka	3.65%	10.9%	0%
logstash	1.83%	5.80%	0%
victoriametrics	2.15%	7.33%	0%

Prometheus Endpoint Status

Time	Prometheus Endpoint	Heartbeat
2024-09-17 15:30:08	ctdb-exporter - leader1:9727	1.00
2024-09-17 15:30:08	ctdb-exporter - leader2:9727	1.00
2024-09-17 15:30:08	ctdb-exporter - leader3:9727	1.00
2024-09-17 15:30:08	gluster-exporter - leader1:9713	1.00
2024-09-17 15:30:08	gluster-exporter - leader2:9713	1.00
2024-09-17 15:30:08	gluster-exporter - leader3:9713	1.00

AIOps implementation in HPCM

More insight = better chance to uncover issues before they turn into failures

- **AIOps Dashboards** to displays anomalies
  - Facility metrics
    - CDU (cooling distribution unit),
    - CRC (cooling rack controller)
  - Visualization panel for CPU and GPU temperature metrics
  - Slingshot telemetry
    - Anomaly metrics using Grafana visualization panel

- **Technical Highlights**
  - Container-based deployment
  - Real-time and offline anomaly detection, prediction for time-series monitoring data
  - Uni-variate and Multi-variate metrics support
  - Grafana visualization support

Traditional approach to monitoring

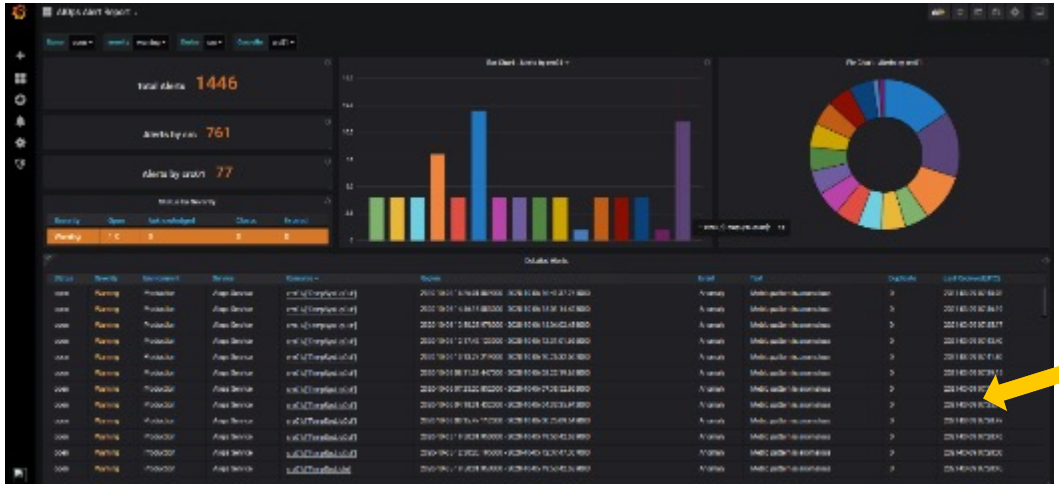
- Relying on thresholds to see issues
- The bigger the system, the more data administrators need to sift through and analyze
- Too many false alarms mean real issues are often overlooked

AIOps for System Infrastructure

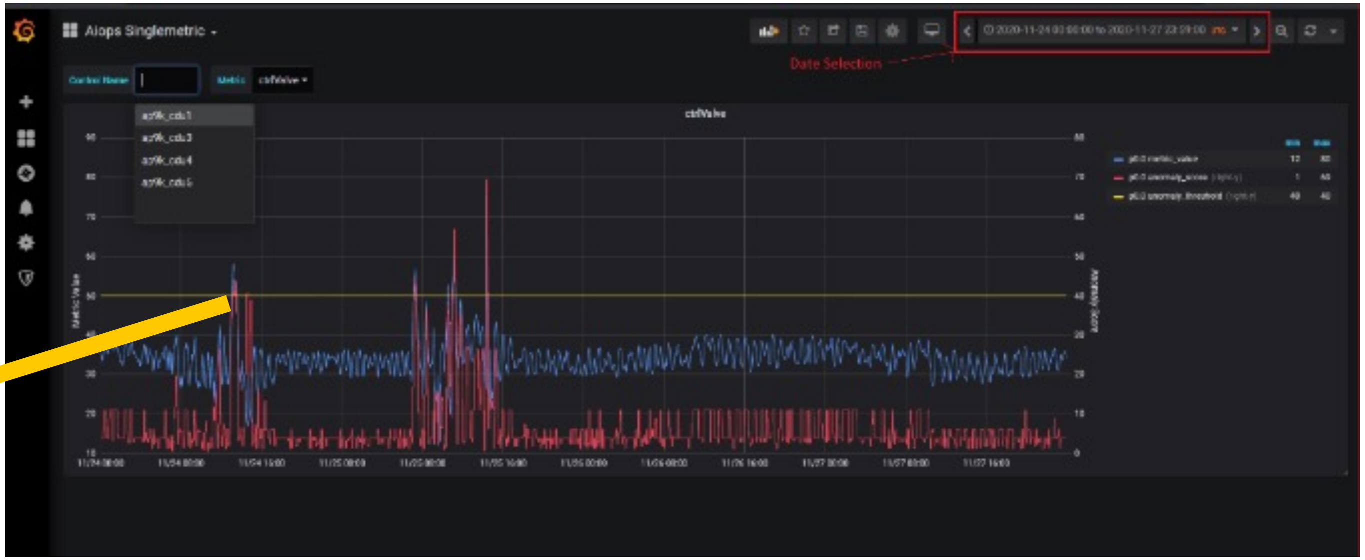
- Takes a predictive approach based on historical patterns of behavior
- AIOps uses machine learning and deep learning technologies to identify and report trends

AIOps dashboards in Grafana

More insight = better chance to uncover issues before they turn into failures



The **AIOps alert report** dashboard displays notifications of anomalies for cooling hardware. The pie chart shows where in the system alerts come from.



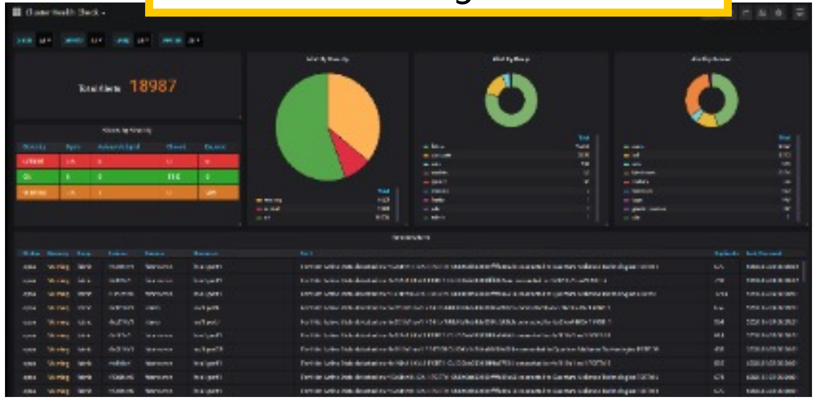
AIOps single metric dashboard contains plots of metric data values **(blue line),** anomaly scores for the monitored metric, which in this case, is the CDU valve position **(red line)**. An alert is generated (and displayed on the dashboard) when the anomaly score exceeds the anomaly threshold **(yellow line).**
The alert expires when additional alerts are not generated during a predefined period of time.

AIOps multimetric dashboard displays the anomaly in a raw metric plot (one that is selected by the user), along with an anomaly score and its threshold for a correlated group of metrics.

Multi-pane Dashboards

Highly scalable and customizable live system monitoring dashboards

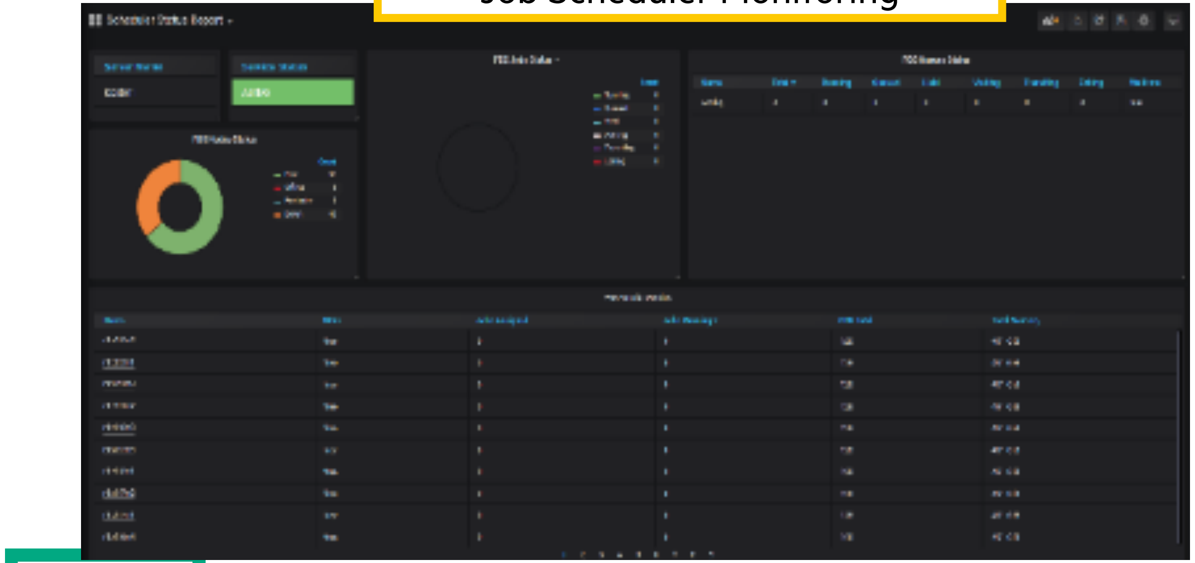
Unified Alerting Dashboard



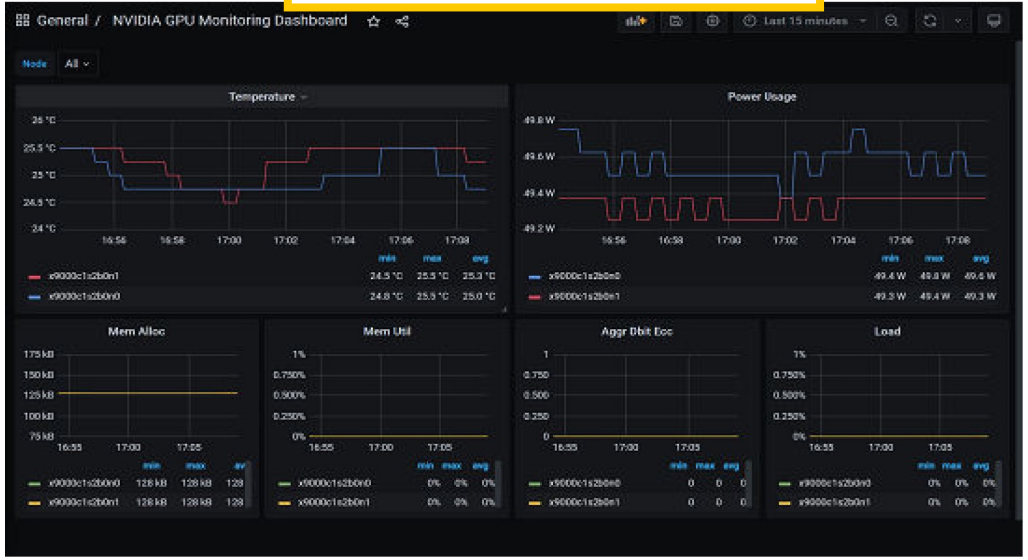
CDU Dashboard

System Monitoring Dashboard

Job Scheduler Monitoring

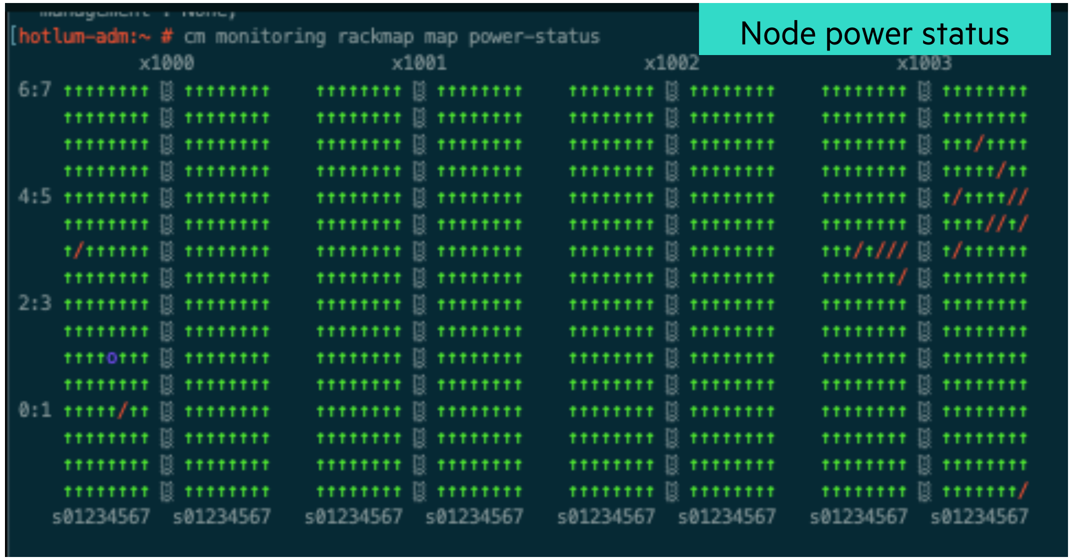


GPU Dashboard



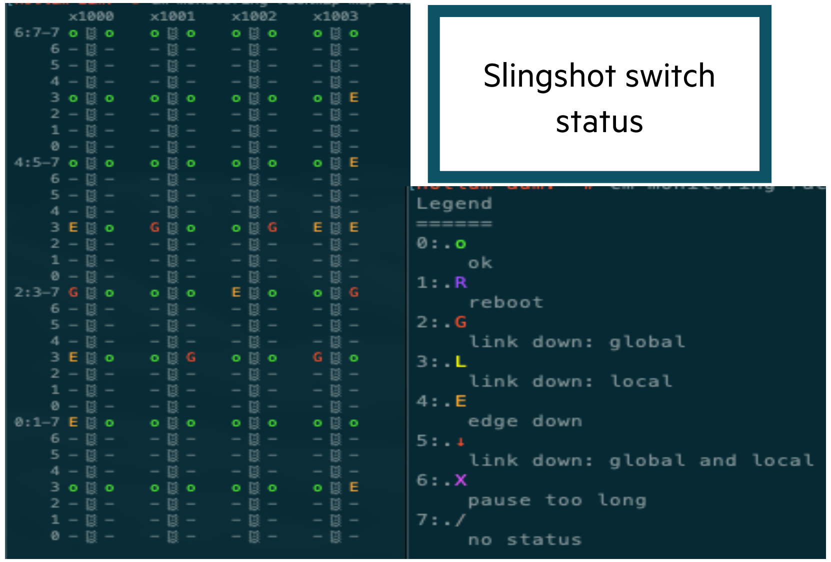
Visualization with CLI – Rackmap – HPCM

- Interactive rack layout – Displays entire rack in a single CLI view with each node clearly defined
- Real time health indicators- Color coded node icons reflect status and update real time
- Network link overlay – Visualize upstream/downstream port health and link errors directly on the rack map



Node power status

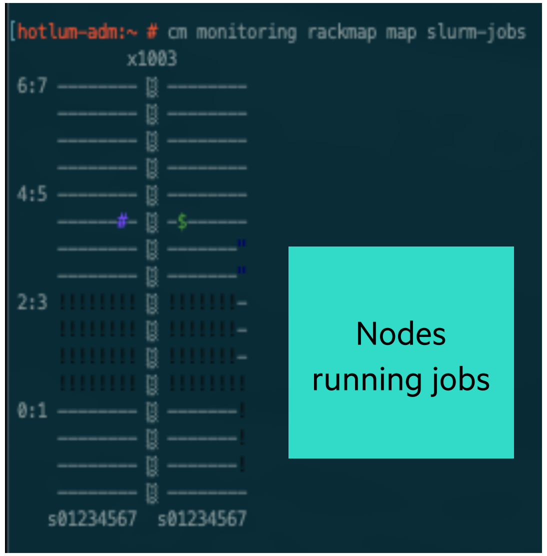
```
[hotlum-adm:~ # cm monitoring rackmap map power-status
              x1000                      x1001                      x1002                      x1003
6:7 ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑/↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑/↑↑
4:5 ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑/↑↑↑↑//
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑//↑/
    ↑/↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑/↑/// ▯ ↑/↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑/ ▯ ↑↑↑↑↑↑↑↑
2:3 ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑O↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
0:1 ↑↑↑↑↑/↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑
    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑↑    ↑↑↑↑↑↑↑↑ ▯ ↑↑↑↑↑↑↑/
   s01234567   s01234567  s01234567   s01234567  s01234567   s01234567  s01234567   s01234567
```



Slingshot switch status

```
        x1000    x1001    x1002    x1003
6:7-7  o ▯ o    o ▯ o    o ▯ o    o ▯ o
    6  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    5  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    4  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    3  o ▯ o    o ▯ o    o ▯ o    o ▯ E
    2  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    1  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    0  - ▯ -    - ▯ -    - ▯ -    - ▯ -
4:5-7  o ▯ o    o ▯ o    o ▯ o    o ▯ E
    6  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    5  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    4  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    3  E ▯ o    G ▯ o    o ▯ G    E ▯ E
    2  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    1  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    0  - ▯ -    - ▯ -    - ▯ -    - ▯ -
2:3-7  G ▯ o    o ▯ o    E ▯ o    o ▯ G
    6  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    5  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    4  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    3  E ▯ o    o ▯ G    o ▯ o    G ▯ o
    2  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    1  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    0  - ▯ -    - ▯ -    - ▯ -    - ▯ -
0:1-7  E ▯ o    o ▯ o    o ▯ o    o ▯ o
    6  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    5  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    4  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    3  o ▯ o    o ▯ o    o ▯ o    o ▯ E
    2  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    1  - ▯ -    - ▯ -    - ▯ -    - ▯ -
    0  - ▯ -    - ▯ -    - ▯ -    - ▯ -
```

```
Legend
======
0:.o
    ok
1:.R
    reboot
2:.G
    link down: global
3:.L
    link down: local
4:.E
    edge down
5:.↓
    link down: global and local
6:.X
    pause too long
7:./
    no status
```



```
[hotlum-adm:~ # cm monitoring rackmap map slurm-jobs
            x1003
6:7 -------- ▯ --------
    -------- ▯ --------
    -------- ▯ --------
    -------- ▯ --------
4:5 -------- ▯ --------
    -------#- ▯ -$------
    -------- ▯ -------"
    -------- ▯ -------"
2:3 !!!!!!!! ▯ !!!!!!!-
    !!!!!!!! ▯ !!!!!!!-
    !!!!!!!! ▯ !!!!!!!-
    !!!!!!!! ▯ !!!!!!!!
0:1 -------- ▯ -------!
    -------- ▯ -------!
    -------- ▯ -------!
    -------- ▯ --------
   s01234567   s01234567
```

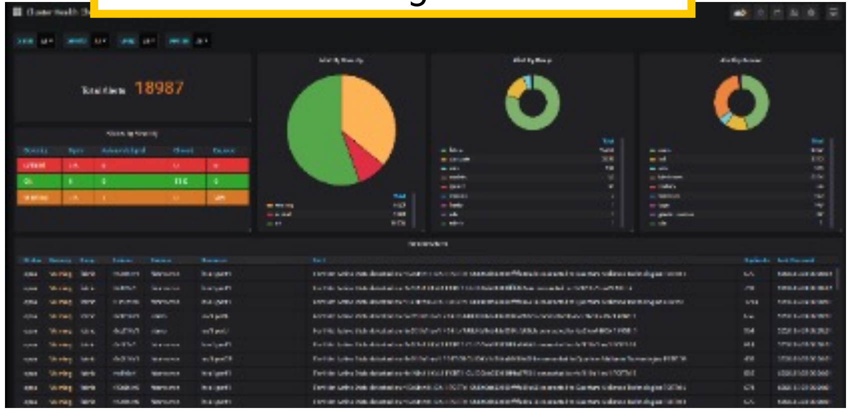
Nodes running jobs

Real-Time Cluster Health Monitoring

Visualization dashboards

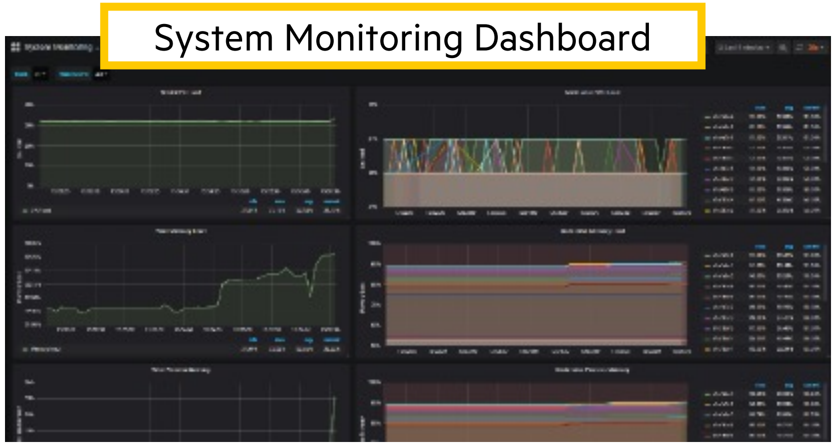
- Cluster Health at-a-Glance—Single Pane view for the complete cluster Health Status
- Live System Monitoring—Dashboards for key metrics like Power, Cooling, CPU, Memory, Disk, Fabric, Gluster, Job Scheduler monitoring metrics
- Scalable—Highly scalable data pipeline at the backend
- Customizable—Create new dashboards easily

Unified Alerting Dashboard

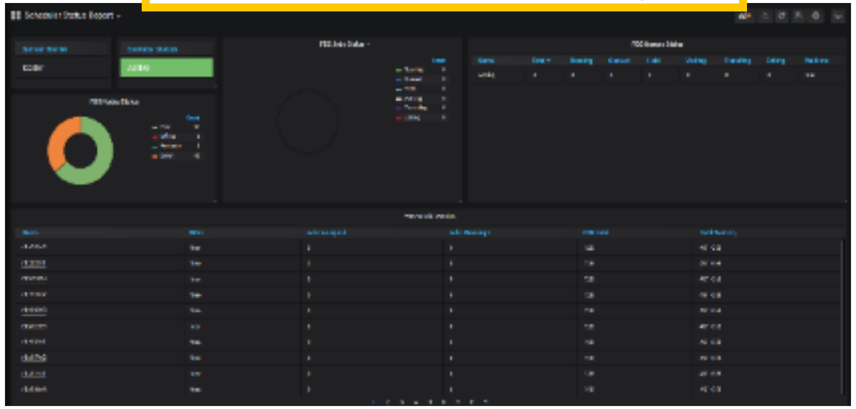


CDU Dashboard

System Monitoring Dashboard

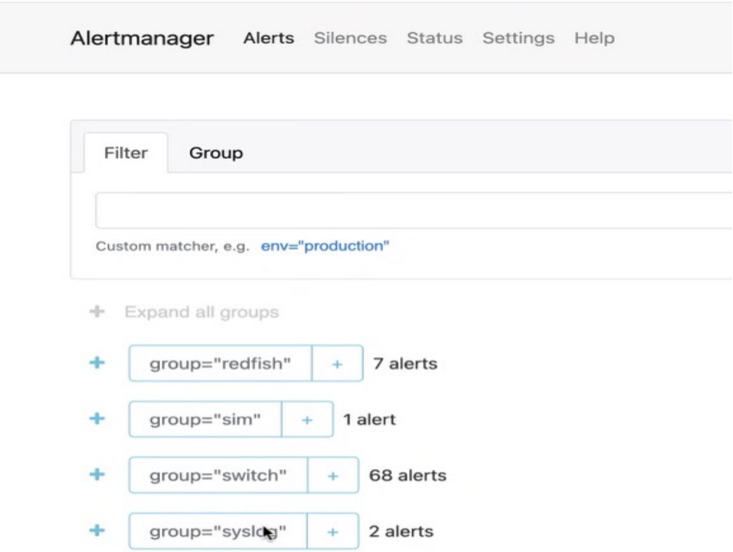


Job Scheduler Monitoring

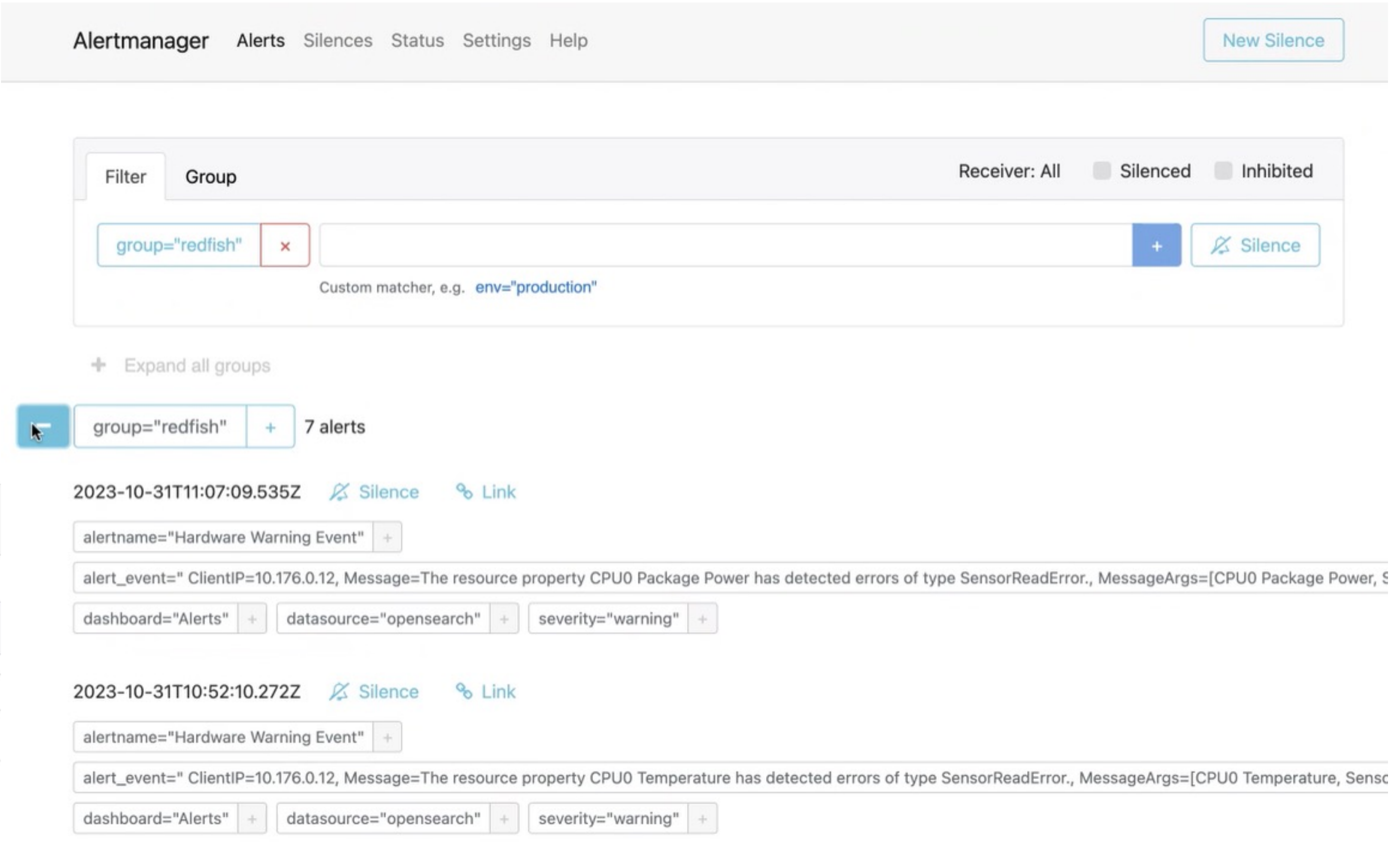


View and React to Alerts

- Collection of system-wide alerts enabled through Alertmanager
- Alertmanager sends the alerts (email, Slack) to OpenSearch repository and provides instant visualization
- Real-time alerts management is part of the HPE Performance Cluster Manager health check capability



```
group="redfish"   7 alerts
group="sim"       1 alert
group="switch"    68 alerts
group="syslog"    2 alerts
```



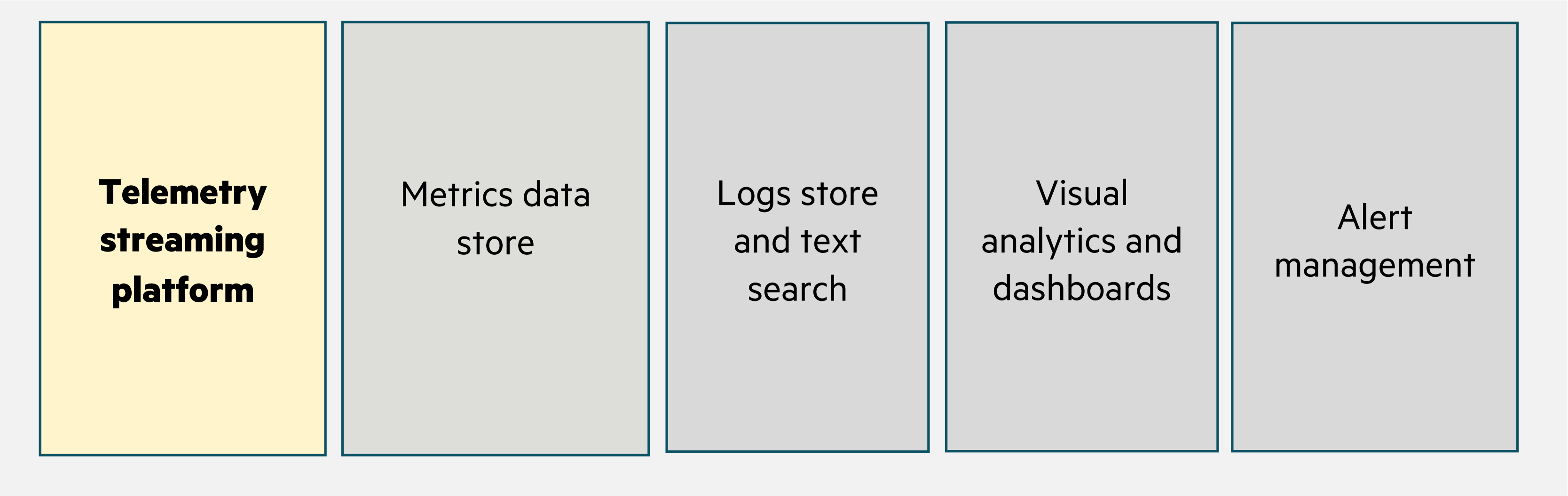
```
group="redfish" ×

group="redfish"   7 alerts

2023-10-31T11:07:09.535Z   Silence   Link
alertname="Hardware Warning Event"
alert_event=" ClientIP=10.176.0.12, Message=The resource property CPU0 Package Power has detected errors of type SensorReadError., MessageArgs=[CPU0 Package Power, S
dashboard="Alerts"   datasource="opensearch"   severity="warning"

2023-10-31T10:52:10.272Z   Silence   Link
alertname="Hardware Warning Event"
alert_event=" ClientIP=10.176.0.12, Message=The resource property CPU0 Temperature has detected errors of type SensorReadError., MessageArgs=[CPU0 Temperature, Sensc
dashboard="Alerts"   datasource="opensearch"   severity="warning"
```

Alert signifying sensor reading error in Cooling Distribution Unit (CDU)

Monitoring key technologies – Telemetry Streaming

Kafka – the industry standard **distributed data and event streaming platform**

Features – Performant, horizontally scalable, reliable data transport mechanism that enables connections from a wide variety of telemetry producers to a wide variety of consumers. Ability to use telemetry in near real time while in Kafka.



Telemetry streaming platform

Metrics data store

Logs store and text search

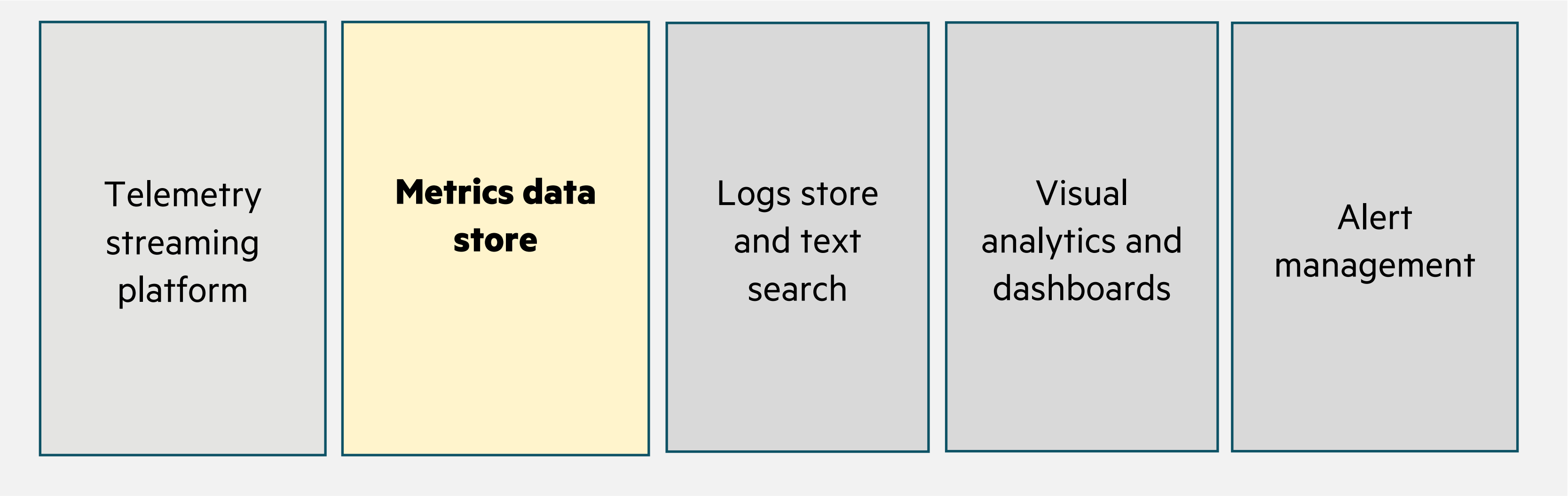
Visual analytics and dashboards

Alert management

Monitoring key technologies – Metrics Data Store

VictoriaMetrics – Timeseries database and service monitoring solution

Features – Horizontally scalable, space and query efficient database with built in features for monitoring services using Prometheus exporters. HPE provides a custom tool for Kafka to VictoriaMetrics (flow).



Telemetry streaming platform

Metrics data store

Logs store and text search

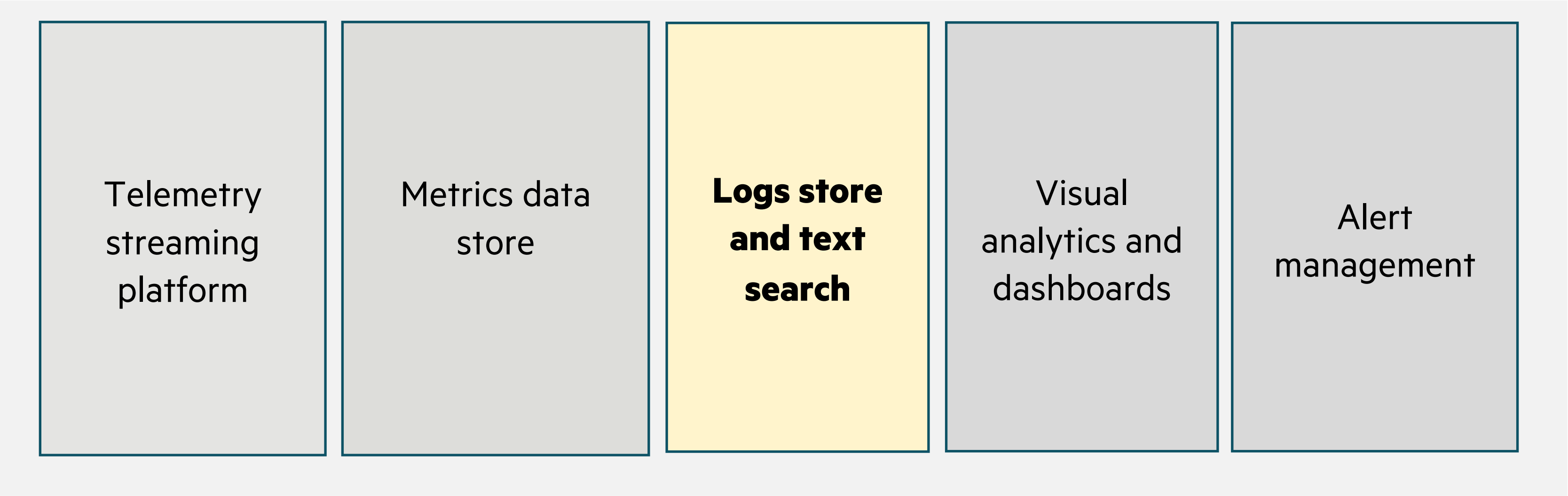
Visual analytics and dashboards

Alert management

Monitoring key technologies – Logs Store

OpenSearch – database for text data

Features – Horizontally scalable full text search with built in analytics and visualizations. Includes tools for managing log ingestion and regularizing data.



Telemetry streaming platform

Metrics data store

Logs store and text search

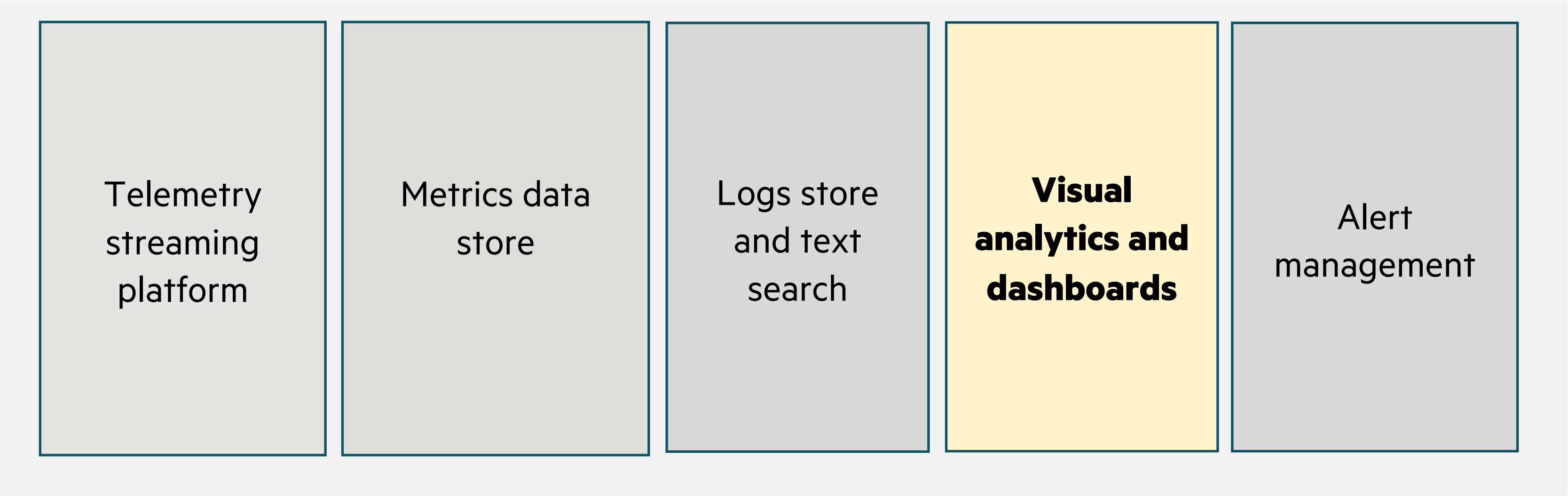
Visual analytics and dashboards

Alert management

Monitoring key technologies – Dashboards

Grafana – Industry leading **visual analytics platform**

Features – Used by most HPC customers. Integrates well with both databases, used as alerting engine for metrics, extensible via plugins which monitoring teams have already authored



Telemetry streaming platform

Metrics data store

Logs store and text search

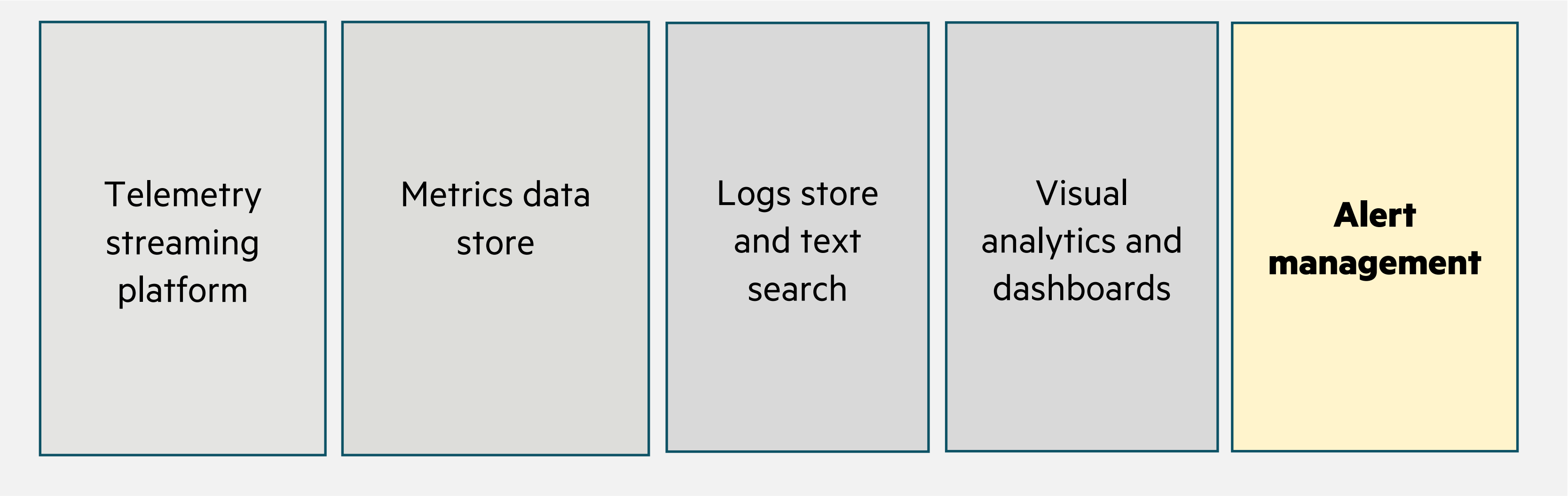
Visual analytics and dashboards

Alert management

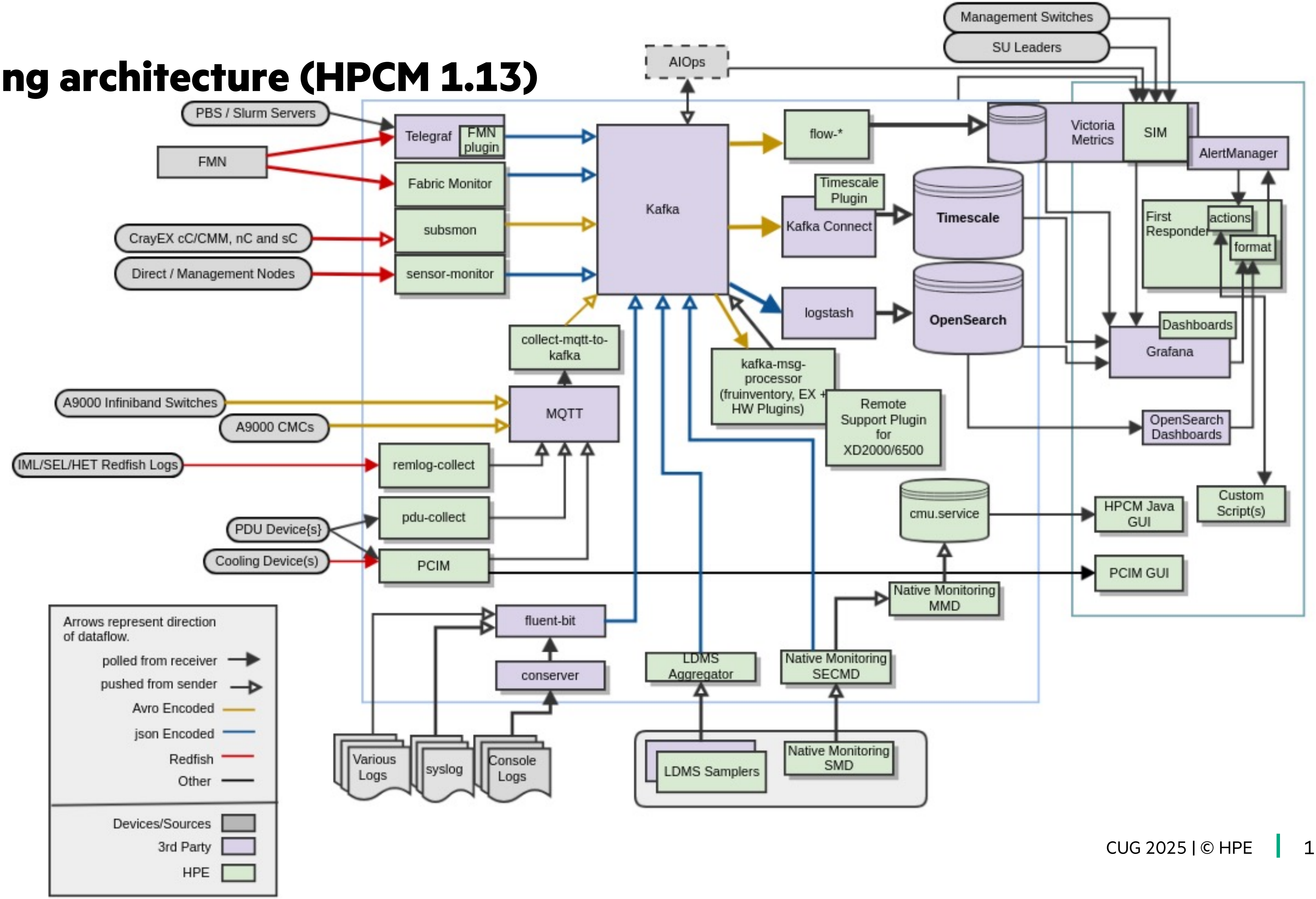
Monitoring key technologies – Alert Management

Alertmanager – Manages alerts sent by client applications for grouping, deduplication, silencing, etc.

Features – Integrates well with other core technologies. Provides many built-in notification methods. Includes webhook feature for more complex processing. Part of the larger integrated solution for alert lifecycle management


Telemetry streaming platform
Metrics data store
Logs store and text search
Visual analytics and dashboards
Alert management

Monitoring architecture (HPCM 1.13)



HPCM architecture differences

HPCM 1.10	**HPCM 1.11**	**HPCM 1.12**	**HPCM 1.13**
connect/Timescale	connect/Timescale	connect/Timescale & Flow/VictoriaMetrics (VM dashboards supported by patches)	flow/VictoriaMetrics
Prometheus	Prometheus	VictoriaMetrics	VictoriaMetrics
Filebeat	Filebeat	Filebeat	Fluentbit
WLM via telegraf (patch 11796 else remlog-collect)	WLM via telegraf	WLM via telegraf	WLM via telegraf
Unified alerting (alertman) introduced. Alerta/Elastalert still available	Unified alerting (alertman) Alerta/Elastalert still available	Unified alerting (alertman) Alerta/Elastalert removed	Unified alerting (alertman)

HPCM CLI differences

ⓘ **Do not use "cm monitoring setup" with <= 1.12**

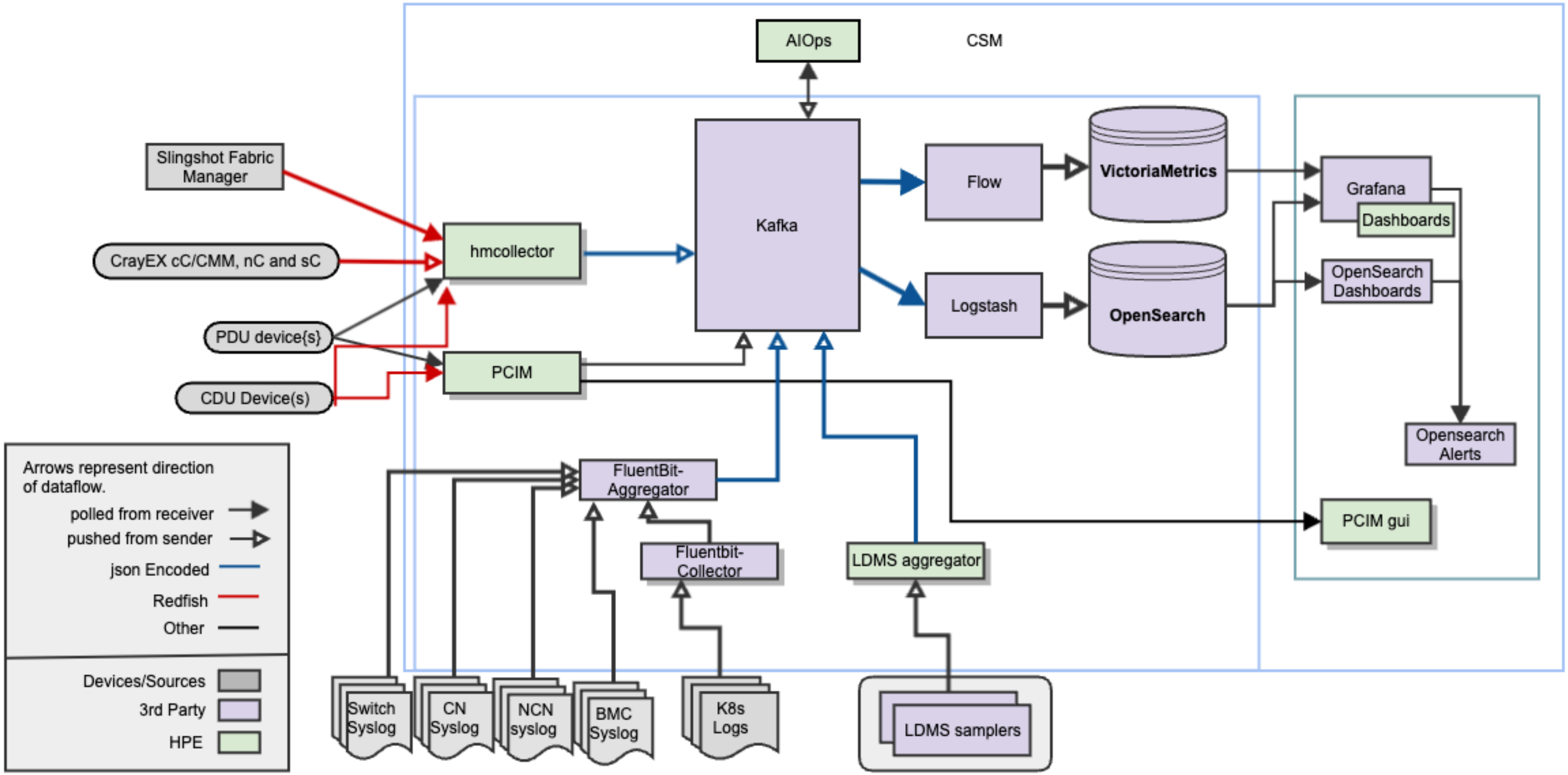
HPCM 1.10	**HPCM 1.11**	**HPCM 1.12**	**HPCM 1.13**
N/A	`cm support moncollect`	`cm support moncollect`	`cm support moncollect`
N/A	`cm monitoring setup` (not documented nor tested – **do not use**)	`cm monitoring setup` (not documented nor tested – **do not use**)	`cm monitoring setup|import|export|teardown|status`
N/A	N/A	`cm logs`	`cm logs`
N/A	N/A	`cm telemetry` (patch needed)	`cm telemetry`
`cm monitoring elk`	`cm monitoring elk`	`cm monitoring elk`	`cm monitoring logstash||opensearch`
`cm monitoring kafka enable || start`	`cm monitoring kafka enable || start`	`cm monitoring kafka enable || start`	`cm monitoring kafka start` for just kafka and no persistence methods

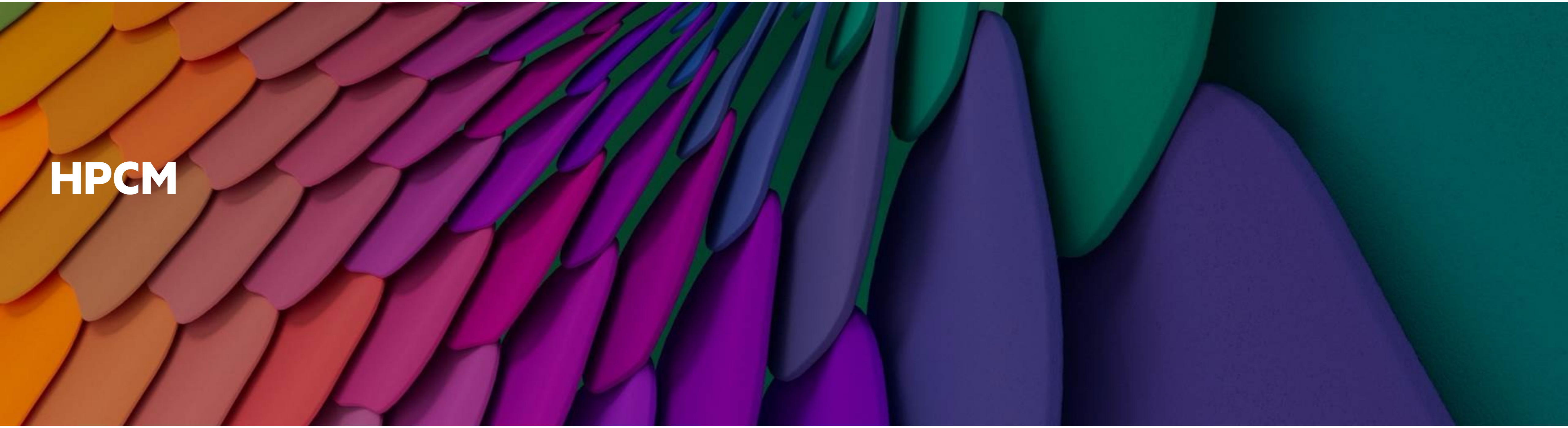
HPCM monitoring general release current patches

HPCM 1.10	HPCM 1.11	HPCM 1.12	HPCM 11.3
11796: monitoring and clusterhealth updates	11821: recommended pdu-collect update	11830: VictoriaMetrics and other monitoring updates	11850: HPCM 1.13: recommended monitoring updates
11809: optional grafana-dashboards update	11824: recommended monitoring, alerting and clusterhealth updates	11838: Grafana dashboard updates for VictoriaMetrics	
11820: recommended pdu-collect update			
11822: recommended pdu-collect update			

>=1.11 patches have an rpm named SG00XXXXX_info_hpcm if installed (update would not pull it in) on the admin plus rpms may be applicable to leaders etc. Refer to the patch release notes in /opt/clmgr/doc/

Monitoring architecture (CSM 1.6.1/SMA 10.15)





HPCM

HPCM

Monitoring Configuration

Kafka and the Consumers: VictoriaMetrics, TimescaleDB, OpenSearch

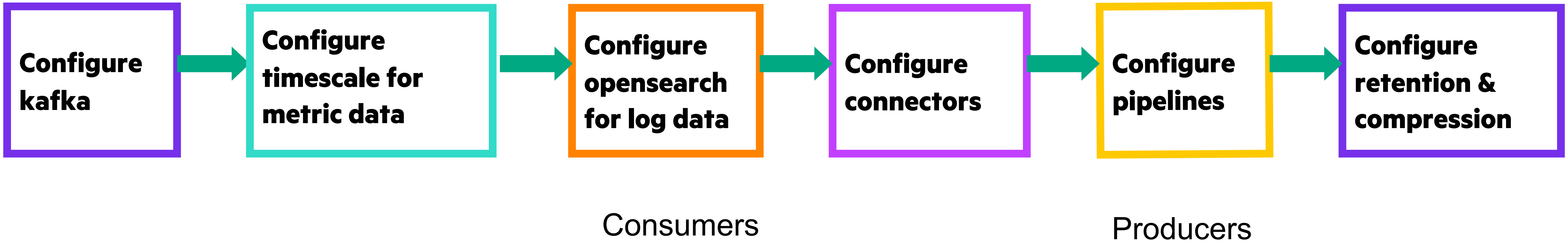
Producers: Native Monitoring, Power and Cooling, Slingshot, Workload Manager

Alerting, SIM and rackmap

AIOps

Monitoring Configuration

Monitoring architecture configuration flow prior to HPCM 1.13



Configure kafka

Configure timescale for metric data

Configure opensearch for log data

Configure connectors

Configure pipelines

Configure retention & compression

Consumers

Producers

Once complete, star the dashboards relevant to the system

Monitoring architecture configuration flow with HPCM 1.13

ⓘ Do not use "cm monitoring setup" with <= 1.12

```
cm monitoring setup [ -n nodes ] [ --pipelines crayex_hardware slingshot_hardware native ldms ] [ retention options ]
```

e.g. # cm monitoring setup -n admin,leader* --pipelines crayex_hardware slingshot_hardware native --kafka-retention-days 3 --opensearch-retention-days 30 --victoriametrics-retention-days 30

Can use –n to put the bulk of monitoring data on other nodes rather than admin and leaders.

ⓘ Conservative defaults: Kafka=1day, OpenSearch=7days, VictoriaMetrics=7days

Monitoring architecture configuration flow with 1.13

```
# cm monitoring setup -p native
Successfully enabled the dashboards:
node_diagnostics.json
Successfully enabled the dashboards:
amd_mi250x_gpu_monitoring.json
nvidia_gpu_monitoring.json
system_monitoring.json
finished!
```

- Various messages will be displayed such as:

```
setting up monitoring...
setting up Zookeeper...
setting up victoriametrics on: admin...
```

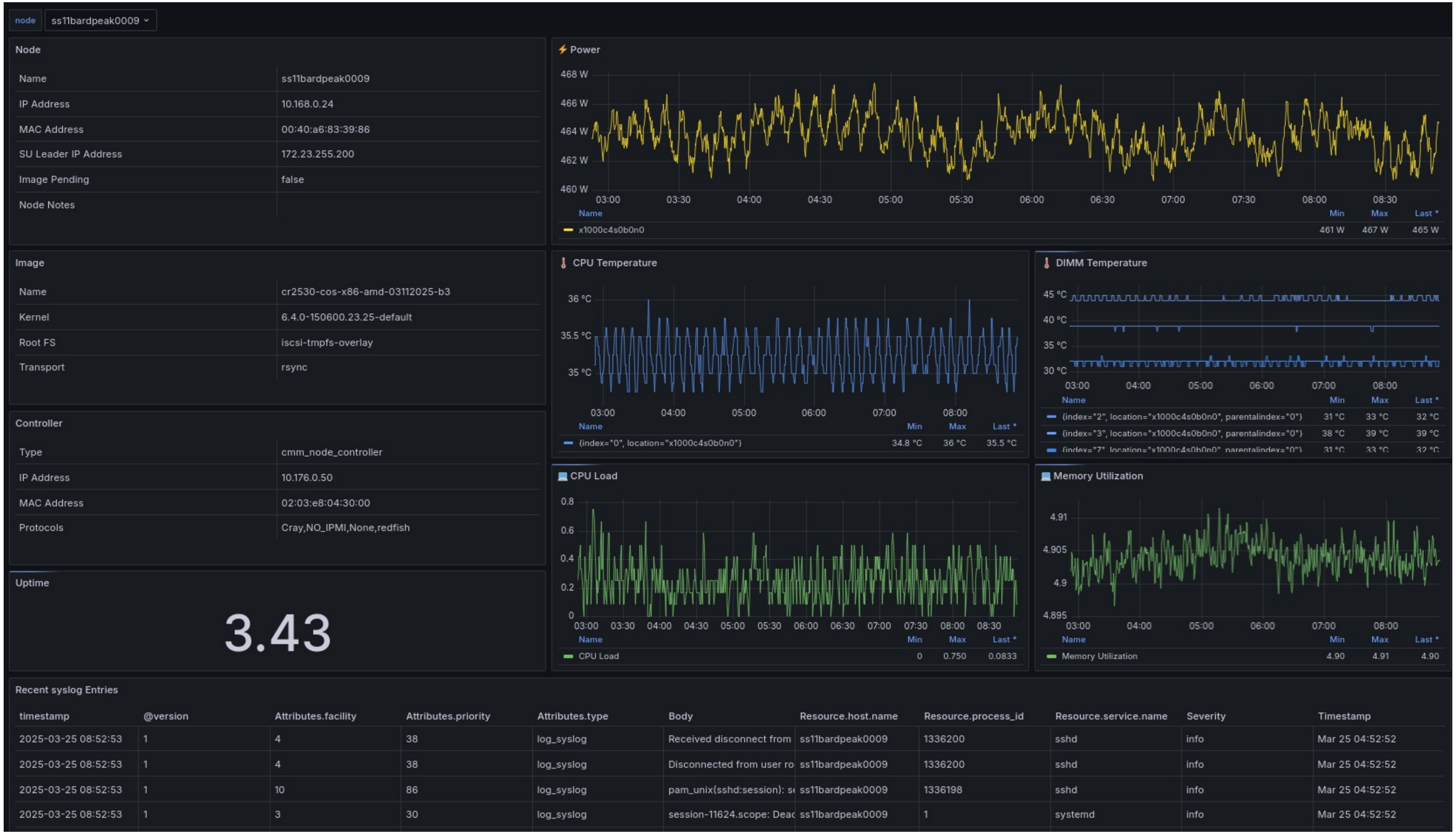
The output updates/refreshes the same line with different text

-v will change this behaviour

- Errors or warnings will persist in the output:

```
ensure all predefined topics are created and configured...[W] failed to create
topic pcm-monitoring: KafkaError{code=TOPIC_ALREADY_EXISTS,val=36,str="Topic
'pcm-monitoring' already exists."}
```

1.13 with 11850: New node diagnostics dashboard (native)



Node

Name	ss11bardpeak0009
IP Address	10.168.0.24
MAC Address	00:40:a6:83:39:86
SU Leader IP Address	172.23.255.200
Image Pending	false
Node Notes	

Image

Name	cr2530-cos-x86-amd-03112025-b3
Kernel	6.4.0-150600.23.25-default
Root FS	iscsi-tmpfs-overlay
Transport	rsync

Controller

Type	cmm_node_controller
IP Address	10.176.0.50
MAC Address	02:03:e8:04:30:00
Protocols	Cray,NO_IPMI,None,redfish

Uptime

3.43

⚡ Power

Name	Min	Max	Last *
x1000c4s0b0n0	461 W	467 W	465 W

🌡 CPU Temperature

Name	Min	Max	Last *
{index="0", location="x1000c4s0b0n0"}	34.8 °C	36 °C	35.5 °C

🌡 DIMM Temperature

Name	Min	Max	Last *
{index="2", location="x1000c4s0b0n0", parentalindex="0"}	31 °C	33 °C	32 °C
{index="3", location="x1000c4s0b0n0", parentalindex="0"}	38 °C	39 °C	39 °C
{index="7", location="x1000c4s0b0n0", parentalindex="0"}	31 °C	33 °C	32 °C

💻 CPU Load

Name	Min	Max	Last *
CPU Load	0	0.750	0.0833

💻 Memory Utilization

Name	Min	Max	Last *
Memory Utilization	4.90	4.91	4.90

Recent syslog Entries

timestamp	@version	Attributes.facility	Attributes.priority	Attributes.type	Body	Resource.host.name	Resource.process_id	Resource.service.name	Severity	Timestamp
2025-03-25 08:52:53	1	4	38	log_syslog	Received disconnect from	ss11bardpeak0009	1336200	sshd	info	Mar 25 04:52:52
2025-03-25 08:52:53	1	4	38	log_syslog	Disconnected from user ro	ss11bardpeak0009	1336200	sshd	info	Mar 25 04:52:52
2025-03-25 08:52:53	1	10	86	log_syslog	pam_unix(sshd:session): s	ss11bardpeak0009	1336198	sshd	info	Mar 25 04:52:52
2025-03-25 08:52:53	1	3	30	log_syslog	session-11624.scope: Deac	ss11bardpeak0009	1	systemd	info	Mar 25 04:52:52

Kafka and consumers: Old short story

```
cm monitoring kafka enable and start

cm monitoring timescaledb enable and start

cm monitoring elk enable and start
```

Kafka and timescaledb use zookeeper to maintain a concept of cluster when using SU leaders but this is usually transparent to the user

- If there are SU leaders:

```
kfka-dist-setup
cm monitoring timescaledb node add <options>
elk-dist-setup
```

- In 1.10 and higher are not enabled by default so enable the ones relevant to the system (VictoriaMetrics uses flow):

```
cm monitoring connect enable --name <name>
```

- Most pipelines (the producers) are not generally configured at this point

Producers: Old short story

Probably need to add a gpu type

```
cm monitoring native enable and start

cm monitoring native metrics add -g slingshot -N <Max # NICs> and restart

systemctl enable and start pcim

cm monitoring dashboard grafana set --cdu|--cdu_ex2500 enable
```

- Add cooling device other than Cray EX CDUs which are detected by default

SU-leaders: sensor-processor

```
systemctl enable and start sensor-monitor

cm monitoring slingshot enable and set <options> and start ##slingshot
changes/differences needed here

cm monitoring dashboard grafana set --slingshot enable
```

- Double check FMN configuration

If changes are made on the FMN, new metrics with inappropriate compression can be created and consume disk

```
cm node zypper|dnf -n <fmn> install slingshot-fabric-check
```

Producers: Old short story

- Set number of switches and switch groups in config file
- `systemctl enable` and `start` 3 services and timers on the fmn after installing rpm
- Curl commands to enable dashboards
- Install hpe-telegraf and telegraf on the slurm controller

`cm monitoring slurm enable <options>` **and** `start`

- For slurm power dependent on hardware, configure the plugin config in slurm and HPCM
  - Configure `/opt/clmgr/wlm-mon/conf/wlm-mon.yml`
- Configure tsdb retention and **compression** after each stage **particularly slingshot pipelines**:

```
for i in slingshot cooldev pcm cray pdu disk; do cm monitoring timescaledb
retention --category  $i --interval 7d ;cm monitoring timescaledb compression --
category  $i --interval 1d ; done
```

IMPORTANT: tsdb compressions save >90% disk space

Alerting and SIM: Old short story

```
cm monitoring alerting enable
```

```
cm monitoring alerting opensearch or grafana --enable-rule <appropriate rules>
```

```
cm monitoring alerting route email --from <email> --to <email> --smtp <smtp.server:25> --alert-group <group>
```

```
cm sim enable and start and add {--service-group monitoring-services|suleader-services}
```

```
cm monitoring rackmap map component-drift or power or cpu-temperature or slingshot-switch-status -l
```

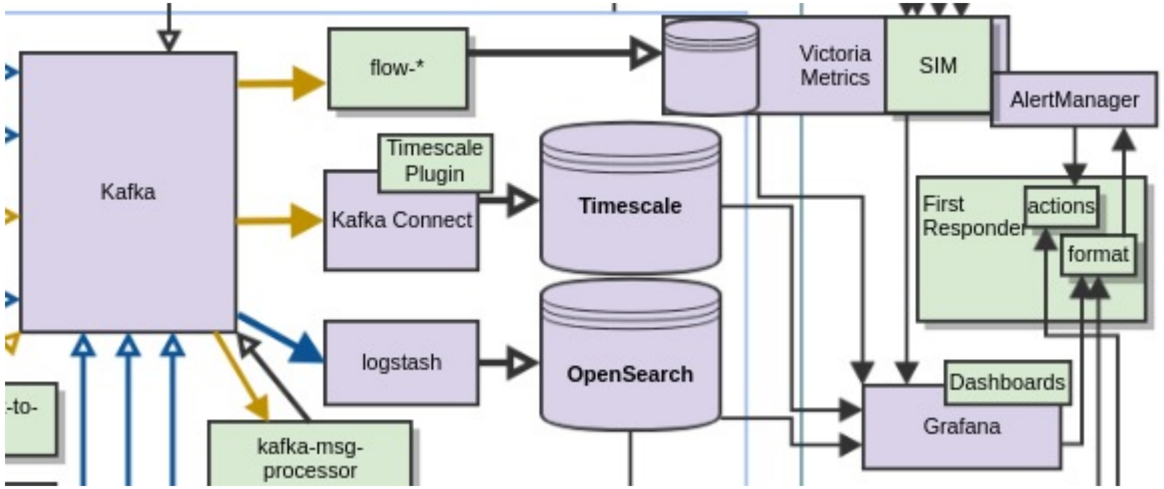
Back to the future: 1.13 high level simplification

- One initial command
- Easy teardown
- High level status
- Configuration import/export via a file

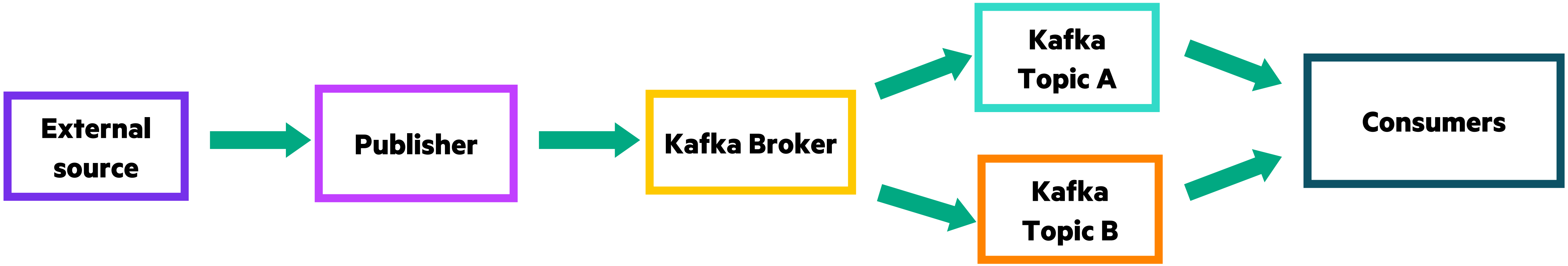
All under "cm monitoring setup|teardown|status|import|export"

```
# cm monitoring status
=== Zookeeper ===
Zookeeper has not been setup, no status to show
=== Kafka ===
Kafka has not been setup, no status to show
=== OpenSearch ===
OpenSearch has not been setup, no status to show
=== Logstash ===
Logstash has not been setup, no status to show
=== FluentBit ===
Fluentbit has not been setup, no status to show
=== VictoriaMetrics ===
VictoriaMetrics has not been setup, no status to show
=== Flow ===
Flow has not been setup, no status to show
=== MQTT ===
MQTT has not been setup, no status to show
=== Subsmon ===
subsmon has not been setup, no status to show
=== Grafana ===
Grafana has not setup, no status to show
=== SIM ===
SIM has not been setup, no status to show
=== Alerting ===
Alerting has not been setup, no status to show
```

Kafka and the Consumers



Kafka Terminology and Simplification


External source

Publisher

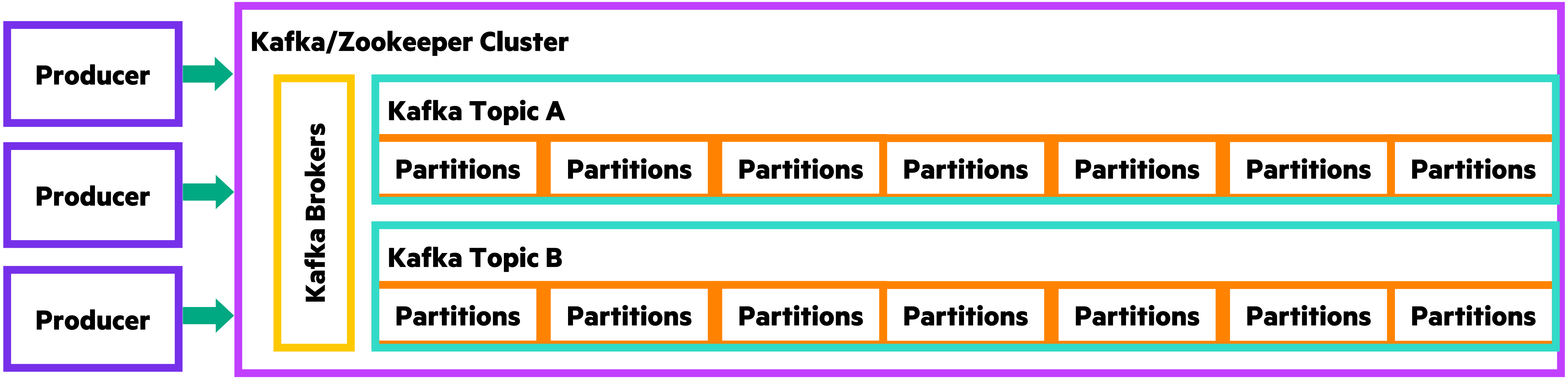
Kafka Broker

Kafka Topic A

Kafka Topic B

Consumers

Kafka Terminology and Simplification



ⓘ **Multiple partitions for systems with SU leaders; If a broker fails the consumer can use partition replicas on the other brokers**

A kafka topic leader is not synonymous with SU leader

Kafka

- The broker service is confluent-kafka.service running on admin & leaders
- Kafka is about large volume event streaming in a scaleable manner in a fault tolerant cluster providing storage for a shorter retention period (1.13 default is 24hrs for all topics)
- confluent-schema-registry.service: manages access to avro schema. Runs on the admin.
- confluent-kafka-rest.service: provides REST API for kafka. Provides an API to query, delete topics etc.
- ksldb was removed in 1.9

Kafka

- confluent-zookeeper.service on admin and a subset of leaders is a distributed configuration store
  - Aside from the admin the other instances are assigned to 2 other leaders
  - This used to be random but is now the first 2 leaders (odd number needed and we use 3)
    - There is a cluster ID
    - A broker ID for the admin and leaders
    - Zookeeper elects a “leader” (not synonymous with SU leader) for each topic
    - The other brokers are replicas
- kafka-msg-processor runs on admin, leader for filtering crayex telemetry data, processing fruinventory data and for the autocase feature specific to XD2000/6500
- OpenSearch and TimeScaleDB provide persistent storage
    - confluent-kafka-connect. cm monitoring connect status and /var/log/kafka/connect.log (timescale)
    - Logstash: started with cm monitoring elk start and /var/log/logstash/ (opensearch)

Kafka

- List the topics

```
kafka-topics --bootstrap-server admin:9092 --list
```

- Logs: `/var/log/kafka` and `/var/log/confluent`

```
cm monitoring kafka status

cm monitoring kafka status -v
```

- Two command line ways to view consumed data, depending on format:

  `kafka-console-consumer` or **`kafka-avro-console-consumer`**

- Example with avro

```
# kafka-avro-console-consumer --bootstrap-server admin:9092 --topic metric_cooldev_craycdu12 --max-messages=1
{"name":"shinercdu","timestamp":1678452287000,"device_type":"CCDU","CDU_Current_Phase_1":{"float":0.0},"CDU_Current_
Phase_2":{"float":0.0},"CDU_Current_Phase_3":{"float":0.0},"VFD1_Current":{"float":6.9},"VFD2_Current":{"float":6.2}
,"VFD1_RunTime_Energy_Counter":{"int":8913},"VFD2_RunTime_Energy_Counter":{"int":8875},"Relative_Humidity":{"float":
<snip />
```

Kafka disk space usage – retention period

- HPCM 1.13: Retention now setup with the initial cm monitoring setup command (default or options) all topics now use the reduced template
- For older:
- Retention periods are set in templates to make topic configuration persistent in case it needs to be re-created for whatever reason. CrayEX used the reduced template but other topics were kept for longer.
- /etc/kafka/topics/templates/reduced.template:

  ```
  retention.ms=43200000
  ```

  Specific topics will use the template e.g.

  ```
  /etc/kafka/topics/crayex_telemetry.topic:template=reduced.template
  ```

  If you have something different to that defined in /etc/kafka/server.properties or template it can be seen with:

  ```
  kafka-topics --bootstrap-server admin:9092 --describe
  ```

Kafka disk space usage – retention period

- 1.13 default is 1 day
  - "cm monitoring setup" option
- Previously and in CSM:
  - crayex_telemetry is 24hrs and all others 168hrs (7 days)
  - `log.retention.hours` in `/etc/kafka/server.properties` followed by a kafka restart
- Log retention can also be configured based on size e.g., log.retention.bytes. /etc/kafka/server.properties can be altered for the global settings

Topic level retention can be changed while running

```
# kafka-configs --bootstrap-server admin:9092 --entity-type topics --alter --
entity-name crayex_telemetry --add-config retention.ms=43200000
```

This example uses 43200000ms (12hrs)

```
  kafka-topics --bootstrap-server admin:9092 --describe --topic crayex_telemetry
```

Kafka disk space usage – retention period

```
admin:/etc/kafka # grep ^log.retention.hours server.properties
log.retention.hours=168
admin:/etc/kafka # vi server.properties
admin:/etc/kafka # grep ^log.retention.hours server.properties
log.retention.hours=60
admin:/etc/kafka # grep "retention.ms" topics/templates/reduced.template
retention.ms=86400000
admin:/etc/kafka # grep ^template topics/crayex_telemetry.topic
template=reduced.template
admin:/etc/kafka # cm monitoring kafka restart
Running restart command for kafka services
Running restart command for confluent-zookeeper services
<truncated for brevity>
```

What data flows by default?

- HPCM 1.13: Some pipelines configured automatically with cm monitoring setup (option needed for some and some still manual)
- Previously: If remlog-collect is enabled, then **iLO/BMC redfish logs** will start to flow to kafka.
- The only other pipelines enabled by default are those using subsmon when kafka is enabled and that will configure redfish **subscriptions to nC, cC, sC** and **logs** via logstash when elk is enabled.

```
# systemctl status remlog-collect | cat
remlog-collect.service - HPCM Remote log collector
     Loaded: loaded (/usr/lib/systemd/system/remlog-collect.service; enabled; vendor preset: disabled)
     Active: active (running) since Tue 2023-12-05 09:45:09 CST; 21h ago
   Main PID: 36759 (RemLogCollect /)
      Tasks: 15
     CGroup: /system.slice/remlog-collect.service
             └─ 36759 "RemLogCollect /opt/clmgr/remlog-collect/tlib/twistd -o -n --pidfile= -y 
/opt/clmgr/remlog-collect/sacmain.tac"
Dec 05 09:45:09 admin systemd[1]: Started HPCM Remote log collector.
```

VictoriaMetrics

- Timescale is deprecated for time series data
- Replaces Prometheus as the technology behind SIM
- 1.12 must have latest patches to get the new dashboards etc.
- Timescale dropped multi-node support
- VictoriaMetrics stores time series data in MergeTree-like data structures
- Better out of the box on disk utilisation

```
# cm monitoring victoria status -v
node         vmstorage
|            |     vminsert
|            |     |     vmselect
admin        OK    OK    OK
leader1      OK    OK    OK
leader2      OK    OK    OK
leader3      OK    OK    OK
```

VictoriaMetrics

- Drop in replacement for Prometheus – exporters are the same

```
# cm sim status
Running is-active for vmagent service : vmagent.service
admin: active
Running is-active for vmalert service : vmalert.service
admin: active
Running is-active for alertmanager service : alertmanager.service
admin: active
Running is-active for core-services service: node_exporter.service
admin: active
leader1: active
leader2: active
leader3: active … <truncated>
```

VictoriaMetrics

- Logs to journald

- **`cm support moncollect`** will capture

```
for VN in $(cat /opt/clmgr/etc/victoria-metrics-node.lst); do
     ssh $VN 'systemctl status vmstorage --no-pager -l' > ${VN}_systemctl_vmstorage
     ssh $VN 'journalctl --no-pager -l -xu vmstorage' > ${VN}_journal_vmstorage
     ssh $VN 'systemctl status vminsert --no-pager -l' > ${VN}_systemctl_vminsert
     ssh $VN 'journalctl --no-pager -l -xu vminsert' > ${VN}_journal_vminsert
     ssh $VN 'systemctl status vmselect --no-pager -l' > ${VN}_systemctl_vmselect
     ssh $VN 'journalctl --no-pager -l -xu vmselect' > ${VN}_journal_vmselect
done
```

- Uses flow-*.service not confluent-kafka-connect connectors

```
# cm monitoring flow metrics slingshot-perf
flow-slingshot-perf
       flow_consumed_messages: 659745099679300
       flow_transformed_messages: 659745099679300
       flow_samples_parsed: 659745099679300
       flow_samples_written: 659745099679300
       flow_transform_errors: 0000
       flow_write_errors: 0000
       flow_flush_errors: 0000
       flow_reconnect_attempts: 0000
       flow_line_too_long: 0000
```

ⓘ No option will list them all

VictoriaMetrics

- List metrics: cm telemetry list -a

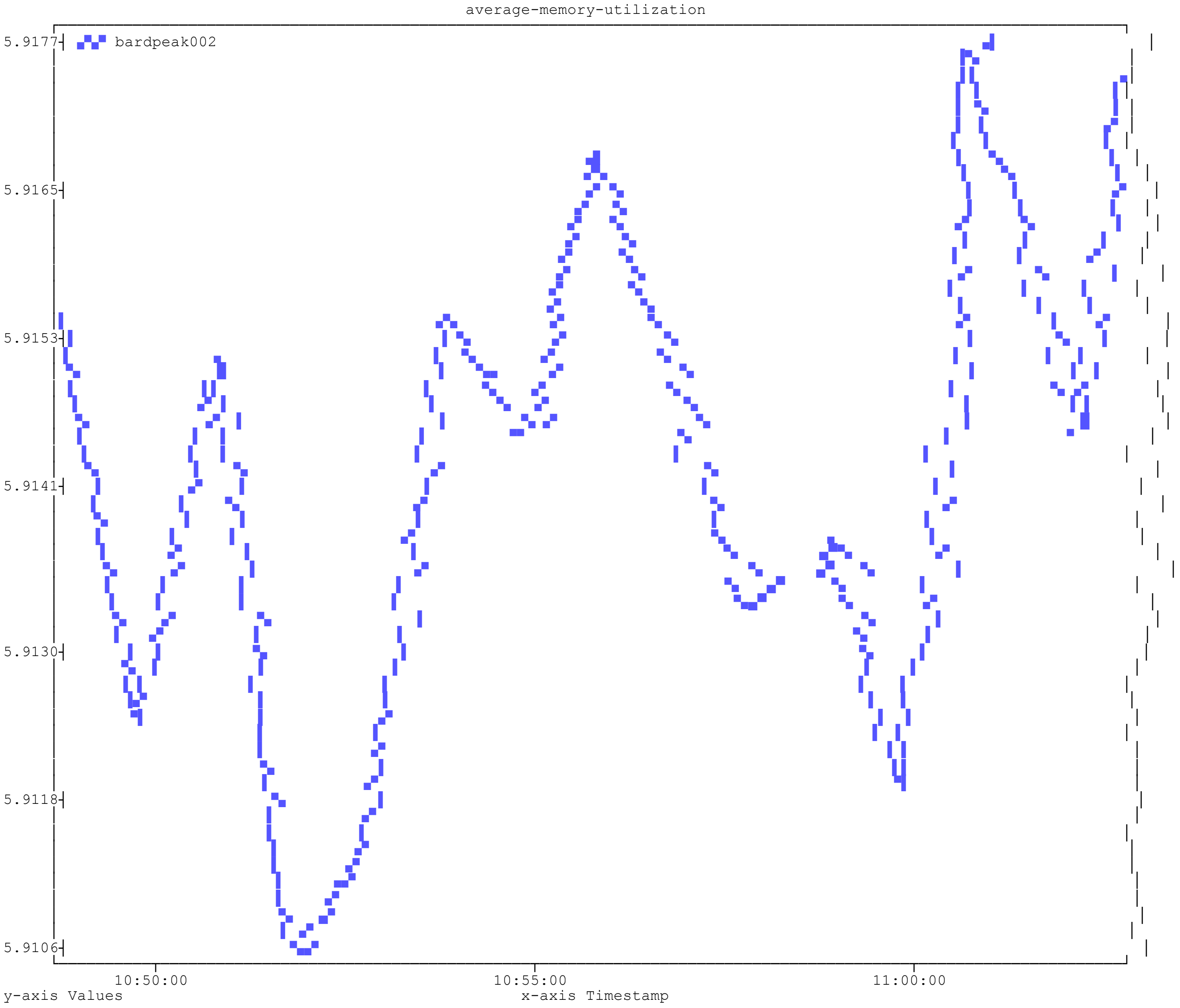
Without –a it just lists "custom" metrics

- Query metrics by many options (see the help) such as node regexp, min, max, eaverage, time period, duration etc. as plain text, csv, json, chart etc.

```
# cm telemetry query -n bardpeak002 average-memory-utilization
timestamp           host        values
2025-02-28T10:48:31 bardpeak002 5.9181
2025-02-28T10:49:31 bardpeak002 5.9136
2025-02-28T10:50:31 bardpeak002 5.9132
2025-02-28T10:51:31 bardpeak002 5.913
2025-02-28T10:52:31 bardpeak002 5.9107
2025-02-28T10:53:31 bardpeak002 5.9146
2025-02-28T10:54:31 bardpeak002 5.9148
2025-02-28T10:55:31 bardpeak002 5.9164
2025-02-28T10:56:31 bardpeak002 5.9156
2025-02-28T10:57:31 bardpeak002 5.9127
2025-02-28T10:58:31 bardpeak002 5.914
2025-02-28T10:59:31 bardpeak002 5.9119
2025-02-28T11:00:31 bardpeak002 5.9159
2025-02-28T11:01:31 bardpeak002 5.9166
2025-02-28T11:02:31 bardpeak002 5.9157
```

VictoriaMetrics

```
cm telemetry query -n bardpeak002 -o chart average-memory-utilization
```



TimescaleDB

- Timescale is for time-series data and runs on admin and leaders
- Timescale is largely deprecated in 1.12 once patches are applied
- It will not be there in a future release
- It partitions tables on a time range; these partitions are called chunks
- Its core is based on postgres

TimescaleDB

- One node is the “access” node and acts a gateway to all read and write queries
- Others are “data” nodes which store the data and service queries
- Data replication occurs between data nodes
- Patroni, zookeeper and postgres streaming replication maintain access replicas running on two other nodes to handle a failure of the access node
- HA Proxy is used so that queries always go to the access primary or, if it fails, one of the access replicas
- Timescale has compression and retention built-in – default retention = 30 days

TimescaleDB

```
# cm monitoring timescaledb status
Data Node Status
ld01 - postgres: active connection: success pingable: True
ld02 - postgres: active connection: success pingable: True
ld03 - postgres: active connection: success pingable: True
Access Node Status
admin - patroni: active role: leader postgres: running lag: none connect: success
Zookeeper Status
zookeeper: active
HAProxy Status
haproxy: active
connect: success
monitoringdb Version
1.5
```

TimescaleDB

```
# cm monitoring timescaledb show --metrics
 name                                         | category  | type   | timestamp scale |
compression interval (sec) | retention interval (sec)
-------------------------------------------------------------------------------------
 Actuator_2_Feedback_Position                 | cooldev   | FLOAT8 | 1000            |
604800                     | 2592000
 CDU_Current_Phase_1                          | cooldev   | FLOAT8 | 1000            |
604800                     | 2592000
 CDU_Current_Phase_2                          | cooldev   | FLOAT8 | 1000            |
604800                     | 2592000
 CDU_Current_Phase_3                          | cooldev   | FLOAT8 | 1000            |
604800                     | 2592000
 CDU_Power                                    | cooldev   | FLOAT8 | 1000            |
604800                     | 2592000
<snip />
# ls /opt/clmgr/postgresql/var/lib/pgsql/14/data/log/
postgresql-Fri.log  postgresql-Mon.log  postgresql-Sat.log  postgresql-Sun.log  postgresql-
Thu.log  postgresql-Tue.log  postgresql-Wed.log
Check /var/log/messages for haproxy and patroni
# psql -h admin -p 5434 -U postgres -d monitoringdb
```

Timescale disk space usage – retention and compression

- Timescale has compression and retention built-in – default retention = 30 days
- The compression interval and retention interval can be changed using cm monitoring timescaledb with the following 2 options:

compression: Adjust compression policy for metric(s) stored in Timescaledb

retention: Adjust retention policy for metric(s) stored in Timescaledb

View the current settings with:

```
# cm monitoring timescaledb show --metrics
 name                    | category | type   | timestamp scale| compression interval (sec)| retention (sec)
-----------------------------------------------------------------------------------------------
 CrayTelemetry.Current | cray     | FLOAT8 | 1000           | 604800                    | 2592000
 CrayTelemetry.Energy  | cray     | FLOAT8 | 1000           | 604800                    | 2592000
```

Timescale disk space usage – retention and compression

```
admin:~ # for i in slingshot cooldev pcm cray pdu disk; do cm monitoring
timescaledb retention --category  $i --interval 7d ;cm monitoring
timescaledb compression --category  $i --interval 1d ; done
```

- Valid units are d (day), w (week), m (month)
- Compression- its developers say it can "achieve 90%+ storage efficiencies".

- Check the categories to list in the above:

```
cm monitoring timescaledb show --categories
```

- Important: Metrics are only created as they come in once pipelines are configured
  - If you manually change the slingshot FMN configuration, for example, you will have to configure the retention/compression for those
  - As monitoring is configured you will need to repeat the above

Slingshot metrics are the big hitter for disk space

Connectors

- Unlike logstash for ELK (now opensearch in the CLI also), configuration is needed for the connectors
  - There are many which are not needed – make your choices:

tsdb-aiops-* are not used by most sites

tsdb-disk-stats – Are you going to use SIM?

tsdb-pcm-monitoring or **tsdb-ldms-monitoring** – native has more like Slingshot NIC and GPUs

tsdb-metric_cooldev – Do you have supported CDUs, RDHX…? (see the release notes /opt/clmgr/doc)

tsdb-slurm or tsdb-pbs – depending on scheduler

tsdb-pdu – Do you have supported PDUs (see the release notes /opt/clmgr/doc)

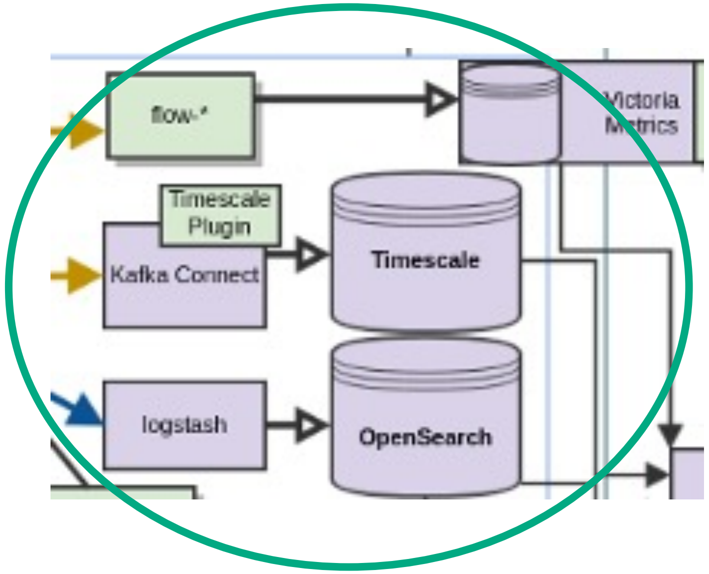
tsdb-slingshot, tsdb-slingshot-diag-perf, tsdb-slingshot-fabric-check, tsdb-slingshot-hardware

tsdb-cray-crayex_telemetry

Kafka Connect

```
admin:~ # for i in tsdb-disk-stats tsdb-metric_cooldev
tsdb-pcm-monitoring tsdb-pdu tsdb-slingshot tsdb-
slingshot-diag-perf tsdb-slingshot-fabric-check tsdb-
slingshot-hardware tsdb-slurm tsdb-cray-
crayex_telemetry; do cm monitoring connect enable --
name $i ; done
```

```
admin:~ # clush -bw 'admin,leader*' 'systemctl restart
confluent-kafka-connect'
```

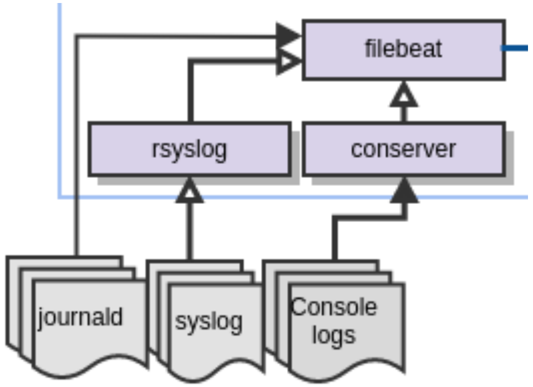


ELK or OpenSearch?

- The persistence method for logs is OpenSearch
  - OpenSearch is still called ELK under the tooling (despite the move from ElasticSearch) until 1.13

    `cm monitoring elk enable|start` will

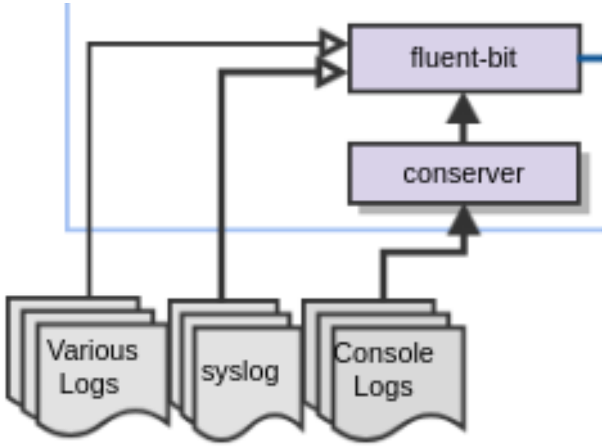
    - enable/start filebeat which captures syslog, console and journald
    - enable/start logstash to get data from kafka to opensearch



ELK or OpenSearch?

1.13: OpenSearch is called opensearch under the tooling and setup with “cm monitoring setup”

- Uses fluent-bit not filebeat which captures syslog, console and more (than before)
- Uses logstash to get data from kafka to opensearch



Log related change in 1.13: fluent-bit

```
# cm monitoring opensearch status
Opensearch Service Status

node         ping  opensearch started
|            |     |     opensearch enabled
|            |     |     |     cluster status
|            |     |     |     |     node.name
|            |     |     |     |     |     network.host
|            |     |     |     |     |     |     discovery.seed_hosts
admin        OK    OK    OK    OK    OK    OK    OK
leader1      OK    OK    OK    OK    OK    OK    OK
leader2      OK    OK    OK    OK    OK    OK    OK
leader3      OK    OK    OK    OK    OK    OK    OK

Opensearch Dashboard Service Status

node     opensearch-dashboard started
|        |  opensearch-dashboard enabled
|        |  |  opensearch DB connection
admin    OK OK OK
```

OpenSearch disk space usage – ISM policy

Index State Management Policy

- Pre 1.13:

```
# cm monitoring elk set policy <options>
```

- It only applies to indices created after it has been set so will need to manually delete older indices from that day and before.

- With 1.13 it is set up initially and can be changed with cm monitoring setup options.
- Manual index clean-up:

```
for IND in $(curl -s http://admin:9200/_cat/indices?v | grep 2024.12 | awk
'{print $3}'); do echo $IND; curl -X DELETE admin:9200/$IND ; done
```

Search OpenSearch

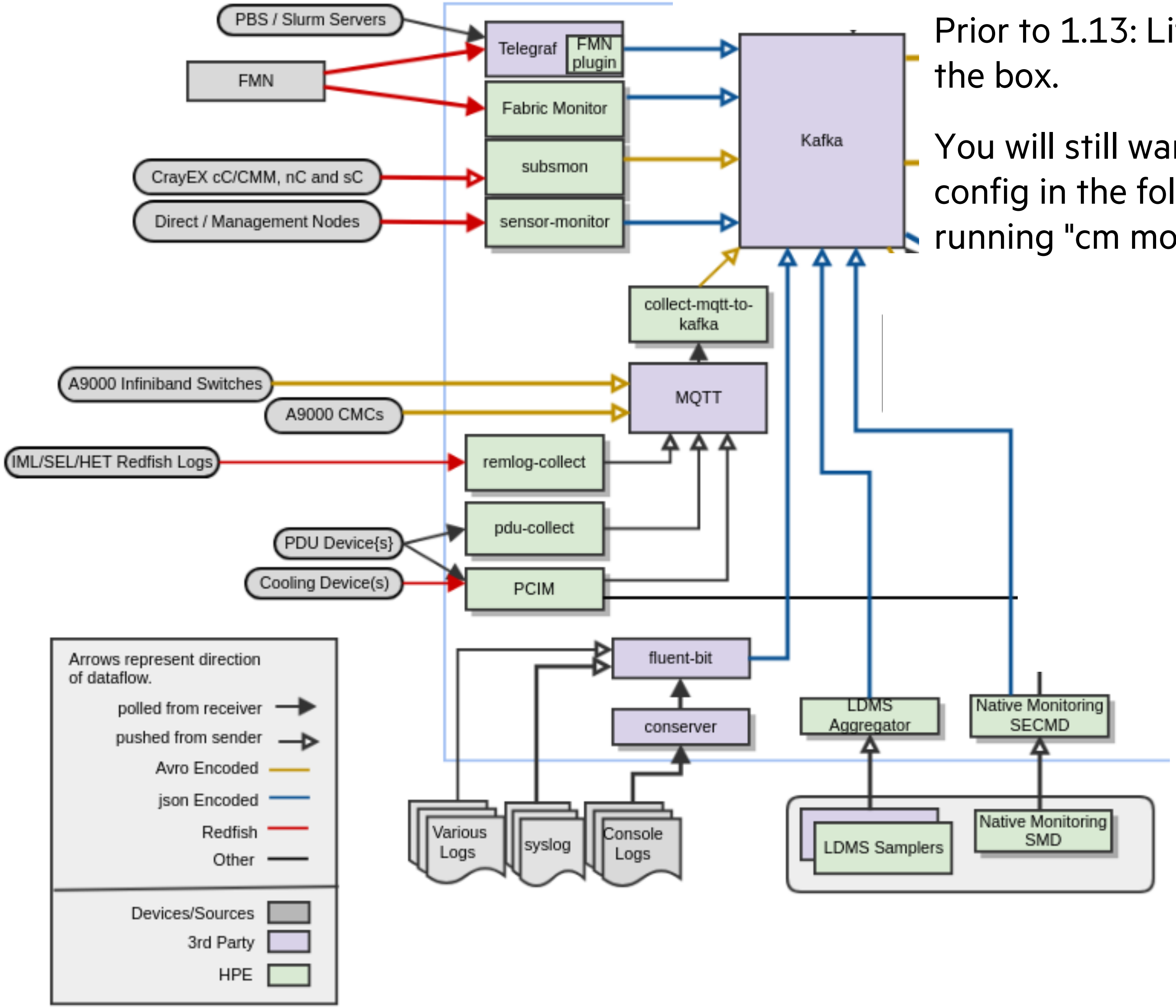
```
# cm logs -h
usage: cm logs [-h] [-t TIME] [-e TIME] [-d DURATION]
       [-q QUERY] [-n NODES]
       [-o {table,json,csv}] [-f] [--debug] [--utctime]

       {all,powerservice,ctdb,console,controller,syslog,nativemon,cmu,http,
          configuration,kafkacollection,cm,inventory,glusterfs,ldms,alert,
          victoriametrics,no_all,no_powerservice,no_ctdb,no_console,
          no_controller,no_syslog,no_nativemon,no_cmu,no_http,
          no_configuration,no_kafkacollection,no_cm,no_inventory,
          no_glusterfs,no_ldms,no_alert,no_victoriametrics}
<truncated for brevity>
```

ⓘ **syslog: Install patch 11850 for HPCM 1.13**

Search OpenSearch

```
# cm logs all -d 5s -n leader1
time                    host          type              msg
20250304 04:32:15.034   leader1       cli          I   [cli.c:840:main] 0-cli: Started running gluster with version 9.3
20250304 04:32:15.034   leader1       cli          I   [MSGID: 101190] [event-epoll.c:670:event_dispatch_epoll_worker] 0-
epoll: Started thread with index [{index=1}]
20250304 04:32:15.034   leader1       opt-clmgr-shared_stora>  I  [MSGID: 108031] [afr-
common.c:3203:afr_local_discovery_cbk] 0-cm_shared-replicate-0: selecting local read_child cm_shared-client-0
20250304 04:32:15.034   leader1       log              leader1 ctdb-eventd[7504]: 41.verify_mounts: gluster and bind mounts
seem OK.
20250304 04:32:15.034   leader1       ldms_watcher     x3000c0s33b4n0: error, cray-ldms not installed on x3000c0s33b4n0
20250304 04:32:15.034   leader1       cli          I   [MSGID: 101190] [event-epoll.c:670:event_dispatch_epoll_worker] 0-
epoll: Started thread with index [{index=0}]
20250304 04:32:15.034   leader1       cli          I   [input.c:31:cli_batch] 0-: Exiting with: 0
20250304 04:32:15.034   leader1       log              leader1 ctdb-eventd[7504]: 41.verify_mounts: mount checker monitor
script, goods after fail counter: 0, fail counter: 0
20250304 04:32:15.034   leader1       log              leader1 ctdb-eventd[7504]: 61.conserver: conserver monitor check.
20250304 04:32:16.034   leader1       SmallMonitoringDaemon_>     <1> [CMUslaveListener  ]  Action <cpuload> is already ON
!I don't do anything special MonitSlChangeActionStatus
20250304 04:32:16.034   leader1       log              leader1 ctdb-eventd[7504]: 62.cm_nfs: NFS checker monitor script,
goods after fail counter: 0, fail counter: 0
20250304 04:32:16.034   leader1       log              leader1 ctdb-eventd[7504]: 62.cm_nfs: Performing NFS mount command
checks to verify the server.
20250304 04:32:16.034   leader1       log              leader1 ctdb-eventd[7504]: 62.rsyslog: Connection to 127.0.0.1 514
port [tcp/shell] succeeded!
20250304 04:32:16.034   leader1       log              leader1 ctdb-eventd[7504]: 81.haproxy_up: haproxy monitor check.
20250304 04:32:16.034   leader1       log              leader1 ctdb-eventd[7504]: 62.cm_nfs: NFS services seem OK (ganesha).
20250304 04:32:16.034   leader1       log              leader1 ctdb-eventd[7504]: 62.rsyslog: rsyslog service monitor check.
20250304 04:32:16.034   leader1       log              leader1 ctdb-eventd[7504]: 63.cm_iscsi: kernel LIO iscsi service
monitor check.
```

Producers

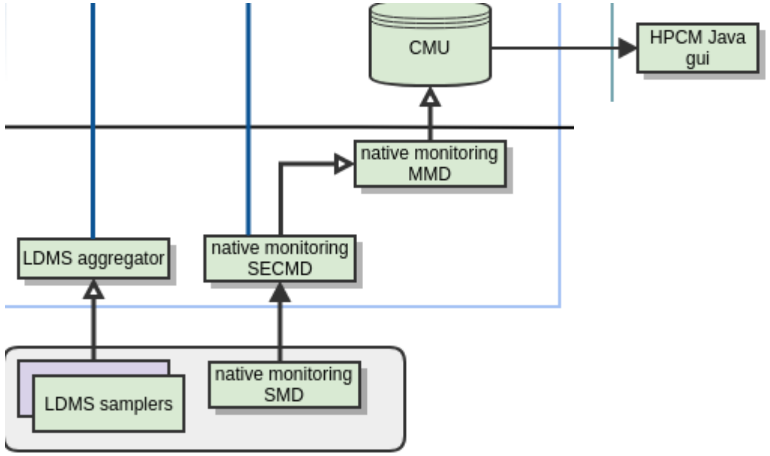


Prior to 1.13: Little was configured out of the box.

You will still want to consider custom config in the following slides prior to running "cm monitoring setup"

Node Level Monitoring

- Do you want to use LDMS or native monitoring? Both are an option to "cm monitoring setup"
- Most sites go with native because:
- More dashboards are provided
- Native monitoring populates the java GUI
- More metrics are collected out of the box
- Native retries starts of appropriate daemons
- pdsh timeouts to compute can cause issues with LDMS (if many nodes are down it won't start cluster wide)
- LDMS requires non-default rpms:
  - admin,leader*: cray-ldms,cray-ldms-store-kafka
  - Compute: cray-ldms,cray-ldms-cray_cxi,cray-cxi (latter 2 for slingshot)



Native Monitoring

- "System Monitoring", "AMD MI250X GPU Monitoring Dashboard", "NVIDIA GPU Monitoring Dashboard”
  - `MainMonitoringDaemon` (Main or MMD) runs on the admin (as well as Sec and SMD for its NE (network entity/group)
  - ssh to some nodes (elected from ones booted as can be seen in logs under /opt/clmgr/log and starts the `SecondaryServerMonitoringDaemon` (Sec)
  - Sec connects to other nodes in its network group to start the `SmallMonitoringDaemon` (SMD).
  - For hierarchical clusters these network groups a based on the rack leader
    - On ICE, the Sec runs on the rack leader and helps the admin with data for its rack
  - Customisable to run any command on a node to collect metrics:

    `/opt/clmgr/etc/ActionAndAlertsFile.txt` (AAA file)

Native Monitoring – Config Considerations

- **Decisions: user, any ssh restrictions, frequency**
- **Before** starting anything consider your configuration!

  `/opt/clmgr/etc/cmuserver.conf`

  - (there are more in the same section of the file e.g. `CMU_NONROOT_USER_ACCOUNT_KEY_TYPE`):

```
admin# grep ^CMU_MONIT /opt/clmgr/etc/cmuserver.conf
CMU_MONITORING_SYNCHRO=true
CMU_MONITORING=on
CMU_MONITORING_USER=root
CMU_MONITORING_USER_UID=default
CMU_MONITORING_USER_GID=default
CMU_MONITORING_INTERVAL=5
CMU_MONITORING_MEMLOCK=off
CMU_MONITORING_PRIORITY=0
CMU_MONITORING_HISTORY_FILES=300
CMU_MONITORING_STATUS_CHK=0
```

Native Monitoring – Dedicated User

```
# local user account on compute nodes to run CMU monitoring agents
# if 'root' then legacy mode: monitoring agents run as root user
# otherwise, the Administrator needs to make sure that the relevant
# CMU_MONITORING_USER settings are correct here below, save and exit,
# and then run the following command:
#
#  /opt/clmgr/tools/cm_config_nonroot_mon_user -c
#
# This command will create this user account in /opt/clmgr/users/hpemon/
# and create and synchronize user ssh keys between the admin node and all
# of the existing HPCM images (except for autoinstall images).
# The last step to enable a non-root monitoring user is to restart
# monitoring and redeploy the updated image to the compute nodes.
# NOTE #1: Do not create this user account beforehand, HPCM will create it.
# NOTE #2: Make sure to rerun the 'cm_config_nonroot_mon_user -c' command
#          whenever a new image is created and before it is deployed.
```

- Known bug in `/opt/clmgr/tools/cmu_mon_ssh_wrapper` fixed in 1.11

  `keypath=/opt/clmgr/etc/$user/.ssh/id_$key`

  - Needs to be:

  `keypath=/opt/clmgr/users/hpemon/$user/.ssh/id_$key`

Native Monitoring

- On 1.13, it is an option to the 1 setup command

```
# cm monitoring setup -p native
Successfully enabled the dashboards:
node_diagnostics.json
Successfully enabled the dashboards:
amd_mi250x_gpu_monitoring.json
nvidia_gpu_monitoring.json
system_monitoring.json
finished!
```

Native Monitoring

- 1.10 separated out a systemctl service cmu from cmdb – just on the admin
  - Enable native HPCM monitoring either globally or per-node using -n
  - Make sure compute nodes have Slingshot and GPU software installed

```
admin:~ # cm monitoring native enable
admin:~ # cm monitoring native start
Adjusting nodes in network group admin
Adjusting nodes in network group rack8000
monitoring daemon started
```

ⓘ **Now done by cm monitoring setup**

- <1.12 needs the timescale connector:

```
admin:~ # cm monitoring connect enable --name tsdb-pcm-monitoring
```

- Status really needs to be verified on nodes:

```
ps -elf| grep Monit
admin:~ # cm monitoring native status
Running
```

ⓘ **1.12 patched/1.13: flow/Victoria Metrics**

Native Monitoring - The Sec (sometimes termed aggregator)

- Where should the Sec run?

```
cm monitoring native set -p <priority> -n <node>
```

Meaning	Priority
The node can never become the aggregator node. i.e. Does not run the Sec	-1
The node can run the Sec if higher priority nodes are unavailable.	0
Higher pri nodes chosen first e.g. 10 chosen over 5	1 to n

ⓘ **Value stored in the DB so not reliant on the cmu service**

Native Monitoring – Additional metrics

- `cm monitoring native metrics add -g <group> [-N <Max # NICs>]`

can be used to add groups of metrics

- This is in addition to anything added to the AAA file manually
- <1.12: Once you have added all metrics and data is flowing review your timescale retention/compression!

Meaning	Group
AMD GPU metrics	gpu-amd
INTEL GPU metrics	gpu-intel
NVIDIA GPU metrics	gpu-nvidia
Metrics for each Slingshot NIC	slingshot

Need to add the slingshot group even with cm monitoring setup –pipelines slingshot_hardware native

Native Monitoring – Additional metrics

```
admin:~ # cm monitoring native metrics add -g slingshot –N=4

You are about to update the HPCM ActionsAndAlerts.txt file with metrics for
monitoring slingshot devices.
Continue? [y/N] y

Slingshot monitoring successfully configured.

Copy of original /opt/clmgr/etc/ActionAndAlertsFile.txt can be found in
/opt/clmgr/etc/ActionAndAlertsFile.txt_before_cm_config_slingshot

Please restart HPCM monitoring to enable these changes.

admin:~ # cm monitoring native restart
initiating monitoring shutdown...
Running: SendUdpMessage -client -host 127.0.0.1 -port 48559 -haltAll
checking every 5 seconds if monitoring is stopped...
starting monitoring...
Adjusting nodes in network group rack8000
Adjusting nodes in network group admin
monitoring daemon started
```

Native Monitoring – tsdb retention/compression

- <1.12: Check if the SMD is running on the nodes and if so:

```
admin:~ # for i in slingshot cooldev pcm cray pdu disk; do cm
monitoring timescaledb retention --category  $i --interval 7d ;cm
monitoring timescaledb compression --category  $i --interval 1d ;
done
```

- Valid units are d (day), w (week), m (month)
- This could be done for just the pcm category but, I like to regularly make sure all the catgories are covered see `cm monitoring timescaledb show --categories`

Native Monitoring Troubleshooting

- Look for the daemons on nodes as the "cm monitoring native status" only checks for the MMD on the admin node

```
# ps -elf | grep Monit

1 S root     2266671       1  0  80   0 - 287499 -      Mar20 ?        00:04:02
/opt/clmgr/bin/MainMonitoringDaemon -a /opt/clmgr/etc/ActionAndAlertsFile.txt -m
/opt/clmgr/etc/MetaActionFile.txt -h 172.23.0.1 -s 1 -L 0 -r 0 -k 1 -b CMDB -t 5000000 -d 1 -e 1 -f 1 -R 0

1 S root     2267051       1  0  80   0 - 146205 -      Mar20 ?        00:00:06
/opt/clmgr/bin/SecondaryServerMonitoringDaemon -h 172.23.0.1 -S 172.23.0.1 -o 49141 -O 50303 -i 48558 -n
/opt/clmgr/etc/NodesList.txt -a /opt/clmgr/etc/ActionAndAlertsFile.txt -t 5000000 -e 1 -f 1 -s 1 -L 0 -r 0
-p 172.23.0.1 -k 1 -b admin:9092

1 S root     2267389       1  0  80   0 - 84681 -       Mar20 ?        00:00:11
/opt/clmgr/bin/SmallMonitoringDaemon -h 172.23.0.1 -o 48560 -O 49722 -i 48557 -a
/opt/clmgr/etc/ActionAndAlertsFile.txt -t 5000000 -M 172.23.0.1 -f 1 -s 1 -L 0 -r 0 -p 172.23.0.1
```

```
# systemctl status cmu
  cmu.service - Cluster Manager Native Monitoring
    Loaded: loaded (/usr/lib/systemd/system/cmu.service; enabled; preset: disabled)
    Active: active (running) since Wed 2025-03-26 12:19:58 EDT; 1 month 3 days ago
     Tasks: 2
       CPU: 1d 32min 42.259s
    CGroup: /system.slice/cmu.service
            ├─  14597 /bin/bash /opt/clmgr/tools/pcm_cluster_ping
            └─2833874 sleep 29
# cm monitoring native status
Running
```

Native Monitoring Troubleshooting

- If daemons are not starting see if you can start them manually with the correct IPs and check the errors both in the terminal and logs

```
root@leader2 ~]# LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/opt/clmgr/lib/
/opt/clmgr/bin/SecondaryServerMonitoringDaemon -h 172.20.0.1 -S 10.64.0.68 -o
49142 -O 50304 -i 48558 -n /opt/clmgr/etc/NodesList.txt -a
/opt/clmgr/etc/ActionAndAlertsFile.txt -t 5000000 -e 1 -f 1 -s 1 -L 0 -r 0 -p
10.64.0.68 -k 2 -b admin:9092
```

Native Monitoring Troubleshooting

- The debug level can be increased to 6 to get maximum verbosity:

  ```
  # grep RING_DEBUG /opt/clmgr/etc/cmuserver.conf
  CMU_MAIN_MONITORING_DEBUG_LEVEL=1
  CMU_SEC_MONITORING_DEBUG_LEVEL=1
  CMU_SMD_MONITORING_DEBUG_LEVEL=1
  ```

- Change the debug level in cmuserver.conf and reduce the debug level again after troubleshooting. Restart native monitoring (cm monitoring native restart) to generate debug logs.
- The admin node has logs for the 3 daemons as it runs all 3
  - There should be one node per network group (or NE - network entity) running the Sec
  - It runs through the nodes in the NE sequentially
  - Logs are moved to *.bak every time the daemons are re-started so you have logs from this instance and the previous instance

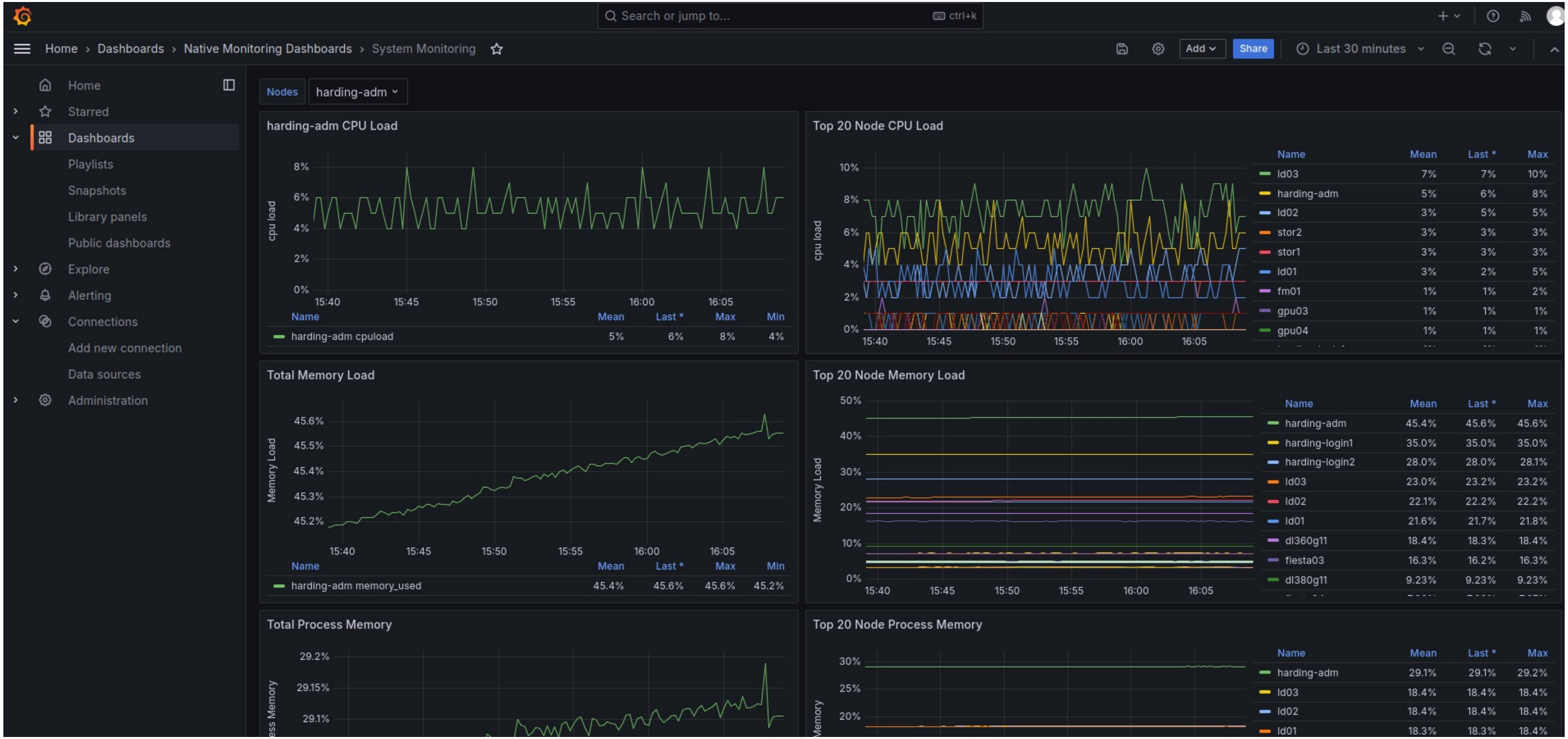
  ```
  MainMonitoringDaemon_<admin>.log, MainMonitoringDaemon_<admin.log.bak,
  SecondaryServerMonitoring_<node>.log, SecondaryServerMonitoring_<node>.log.bak,
  SmallMonitoringDaemon_<node>.log, SmallMonitoringDaemon_<node>.log.bak
  ```

- There are also some logs for start and stop with smd or sec in the names for issues stopping or starting the process

Native Monitoring

```
# cm monitoring native metrics show -n service0            #or cm telemetry in 1.13
 service0 : time = 1676012585
 service0 : __cm_monitoring_state__ = 5
 service0 : kernel_version = 5.14.21-150400.24.21-default
 service0 : cpuload = 0
 service0 : memory_used = 1.641557
 service0 : process_memory = 1.270441
 service0 : page_cache = 0.374057
 service0 : buffer_cache = 0.002942
 service0 : uptime = 7.881000
 <snip />
```

Native Monitoring Dashboards – System Monitoring



harding-adm CPU Load

Name	Mean	Last *	Max	Min
harding-adm cpuload	5%	6%	8%	4%

Top 20 Node CPU Load

Name	Mean	Last *	Max
ld03	7%	7%	10%
harding-adm	5%	6%	8%
ld02	3%	5%	5%
stor2	3%	3%	3%
stor1	3%	3%	3%
ld01	3%	2%	5%
fm01	1%	1%	2%
gpu03	1%	1%	1%
gpu04	1%	1%	1%

Total Memory Load

Name	Mean	Last *	Max	Min
harding-adm memory_used	45.4%	45.6%	45.6%	45.2%

Top 20 Node Memory Load

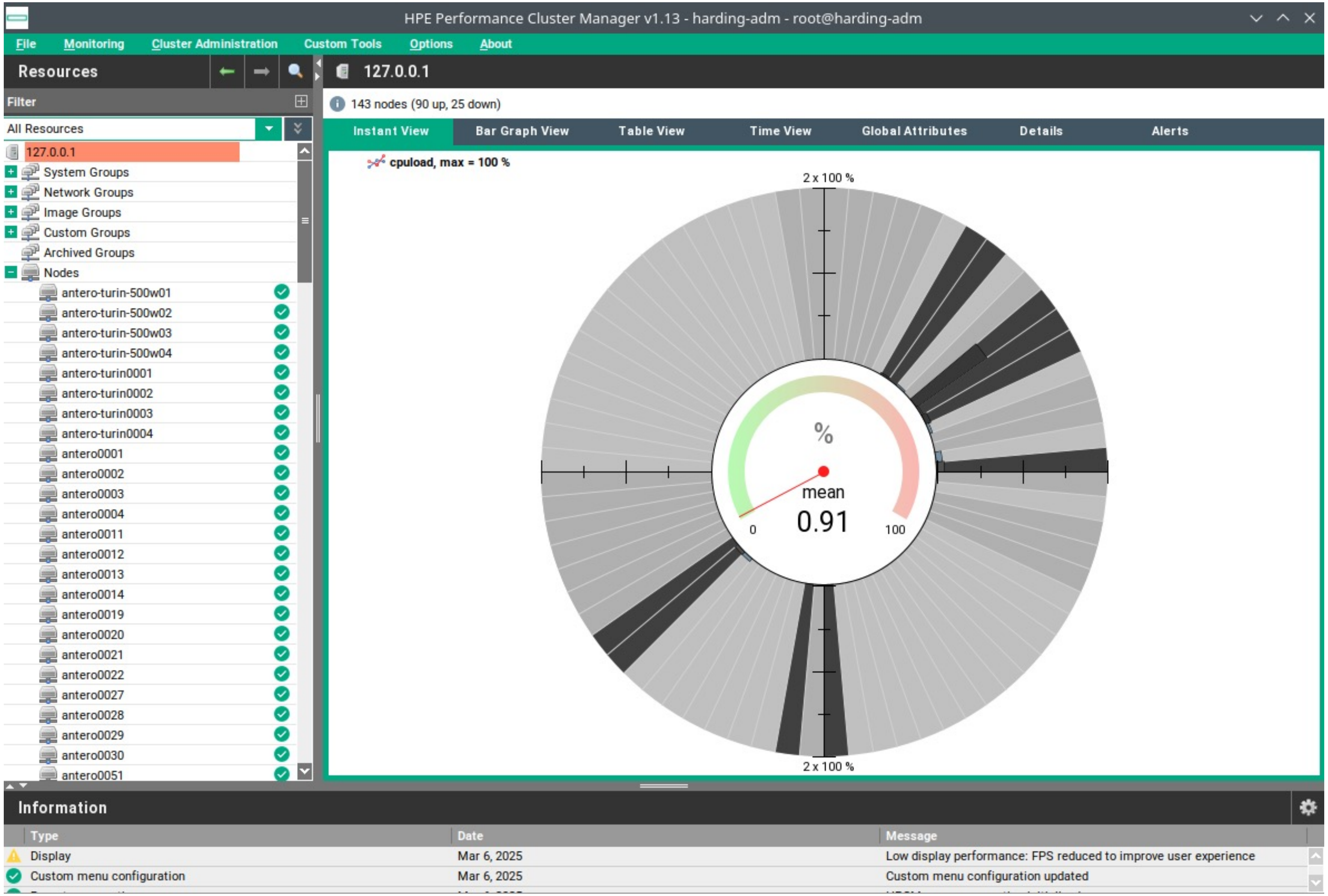
Name	Mean	Last *	Max
harding-adm	45.4%	45.6%	45.6%
harding-login1	35.0%	35.0%	35.0%
harding-login2	28.0%	28.0%	28.1%
ld03	23.0%	23.2%	23.2%
ld02	22.1%	22.2%	22.2%
ld01	21.6%	21.7%	21.8%
dl360g11	18.4%	18.3%	18.4%
fiesta03	16.3%	16.2%	16.3%
dl380g11	9.23%	9.23%	9.23%

Total Process Memory

Top 20 Node Process Memory

Name	Mean	Last *	Max
harding-adm	29.1%	29.1%	29.2%
ld03	18.4%	18.4%	18.4%
ld02	18.4%	18.4%	18.4%
ld01	18.3%	18.3%	18.4%

Native Monitoring



143 nodes (90 up, 25 down)

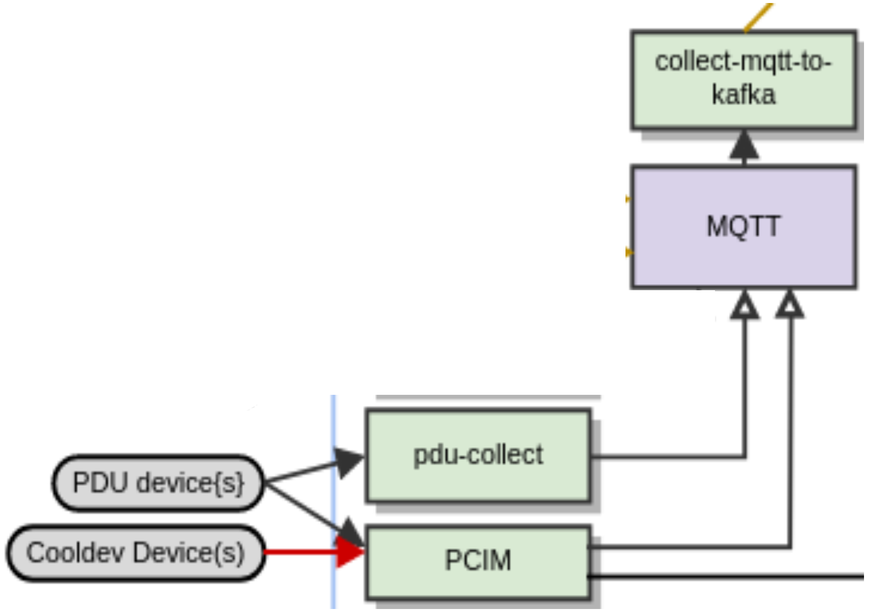
Type	Date	Message
Display	Mar 6, 2025	Low display performance: FPS reduced to improve user experience
Custom menu configuration	Mar 6, 2025	Custom menu configuration updated

Power and Cooling

- PCIM – Power and Cooling Infrastructure Monitor
- pdu-collect
  - PDU data is collected by both PCIM and pdu-collect currently
  - Both feed into the same kafka topic
  - In the future: just pdu-collect
  - Primary poll: pdu-collect 2s
  - Secondary poll: PCIM 20s

Not enabled by "cm monitoring setup"

Use systemctl!



/opt/clmgr/log/pdu-collect.log and /opt/**cmu**/pcim/log

Power and Cooling – Supported Devices

- Supported devices listed in the release notes (/opt/clmgr/doc)
  - A power distribution unit (PDU) reads AC power and energy measurements on cluster rack-level power domains
  - For the AC power measurement feature to function, the cluster must have one or more of the following PDUs:
    - Server Technology Sentry3
    - Server Technology Sentry4
    - 880459-B21 (Raritan) HPE Mtrd 3P 39.9kVA/60A 48A/277V FIO PDU
    - PX-5946V-F5V2 (Raritan) HPE Mtrd 3P 17.3kVA/48A 9brkr PDU
    - P9R82A HPE G2 Metered 3Ph 17.3kVA/60309 4-wire 48A/208V
    - P9R84A HPE G2 Metered 3Ph 22kVA/60309 5-wire 32A/230V
  - For more details on power management see the HPE Performance Cluster Manager Power Consumption Management Guide

ⓘ 1.13: Adds pdu-collect support for Enlogic (HPE) G3 PDUs

Power and Cooling – PCIM Supported Devices

- The HPE Power and Cooling Infrastructure Monitor provides insight into the state of the hardware related to the power and water-cooling components of an HPE water-cooled solution
- Supported devices include the following:
  - HPE Apollo 9000 CDU (Cooling Distribution Unit)
  - HPE Apollo 9000 Chassis (Power Supplies and Switches)
  - HPE Cray EX CDU (1.2 MW and 1.6 MW)
  - Apollo DLC Passive CDU (for A2k and A6500 clusters)
  - HPE SGI 8600 CDU
  - ARCS (Adaptive Rack Cooling System)
  - SGI 8600 CRC (Cooling Rack Controller)
  - Motivair RDHX (Rear Door Heat Exchanger)
  - Raritan PDUs (Power Distribution Unit)
  - HPE Cray EX VCDU (Virtual Cooling Distribution Unit)
  - HPE PDUs
  - ServerTech Cray ClusterStor Switch 63A 400V PDU (R4M34A)
  - ServerTech Cray ClusterStor Switch 60A 415V PDU (R4M35A)

PCIM – Power and Cooling Infrastructure Manager

```
# systemctl enable pcim
```

```
# systemctl start pcim
```

- <1.12: Remember the connector from earlier:

```
# cm monitoring connect enable tsdb-metric_cooldev
```

```
# cm monitoring flow create metric_cooldev
enable all units for flow-metric_cooldev...
```

ⓘ **1.12 patched/1.13: flow & Victoria Metrics**

```
# cm monitoring flow create pdu
enable all units for flow-pdu...
```

- Cray EX: configures the CDUs and VCDUs automatically during the installation process when `cm node update config --sync pcim -n admin` is run

PCIM – Power and Cooling Infrastructure Manager

Cray EX: configures the CDUs and VCDUs automatically during the installation process when `cm node update config --sync pcim -n admin` is run

Cray XD, Apollo 9000, HPE Apollo DLC CDUs, Adaptive rack cooling systems (ARCS)

Components, Rear-door heat exchangers (RXHX): `cm cooldev` used to add them

SGI 8600/ICE CDUs and CRCs, supported PDUs: `/opt/cmu/pcim/configure_snmp_device`

- <1.12: Once metrics are flowing, set your tsdb retention/compression!
- `/opt/cmu/pcim/tools/get_metric_data` to see whether PCIM is getting metrics from devices

PCIM and pdu-collect dashboards

- <1.13: `cm monitoring dashboard grafana set --cdu enable or --cdu_ex2500 enable`

```
# cm monitoring grafana dashboard enable pdu
# cm monitoring grafana dashboard enable cdu
```

PCIM and pdu-collect dashboards and flow

```
# cm monitoring grafana dashboard enable pdu sentry
Successfully enabled the dashboards:
sentry_pdu_monitoring.json
```

```
# systemctl restart grafana-server.service
```

- Configure each PDU e.g. Sentry web interface and set the ro community string to ‘public’
  - make sure to enable snmp support, you will usually need to reboot the management module before you can use it
- Enable flow for cdu and pdu data

```
# cm monitoring flow create metric_cooldev
enable all units for flow-metric_cooldev...
```

```
# cm monitoring flow create pdu
enable all units for flow-pdu...
```

PCIM

```
# kafka-avro-console-consumer --bootstrap-server admin:9092 --topic metric_cooldev_pdu --max-messages=1
{"name":"pdu01","timestamp":1741274314179,"device_type":"STPDU","PDU_Input_Power":{"float":1083.0},
"PDU_Input_VA":null,"PDU_Input_Power_Factor":{"float":96.0},"PDU_Active_Energy":{"float":310330.0},
"PDU_Line_1_Current":{"float":2.71},"PDU_Line_1_Energy":{"float":213012.0},
"PDU_Line_1_Power":{"float":540.0},"PDU_Line_1_Power_Factor":{"float":96.0},
"PDU_Line_1_VA":null,"PDU_Line_1_Voltage":{"float":209.0},"PDU_Line_2_Current":{"float":1.85},
"PDU_Line_2_Energy":{"float":84693.0},"PDU_Line_2_Power":{"float":377.0},
"PDU_Line_2_Power_Factor":{"float":97.0},"PDU_Line_2_VA":null,"PDU_Line_2_Voltage":{"float":209.7},
"PDU_Line_3_Current":{"float":0.74},"PDU_Line_3_Energy":{"float":12625.0},
"PDU_Line_3_Power":{"float":142.0},"PDU_Line_3_Power_Factor":{"float":92.0},
"PDU_Line_3_VA":null,"PDU_Line_3_Voltage":{"float":210.5},"PDU_Temperature_1":null,
"PDU_Temperature_2":null,"PDU_Rel_Humidity":null}
# acpower --pdu-metrics
{
   "pdu01": {
       "PDU_Input_Power": 1055.0,
       "PDU_Input_Power_Factor": 96.0,
       "PDU_Active_Energy": 310310.0,

<truncated for brevity />
```

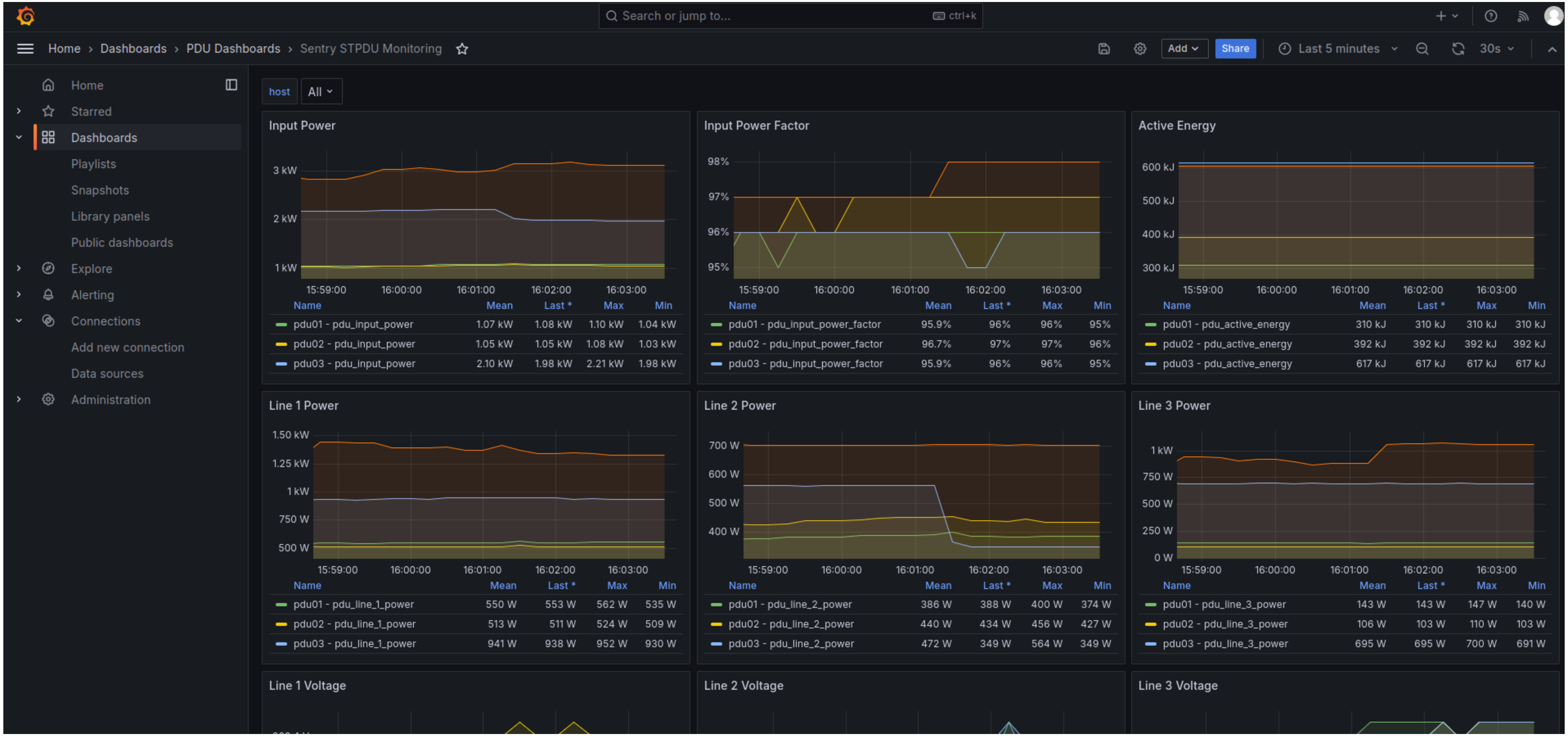
PCIM – PDUs

- This is planned to be retired

```
admin:~ # /opt/cmu/pcim/configure_snmp_device -n x3000-pdu0 -i 172.24.253.200 -v
1 -c public

admin:~ # cm cooldev cdu show

x8000cdu1 10.176.0.1

x8000cdu0 10.176.0.1

admin:~ # cm cooldev cdu|rdhx|arcs add --name <NAME> --type <2000|9000> --ip <IP>
--mac <MAC>
```

- Use `/opt/cmu/pcim/tools/find_chassis.pl -r` to create `/opt/cmu/pcim/config/.pcimchassis.conf` with IP to chassis mappings pulling information from the DB

PDU Dashboards – Sentry STPPDU Monitoring



Input Power

Name	Mean	Last *	Max	Min
pdu01 - pdu_input_power	1.07 kW	1.08 kW	1.10 kW	1.04 kW
pdu02 - pdu_input_power	1.05 kW	1.05 kW	1.08 kW	1.03 kW
pdu03 - pdu_input_power	2.10 kW	1.98 kW	2.21 kW	1.98 kW

Input Power Factor

Name	Mean	Last *	Max	Min
pdu01 - pdu_input_power_factor	95.9%	96%	96%	95%
pdu02 - pdu_input_power_factor	96.7%	97%	97%	96%
pdu03 - pdu_input_power_factor	95.9%	96%	96%	95%

Active Energy

Name	Mean	Last *	Max	Min
pdu01 - pdu_active_energy	310 kJ	310 kJ	310 kJ	310 kJ
pdu02 - pdu_active_energy	392 kJ	392 kJ	392 kJ	392 kJ
pdu03 - pdu_active_energy	617 kJ	617 kJ	617 kJ	617 kJ

Line 1 Power

Name	Mean	Last *	Max	Min
pdu01 - pdu_line_1_power	550 W	553 W	562 W	535 W
pdu02 - pdu_line_1_power	513 W	511 W	524 W	509 W
pdu03 - pdu_line_1_power	941 W	938 W	952 W	930 W

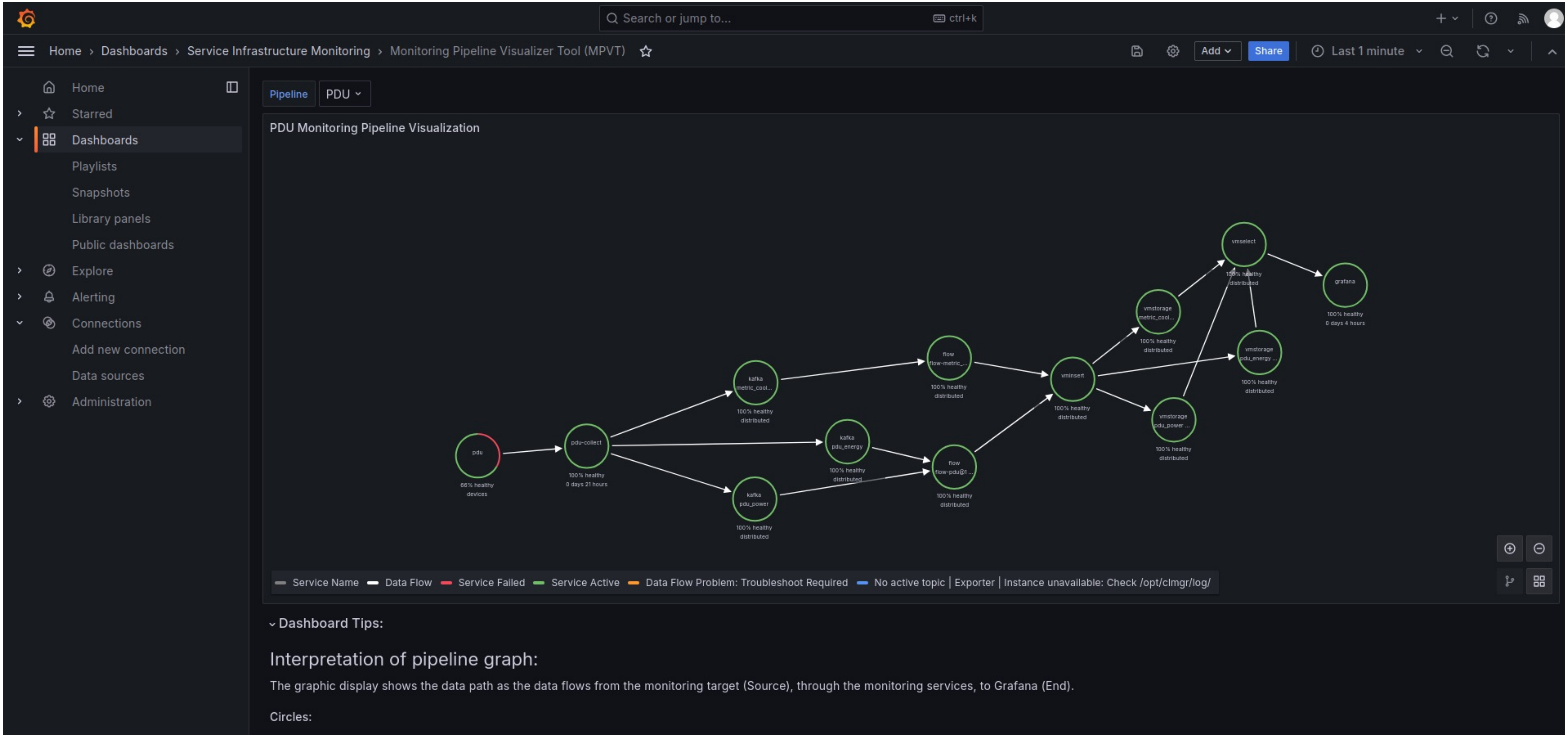
Line 2 Power

Name	Mean	Last *	Max	Min
pdu01 - pdu_line_2_power	386 W	388 W	400 W	374 W
pdu02 - pdu_line_2_power	440 W	434 W	456 W	427 W
pdu03 - pdu_line_2_power	472 W	349 W	564 W	349 W

Line 3 Power

Name	Mean	Last *	Max	Min
pdu01 - pdu_line_3_power	143 W	143 W	147 W	140 W
pdu02 - pdu_line_3_power	106 W	103 W	110 W	103 W
pdu03 - pdu_line_3_power	695 W	695 W	700 W	691 W

SIM – Monitoring Pipeline Visualizer Tool (MPVT)



PDU Monitoring Pipeline Visualization

▬ Service Name ▬ Data Flow 🟥 Service Failed 🟩 Service Active 🟧 Data Flow Problem: Troubleshoot Required 🟦 No active topic | Exporter | Instance unavailable: Check /opt/clmgr/log/

⌄ Dashboard Tips:

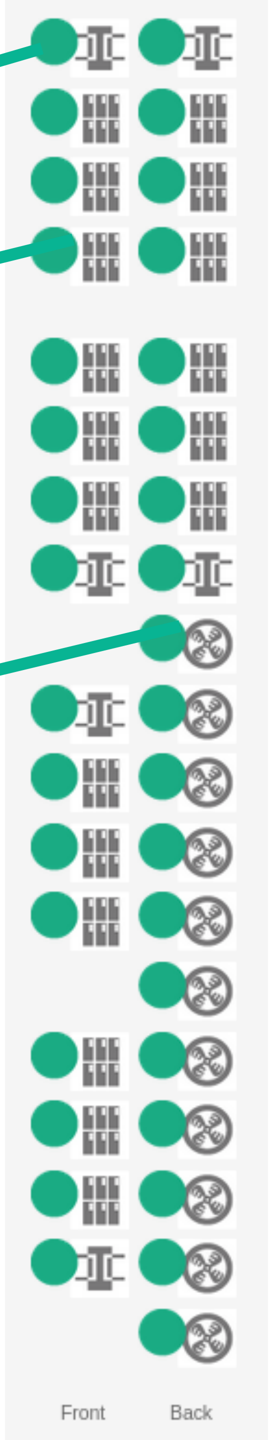
Interpretation of pipeline graph:

The graphic display shows the data path as the data flows from the monitoring target (Source), through the monitoring services, to Grafana (End).

Circles:

PCIM – Web GUI

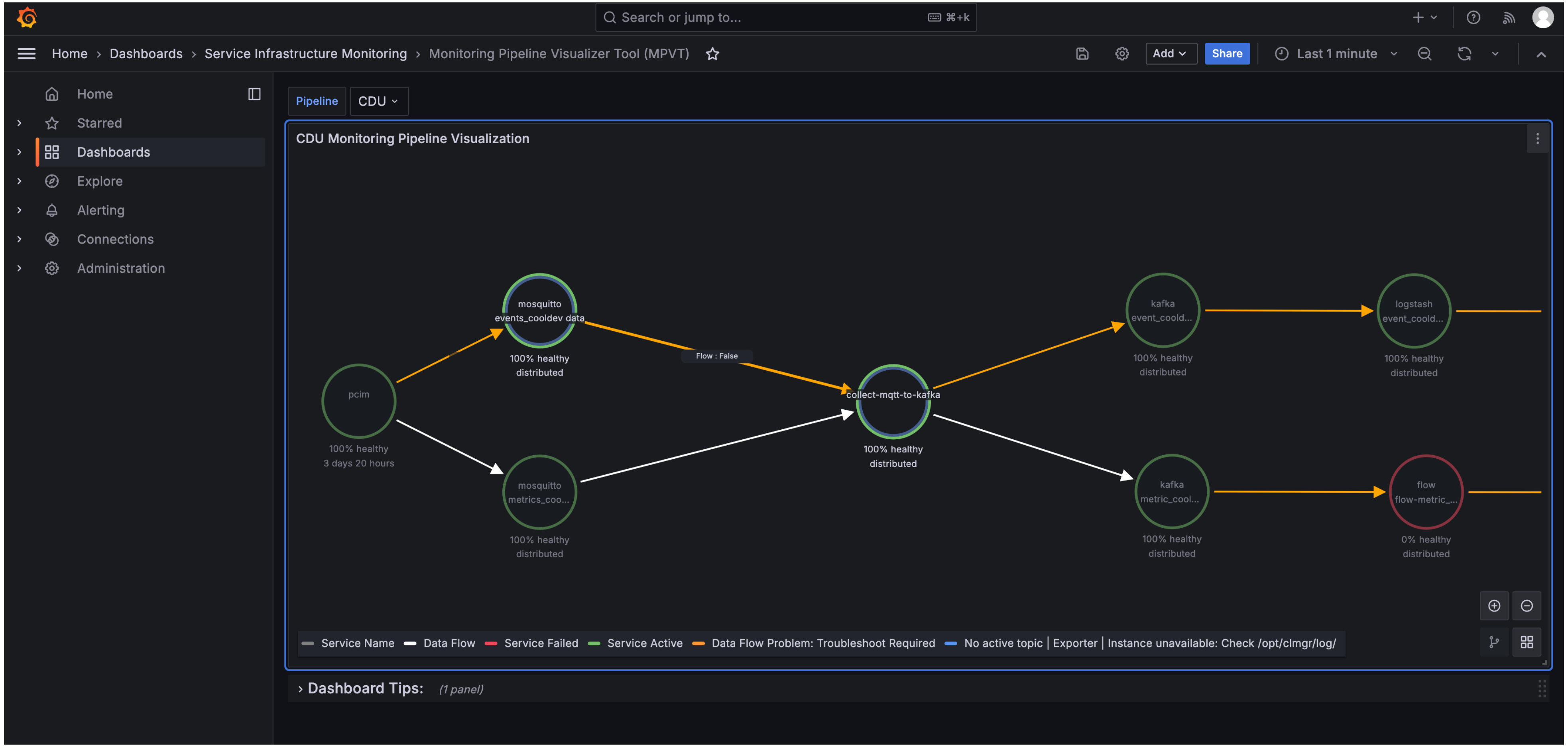
```
# cat /opt/cmu/pcim/layout.txt
x1005cdu x1105cdu
x1005 x1105
x1006 x1106
x1007 x1107

x1008 x1108
x1009 x1109
x1010 x1110
x1010cdu x1110cdu
   x3100rdhx
x1011cdu x3101rdhx
x1011 x3102rdhx
x1012 x3103rdhx
x1013 x3104rdhx
     x3105rdhx
x1014 x3105rdhx
x1015 x3106rdhx
x1016 x3107rdhx
x1016cdu x3108rdhx
   x3109rdhx
```



The CDU monitoring pipeline – Troubleshooting

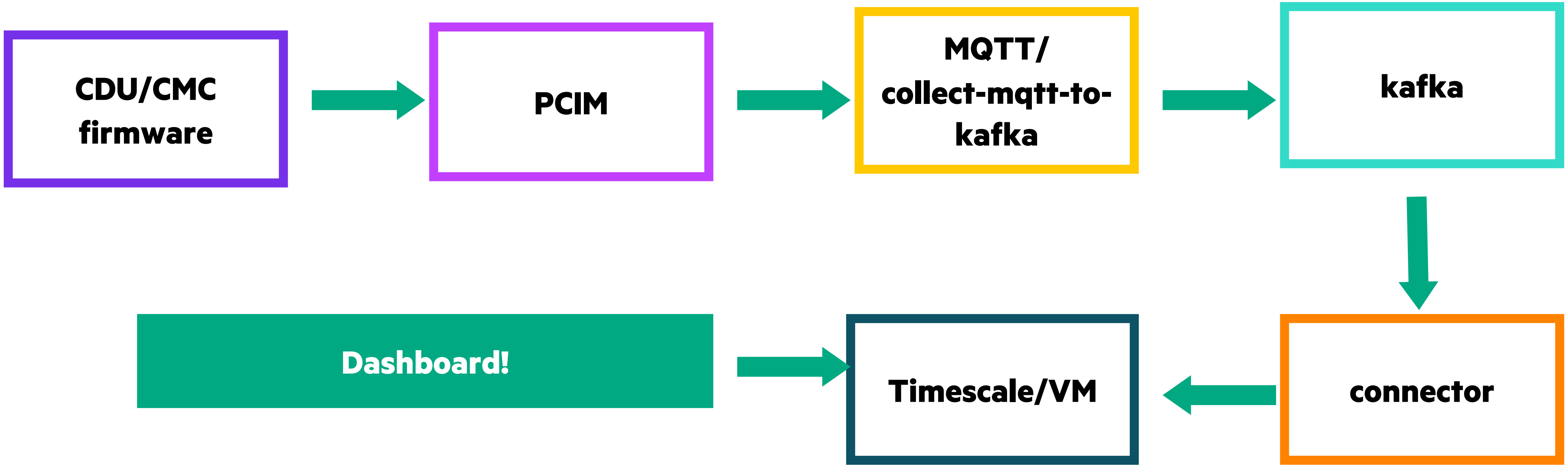
- Tech preview in 1.12! Monitoring Pipeline Visualisation Tool as part of SIM (more later on SIM):

```
cm sim add --service-group mpvt-service
```



The CDU monitoring pipeline – Troubleshooting

Start troubleshooting at the start of the pipeline



The CDU monitoring pipeline – Troubleshooting

```
x9000c1:> grep . /var/volatile/cec/cdu/plc/*
/var/volatile/cec/cdu/plc/actuator1_fb:4294967272
/var/volatile/cec/cdu/plc/actuator2_fb:35
/var/volatile/cec/cdu/plc/belimo_valve_output_volts:4.8
<snip />
```

- If it is CDU power monitoring check :

  ```
  grep . /var/volatile/cec/cdu*/power_mon/*
  ```

  - Note: EX2500 uses 4U-F version CDU (Model Code: 04205)
    - The 4U-F does not have an internal power meter

The CDU monitoring pipeline – Troubleshooting

```
cm cooldev cdu|rdhx|arcs show

systemctl status pcim

/opt/cmu/pcim/log

/opt/cmu/pcim/tools/get_metric_data

systemctl status mosquito

mosquitto_sub -t \# -v | grep cdu

systemctl status collect-mqtt-to-kafka

/opt/clmgr/log/collect-mqtt-to-kafka/collect-mqtt-to-kafka.log
/var/log/messages

kafka-avro-console-consumer --bootstrap-server admin:9092 --topic
metric_cooldev_craycdu12 –max-messages=1|jq

cm monitoring kafka status -v

/var/log/kafka/ and /var/log/confluent/
```

The CDU monitoring pipeline – Troubleshooting

```
cm monitoring connect status | cm monitoring flow status
/var/log/kafka/connect.log | journalctl -xeu flow-<topic>

cm monitoring timescaledb status | cm monitoring victoria status
/opt/clmgr/postgresql/var/lib/pgsql/14/data/log/ | journalctl
-xeu vmstorage|vminsert|vmselect
/var/log/messages
psql -h admin -p 5434 -U postgres -d monitoringdb
```

or see next slide

The CDU monitoring pipeline – Troubleshooting Timescaledb

```
# cm monitoring timescaledb show --metrics | grep cooldev
Actuator_2_Feedback_Position                 | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CDU_Current_Phase_1                          | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CDU_Current_Phase_2                          | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CDU_Current_Phase_3                          | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CDU_Power                                    | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CDU_Voltage_Phase_1_2                        | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CDU_Voltage_Phase_2_3                        | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CDU_Voltage_Phase_3_1                        | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
CWV_Valve_Actuator_Voltage                   | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
PLC_Temperature                              | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
PLC_to_VFD_Voltage                           | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
Primary_Facility_Flow                        | cooldev  | FLOAT8 | 1000            | 604800      | 2592000
<truncated for brevity>
# cm monitoring timescaledb query --metric Primary_Facility_Flow
timestamp     | location | value
--------------------------------
1714480620000 | x9000cdu | 5.4   <truncated for brevity>
```

The CDU monitoring pipeline – Troubleshooting VictoriaMetrics

```
# cm telemetry list -a| grep -i ^cdu
cdu_current_phase_1
cdu_current_phase_2
cdu_current_phase_3
cdu_power
cdu_voltage_phase_1_2
cdu_voltage_phase_2_3
cdu_voltage_phase_3_1
# cm telemetry query cdu_power
timestamp           device_type   host     name     values
2025-03-03T13:49:17 CCDU          x9000cdu x9000cdu 535
2025-03-03T13:50:17 CCDU          x9000cdu x9000cdu 535
2025-03-03T13:51:17 CCDU          x9000cdu x9000cdu 535
2025-03-03T13:52:17 CCDU          x9000cdu x9000cdu 537
2025-03-03T13:53:17 CCDU          x9000cdu x9000cdu 535
2025-03-03T13:54:17 CCDU          x9000cdu x9000cdu 534
2025-03-03T13:55:17 CCDU          x9000cdu x9000cdu 536
2025-03-03T13:56:17 CCDU          x9000cdu x9000cdu 536
2025-03-03T13:57:17 CCDU          x9000cdu x9000cdu 536
2025-03-03T13:58:17 CCDU          x9000cdu x9000cdu 536
2025-03-03T13:59:17 CCDU          x9000cdu x9000cdu 538
2025-03-03T14:00:17 CCDU          x9000cdu x9000cdu 536
2025-03-03T14:01:17 CCDU          x9000cdu x9000cdu 537
2025-03-03T14:02:17 CCDU          x9000cdu x9000cdu 534
2025-03-03T14:03:17 CCDU          x9000cdu x9000cdu 531
```

Node sensor information

- **Consider your hardware: SGI 8600/ICE, Cray EX, other iLO/BMC**
  - Nearly every site will require sensor-monitor and its helper sensor-processor whereas many will not require HET for SGI hardware or subsmon for Cray EX

```
# systemctl enable --now sensor-monitor.service
```

ⓘ **Not enabled by cm monitoring setup**

```
# cm monitoring flow create sensormon
```

- SU leaders?

```
# cm node zypper|dnf --repos Cluster-Manager-1.13-sles15sp6-x86_64 -n "leader*" install
sensor-processor
```

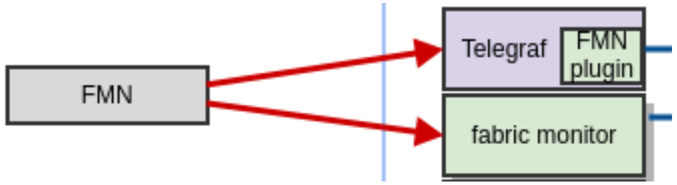
```
# pdsh -w ‘admin,leader*’ systemctl enable --now sensor-processor.service
```

- EX?
  - subsmon will establish a subscription on the controller to the admin or SU leader alias automatically as it was enabled with kafka
  - With 1.13, `cm monitoring setup -p crayex_hardware` will enable it

```
# systemctl restart subsmon
```

Slingshot

- The architecture diagram is a simplification and there are multiple pipelines involved
- Monitoring pipeline visualisation tool also does not give the full story?

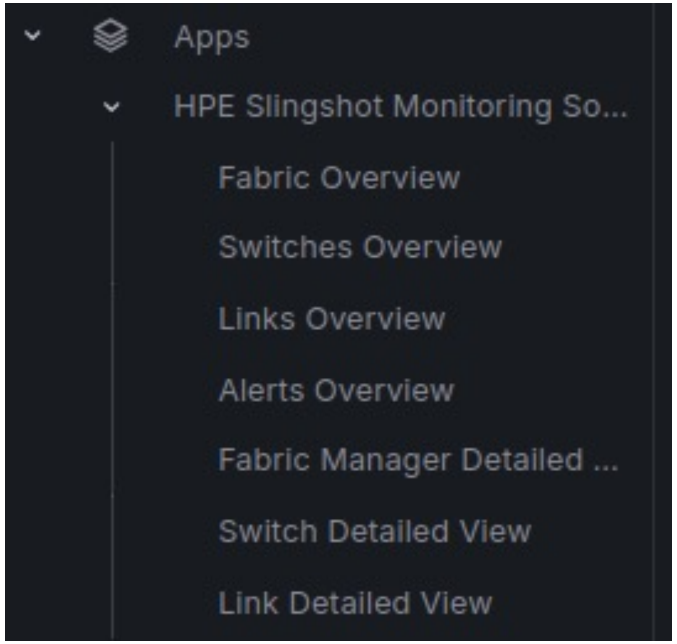


- "Alerts" dashboards become relevant once these multiple pipelines are configured and alerting is setup
  - graphs are an endpoint and everything involved in the pipeline must be configured first
- With 1.10: CHC and alerting are split out, so new dashboard is “Slingshot Alerts” (See later)
- In 1.13, metrics dashboards are still using Timescale rather than Victoria Metrics

  1 pipeline: Node level data comes from native monitoring
- 1 pipeline: subsmon (kafka enable <=1.12 and cm monitoring setup –p crayex_hardware" - sets redfish subscriptions on the switches providing switch hardware telemetry (slingshot_CraySwitchHardwareTelemetry)

Slingshot – SMS

- Slingshot Monitoring Software (SMS) largely replaces HPCM slingshot monitoring in 1.13 but it can be made to work
- There is another presentation at CUG on SMS
- SMS is a grafana app not dashboards
- Fabric AIOps (may be renamed as Slingshot AIOps and NOT HPCM Slingshot AIOps) provides other functionality and has overlap with HPCM
- Network Status from SMS: Slingshot Group Status, Slingshot Network Summary, Slingshot Switch Status
- Fabric Performance from HPCM: Slingshot Bandwidth, PausePercent, IfHCInOctets/IfHCOutOctets, rxBroadcastPkts/txBroadcastPkts, rxMulticastPkts/txMulticastPkts, rxPauseFrames/txPauseFrames
- Fabric Hardware Status (currently HPCM): Current, Power, Rotational, Temperature, Voltage
- Fabric Quality Performance (SMS): Bit Error Rate, Slingshot Routing and Hard Switch Errors



Slingshot – SMS

- 1.12 with SMS needs 11830
- cm node zypper -n admin install hpe-slingshotmonitoringsoftware-app
- Administration > Plugins and data > Plugins on the left side menu
- Select the HPE Slingshot Monitoring Software
- Optional: Slingshot AIOps URL label
  - Type the URL where the HPE Slingshot AIOps service is running
- Create user for Fabric Manger Data source
  - Log into FMN and run

```
# fmn-create-user -n sms-admin -r slingshot-admin
```

  - If this fails, please run the following

```
# fmn-enable-secure-mode --reauth-only --admin-credential root:initial0
```

- Click Add a Fabric Manager Data Source
  - Enter the detail from the FMN

```
fm01:~ # fmn-show-user -n sms-admin
{
  "clientId": "sms-admin",
  "secret": "B5Lcqn2N5HaXWBYwWnECXMYJCxr3kGNx"
}
```

Slingshot – SMS

- The FMN needs node_expoter to be installed and started

```
# cm image|node zypper -i fmn-sles15sp6 -repos Cluster-Manager-1.13-sles15sp6-x86_64 install node_exporter
# ssh fm01 systemctl start node_exporter.service
```

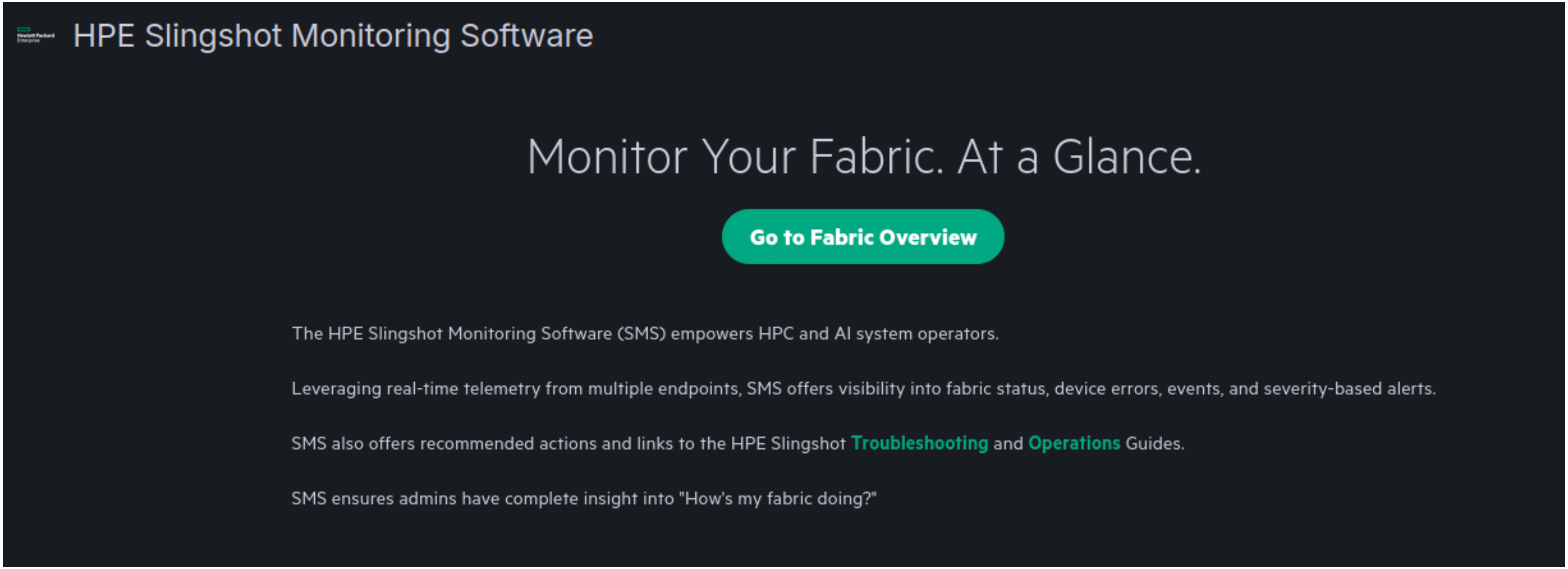
- Edit the following file to add the FMN :

```
/etc/victoriametrics/vmagent/vmagent.yml
- job_name: node-exporter
   static_configs:
   - targets:
     - ld03:9100
     - ld01:9100
     - ld02:9100
     - admin:9100
     - fm01:9100
```

- Restart SIM services

```
# cm sim service restart
```

Slingshot Monitoring Software



HPE Slingshot Monitoring Software

Monitor Your Fabric. At a Glance.

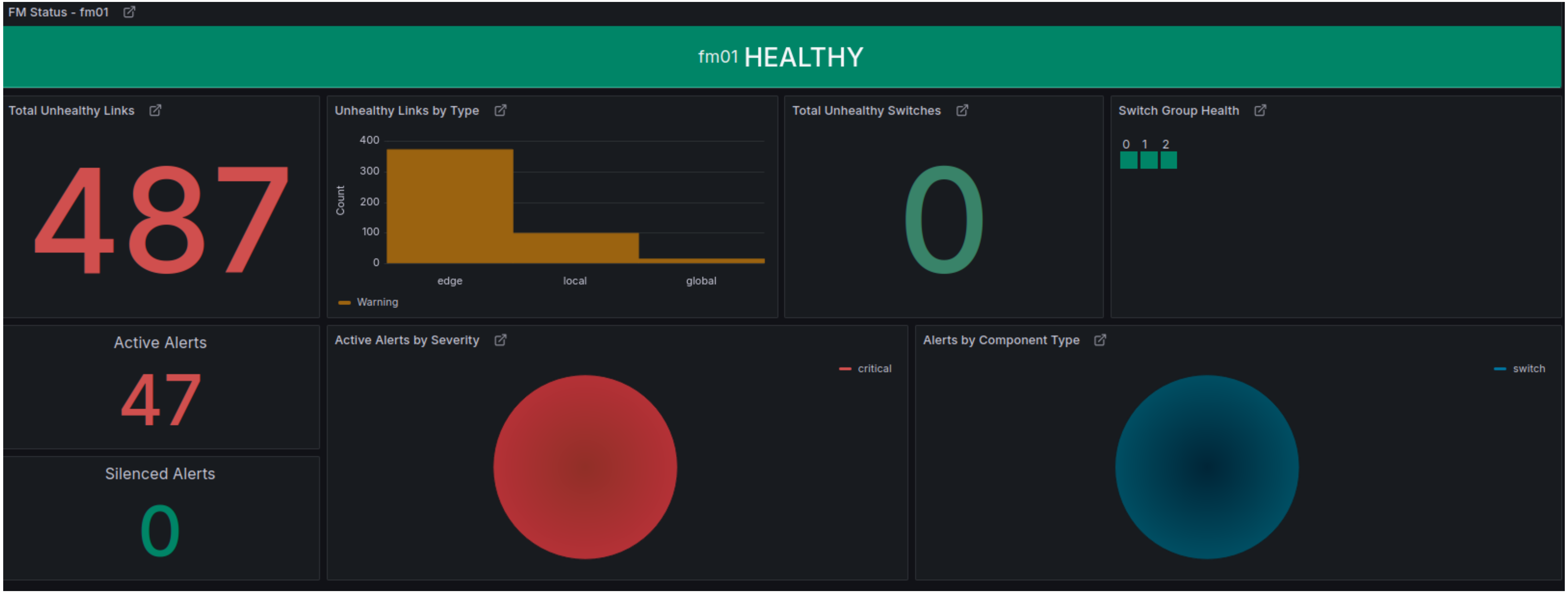
The HPE Slingshot Monitoring Software (SMS) empowers HPC and AI system operators.

Leveraging real-time telemetry from multiple endpoints, SMS offers visibility into fabric status, device errors, events, and severity-based alerts.

SMS also offers recommended actions and links to the HPE Slingshot Troubleshooting and Operations Guides.

SMS ensures admins have complete insight into "How's my fabric doing?"

Slingshot Fabric Manager Status



FM Status - fm01

fm01 HEALTHY

Total Unhealthy Links

487

Total Unhealthy Switches

0

Active Alerts

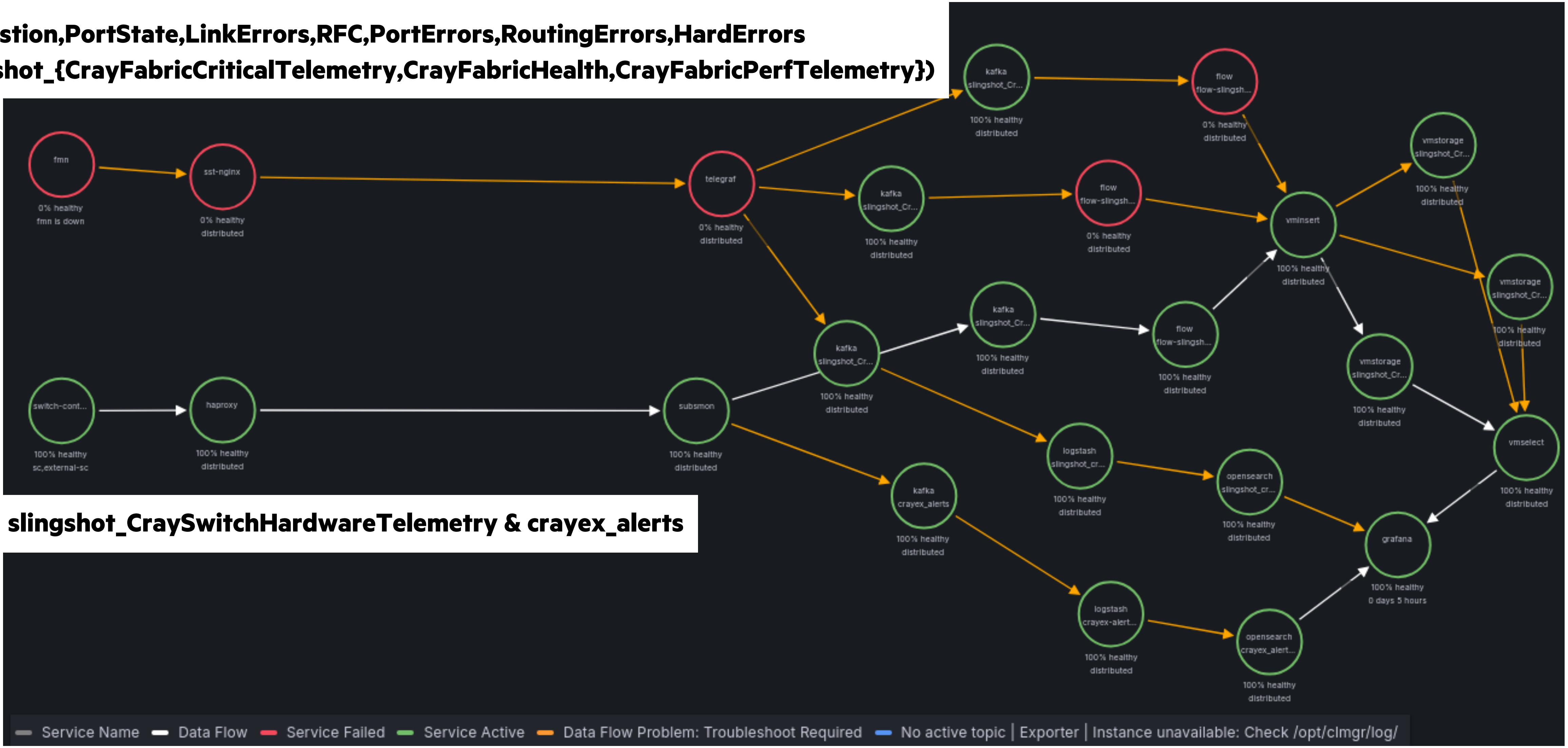
47

Silenced Alerts

0

Slingshot Monitoring Pipeline Visualization

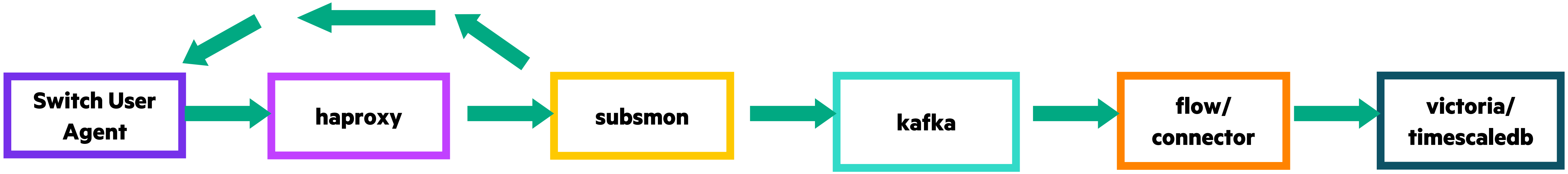
Congestion,PortState,LinkErrors,RFC,PortErrors,RoutingErrors,HardErrors (slingshot_{CrayFabricCriticalTelemetry,CrayFabricHealth,CrayFabricPerfTelemetry})



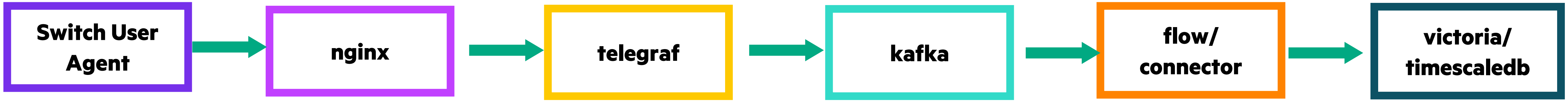
slingshot_CraySwitchHardwareTelemetry & crayex_alerts

Slingshot Monitoring

- **Individual pipelines for which the source needs to be working before even contemplating grafana**
  - 6 main different pipelines involved in slingshot - Multiple topics!
- **Switch hardware metrics such as voltages via redfish subscription (slingshot_CraySwitchHardwareTelemetry & crayex_alerts)**
- **Configured by cm monitoring setup options**



Congestion,PortState,LinkErrors,RFC,PortErrors,RoutingErrors,HardErrors (slingshot_{CrayFabricCriticalTelemetry,CrayFabricHealth,CrayFabricPerfTelemetry})



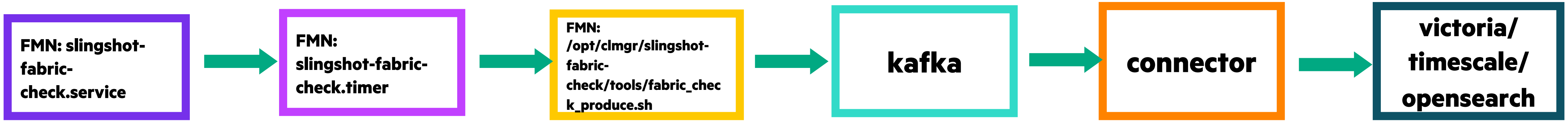
Currently these go to opensearch also!

Slingshot Monitoring

Individual pipelines for which the source to be working before even contemplating grafana and the steps in between

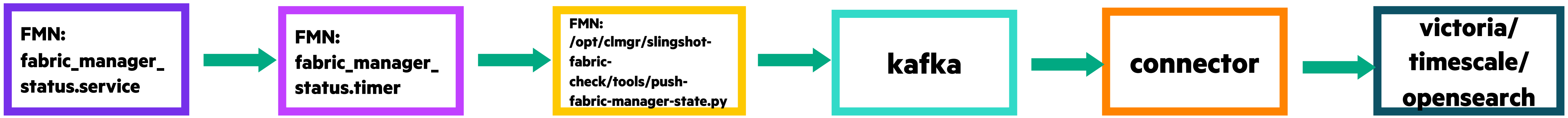
ⓘ **Varies: 1.13 and patched 1.12 use opensearch**

Slingshot - Switch performance (slingshot-perf)



FMN: slingshot-fabric-check.service → FMN: slingshot-fabric-check.timer → FMN: /opt/clmgr/slingshot-fabric-check/tools/fabric_check_produce.sh → kafka → connector → victoria/ timescale/ opensearch

Slingshot – Switches online|offline, fabric online|offline, edge online|offline (slingshot-fabric-manager-state)



FMN: fabric_manager_status.service → FMN: fabric_manager_status.timer → FMN: /opt/clmgr/slingshot-fabric-check/tools/push-fabric-manager-state.py → kafka → connector → victoria/ timescale/ opensearch

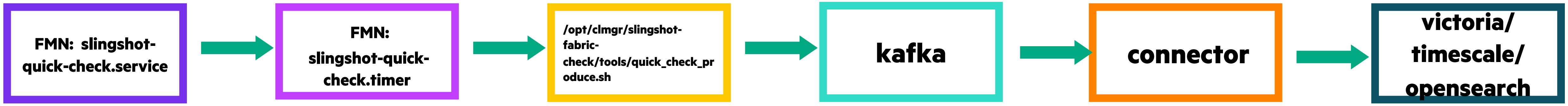
ⓘ **Service name uses underscore**

Slingshot Monitoring

Individual pipelines for which the source to be working before even contemplating grafana and the steps in between

Slingshot – slingshot-switch-state-live (a more frequent check on a subset of metrics)

ⓘ **Varies: 1.13 and patched 1.12 use opensearch**



FMN: slingshot-quick-check.service

FMN: slingshot-quick-check.timer

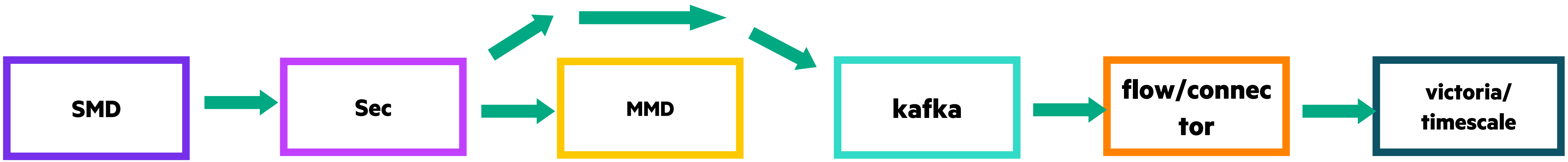
/opt/clmgr/slingshot-fabric-check/tools/quick_check_produce.sh

kafka

connector

victoria/ timescale/ opensearch

Slingshot – 200Gbps NIC (Native monitoring – so configured on previous slide with the metric group)



SMD

Sec

MMD

kafka

flow/connector

victoria/ timescale

Slingshot Monitoring

```
# cm monitoring kafka health -s T topics | head -n 6
                                       partitions
                                       |  partitions broken
                                       |  |      replication
                                       |  |      |   avro encoded
                                       |  |      |   |     msg/min
                                       |  |      |   |     |  messages
```

<1.13: cm monitoring advanced kafka health

```
# cm monitoring kafka health -s T topics | grep -i slingshot
slingshot-fabric-manager-state         1  0     2/3  -    0.0        0
slingshot-link-state                   1  0     2/3  -    0.0        0
slingshot-perf                        10  0   20/30  -    0.0        0
slingshot-switch-state                 1  0     2/3  -    0.0        0
slingshot-switch-state-live            1  0     2/3  -    0.0        0
slingshot_CrayFabricCriticalTelemetry  1  0     2/3  -  317.0  1750953
slingshot_CrayFabricHealth             1  0     2/3  -    4.0    75915
slingshot_CrayFabricPerfTelemetry      1  0     2/3  - 1500.9  8310856
slingshot_CrayFabricTelemetry          1  0     2/3  -    0.0        0
slingshot_CraySwitchHardwareTelemetry 10  0   20/30  Y  764.7  6298074
slingshot_joblevel                     1  0     2/3  -    0.0        0
slingshot_joblevel_congestion          1  0     2/3  -    0.0        0
```

Slingshot Monitoring

- **At this point we only have one dashboard for slingshot and more configuration is needed:**

```
# cm monitoring grafana list
cdu: ['crayex-cdu-monitoring']
native: ['amd-mi250x-gpu-monitoring', 'nvidia-gpu-monitoring', 'system-
monitoring']
hardware: ['crayex-rack-power', 'hardware-monitoring', 'rectifier-check']
pdu: ['sentry-pdu-monitoring']
ldms: ['ldms']
slingshot: ['alertmanager-slingshot-alerts']
```

- **Similarly with flow:**

```
# cm monitoring flow list | grep ^sling
slingshot-critical........... DISABLED
slingshot-diag-perf.......... DISABLED
slingshot-fabric-telemetry... DISABLED
slingshot-hardware........... ENABLED
slingshot-perf............... DISABLED
slingshot-switch-state....... DISABLED
slingshot-switch-state-live.. DISABLED
```

  - subsmon is running so redfish subscriptions may be established to all sCs, nCs and cCs.

Slingshot – Configuration for data sent to telegraf

- Congestion,PortState,LinkErrors,RFC,PortErrors,RoutingErrors,HardErrors

```
# cm monitoring slingshot set -c config-Mar-2025 --listener \
su-aliases.head.cm.hpc.amslabs.hpecorp.net --fmn fm01 –t max
Adding config-Mar-2025 as a new configuration with fmn as fm01 and listener as su-
aliases.head.cm.hpc.amslabs.hpecorp.net to Slingshot Telemetry Configuration.
```

- Important options:

```
-pr PERIODICITY, --periodicity PERIODICITY
                       Enter the periodicity value for the telemetry collection. A value of '0' will
disable telemetry collection.(Default:60)
                       For Slingshot 2.2 onwards, periodicity sets HeartBeat periodicity.
 -t TELEMETRY, --telemetry TELEMETRY
                       Enter the telemetry collection level. A value of 'basic' gets few basic telemetry
metrics.
                       A value of 'vital' gets important telemetry metrics
                       A value of 'max' gets all possible telemetry metrics.
                       (Default will be: 'vital')
```

Slingshot – Configuration for data sent to telegraf

- The previous slide command resulted in:

```
fm01:~ # fmn-show-telemetry-config -a
{
    "/telemetry/configurations/hpcm_config": {
        "categories": {
            "CrayFabricCriticalTelemetry": {
                "HardErrors": {
                    "periodicity": 60.0
                },
                "RoutingErrors": {
                    "periodicity": 60.0
                }
            },
            "CrayFabricHealth": {
                "all": {
                    "severity": "CRITICAL"
                }
            },
            "CrayFabricPerfTelemetry": {
                "Congestion": {
                    "periodicity": 60.0
                },
                "PauseDetails": {
                    "periodicity": 60.0
                }
            }
        },
        "collector": "http://su-aliases.head.cm.hpc.amslabs.hpecorp.net:9400",
        "enable": true,
        "eventsFailureRetries": 3,
        "heartbeatEnable": true,
        "heartbeatPeriodicity": 60.0,
        "name": "hpcm_config"
    }
}
```

Slingshot – Configuration for data sent to telegraf

- Using -t max is not the max for the FMN but the max that grafana will plot (output for <=SS 2.2 else use previous command):

```
fm01:~ # fmctl get /telemetry/configurations/hpcm_config
DOCUMENT/KEY           SUBKEY                                                   VALUE
categories             CrayFabricHealth.all.severity                            CRITICAL
categories             CrayFabricPerfTelemetry.Congestion.periodicity           60
categories             CrayFabricPerfTelemetry.PauseDetails.periodicity         60
categories             CrayFabricPerfTelemetry.CongestionDetails.periodicity    60
categories             CrayFabricPerfTelemetry.RFC3635.periodicity              60
categories             CrayFabricCriticalTelemetry.HardErrors.periodicity       60
categories             CrayFabricCriticalTelemetry.RoutingErrors.periodicity    60
categories             CrayFabricCriticalTelemetry.PortErrors.periodicity       60
collector              http://su-aliases.head.cm.hpc.amslabs.hpecorp.net:9400
documentSelfLink                                                      /telemetry/configurations/hpcm_config
enable                                                                           true
eventsFailureRetries                                                             3
heartbeatEnable                                                                  true
heartbeatPeriodicity                                                             60
name                                                                             hpcm_config
```

- Slingshot produces a massive amount of data
  - Default periodicity is 60
    - Using 120 would obviously halve that
  - Note: It is recommended that the FMN uses UTC as the switch controller timezone cannot be changed from UTC

Slingshot – Configuration for data sent to telegraf

```
# cm monitoring slingshot enable
Enabled and started slingshot-heartbeat.service on admin node.

Enabled and started Slingshot Telemetry Agent on all leader nodes.

Enabled and started nginx on all leader nodes.

Enabled and started nginx_exporter on all leader nodes.
# cm monitoring flow create slingshot-critical
enable all units for flow-slingshot-critical...
# cm monitoring flow create slingshot-perf
enable all units for flow-slingshot-perf...
# cm monitoring flow create slingshot-fabric-telemetry
enable all units for flow-slingshot-fabric-telemetry...
```

Hard, port and routing errors

RFC3635, BW, pause percent

<= 2.1 LinkErrors.*

Slingshot – State info from the FMN and switch performance (1.13)

Not /var/lib/grafana/dashboards

```
# cm monitoring grafana dashboard enable slingshot
Successfully enabled the dashboards:
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_Congestion_rxPausePercent_txPausePercent.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_routing_hard_errors.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_telemetry_main.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_bandwidth.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_bit_error_rate.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_current.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_group_status.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_network_summary.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_power.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_rfc3635_ifhcin_outoctets.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_rotational.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_rxbroad_txbroadcastpkts.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_rxcongestion.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_rxmulti_txmulticastpkts.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_rxpause_txpauseframes.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_rxucast_txucastpkts.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_switch_status.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_temperature.json
/opt/clmgr/slingshot-monitoring/dashboards/slingshot_voltage.json
```

Slingshot – Configuration for data sent to telegraf

```
# cm monitoring slingshot status

Slingshot Telemetry Monitoring status: enabled

Dependent Service Status:

+--------------------------------------+----------------+-------------+
|               Service                |    Running     | Not Running |
+--------------------------------------+----------------+-------------+
|              sst-nginx               | ld03,ld01,ld02 |             |
| Slingshot Telemetry Agent (telegraf) | ld03,ld01,ld02 |             |
+--------------------------------------+----------------+-------------+

Note:
All the above dependent services must be running for Slingshot telemetry monitoring to stream the data to
kafka.
If telegraf is running and Slingshot Telemetry Monitoring is disabled, no slingshot data is collected

Additional Service Status:

+-------------------------------------------+----------+
|                  Service                  |  Status  |
+-------------------------------------------+----------+
|        Slingshot Heartbeat service        | Enabled  |
| Slingshot Job Level Congestion Monitoring | Disabled |
+-------------------------------------------+----------+
```

Slingshot – State info from the FMN and switch performance

- We now have Slingshot Monitoring Software (SMS)
- Two state related service: which fire on a timer – one more frequent
  - slingshot_manager_status
  - slingshot-quick-check

ⓘ **Service name uses underscore**

```
# cm node zypper --repos Cluster-Manager-1.13-sles15sp6-x86_64 -n 'fm*' install slingshot-fabric-check
# ssh fm01
Last login: Mon Mar 10 05:34:25 2025 from 172.23.0.1
fm01:~ # /opt/clmgr/slingshot-fabric-check/get_groups_switches.sh
Number of groups = 3

Group 0 has 16 switches
Group 1 has 16 switches
Group 2 has 1 switches
fm01:~ # vi /opt/clmgr/etc/slingshot-fabric-check.conf
fm01:~ # egrep '^GROU|^SWI' /opt/clmgr/etc/slingshot-fabric-check.conf
GROUPS="3"
SWITCHES="16"
```

Slingshot – State info from the FMN and switch performance

```
fm01:~ # systemctl enable --now slingshot-fabric-check.service fabric_manager_status.service slingshot-quick-
  check.service slingshot-fabric-check.timer slingshot-quick-check.timer
Created symlink /etc/systemd/system/default.target.wants/slingshot-fabric-check.service →
  /usr/lib/systemd/system/slingshot-fabric-check.service.
Created symlink /etc/systemd/system/default.target.wants/fabric_manager_status.service →
  /usr/lib/systemd/system/fabric_manager_status.service.
Created symlink /etc/systemd/system/default.target.wants/slingshot-quick-check.service →
  /usr/lib/systemd/system/slingshot-quick-check.service.
Created symlink /etc/systemd/system/timers.target.wants/slingshot-fabric-check.timer →
  /usr/lib/systemd/system/slingshot-fabric-check.timer.
Created symlink /etc/systemd/system/timers.target.wants/slingshot-quick-check.timer →
/usr/lib/systemd/system/slingshot-quick-check.timer.
fm01:~ # logout
Connection to fm01 closed.
```

Slingshot – State info from the FMN and switch performance (previously)

```
# curl -k -X POST -u admin:admin -H 'Content-Type: application/json' --data-binary
'@./fabric-summary.json' https://admin/grafana/api/dashboards/db
"folderUid":"","id":59,"slug":"fabric-summary","status":"success","uid":"fabric-
summary","url":"/grafana/d/fabric-summary/fabric-summary","version":1}
# curl -k -X POST -u admin:admin -H 'Content-Type: application/json' --d
ata-binary '@./group-health.json' https://admin/grafana/api/dashboards/db
{"folderUid":"","id":60,"slug":"group-summary","status":"success","uid":"group-
health","url":"/grafana/d/group-health/group-summary","version":1}
# curl -k -X POST -u admin:admin -H 'Content-Type: application/json' --data-binary
'@./switch-overview.json' https://admin/grafana/api/dashboards/db
{"folderUid":"","id":61,"slug":"switch-summary","status":"success","uid":"switch-
overview","url":"/grafana/d/switch-overview/switch-summary","version":1}
# curl -k -X POST -u admin:admin -H 'Content-Type: application/json' --data-binary
'@./switch-performance.json' https://admin/grafana/api/dashboards/db
{"folderUid":"","id":62,"slug":"switch-low-level-performance-
metrics","status":"success","uid":"switch-performance","url":"/grafana/d/switch-
performance/switch-low-level-performance-metrics","version":1}
```

Dashboards still use timescale so would need timescale and connectors enabling

Slingshot – State info from the FMN and switch performance (previously)

```
# cm monitoring timescaledb enable
# cm monitoring timescaledb start
# cm monitoring timescaledb node add --data-node -n 'ld*'
# cm monitoring connect enable --name tsdb-slingshot-diag-perf
# systemctl enable --now confluent-kafka-connect
# cm monitoring connect enable --name tsdb-slingshot-diag-perf
# cm monitoring connect enable --name tsdb-slingshot-fabric-check
# cm monitoring timescaledb compression --category slingshot --interval 1d
# cm monitoring timescaledb retention --category slingshot --interval 7d
```

Setting retention and especially compression is very important

Slingshot – State info from the FMN and switch performance

- Some dashboards are not in the slingshot folder nor are linked from the Slingshot Monitoring Main Dashboard: Home > Dashboards > Switch Low Level Performance Metrics
- Some panels on the dashboards require the following command before viewing e.g. Slingshot Network Summary

```
# cm health report slingshot refresh
Detailed logs are available here: /var/clustertest/logs/portstate-collector-20250310-1106.log
Recreating slingshot_portstate index...
Recreating slingshot_switchstate index...
INFO: fabric switch credentials is not provided and hence using default credentials (root/initial0)
Getting fabric template from fmn...
Processing fabric template...
Getting cable info...
Getting switch port info...
Getting switch status info...
Populating slingshot_portstate index...
Populating slingshot_switchstate index...
Checking for errors/warnings...
DB - slingshot_portstate got refreshed
DB - slingshot_switchstate got refreshed
Done !
```

Needed before using some of SMS app pages also!

slingshot_CraySwitchHardwareTelemetry – Troubleshooting

- Ping the switch controller?
- List subscriptions:

```
rest_agent_tool -b x3000c0r37b0 -u EventService/Subscriptions
```

- Check destination of those listed:

```
rest_agent_tool -b x3000c0r37b0 -u EventService/Subscriptions/X
```

`systemctl status hmpad` on the switch

```
systemctl status haproxy
/var/log/messages
pdsh -g su-leader systemctl status subsmon-worker@* | grep Active -B 2
/opt/clmgr/log/subsmon-*.log
journalctl -u subsmon-worker@*
```

`curl -s admin:11890/metrics | grep subs` (worker ports 11890-5)

slingshot_CraySwitchHardwareTelemetry – Troubleshooting

```
kafka-avro-console-consumer --bootstrap-server admin:9092 --topic slingshot_CraySwitchHardwareTelemetry --max-messages=1|jq
```

```
cm monitoring kafka status [-v]
```

`/var/log/kafka/` and `/var/log/confluent/`

- Kafka topics from Redfish subscription:

```
slingshot_CraySwitchHardwareTelemetry: CrayTelemetry.Power, CrayTelemetry.Voltage,
CrayTelemetry.Current,CrayTelemetry.Temperature
```

```
crayex_alerts: CrayAlerts.1.0.HsnLinkDownDetected, CrayAlerts.1.0.HsnLinkUpDetected,
CrayAlerts.1.0.HsnLinkFlapDetected, CrayAlerts.1.0.HsnLinkErrorDetected,
CrayAlerts.1.0.HsnTransceiverInstalled, CrayAlerts.1.0.ResourcePowerStateChanged
```

slingshot_CraySwitchHardwareTelemetry – Troubleshooting

- Previously both Timescale and OpenSearch have been used for this
  - Now flow/VM

```
cm monitoring flow status

cm telemetry query -n <switch> -d 1m
current_asic_voltageregulator_input|current_asic_voltageregulator_output|current_powersupplysubsystem_voltageregulator_input|power_asic_voltageregulator_input|power_asic_voltageregulator_output|power_powersupplysubsystem_voltageregulator_input|temperature_asic_voltageregulator|voltage_asic|voltage_asic_voltageregulator_input|voltage_asic_voltageregulator_output|voltage_networkingdevice|voltage_powersupplysubsystem_voltageregulator_input|voltage_powersupplysubsystem_voltageregulator_output
```

Congestion,PortState,LinkErrors,RFC,PortErrors,RoutingErrors,HardErrors

- HTTP streaming from the switches
- HMSCollector setting on FMN as configured by

  `cm monitoring slingshot config`: SU leader alias or admin port 9400 for nginx
- FM: `fmctl get /telemetry/configurations/hpcm_config` (name of config used in previous command)
- FM: `systemctl status fabric-manager.service`
- `systemctl status sst-nginx and telegraf` (if leader nodes are in use those will run there)

Congestion,PortState,LinkErrors,RFC,PortErrors,RoutingErrors,HardErrors

```
kafka-console-consumer --bootstrap-server admin:9092 --topic
slingshot_CrayFabricTelemetry --max-messages=1|jq
```

```
cm monitoring kafka status –v
```

```
/var/log/kafka/ and /var/log/confluent/
```

- Now moved to flow/victoria:

```
cm monitoring flow status
```

```
cm telemetry query -n x1000c1r3 -d 1m porterrors_pcs_corrected_cw_local
timestamp           devicespecificcontext
|                   |     index
|                   |     |  location       messagecategory        parentalindex
|                   |     |  |              |                      |  physicalcontext       sstvalue
|                   |     |  |              |                      |  |                     |  subindex
|                   |     |  |              |                      |  |                     |  |  switch
 values
2025-03-10T12:32:17 local 5  x1000c1r3j10p0 CrayFabricCriticalTel> 0  PortErrors.pcs_corre> 0  12
x1000c1r3  164571946
2025-03-10T12:32:17 local 5  x1000c1r3j10p1 CrayFabricCriticalTel> 0  PortErrors.pcs_corre> 0  13
x1000c1r3  324270785
2025-03-10T12:32:17 local 5  x1000c1r3j11p0 CrayFabricCriticalTel> 0  PortErrors.pcs_corre> 0  15
x1000c1r3  550423150
2025-03-10T12:32:17 local 5  x1000c1r3j11p1 CrayFabricCriticalTel> 0  PortErrors.pcs_corre> 0  14 x1000c1r3
1414061261
```

Slingshot – Switch performance – Troubleshooting

```
systemctl status slingshot-fabric-check.service
```

```
systemctl status slingshot-fabric-check.timer
```

```
/opt/clmgr/slingshot-fabric-check/logdir/
```

```
kafka-console-consumer --bootstrap-server admin:9092 --topic slingshot-perf --max-messages=1
```

```
 IfInOctets, IfOutOctets, cfrx_rx_pause_pfc_cycles_00, cfrx_rx_pause_pfc_cycles_01,
 cftx_tx_pause_pfc_cycles_00, cftx_tx_pause_pfc_cycles_01, ifct_not_blocked_a,
 ifct_blocked_for_egress_fe_a, ifct_blocked_for_congestion_a, ifct_blocked_for_bandwidth_a,
 ifct_blocked_for_upstream_fe_a, ifct_blocked_for_incast_a, ifct_blocked_until_empty_a,
 ifct_blocked_other_reason, frf_empty_route_cntr, pcs_corrected_cw, pcs_uncorrected_cw,
 ifct_discard_acks_a, ifct_error_acks_a, ifct_flow_timeouts, ofct_flow_timeouts, llr_tx_replay_event,
 llr_rx_replay_event, llr_tx_poisoned_lossless, llr_rx_poisoned_lossless, ifct_over_injection_limit_a,
 ifct_blocking_over_il_a, ifct_blocking_for_redirect_a, ifct_redirect_acks_below_ecat_a,
 ifct_redirect_acks_above_ecat_a, ibuf_ibuf_full, frf_empty_route_uf_cntr, frf_empty_route_edge_cntr,
 cfrx_rx_pause_pfc_cycles_06, cfrx_rx_pause_pfc_cycles_07, cftx_tx_pause_pfc_cycles_06,
 cftx_tx_pause_pfc_cycles_07, ofct_cycles_n_flows_allocated_0, ibuf_ifct_disc, ifct_hdr_always_abort,
 ifct_intr
```

```
cm monitoring kafka status -v
```

```
/var/log/kafka/ and /var/log/confluent/
```

Slingshot – Kafka topics through http streaming from switches

- Dashboards use timescaledb still but the data is also in VictoriaMetrics

`confluent-kafka-connect` runs on the admin and leaders and is needed to get data into timescaledb

```
/var/log/kafka/connect.log
```

```
cm monitoring connect status
```

Slingshot – Switches online|offline, fabric online|offline, edge online|offline

```
systemctl status fabric_manager_status.service
```

```
systemctl status fabric_manager_status.timer
```

```
/opt/clmgr/slingshot-fabric-check/logdir/
```

```
kafka-console-consumer --bootstrap-server admin:9092 --topic slingshot-fabric-manager-state --max-messages=1
```

```
cm monitoring kafka status -v
```

```
/var/log/kafka/ and /var/log/confluent/
```

- `confluent-kafka-connect` runs on the admin and leaders and is needed to get data into timescaledb

```
/var/log/kafka/connect.log
```

```
cm monitoring connect status | jq
```

Slingshot – slingshot-switch-state-live (a more frequent check on a subset of metrics)

```
systemctl status  slingshot-quick-check.service
systemctl status  slingshot-quick-check.timer
```

`/opt/clmgr/slingshot-fabric-check/logdir/`

```
kafka-console-consumer --bootstrap-server admin:9092 --topic slingshot-switch-state-live –max-messages=1
cm monitoring kafka status -v
```

`/var/log/kafka/` and `/var/log/confluent/`

- `confluent-kafka-connect` service runs on the admin and leaders and is needed to get data into timescaledb

`/var/log/kafka/connect.log`

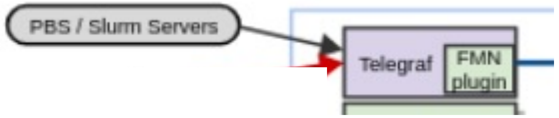
```
cm monitoring connect status | jq
```

Workload manager

Consider: Slurm or PBS

- We will cover Slurm as that is the most common
- We will not cover installing and setting up Slurm
- Basic Slurm installation and configuration can be done by following the step by step guide in `cm wlm help slurm`

>=1.10 with patch 11796



Both WLMs cover 2 aspects:

- Operations
- Power

Workload manager

1.11 now has the following but we will discuss the manual config this sets up:

```
# cm monitoring slurm --help
usage: cm monitoring slurm [-h] {disable,enable,status,update} ...
positional arguments:
 {disable,enable,status,update}
   disable             Disable Slurm Monitoring
   enable              Enable Slurm Monitoring
   status              Show status of Slurm Monitoring
   update              Update configuration of Slurm Monitoring
options:
 -h, --help            show this help message and exit
```

Workload manager

```
# cm monitoring slurm status
Dependent Service Status:

Service                     Status
------------------------------
slurmctld                   Running
telegraf                    Not Running
grafana                     Running
hpe-telegraf rpm            Installed
telegraf rpm                Installed
wlm-monitoring rpm          Installed

Slurm Monitoring is Disabled.

Slurm Configuration:
Slurm Server Node = localhost
Install Path = /usr/bin
Timeout Value in seconds = 20
Polling interval Value in seconds = 60
```

ⓘ **--server-name to specify the slurm server**

Slurm operations configuration

- Needs hpe-telegraf and the telegraf RPMs installed

```
cm monitoring slurm enable [--install-path path] [--timeout response] [--interval polling] [--server-name slurm_host]
```

- This will:

```
slurm# cp /etc/telegraf/telegraf.d/slurm.disable /etc/telegraf/telegraf.d/slurm.conf
```

- --install-path path is the path to sinfo/scontrol which is in that file

```
commands = ["/opt/clmgr/bin/cm-python3 /opt/clustertest/bin/chc_slurm_mon.py
/usr/bin"]
```

Previously: **`cm monitoring connect enable --name tsdb-slurm`**

- Now: `cm monitoring flow create slurm`

Slurm operations configuration

- Ensure the interval and timeout settings are sufficient in `/etc/telegraf/telegraf.d/slurm.conf` or `pbs.conf`
- PBS does not have a "cm monitoring" command line
  - Interval period should always be greater than the timeout
- The script name for PBS is `chc_pbs_mon.py`
- The path on >=1.11 is `/opt/clmgr/wlm-mon/bin`

```
slurm# time /opt/clustertest/bin/chc_slurm_mon.py /usr/bin
```

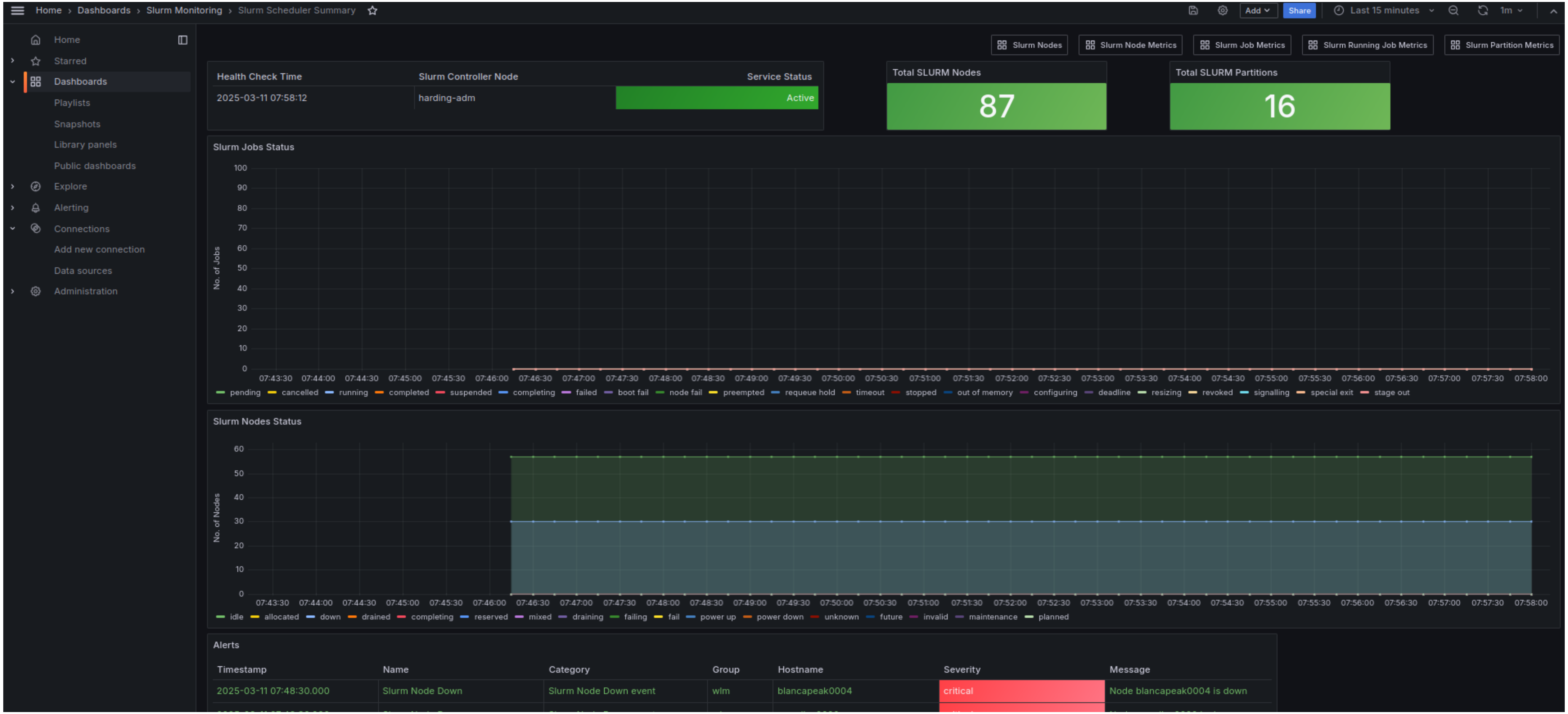
Do this during load and multiple times at different times to choose an appropriate value.

Slurm operations configuration

- Slurm on the admin node example:

```
# systemctl enable --now telegraf.service
Created symlink /etc/systemd/system/multi-user.target.wants/telegraf.service →
/usr/lib/systemd/system/telegraf.service.
# cm monitoring slurm enable
Dependent Service Status:

Service                    Status
------------------------------
slurmctld                  Running
telegraf                   Running
grafana                    Running
hpe-telegraf rpm           Installed
telegraf rpm               Installed
wlm-monitoring rpm         Installed

Slurm Monitoring is enabled.
Enabled the slurm dashboards successfully to Grafana.
Enabled the Slurm Alert Rules successfully.
# cm monitoring flow create slurm
enable all units for flow-slurm...
```

Slurm operations configuration – Slurm Scheduler Summary



Health Check Time	Slurm Controller Node	Service Status
2025-03-11 07:58:12	harding-adm	Active

Total SLURM Nodes

87

Total SLURM Partitions

16

Slurm Jobs Status

Slurm Nodes Status

Alerts

Timestamp	Name	Category	Group	Hostname	Severity	Message
2025-03-11 07:48:30.000	Slurm Node Down	Slurm Node Down event	wlm	blancapeak0004	critical	Node blancapeak0004 is down

Slurm power monitoring configuration

```
# egrep -v '^#|^    #|^  #' /opt/clmgr/wlm-mon/conf/wlm-mon.yml
slurm:
 power_api_timeout: 20

 node_power_report:
   jobmonitor_power_module: True
   clmgr_power_module: False

 job_power_report:
   pm_counters_plugin: False
   spank_plugin: True
   spank_plugin_type:
     jobmonitor_power_module: True
     clmgr_power_module: False

 job_energy_overhead_report:
   conversion_ratio: 1.0
   static_power: 0
   total_nodes: 0

pbs:
 job_monitor_module: False
 power_api_timeout: 20
```

ⓘ **jobmonitor, clmgr-power or pm_counters as sources dependent on hardware and software**

Slurm power monitoring configuration – jobmonitor (>1.10 and patch)

- On EX and clusters that do not have SU leaders (but not HPE Apollo 9000) for job-level energy and power consumption monitoring:
  - use the jobmonitor service
- EX: also supports the USS pm_counters (bpmcdmod kernel module)

```
# systemctl enable --now jobmonitor.service
Created symlink /etc/systemd/system/multi-
user.target.wants/jobmonitor.service →
/usr/lib/systemd/system/jobmonitor.service.
Created symlink /etc/systemd/system/cmdb.service.wants/jobmonitor.service →
/usr/lib/systemd/system/jobmonitor.service.

slurm# grep ^Pl /etc/slurm/slurm.conf
PlugStackConfig=/opt/clmgr/wlm-mon/conf/jobmonitor_plugstack.conf
```

ⓘ **<1.11 path is /opt/clmgr/etc**

Slurm power monitoring configuration – jobmonitor (>1.10 and patch)

```
# rpm -q slurm-slurmctld
slurm-slurmctld-24.05.4-20241127112017_fec88a311c6c.x86_64
# vi /etc/slurm/slurm.conf
# grep ^Pl /etc/slurm/slurm.conf
PlugStackConfig=/opt/clmgr/wlm-mon/conf/jobmonitor_plugstack.conf
# vi /opt/clmgr/wlm-mon/conf/jobmonitor_plugstack.conf
# grep -v ^# /opt/clmgr/wlm-mon/conf/jobmonitor_plugstack.conf
optional /opt/clmgr/wlm-mon/lib/jobmonitor_slurm-24.05_power_plugin.so 172.23.0.1 4442
# systemctl restart slurmctld

# pdsh -g compute systemctl restart slurmd.service
```

ⓘ **<1.11 path is /opt/clustertest/lib**

Slurm power monitoring configuration – power API

```
# rpm -q slurm-slurmctld
slurm-slurmctld-24.05.4-20241127112017_fec88a311c6c.x86_64
# vi /etc/slurm/slurm.conf
# grep ^Pl /etc/slurm/slurm.conf
PlugStackConfig=/opt/clmgr/wlm-mon/conf/plugstack.conf
# vi /opt/clmgr/wlm-mon/conf/plugstack.conf
# grep -v ^# /opt/clmgr/wlm-mon/conf/plugstack.conf
optional /opt/clmgr/wlm-mon/lib/chc_slurm-24.05_power_plugin.so
172.23.0.1 8888
# systemctl restart slurmctld

# pdsh -g compute systemctl restart slurmd.service
```

ⓘ **<1.11 path is /opt/clustertest/lib and /opt/clmgr/etc**

Slurm power monitoring configuration – power API (clmgr-power)

```
# egrep -v '^#|^    #|^  #' /opt/clmgr/wlm-mon/conf/wlm-mon.yml | grep -A 4
"spank_plugin:"
   spank_plugin: True

   spank_plugin_type:
     jobmonitor_power_module: False
     clmgr_power_module: True
# vi /opt/clmgr/etc/clmgr-power.conf
# grep ^node_power_monitoring_active /opt/clmgr/etc/clmgr-power.conf
node_power_monitoring_active = True
# systemctl restart clmgr-power
```

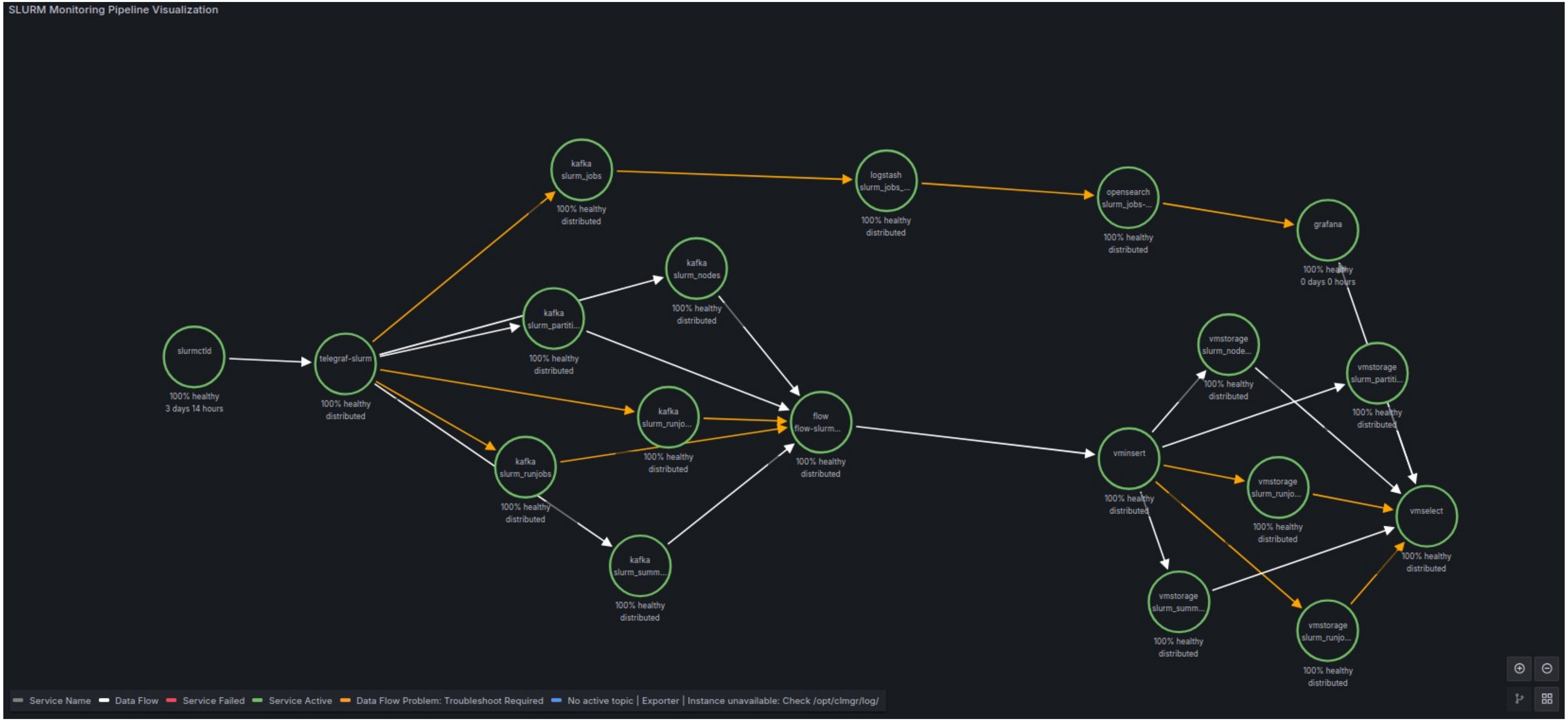
<1.11 path is /opt/clmgr/etc

Older versions may not have this file as the only possibility was to use the power service

Slingshot/Slurm job level congestion

```
# cm health report slingshot refresh
Detailed logs are available here: /var/clustertest/logs/portstate-collector-20250401-0547.log
Recreating slingshot_portstate index...
Recreating slingshot_switchstate index...
INFO: fabric switch credentials is not provided and hence using default credentials
(root/initial0)
Getting fabric template from fmn...
Processing fabric template...
Getting cable info...
Getting switch port info...
Getting switch status info...
Populating slingshot_portstate index...
Populating slingshot_switchstate index...
Checking for errors/warnings...
DB - slingshot_portstate got refreshed
DB - slingshot_switchstate got refreshed
Done !
# cm monitoring slingshot enable --name joblevel-congestion
Slingshot Joblevel Congestion Monitoring and it's related dashboards have been enabled.
```

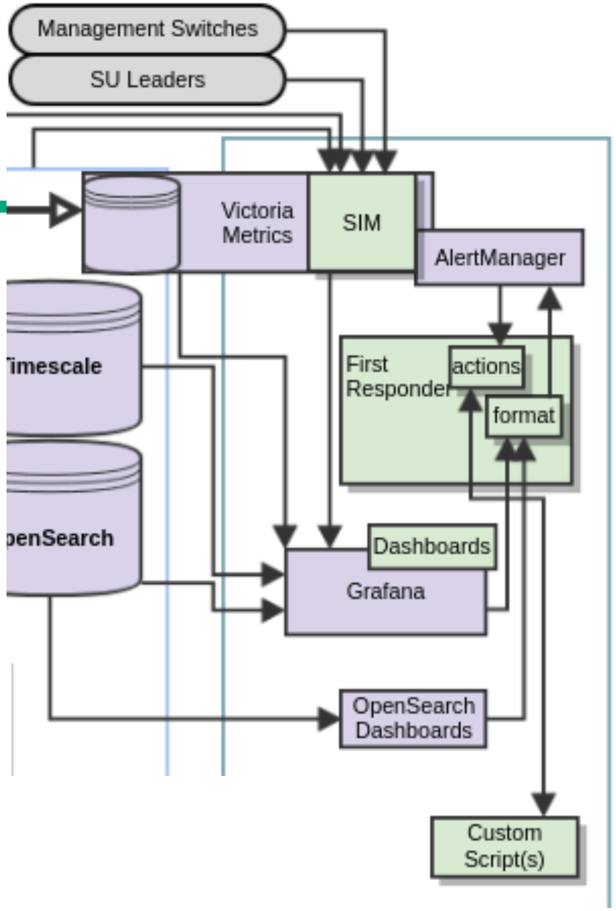
Slurm MPVT


SLURM Monitoring Pipeline Visualization
slurmctld
100% healthy
3 days 14 hours
telegraf-slurm
100% healthy
distributed
kafka
slurm_jobs
100% healthy
distributed
kafka
slurm_partiti...
100% healthy
distributed
kafka
slurm_nodes
100% healthy
distributed
kafka
slurm_runjo...
100% healthy
distributed
kafka
slurm_runjobs
100% healthy
distributed
kafka
slurm_summ...
100% healthy
distributed
flow
flow-slurm...
100% healthy
distributed
logstash
slurm_jobs_...
100% healthy
distributed
opensearch
slurm_jobs-...
100% healthy
distributed
grafana
100% healthy
0 days 0 hours
vminsert
100% healthy
distributed
vmstorage
slurm_node...
100% healthy
distributed
vmstorage
slurm_partiti...
100% healthy
distributed
vmstorage
slurm_runjo...
100% healthy
distributed
vmstorage
slurm_summ...
100% healthy
distributed
vmstorage
slurm_runjo...
100% healthy
distributed
vmselect
100% healthy
distributed
Service Name
Data Flow
Service Failed
Service Active
Data Flow Problem: Troubleshoot Required
No active topic | Exporter | Instance unavailable: Check /opt/clmgr/log/

The slurm monitoring pipeline

```
kafka-console-consumer --bootstrap-server admin:9092 --topic slurm_jobs --max-messages=1 | jq
cm monitoring kafka status -v
/var/log/kafka/ and /var/log/confluent/
systemctl status fluent-bit
journalctl –xeu fluent-bit
cm monitoring opensearch status -v
curl -s http://admin:9200/_cat/indices
cm monitoring victoria status –v
cm telemetry list -a | grep ^slurm and then query
```

Alerting, SIM and rackmap



Service Infrastructure Monitoring is monitoring of the monitoring (and other infrastructure) and best practice

Alerting, SIM and rackmap

cm monitoring alerting enable (1.10 with patch and 1.1{1,2})

cm monitoring setup with 1.13

cm monitoring alerting status

cm monitoring alerting opensearch or grafana --enable-rule <appropriate rules>

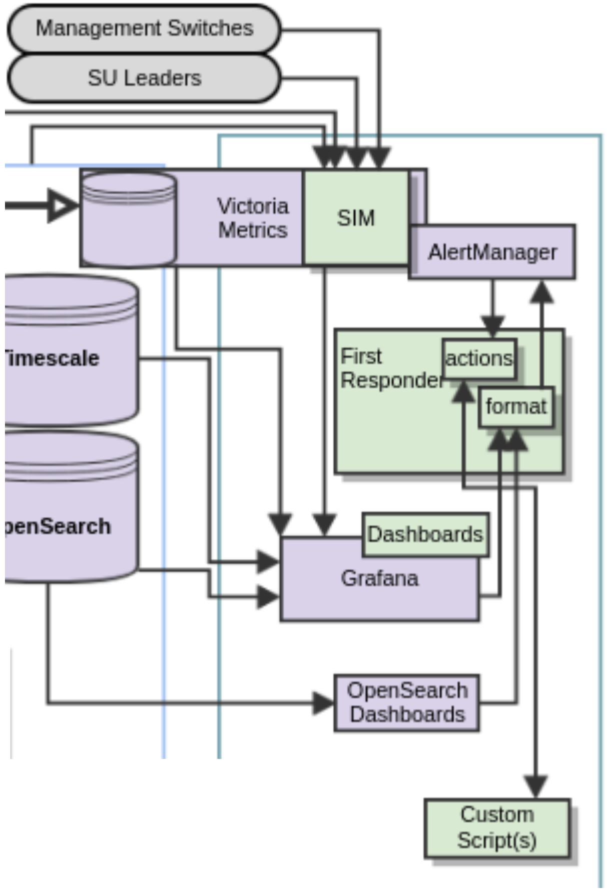
cm monitoring alerting route email --from <email> --to <email> --smtp <smtp.server:25> --alert-group <group>

<1.13: cm sim enable **and** start **and** add {--service-group monitoring-services|suleader-services}

>=1.11: cm monitoring rackmap map component-drift **or** power **or** cpu-temperature **or** slingshot-switch-status -l

Alerts

- Dependent on everything which has gone before
- Unified alerting using alertman in >=1.10
- `cm heath alertman` is a single interface to
  - OpenSearch data
  - Grafana (Timescale data (removed in 1.13) but still used for VM telemetry)
  - Victoria Metrics (vmalert)
- Alertmanager is about live alerts but is persisted to opensearch



Unified Alerting status

```
# cm monitoring alerting status
Alerting is enabled.

Dependent services status:
               service status
               |            api status
opensearch      OK           OK (200)
grafana-server  OK           OK (200)
alertmanager    OK           OK (200)
first_responder OK           OK (200)

Default Alert Rules Status:
                                   datasource
                                   |          rule-engine
                                   |          |          state      alert-group      notification
CDU Alert                           OS         opensearch Enabled    cooldev          None
IML Alert                           OS         opensearch Enabled    iml              None
Rasdaemon Alert                     OS         opensearch Enabled    rasdaemon        None
Syslog Alert                        OS         opensearch Enabled    syslog           None
Hardware Alert                      OS         opensearch Enabled    redfish,link     None
Native Alert                        OS         opensearch Enabled    native           None
Node Down                           OS         opensearch Enabled    node_status      None
Node UP                             OS         opensearch Enabled    node_status      None
Slingshot Cassini Alert             OS         opensearch Enabled    fabric-nic       None
Slingshot Cassini Up Alert          OS         opensearch Enabled    fabric-nic       None
Slingshot Health Alert              OS         opensearch Enabled    link,switch,port None
Slingshot Rosetta Alert             OS         opensearch Enabled    switch           None
Slingshot Joblevel Monitoring Alert OS         opensearch Enabled    joblevel         None
Slingshot Switch Down               OS         opensearch Enabled    switch           None
Slingshot Switch Up                 OS         opensearch Enabled    switch           None
```

Unified Alerting status continued

```
Extra Alert Rules Status:
                                                   datasource
                                                   |          rule-engine
                                                   |          |       state      alert-group
                                                   |          |       |          |          notification
Primary Facility Flow Critical                      VM         grafana Disabled   cooldev    None
Primary Facility Flow Warning                       VM         grafana Disabled   cooldev    None
Primary Facility Return Water Pressure Critical     VM         grafana Disabled   cooldev    None
Primary Facility Return Water Pressure Warning      VM         grafana Disabled   cooldev    None
Primary Facility Supply Water Pressure Critical     VM         grafana Disabled   cooldev    None
Primary Facility Supply Water Pressure Warning      VM         grafana Disabled   cooldev    None
Secondary Cabinet Return Water Pressure Critical    VM         grafana Disabled   cooldev    None
Secondary Cabinet Return Water Pressure Warning     VM         grafana Disabled   cooldev    None
Secondary Cabinet Supply Water Pressure Critical    VM         grafana Disabled   cooldev    None
Secondary Cabinet Supply Water Pressure Warning     VM         grafana Disabled   cooldev    None
Secondary Cabinet Supply Water Temperature Critical VM         grafana Disabled   cooldev    None
Secondary System Differential Pressure Critical     VM         grafana Disabled   cooldev    None
Secondary System Differential Pressure Warning      VM         grafana Disabled   cooldev    None
Supply Filter A Differential Pressure Critical      VM         grafana Disabled   cooldev    None
Supply Filter A Differential Pressure Warning       VM         grafana Disabled   cooldev    None
Supply Filter B Differential Pressure Critical      VM         grafana Disabled   cooldev    None
Supply Filter B Differential Pressure Warning       VM         grafana Disabled   cooldev    None
Amd GPU Power                                       VM         grafana Disabled   gpu        None
Amd GPU Temperature                                 VM         grafana Disabled   gpu        None
Nvidia GPU Power                                    VM         grafana Disabled   gpu        None
Nvidia GPU Temperature                              VM         grafana Disabled   gpu        None
PDU Active Energy                                   VM         grafana Disabled   pdu        None
PDU Input Power                                     VM         grafana Disabled   pdu        None
PDU Input Power Factor                              VM         grafana Disabled   pdu        None
PDU Input VA                                        VM         grafana Disabled   pdu        None
Slingshot Congestion rxPausePercent                 VM         grafana Disabled   port       None
Slingshot Congestion txPausePercent                 VM         grafana Disabled   port       None
```

Unified Alerting status continued

```
                      datasource
                      |          rule-engine
                      |          |       state      alert-group
                      |          |       |          |          notification
connect alerts        VM         vmalert Enabled    sim        None
elasticsearch alerts  VM         vmalert Enabled    sim        None
flexfabric alerts     VM         vmalert Enabled    sim        None
kafka alerts          VM         vmalert Enabled    sim        None
node alerts           VM         vmalert Enabled    sim        None
zookeeper alerts      VM         vmalert Enabled    sim        None
Notes:
Datasource: OS - Opensearch and VM - VictoriaMetrics.
rule-engine: opensearch - logs (OpenSearch), grafana - telemetry (VictoriaMetrics), vmalert - SIM
(VictoriaMetrics).
Notification: Email - Routed the alerts for the alert group of the rule to Email.
Alert rules file path:
OS - /opt/clmgr/alerting/opensearch-alerting/alert_rules/
VM - /opt/clmgr/alerting/grafana-alerting/alert_rules/
SIM - /opt/clmgr/cm-sim-admin/rules/
Custom Alert rules file path: /opt/clmgr/alerting/grafana-alerting/custom_alert_rules/
```

Unified Alerting

```
# cm health alertman -s

+---------------+--------+
| Alert Status  | Count  |
+---------------+--------+
| Critical      | 98     |
| Warnings      | 347    |
|               |        |
|               |        |
| Active        | 445    |
+---------------+--------+

+-------------+----------+-------------------------------+
| Group       | Severity | Alerts                        |
+-------------+----------+-------------------------------+
| cooldev     | ok       | critical : 0, warning : 0     |
| fabric-nic  | ok       | critical : 0, warning : 0     |
| gpu         | ok       | critical : 0, warning : 0     |
| iml         | ok       | critical : 0, warning : 0     |
| joblevel    | ok       | critical : 0, warning : 0     |
| link        | ok       | critical : 0, warning : 0     |
| native      | ok       | critical : 0, warning : 0     |
| node_status | ok       | critical : 0, warning : 0     |
| pdu         | ok       | critical : 0, warning : 0     |
| port        | ok       | critical : 0, warning : 0     |
| rasdaemon   | ok       | critical : 0, warning : 0     |
| redfish     | critical | critical : 5, warning : 2     |
| sim         | critical | critical : 3, warning : 0     |
| syslog      | critical | critical : 60, warning : 333  |
| switch      | warning  | critical : 0, warning : 12    |
| wlm         | critical | critical : 30, warning : 0    |
| others      | ok       | critical : 0, warning : 0     |
+-------------+----------+-------------------------------+
```

Unified Alerting

```
# cm monitoring alerting route email --from no-one@org.com --to myemail@org.com --smtp
smtp.org.com:25 --alert-group iml
# ls /opt/clmgr/alerting/opensearch-alerting/alert_rules/
CDU  hw_events  IML  native_monitoring  node_status  Rasdaemon  SAMPLE.yml.reference  slingsh
ot  Syslog
# ls /opt/clmgr/alerting/grafana-alerting/alert_rules/
cooldev  gpu  pdu  slingshot  wlm
# ls /opt/clmgr/cm-sim-admin/rules/
additional_rules.yml  connect.yml  elasticsearch.yml  flexfabric.yml  kafka.yml  node.yml  zo
okeeper.yml
```

- For telemetry metrics, 1.13 has commands for creating rules rather than writing yml:
- These existing rules can be modified with `cm monitoring alerting grafana update`
  Custom rules can be created with `cm monitoring alerting grafana create`

- For opensearch, create the yml and enable the rule

Unified Alerting

```
# cm health alertman silence -h
usage: cm health alertman silence [-h] {add|query|expire} ...

positional arguments:
 add           Silence the alertmanager alerts.
 query         Query silenced alerts from alertmanager.
 expire        Expire the silenced alerts.

options:
 -h , --help   show this help message and exit.

The 'cm health alertman silence' command provides the following functionalities for managing
silences in the alertmanager:

 1. **Add a silence**: You can silence specific alerts based on matchers.
 2. **Query silenced alerts**: Retrieve a list of currently silenced alerts.
 3. **Expire silenced alerts**: Expire a specific silence and restore the alerts.

<truncated for brevity />
```

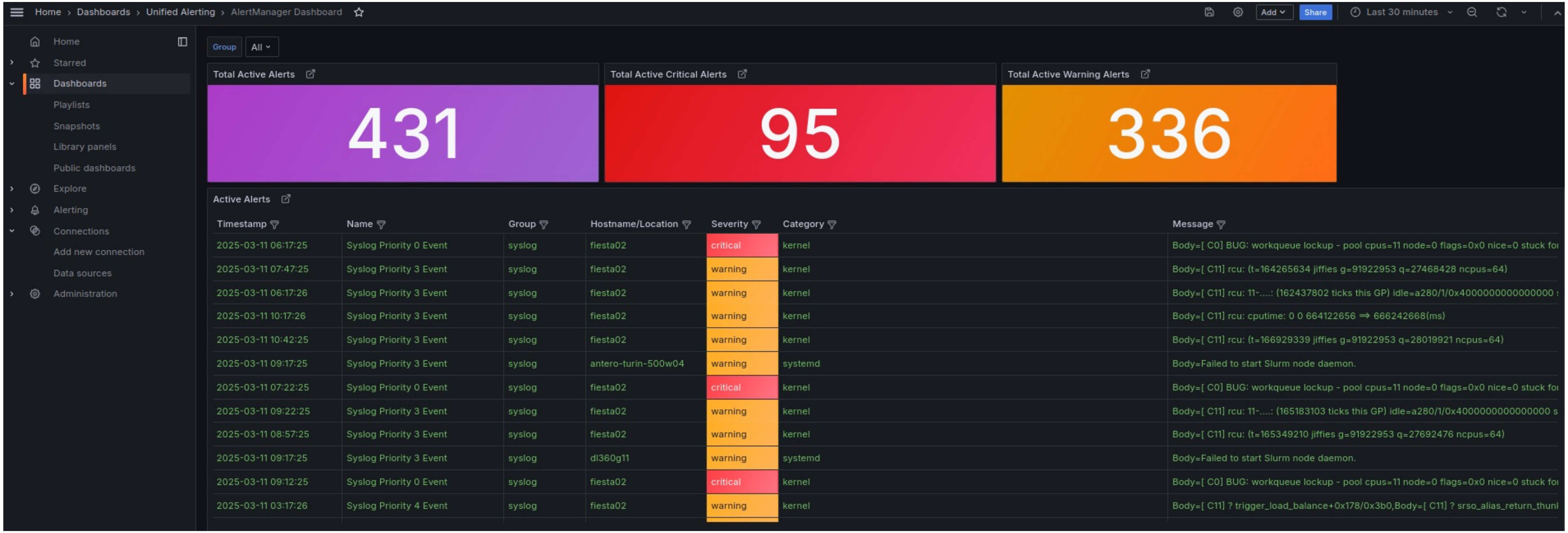
Unified Alerting – first responder actions

/opt/clmgr/first-responder/action/sample_action.yml contents:

```
# yaml-language-server: $schema=./action.schema.json
# install redhat.vscode-yaml for validation and code complete
# copy below contents and create a new yaml file and update the action name, match_labels and the operations
Action:
 name: <Action name>
 match_labels:
   <label>: <value>
 operations:
   firing:
     < command 2 >
   cleared:
     < command 3 >
---
# Step 1: Stop and Disable first_responder service
# Step2: In '/usr/lib/systemd/system/first_responder.service' check and update the following line
 # ExecStart=/opt/clmgr/bin/cm-python3 /opt/clmgr/first-responder/lib/first_responder.py --action_required
True
# Step3: Reload and restart the service with following commands
 # systemctl daemon-reload
 # systemctl enable first_responder.service
 # systemctl start first_responder.service
# Step4: Route the alerts to webhook end point in alertmanager and restart alertmanager
```

e.g.
group: cooldev
severity: critical
event_type:

Unified Alerting – AlertManager Dashboard



Total Active Alerts 431

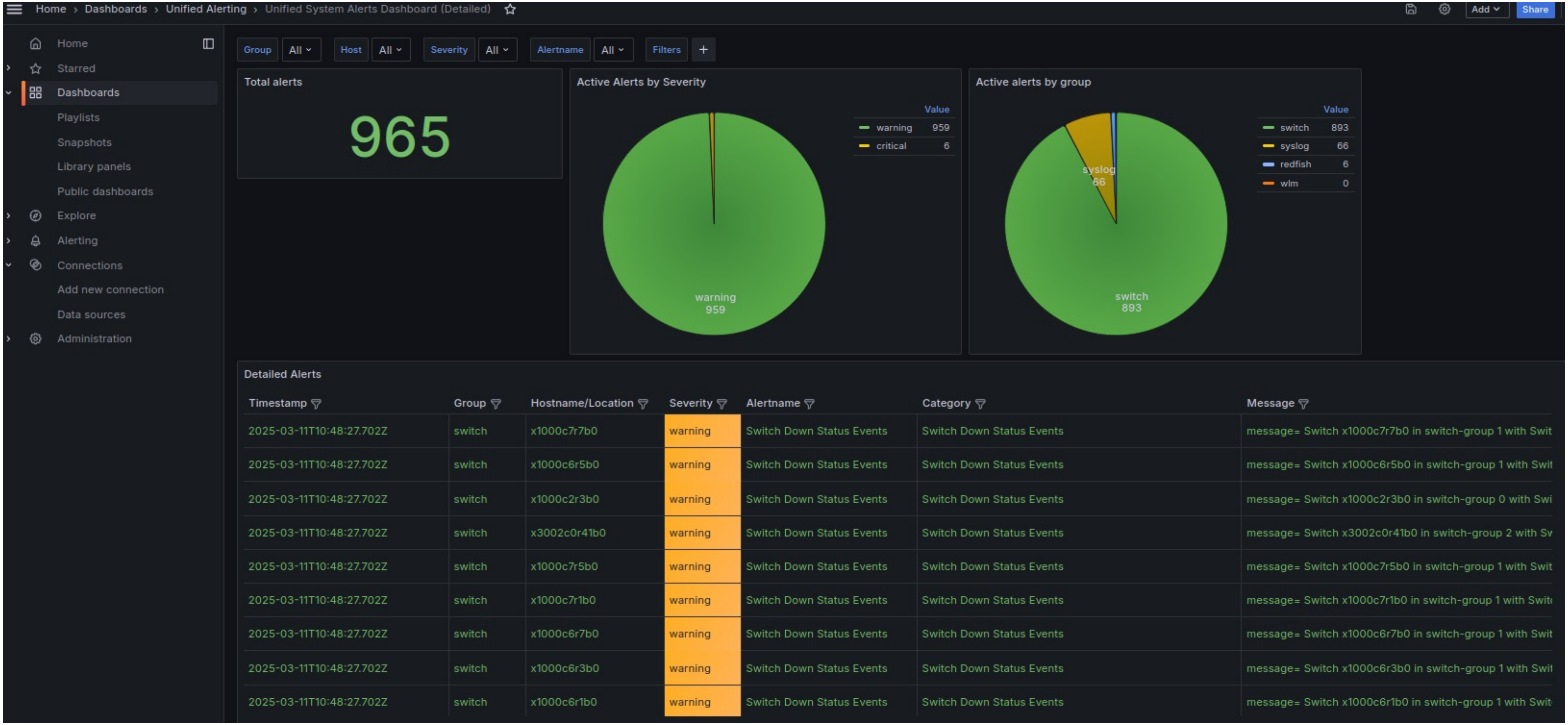
Total Active Critical Alerts 95

Total Active Warning Alerts 336

Active Alerts

Timestamp	Name	Group	Hostname/Location	Severity	Category	Message
2025-03-11 06:17:25	Syslog Priority 0 Event	syslog	fiesta02	critical	kernel	Body=[C0] BUG: workqueue lockup - pool cpus=11 node=0 flags=0x0 nice=0 stuck fo
2025-03-11 07:47:25	Syslog Priority 3 Event	syslog	fiesta02	warning	kernel	Body=[C11] rcu: (t=164265634 jiffies g=91922953 q=27468428 ncpus=64)
2025-03-11 06:17:26	Syslog Priority 3 Event	syslog	fiesta02	warning	kernel	Body=[C11] rcu: 11-....: (162437802 ticks this GP) idle=a280/1/0x4000000000000000
2025-03-11 10:17:26	Syslog Priority 3 Event	syslog	fiesta02	warning	kernel	Body=[C11] rcu: cputime: 0 0 664122656 ⇒ 666242668(ms)
2025-03-11 10:42:25	Syslog Priority 3 Event	syslog	fiesta02	warning	kernel	Body=[C11] rcu: (t=166929339 jiffies g=91922953 q=28019921 ncpus=64)
2025-03-11 09:17:25	Syslog Priority 3 Event	syslog	antero-turin-500w04	warning	systemd	Body=Failed to start Slurm node daemon.
2025-03-11 07:22:25	Syslog Priority 0 Event	syslog	fiesta02	critical	kernel	Body=[C0] BUG: workqueue lockup - pool cpus=11 node=0 flags=0x0 nice=0 stuck fo
2025-03-11 09:22:25	Syslog Priority 3 Event	syslog	fiesta02	warning	kernel	Body=[C11] rcu: 11-....: (165183103 ticks this GP) idle=a280/1/0x4000000000000000 s
2025-03-11 08:57:25	Syslog Priority 3 Event	syslog	fiesta02	warning	kernel	Body=[C11] rcu: (t=165349210 jiffies g=91922953 q=27692476 ncpus=64)
2025-03-11 09:17:25	Syslog Priority 3 Event	syslog	dl360g11	warning	systemd	Body=Failed to start Slurm node daemon.
2025-03-11 09:12:25	Syslog Priority 0 Event	syslog	fiesta02	critical	kernel	Body=[C0] BUG: workqueue lockup - pool cpus=11 node=0 flags=0x0 nice=0 stuck fo
2025-03-11 03:17:26	Syslog Priority 4 Event	syslog	fiesta02	warning	kernel	Body=[C11] ? trigger_load_balance+0x178/0x3b0,Body=[C11] ? srso_alias_return_thun

Unified Alerting – Univified System Alerts Dashboard (Detailed)



Total alerts

965

Active Alerts by Severity

	Value
warning	959
critical	6

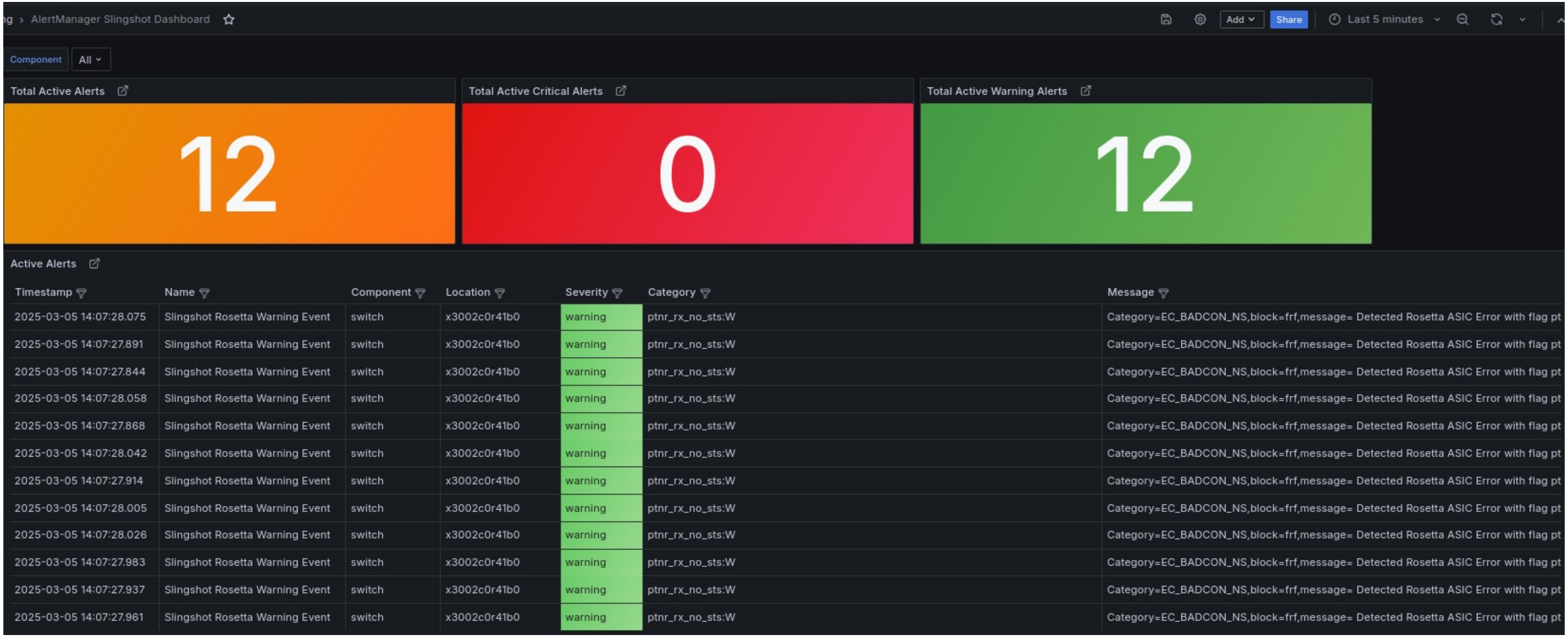
Active alerts by group

	Value
switch	893
syslog	66
redfish	6
wlm	0

Detailed Alerts

Timestamp	Group	Hostname/Location	Severity	Alertname	Category	Message
2025-03-11T10:48:27.702Z	switch	x1000c7r7b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c7r7b0 in switch-group 1 with Swit
2025-03-11T10:48:27.702Z	switch	x1000c6r5b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c6r5b0 in switch-group 1 with Swit
2025-03-11T10:48:27.702Z	switch	x1000c2r3b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c2r3b0 in switch-group 0 with Swi
2025-03-11T10:48:27.702Z	switch	x3002c0r41b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x3002c0r41b0 in switch-group 2 with Sv
2025-03-11T10:48:27.702Z	switch	x1000c7r5b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c7r5b0 in switch-group 1 with Swit
2025-03-11T10:48:27.702Z	switch	x1000c7r1b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c7r1b0 in switch-group 1 with Swit
2025-03-11T10:48:27.702Z	switch	x1000c6r7b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c6r7b0 in switch-group 1 with Swit
2025-03-11T10:48:27.702Z	switch	x1000c6r3b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c6r3b0 in switch-group 1 with Swit
2025-03-11T10:48:27.702Z	switch	x1000c6r1b0	warning	Switch Down Status Events	Switch Down Status Events	message= Switch x1000c6r1b0 in switch-group 1 with Swit

Unified Alerting – AlertManager Slingshot Dashboard



Total Active Alerts

12

Total Active Critical Alerts

0

Total Active Warning Alerts

12

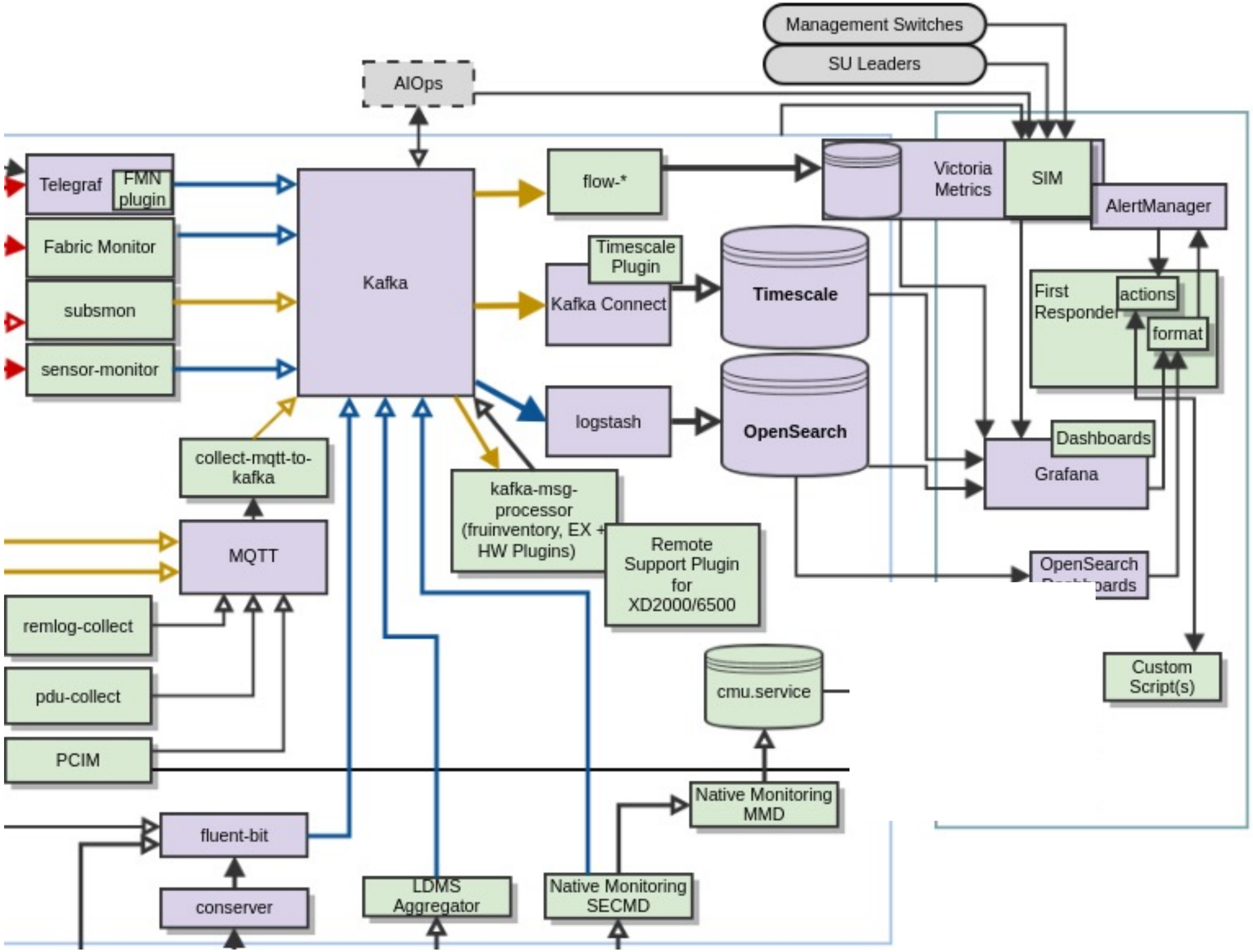
Active Alerts

Timestamp	Name	Component	Location	Severity	Category	Message
2025-03-05 14:07:28.075	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:27.891	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:27.844	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:28.058	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:27.868	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:28.042	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:27.914	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:28.005	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:28.026	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:27.983	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:27.937	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt
2025-03-05 14:07:27.961	Slingshot Rosetta Warning Event	switch	x3002c0r41b0	warning	ptnr_rx_no_sts:W	Category=EC_BADCON_NS,block=frf,message= Detected Rosetta ASIC Error with flag pt

Service Infrastructure Monitor - SIM

- The service infrastructure monitor (SIM)
  - shows information about the health of your infrastructure services
  - includes metrics, dashboards, and alerts that help you understand how HPCM core services, monitoring services and su-leader services and resources are operating
  - supports AIOPs
  - displays any alerts that arise including disk space issues if retention/compression has not been configure appropriately
    - If you do have full disks please refer to the article on the portal
- The cluster manager includes tools and commands to help you monitor the services in Grafana
- **It is best practice to enable SIM.** With 1.13 and `cm monitoring setup`, it is enabled by default Otherwise: `cm sim enable and start`

Service Infrastructure Monitor - SIM



Service Infrastructure Monitor - SIM

- By default with `cm monitoring setup`:

```
# cm sim status | head
Running is-active for vmagent service : vmagent.service

admin: active

Running is-active for vmalert service : vmalert.service

admin: active

Running is-active for alertmanager service : alertmanager.service

# cm sim status | tail -n 1
Services enabled in SIM are: core-services | monitoring-services | suleader-services
```

- Exporters are systemctl services/timers running on nodes such as the leaders:

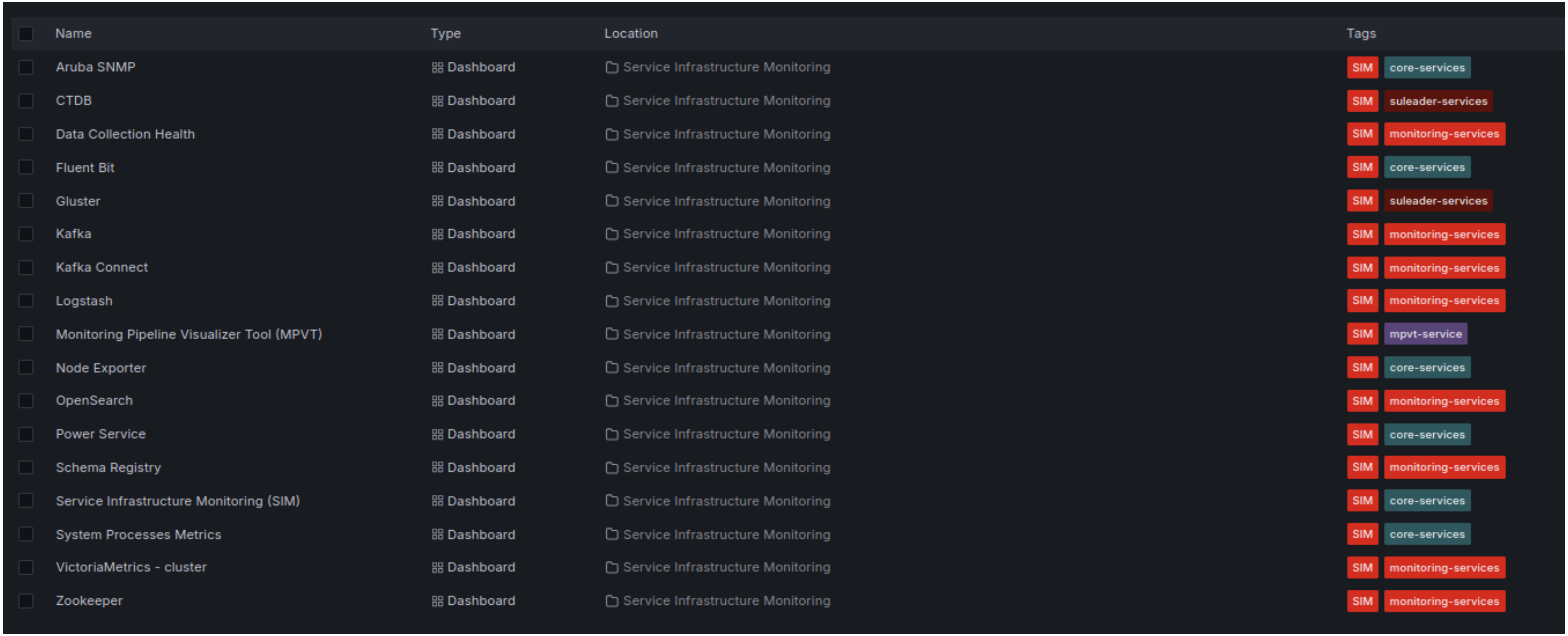
```
Running is-active for core-services service: disk_exporter.timer

ld01: active
ld02: active
ld03: active
```

Service Infrastructure Monitor - SIM

- Add MPVT (monitoring pipe visualisation tool) and AIOps (artificial intelligence for operations) if required:

```
# cm sim add --service-group mpvt-service
Running enable for mpvt-service service: kafka_exporter.service
Running enable for mpvt-service service: mqtt_exporter.service
Running enable for mpvt-service service: mosquitto_exporter.service
Running enable for mpvt-service service: logstash_exporter.service
Running enable for mpvt-service service: haproxy_exporter.service
Running start for mpvt-service service: kafka_exporter.service
Running start for mpvt-service service: mqtt_exporter.service
Running start for mpvt-service service: mosquitto_exporter.service
Running start for mpvt-service service: logstash_exporter.service
Running start for mpvt-service service: haproxy_exporter.service
Updating mpvt configuration file
Running enable for mpvt service : mpvt.service
Running start for mpvt service : mpvt.service
Adding respective dashboard in to grafana
Services enabled in SIM are: core-services | monitoring-services | suleader-services | mpvt-service
```

Service Infrastructure Monitor – SIM Dashboards



Name	Type	Location	Tags
Aruba SNMP	Dashboard	Service Infrastructure Monitoring	SIM core-services
CTDB	Dashboard	Service Infrastructure Monitoring	SIM suleader-services
Data Collection Health	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services
Fluent Bit	Dashboard	Service Infrastructure Monitoring	SIM core-services
Gluster	Dashboard	Service Infrastructure Monitoring	SIM suleader-services
Kafka	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services
Kafka Connect	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services
Logstash	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services
Monitoring Pipeline Visualizer Tool (MPVT)	Dashboard	Service Infrastructure Monitoring	SIM mpvt-service
Node Exporter	Dashboard	Service Infrastructure Monitoring	SIM core-services
OpenSearch	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services
Power Service	Dashboard	Service Infrastructure Monitoring	SIM core-services
Schema Registry	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services
Service Infrastructure Monitoring (SIM)	Dashboard	Service Infrastructure Monitoring	SIM core-services
System Processes Metrics	Dashboard	Service Infrastructure Monitoring	SIM core-services
VictoriaMetrics - cluster	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services
Zookeeper	Dashboard	Service Infrastructure Monitoring	SIM monitoring-services

cm monitoring rackmap

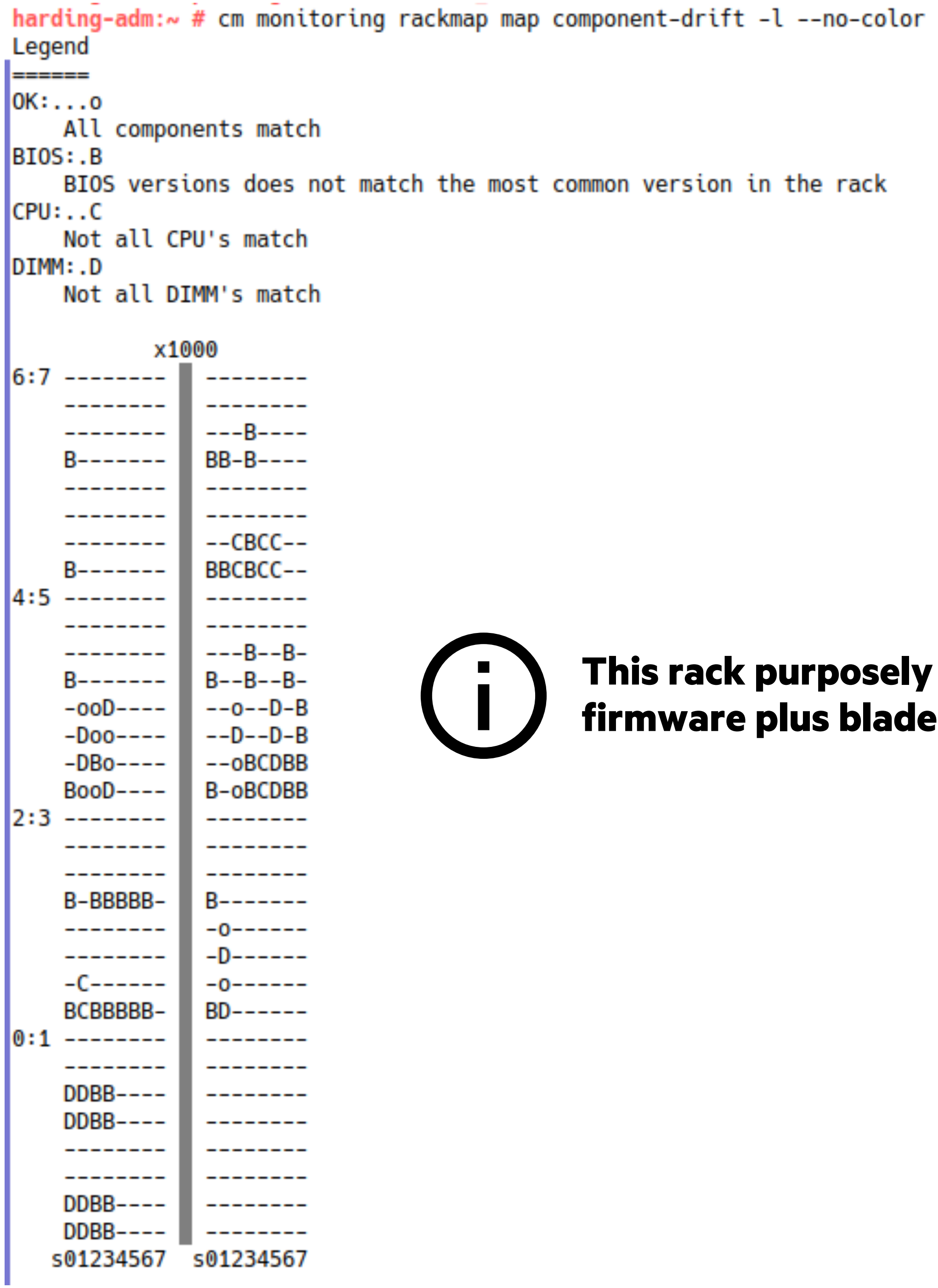
Terminal mapping of **component-drift, power, slingshot-switch-status and cpu-temperature**

- -l will prepend a legend
- -s summary
- --no-color
- --interactive
- -b to select a blade type

ⓘ **>=1.11**

```
# cm monitoring rackmap map -l  cpu-temperature Legend

======

0:.o

   Node temperature is within acceptable range of the median
temperature of other nodes in its rack.

1:./

   Node temperature deviates from the median temperature of
other nodes in its rack.

2:.X

   Node temperature significantly deviates from the median
temperature of other nodes in its rack.
```

cm monitoring rackmap



```
harding-adm:~ # cm monitoring rackmap map component-drift -l --no-color
Legend
======
OK:...o
    All components match
BIOS:.B
    BIOS versions does not match the most common version in the rack
CPU:..C
    Not all CPU's match
DIMM:.D
    Not all DIMM's match

           x1000
6:7 --------   --------
    --------   --------
    --------   ---B----
    B-------   BB-B----
    --------   --------
    --------   --------
    --------   --CBCC--
    B-------   BBCBCC--
4:5 --------   --------
    --------   --------
    --------   ---B--B-
    B-------   B--B--B-
    -ooD----   --o--D-B
    -Doo----   --D--D-B
    -DBo----   --oBCDBB
    BooD----   B-oBCDBB
2:3 --------   --------
    --------   --------
    --------   --------
    B-BBBBB-   B-------
    --------   -o------
    --------   -D------
    -C------   -o------
    BCBBBBB-   BD------
0:1 --------   --------
    --------   --------
    DDBB----   --------
    DDBB----   --------
    --------   --------
    --------   --------
    DDBB----   --------
    DDBB----   --------
   s01234567  s01234567
```

This rack purposely has varied hardware and firmware plus blades not available

cm support moncollect

1.11 introduced a command for log, status and configuration information pertaining to monitoring to aid support in troubleshooting issues

```
admin:~ # cm support moncollect
This command takes a long time to run as it observes data flow across different parts of the
monitoring architecture to aid support in troubelshooting.
Create early kafka topic numbers plus opensearch to assess if data is flowing.
Collecting generic linux background information...
Starting with several command outputs including commands which will take a while to run such as rpm
-Va
Collecting monitoring log files...
Collecting monitoring configuration...
Collecting monitoring status...this may take a while.
Querying metrics in tsb. This will take a while...
Checking for network traffic to nginx/telegraf.
Checking if cooldev data is flowing through mosquitto. This will take at least 30s.
Checking if this is an ICE system...
No ICE rack leader nodes detected.
Checking if this system has SU leaders...
Scalable Unit leader nodes detected so collecting data!
Creating tar ball and compressing data. This may take some time.
Monitoring tar ball is available here: /var/tmp/cm-support-monitoring-2023-12-05T0655CST.tar.xz
```

AIOps (Artificial Intelligence for Operations)

Thresholds and dashboards don't sufficiently identify anomalies - Anomalies can be based on multiple metrics

AIOps provides:

- Anomaly Detection (single metric and multi-metric models)
- Forecasting
- Preventative maintenance

Command	Info returned
`cm aiops status`	A status summary for all data types
`cm aiops cooldev status`	A status summary for cooling devices
`cm aiops fabric status`	A status summary for HPE Slingshot anomaly detection
`cm aiops forecast status`	A status summary for anomaly forecasts using the metrics from cooling distribution units (CDUs)
`cm aiops it status`	A status summary for hardware telemetry metrics on HPE CraySupercomputing EX clusters

AIOps (Artificial Intelligence for Operations)

- For slingshot_CraySwitchHardwareTelemetry, AIOps supports the following metrics:
- ASIC
- NetworkingDevice
- SystemBoard
- VoltageRegulator
- Chassis
- PowerSupply

- For slingshot_CrayFabricPerfTelemetry, AIOps supports the following metrics:
- rxPausePercent
- txPausePercent
- rxCongestion

AIOps (Artificial Intelligence for Operations)

```
# cm repo add ./aiops-1.13-cd1-media-suse-x86_64.iso
Mounting ISO file loopback...
 Running: cp -a /tmp/dr6__HPzDk /opt/clmgr/repos/cm/Cluster-Manager-AIOps-1.13-
sles15sp6-x86_64
Exporting repository for use with yume....
Exporting /opt/clmgr/repos/cm/Cluster-Manager-AIOps-1.13-sles15sp6-x86_64 through
httpd, http://harding-adm/repo/opt/clmgr/repo
s/cm/Cluster-Manager-AIOps-1.13-sles15sp6-x86_64
Updating default rpm lists...
Updating: /opt/clmgr/image/rpmlists/generated/generated-sles15sp6.rpmlist
Updating: /opt/clmgr/image/rpmlists/generated/generated-ice-sles15sp6.rpmlist
Updating: /opt/clmgr/image/rpmlists/generated/generated-lead-sles15sp6.rpmlist
Updating: /opt/clmgr/image/rpmlists/generated/generated-admin-sles15sp6.rpmlist
```

AIOps (Artificial Intelligence for Operations)

```
# cm node zypper --repos Cluster-Manager-1.13-sles15sp6-x86_64,Cluster-Manager-AIOps-1.13-sles15sp6-x86_64 -n adm
in install aiops-config aiops-service

# cm aiops enable
Running: /etc/opt/sgi/conf.d/80-aiops-configure
Loading container tar /opt/clmgr/aiops/images/aiops_service.1.13.tar.xz into podman. This may take a few minutes...
Getting image source signatures
Copying blob 0e4110d1f6ba done   |
Copying blob a30a5965a4f7 done   |
Copying blob 2693d194688d done   |
Copying blob 210189007693 done   |
Copying blob 4f37089fe9b0 done   |
Copying blob 60760791dba9 done   |
Copying config ea180aacd5 done   |
Writing manifest to image destination
Loaded image: localhost/aiops:1.13.250221T153052
Running: systemctl enable aiops-alert-processor.service
Created symlink /etc/systemd/system/multi-user.target.wants/aiops-alert-processor.service →
/usr/lib/systemd/system/aiops-alert-processor.service.
Running: systemctl enable aiops.service
Created symlink /etc/systemd/system/multi-user.target.wants/aiops.service → /usr/lib/systemd/system/aiops.service.
Created symlink /etc/systemd/system/multi-user.target.wants/aiops-mlflow.service → /usr/lib/systemd/system/aiops-
mlflow.service
```

AIOps (Artificial Intelligence for Operations)

```
# cm aiops start
Enabling flow service metrics_aiops_ad_fabric_perf
writing unit file for service flow-metrics_aiops_ad_fabric_perf at /usr/lib/systemd/system/flow-
metrics_aiops_ad_fabric_perf@.s
enable all units for flow-metrics_aiops_ad_fabric_perf...Enabling flow service metrics_aiops_pr_forecast
writing unit file for service flow-metrics_aiops_pr_forecast at /usr/lib/systemd/system/flow-
metrics_aiops_pr_forecast@.service
enable all units for flow-metrics_aiops_pr_forecast...Enabling flow service metrics_aiops_ad_crayex_it
writing unit file for service flow-metrics_aiops_ad_crayex_it at /usr/lib/systemd/system/flow-
metrics_aiops_ad_crayex_it@.servi
enable all units for flow-metrics_aiops_ad_crayex_it...Enabling flow service metrics_aiops_ad_fabric_temp
writing unit file for service flow-metrics_aiops_ad_fabric_temp at /usr/lib/systemd/system/flow-
metrics_aiops_ad_fabric_temp@.s
enable all units for flow-metrics_aiops_ad_fabric_temp...Starting flow service for metrics_aiops_ad_cooldev
writing unit file for service flow-metrics_aiops_ad_cooldev at /usr/lib/systemd/system/flow-
metrics_aiops_ad_cooldev@.service..
enable all units for flow-metrics_aiops_ad_cooldev...activated aiops cooldev models
activated aiops it models
activated aiops fabric models
activated aiops forecast models
Running: systemctl start aiops.service
Running: systemctl start aiops-alert-processor.service
```

ⓘ 24 hours before data is in dashboards as defined by 'warmup_period_hours' field in /opt/clmgr/aiops/models/model_repository.yaml and /opt/clmgr/aiops/models/model_repository_fabric.yaml

AIOps (Artificial Intelligence for Operations)

```
# cm sim add --service-group aiops-services
Running enable for aiops-services service: cadvisor.service

Running start for aiops-services service: cadvisor.service

Adding respective dashboard into grafana

Services enabled in SIM are: core-services | monitoring-services | suleader-services
| aiops-services | mpvt-service
```

AIOps (Artificial Intelligence for Operations) - 1.13 bug

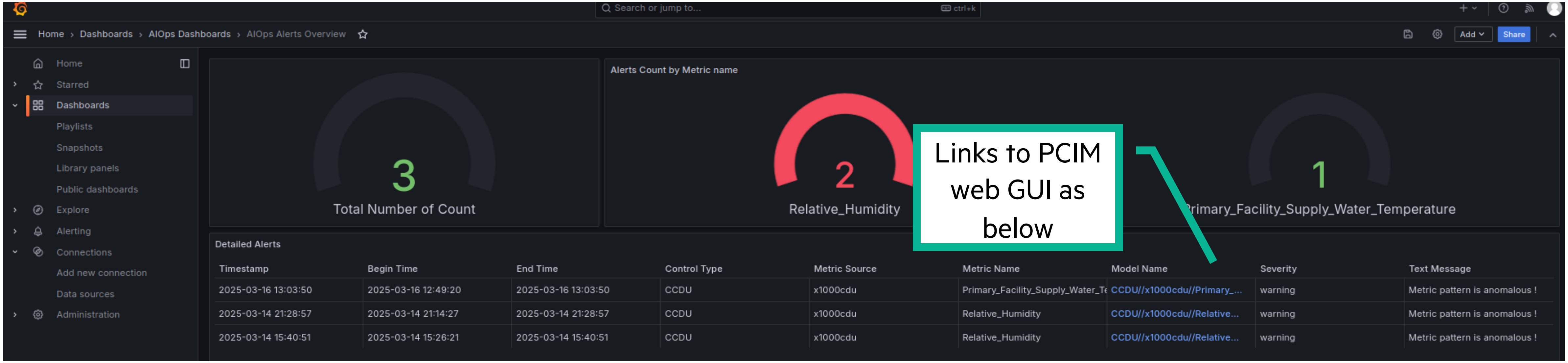
- 1.13: slingshot fabric perf telemetry issue due to a timestamp change (s -> ms) in the latest slingshot version (2.3)

- To work around the issue, add line 75:

```
# grep -nA 5 slingshot_CrayFabricPerfTelemetry:$ /opt/clmgr/aiops/config/aiops.yaml
71:     slingshot_CrayFabricPerfTelemetry:
72-       __type__: codecs.MultiCodec
73-       codecs:
74-       - __type__: codecs.JsonCodec
75-       - __type__: codecs.TimestampCodec
76-       - __type__: codecs.FabricPerfCodec
```

- …and then stop then start aiops
  - Note: Dashboards are disabled during the stop and start so you will see dashboard UID warning in the corner if you do have them open

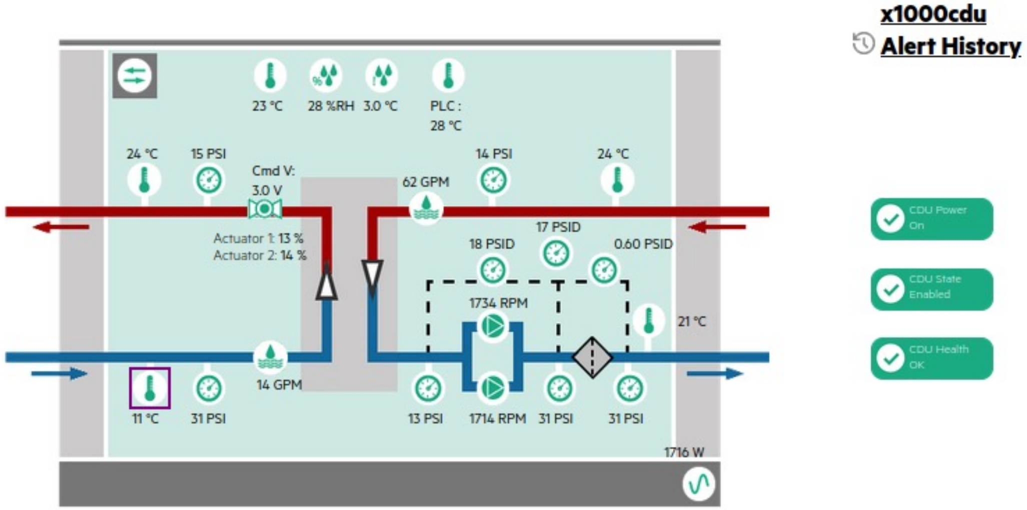
AIOps (Artificial Intelligence for Operations) – Alerts Overview



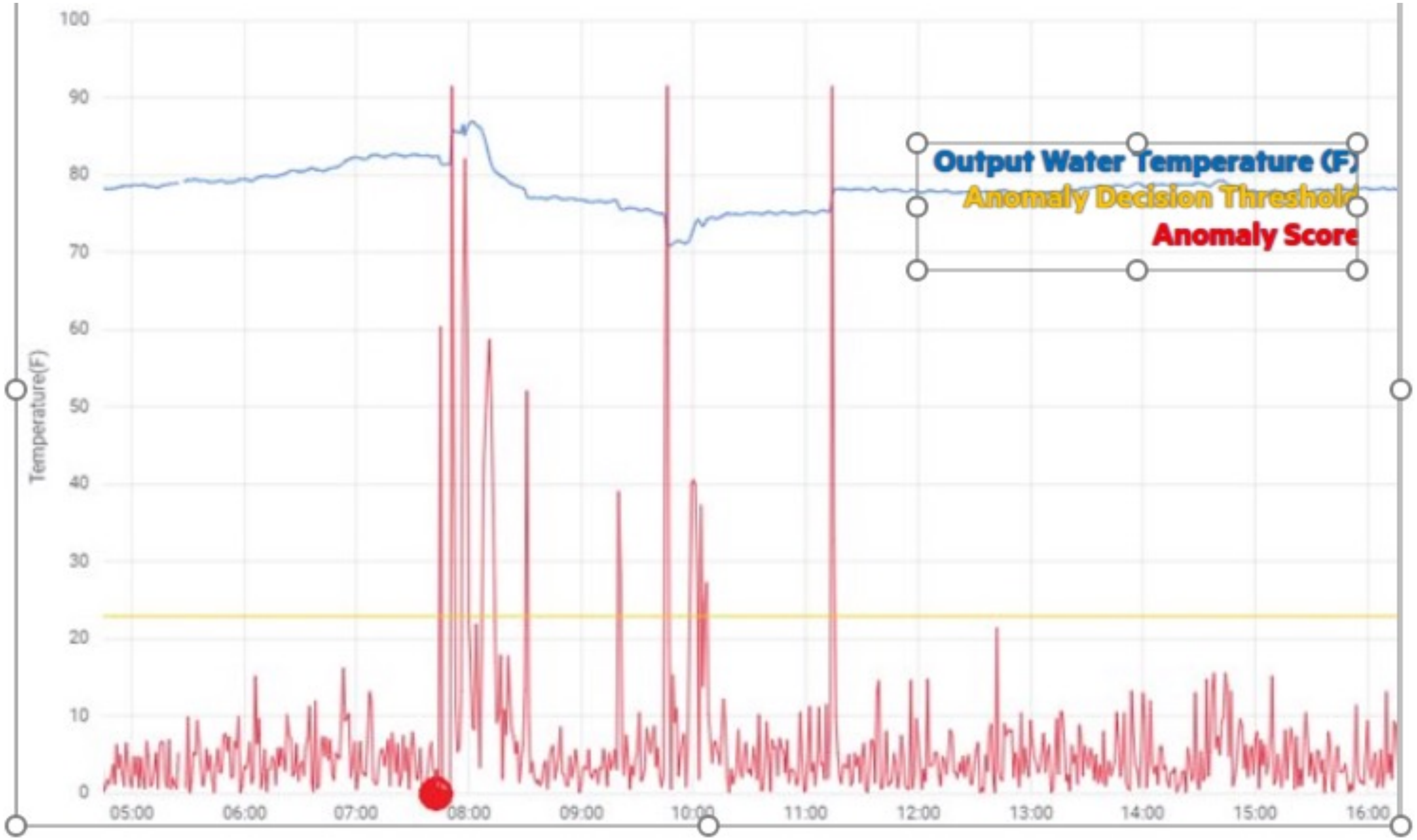
Links to PCIM web GUI as below

Detailed Alerts

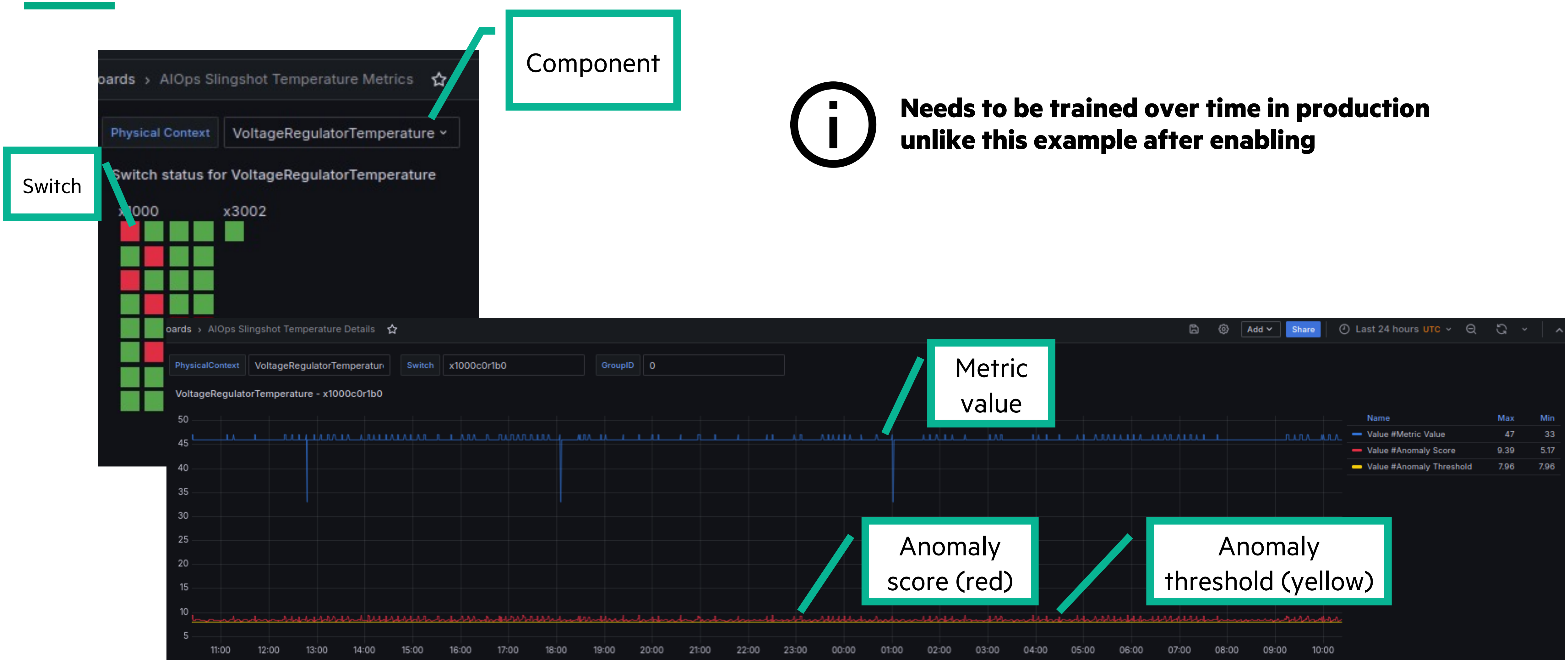
Timestamp	Begin Time	End Time	Control Type	Metric Source	Metric Name	Model Name	Severity	Text Message
2025-03-16 13:03:50	2025-03-16 12:49:20	2025-03-16 13:03:50	CCDU	x1000cdu	Primary_Facility_Supply_Water_Te	CCDU//x1000cdu//Primary_...	warning	Metric pattern is anomalous !
2025-03-14 21:28:57	2025-03-14 21:14:27	2025-03-14 21:28:57	CCDU	x1000cdu	Relative_Humidity	CCDU//x1000cdu//Relative...	warning	Metric pattern is anomalous !
2025-03-14 15:40:51	2025-03-14 15:26:21	2025-03-14 15:40:51	CCDU	x1000cdu	Relative_Humidity	CCDU//x1000cdu//Relative...	warning	Metric pattern is anomalous !



AIOps (Artificial Intelligence for Operations) – Anomaly score



AIOps – Clickable drill down to components



Component

Switch

Metric value

Anomaly score (red)

Anomaly threshold (yellow)

oards › AIOps Slingshot Temperature Metrics ☆

Physical Context | VoltageRegulatorTemperature ˅

Switch status for VoltageRegulatorTemperature

x1000 x3002

oards › AIOps Slingshot Temperature Details ☆

PhysicalContext: VoltageRegulatorTemperatur | Switch: x1000c0r1b0 | GroupID: 0

Add ˅ | Share | Last 24 hours UTC ˅

VoltageRegulatorTemperature - x1000c0r1b0

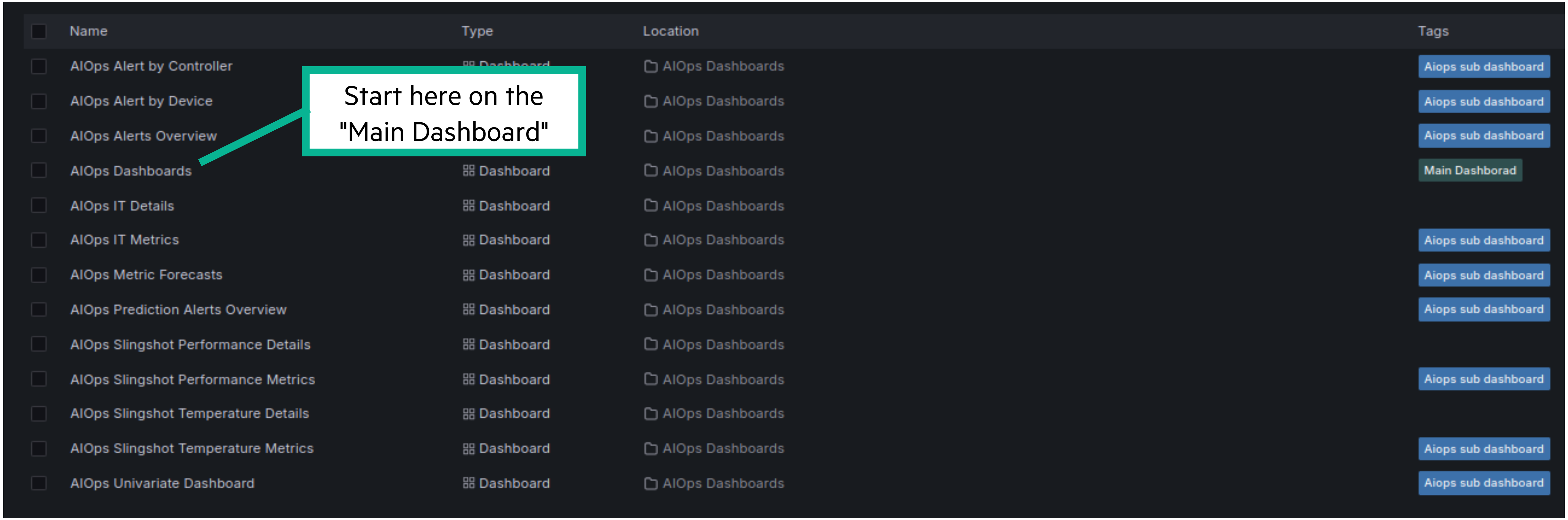
Name	Max	Min
Value #Metric Value	47	33
Value #Anomaly Score	9.39	5.17
Value #Anomaly Threshold	7.96	7.96

Needs to be trained over time in production unlike this example after enabling

AIOps – Clickable drill down to components

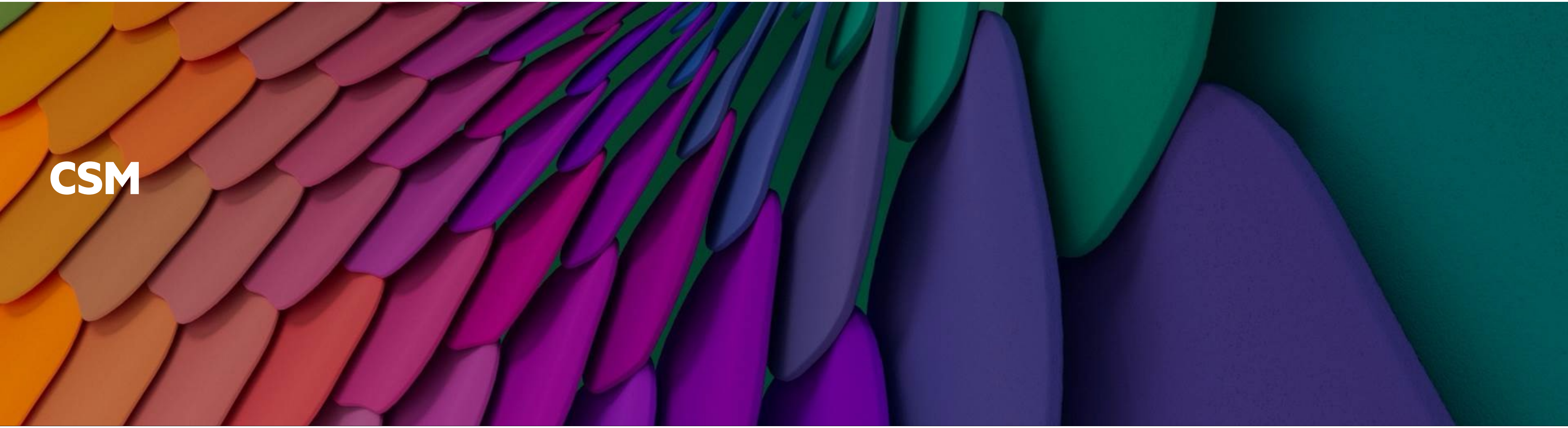
Dashboards ending in the word "Details" may contain "Incorrect query format" when accessed directly from the AIOps Dashboards folder

They are meant to be accessed via their respective "metrics" dashboard as a drill down to a specific component which is flagged there



☐	Name	Type	Location	Tags
☐	AIOps Alert by Controller	Dashboard	AIOps Dashboards	Aiops sub dashboard
☐	AIOps Alert by Device		AIOps Dashboards	Aiops sub dashboard
☐	AIOps Alerts Overview		AIOps Dashboards	Aiops sub dashboard
☐	AIOps Dashboards	Dashboard	AIOps Dashboards	Main Dashborad
☐	AIOps IT Details	Dashboard	AIOps Dashboards	
☐	AIOps IT Metrics	Dashboard	AIOps Dashboards	Aiops sub dashboard
☐	AIOps Metric Forecasts	Dashboard	AIOps Dashboards	Aiops sub dashboard
☐	AIOps Prediction Alerts Overview	Dashboard	AIOps Dashboards	Aiops sub dashboard
☐	AIOps Slingshot Performance Details	Dashboard	AIOps Dashboards	
☐	AIOps Slingshot Performance Metrics	Dashboard	AIOps Dashboards	Aiops sub dashboard
☐	AIOps Slingshot Temperature Details	Dashboard	AIOps Dashboards	
☐	AIOps Slingshot Temperature Metrics	Dashboard	AIOps Dashboards	Aiops sub dashboard
☐	AIOps Univariate Dashboard	Dashboard	AIOps Dashboards	Aiops sub dashboard

Start here on the "Main Dashboard"



CSM

CSM

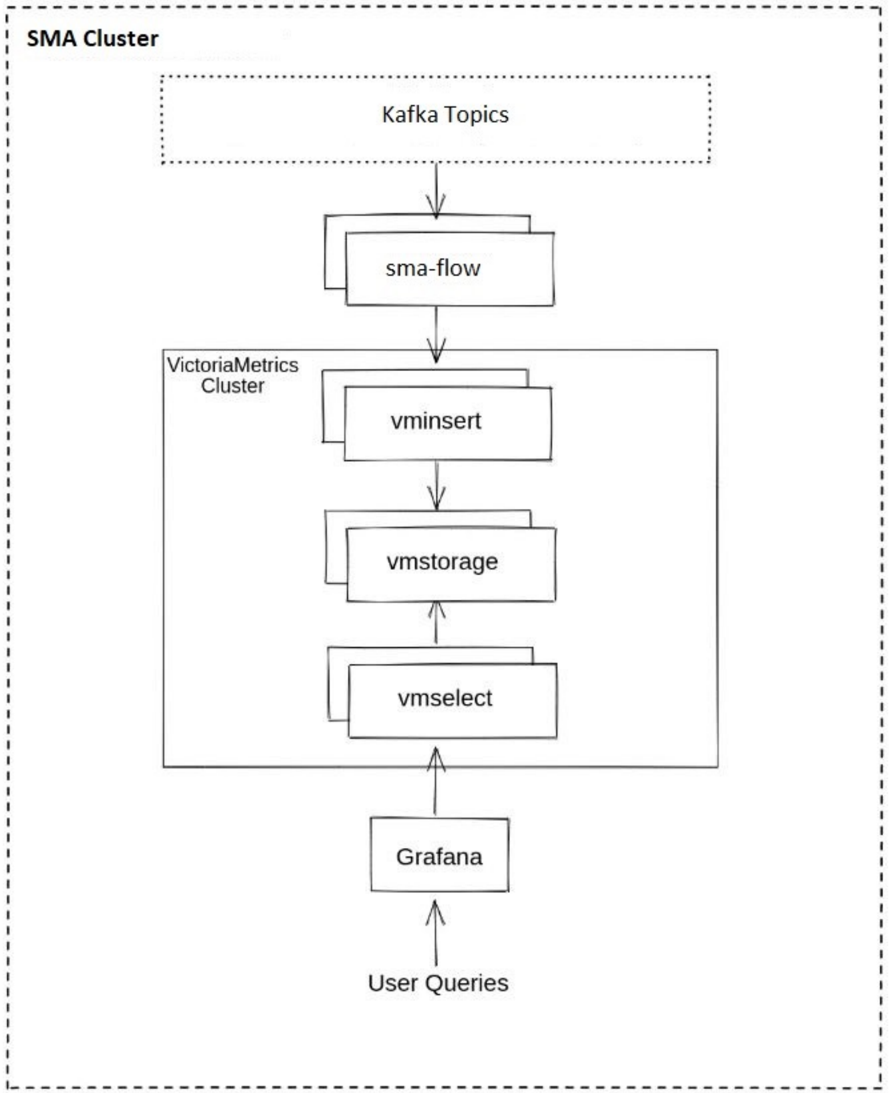
- Monitoring Configuration
- AIOps Configuration
- Alerting Configuration
- System Management Health Monitoring
- SMA Monitoring
- Logs

Monitoring Configuration

- VictoriaMetrics
- Kafka
- Log Handling
- OpenSearch
- LDMS
- Slingshot Fabric Manager

SMA VictoriaMetrics

- VictoriaMetrics
  - High-performance, cost-effective, and scalable time-series database used as long-term storage for Prometheus
  - Superior data compression and high-speed data ingestion
  - Source code: https://github.com/VictoriaMetrics/VictoriaMetrics
- VictoriaMetrics consists of three different services
  - vmstorage writes telemetry data to disk
  - vminsert handles write requests
  - vmselect handles read requests
- CSM 1.6 migrated from SMA PostgreSQL (pmdb) to VictoriaMetrics
  - Performance
    - VictoriaMetrics handles high ingestion rates and large query loads better than PostgreSQL
  - Scalability
    - Native support for clustering and distributed setups
  - Resource Efficiency
    - Lower CPU and memory usage for time-series workloads
  - Simplified Maintenance
    - Easier to manage in large-scale environments



Tuning SMA VictoriaMetrics

- Is VictoriaMetrics cluster storage nearly exhausted?
  - When storage is completely exhausted, VictoriaMetrics becomes inaccessible, requiring a manual recovery

```
ncn# kubectl exec -it -n sma sma-vm-cluster-vmstorage-0 -- df | egrep "Used|storage"
Filesystem  Size       Used       Available Use% Mounted on
/dev/rbd5   3450260472 3450250472 2589       98% /storage
ncn# kubectl exec -it -n sma sma-vm-cluster-vmstorage-1 -- df | egrep "Used|storage"
Filesystem  Size       Used       Available Use% Mounted on
/dev/rbd30  3450260472 3450250472 2589       98% /storage
```

- Increase the vmstorage PVC size
  - Check current size

```
ncn# kubectl -n sma get pvc vmstorage-volume-sma-vm-cluster-vmstorage-{0,1} -o json \
| jq '.items[0,1].status.capacity.storage’
"3344Gi"
"3344Gi"
```

  - Increase the PVC size
  - Observe the PVC resize in progress
  - Update vminsert and vmselect
  - Check that more free space is available

Tuning SMA Kafka

- Amount of storage allocated to Kavka via a Persistent Volume Claim (PVC) can be changed
  - Resizing does not require restarting the Kafka pods and does not result in data loss
- Resize Kafka PVC for each of three pods

```
ncn# kubectl -n sma edit pvc data-cluster-kafka-0
ncn# kubectl -n sma edit pvc data-cluster-kafka-1
ncn# kubectl -n sma edit pvc data-cluster-kafka-2
```

- Adjust Kafka pruning parameters
  - Disable purging of Kafka data due to size limits when there is ample storage for a 4 hour retention time
  - Reduce the amount of data stored for a specific kafka topic if storage is insufficient for that data
  - Both disk usage limit and retention policy data pruning are enabled by default

Log Handling – Fluent Bit

- Fluent Bit is a super fast, lightweight, and highly scalable logging, metrics, and traces processor and forwarder
- Configure Forwarding Log Messages from Fluent Bit to an External syslog Server
- Resize Fluent Bit Aggregator PVC
  - The fluentbit-aggregator can store data while a log target is unavailable
    - OpenSearch, kafka, and external log servers, which may be off-line or overwhelmed by log volume
- Enable SMA Audit Log Forwarding
  - Audit logs from all management nodes can be forwarded to an external server
  - Forwarding Fluent Bit audit logs to an external system enables centralized control and analysis and prevents log tampering
- Change the fluentbit-aggregator replicaCount or PVC Size
  - The Fluent Bit aggregator uses 3 replicas
    - To increase this there must be more than 3 eligible worker nodes
  - Increasing the per-replica PVC size can be used to increase the on-disk buffer
    - Must have sufficient available ceph disk space to allocate the additional space for the claims

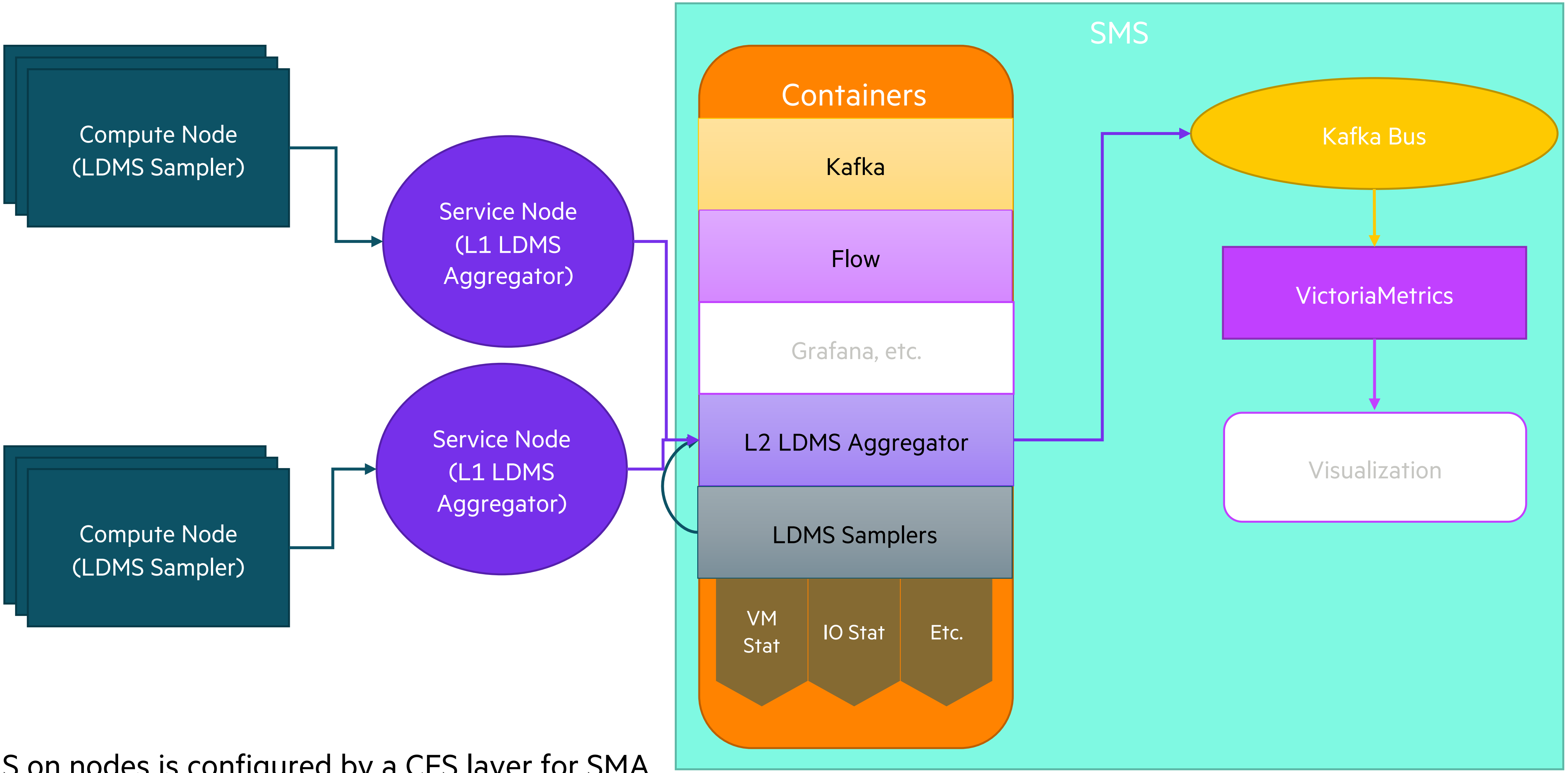
Log Handling – Aggregation and Filtering

- Log aggregation
  - Collects logs from host operating systems and containerized microservices
  - Fluentbit-collector service collects logs from each management node and stores them in OpenSearch
  - Fluentbit-collector forwards data to the fluentbit-aggregator services which write the data to Kafka and to OpenSearch
    - Pre-CSM 1.6 this was rsyslog-collector and rsyslog-aggregator
- Log filtering
  - SMA collects logs from all system nodes and from all management service containers, resulting in a very large total volume of collected logs
  - While the tools for collecting, storing, and searching logs are designed to scale with the size of the system, all of these tools function better when dealing with a smaller volume of log messages
  - If there exist a small list of log messages that are very common, but not useful for maintaining the system, they can be excluded from log collection
  - Fluentbit includes native filtering functions for discarding log messages matching regular expressions
  - If the lines to be excluded are from syslog, they must be filtered out by the fluentbit-aggregator
  - Lines from kubernetes pods can be filtered by either the fluentbit-aggregator or the fluentbit-collector

Tuning SMA OpenSearch

- Change the Number of OpenSearch Pods
  - Likely the same as the number of worker nodes
- Change the Number of OpenSearch Shards for Each Index
  - Same as number of OpenSearch nodes
  - Number of replicas: Commonly 1, but might need to be increased
- Change the OpenSearch Log Volume Size
  - Initial size set based on best practices and number of nodes at installation
    - Increase when
      - Number of nodes changes
      - Log volume is unexpectedly high
      - Desired retention period is longer than is allowed by the available storage
- Identify and Remedy Unassigned OpenSearch Shards
  - Opensearch may exceed the java heap space available to the pod
    - If this happens data written by fluentbit will be retried
    - But index replicas shard allocation is not automatically healed and must be reallocated manually
- Set Disk Usage Limits for OpenSearch
  - Configure how OpenSearch logs will be pruned
    - This can help prevent OpenSearch from consuming all the storage in its PVC when incoming log data volume spikes and when the default disk usage limits are insufficient to delete old data fast enough to prevent OpenSearch storage from completely filling up
- Change the OpenSearch Index Retention Policy
  - Adjust the min_index_age to retain data for longer
- Manually remove old data from OpenSearch

Lightweight Distributed Metric Service (LDMS)



- LDMS on nodes is configured by a CFS layer for SMA

Tuning SMA LDMS

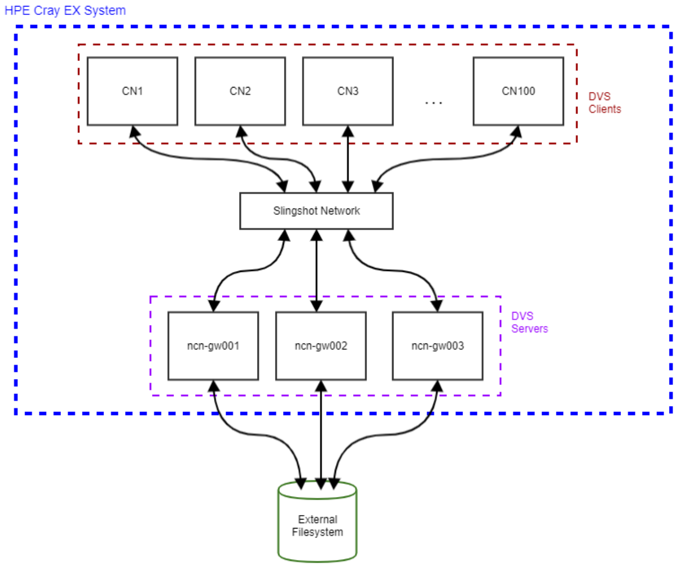
- LDMS data can be collected from all compute nodes or a subset
  - Ensure that LDMS compute node aggregation service is running on worker nodes
- LDSM data can be collected from NCNs (non-compute-nodes)
  - Ensure that LDMS NCN aggregation service is running on worker nodes
- Aggregation pods

```
ncn# kubectl get pods -o wide -A | grep ldms-aggr
sma sma-ldms-aggr-compute-0 2/2 Running 0  4d10h   10.38.0.76    ncn-w001
sma sma-ldms-aggr-ncn-0     2/2 Running 0  4d9h    10.38.0.97    ncn-w001
```

- Can be extended with additional LDMS samplers
  - SMA Admin guide
    - NVIDIA datacenter GPU manager (DCGM) can collect detailed information from NVIDIA GPUs
    - Add Customer Provided Samplers to LDMS v4 Configuration

LDMS Extension – DVS Sampling

- An LDMS sampler could be created to collect DVS information on client or server
- DVS records aggregate and per-mount-point statistics in several files in /sys/kernel/debug
  - Enables performance and root-cause analysis
  - Reports statistics for client and server nodes
- /sys/kernel/dvs/statistics
  - Time spent queued on the server
  - Time spent being processed by the server
  - Time spent in the underlying filesystem
  - Time spent on the network and in network transport software
- /sys/kernel/debug/dvs/stats
  - Aggregate statistcs which cannot be correlated to a specific DVS mount point
  - More interesting on the DVS server
- /sys/kernel/debug/dvs/mounts/NNN/stats
  - Per-mount statistics for the node
  - NNN is incremented when the node mounts any DVS filesystem to uniquely identify the mount
- /sys/kernel/debug/dvsipc/stats
  - DVS interprocess communication (IPC) bytes transferred and received, NAK counts, and message counts by type and size



LDMS Data Missing

- If SMA dashboards lack data from a node, check whether data from that node is available via Kafka
  - This example looks for data from a hostname “lnet01” in one of the cluster-kafka pods
    – Be ready to interrupt the command with a control-C since there could be much output

```
ncn# kubectl exec -it -n sma cluster-kafka-0 -- bash
kafka@cluster-kafka-0# /opt/kafka/bin/kafka-console-consumer.sh --bootstrap-server cluster-kafka-
bootstrap.sma.svc.cluster.local:9092 --topic cray-node --from-beginning | grep lnet01
```

– Sample of expected data: name of metrics will vary

```
{"metric":{"name":"cray_storage.cray_vmstat.cpu_wa","dimensions":{"product":"shasta","system":"ncn","service":"ldms","component":"c
ray_vmstat","hostname":"lnet01","cname":"x3000c0s13b0n0","job_id":"0"},"timestamp":1745479130110,"value":0},"meta":{"tenantId":"987
359e09cd74e56a5289a693b3b8875","region":"RegionOne"},"creation_time":2466321560063209057}

control-C
kafka@cluster-kafka-0# exit
```

- Check Kafka for another node
  - If no node data is present via Kafka, check health of cray-hms-hmcollector-ingress pods and inspect pod logs

```
ncn# kubectl get pods -n services | grep hmcollector
cray-hms-hmcollector-ingress-76b8d75767-86hbp                   2/2     Running     0               7d4h
cray-hms-hmcollector-ingress-76b8d75767-jjlbd                   2/2     Running     0               7d4h
cray-hms-hmcollector-ingress-76b8d75767-ndrjp                   2/2     Running     0               7d3h
cray-hms-hmcollector-poll-589c457778-km6d6                      2/2     Running     0               7d4h
ncn# kubectl logs -f -n services cray-hms-hmcollector-ingress-76b8d75767-86hbp
```

- Is there a problem with LDMS services on the node?

```
lnet01# systemctl status ldmsd-bootstrap.service
lnet01# systemctl status ldmsd@mellanox.service
lnet01# systemctl status ldmsd@ncn.service
lnet01# systemctl status ldmsd.service
```

  - Are they missing configuration files (pulled from Ceph/S3)?

- Is the LDMS rpm missing from the node?

```
lnet01# rpm -q cray-ldms
```

  - If missing, update CFS configuration, rebuild image, reboot node from new image

Monasca email notifications

- SMA monitors metric data that is transmitted on the main telemetry bus
  - Provides a way to notify users when select metric data is outside of normal operating values
  - Includes several pre-defined alarms
- SMA configmap `sma-monasca-alarms-configdata-cm`
  - email_destination
  - sendmail_server
  - email_source

```
ncn# kubectl -n sma edit cm sma-monasca-alarms-configdata-cm
ncn# kubectl -n sma describe cm sma-monasca-alarms-configdata-cm
```

- Changes require deleting pods and job and creating new job

```
ncn# kubectl -n sma delete pod -l component=notification
ncn# kubectl -n sma delete job -l component=alarms-init-job
ncn# kubectl -n sma delete pod -l component=alarms-init-job
ncn# vi alarms-init-job.yaml
ncn# kubectl -n sma apply -f alarms-init-job.yaml
```

Monasca is in CSM 1.6/SMA 1.0, but removed in CSM 1.7/SMA 1.11

Monasca local alarms

- Local alarms can be created that send email notifications
  - Create local alarm definitions

    ```
    ncn# vi customer-alarms-configmap.yaml
    ```

  - Deploy configmap

    ```
    ncn-# kubectl -n sma apply -f customer-alarms-configmap.yaml
    ```

  - Create job definition

    ```
    ncn# vi customer-alarms-init-job.yaml
    ```

  - Execute the SMA alarm initialization job

    ```
    ncn# kubectl -n sma apply -f customer-alarms-init-job.yaml
    ```

  - Verify job succeeds

    ```
    ncn# kubectl -n sma get po -l component=customer-alarms-init-job
    NAME READY STATUS RESTARTS AGE
    customer-alarms-init-job-dtrw5 0/1 Completed 0 5m
    ```

CLI for alarms

- List the state of all defined alarms

```
ncn# kubectl -n sma exec -it sma-monasca-agent-0 -c collector -- sh -c 'monasca alarm-list'
+--------------------------------------+--------------------------------------+------------------+
| id                                   | alarm_definition_id             | alarm_definition_name |
+--------------------------------------+--------------------------------------+------------------+
| 0881af14-5659-4468-813a-d99ac7f415c5 | cd72e681-995c-4f0a-9d29-c6a0a0e0dde8 | validation1Alarm |
| 64bbb62d-3cb1-466c-bed3-e7012f742683 | cd72e681-995c-4f0a-9d29-c6a0a0e0dde8 | validation1Alarm |
| b571cb91-2e1f-486e-9dc7-8b1e112cb530 | cd72e681-995c-4f0a-9d29-c6a0a0e0dde8 | validation1Alarm |
+--------------------------------------+--------------------------------------+------------------+
```

- List all defined alarms

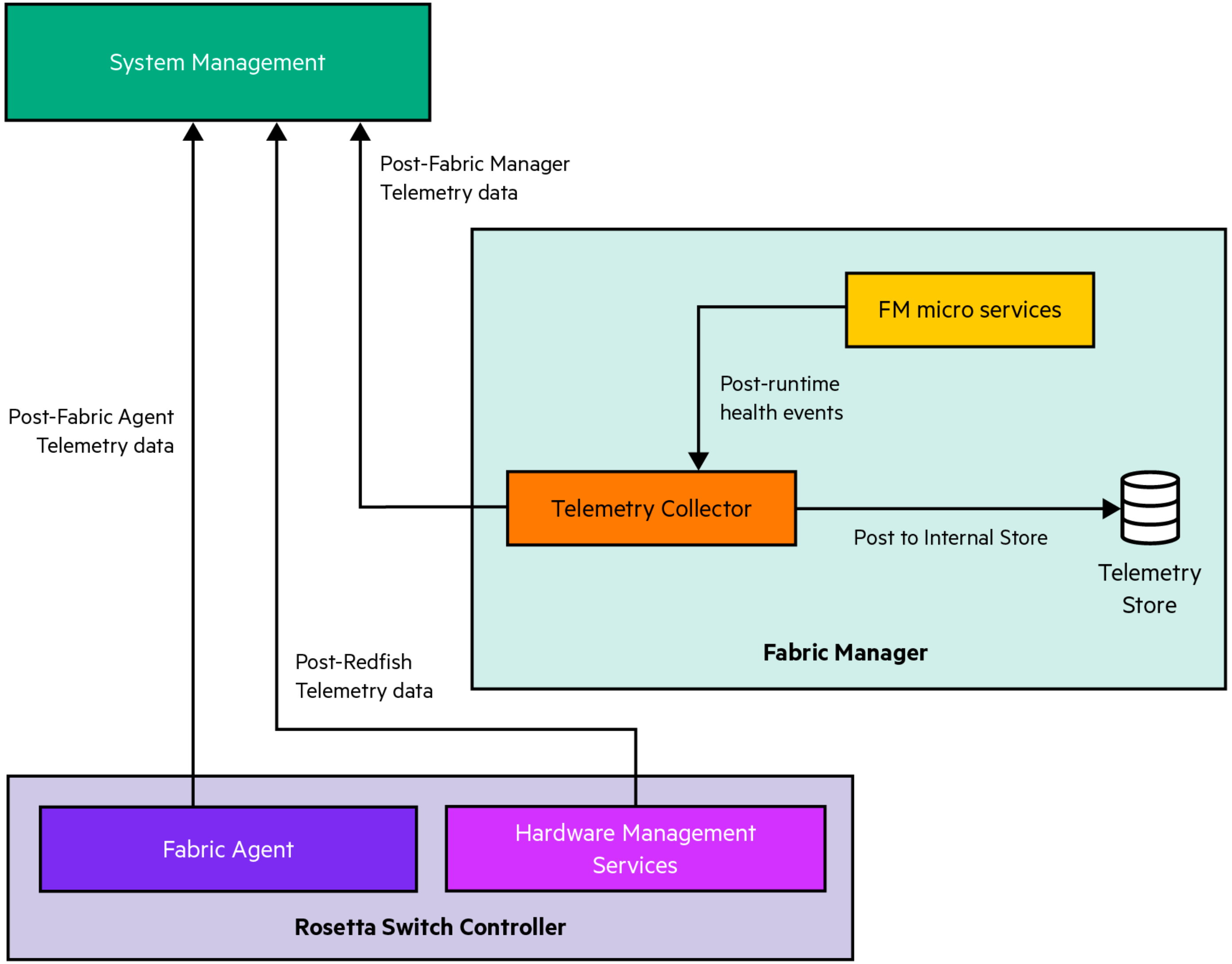
```
ncn# kubectl -n sma exec -it sma-monasca-agent-0 -c collector -- sh -c 'monasca alarm-definition-list'
+------------------------+--------------------------------------+------------------------------------------------------------------+
| name                   | id                                   | expression                                                       |
+------------------------+--------------------------------------+------------------------------------------------------------------+
| validation1Alarm       | 791d8c5a-f217-456c-9223-53976dfc5cdf | last(kubelet.health_status) < 0                                  |
| metricsTestAlarm       | a8ae3ac5-de01-4019-b1bb-090faf9c8c51 | avg(cray_test.other_test) < 20                                   |
| validation2Alarm       | ccc9cbb1-ad79-49a6-94ff-30c465a4479b | last(monasca.thread_count) < 0                                   |
| Critical Redfish Event | d98b6394-28b6-4ff2-a0d3-214fe5dc636e | count(dmtf.redfish_event{severity=Critical}, deterministic) >= 1 |
| vmstatTestAlarm        | e11d3234-3bdb-4318-8d5b-3cee64affb7f | avg(cray_test.vmstat_test) < 20                                  |
+------------------------+--------------------------------------+------------------------------------------------------------------+
```

- List all defined notifications

```
ncn# kubectl -n sma exec -it sma-monasca-agent-0 -c collector -- sh -c 'monasca notification-list'
+----------------+--------------------------------------+---------+---------------------------------------------------------------+
| name           | id                                   | type    | address                                                       |
+----------------+--------------------------------------+---------+---------------------------------------------------------------+
| defaultEmail   | 5f326486-5667-4afe-999f-1a65fe9ca7b0 | EMAIL   | email-distribution-list@customer.com                          |
| defaultWebhook | 8bb6cd0c-5674-42c4-aab1-cf1db78e164d | WEBHOOK | http://sma-alerta.sma.svc.cluster.local:8080/webhooks/monasca |
+----------------+--------------------------------------+---------+---------------------------------------------------------------+
```

Slingshot Fabric Manager

- Sources for events and metrics of the Slingshot switches and Fabric Manager
  - Fabric Agent (FA) running inside Rosetta switch controller
    - Periodically collects Rosetta counters, congestion metrics, and so on
    - Generates health events when it detects info, warning, or error condition from micro-services inside the FA
  - Hardware Management Services running inside Rosetta switch controller
    - Periodically collects switch environmental metrics such as voltage, power, temperature, and so on
    - Generates switch environmental or components Redfish events
  - Fabric Manager running inside Management Nodes generates health events when it detects an info, warning, or error condition from micro-services running inside the Fabric Manager


System Management
Post-Fabric Manager Telemetry data
FM micro services
Post-runtime health events
Post-Fabric Agent Telemetry data
Telemetry Collector
Post to Internal Store
Telemetry Store
Fabric Manager
Post-Redfish Telemetry data
Fabric Agent
Hardware Management Services
Rosetta Switch Controller

Slingshot Fabric Manager – Sub-topology

- Telemetry configuration is flexible at sub-topology level
  - Creating multiple telemetry configurations in the Fabric Manager
  - Categorizing and filtering only a subset of switches and group them according to their role
  - Configuring metrics or events selectively at category and subcategory level
  - Streaming any metric or event with custom periodicity
  - Filtering events by their severity
  - Streaming telemetry data to multiple collectors and multiple custom configurations
  - Enabling or disabling telemetry configurations based on requirements
  - Monitoring exclusively health of Fabric Agent

Slingshot Fabric Manager – Granular Control

- Monitoring at sub-topology level
  - Set telemetry data collection for specific subset of switches
  - Distinct telemetry configurations for each subset
  - Provides ability to get data based on
    - Switches' role in the system
    - Dragonfly group ID
    - Cabinet location
- Customized monitoring
  - Enable or disable metrics or events
    - Collect and analyze required data
    - Reduce overhead
    - Improve efficiency
  - Streamlines monitoring processes
  - Optimizes resource utilization
  - Prioritizes data collection from operational requirements
- Reliable transport of critical events and data streaming
  - Monitor Fabric Agent telemetry source health and flow of Fabric Agent telemetry using heartbeat
- Resource optimization and scalability management
  - Prioritize critical congestion metrics at higher rate
  - Lesser frequency for RFC statistics
  - Can distribute workload across multiple collectors
- Event filtering
  - Severity level
  - Event categories or subcategories
- Troubleshooting and diagnostics
  - Enable optional Fabric Agent metrics while looking for root cause of issue or analyzing system performance
    - RFC statistics
    - Routing, hard error, port error, counters
- Co-existence of multiple monitoring solutions
  - Can integrate and use multiple monitoring tools by streaming telemetry to multiple collectors based on their requirements and capabilities

Slingshot Fabric Manager - Configuration

- Enter the CSM slingshot-fabric-manager pod

```
ncn# podname=$(kubectl get pods -n services | grep slingshot-fabric-manager | awk '{print $1}’)
ncn# kubectl exec -it -n services $podname -- /bin/bash
```

- Get CSM Base Domain Name

```
slingshot-fabric-manager> env | grep BASE_DOMAIN
system.domain.com
```

- Create default telemetry configuration

```
slingshot-fabric-manager> fmn-create-telemetry-config --name csm-config --default
```

- Stream telemetry to CSM collector

```
slingshot-fabric-manager> fmn-update-telemetry-config --name csm-config \
--collector http://hmcollector.hmnlb.system.domain.com
```

- Enable data streaming from Fabric Manager

```
slingshot-fabric-manager> fmctl update /fabric/topology-policies/template-policy \
fabricPropertyMap.HMSCollector=http://hmcollector.hmnlb.system.domain.com
```

- Enable named telemetry configuration

```
slingshot-fabric-manager> fmn-update-telemetry-config --name csm-config --enable
```

Slingshot Fabric Manager – show telemetry config

- What is in the default configuration from previous slide?

```
slingshot-fabric-manager> fmn-show-telemetry-config --name csm-config
{
    "categories": {
        "CrayFabricHealth": {
            "all": {
                "severity": "CRITICAL"
            }
        },
        "CrayFabricPerfTelemetry": {
            "Congestion": {
                "periodicity": 60.0
            }
        }
    },
    "collector": "http://hmcollector.hmnlb.system.domain.com",
    "enable": false,
    "eventsFailureRetries": 3,
    "heartbeatEnable": true,
    "heartbeatPeriodicity": 60.0,
    "name": "csm-config"
}
```

Slingshot Fabric Manager – show telemetry config

- Which configurations have been created

```
  slingshot-fabric-manager> fmn-show-telemetry-config --list
  +----------------------------------------+----------+
  |             Configurations             |  State   |
  +----------------------------------------+----------+
  |  /telemetry/configurations/csm-config  |  ACTIVE  |
```

- Show supported configuration parameters

```
slingshot-fabric-manager> fmn-show-telemetry-config --supported-parameters
```

- To display list of supported metrics or events for each category or subcategory

```
slingshot-fabric-manager> fmn-show-telemetry-config --supported-metrics
slingshot-fabric-manager> fmn-show-telemetry-config --supported-events
```

- To show telemetry configurations on switches

```
slingshot-fabric-manager> fmn-show-telemetry-config --switch-list x3000c0r39b0,x3000c0r40b0 –detail
```

- If labels have been assigned to switches

```
slingshot-fabric-manager> fmn-show-telemetry-config --switch-label-list Gateway
```

Slingshot Fabric Manager – Sample Configurations

- Several sample telemetry configurations are available as JSON files **`/opt/slingshot/examples/telemetry/`**

  ```
  all-telemetry-config.json
  critical-events-fabric-manager.json
  fabric-manager-csm-config.json
  switch-list-config.json
  ```

- Right side shows all-telemetry enabled (in two columns to fit page)
  - Very verbose for CrayFabricHealth using INFO level
  - Notice that the periodicity can be adjusted for each telemetry metric
- Create all-config telemetry configuration

  ```
  fmn-create-telemetry-config --file
    /opt/slingshot/examples/telemetry/all-telemetry-
    config.json --name all-telemetry-config
  fmn-update-telemetry-config --name all-config --collector
    http://hmcollector.hmnlb.system.domain.com
  fmn-update-telemetry-config --name all-config --enable
  ```

```
{
 "categories": {
  "CrayFabricHealth": {
   "all": {
    "severity": "INFO"
   }
  },
  "CrayFabricPerfTelemetry": {
   "Congestion": {
    "periodicity": 60
   },
   "RFC2819": {
    "periodicity": 60
   },
   "RFC4188": {
    "periodicity": 60
   },
   "RFC1213": {
    "periodicity": 60
   },
   "RFC3635": {
    "periodicity": 60
   },
   "RFC2863": {
    "periodicity": 60
   },
   "PauseDetails": {
    "periodicity": 60
   },
   "CongestionDetails": {
    "periodicity": 60
   }
  },
  "CrayFabricCriticalTelemetry": {
   "HardErrors": {
    "periodicity": 60
   },
   "RoutingErrors": {
    "periodicity": 60
   },
   "PortErrors": {
    "periodicity": 60
   },
   "PortErrorsDetails": {
    "periodicity": 60
   }
  }
 },
 "collector": "http://any.valid.collector",
 "enable": false,
 "eventsFailureRetries": 3,
 "heartbeatEnable": true,
 "heartbeatPeriodicity": 60,
 "name": "/telemetry/configurations/all-telemetry-config"
}
```

AIOps Configuration

AIOps

- Typical monitoring systems are based on thresholds
  - IT operations require administrators to monitor dashboards
  - The dashboards consolidate data from multiple monitoring systems based on established thresholds
- AIOps offers the following features:
  - Anomaly detection and processing
    - AIOps issues notifications for critical anomalies detected in the metrics derived from the cooling distribution units (CDUs)
    - AIOps simplifies data center management by reducing the number of false alarms, surfacing only anomalous results, limiting the number of dashboards needed, and providing other features
  - Default cooling device monitoring
    - Rather than rely on established thresholds, the default AIOps cooling device monitor uses dynamic thresholds for monitoring cooling devices
    - These dynamic thresholds are calculated automatically and are based on the latest data used to train the AI models
    - The data from the cooling systems can change over time for a number of reasons, and this approach makes alerting relevant to the latest data
  - Alert processing
    - You can display AIOps data in Grafana
    - Within Grafana, AIOps provides several dashboards in JSON format

AIOps for CDUs

- Forecasting - communicate metric patterns and value 16 minutes into the future
  - AIOps forecasting uses metric values derived from the cooling distribution units (CDUs)
  - These values appear in the AIOps Anomaly Forecast dashboard
  - The cluster manager creates this dashboard based on historical cluster data
  - Forecasting is a powerful method for predicting future trends and values in time-series data
  - The AIOps time series forecasting technique assumes that past trends can predict future events
  - The technique uses historical data to predict future values
- Failure prediction - communicate expected CDU sensor failures up to 30 minutes before the failure occurs
  - Failures in cooling devices can have critical implications on cluster efficiency and reliability
  - In high-performance computing clusters, cooling is essential to prevent the overheating of the high-powered computing components
  - Cooling ensures optimal performance and longevity
  - AIOps uses long short-term memory (LSTM) models to learn about events and anomalies that lead up to failures in cooling devices using historical data provided by the cluster
  - AIOps attempts to generate alerts before failures happen during inference
  - These alerts can be viewed in `AIOps Alert Overview Grafana` dashboard

Start AIOps monitoring

- Check status of AIOps

```
ncn# cm aiops status
sma-aiops is not running
```

- Start AIOps

```
ncn# cm aiops start
Deployment scaled up successfully
```

- AIOps anomaly detection models may need to collect statistics on monitored metrics before producing results
  - Up to 24 hours before results from the anomaly detection models are available
- Start HPE Slingshot anomaly detection

```
ncn# cm aiops fabric start
```

- Start CDU metric forecasts

```
ncn# cm aiops forecast start
```

Start AIOps trainer

- Enable CDU failure prediction
  - Check status of AIOps

```
ncn# cm aiops trainer status
SMA AIOps Trainer is not running
```

  - Start AIOps
    - Must specify one to seven days of training

```
ncn# cm aiops trainer start -dur 7
Deployment scaled up successfully
```

  - Check status of AIOps

```
ncn# cm aiops trainer status
SMA AIOps Trainer is running
```

Alerting Configuration

- VictoriaMetrics alerts
- Alertmanager
- Monasca Alarms and Notifications
- cm health alert
- sat sensors

VictoriaMetrics Health Checks

- VictoriaMetrics alerts provide coverage across infrastructure and platform
- Coarse-grained and comprehensive, as opposed to fine-grained and exhaustive
- Supports preventive and diagnostic use cases

NON-COMPUTE NODES	UTILITY STORAGE	CONTAINER ORCHESTRATION	SERVICE MESH	WORKLOADS
• CPU and memory utilization • Local storage utilization • Network I/O errors and latency • Clock skew	• Ceph status • Storage utilization • Disk I/O errors and latency	• Kubernetes status • API errors • CPU and memory overcommitments	• Istio status • Service availability • Service request rates • Service response statuses and latency	• Status of pods, deployments, stateful sets, daemon sets, jobs • CPU, memory, network, and storage utilization and errors

Retrieving Alerts from VictoriaMetrics

```
ncn# kubectl -n sysmgmt-health get svc vmalert-vms
NAME          TYPE        CLUSTER-IP      EXTERNAL-IP   PORT(S)    AGE
vmalert-vms   ClusterIP   10.23.224.215   <none>        8080/TCP   7d1h
ncn# curl -s http://10.23.224.215:8080/api/v1/alerts \
| jq -j '.data' | grep alertname | grep -v description |sort | uniq -c
     10           "alertname": "CPUThrottlingHigh",
      4           "alertname": "etcdHighNumberOfFailedGRPCRequests",
      1           "alertname": "IstioHighRequestLatency",
      1           "alertname": "IstioLatency99Percentile",
      2           "alertname": "KubeDeploymentReplicasMismatch",
      1           "alertname": "KubeDeploymentRolloutStuck",
    218           "alertname": "KubeJobFailed",
      2           "alertname": "KubeJobNotCompleted",
     58           "alertname": "KubePersistentVolumeInodesFillingUp",
      2           "alertname": "KubeStatefulSetReplicasMismatch",
      2           "alertname": "KubeStatefulSetUpdateNotRolledOut",
     12           "alertname": "MetalLBBGPSessionDown",
      1           "alertname": "NodeCpuUsageTooHigh",
      1           "alertname": "NodeCpuUsageWarning",
     14           "alertname": "NodeSystemdServiceFailed",
      2           "alertname": "NodeTextFileCollectorScrapeError",
      3           "alertname": "OpensearchPVCVolumeFullCritical",
      3           "alertname": "OpensearchPVCVolumeFullWarning",
      4           "alertname": "PodReadinessProbeFailure",
      3           "alertname": "PostgresqlHighRollbackRate",
      4           "alertname": "RequestErrorsToAPI",
      1           "alertname": "SwitchPortDown",
      8           "alertname": "TargetDown",
      4           "alertname": "TooManyLogs",
      4           "alertname": "TooManyScrapeErrors",
      2           "alertname": "Watchdog",
```

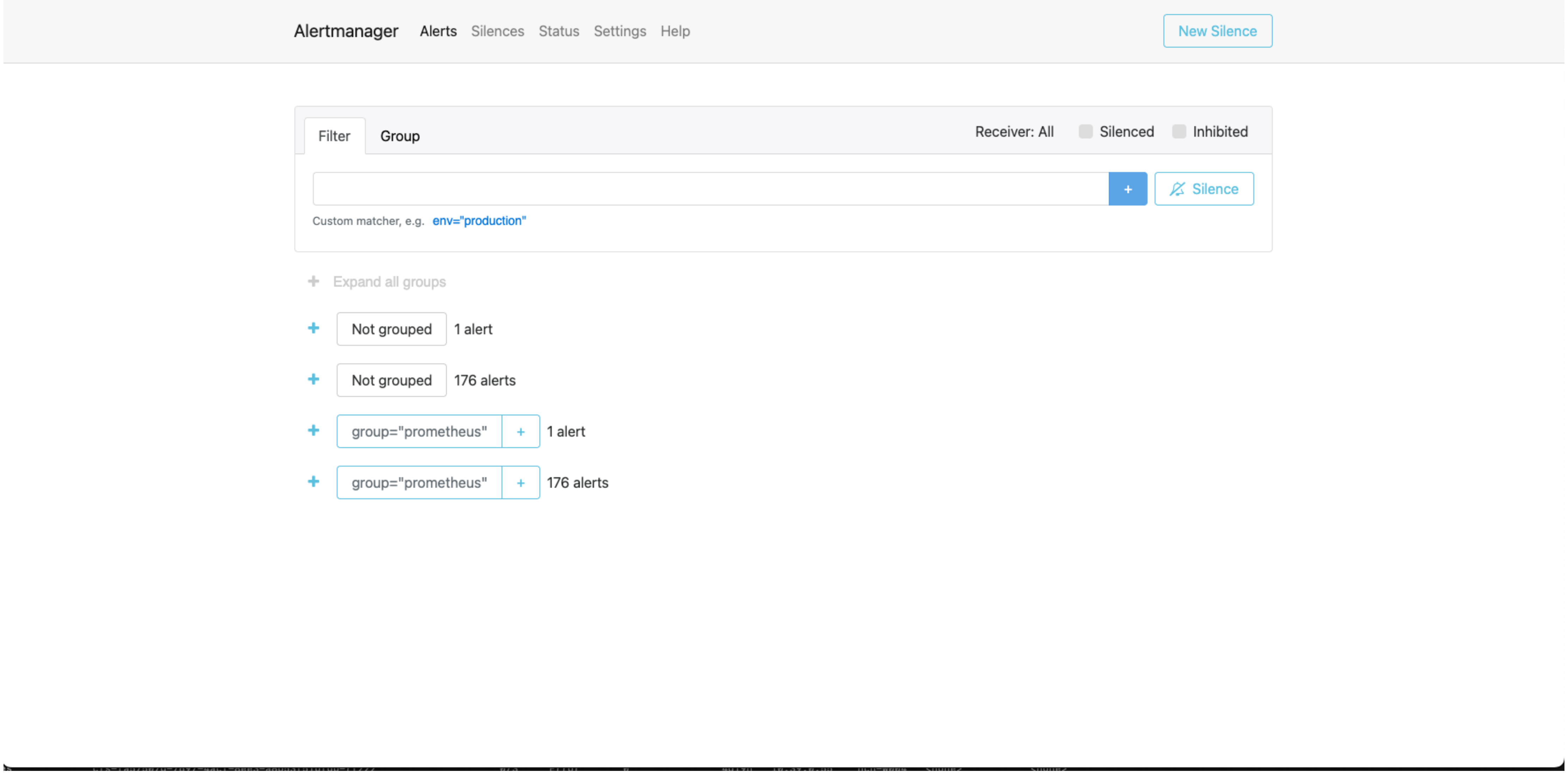
Retrieving Alerts from VictoriaMetrics - CPUThrottlingHigh

```
ncn# curl -s http://10.23.224.215:8080/api/v1/alerts | jq -j '.data.alerts \
| map(select(.labels.alertname == CPUThrottlingHigh")) | max_by(.activeAt)'
{
  "state": "pending",
  "name": "CPUThrottlingHigh",
  "value": "0.7037037",
  "labels": {
    "alertgroup": "kubernetes-resources",
    "alertname": "CPUThrottlingHigh",
    "container": "cray-k8s-encryption",
    "group": "prometheus",
    "namespace": "kube-system",
    "pod": "cray-k8s-encryption-k44xx",
    "severity": "info"
  },
  "annotations": {
    "description": "\"70.37%\" throttling of CPU in namespace \"kube-system\" for container \"cray-k8s-encryption\" in pod  \"cray-k8s-encryption-
 k44xx\".\n",
    "summary": "Processes experience elevated CPU throttling."
  },
  "activeAt": "2025-04-23T12:24:30Z",
  "id": "7013044465115556713",
  "rule_id": "13184710450506974973",
  "group_id": "6173127222314601989",
  "expression": "sum(increase(container_cpu_cfs_throttled_periods_total{container!=\"\", }[5m])) by (container, pod,
 namespace)\n  /\nsum(increase(container_cpu_cfs_periods_total{}[5m])) by (container, pod, namespace)\n  > ( 25 / 100 )",
  "source": "http://vmalert-vms-7f649f7d98-ch255:8080/vmalert/alert?group_id=6173127222314601989&alert_id=7013044465115556713",
  "restored": false,
  "stabilizing": false
}
```

Retrieving Alerts from VictoriaMetrics - KubeJobFailed

```
ncn# curl -s http://10.23.224.215:8080/api/v1/alerts | jq -j '.data.alerts' | head -35
[
  {
    "state": "firing",
    "name": "KubeJobFailed",
    "value": "1",
    "labels": {
      "alertgroup": "kubernetes-apps",
      "alertname": "KubeJobFailed",
      "cluster": "cluster-name",
      "condition": "true",
      "container": "kube-state-metrics",
      "endpoint": "http",
      "instance": "10.34.0.81:8080",
      "job": "kube-state-metrics",
      "job_name": "cfs-87439415-9383-4830-9e27-ccc6491d45f7",
      "namespace": "services",
      "pod": "cray-sysmgmt-health-kube-state-metrics-68b46b54b7-j747x",
      "prometheus": "sysmgmt-health/vms",
      "service": "cray-sysmgmt-health-kube-state-metrics",
      "severity": "warning"
    },
    "annotations": {
      "description": "Job services/cfs-87439415-9383-4830-9e27-ccc6491d45f7 failed to complete. Removing failed job after investigation should clear this alert.",
      "runbook_url": "https://runbooks.prometheus-operator.dev/runbooks/kubernetes/kubejobfailed",
      "summary": "Job failed to complete."
    },
    "activeAt": "2025-04-18T00:07:15Z",
    "id": "10140759166439655108",
    "rule_id": "10792094807988496288",
    "group_id": "10587253488064982662",
    "expression": "kube_job_failed{job=\"kube-state-metrics\", namespace=~\".*\"}  > 0",
    "source": "http://vmalert-vms-7f649f7d98-ch255:8080/vmalert/alert?group_id=10587253488064982662&alert_id=10140759166439655108",
    "restored": false,
    "stabilizing": false
  },
```

Alertmanager



Custom matcher, e.g. env="production"

+ Not grouped 1 alert

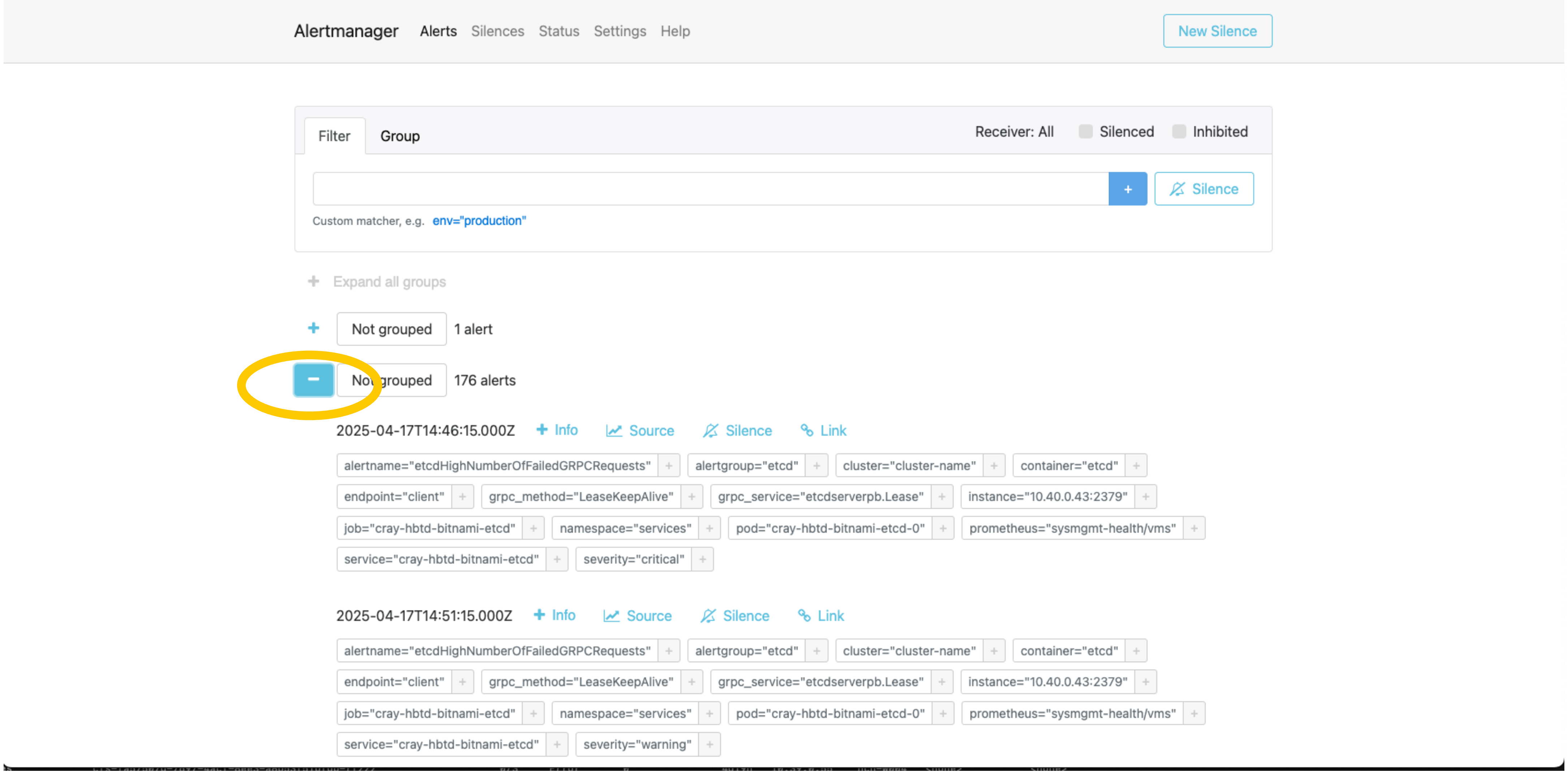
+ Not grouped 176 alerts

+ group="prometheus" + 1 alert

+ group="prometheus" + 176 alerts

https://alertmanager.cmn.SYSTEM_DOMAIN_NAME/

Alertmanager - Expanded



2025-04-17T14:46:15.000Z

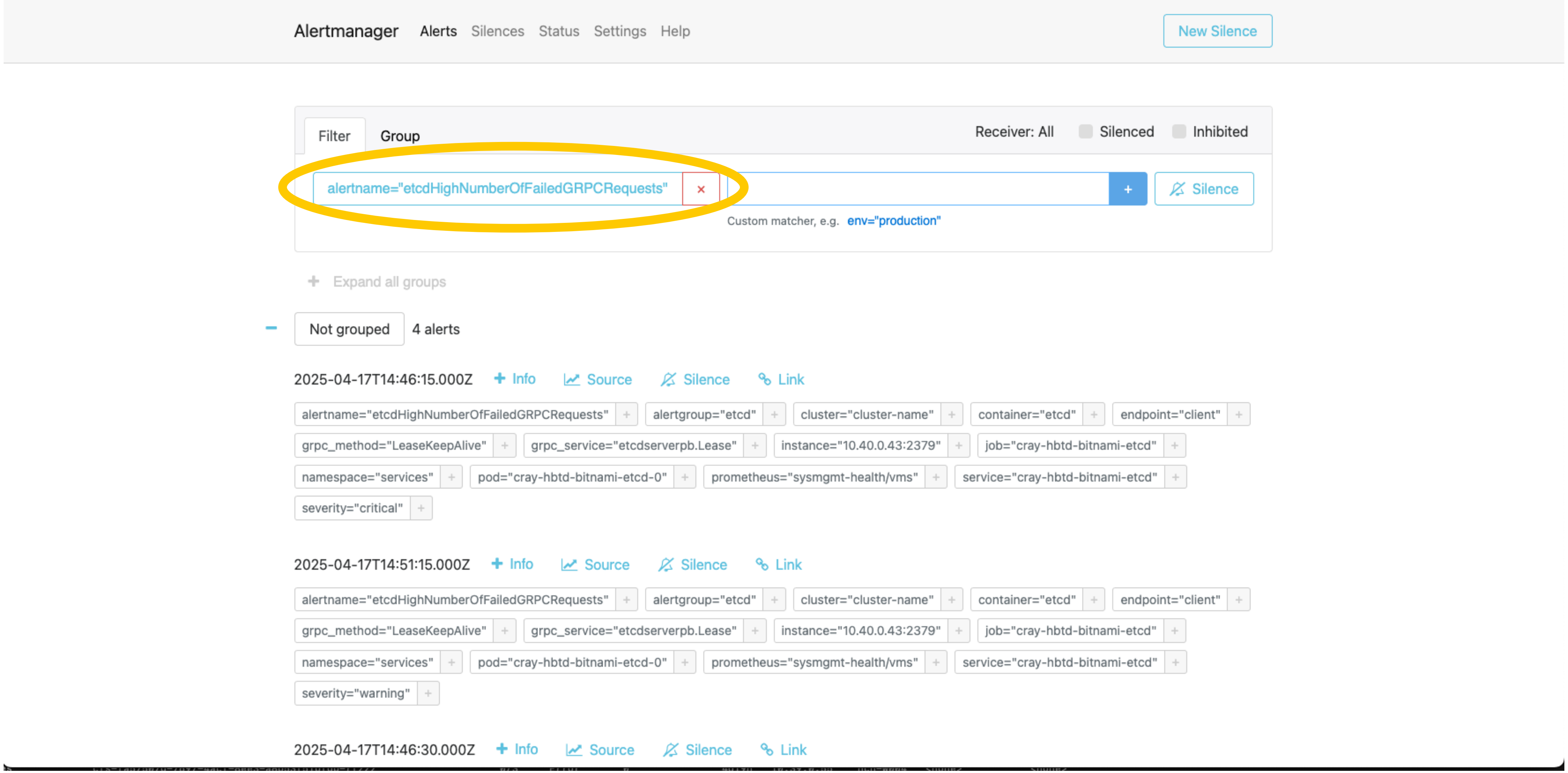
alertname="etcdHighNumberOfFailedGRPCRequests" alertgroup="etcd" cluster="cluster-name" container="etcd" endpoint="client" grpc_method="LeaseKeepAlive" grpc_service="etcdserverpb.Lease" instance="10.40.0.43:2379" job="cray-hbtd-bitnami-etcd" namespace="services" pod="cray-hbtd-bitnami-etcd-0" prometheus="sysmgmt-health/vms" service="cray-hbtd-bitnami-etcd" severity="critical"

2025-04-17T14:51:15.000Z

alertname="etcdHighNumberOfFailedGRPCRequests" alertgroup="etcd" cluster="cluster-name" container="etcd" endpoint="client" grpc_method="LeaseKeepAlive" grpc_service="etcdserverpb.Lease" instance="10.40.0.43:2379" job="cray-hbtd-bitnami-etcd" namespace="services" pod="cray-hbtd-bitnami-etcd-0" prometheus="sysmgmt-health/vms" service="cray-hbtd-bitnami-etcd" severity="warning"

https://alertmanager.cmn.SYSTEM_DOMAIN_NAME/

Alertmanager – Filter on alertname - etcdHighNumberOfFailedGRPCRequests



alertname="etcdHighNumberOfFailedGRPCRequests" ×

Custom matcher, e.g. env="production"

Not grouped 4 alerts

2025-04-17T14:46:15.000Z + Info Source Silence Link

alertname="etcdHighNumberOfFailedGRPCRequests" alertgroup="etcd" cluster="cluster-name" container="etcd" endpoint="client" grpc_method="LeaseKeepAlive" grpc_service="etcdserverpb.Lease" instance="10.40.0.43:2379" job="cray-hbtd-bitnami-etcd" namespace="services" pod="cray-hbtd-bitnami-etcd-0" prometheus="sysmgmt-health/vms" service="cray-hbtd-bitnami-etcd" severity="critical"

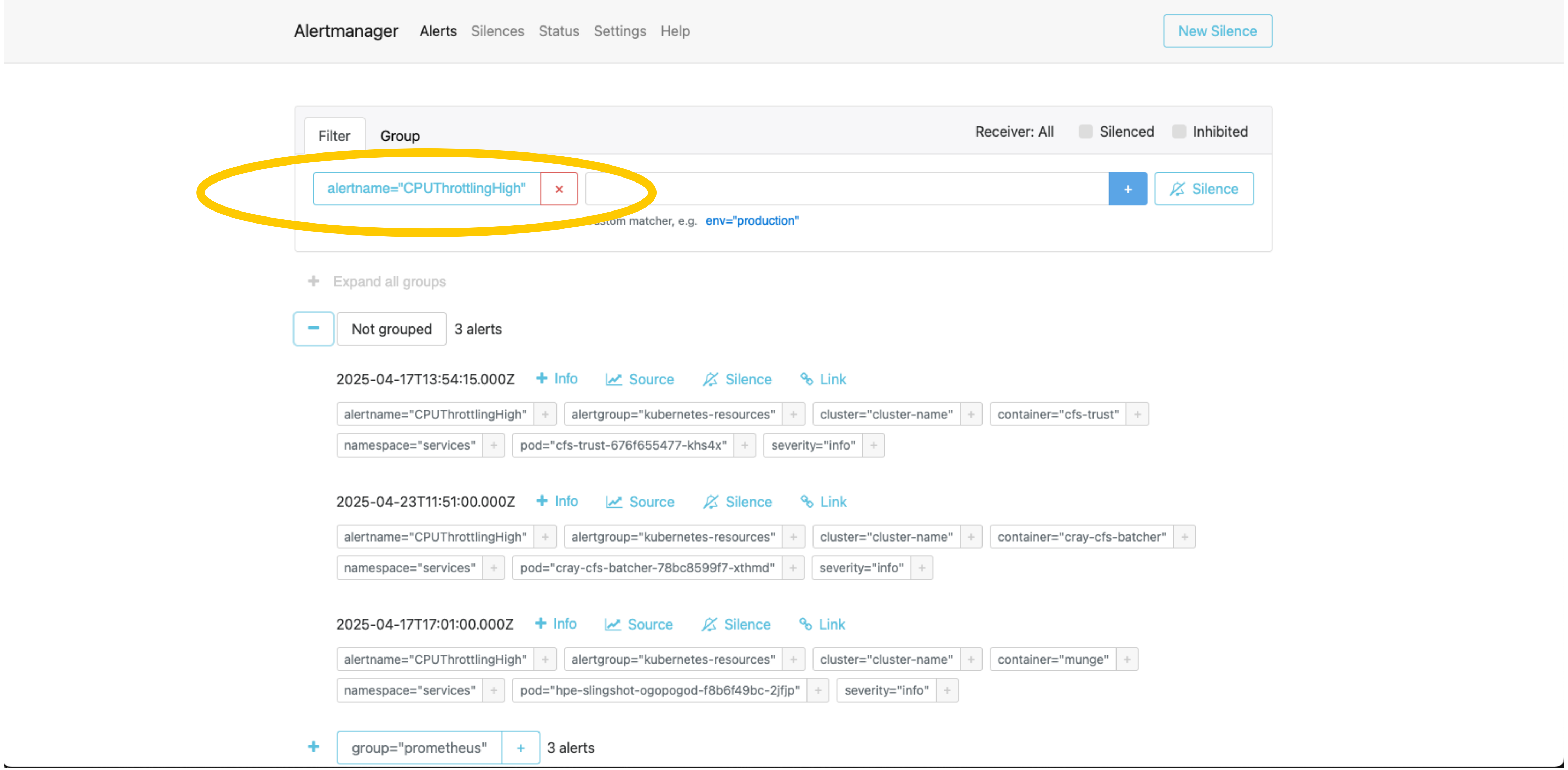
2025-04-17T14:51:15.000Z + Info Source Silence Link

alertname="etcdHighNumberOfFailedGRPCRequests" alertgroup="etcd" cluster="cluster-name" container="etcd" endpoint="client" grpc_method="LeaseKeepAlive" grpc_service="etcdserverpb.Lease" instance="10.40.0.43:2379" job="cray-hbtd-bitnami-etcd" namespace="services" pod="cray-hbtd-bitnami-etcd-0" prometheus="sysmgmt-health/vms" service="cray-hbtd-bitnami-etcd" severity="warning"

2025-04-17T14:46:30.000Z + Info Source Silence Link

https://alertmanager.cmn.SYSTEM_DOMAIN_NAME/

Alertmanager – Filter on alertname - CPUThrottlingHigh



alertname="CPUThrottlingHigh" ×

...stom matcher, e.g. env="production"

Not grouped 3 alerts

2025-04-17T13:54:15.000Z + Info Source Silence Link

alertname="CPUThrottlingHigh" + alertgroup="kubernetes-resources" + cluster="cluster-name" + container="cfs-trust" +
namespace="services" + pod="cfs-trust-676f655477-khs4x" + severity="info" +

2025-04-23T11:51:00.000Z + Info Source Silence Link

alertname="CPUThrottlingHigh" + alertgroup="kubernetes-resources" + cluster="cluster-name" + container="cray-cfs-batcher" +
namespace="services" + pod="cray-cfs-batcher-78bc8599f7-xthmd" + severity="info" +

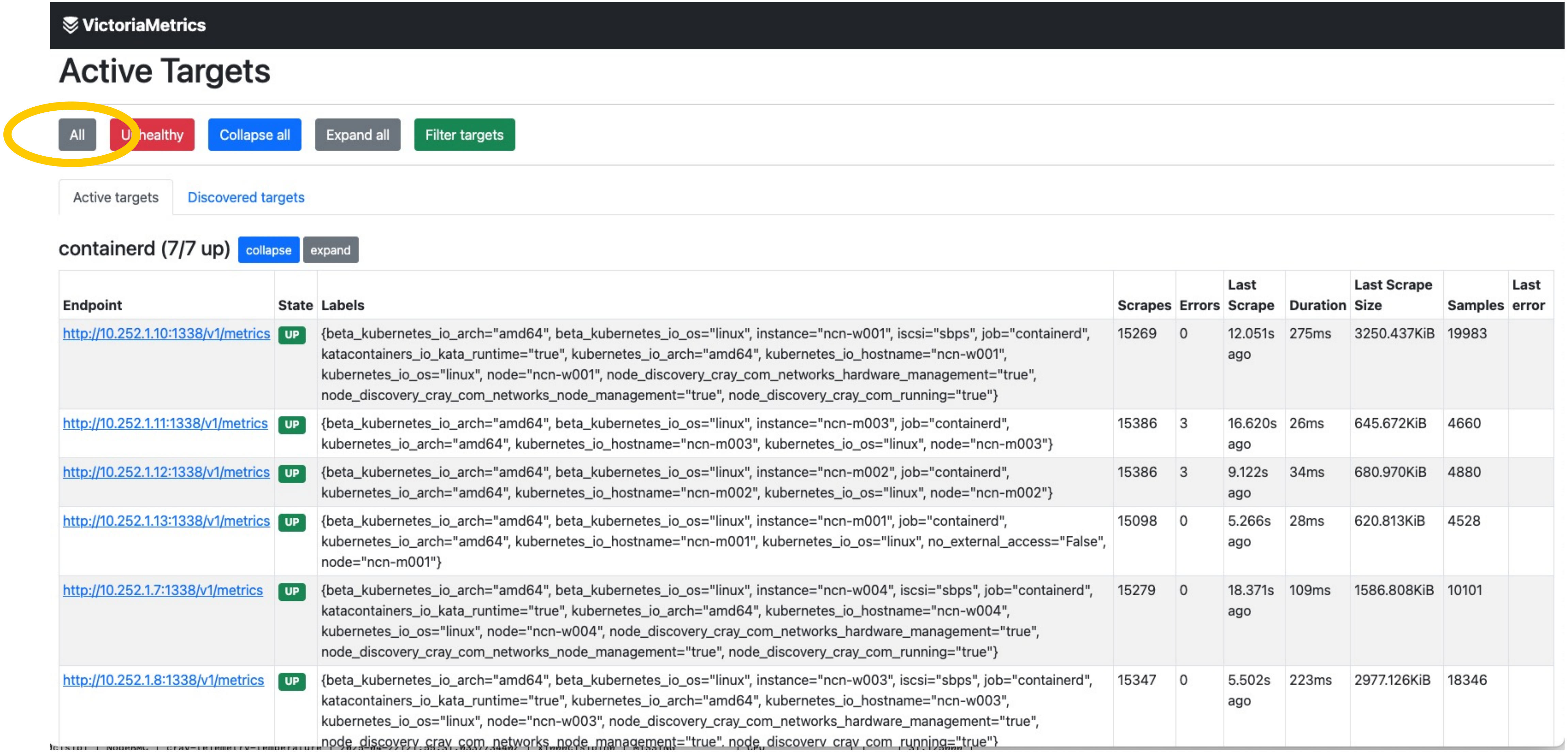
2025-04-17T17:01:00.000Z + Info Source Silence Link

alertname="CPUThrottlingHigh" + alertgroup="kubernetes-resources" + cluster="cluster-name" + container="munge" +
namespace="services" + pod="hpe-slingshot-ogopogod-f8b6f49bc-2jfjp" + severity="info" +

+ group="prometheus" + 3 alerts

https://alertmanager.cmn.SYSTEM_DOMAIN_NAME/

VictoriaMetrics Agent – Alerts (all)



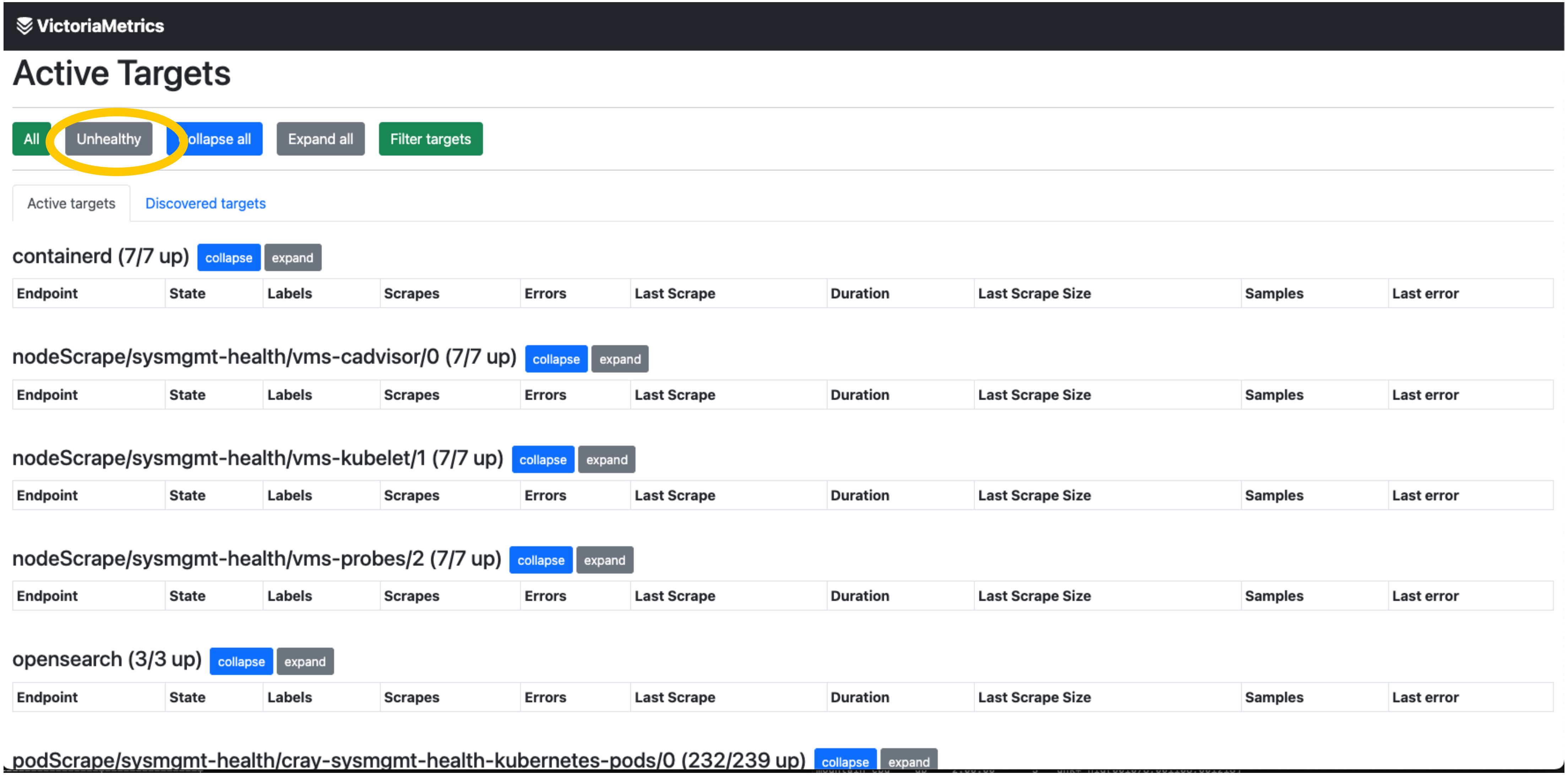
Active Targets

containerd (7/7 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error
http://10.252.1.10:1338/v1/metrics	UP	{beta_kubernetes_io_arch="amd64", beta_kubernetes_io_os="linux", instance="ncn-w001", iscsi="sbps", job="containerd", katacontainers_io_kata_runtime="true", kubernetes_io_arch="amd64", kubernetes_io_hostname="ncn-w001", kubernetes_io_os="linux", node="ncn-w001", node_discovery_cray_com_networks_hardware_management="true", node_discovery_cray_com_networks_node_management="true", node_discovery_cray_com_running="true"}	15269	0	12.051s ago	275ms	3250.437KiB	19983	
http://10.252.1.11:1338/v1/metrics	UP	{beta_kubernetes_io_arch="amd64", beta_kubernetes_io_os="linux", instance="ncn-m003", job="containerd", kubernetes_io_arch="amd64", kubernetes_io_hostname="ncn-m003", kubernetes_io_os="linux", node="ncn-m003"}	15386	3	16.620s ago	26ms	645.672KiB	4660	
http://10.252.1.12:1338/v1/metrics	UP	{beta_kubernetes_io_arch="amd64", beta_kubernetes_io_os="linux", instance="ncn-m002", job="containerd", kubernetes_io_arch="amd64", kubernetes_io_hostname="ncn-m002", kubernetes_io_os="linux", node="ncn-m002"}	15386	3	9.122s ago	34ms	680.970KiB	4880	
http://10.252.1.13:1338/v1/metrics	UP	{beta_kubernetes_io_arch="amd64", beta_kubernetes_io_os="linux", instance="ncn-m001", job="containerd", kubernetes_io_arch="amd64", kubernetes_io_hostname="ncn-m001", kubernetes_io_os="linux", no_external_access="False", node="ncn-m001"}	15098	0	5.266s ago	28ms	620.813KiB	4528	
http://10.252.1.7:1338/v1/metrics	UP	{beta_kubernetes_io_arch="amd64", beta_kubernetes_io_os="linux", instance="ncn-w004", iscsi="sbps", job="containerd", katacontainers_io_kata_runtime="true", kubernetes_io_arch="amd64", kubernetes_io_hostname="ncn-w004", kubernetes_io_os="linux", node="ncn-w004", node_discovery_cray_com_networks_hardware_management="true", node_discovery_cray_com_networks_node_management="true", node_discovery_cray_com_running="true"}	15279	0	18.371s ago	109ms	1586.808KiB	10101	
http://10.252.1.8:1338/v1/metrics	UP	{beta_kubernetes_io_arch="amd64", beta_kubernetes_io_os="linux", instance="ncn-w003", iscsi="sbps", job="containerd", katacontainers_io_kata_runtime="true", kubernetes_io_arch="amd64", kubernetes_io_hostname="ncn-w003", kubernetes_io_os="linux", node="ncn-w003", node_discovery_cray_com_networks_hardware_management="true", node_discovery_cray_com_networks_node_management="true", node_discovery_cray_com_running="true"}	15347	0	5.502s ago	223ms	2977.126KiB	18346	

https://vmagent.cmn.SYSTEM_DOMAIN_NAME/targets

VictoriaMetrics Agent – Alerts (only Unhealthy)



Active Targets

containerd (7/7 up) collapse expand

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

nodeScrape/sysmgmt-health/vms-cadvisor/0 (7/7 up) collapse expand

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

nodeScrape/sysmgmt-health/vms-kubelet/1 (7/7 up) collapse expand

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

nodeScrape/sysmgmt-health/vms-probes/2 (7/7 up) collapse expand

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

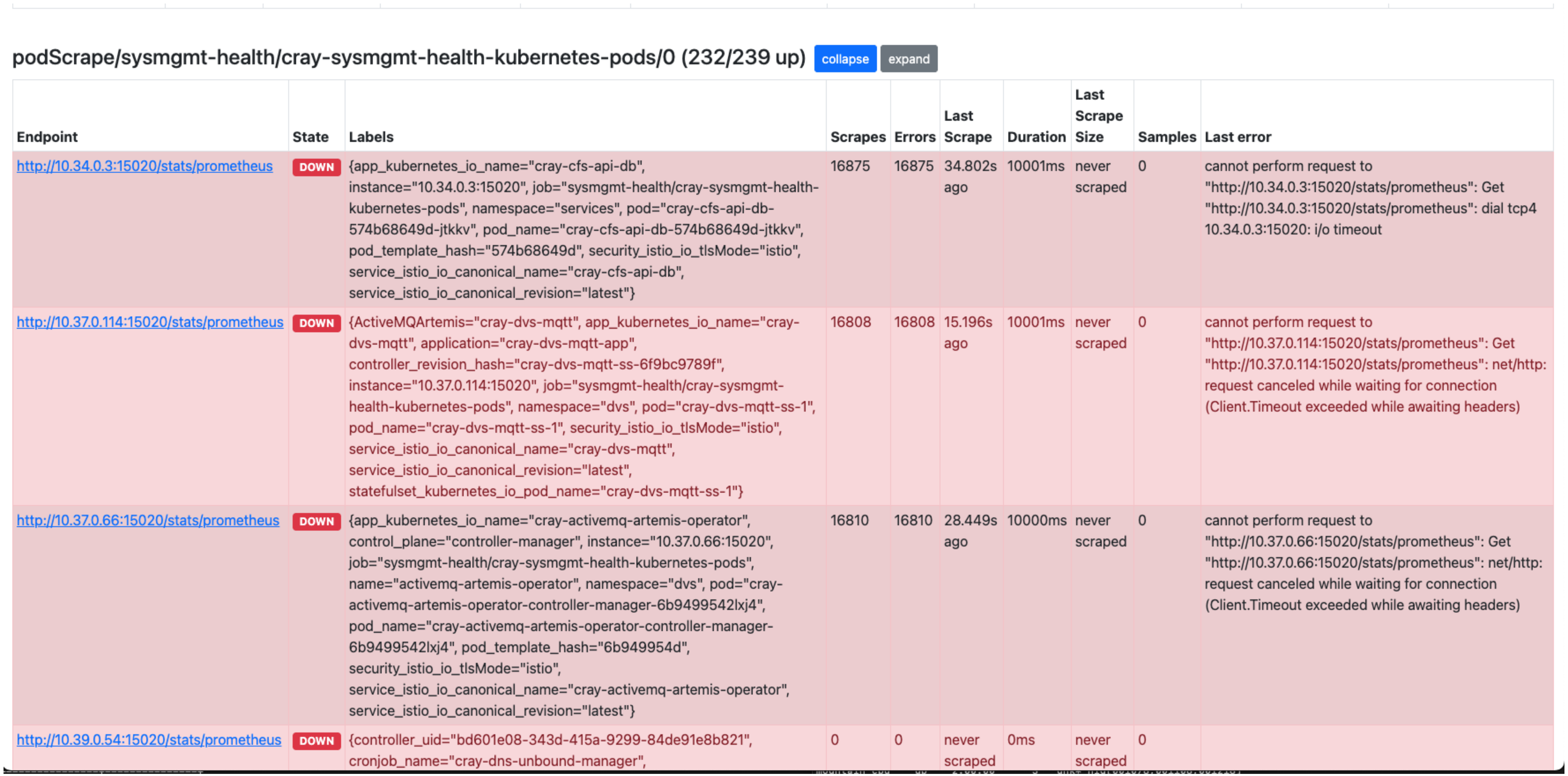
opensearch (3/3 up) collapse expand

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

podScrape/sysmgmt-health/cray-sysmgmt-health-kubernetes-pods/0 (232/239 up) collapse expand

https://vmagent.cmn.SYSTEM_DOMAIN_NAME/targets

VictoriaMetrics Agent – Alerts (Unhealthy Alert)

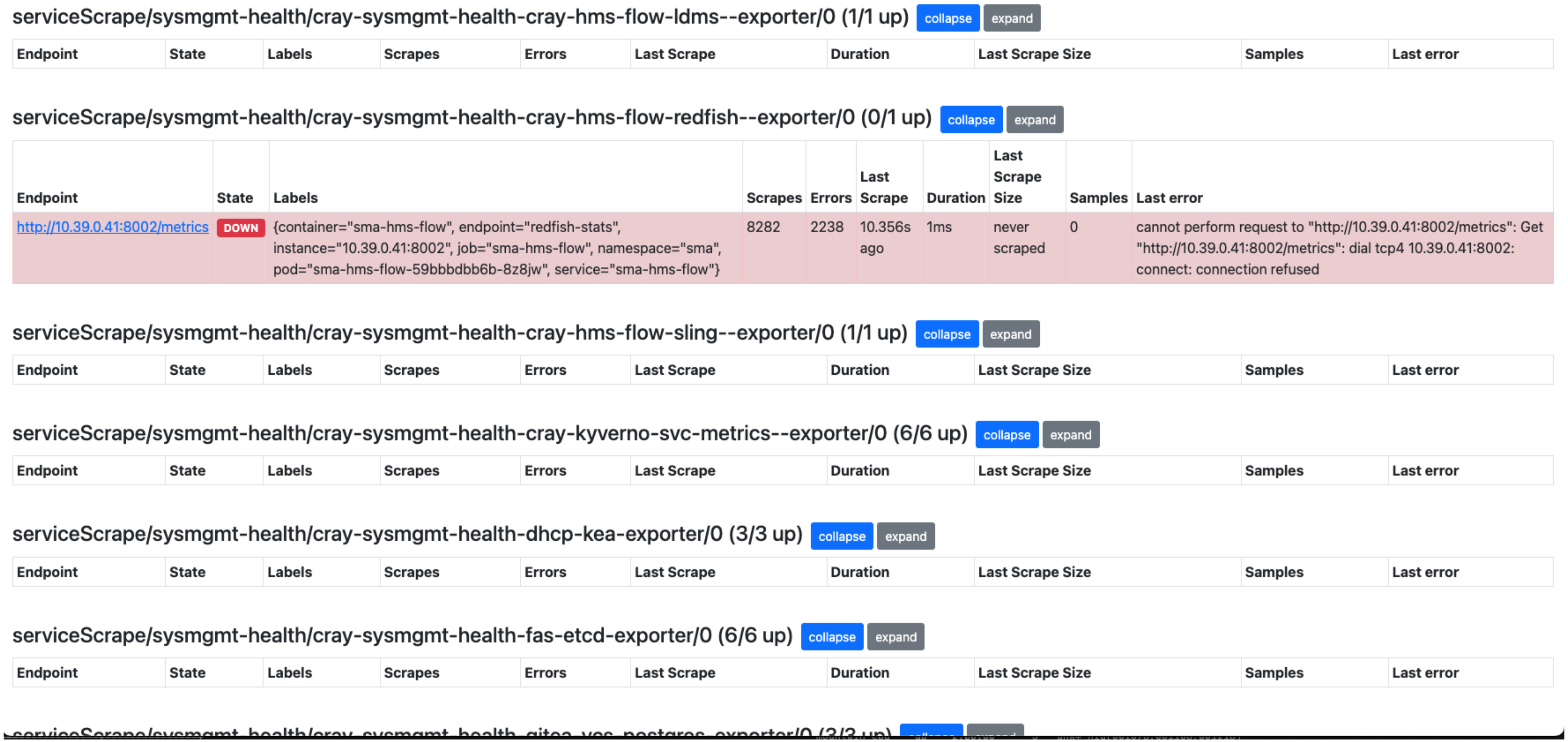


podScrape/sysmgmt-health/cray-sysmgmt-health-kubernetes-pods/0 (232/239 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error
http://10.34.0.3:15020/stats/prometheus	DOWN	{app_kubernetes_io_name="cray-cfs-api-db", instance="10.34.0.3:15020", job="sysmgmt-health/cray-sysmgmt-health-kubernetes-pods", namespace="services", pod="cray-cfs-api-db-574b68649d-jtkkv", pod_name="cray-cfs-api-db-574b68649d-jtkkv", pod_template_hash="574b68649d", security_istio_io_tlsMode="istio", service_istio_io_canonical_name="cray-cfs-api-db", service_istio_io_canonical_revision="latest"}	16875	16875	34.802s ago	10001ms	never scraped	0	cannot perform request to "http://10.34.0.3:15020/stats/prometheus": Get "http://10.34.0.3:15020/stats/prometheus": dial tcp4 10.34.0.3:15020: i/o timeout
http://10.37.0.114:15020/stats/prometheus	DOWN	{ActiveMQArtemis="cray-dvs-mqtt", app_kubernetes_io_name="cray-dvs-mqtt", application="cray-dvs-mqtt-app", controller_revision_hash="cray-dvs-mqtt-ss-6f9bc9789f", instance="10.37.0.114:15020", job="sysmgmt-health/cray-sysmgmt-health-kubernetes-pods", namespace="dvs", pod="cray-dvs-mqtt-ss-1", pod_name="cray-dvs-mqtt-ss-1", security_istio_io_tlsMode="istio", service_istio_io_canonical_name="cray-dvs-mqtt", service_istio_io_canonical_revision="latest", statefulset_kubernetes_io_pod_name="cray-dvs-mqtt-ss-1"}	16808	16808	15.196s ago	10001ms	never scraped	0	cannot perform request to "http://10.37.0.114:15020/stats/prometheus": Get "http://10.37.0.114:15020/stats/prometheus": net/http: request canceled while waiting for connection (Client.Timeout exceeded while awaiting headers)
http://10.37.0.66:15020/stats/prometheus	DOWN	{app_kubernetes_io_name="cray-activemq-artemis-operator", control_plane="controller-manager", instance="10.37.0.66:15020", job="sysmgmt-health/cray-sysmgmt-health-kubernetes-pods", name="activemq-artemis-operator", namespace="dvs", pod="cray-activemq-artemis-operator-controller-manager-6b9499542lxj4", pod_name="cray-activemq-artemis-operator-controller-manager-6b9499542lxj4", pod_template_hash="6b949954d", security_istio_io_tlsMode="istio", service_istio_io_canonical_name="cray-activemq-artemis-operator", service_istio_io_canonical_revision="latest"}	16810	16810	28.449s ago	10000ms	never scraped	0	cannot perform request to "http://10.37.0.66:15020/stats/prometheus": Get "http://10.37.0.66:15020/stats/prometheus": net/http: request canceled while waiting for connection (Client.Timeout exceeded while awaiting headers)
http://10.39.0.54:15020/stats/prometheus	DOWN	{controller_uid="bd601e08-343d-415a-9299-84de91e8b821", cronjob_name="cray-dns-unbound-manager",	0	0	never scraped	0ms	never scraped	0	

https://vmagent.cmn.SYSTEM_DOMAIN_NAME/targets

VictoriaMetrics Agent – Alerts (Unhealthy sma-hms-flow)



serviceScrape/sysmgmt-health/cray-sysmgmt-health-cray-hms-flow-ldms--exporter/0 (1/1 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

serviceScrape/sysmgmt-health/cray-sysmgmt-health-cray-hms-flow-redfish--exporter/0 (0/1 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error
http://10.39.0.41:8002/metrics	DOWN	{container="sma-hms-flow", endpoint="redfish-stats", instance="10.39.0.41:8002", job="sma-hms-flow", namespace="sma", pod="sma-hms-flow-59bbbdbb6b-8z8jw", service="sma-hms-flow"}	8282	2238	10.356s ago	1ms	never scraped	0	cannot perform request to "http://10.39.0.41:8002/metrics": Get "http://10.39.0.41:8002/metrics": dial tcp4 10.39.0.41:8002: connect: connection refused

serviceScrape/sysmgmt-health/cray-sysmgmt-health-cray-hms-flow-sling--exporter/0 (1/1 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

serviceScrape/sysmgmt-health/cray-sysmgmt-health-cray-kyverno-svc-metrics--exporter/0 (6/6 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

serviceScrape/sysmgmt-health/cray-sysmgmt-health-dhcp-kea-exporter/0 (3/3 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

serviceScrape/sysmgmt-health/cray-sysmgmt-health-fas-etcd-exporter/0 (6/6 up)

Endpoint	State	Labels	Scrapes	Errors	Last Scrape	Duration	Last Scrape Size	Samples	Last error

serviceScrape/sysmgmt-health/cray-sysmgmt-health-gitea-vcs-postgres-exporter/0 (3/3 up)

https://vmagent.cmn.SYSTEM_DOMAIN_NAME/targets

Viewing Alerts

- Alerting
  - View alerts from the command line

    ```
    cm health alertman
    ```
  - View alerts with opensearch dashboard

    ```
    https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME
    ```
  - Default Alerts: CDU, Slingshot Cassini, Slingshot Port Flap, Slingshot Rosetta Error, Slingshot FabricHealthTelemetry Events

- Alerta removed from CSM 1.6

cm health alertman

```
ncn# cm health alertman -s

+---------------+--------+
| Alert Status  | Count  |
+---------------+--------+
| Critical      | 76     |
| Warnings      | 269    |
| Active        | 351    |
+---------------+--------+
```

```
+------------------+-----------+------------------------------+
| Group            | Severity  | Alerts                       |
+------------------+-----------+------------------------------+
| compute          | ok        | critical : 0, warning : 0    |
| fabric           | ok        | critical : 0, warning : 0    |
| slingshothsn     | ok        | critical : 0, warning : 0    |
| slingshotswitch  | ok        | critical : 0, warning : 0    |
| prometheus       | critical  | critical : 45, warning : 126 |
| aiops            | ok        | critical : 0, warning : 0    |
| crayalerts       | ok        | critical : 0, warning : 0    |
| cooldev          | ok        | critical : 0, warning : 0    |
•  +------------------+-----------+------------------------------+
```

```
ncn# cm health alertman query
ID        STATUS    SEVERITY   GROUP    ENV           SERVICE      RESOURCE            EVENT          VALUE  DESCRIPTION                               DUPL  LAST RECEIVED
--------  --------  ---------- -------  ------------  -----------  ------------------  ------------  -------  ----------------------------------------  ----  -------------------
2809d804  open      critical   compute  x3700c0r41b0  SensorEvent  dmtf.redfish_event  PSU1-Voltage        0  Sensor _PSU1 Voltage_ reading of 0 _V_ is  0     2024/03/05 19:13:21
                                                                                                            below the 11.16 lower critical threshold.
85082bef  open      critical   compute  x3700c0r39b0  SensorEvent  dmtf.redfish_event  PSU1-Voltage        0  Sensor _PSU1 Voltage_ reading of 0 _V_ is  0.    2024/03/05 19:17:35
                                                                                                            below the 11.16 lower critical threshold.
```

- Manage alerts from many sources: Alertmanager, Monasca, Slingshot
  - Looks for events in the data
  - Constantly analyzes each event
  - Alerts the user regarding the event
  - Stores the event in the alert dashboard
- Manage the life cycle of alerts
  - Retrieve alerts
  - Process alerts
  - Close alerts
  - Disable during maintenance periods and re-enable after maintenance ends

Check sensors

- Obtain sensor readings from BMCs (ChassisBMC, NodeBMC, RouterBMC)
  - Limit the telemetry topics queried to the Kafka topics listed
  - The default is to query all topics:
    - cray-telemetry-temperature, cray-telemetry-voltage, cray-telemetry-power, cray-telemetry-energy, cray-telemetry-fan, cray-telemetry-pressure

```
ncn-m# sat sensors -x x1003c2s6b1 -t NodeBMC -b 2 --timeout 10 --topic cray-telemetry-temperature
Telemetry data being collected for x1003c2s6b1
Please be patient...
Waiting for metrics for all requested xnames from cray-telemetry-temperature.
Receiving metrics from stream: cray-telemetry-temperature...
Telemetry data received from cray-telemetry-temperature for all requested xnames.
+-------------+---------+----------------------------+--------------------------------+-------------+----------+-----------+---------+------+
| xname       | Type    | Topic                      | Timestamp                      | Location | Parental Context | Physical Context | Index |Value |
+-------------+---------+----------------------------+--------------------------------+-------------+----------+-----------+---------+------+
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:57.079525696Z | x1003c2s6b1n0 | Chassis | VoltageRegulator | 0 | 55.4 |
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:56.585058025Z | x1003c2s6b1n0 | Chassis | VoltageRegulator | 2 | 45.8 |
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:57.081500532Z | x1003c2s6b1n1 | Chassis | VoltageRegulator | 0 | 51.2 |
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:56.580577726Z | x1003c2s6b1n1 | Chassis | VoltageRegulator | 2 | 45.8 |
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:57.072975044Z | x1003c2s6b1n0 | MISSING | CPU       | 0    | 30.875000 |
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:57.072913765Z | x1003c2s6b1n0 | MISSING | CPU       | 1    | 26.500000 |
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:57.073033042Z | x1003c2s6b1n1 | MISSING | CPU       | 0    | 29.750000 |
| x1003c2s6b1 | NodeBMC | cray-telemetry-temperature | 2022-04-01T18:17:57.073074561Z | x1003c2s6b1n1 | MISSING | CPU       | 1    | 27.500000 |
+-------------+---------+----------------------------+--------------------------------+-------------+----------+-----------+---------+------+
```

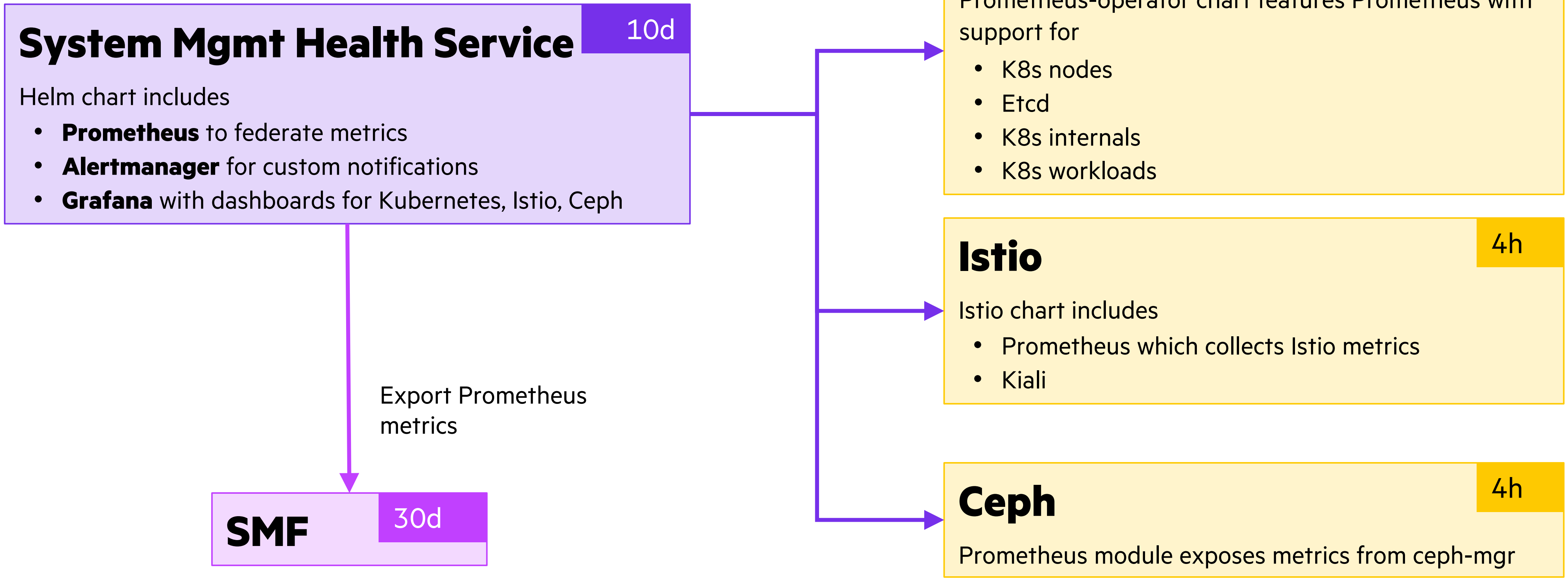
System management health monitoring

- System management health
- Grafana
- Sample Dashboards
- Kiali

System Management Health Service

Is the system healthy?

- Independent from the System Monitoring Framework (SMA)
    - Abbreviated on later slides
        - System Mgmt Health
- Does not monitor computes!



System Mgmt Health Service

10d

Helm chart includes

- **Prometheus** to federate metrics
- **Alertmanager** for custom notifications
- **Grafana** with dashboards for Kubernetes, Istio, Ceph

Export Prometheus metrics

SMF

30d

Kubernetes

4h

Prometheus-operator chart features Prometheus with support for

- K8s nodes
- Etcd
- K8s internals
- K8s workloads

Istio

4h

Istio chart includes

- Prometheus which collects Istio metrics
- Kiali

Ceph

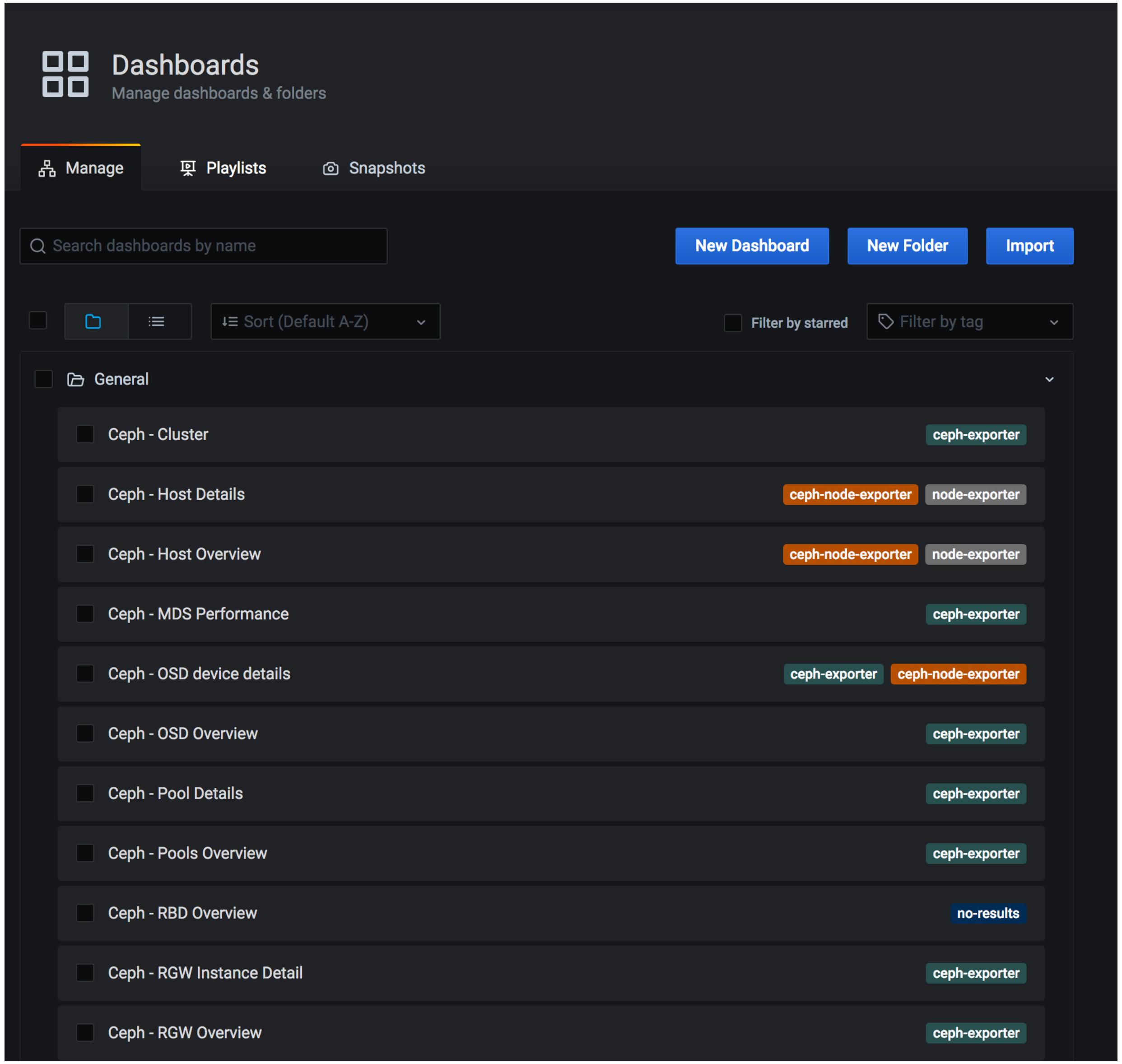
4h

Prometheus module exposes metrics from ceph-mgr

System Mgmt Health Grafana Dashboards

- Uses Keycloak authentication/authorization
- Secured with TLS sharing cluster certificate bundle
- Over 60 included dashboards
  - CANU
  - Ceph
  - CoreDNS
  - Etcd
  - ETCD Clusters
  - FluentBit
  - Goss tests
  - iSCSI
  - Istio
  - IUF
  - Kafka
  - Kea-dhcp
  - Kubernetes
  - Kyverno
  - Logstash
  - Node Exporter
  - Nodes
  - OpenSearch
  - PostgreSQL
  - Prometheus
  - SMA-Flow
  - SMARTMON
  - SNMP
  - VictoriaMetrics
  - Zookeeper

https://grafana.cmn.SYSTEM_DOMAIN_NAME/dashboards

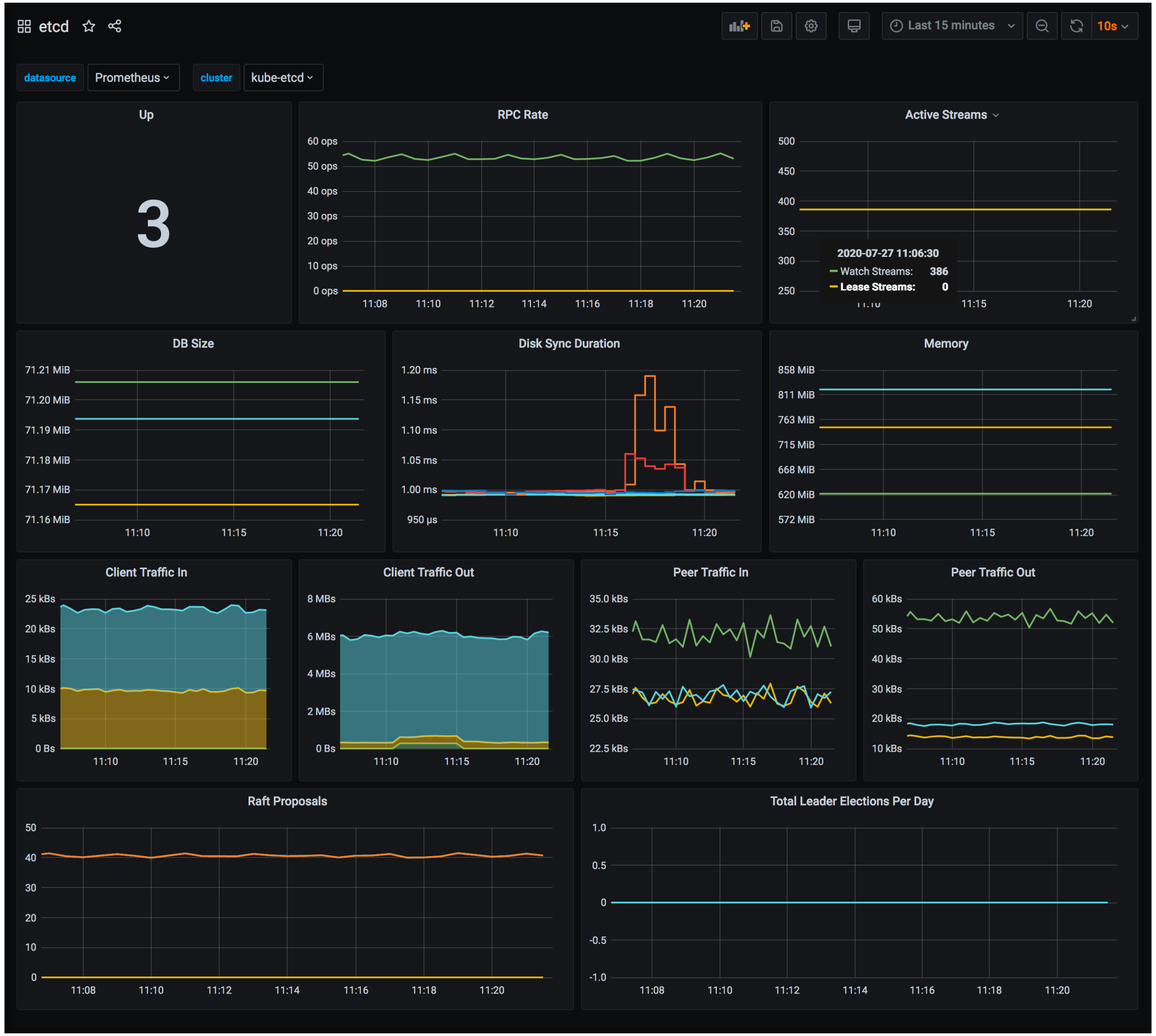


General

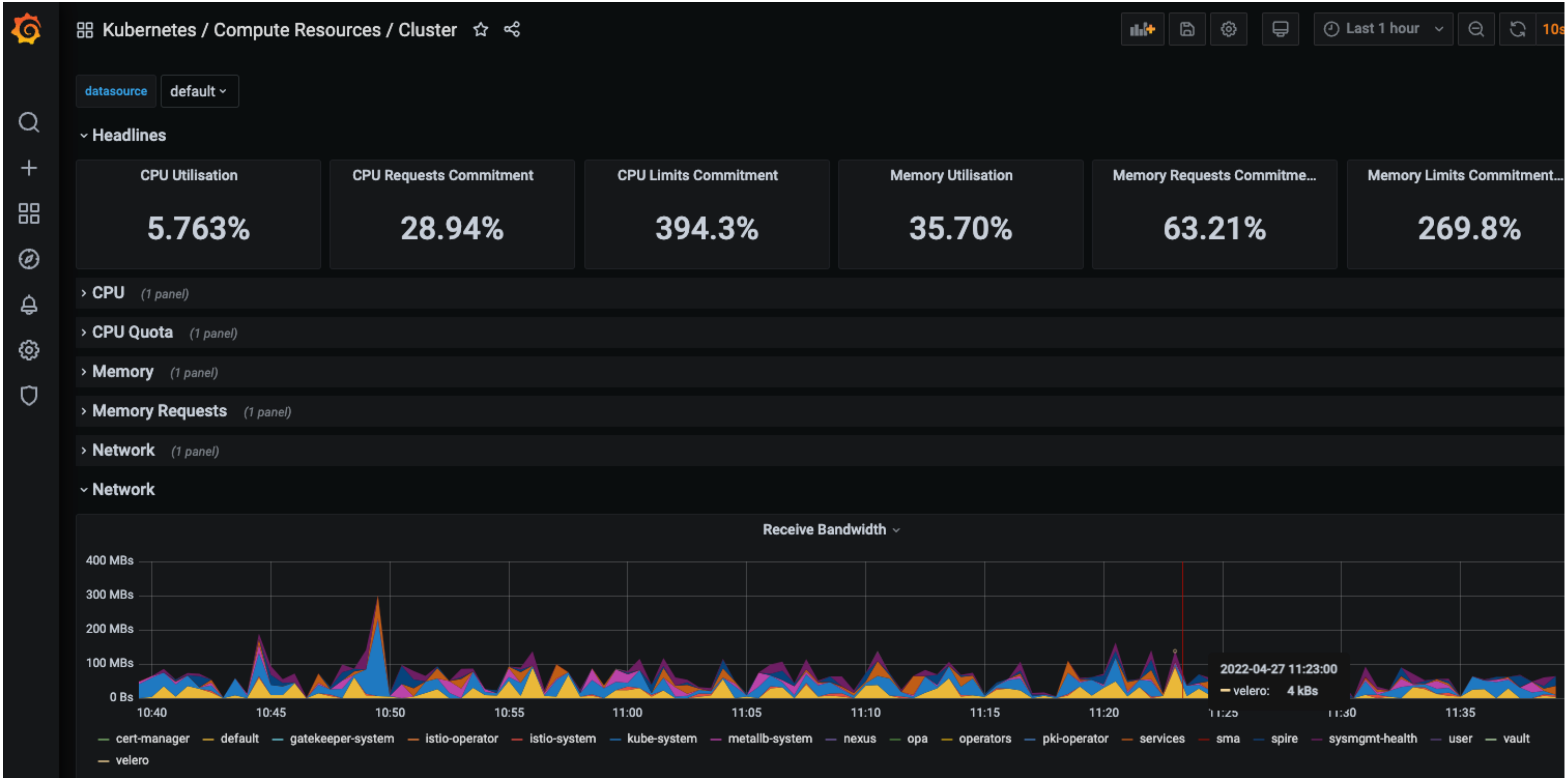
Ceph - Cluster	ceph-exporter
Ceph - Host Details	ceph-node-exporter node-exporter
Ceph - Host Overview	ceph-node-exporter node-exporter
Ceph - MDS Performance	ceph-exporter
Ceph - OSD device details	ceph-exporter ceph-node-exporter
Ceph - OSD Overview	ceph-exporter
Ceph - Pool Details	ceph-exporter
Ceph - Pools Overview	ceph-exporter
Ceph - RBD Overview	no-results
Ceph - RGW Instance Detail	ceph-exporter
Ceph - RGW Overview	ceph-exporter

System Mgmt Health Grafana Dashboards: ETCD

- Nodes up (quorum)
- RPC Rate
- Active Streams
- DB Size
- Disk Sync Duration
- Memory
- Client Traffic in
- Client Traffic Out
- Peer Traffic In
- Peer Traffic Out
- Raft proposals
- Total Leader Elections Per day



System Mgmt Health Grafana Dashboards: Kubernetes Cluster



Headlines

CPU Utilisation	CPU Requests Commitment	CPU Limits Commitment	Memory Utilisation	Memory Requests Commitme...	Memory Limits Commitment...
5.763%	28.94%	394.3%	35.70%	63.21%	269.8%

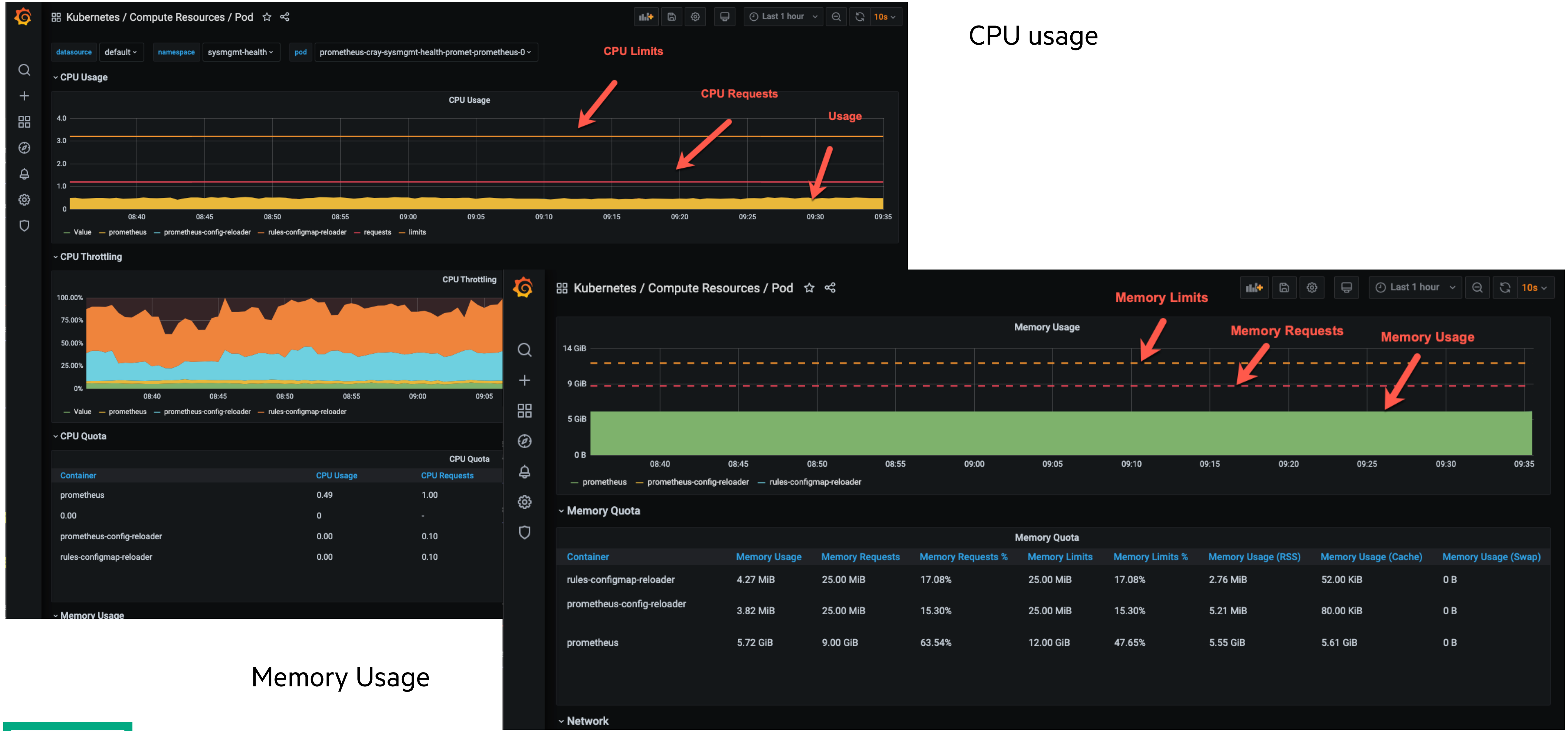
- CPU (1 panel)
- CPU Quota (1 panel)
- Memory (1 panel)
- Memory Requests (1 panel)
- Network (1 panel)

Network

Receive Bandwidth

System Mgmt Health Grafana : Kubernetes pod Requests and Limits

CPU usage



CPU Quota

Container	CPU Usage	CPU Requests
prometheus	0.49	1.00
0.00	0	-
prometheus-config-reloader	0.00	0.10
rules-configmap-reloader	0.00	0.10

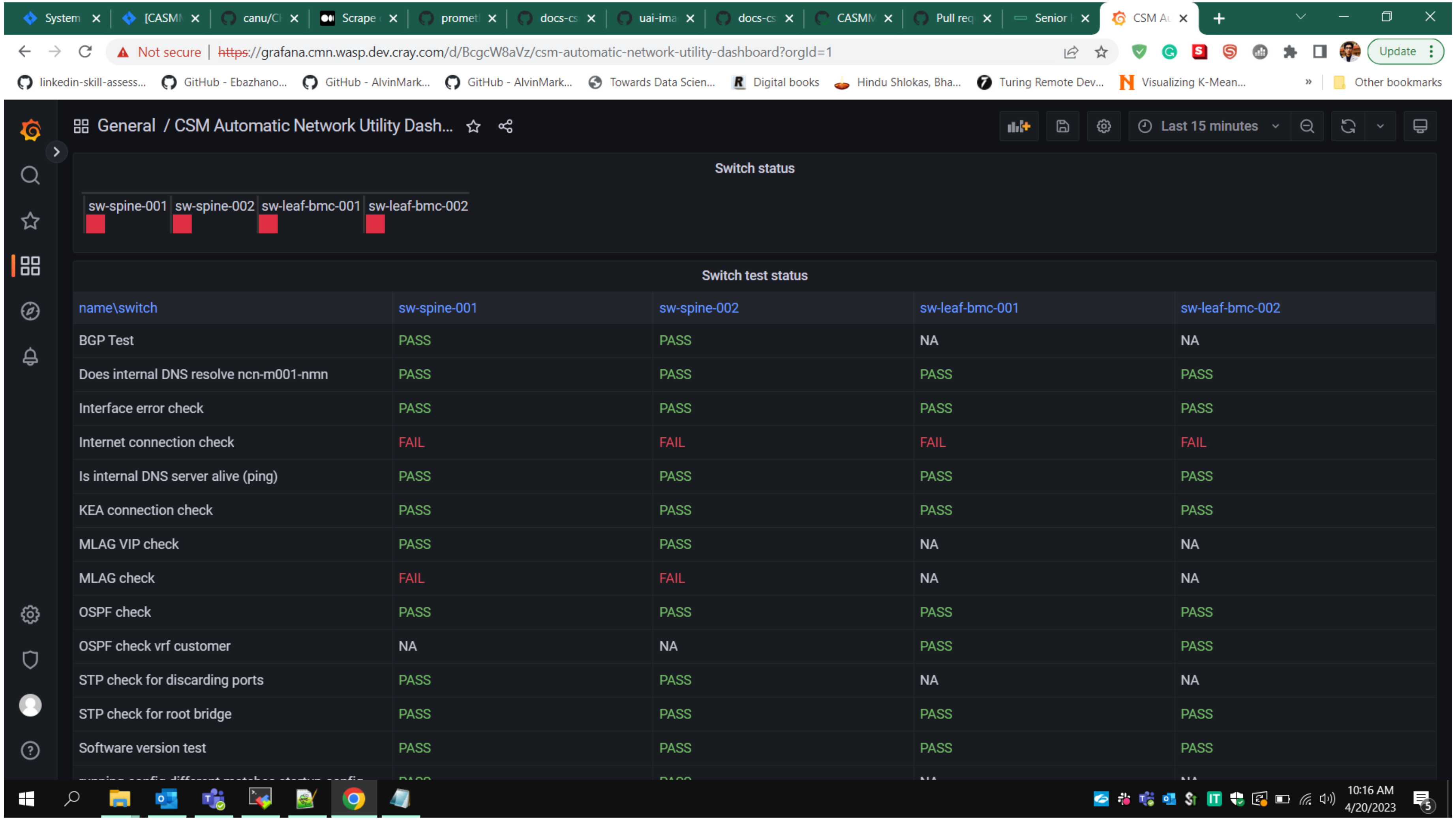
Memory Quota

Container	Memory Usage	Memory Requests	Memory Requests %	Memory Limits	Memory Limits %	Memory Usage (RSS)	Memory Usage (Cache)	Memory Usage (Swap)
rules-configmap-reloader	4.27 MiB	25.00 MiB	17.08%	25.00 MiB	17.08%	2.76 MiB	52.00 KiB	0 B
prometheus-config-reloader	3.82 MiB	25.00 MiB	15.30%	25.00 MiB	15.30%	5.21 MiB	80.00 KiB	0 B
prometheus	5.72 GiB	9.00 GiB	63.54%	12.00 GiB	47.65%	5.55 GiB	5.61 GiB	0 B

Memory Usage

System Mgmt Health Grafana Dashboards: CANU Dashboard

- CSM Automatic Network Utility (CANU) Dashboard
  - Shows results from CANU tests on management network switches
- VictoriaMetrics
  - Search for canu_test


Not secure | https://grafana.cmn.wasp.dev.cray.com/d/BcgcW8aVz/csm-automatic-network-utility-dashboard?orgId=1

General / CSM Automatic Network Utility Dash...

Last 15 minutes

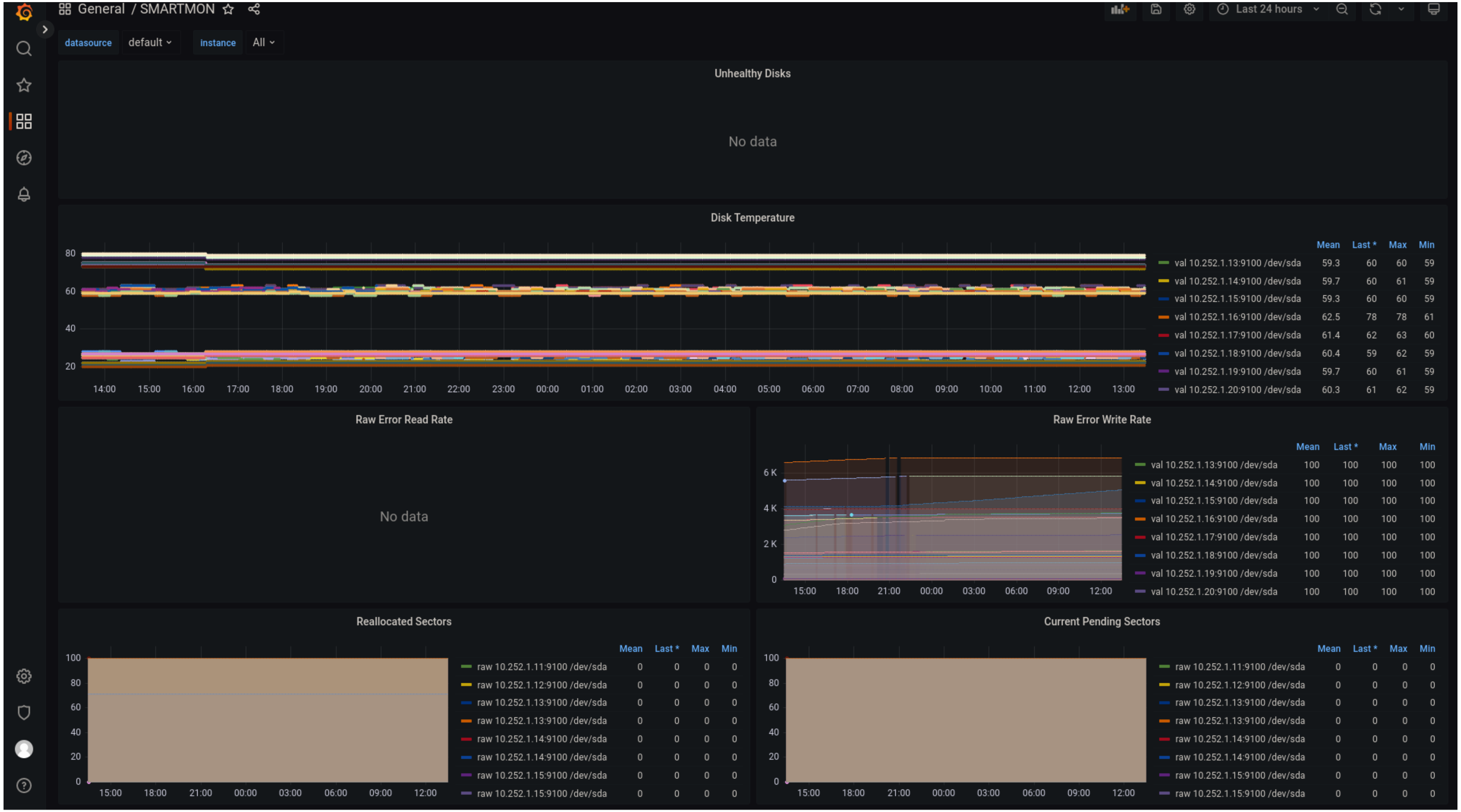
Switch status

sw-spine-001	sw-spine-002	sw-leaf-bmc-001	sw-leaf-bmc-002
🟥	🟥	🟥	🟥

Switch test status

name\switch	sw-spine-001	sw-spine-002	sw-leaf-bmc-001	sw-leaf-bmc-002
BGP Test	PASS	PASS	NA	NA
Does internal DNS resolve ncn-m001-nmn	PASS	PASS	PASS	PASS
Interface error check	PASS	PASS	PASS	PASS
Internet connection check	FAIL	FAIL	FAIL	FAIL
Is internal DNS server alive (ping)	PASS	PASS	PASS	PASS
KEA connection check	PASS	PASS	PASS	PASS
MLAG VIP check	PASS	PASS	NA	NA
MLAG check	FAIL	FAIL	NA	NA
OSPF check	PASS	PASS	PASS	PASS
OSPF check vrf customer	NA	NA	PASS	PASS
STP check for discarding ports	PASS	PASS	NA	NA
STP check for root bridge	PASS	PASS	PASS	PASS
Software version test	PASS	PASS	PASS	PASS
running config different matches startup config	PASS	PASS	NA	NA

System Mgmt Health Grafana Dashboards: SMARTMON



Unhealthy Disks

No data

Disk Temperature

	Mean	Last *	Max	Min
🟩 val 10.252.1.13:9100 /dev/sda	59.3	60	60	59
🟨 val 10.252.1.14:9100 /dev/sda	59.7	60	61	59
🟦 val 10.252.1.15:9100 /dev/sda	59.3	60	60	59
🟧 val 10.252.1.16:9100 /dev/sda	62.5	78	78	61
🟥 val 10.252.1.17:9100 /dev/sda	61.4	62	63	60
🟦 val 10.252.1.18:9100 /dev/sda	60.4	59	62	59
🟪 val 10.252.1.19:9100 /dev/sda	59.7	60	61	59
🟪 val 10.252.1.20:9100 /dev/sda	60.3	61	62	59

Raw Error Read Rate

No data

Raw Error Write Rate

	Mean	Last *	Max	Min
🟩 val 10.252.1.13:9100 /dev/sda	100	100	100	100
🟨 val 10.252.1.14:9100 /dev/sda	100	100	100	100
🟦 val 10.252.1.15:9100 /dev/sda	100	100	100	100
🟧 val 10.252.1.16:9100 /dev/sda	100	100	100	100
🟥 val 10.252.1.17:9100 /dev/sda	100	100	100	100
🟦 val 10.252.1.18:9100 /dev/sda	100	100	100	100
🟪 val 10.252.1.19:9100 /dev/sda	100	100	100	100
🟪 val 10.252.1.20:9100 /dev/sda	100	100	100	100

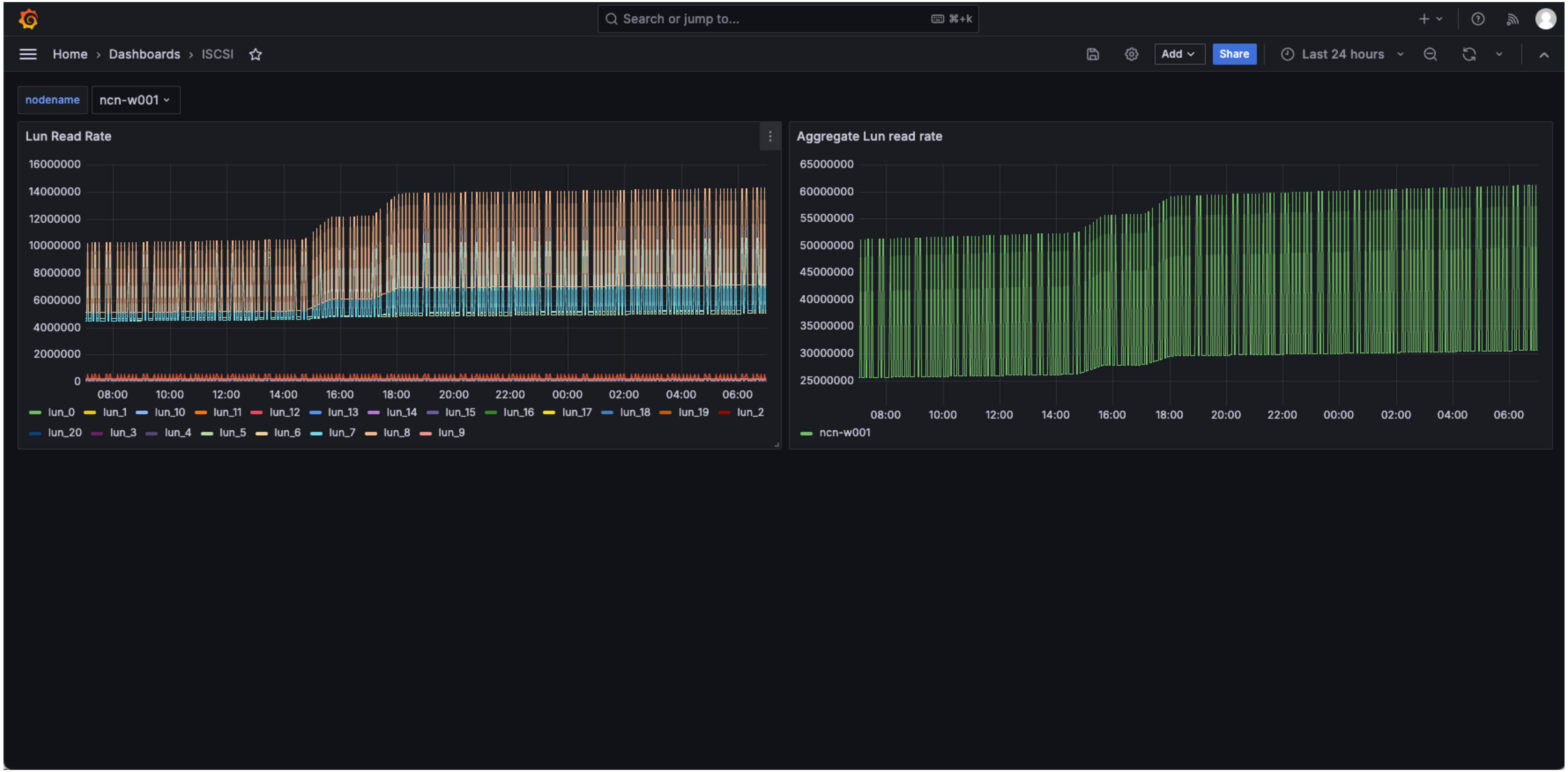
Reallocated Sectors

	Mean	Last *	Max	Min
🟩 raw 10.252.1.11:9100 /dev/sda	0	0	0	0
🟨 raw 10.252.1.12:9100 /dev/sda	0	0	0	0
🟦 raw 10.252.1.13:9100 /dev/sda	0	0	0	0
🟧 raw 10.252.1.13:9100 /dev/sda	0	0	0	0
🟥 raw 10.252.1.14:9100 /dev/sda	0	0	0	0
🟦 raw 10.252.1.14:9100 /dev/sda	0	0	0	0
🟪 raw 10.252.1.15:9100 /dev/sda	0	0	0	0
🟪 raw 10.252.1.15:9100 /dev/sda	0	0	0	0

Current Pending Sectors

	Mean	Last *	Max	Min
🟩 raw 10.252.1.11:9100 /dev/sda	0	0	0	0
🟨 raw 10.252.1.12:9100 /dev/sda	0	0	0	0
🟦 raw 10.252.1.13:9100 /dev/sda	0	0	0	0
🟧 raw 10.252.1.13:9100 /dev/sda	0	0	0	0
🟥 raw 10.252.1.14:9100 /dev/sda	0	0	0	0
🟦 raw 10.252.1.14:9100 /dev/sda	0	0	0	0
🟪 raw 10.252.1.15:9100 /dev/sda	0	0	0	0
🟪 raw 10.252.1.15:9100 /dev/sda	0	0	0	0

System Mgmt Health Grafana Dashboards: iSCSI

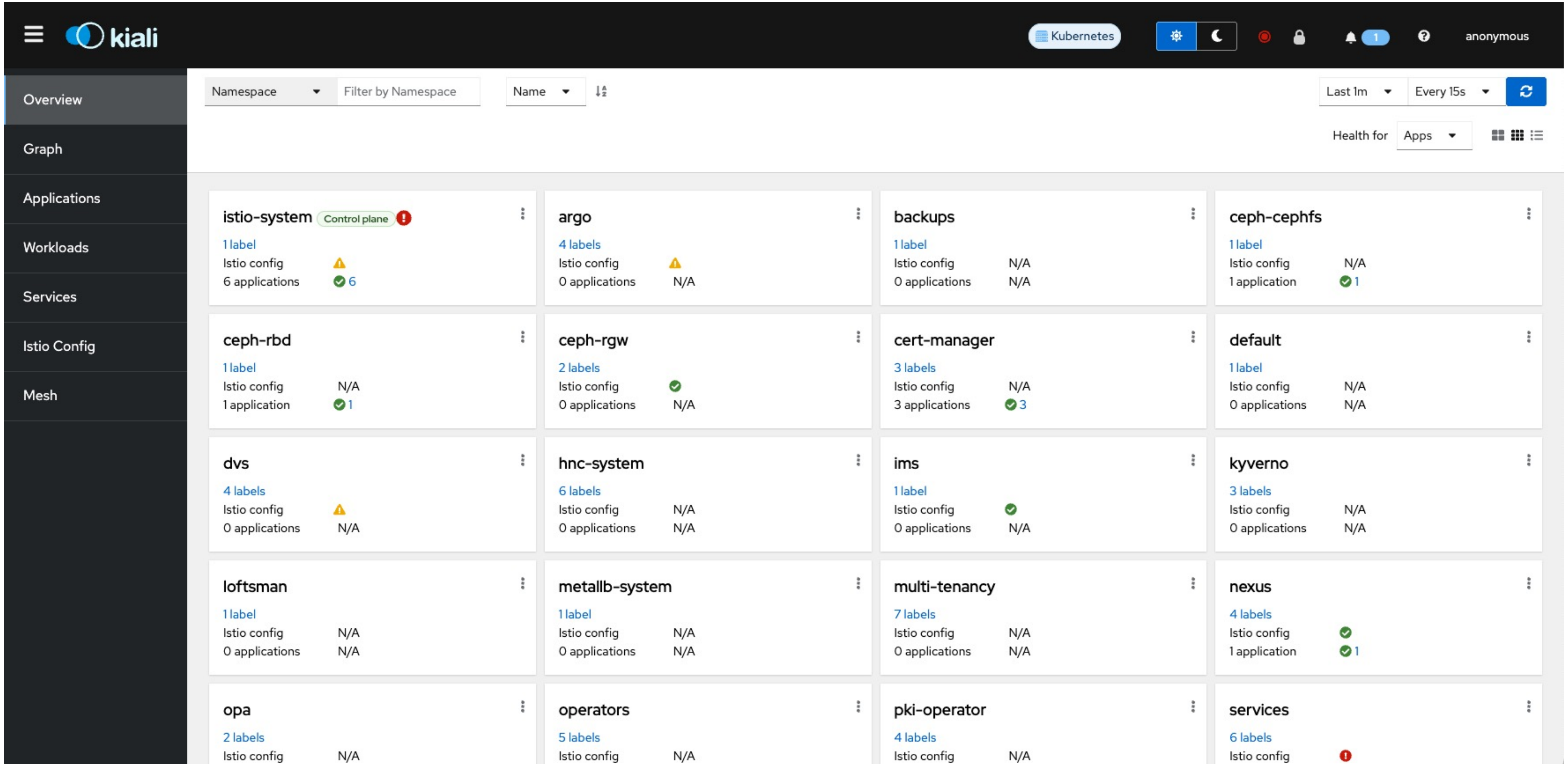


Kiali

- Kiali provides real-time introspection into the Istio service mesh using metrics from prometheus-istio
  - Observability console for Istio with service mesh configuration and validation capabilities
  - Helps you understand the structure and health of your service mesh by monitoring traffic flow to infer the topology and report errors
  - Provides detailed metrics and a basic Grafana integration, which can be used for advanced queries
    - https://kiali-istio.SYSTEM_DOMAIN_NAME/
  - Documentation
    - https://kiali.io/documentation/

Kiali – Overview

- Identify namespaces with issues
- Summary of configuration health, component health and request traffic health
- Offers various filter, sort and presentation options



istio-system Control plane ❗
1 label
Istio config ⚠
6 applications ✓ 6

argo
4 labels
Istio config ⚠
0 applications N/A

backups
1 label
Istio config N/A
0 applications N/A

ceph-cephfs
1 label
Istio config N/A
1 application ✓ 1

ceph-rbd
1 label
Istio config N/A
1 application ✓ 1

ceph-rgw
2 labels
Istio config ✓
0 applications N/A

cert-manager
3 labels
Istio config N/A
3 applications ✓ 3

default
1 label
Istio config N/A
0 applications N/A

dvs
4 labels
Istio config ⚠
0 applications N/A

hnc-system
6 labels
Istio config N/A
0 applications N/A

ims
1 label
Istio config ✓
0 applications N/A

kyverno
3 labels
Istio config N/A
0 applications N/A

loftsman
1 label
Istio config N/A
0 applications N/A

metallb-system
1 label
Istio config N/A
0 applications N/A

multi-tenancy
7 labels
Istio config N/A
0 applications N/A

nexus
4 labels
Istio config ✓
1 application ✓ 1

opa
2 labels
Istio config N/A

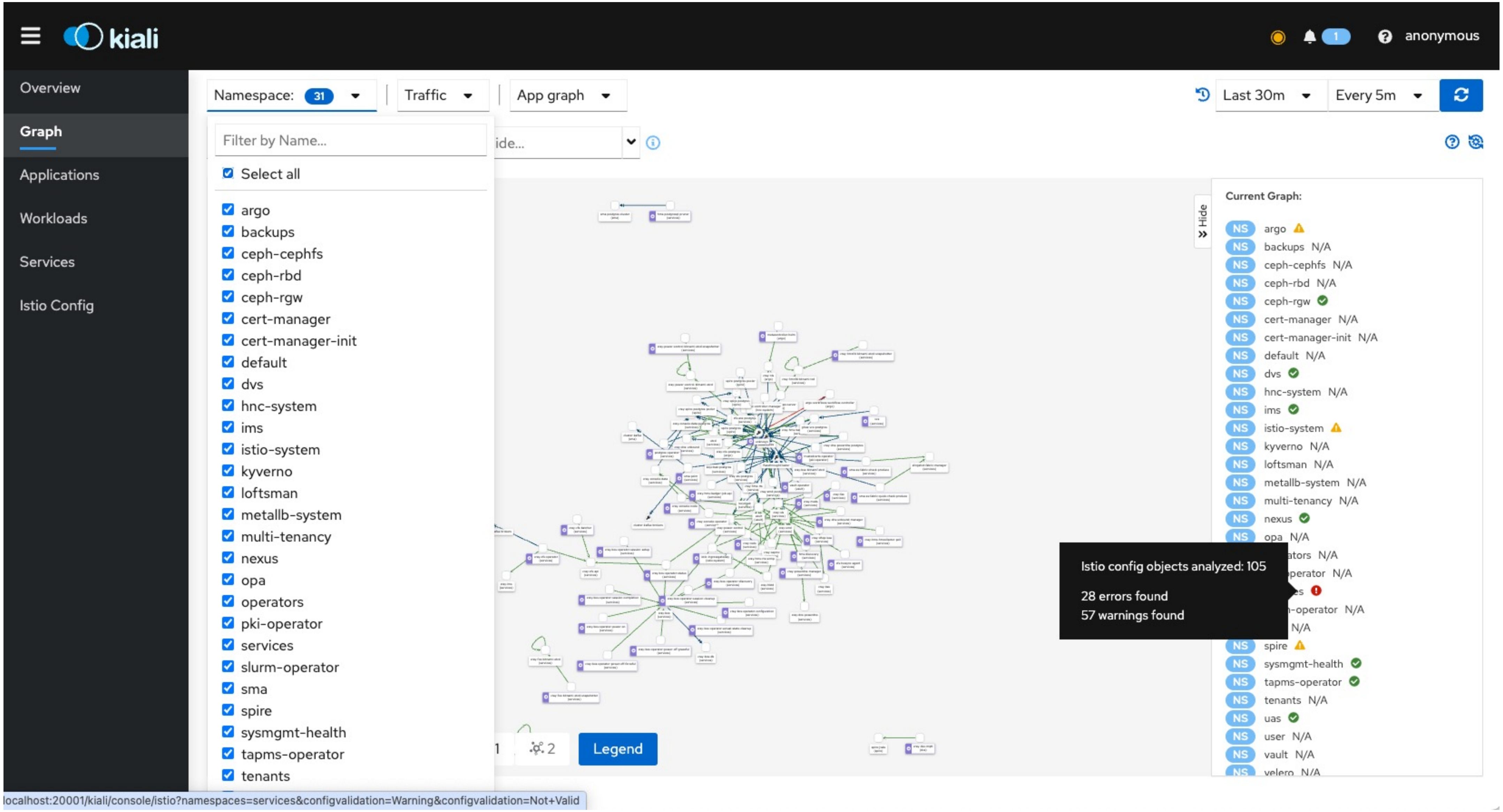
operators
5 labels
Istio config N/A

pki-operator
4 labels
Istio config N/A

services
6 labels
Istio config ❗

https://kiali.cmn.SYSTEM_DOMAIN_NAME

Kiali – Graph health



Namespace: 31

Last 30m

Every 5m

Filter by Name...

☑ Select all

☑ argo
☑ backups
☑ ceph-cephfs
☑ ceph-rbd
☑ ceph-rgw
☑ cert-manager
☑ cert-manager-init
☑ default
☑ dvs
☑ hnc-system
☑ ims
☑ istio-system
☑ kyverno
☑ loftsman
☑ metallb-system
☑ multi-tenancy
☑ nexus
☑ opa
☑ operators
☑ pki-operator
☑ services
☑ slurm-operator
☑ sma
☑ spire
☑ sysmgmt-health
☑ tapms-operator
☑ tenants

Current Graph:

NS argo ⚠
NS backups N/A
NS ceph-cephfs N/A
NS ceph-rbd N/A
NS ceph-rgw ✓
NS cert-manager N/A
NS cert-manager-init N/A
NS default N/A
NS dvs ✓
NS hnc-system N/A
NS ims ✓
NS istio-system ⚠
NS kyverno N/A
NS loftsman N/A
NS metallb-system N/A
NS multi-tenancy N/A
NS nexus ✓
NS opa N/A
ators N/A
perator N/A
es ❗
-operator N/A
N/A
NS spire ⚠
NS sysmgmt-health ✓
NS tapms-operator ✓
NS tenants N/A
NS uas ✓
NS user N/A
NS vault N/A
NS velero N/A

Istio config objects analyzed: 105

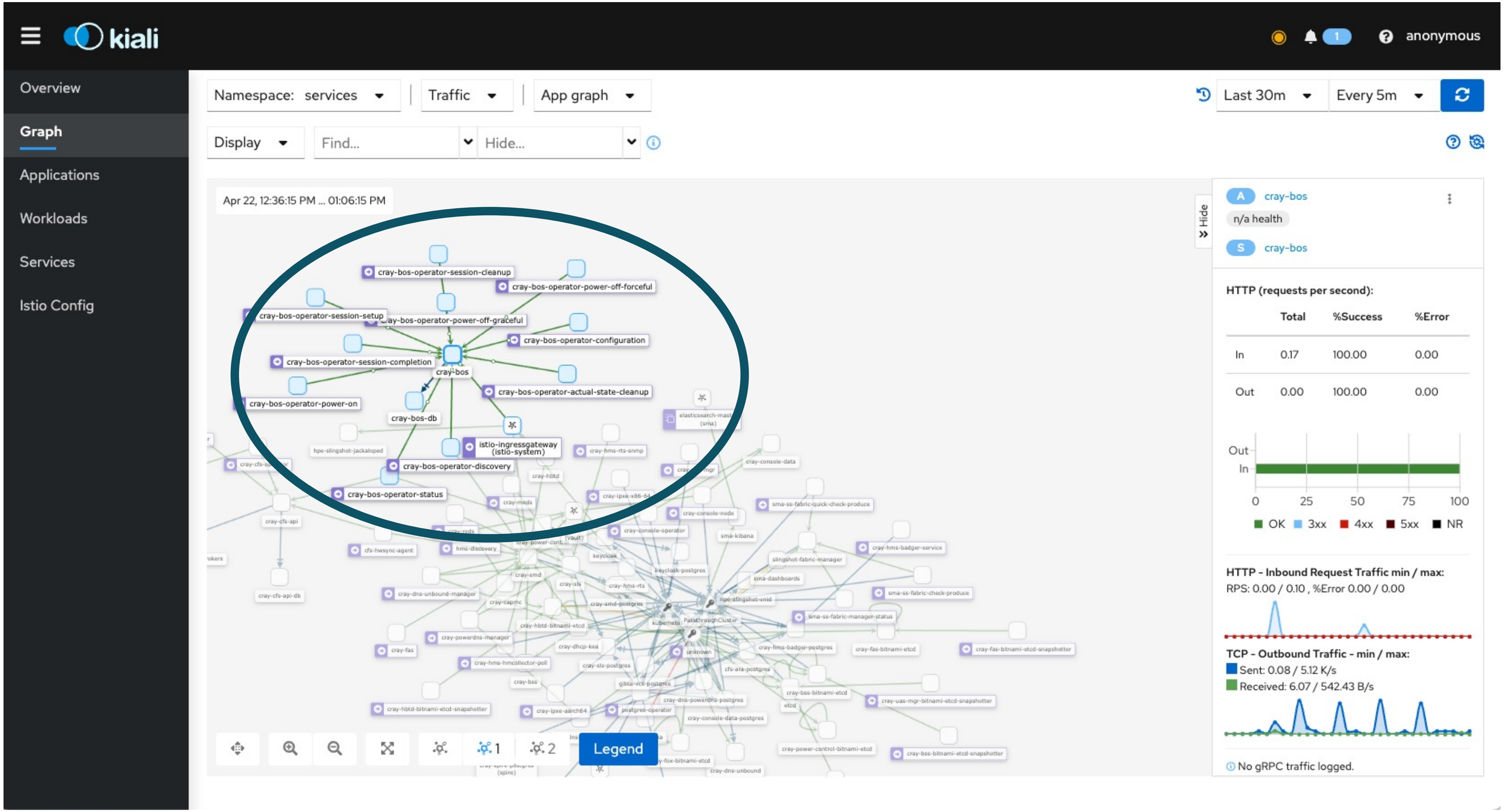
28 errors found

57 warnings found

localhost:20001/kiali/console/istio?namespaces=services&configvalidation=Warning&configvalidation=Not+Valid

https://kiali.cmn.SYSTEM_DOMAIN_NAME

Kiali – cray-bos



kiali
anonymous
Overview
Graph
Applications
Workloads
Services
Istio Config
Display
Hide
cray-bos-operator-session-cleanup
cray-bos-operator-power-off-forceful
cray-bos-operator-session-setup
cray-bos-operator-power-off-graceful
cray-bos-operator-configuration
cray-bos-operator-session-completion
cray-bos
cray-bos-operator-actual-state-cleanup
cray-bos-operator-power-on
cray-bos-db
istio-ingressgateway (istio-system)
cray-bos-operator-discovery
cray-bos-operator-status
Legend
A cray-bos
n/a health
S cray-bos

Namespace: services · Traffic · App graph · Last 30m · Every 5m

Find... · Hide...

Apr 22, 12:36:15 PM ... 01:06:15 PM

HTTP (requests per second):

	Total	%Success	%Error
In	0.17	100.00	0.00
Out	0.00	100.00	0.00

Out
In
0
25
50
75
100
OK
3xx
4xx
5xx
NR

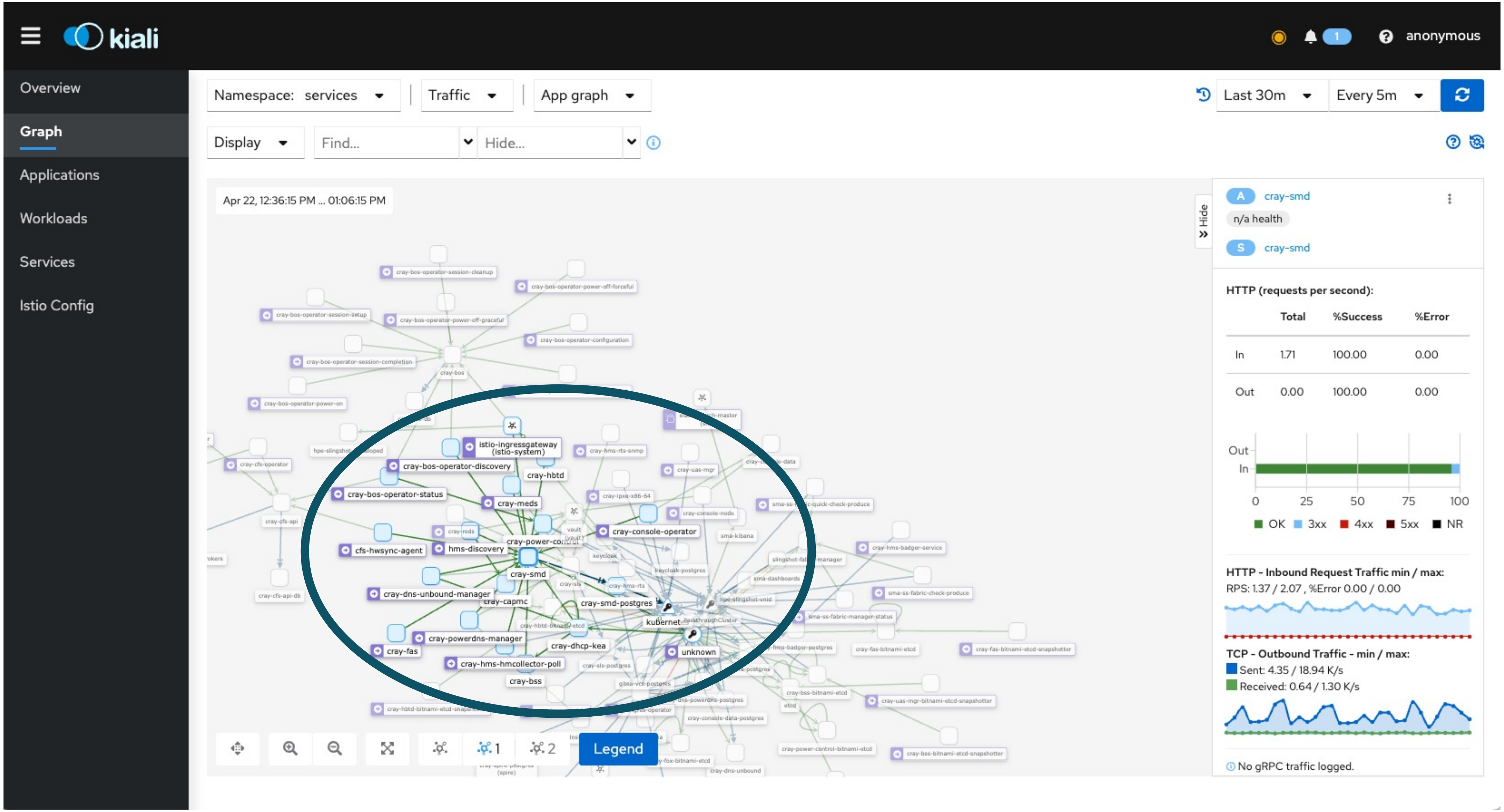
HTTP - Inbound Request Traffic min / max:
RPS: 0.00 / 0.10 , %Error 0.00 / 0.00

TCP - Outbound Traffic - min / max:
Sent: 0.08 / 5.12 K/s
Received: 6.07 / 542.43 B/s

No gRPC traffic logged.

https://kiali.cmn.SYSTEM_DOMAIN_NAME

Kiali – cray-smd



kiali
anonymous
Overview
Graph
Applications
Workloads
Services
Istio Config
Display
Find...
Hide...
Legend
Hide

Namespace: services | Traffic | App graph

Last 30m | Every 5m

Apr 22, 12:36:15 PM ... 01:06:15 PM

cray-bos-operator-session-cleanup
cray-bos-operator-power-off-forceful
cray-bos-operator-session-setup
cray-bos-operator-power-off-graceful
cray-bos-operator-configuration
cray-bos-operator-session-completion
cray-bos
cray-bos-operator-power-on
istio-ingressgateway (istio-system)
cray-bos-operator-discovery
cray-hbtd
cray-bos-operator-status
cray-meds
cray-console-operator
cfs-hwsync-agent
hms-discovery
cray-power-control
cray-smd
cray-dns-unbound-manager
cray-capmc
cray-smd-postgres
kubernetes
cray-powerdns-manager
cray-dhcp-kea
unknown
cray-fas
cray-hms-hmcollector-poll
cray-bss
cray-cfs-operator
cray-cfs-api
cray-cfs-api-db
cray-hms-rts-snmp
cray-uas-mgr
cray-ipxe-x86-64
cray-console-node
sma-kibana
cray-hms-badger-service
sma-ss-fabric-check-produce
sma-ss-fabric-manager-status
cray-fas-bitnami-etcd
cray-fas-bitnami-etcd-snapshotter
cray-uas-mgr-bitnami-etcd-snapshotter
cray-bss-bitnami-etcd
cray-bss-bitnami-etcd-snapshotter
cray-power-control-bitnami-etcd
cray-console-data-postgres
cray-hbtd-bitnami-etcd-snapshotter
cray-sls-postgres
gitea-vcs-postgres
cray-dns-unbound

A cray-smd

n/a health

S cray-smd

HTTP (requests per second):

	Total	%Success	%Error
In	1.71	100.00	0.00
Out	0.00	100.00	0.00

Out
In
0
25
50
75
100
🟩 OK 🟦 3xx 🟥 4xx ■ 5xx ■ NR

HTTP - Inbound Request Traffic min / max:

RPS: 1.37 / 2.07 , %Error 0.00 / 0.00

TCP - Outbound Traffic - min / max:

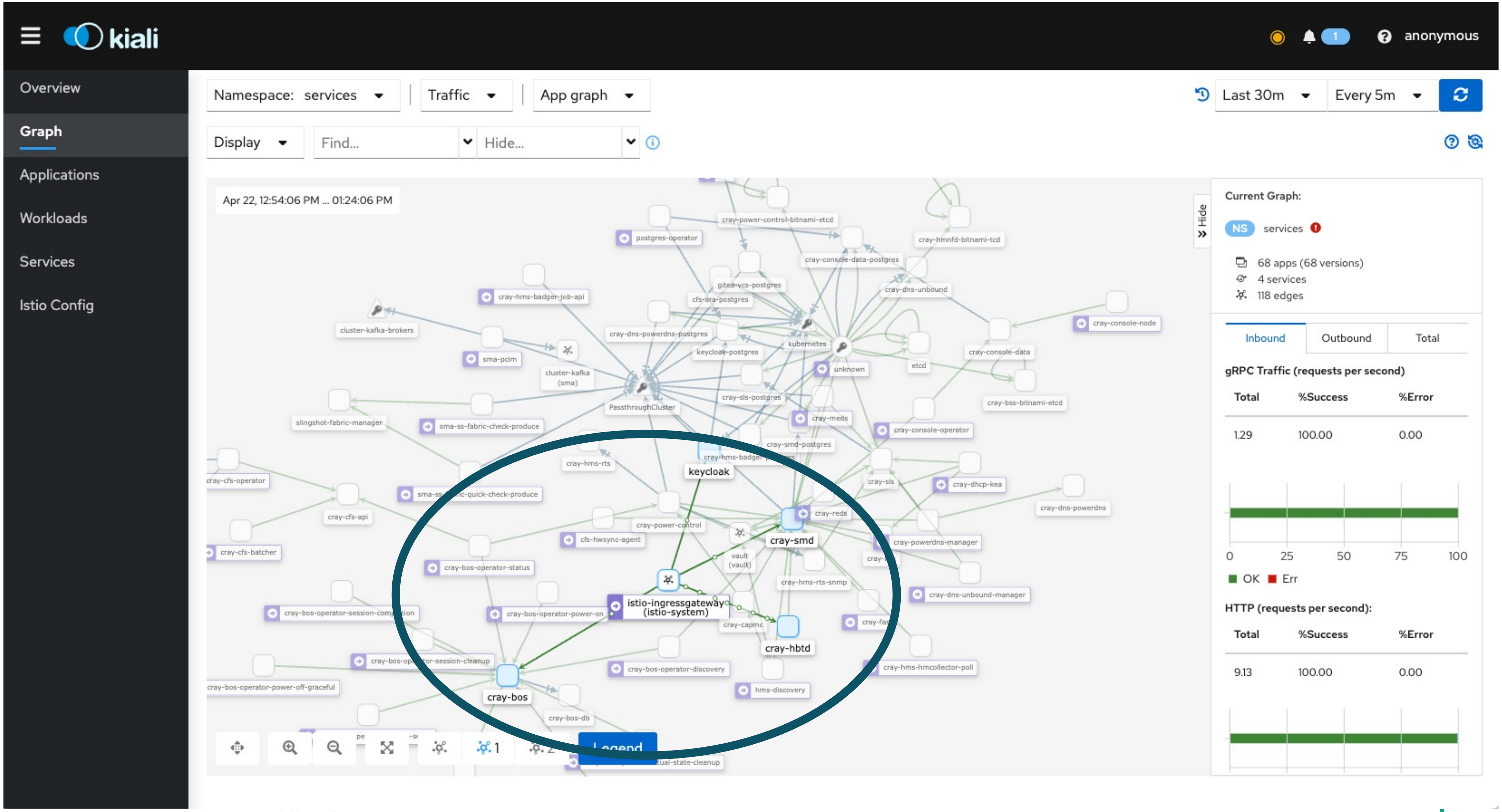
🟦 Sent: 4.35 / 18.94 K/s

🟩 Received: 0.64 / 1.30 K/s

ⓘ No gRPC traffic logged.

https://kiali.cmn.SYSTEM_DOMAIN_NAME

Kiali – istio-ingressgateway



Namespace: services

Traffic

App graph

Last 30m

Every 5m

Display

Find...

Hide...

Apr 22, 12:54:06 PM ... 01:24:06 PM

Current Graph:

NS services

68 apps (68 versions)

4 services

118 edges

Inbound | Outbound | Total

gRPC Traffic (requests per second)

Total	%Success	%Error
1.29	100.00	0.00

0 25 50 75 100

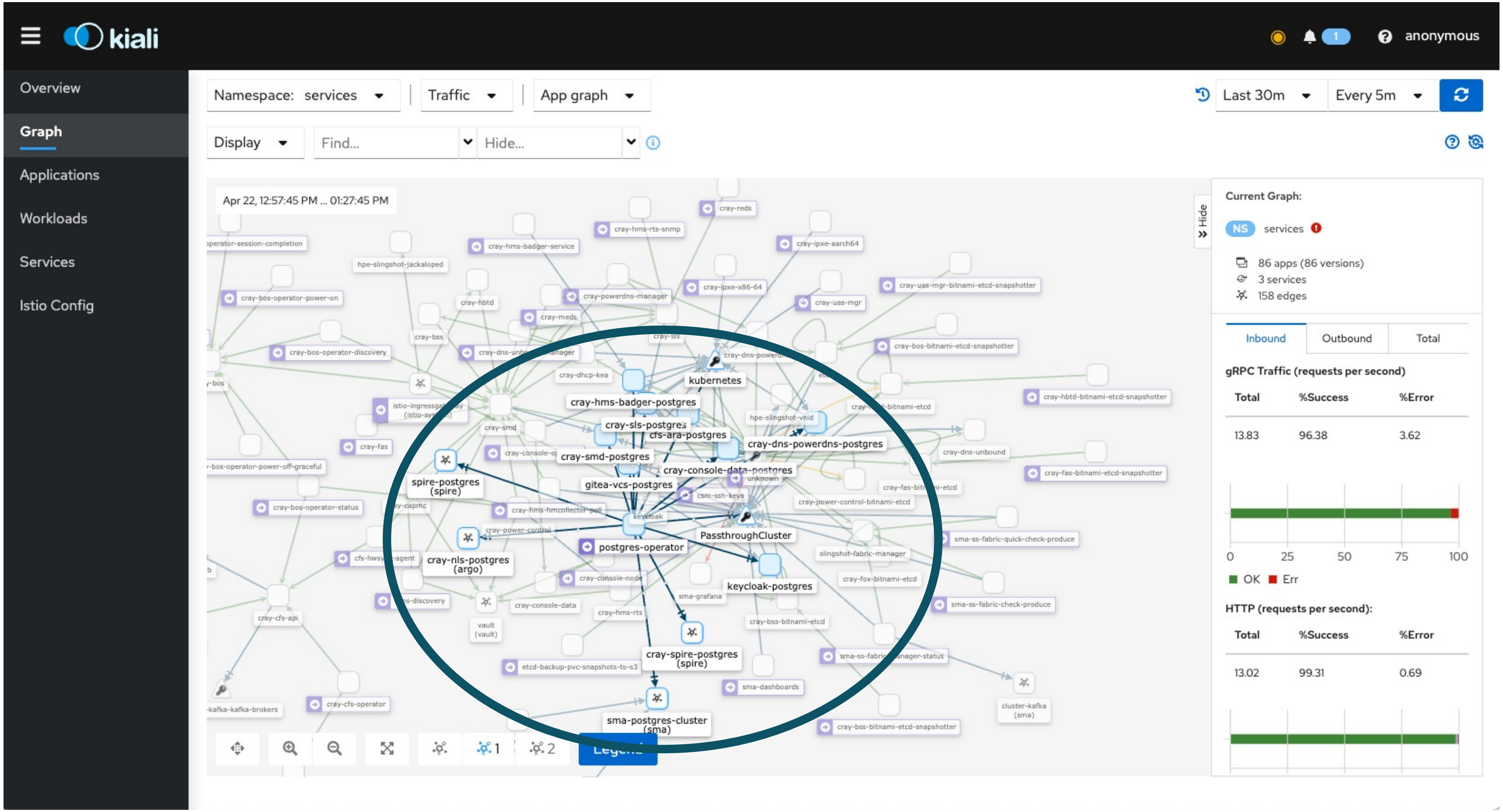
🟩 OK 🟥 Err

HTTP (requests per second):

Total	%Success	%Error
9.13	100.00	0.00

https://kiali.cmn.SYSTEM_DOMAIN_NAME

Kiali – postgres-operator



Namespace: services | Traffic | App graph

Last 30m | Every 5m

Apr 22, 12:57:45 PM ... 01:27:45 PM

Current Graph:

NS services

86 apps (86 versions)
3 services
158 edges

gRPC Traffic (requests per second)

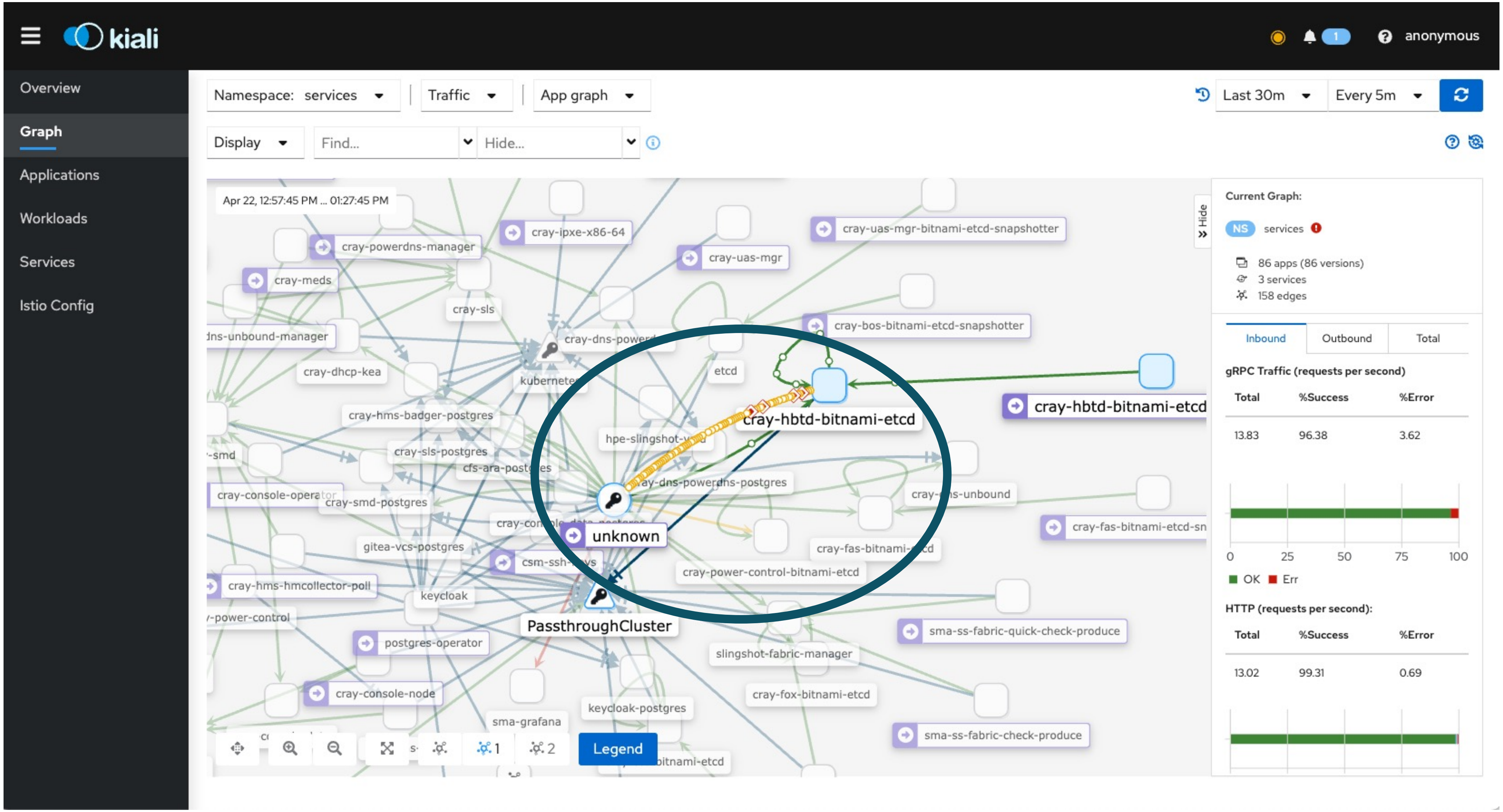
Total	%Success	%Error
13.83	96.38	3.62

HTTP (requests per second):

Total	%Success	%Error
13.02	99.31	0.69

https://kiali.cmn.SYSTEM_DOMAIN_NAME

Kiali



Namespace: services

Traffic

App graph

Last 30m

Every 5m

Apr 22, 12:57:45 PM ... 01:27:45 PM

Current Graph:

NS services

86 apps (86 versions)
3 services
158 edges

Inbound | Outbound | Total

gRPC Traffic (requests per second)

Total	%Success	%Error
13.83	96.38	3.62

🟩 OK 🟥 Err

HTTP (requests per second):

Total	%Success	%Error
13.02	99.31	0.69

https://kiali.cmn.SYSTEM_DOMAIN_NAME

SMA Monitoring

- System Monitoring Framework (SMF)
- SMA Grafana
- SMA Flow

System monitoring Framework

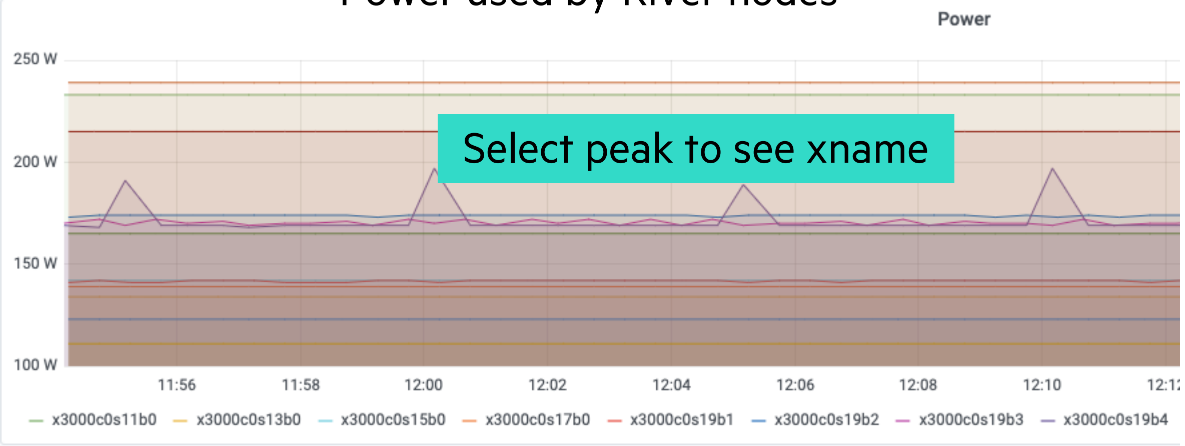
- Tightly-integrated monitoring system
- Provides detailed telemetry information from multiple subsystems:
  - Fabric
  - Environmental
  - Network
  - Storage
  - Operating systems (vmstat and iostat metrics)
- Incorporates the context necessary to understand telemetry data
- Feeds into a common message bus (Kafka), persistence, and minimal UI infrastructure
- SMA alarms and notifications subsystem monitors metric data
  - Provides a way to notify administrators when select metric data is outside of normal operating values
  - SMA includes several pre-defined alarms
  - Can be extended with site defined alarms

SMA-Grafana Dashboards

- Over 40 included dashboards
  - AIOps
  - System CPU, I/O, Kernel, Memory, Processes, Swap
  - Cabinet Controller Sensors
  - CDU Monitoring
  - CPU & GPU Temperatures
  - Fabric Congestion
  - Fabric Errors
  - Fabric Telemetry
  - Node Controller Sensors
  - Overview Details
  - Overview Device I/O Stats
  - PDU Monitoring
  - Redfish Events
  - River Sensors
  - Slingshot
  - Switch Controller Sensors
  - System Monitoring Dashboards

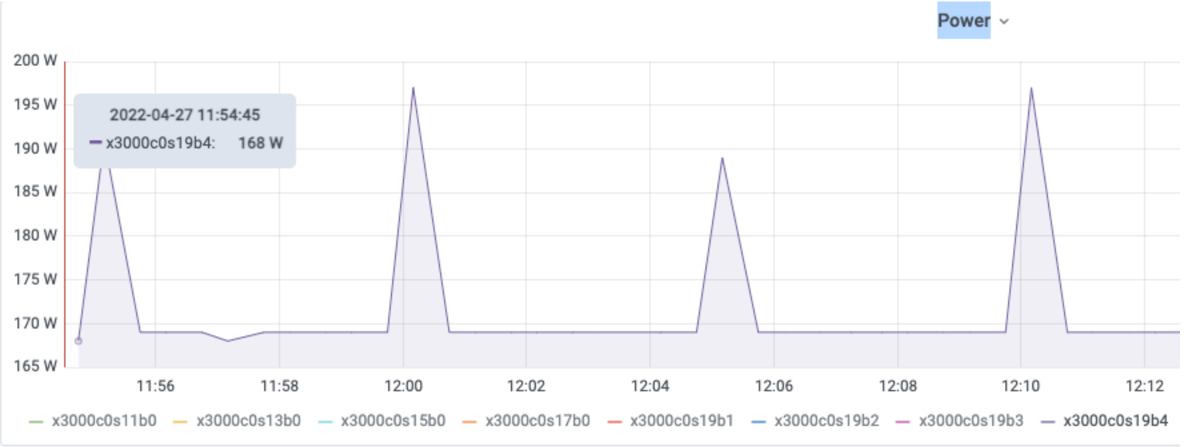
https://sma-grafana.cmn.SYSTEM_DOMAIN_NAME/dashboards

Power used by River nodes



Select peak to see xname

Click on xname to drill into that node



SMA New Grafana dashboards

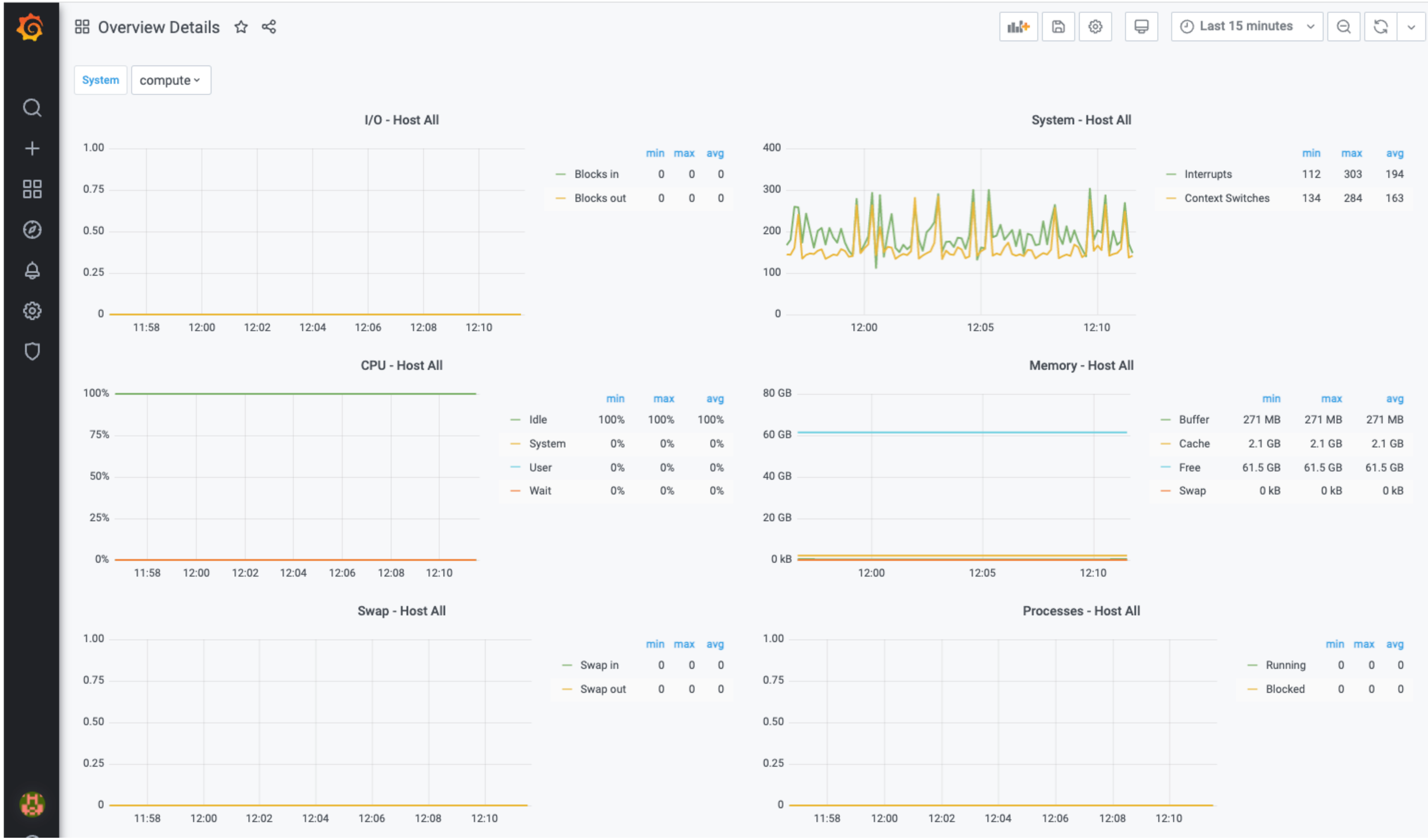
SMA 1.10/CSM 1.6

- Grafana dashboards which formerly used Postgres now use VictoriaMetrics
- Removed dashboards
  - Alerta Dashboard
  - Cluster Health Check
  - Prometheus Alerts Overview

SMA 1.9/CSM 1.5

- See Monitoring Cooling Devices with Artificial Intelligence for IT operations
  - AIOps Anomaly Forecast
  - AIOps Slingshot Physical Context Congestion
  - AIOps Slingshot Physical Context Congestion Details
  - AIOps Slingshot Physical Context Temperature Details
  - AIOps Univariate Dashboard

SMA-Grafana Overview Details



I/O - Host All

	min	max	avg
Blocks in	0	0	0
Blocks out	0	0	0

System - Host All

	min	max	avg
Interrupts	112	303	194
Context Switches	134	284	163

CPU - Host All

	min	max	avg
Idle	100%	100%	100%
System	0%	0%	0%
User	0%	0%	0%
Wait	0%	0%	0%

Memory - Host All

	min	max	avg
Buffer	271 MB	271 MB	271 MB
Cache	2.1 GB	2.1 GB	2.1 GB
Free	61.5 GB	61.5 GB	61.5 GB
Swap	0 kB	0 kB	0 kB

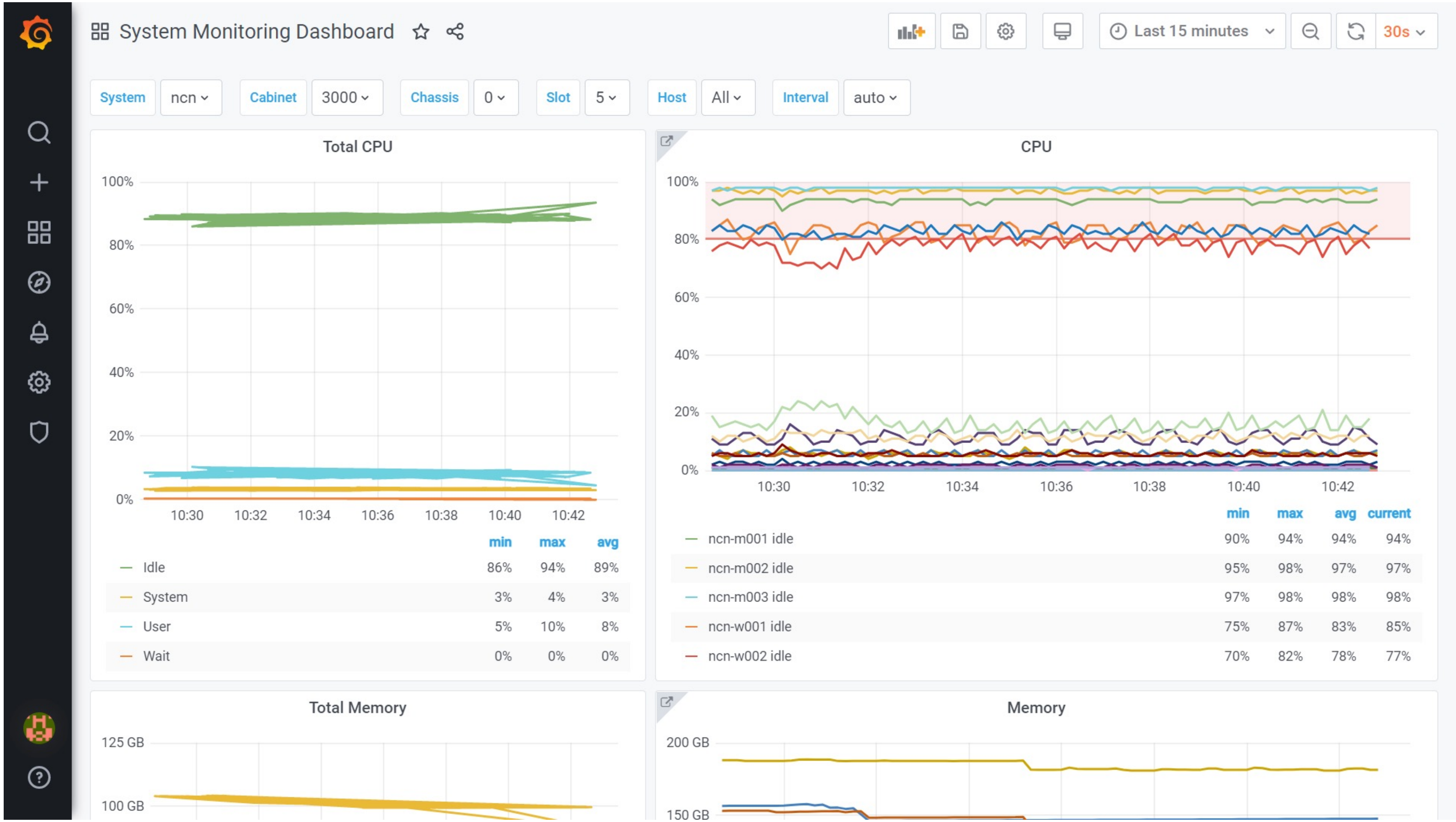
Swap - Host All

	min	max	avg
Swap in	0	0	0
Swap out	0	0	0

Processes - Host All

	min	max	avg
Running	0	0	0
Blocked	0	0	0

SMA-Grafana System Monitoring Dashboard



System Monitoring Dashboard

Last 15 minutes
30s

System ncn
Cabinet 3000
Chassis 0
Slot 5
Host All
Interval auto

Total CPU

	min	max	avg
— Idle	86%	94%	89%
— System	3%	4%	3%
— User	5%	10%	8%
— Wait	0%	0%	0%

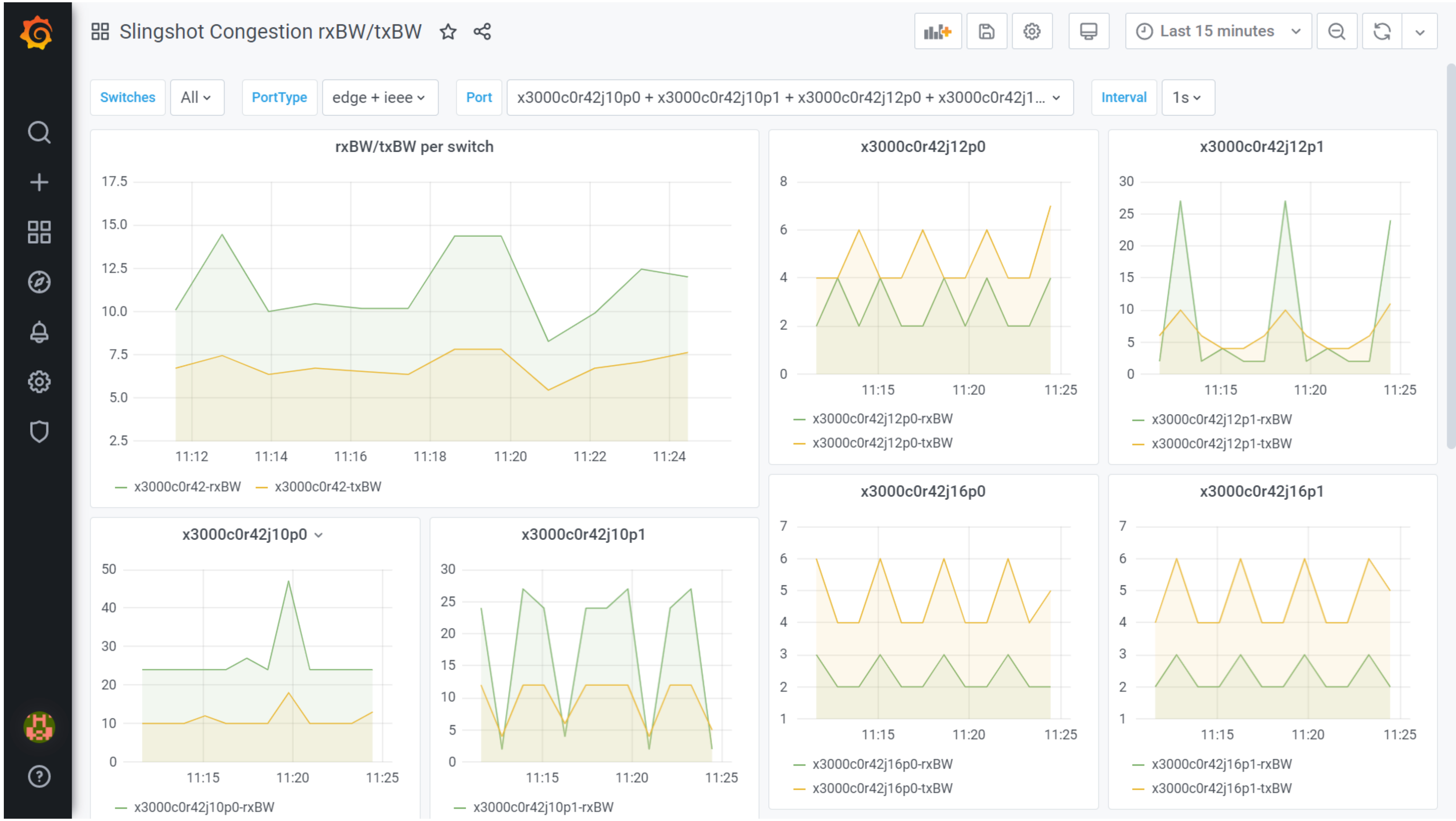
CPU

	min	max	avg	current
— ncn-m001 idle	90%	94%	94%	94%
— ncn-m002 idle	95%	98%	97%	97%
— ncn-m003 idle	97%	98%	98%	98%
— ncn-w001 idle	75%	87%	83%	85%
— ncn-w002 idle	70%	82%	78%	77%

Total Memory

Memory

SMA-Grafana Slingshot Congestion Receive/Transmit Bandwidth



Slingshot Congestion rxBW/txBW

Switches: All

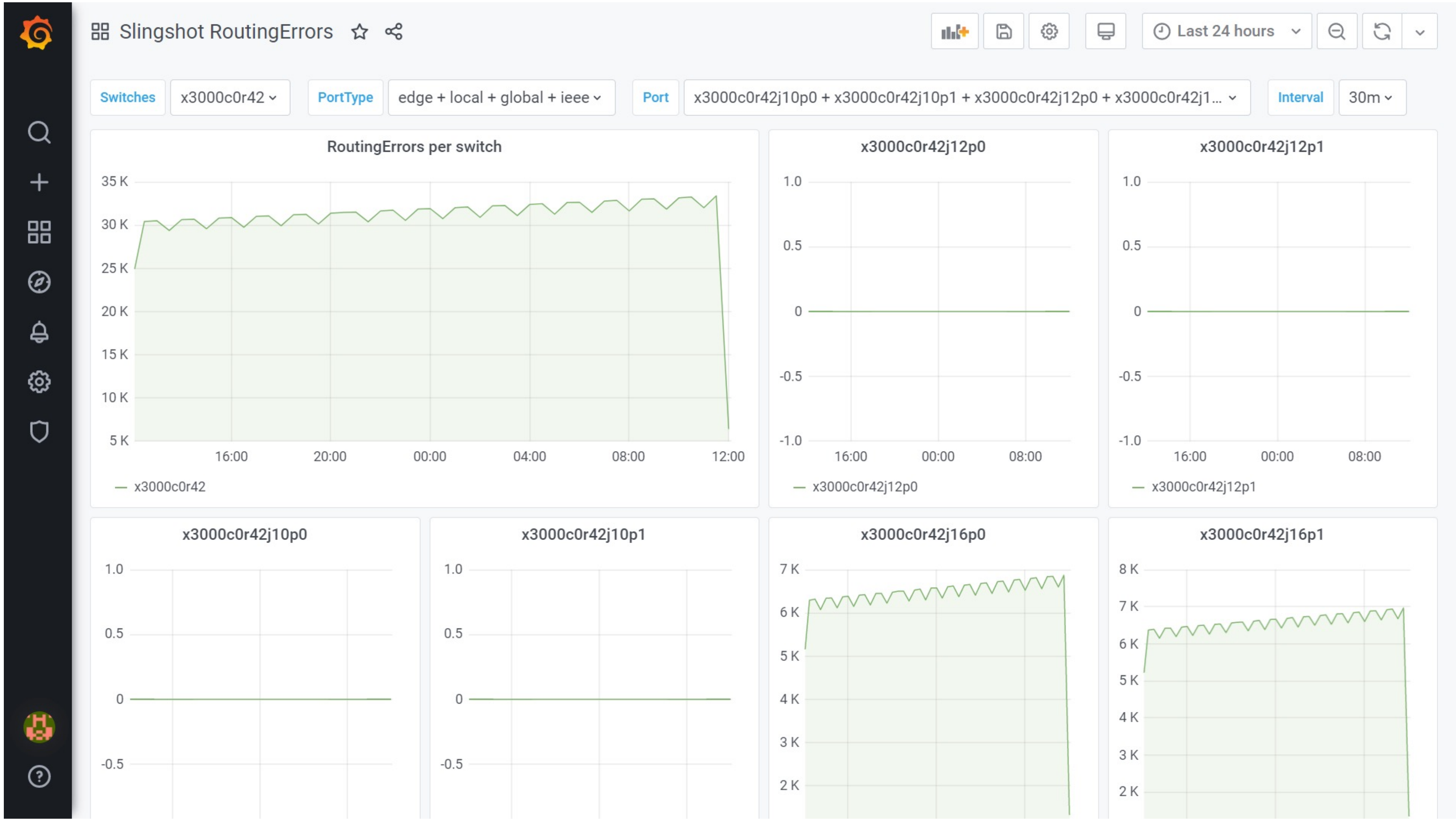
PortType: edge + ieee

Port: x3000c0r42j10p0 + x3000c0r42j10p1 + x3000c0r42j12p0 + x3000c0r42j1...

Interval: 1s

Last 15 minutes

SMA-Grafana Slingshot RoutingErrors



Slingshot RoutingErrors

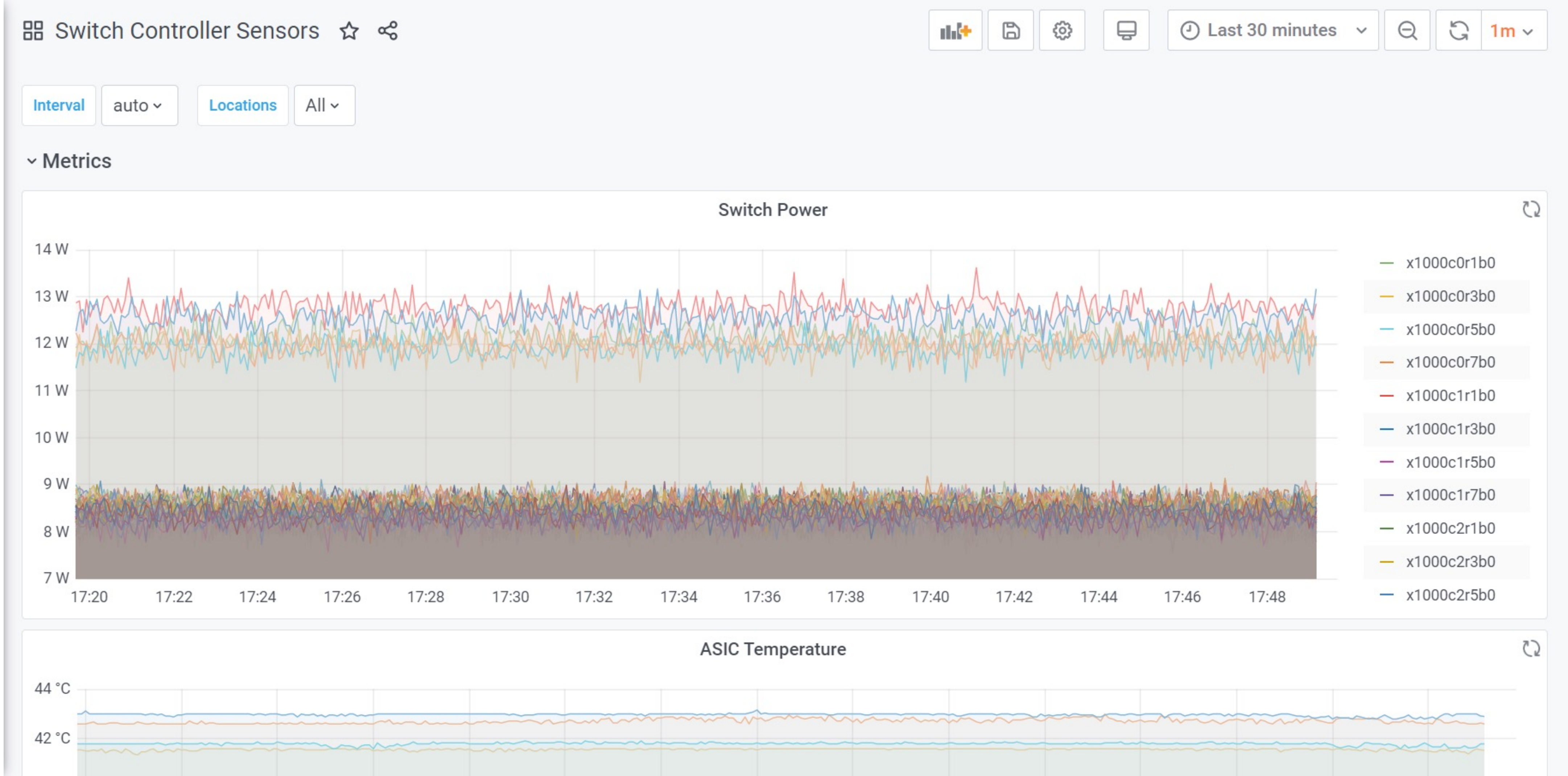
Switches: x3000c0r42

PortType: edge + local + global + ieee

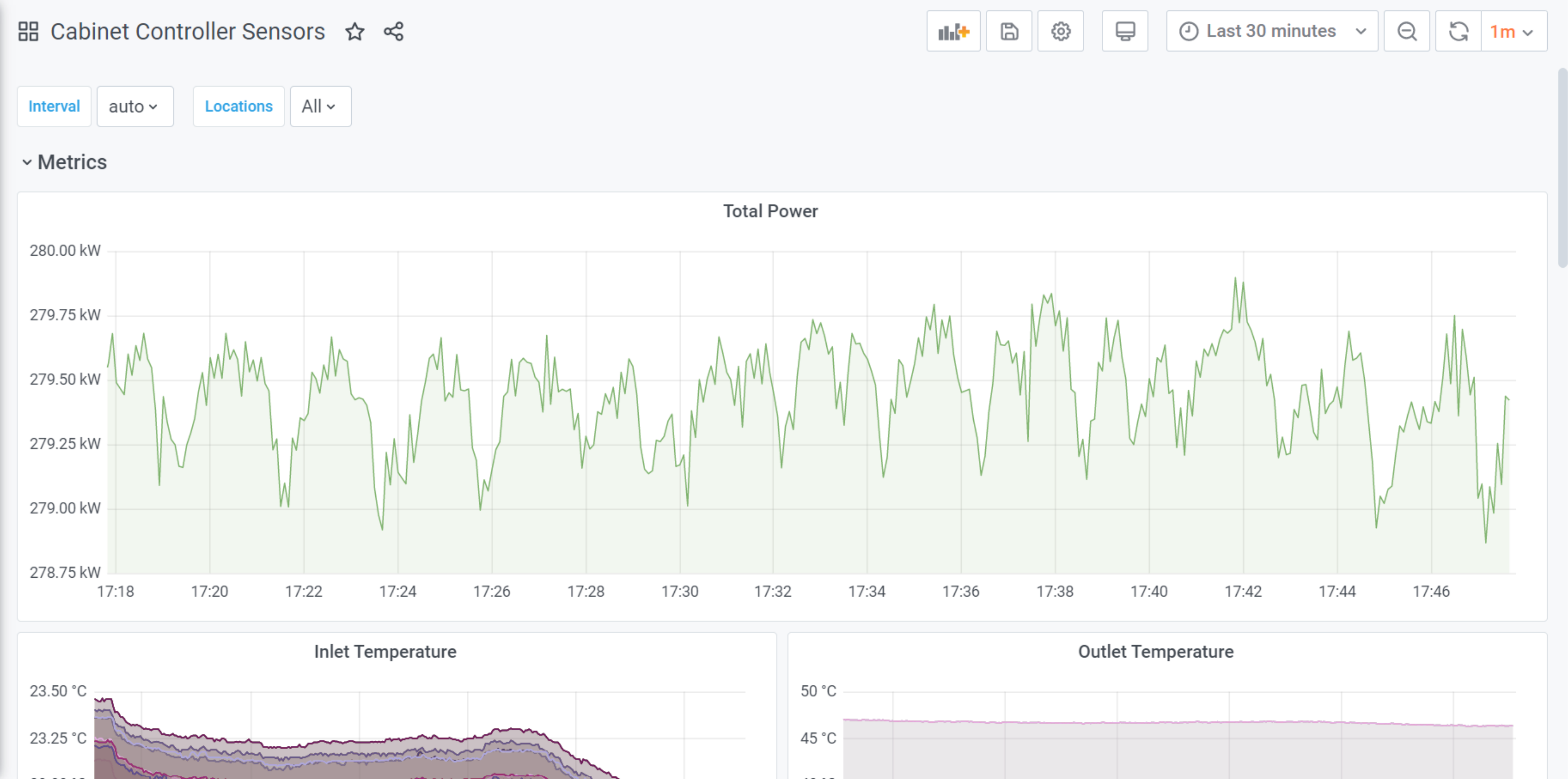
Port: x3000c0r42j10p0 + x3000c0r42j10p1 + x3000c0r42j12p0 + x3000c0r42j1...

Interval: 30m

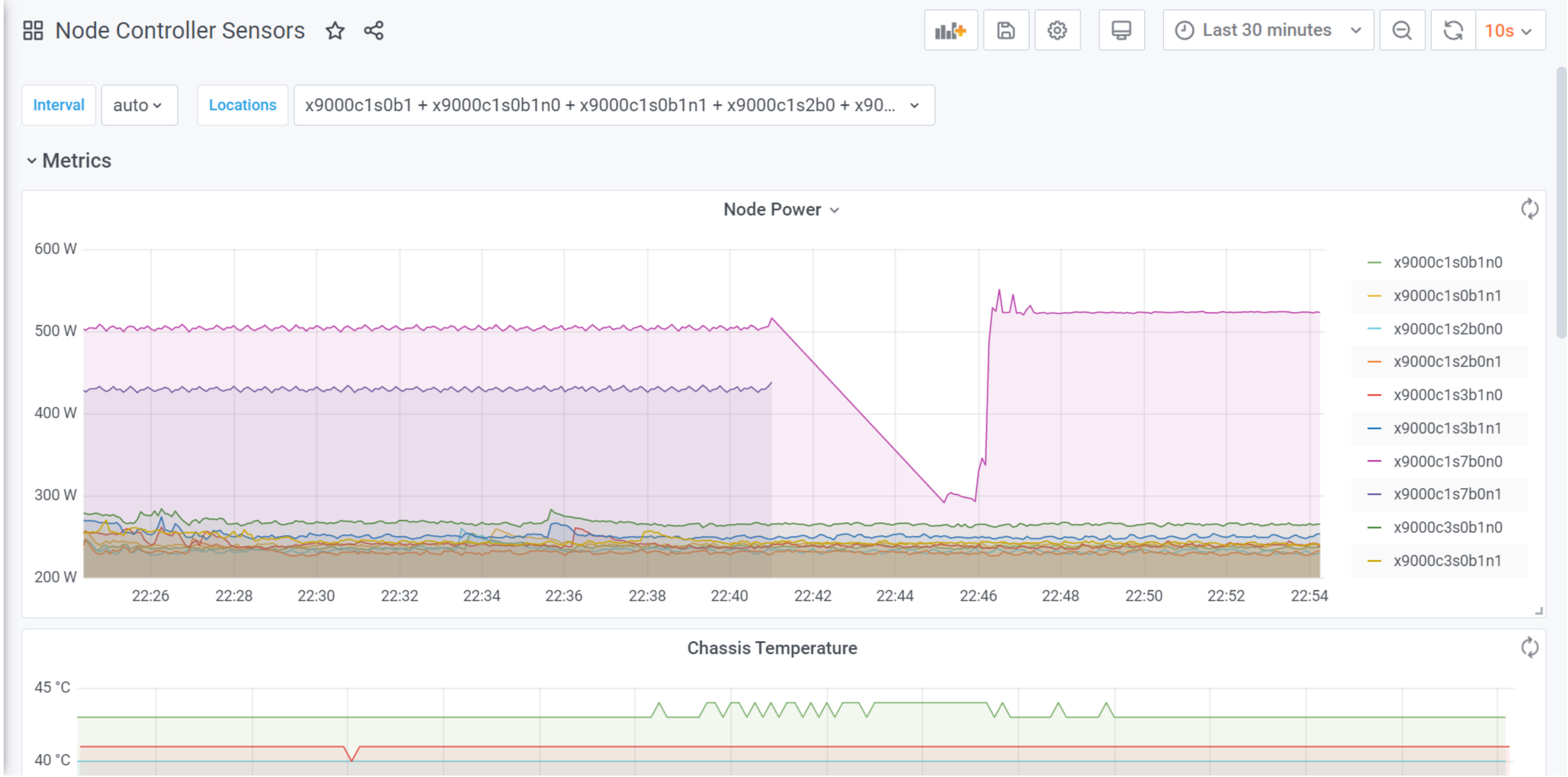
SMA-Grafana Switch Controller Sensors



SMA-Grafana Cabinet Controller Sensors



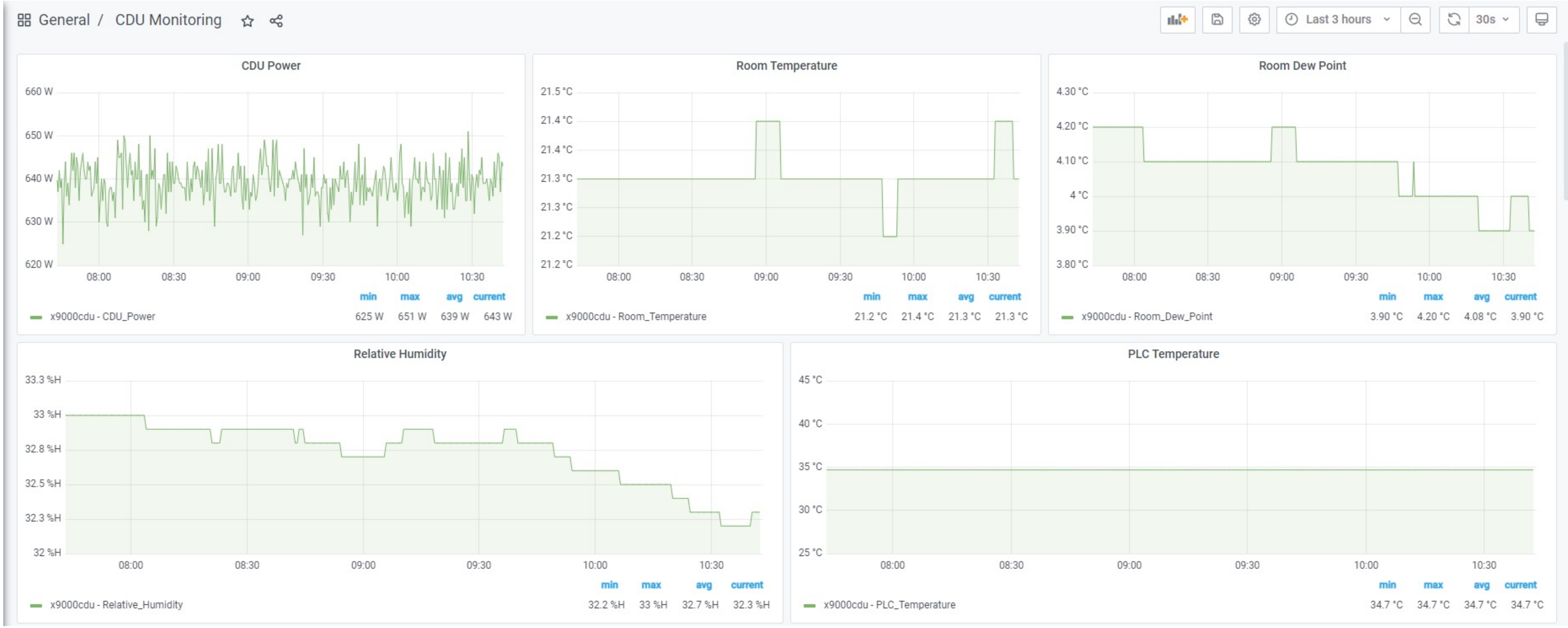
SMA-Grafana Node Controller Sensors



Interval: auto

Locations: x9000c1s0b1 + x9000c1s0b1n0 + x9000c1s0b1n1 + x9000c1s2b0 + x90...

CDU Monitoring



	min	max	avg	current
x9000cdu - CDU_Power	625 W	651 W	639 W	643 W

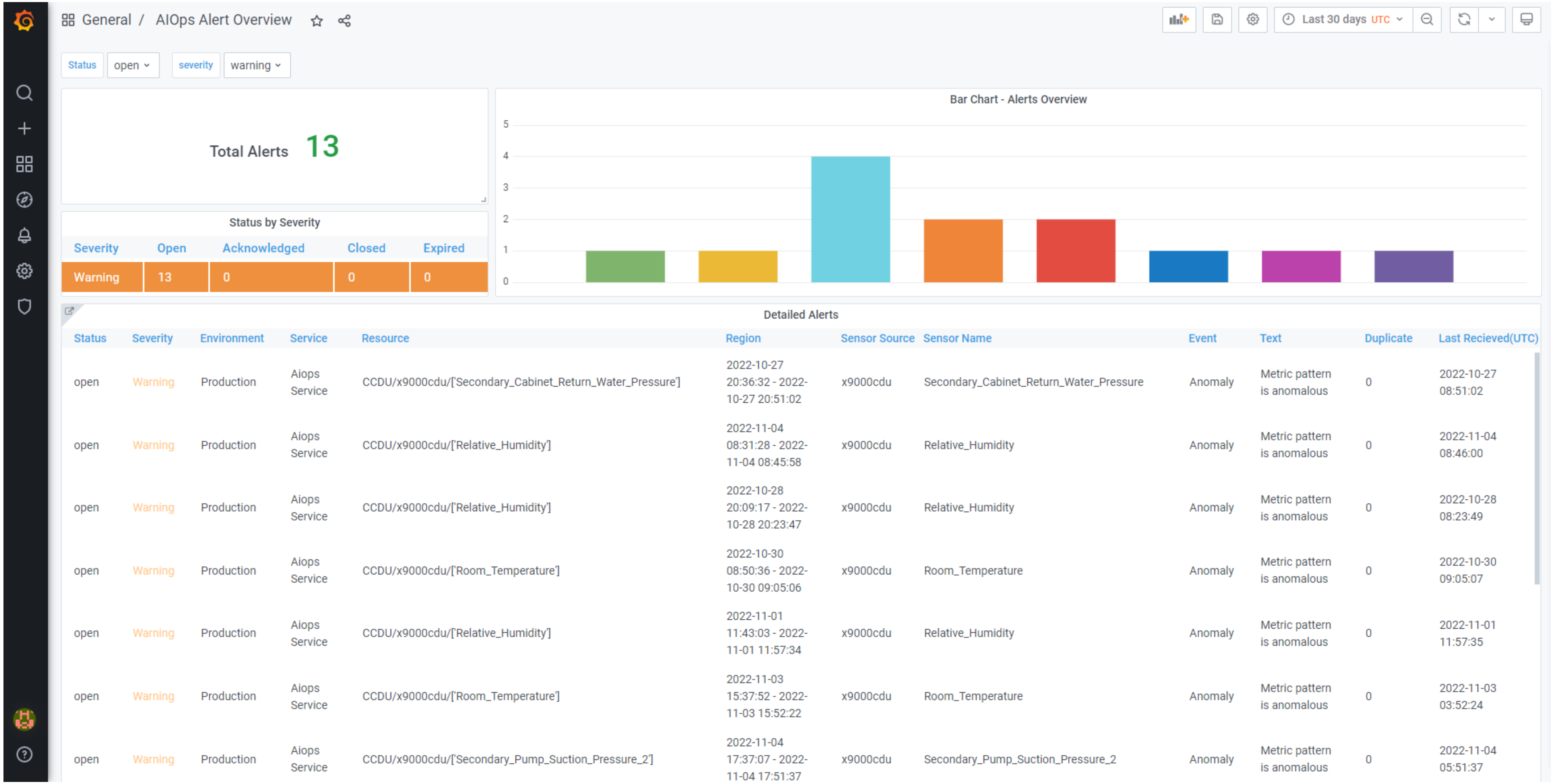
	min	max	avg	current
x9000cdu - Room_Temperature	21.2 °C	21.4 °C	21.3 °C	21.3 °C

	min	max	avg	current
x9000cdu - Room_Dew_Point	3.90 °C	4.20 °C	4.08 °C	3.90 °C

	min	max	avg	current
x9000cdu - Relative_Humidity	32.2 %H	33 %H	32.7 %H	32.3 %H

	min	max	avg	current
x9000cdu - PLC_Temperature	34.7 °C	34.7 °C	34.7 °C	34.7 °C

AIOps Alert Overview



Total Alerts 13

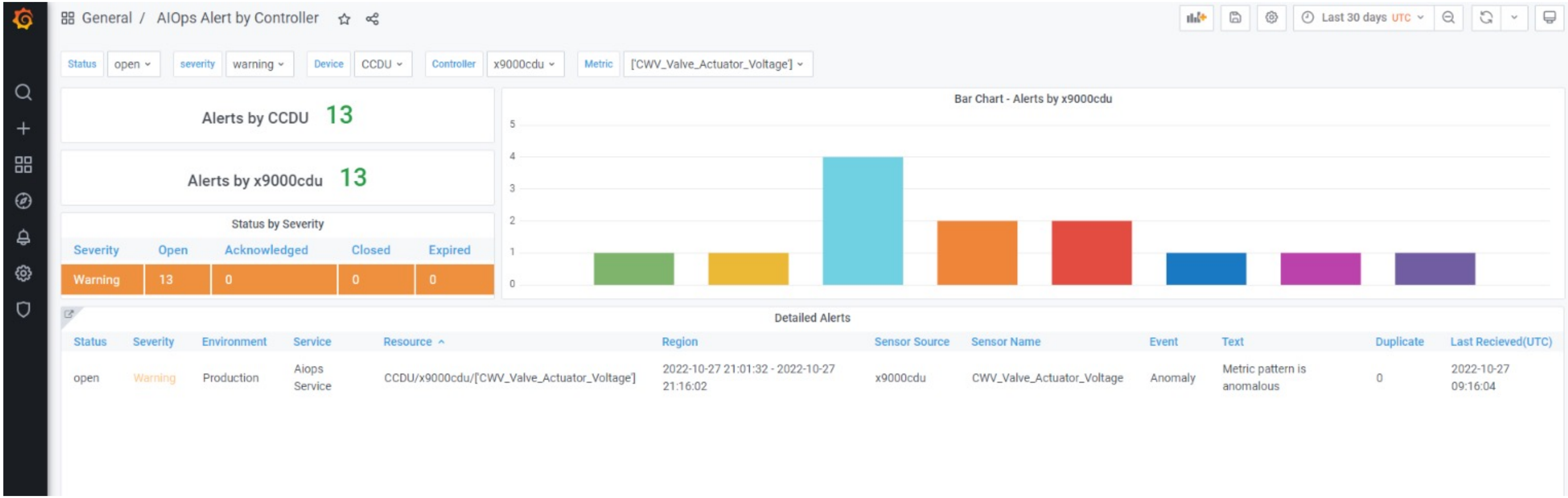
Status by Severity

Severity	Open	Acknowledged	Closed	Expired
Warning	13	0	0	0

Detailed Alerts

Status	Severity	Environment	Service	Resource	Region	Sensor Source	Sensor Name	Event	Text	Duplicate	Last Recieved(UTC)
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['Secondary_Cabinet_Return_Water_Pressure']	2022-10-27 20:36:32 - 2022-10-27 20:51:02	x9000cdu	Secondary_Cabinet_Return_Water_Pressure	Anomaly	Metric pattern is anomalous	0	2022-10-27 08:51:02
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['Relative_Humidity']	2022-11-04 08:31:28 - 2022-11-04 08:45:58	x9000cdu	Relative_Humidity	Anomaly	Metric pattern is anomalous	0	2022-11-04 08:46:00
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['Relative_Humidity']	2022-10-28 20:09:17 - 2022-10-28 20:23:47	x9000cdu	Relative_Humidity	Anomaly	Metric pattern is anomalous	0	2022-10-28 08:23:49
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['Room_Temperature']	2022-10-30 08:50:36 - 2022-10-30 09:05:06	x9000cdu	Room_Temperature	Anomaly	Metric pattern is anomalous	0	2022-10-30 09:05:07
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['Relative_Humidity']	2022-11-01 11:43:03 - 2022-11-01 11:57:34	x9000cdu	Relative_Humidity	Anomaly	Metric pattern is anomalous	0	2022-11-01 11:57:35
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['Room_Temperature']	2022-11-03 15:37:52 - 2022-11-03 15:52:22	x9000cdu	Room_Temperature	Anomaly	Metric pattern is anomalous	0	2022-11-03 03:52:24
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['Secondary_Pump_Suction_Pressure_2']	2022-11-04 17:37:07 - 2022-11-04 17:51:37	x9000cdu	Secondary_Pump_Suction_Pressure_2	Anomaly	Metric pattern is anomalous	0	2022-11-04 05:51:37

AIOps Alert by controller



General / AIOps Alert by Controller
Last 30 days UTC
Bar Chart - Alerts by x9000cdu
5
4
3
2
1
0

Status: open
severity: warning
Device: CCDU
Controller: x9000cdu
Metric: ['CWV_Valve_Actuator_Voltage']

Alerts by CCDU 13

Alerts by x9000cdu 13

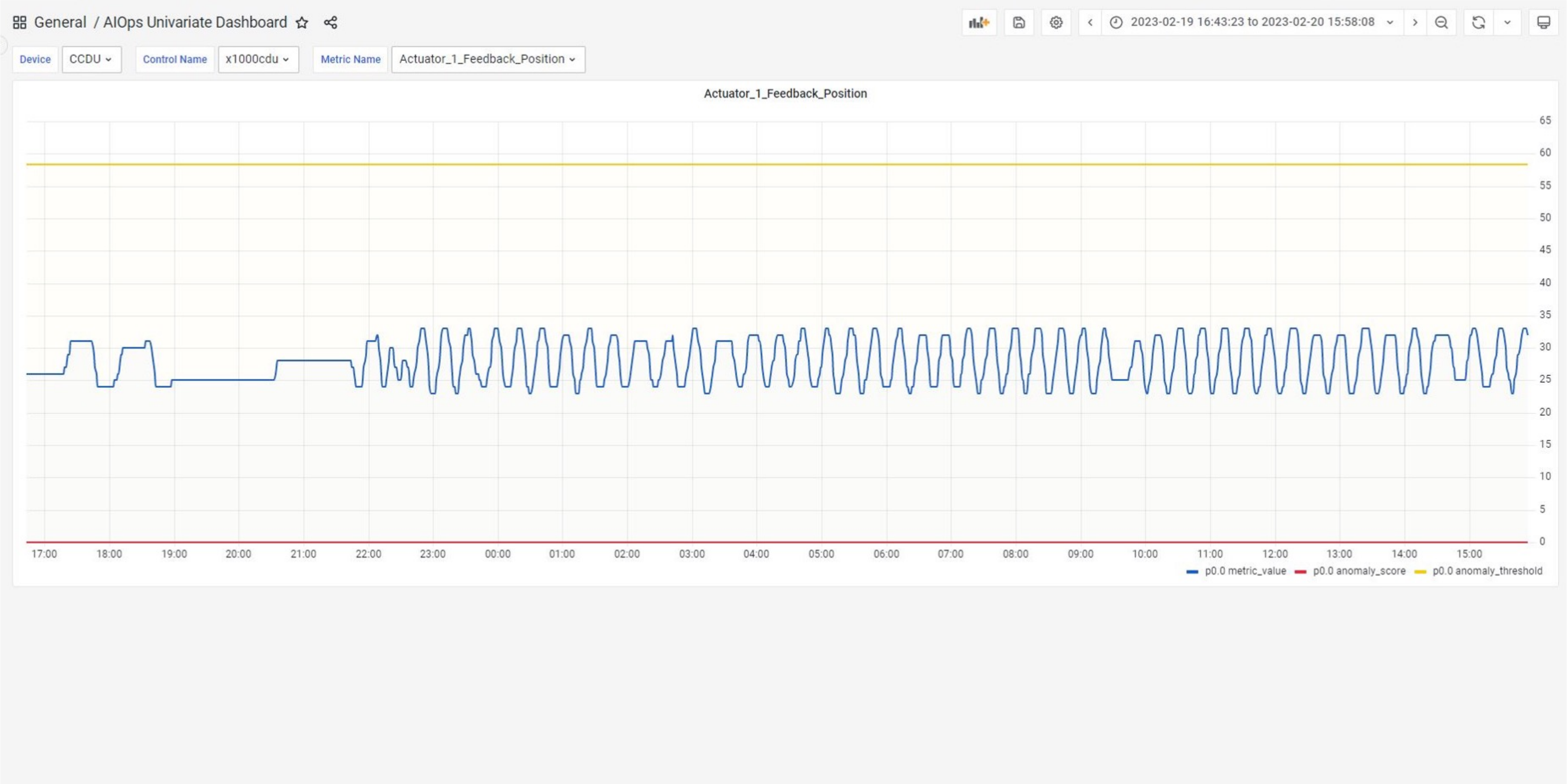
Status by Severity

Severity	Open	Acknowledged	Closed	Expired
Warning	13	0	0	0

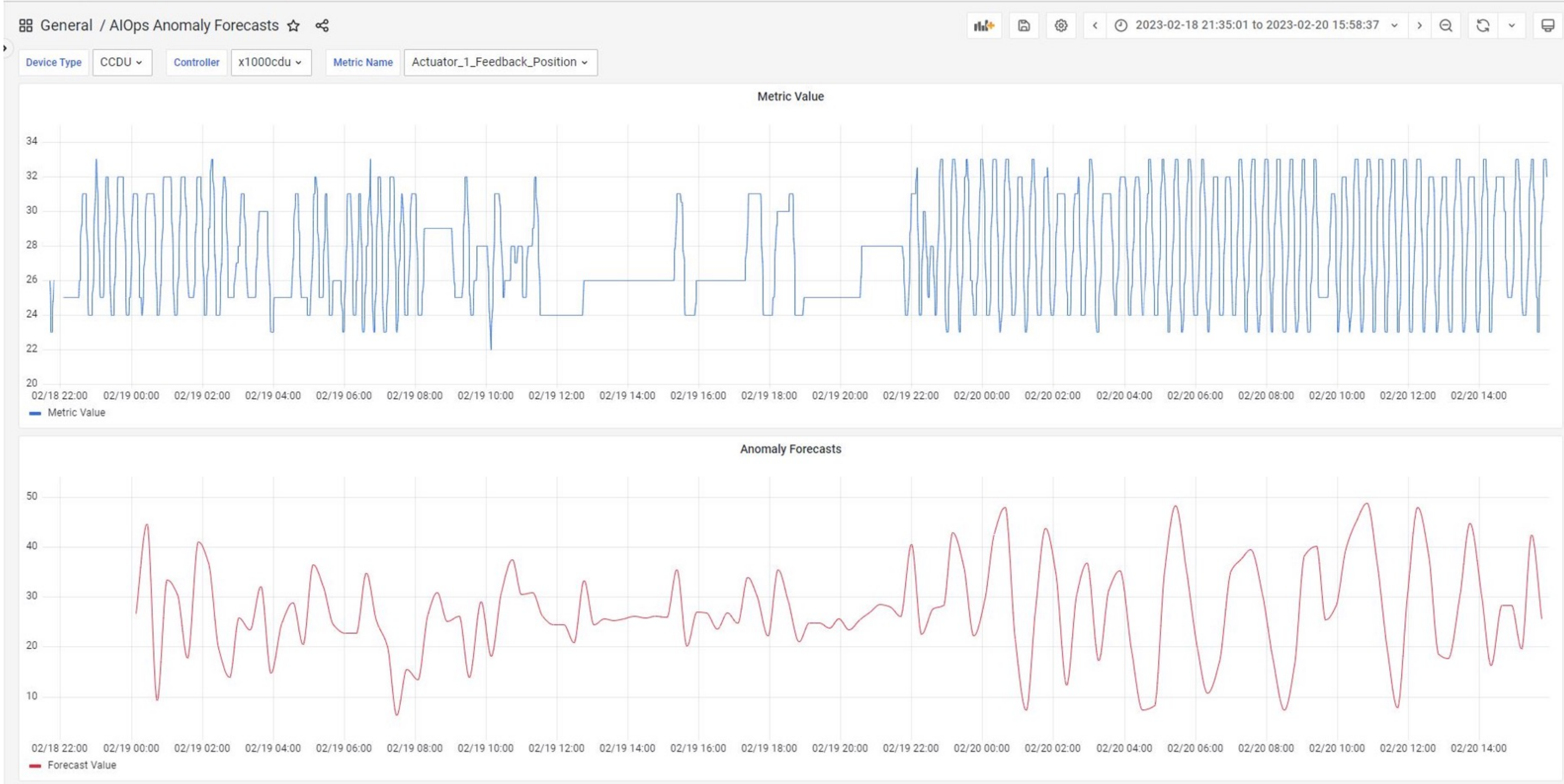
Detailed Alerts

Status	Severity	Environment	Service	Resource	Region	Sensor Source	Sensor Name	Event	Text	Duplicate	Last Recieved(UTC)
open	Warning	Production	Aiops Service	CCDU/x9000cdu/['CWV_Valve_Actuator_Voltage']	2022-10-27 21:01:32 - 2022-10-27 21:16:02	x9000cdu	CWV_Valve_Actuator_Voltage	Anomaly	Metric pattern is anomalous	0	2022-10-27 09:16:04

AIOps Univariate Dashboard



AIOps Anomaly Forecast dashboard



General / AIOps Anomaly Forecasts

2023-02-18 21:35:01 to 2023-02-20 15:58:37

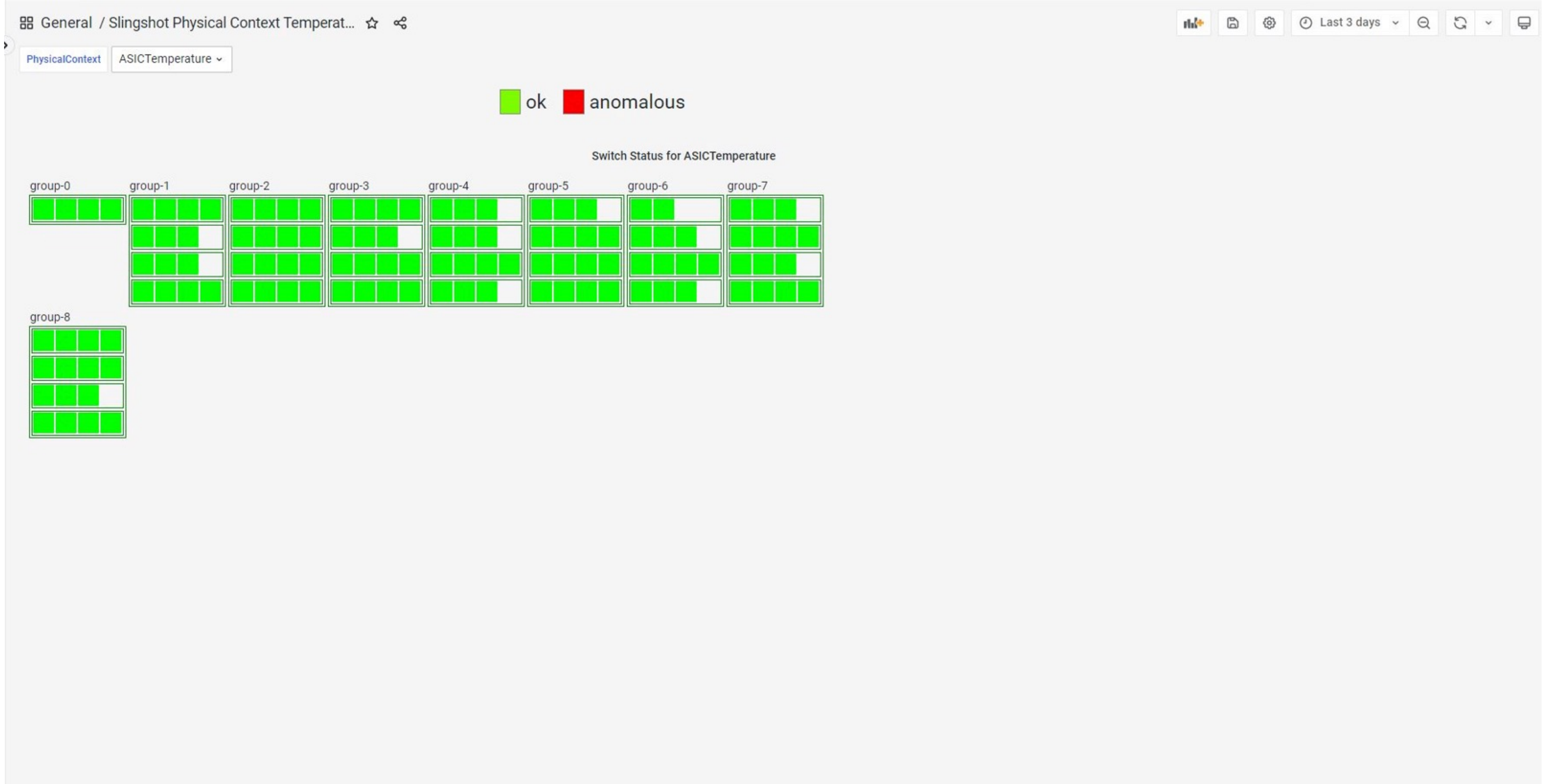
Device Type: CCDU

Controller: x1000cdu

Metric Name: Actuator_1_Feedback_Position

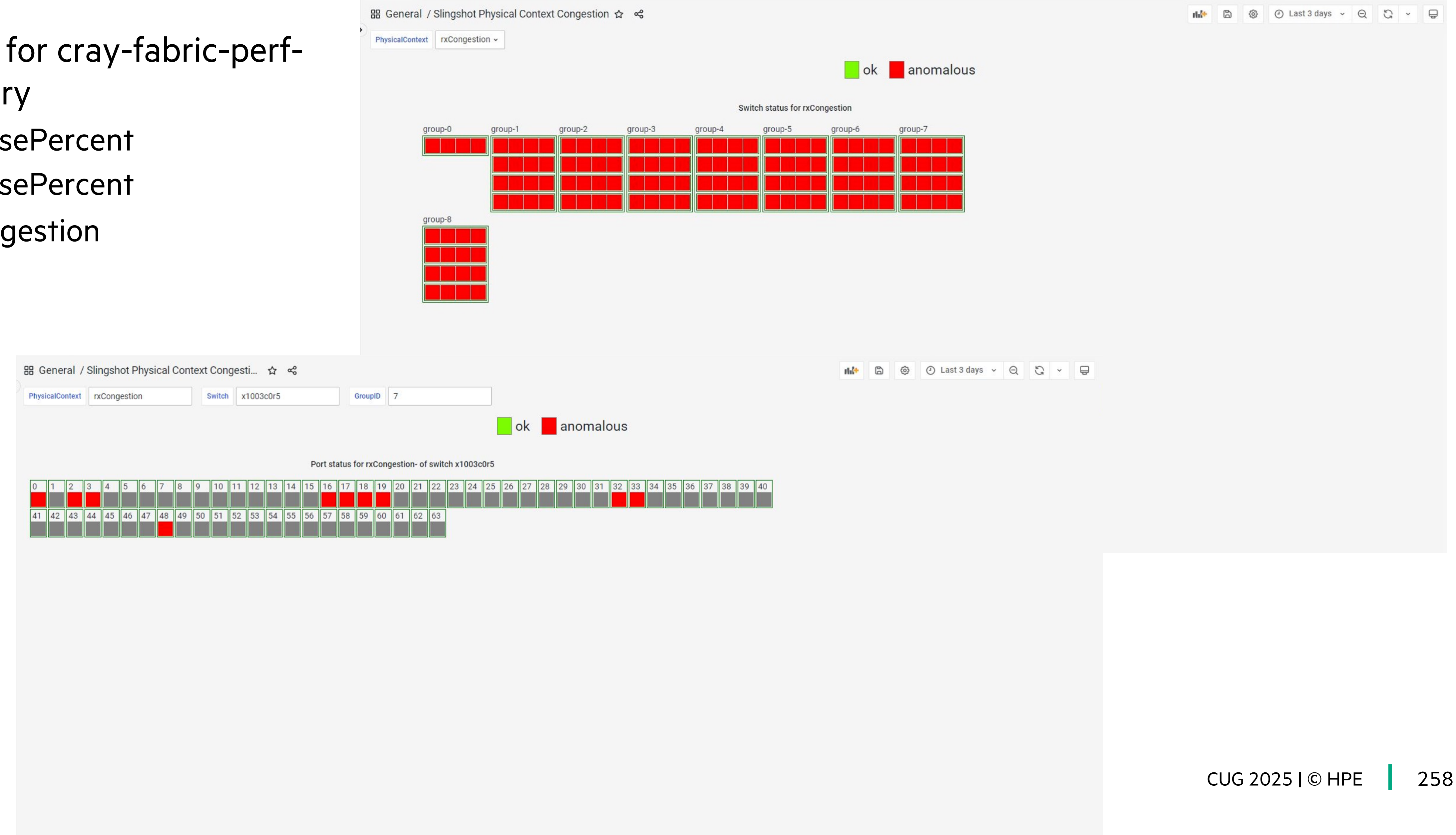
AIOps Anomaly Detection Slingshot Physical Context Temperature

- Metrics used for cray-telemetry-temperature
  - ASIC
  - NetworkingDevice
  - SystemBoard
  - VoltageRegulator
  - Chassis
  - PowerSupply



AIOps Anomaly Detection Slingshot Physical Context Congestion

- Metrics for cray-fabric-perf-telemetry
  - rxPausePercent
  - txPausePercent
  - rxCongestion


General / Slingshot Physical Context Congestion
Last 3 days
group-0
group-1
group-2
group-3
group-4
group-5
group-6
group-7
group-8
General / Slingshot Physical Context Congesti...
Last 3 days

PhysicalContext: rxCongestion

🟩 ok 🟥 anomalous

Switch status for rxCongestion

PhysicalContext: rxCongestion Switch: x1003c0r5 GroupID: 7

🟩 ok 🟥 anomalous

Port status for rxCongestion- of switch x1003c0r5

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
🟥	▨	🟥	🟥	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟥	🟥	🟥	🟥	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟥	🟥	▨	▨	▨	▨	▨	▨	▨

41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
▨	▨	▨	▨	▨	▨	▨	🟥	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

AIOps MLflow

- Mlflow
  - Open-source platform that manages the ML(Machine Learning) lifecycle
    - Including experimentation, reproducibility, deployment, and a central model registry
- MLflow facilitates transparency and standardization when you are training, tuning, and deploying machine learning models
- AIOps use MLFlow to deploy the latest-trained, best model to the production stage and also to manage the AIOps model lifecycle

- Get IP and port for MLflow

```
ncn# kubectl get -n sma service/sma-mlflow
NAME       TYPE      CLUSTER-IP   EXTERNAL-IP PORT(S)  AGE
sma-mlflow ClusterIP 10.31.223.14 <none>      5000/TCP 15d
```

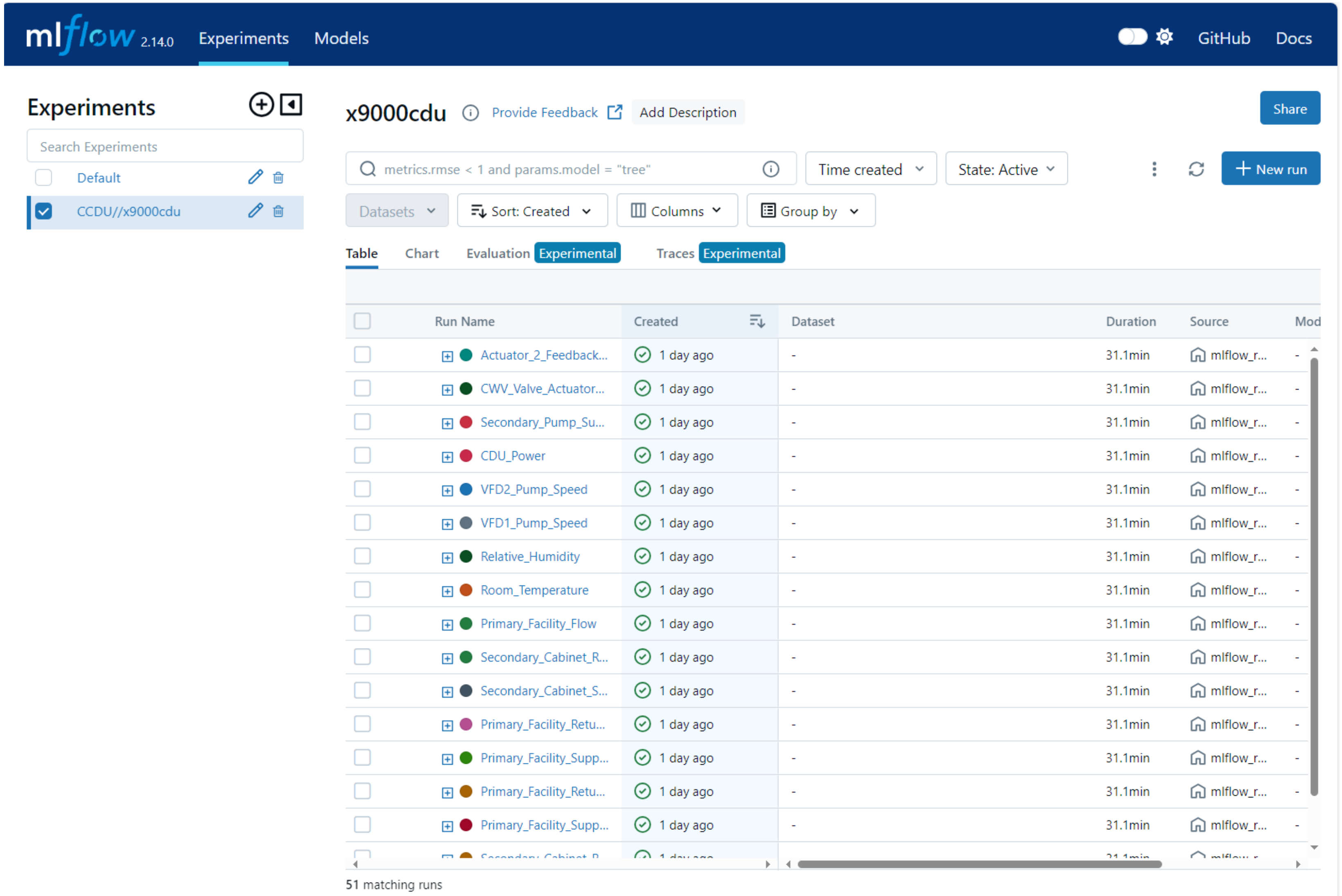
- Replace the cluster IP, port, and system name values in the following example:

```
external$ ssh -L 5000:10.31.223.14:5000 root@SYSTEM_NCN_DOMAIN_NAME
```

- Open a browser on a laptop or workstation and go to http://localhost:5000/

AIOps MLflow Experiments

- MLflow Tracking is organized around the concept of **runs**, which are executions of some piece of data science code, for example, a single python train.py execution
- Each run records metadata (various information about your run such as metrics, parameters, start and end times) and artifacts (output files from the run such as model weights, images, etc)
- An experiment groups together runs for a specific task
- The MLflow API and UI also let you create and search for experiments

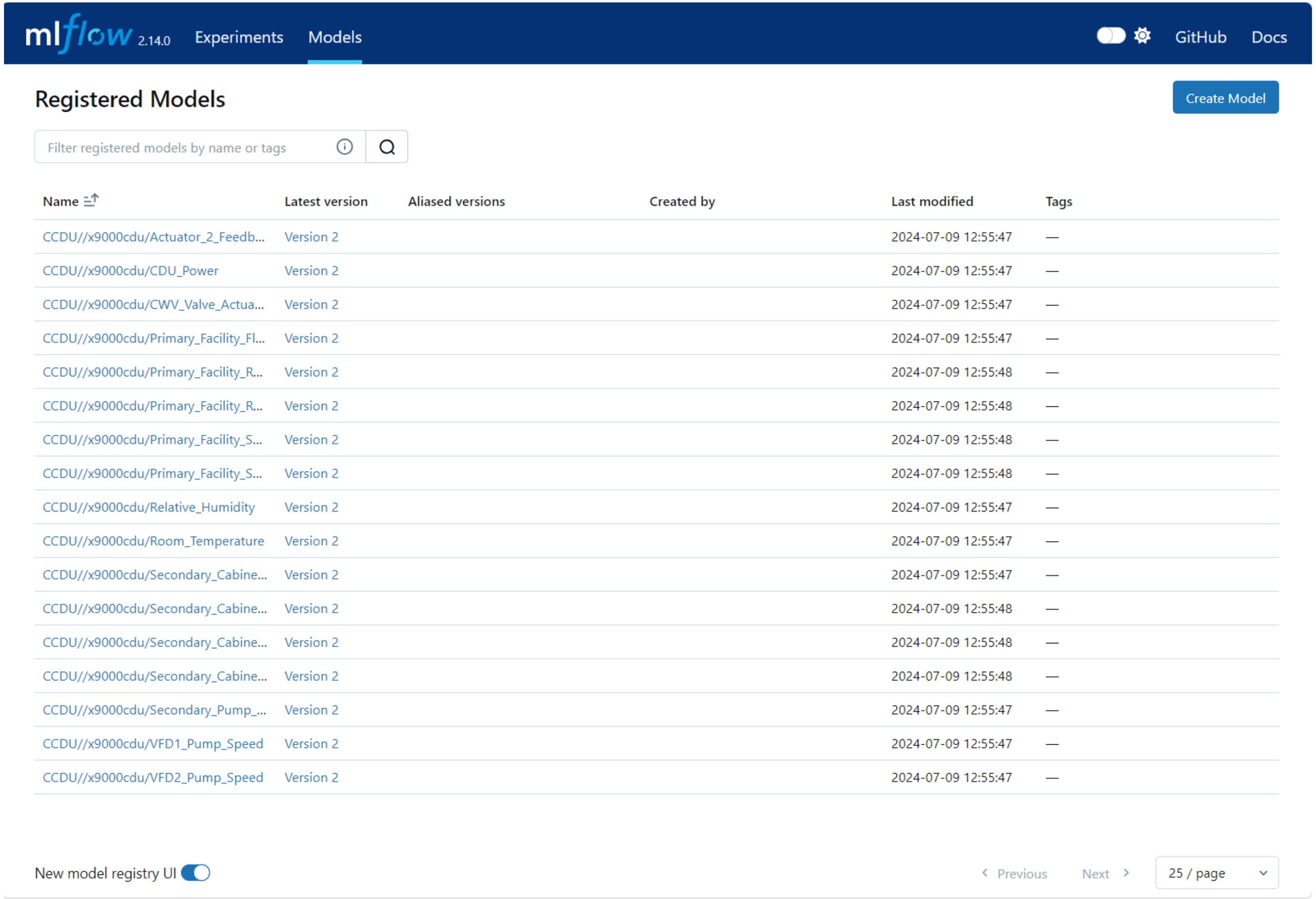


□	Run Name	Created	Dataset	Duration	Source	Mod
□	Actuator_2_Feedback...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	CWV_Valve_Actuator...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Secondary_Pump_Su...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	CDU_Power	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	VFD2_Pump_Speed	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	VFD1_Pump_Speed	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Relative_Humidity	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Room_Temperature	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Primary_Facility_Flow	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Secondary_Cabinet_R...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Secondary_Cabinet_S...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Primary_Facility_Retu...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Primary_Facility_Supp...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Primary_Facility_Retu...	✓ 1 day ago	-	31.1min	mlflow_r...	-
□	Primary_Facility_Supp...	✓ 1 day ago	-	31.1min	mlflow_r...	-

51 matching runs

AIOps MLflow Models

- The MLflow Model Registry
  - centralized model repository
  - a user interface
  - set of APIs that enable you to manage the full lifecycle of MLflow Models

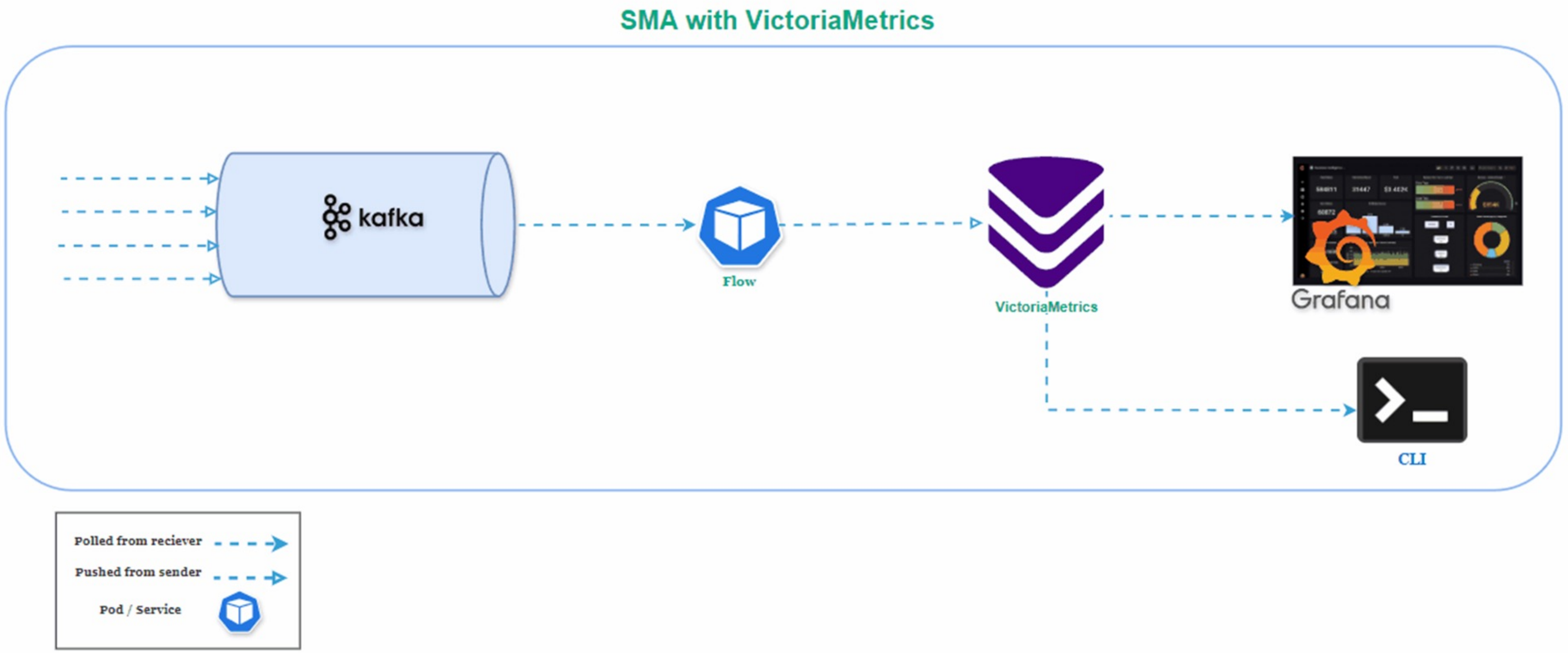


Registered Models

Name	Latest version	Aliased versions	Created by	Last modified	Tags
CCDU//x9000cdu/Actuator_2_Feedb...	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/CDU_Power	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/CWV_Valve_Actua...	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/Primary_Facility_Fl...	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/Primary_Facility_R...	Version 2			2024-07-09 12:55:48	—
CCDU//x9000cdu/Primary_Facility_R...	Version 2			2024-07-09 12:55:48	—
CCDU//x9000cdu/Primary_Facility_S...	Version 2			2024-07-09 12:55:48	—
CCDU//x9000cdu/Primary_Facility_S...	Version 2			2024-07-09 12:55:48	—
CCDU//x9000cdu/Relative_Humidity	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/Room_Temperature	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/Secondary_Cabine...	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/Secondary_Cabine...	Version 2			2024-07-09 12:55:48	—
CCDU//x9000cdu/Secondary_Cabine...	Version 2			2024-07-09 12:55:48	—
CCDU//x9000cdu/Secondary_Cabine...	Version 2			2024-07-09 12:55:48	—
CCDU//x9000cdu/Secondary_Pump_...	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/VFD1_Pump_Speed	Version 2			2024-07-09 12:55:47	—
CCDU//x9000cdu/VFD2_Pump_Speed	Version 2			2024-07-09 12:55:47	—

SMA Flow

- Flow
  - Transforms Kafka messages into records of the Prometheus Exposition Format
  - Feeds the transformed data to Victoria-Metrics
- Replaces PMDB Postgres database persister and the LDMS persister from pre-CSM 1.6
- Supported metrics
  - Telemetry/Slingshot
  - LDMS
  - Redfish



SMA Flow Metrics

```
ncn# kubectl get svc -n sma | grep sma-hms-flow
sma-hms-flow  ClusterIP   10.26.204.25  <none> 8000/TCP,8001/TCP,8002/TCP 5d21h
```

- Telemetry/Slingshot metrics

```
ncn# curl
 http://10.26.204.25:8000/metrics
flow_consumed_messages 167376167
flow_transformed_messages 167376167
flow_samples_parsed 2476797578
flow_samples_written 2476797580
flow_transform_errors 0
flow_write_errors 1
flow_flush_errors 3
flow_reconnect_attempts 0
flow_line_too_long 0
```

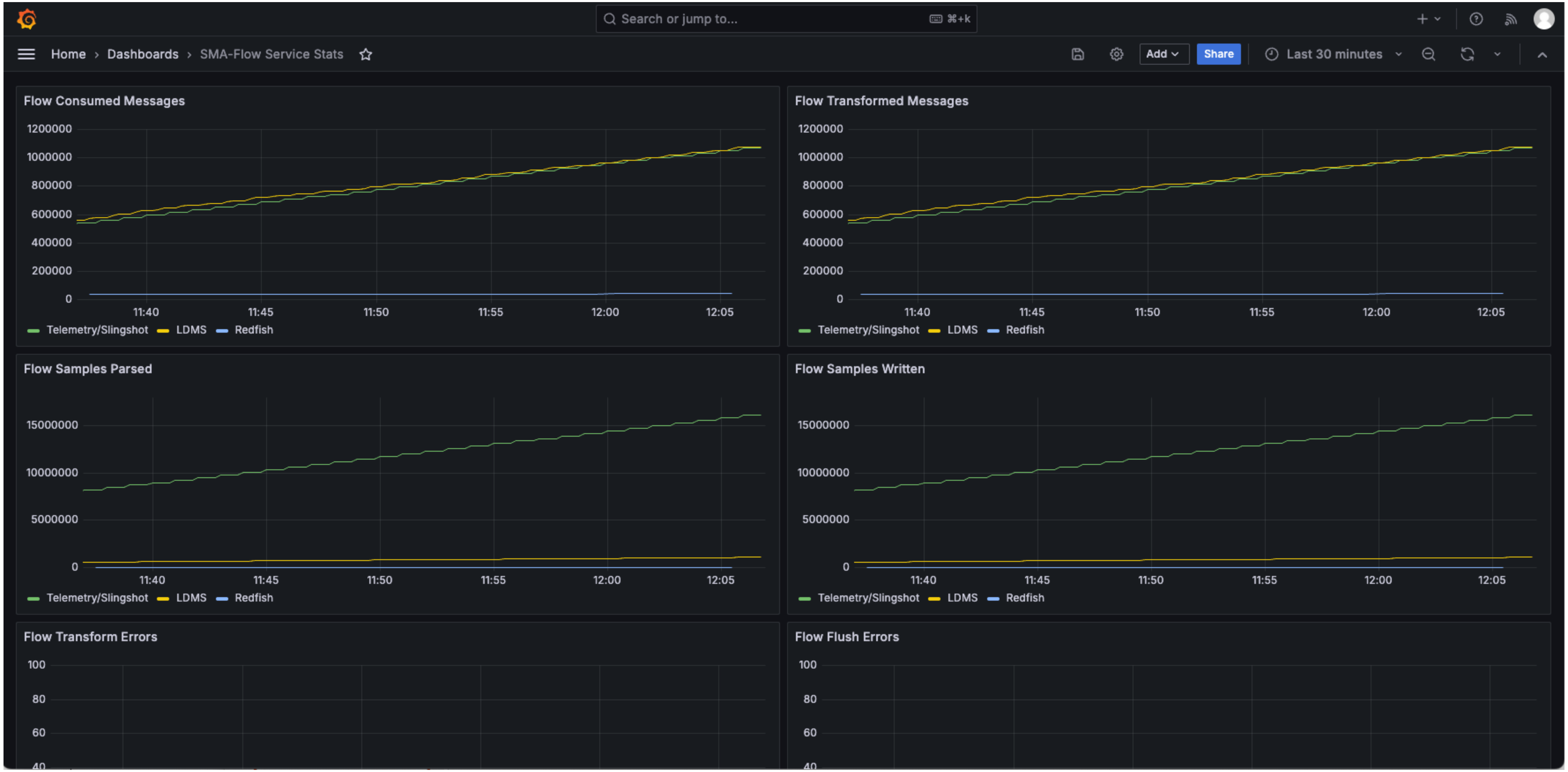
- LDMS metrics

```
ncn# curl
 http://10.26.204.25:8001/metrics
flow_consumed_messages 182959289
flow_transformed_messages 182959289
flow_samples_parsed 182959289
flow_samples_written 182959292
flow_transform_errors 0
flow_write_errors 1
flow_flush_errors 4
flow_reconnect_attempts 0
flow_line_too_long 0
```

- Redfish metrics

```
ncn# curl
 http://10.26.204.25:8002/metrics
flow_consumed_messages 40460
flow_transformed_messages 40460
flow_samples_parsed 40460
flow_samples_written 40460
flow_transform_errors 0
flow_write_errors 1
flow_flush_errors 4
flow_reconnect_attempts 0
flow_line_too_long 0
```

SMA-Flow Service Stats Dashboard



VictoriaMetrics (VMUI)

- vmselect handles read requests
  - vmselect UI can be accessed by using SSH port forwarding
- Use kubectl command to get the SERVICE-IP of sma-vm-cluster-vmselect service

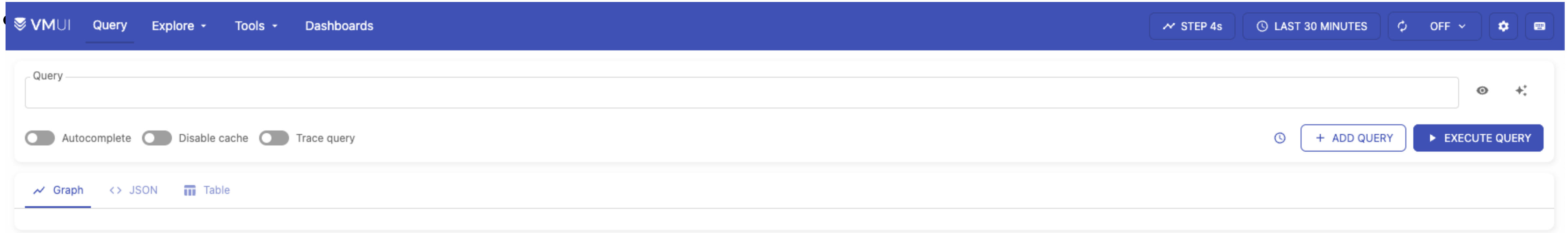
```
ncn# kubectl get svc -n sma sma-vm-cluster-vmselect
NAME                        TYPE        CLUSTER-IP     EXTERNAL-IP   PORT(S)    AGE
sma-vm-cluster-vmselect     ClusterIP   10.17.220.11   <none>        8481/TCP   6d2h
```

- Use sma-vm-cluster-vmselect service name and 8481 port number to port forward

```
ncn# kubectl port-forward -n sma service/sma-vm-cluster-vmselect 8481:8481
```

- Use SSH port-forwarding using the service IP of the service

```
external# ssh -L 8481:10.17.220.11:8481 root@SYSTEM_NCN_DOMAIN_NAME
```



Kafka Producers and Consumers – Controller Telemetry

Topics	Source	Producer	Consumer
cray-telemetry-temperature cray-telemetry-voltage cray-telemetry-power cray-telemetry-energy cray-telemetry-fan cray-telemetry-pressure cray-telemetry-humidity cray-telemetry-liquidflow cray-telemetry-frequency cray-telemetry-powerfactor cray-telemetry-percent cray-telemetry-metrics	Redfish The telemetry sources are: "sC"(Switch Controller) "nC"(Node Controller) "cC"(Cabinet Controller) "River"	Cray-HMS-Collector	Telemetry Topics Filter

Kafka Producers and Consumers – Controller Telemetry Filtered

Topics	Source	Producer	Consumer
cray-telemetry-temperature-filtered cray-telemetry-voltage-filtered cray-telemetry-power-filtered cray-telemetry-energy-filtered cray-telemetry-fan-filtered cray-telemetry-pressure-filtered cray-telemetry-humidity-filtered cray-telemetry-liquidflow-filtered cray-telemetry-frequency-filtered cray-telemetry-powerfactor-filtered cray-telemetry-percent-filtered cray-telemetry-metrics-filtered	Cray HMS Collector	Telemetry Topics Filter	Flow

Kafka Producers and Consumers – Fabric Telemetry, LDMS, CDU

Topics	Source	Producer	Consumer
cray-fabric-telemetry cray-fabric-perf-telemetry cray-fabric-crit-telemetry	Redfish The telemetry sources are: Fabric Telemetry Fabric Performance Fabric Critical	Cray-HMS-Collector	Flow
cray-fabric-health	Redfish The telemetry source is: Fabric Health	Cray-HMS-Collector	Logstash
cray-node	LDMS Collector systemd service	LDMS Aggregator	Flow
cray-dmtf-resource-event	Redfish Events	Cray-HMS-Collector	Flow and Logstash
cray-cdu cray-pdu cray-cdu-event	CDU and PDU devices	PCIM (Power Cooling Infrastructure Manager) Uses SNMP to collect data and event from CDU and PDU devices	Logstash

Kafka Producers and Consumers – AIOps Telemetry, Logs

Topics	Source	Producer	Consumer
aiops-anomaly-detection aiops-anomaly-forecast aiops-anomaly-notifications aiops-fabric-perf-anomaly-detection aiops-fabric-temp-anomaly-detection	AIOps Anomaly Detection and Forecast Models	AIOps	Logstash
alerts	The main sources are Slingshot switches and CDU Alerts There may be other sources	Slingshot Alerting, CDU Alerting and others	Logstash
cray-logs-containers	Kubernetes containers	fluentbit collector	Logstash
cray-logs-syslogs	syslogd on all the nodes	fluentbit aggregator	Logstash

Logs

- Kubernetes logs
- Console logs
- SMA OpenSearch

Kubernetes – View pod logs

- View the pod’s log

```
ncn# kubectl -n NAMESPACE logs PODNAME
```

- Follow log continuously with the -f|–follow=true option

```
ncn# kubectl -n NAMESPACE logs PODNAME -f
```

- You may want to view the logs of more than one pod when there are multiple replicas of the same pod
  - Find the pods using labels

```
ncn# kubectl -n services get pods -l app.kubernetes.io/name=cray-bos
NAME                           READY   STATUS    RESTARTS   AGE
cray-bos-67c5f989f8-kb4rt      2/2     Running   0          18d
cray-bos-67c5f989f8-qf7ff      2/2     Running   0          18d
```

  - Watch their logs

```
ncn# kubectl -n services logs -l app.kubernetes.io/name=cray-bos
```

Containerized console access

- ConMan is a serial console management program designed to support a large number of console devices and simultaneous users
- cray-console uses ConMan for interactive remote console access and console log collection
  - Automatically detects nodes which have been added or removed
  - Shared filesystem in Ceph for all cray-console pods to easily view log data
  - Console log data sent to SMA for other log processing
  - Dynamic autoscaling number of cray-console-node pods for size of system
    - Minimally, two pods are started
    - The number of PODs is scaled on
      - 750 Liquid-cooled nodes and/or 2000 "River" nodes
        - The Liquid-cooled nodes each require an ssh connection, so numbers are different
- Log locations:
  - Logs visible in any `cray-console-node-`***`x`*** pod
  - Node logs: `/var/log/conman/console.XNAME`
  - ConMan daemon logs: `/var/log/conman.log`

```
ncn# kubectl get pods -A |grep cray-console
services              cray-console-data-5cd59677d9-lf4f4
services              cray-console-data-postgres-0
services              cray-console-data-postgres-1
services              cray-console-data-postgres-2
services              cray-console-node-0
services              cray-console-node-1
services              cray-console-operator-7f9894f657-5psn5
```

Console logs with cray-console-node

```
ncn# kubectl get pods -A |grep console-node
services              cray-console-node-0        3/3     Running        1      62d
services              cray-console-node-1        3/3     Running        0      68d
ncn# kubectl -it exec -n services cray-console-node-1 -c cray-console-node -- ls /var/log/conman
console.x1000c0s1b0n0  console.x1000c3s3b0n0   console.x3000c0s20b4n0
console.x1000c0s1b0n1  console.x1000c3s3b0n1   console.x3000c0s23b1n0
console.x1000c0s1b1n0  console.x1000c3s3b1n0   console.x3000c0s23b2n0
console.x1000c0s1b1n1  console.x1000c3s3b1n1   console.x3000c0s23b3n0
console.x1000c0s5b0n0  console.x1000c5s5b0n0   console.x3000c0s23b4n0
console.x1000c0s5b0n1  console.x1000c5s5b0n1   console.x3000c0s25b1n0
console.x1000c0s5b1n0  console.x1000c5s5b1n0   console.x3000c0s25b2n0
console.x1000c0s5b1n1  console.x1000c5s5b1n1   console.x3000c0s25b3n0
console.x1000c0s7b0n0  console.x1000c7s7b0n0   console.x3000c0s25b4n0
```

Each pod sees all the console files, only one cray-console-node pod is managing that node and writing its log file

```
ncn# kubectl -it exec -n services cray-console-node-1 -c cray-console-node -- \
tail -f /var/log/conman/console.x1000c0s1b0n0
```

Can view log without entering pod

```
ncn# kubectl -it exec -n services cray-console-node-1 -c cray-console-node -- /bin/bash
cray-console-node-1-pod# grep –i error /var/log/conman/console.x1000c0s1b0n0
```

Can view log by entering pod

Interactive console example (long)

- To join the console, use `conman -j`
  - Retrieve the `cray-console-operator` pod ID

    ```
    ncn# CONPOD=$(kubectl get pods -n services \
        -o wide|grep cray-console-operator|awk '{print $1}')
    ncn# echo $CONPOD
    cray-console-operator-79bf95964-qpcpp
    ```

  - Set the `XNAME` variable to the xname of the node whose console you wish to open

    ```
    ncn# XNAME=x1000c0s0b0n0
    ```

  - Find the `cray-console-node` pod that is managing that node

    ```
    ncn# NODEPOD=$(kubectl -n services exec $CONPOD -c cray-console-operator \
    -- sh -c "/app/get-node $XNAME" | jq .podname | sed 's/"//g')
    ncn# echo $NODEPOD
    cray-console-node-1
    ```

  - Connect to the node's console using ConMan on the `cray-console-node` pod you found

    ```
    ncn# kubectl exec -it -n services $NODEPOD -- conman -j $XNAME
    <ConMan> Connection to console [x1000c0s0b0] opened.
    nid000001 login:
    ```

- To exit console use `&.` command

Interactive console example (short)

- Bash function to join console

```
ncn# ConsoleJ ()
{
    XNAME=$@;
    CONPOD=$(kubectl get pods -n services -o wide|grep cray-console-operator|awk '{print $1}');
    NODEPOD=$(kubectl -n services -c cray-console-operator exec $CONPOD -- sh -c "/app/get-node $XNAME" | jq .podname | tr -d '"');
    echo conpod = $CONPOD nodepod = $NODEPOD;
    kubectl exec -it -n services $NODEPOD -c cray-console-node -- conman -j $XNAME
}
ncn# ConsoleJ x1000c0s0b0n0
<ConMan> Connection to console [x1000c0s0b0n0] opened.
nid000001 login:
```

- To exit console use `&.` command
- To view the console read-only instead of joining it read-write, use `conman –m $XNAME`

NodeBMC console logs

- In addition to console logs available via conman in cray-console-node pods, check the NodeBMC
  - Console logs for nodes on blades in (liquid-cooled) Olympus cabinets can be accessed from the nodeController (BMC)
    - For node x1000c0s0b1n0, connect to its nodeController

```
ncn# ssh x1000c0s0b1
x1000c0s0b1> cd /var/log/n0
x1000c0s0b1> tail current
x1000c0s0b1> grep -i error current
```

  - Console logs for nodes in (air-cooled) River cabinets can be accessed from the node BMC using a Web GUI to the BMC

SMA OpenSearch

- SMA OpenSearch enables
  - Viewing all logs from all nodes and Kubernetes pods
  - Sorting and searching through log information from multiple sources to help troubleshoot issues
- View and analyze system logs in the web UI
- Access sma-dashboards
  1. Determine the external domain name by running the following command on any NCN:

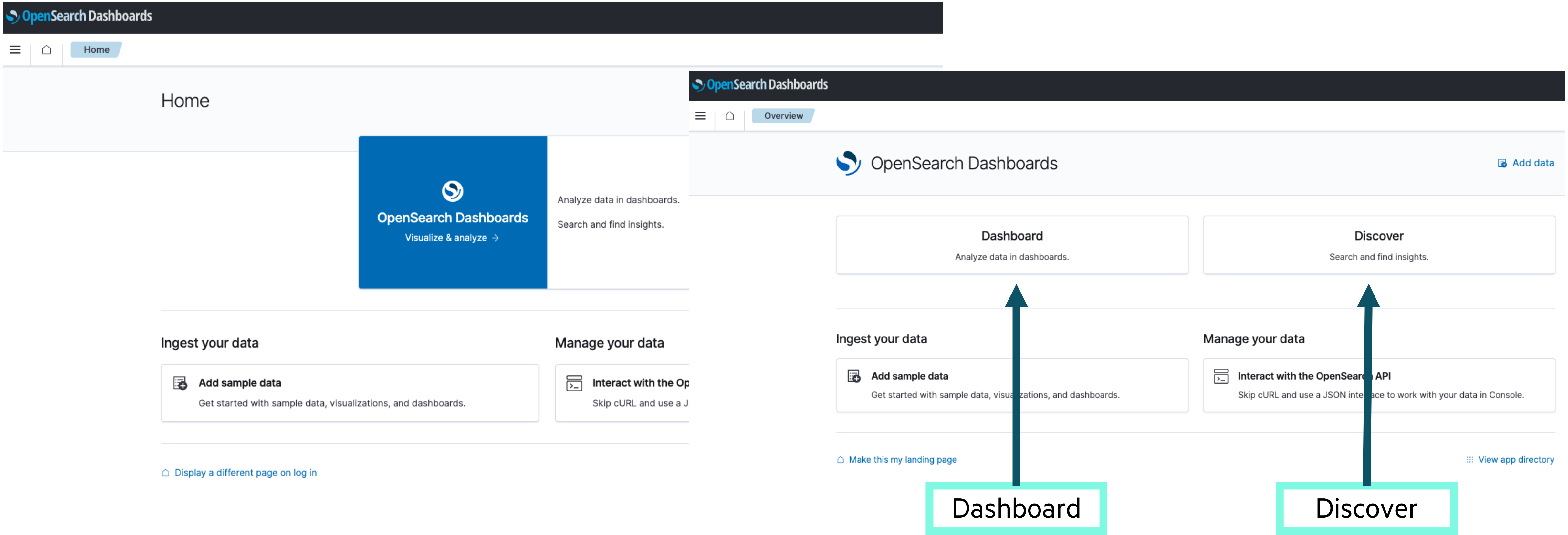
     ```
     ncn-m001# kubectl get secret site-init -n loftsman \
     -o jsonpath='{.data.customizations\.yaml}' | base64 -d | grep "external:"
     external: SYSTEM_DOMAIN_NAME
     ```

  2. Navigate to the following URL in a web browser:

     ```
     https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME/
     ```

  3. Login by entering a valid username and password
- See Opensearch documentation to further explore and analyze the system logs
  - https://opensearch.org/docs/latest/about/

OpenSearch – sma-dashboards



OpenSearch Dashboards

Home

OpenSearch Dashboards
Visualize & analyze →

Analyze data in dashboards.

Search and find insights.

Ingest your data

Add sample data
Get started with sample data, visualizations, and dashboards.

Manage your data

Interact with the Op
Skip cURL and use a J

⌂ Display a different page on log in

OpenSearch Dashboards

OpenSearch Dashboards

Add data

Dashboard
Analyze data in dashboards.

Discover
Search and find insights.

Ingest your data

Add sample data
Get started with sample data, visualizations, and dashboards.

Manage your data

Interact with the OpenSearch API
Skip cURL and use a JSON interface to work with your data in Console.

⌂ Make this my landing page

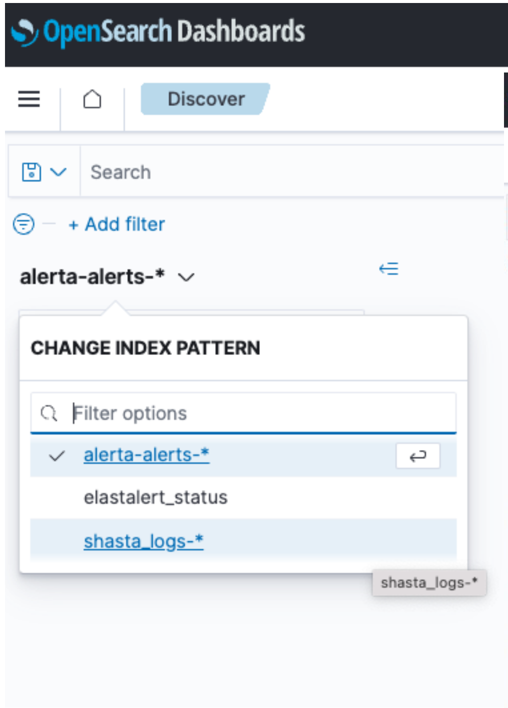
View app directory

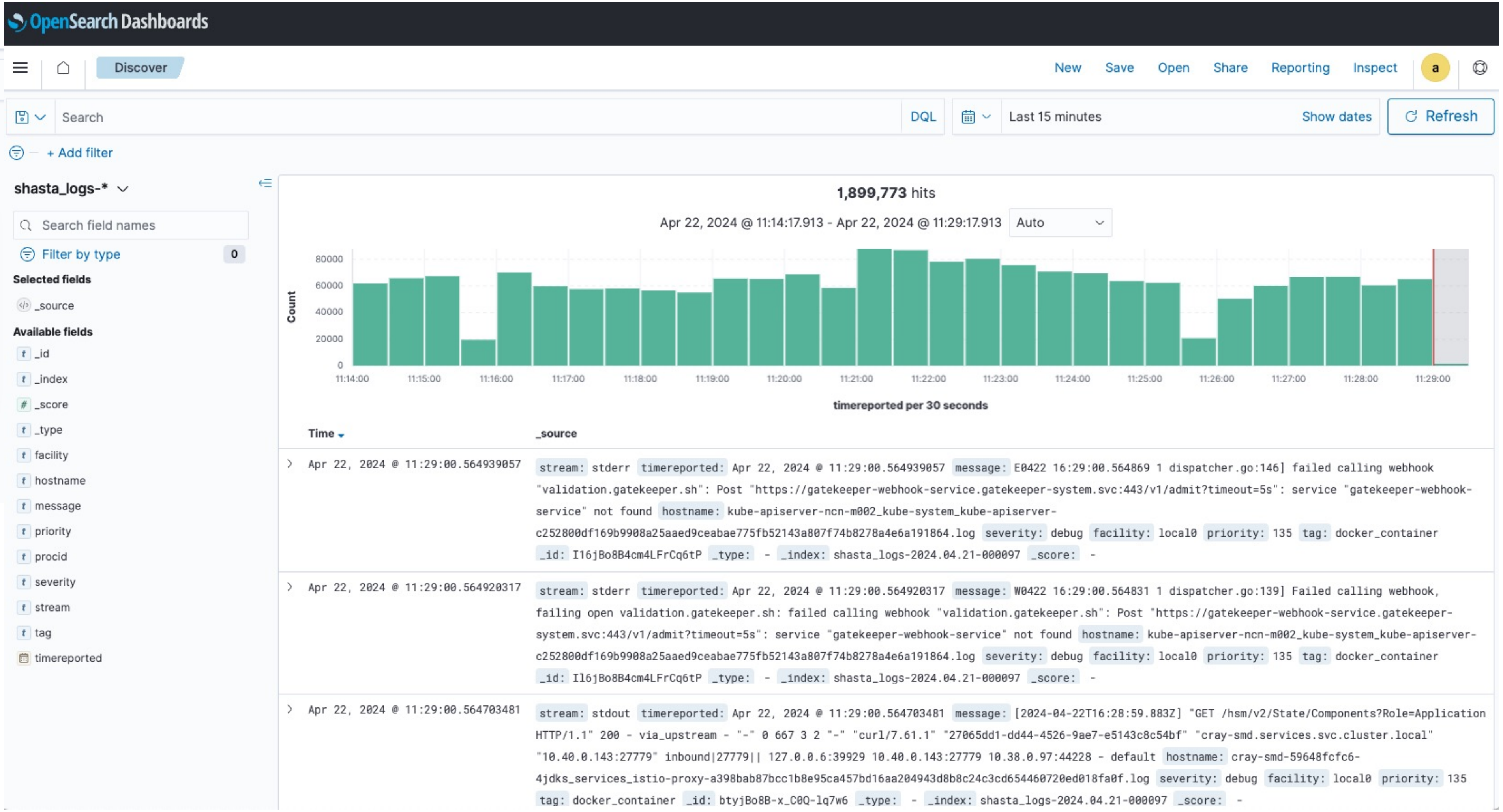
Dashboard

Discover

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

OpenSearch - Discover

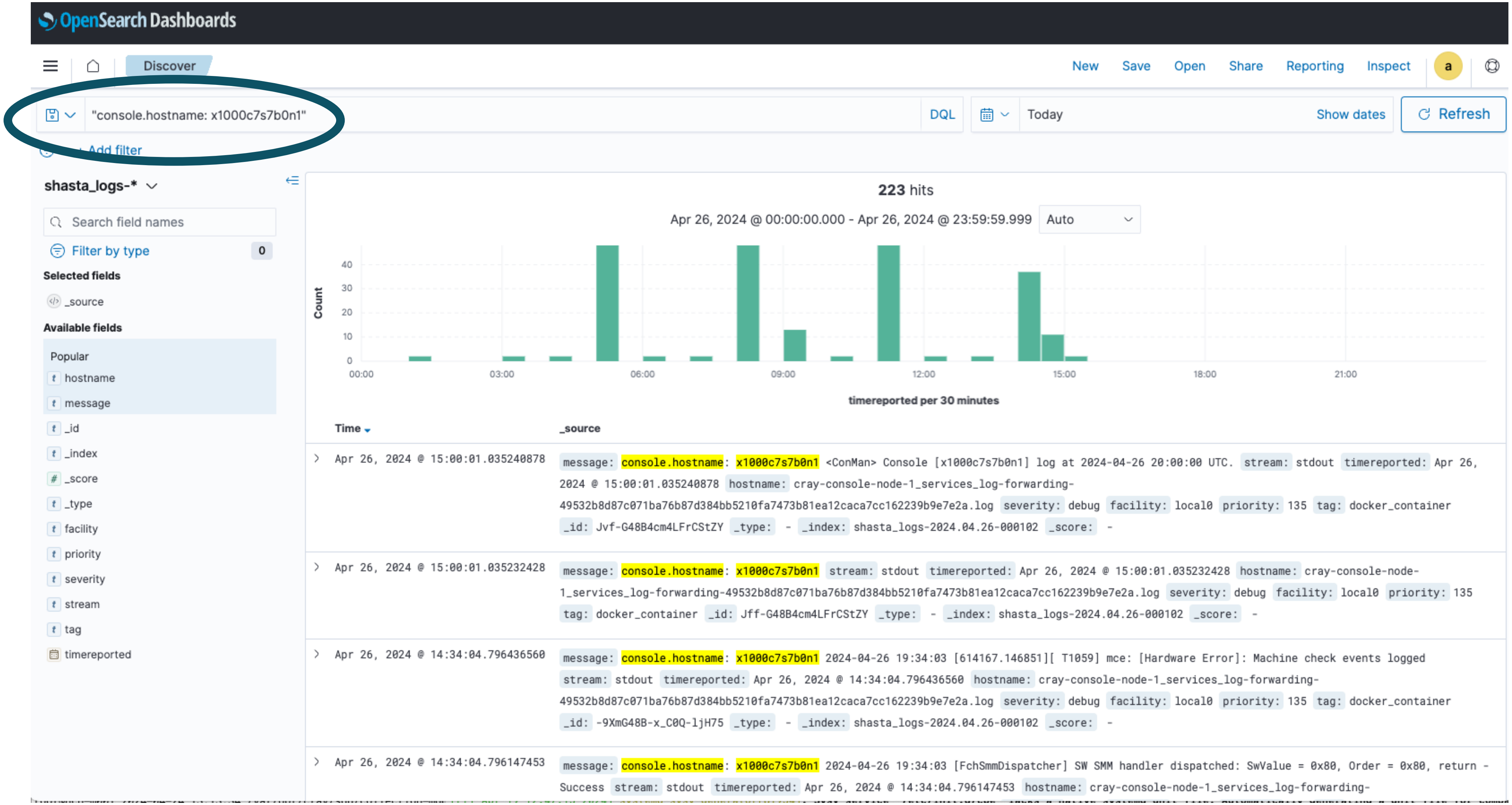




```
Apr 22, 2024 @ 11:29:00.564939057   stream: stderr timereported: Apr 22, 2024 @ 11:29:00.564939057 message: E0422 16:29:00.564869 1 dispatcher.go:146] failed calling webhook
                                    "validation.gatekeeper.sh": Post "https://gatekeeper-webhook-service.gatekeeper-system.svc:443/v1/admit?timeout=5s": service "gatekeeper-webhook-
                                    service" not found hostname: kube-apiserver-ncn-m002_kube-system_kube-apiserver-
                                    c252800df169b9908a25aaed9ceabae775fb52143a807f74b8278a4e6a191864.log severity: debug facility: local0 priority: 135 tag: docker_container
                                    _id: I16jBo8B4cm4LFrCq6tP _type: - _index: shasta_logs-2024.04.21-000097 _score: -

Apr 22, 2024 @ 11:29:00.564920317   stream: stderr timereported: Apr 22, 2024 @ 11:29:00.564920317 message: W0422 16:29:00.564831 1 dispatcher.go:139] Failed calling webhook,
                                    failing open validation.gatekeeper.sh: failed calling webhook "validation.gatekeeper.sh": Post "https://gatekeeper-webhook-service.gatekeeper-
                                    system.svc:443/v1/admit?timeout=5s": service "gatekeeper-webhook-service" not found hostname: kube-apiserver-ncn-m002_kube-system_kube-apiserver-
                                    c252800df169b9908a25aaed9ceabae775fb52143a807f74b8278a4e6a191864.log severity: debug facility: local0 priority: 135 tag: docker_container
                                    _id: I16jBo8B4cm4LFrCq6tP _type: - _index: shasta_logs-2024.04.21-000097 _score: -

Apr 22, 2024 @ 11:29:00.564703481   stream: stdout timereported: Apr 22, 2024 @ 11:29:00.564703481 message: [2024-04-22T16:28:59.883Z] "GET /hsm/v2/State/Components?Role=Application
                                    HTTP/1.1" 200 - via_upstream - "-" 0 667 3 2 "-" "curl/7.61.1" "27065dd1-dd44-4526-9ae7-e5143c8c54bf" "cray-smd.services.svc.cluster.local"
                                    "10.40.0.143:27779" inbound|27779|| 127.0.0.6:39929 10.40.0.143:27779 10.38.0.97:44228 - default hostname: cray-smd-59648fcfc6-
                                    4jdks_services_istio-proxy-a398bab87bcc1b8e95ca457bd16aa204943d8b8c24c3cd654460720ed018fa0f.log severity: debug facility: local0 priority: 135
                                    tag: docker_container _id: btyjBo8B-x_C0Q-1q7w6 _type: - _index: shasta_logs-2024.04.21-000097 _score: -
```

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

OpenSearch – Discover – Console Logs



"console.hostname: x1000c7s7b0n1"

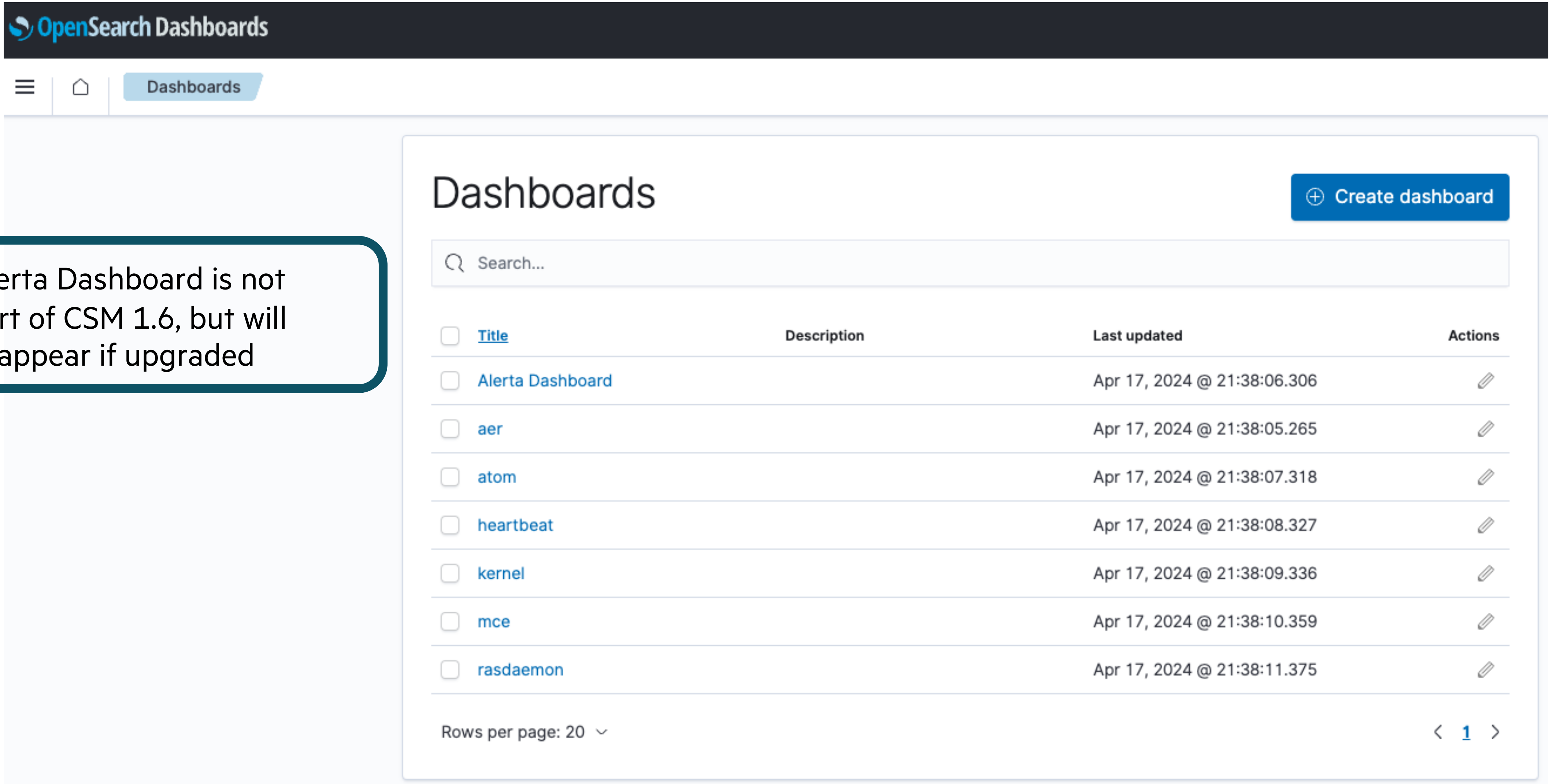
223 hits

Apr 26, 2024 @ 00:00:00.000 - Apr 26, 2024 @ 23:59:59.999

```
Apr 26, 2024 @ 15:00:01.035240878   message: console.hostname: x1000c7s7b0n1 <ConMan> Console [x1000c7s7b0n1] log at 2024-04-26 20:00:00 UTC. stream: stdout timereported: Apr 26, 2024 @ 15:00:01.035240878 hostname: cray-console-node-1_services_log-forwarding-49532b8d87c071ba76b87d384bb5210fa7473b81ea12caca7cc162239b9e7e2a.log severity: debug facility: local0 priority: 135 tag: docker_container _id: Jvf-G48B4cm4LFrCStZY _type: - _index: shasta_logs-2024.04.26-000102 _score: -

Apr 26, 2024 @ 15:00:01.035232428   message: console.hostname: x1000c7s7b0n1 stream: stdout timereported: Apr 26, 2024 @ 15:00:01.035232428 hostname: cray-console-node-1_services_log-forwarding-49532b8d87c071ba76b87d384bb5210fa7473b81ea12caca7cc162239b9e7e2a.log severity: debug facility: local0 priority: 135 tag: docker_container _id: Jff-G48B4cm4LFrCStZY _type: - _index: shasta_logs-2024.04.26-000102 _score: -

Apr 26, 2024 @ 14:34:04.796436560   message: console.hostname: x1000c7s7b0n1 2024-04-26 19:34:03 [614167.146851][ T1059] mce: [Hardware Error]: Machine check events logged stream: stdout timereported: Apr 26, 2024 @ 14:34:04.796436560 hostname: cray-console-node-1_services_log-forwarding-49532b8d87c071ba76b87d384bb5210fa7473b81ea12caca7cc162239b9e7e2a.log severity: debug facility: local0 priority: 135 tag: docker_container _id: -9XmG48B-x_C0Q-1jH75 _type: - _index: shasta_logs-2024.04.26-000102 _score: -

Apr 26, 2024 @ 14:34:04.796147453   message: console.hostname: x1000c7s7b0n1 2024-04-26 19:34:03 [FchSmmDispatcher] SW SMM handler dispatched: SwValue = 0x80, Order = 0x80, return - Success stream: stdout timereported: Apr 26, 2024 @ 14:34:04.796147453 hostname: cray-console-node-1_services_log-forwarding-
```

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

OpenSearch Dashboards



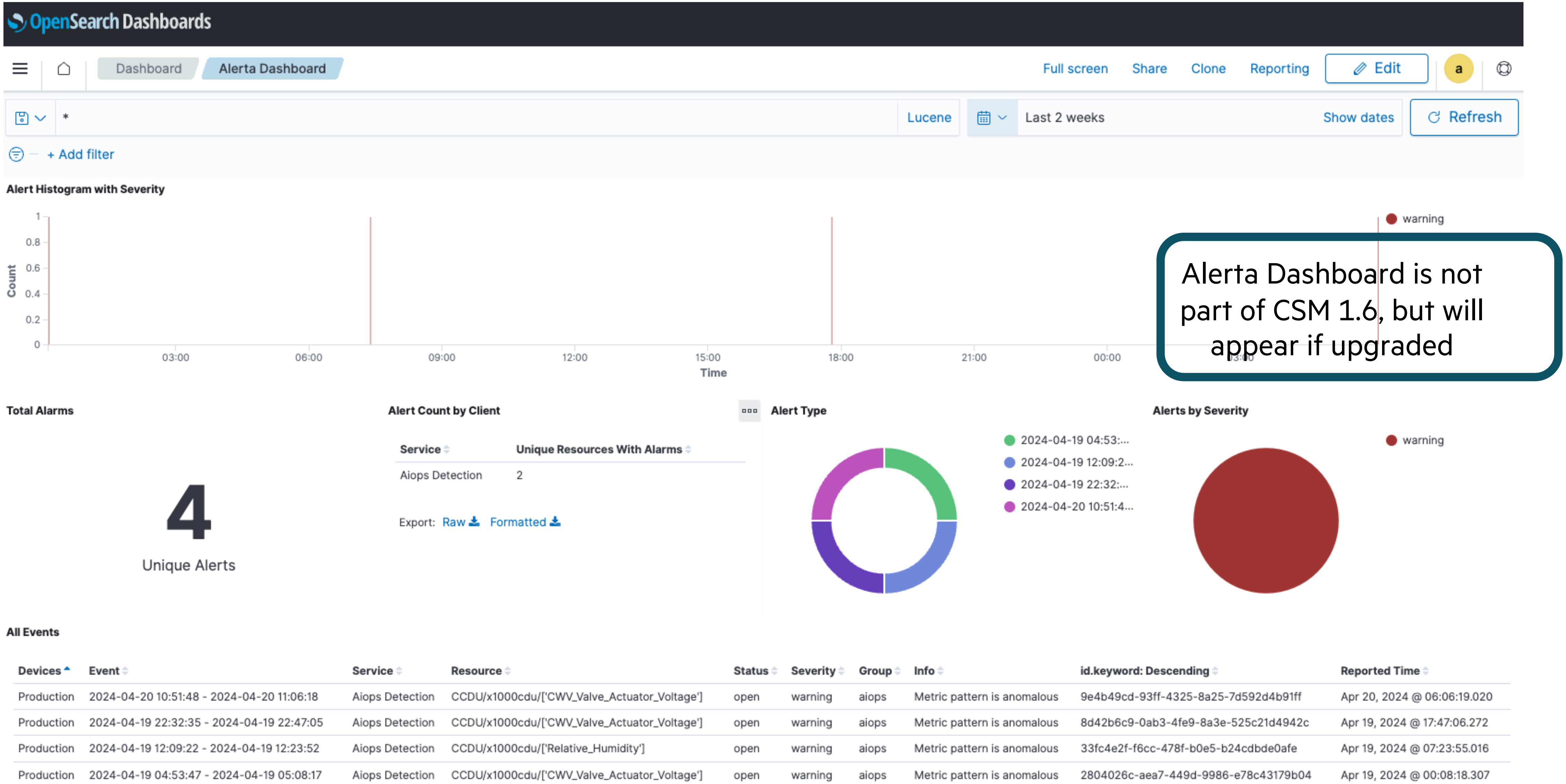
Dashboards

□	Title	Description	Last updated	Actions
□	Alerta Dashboard		Apr 17, 2024 @ 21:38:06.306	✎
□	aer		Apr 17, 2024 @ 21:38:05.265	✎
□	atom		Apr 17, 2024 @ 21:38:07.318	✎
□	heartbeat		Apr 17, 2024 @ 21:38:08.327	✎
□	kernel		Apr 17, 2024 @ 21:38:09.336	✎
□	mce		Apr 17, 2024 @ 21:38:10.359	✎
□	rasdaemon		Apr 17, 2024 @ 21:38:11.375	✎

Alerta Dashboard is not part of CSM 1.6, but will appear if upgraded

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

OpenSearch Dashboards - Alerta dashboard



Total Alarms

4

Unique Alerts

Alert Count by Client

Service	Unique Resources With Alarms
Aiops Detection	2

Export: Raw Formatted

All Events

Devices	Event	Service	Resource	Status	Severity	Group	Info	id.keyword: Descending	Reported Time
Production	2024-04-20 10:51:48 - 2024-04-20 11:06:18	Aiops Detection	CCDU/x1000cdu/['CWV_Valve_Actuator_Voltage']	open	warning	aiops	Metric pattern is anomalous	9e4b49cd-93ff-4325-8a25-7d592d4b91ff	Apr 20, 2024 @ 06:06:19.020
Production	2024-04-19 22:32:35 - 2024-04-19 22:47:05	Aiops Detection	CCDU/x1000cdu/['CWV_Valve_Actuator_Voltage']	open	warning	aiops	Metric pattern is anomalous	8d42b6c9-0ab3-4fe9-8a3e-525c21d4942c	Apr 19, 2024 @ 17:47:06.272
Production	2024-04-19 12:09:22 - 2024-04-19 12:23:52	Aiops Detection	CCDU/x1000cdu/['Relative_Humidity']	open	warning	aiops	Metric pattern is anomalous	33fc4e2f-f6cc-478f-b0e5-b24cdbde0afe	Apr 19, 2024 @ 07:23:55.016
Production	2024-04-19 04:53:47 - 2024-04-19 05:08:17	Aiops Detection	CCDU/x1000cdu/['CWV_Valve_Actuator_Voltage']	open	warning	aiops	Metric pattern is anomalous	2804026c-aea7-449d-9986-e78c43179b04	Apr 19, 2024 @ 00:08:18.307

Alerta Dashboard is not part of CSM 1.6, but will appear if upgraded

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

OpenSearch Dashboards – AER, ATOM, heartbeat

Dashboard	Short description	Long description	Visualization and Search name
aer	AER corrected	Corrected Advanced Error Reporting messages from PCI Express devices on each node	aer-corrected
aer	AER fatal	Fatal Advanced Error Reporting messages from PCI Express devices on each node	aer-fatal
atom	ATOM failures	Application Task Orchestration and Management tests are run on a node when a job finishes. Test failures are logged	atom-failed
atom	ATOM admindown	Application Task Orchestration and Management test failures can result in nodes being marked admindown. An admindown node is not available for job launch	atom-admindown
heartbeat	Heartbeat loss events	Heartbeat loss event messages reported by the hbtd pods that monitor for heartbeats across nodes in the system	heartbeat

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

OpenSearch Dashboards - kernel

Dashboard	Short description	Long description	Visualization and Search name
kernel	Kernel assertions	The kernel software performs a failed assertion when some condition represents a serious fault so the node goes down	kassertions
kernel	Kernel panics	The kernel panics when something is seriously wrong so the node goes down	kernel-panic
kernel	Lustre bugs (LBUGs)	The Lustre software in the kernel stack performs a failed assertion when some condition related to file system logic represents a serious fault so the node goes down	lbug
kernel	CPU stalls	CPU stalls (Read-Copy-Update stalls where software in the kernel stack holds onto memory for too long) are serous conditions that can reduce node performance, and sometimes cause a node to go down. Read-Copy-Update is a vital aspect of kernel performance and rather esoteric	cpu-stall
kernel	Out of memory	An Out Of Memory (OOM) condition has occurred so the kernel must select an expendable process to kill to continue or if there is no expendable process the node usually goes down in some manner	oom

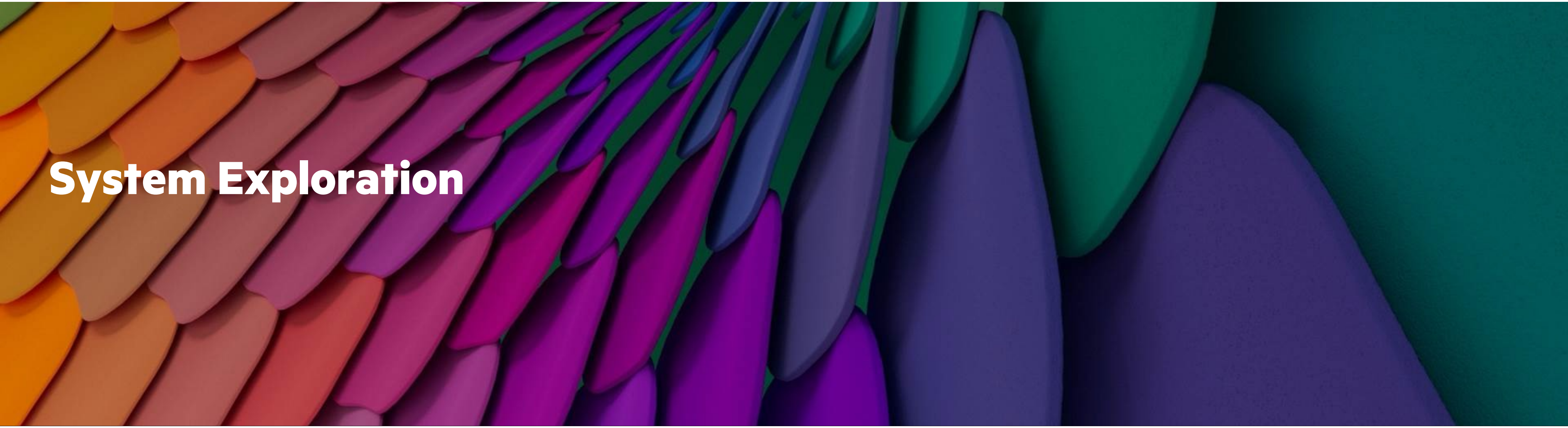
https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

OpenSearch Dashboards – MCE and rasdaemon

Dashboard	Short description	Long description	Visualization and Search name
mce	MCE	Machine Check Exceptions (MCE) are errors detected at the processor level	mce
rasdaemon	rasdaemon errors	Errors from the rasdaemon service on nodes. The rasdaemon service is the Reliability, Availability, and Serviceability Daemon, and it is intended to collect all hardware error events reported by the Linux kernel, including PCI and MCE errors. This may include certain HSN errors in the future	rasdaemon-error
rasdaemon	rasdaemon messages	All messages from the rasdaemon service on nodes	rasdaemon

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

System Exploration



System Exploration

Logs and OpenSearch

Telemetry and Grafana

Alerts and Alertmanager and cm health alertman

Logs and OpenSearch

- Scanning Console Logs
- OpenSearch for Console Logs
- OpenSearch for CSM Pod Logs

Scan console logs – CSM

- HSN NIC PCIe lane degrade messages
- MCEs for memory DIMM errors
- NO-CARRIER on HSN NIC
- FATAL BIOS ERROR COUNT
- How do you find problems across many nodes with CSM?
  - From cray-console-node pod find pattern from a bad node’s console log

    ```
    ncn# kubectl -it exec -n services cray-console-node-0 -c cray-console-node – bash
    cray-console-node> cd /var/log/conman
    cray-console-node> vi console.x1000c2s3b1n0
    ```

  - Search for that pattern in all the console logs in that cabinet

    ```
    cray-console-node> grep pattern console.x1000c2s3b1n0
    cray-console-node> grep pattern console.x1000*
    cray-console-node> exit
    ```

  - Search from outside the cray-console-node pod for all nodes

    ```
    ncn# kubectl -it exec -n services cray-console-node-1 -c cray-console-node \
    -- grep pattern /var/log/conman/console.*
    ```

  - From management node
    - ssh to BMC of liquid-cooled node
    - Find pattern in /var/log/n0/current (for node 0 on the BMC) or /var/log/n1/current (for node 1 on the BMC)
    - Then pdsh to group of BMCs for liquid-cooled nodes to grep for pattern in /var/log/n*/current
  - Run hwtriage tool from management node to analyze each node considered bad
    - It will find MCEs and PCIe lane degrades (and other issues)
    - See CSM Diags Administration Guide
- Query the collected console logs using OpenSearch

Troubleshooting Nodes Not Ready – CSM

- Check node state with HSM or SAT commands for which nodes are not in the READY state to find missing nodes

  ```
  ncn# sat status --filter role=compute --filter state!=ready
  ```

- For any nodes in the OFF state, check the power logs on their BMC (nodecontroller) for power up faults
  - Olympus nodes: (example x1000c0s0b0n1)

    ```
    ncn# ssh x1000c0s0b0
    x1000c0s0b0> egrep "\(partially powered up\)|Stopped at PS|already fully powered up" /var/log/powerfault_up.Node1
    ```

  - Refer to hardware service team if these message patterns are found
- For any nodes in the ON state, check the console logs for each node
  - After inspecting a few console logs to find a pattern, grep for that pattern in all the console logs conman

    ```
    ncn# kubectl -it exec -n services cray-console-node-1 -c cray-console-node -- grep pattern /var/l
    og/conman/console.*
    ```

  - Repeat that grep command but count how many nodes have a problem matching that pattern

    ```
    ncn# kubectl -it exec -n services cray-console-node-1 -c cray-console-node -- grep pattern /var/l
    og/conman/console.* | wc -l
    ```

  - Do the console messages indicate whether the node failed:
    - To get a DHCP response and start downloading ipxe.efi binary? – check cray-kea and cray-ipxe pods
    - To contact BSS for the iPXE boot script once ipxe.efi started running? – check BSS pods and whether the MACaddr used has valid data in BSS
    - To download the boot artifacts (kernel, initrd) from S3? Check for presence of them with "`cray artifacts`" command
    - To configure the HSN NICs as tmp0, tmp1, etc while in Dracut? Check Slingshot fabric manager configuration and edge port health
  - Did the node have any of these errors in the console log? Some are kernel panics, some show node dropping into UEFI shell, and some indicate hardware errors

    ```
    startup.nsh|ernel panic|any other key to continue|Enter for maintenance|Entering emergency mode|query
    intf hsn|WHEA: Detected Memory Error|ASSERT|Shell\>|Unable to get TLV for interface|Machine Check"
    ```

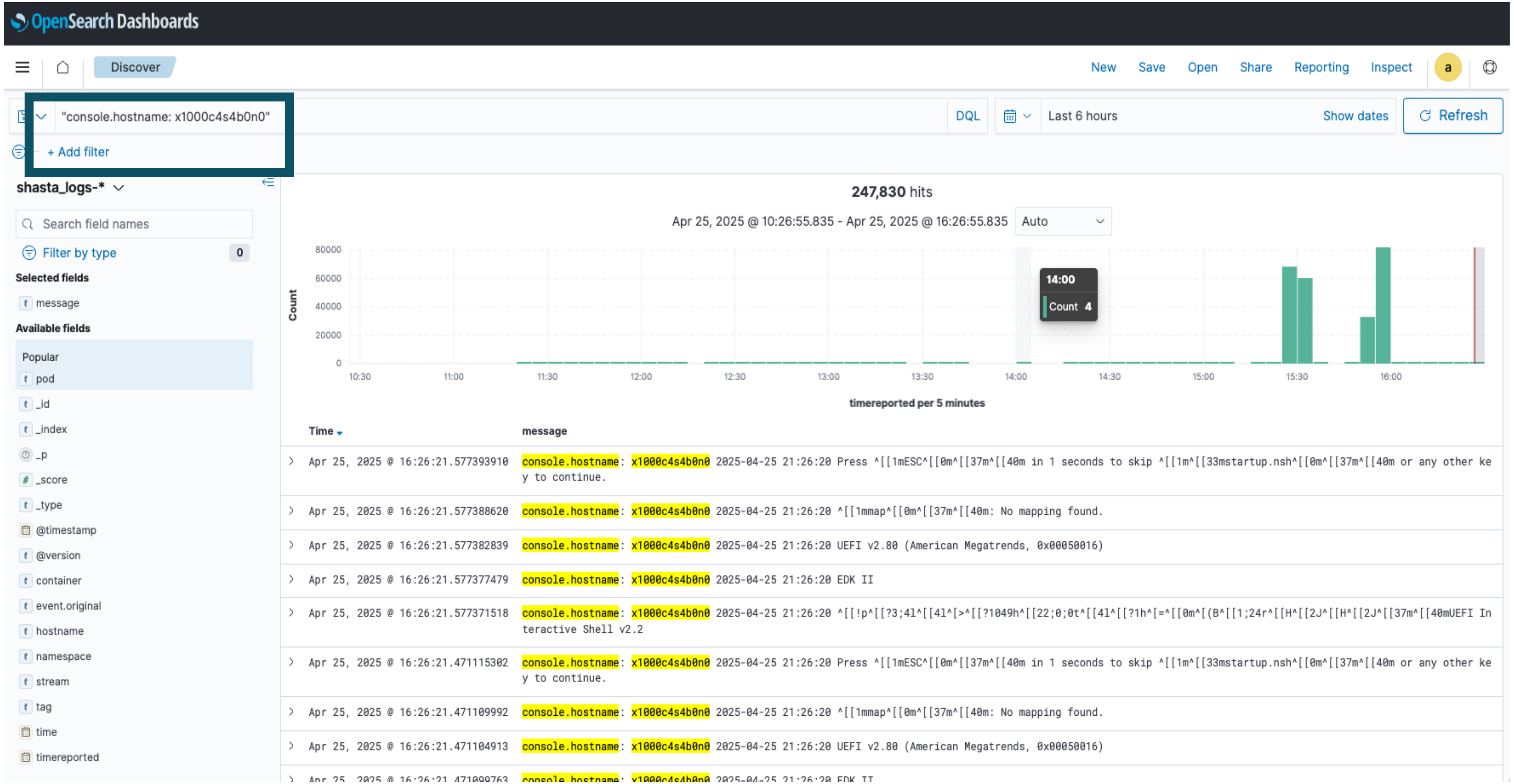
PCIe Lane Degrade

- Console log shows a PCIe lane degrade for HSN NIC1 from Gen4x16 to Gen4x8 on a node
- How often has this happened on this node?

```
2025-04-25 20:58:55 CRAY PCIE Device Information
2025-04-25 20:58:55 BC2:D00:F0 - High-speed NIC0
2025-04-25 20:58:55     VID: 0x17DB
2025-04-25 20:58:55     DID: 0x0501
2025-04-25 20:58:55     Link Width (Capable): x16
2025-04-25 20:58:55     Link Speed (Capable): Gen4
2025-04-25 20:58:55     Link Width (Actual):  x16
2025-04-25 20:58:55     Link Speed (Actual):  Gen4
2025-04-25 20:58:55
2025-04-25 20:58:55 B01:D00:F0 - High-speed NIC1
2025-04-25 20:58:55     VID: 0x17DB
2025-04-25 20:58:55     DID: 0x0501
2025-04-25 20:58:55     Link Width (Capable): x16
2025-04-25 20:58:55     Link Speed (Capable): Gen4
2025-04-25 20:58:55     Link Width (Actual):  x8
2025-04-25 20:58:55     Link Speed (Actual):  Gen4
2025-04-25 20:58:55
2025-04-25 20:58:55 {"NonFatalUefiDegradedPcieDeviceError": ["High-speed NIC1", "0x17DB", "0x0501", "01", "00", "0", "4", "4", "16", "8"]}
2025-04-25 20:58:55 Type:    7
2025-04-25 20:58:55 Flags:  0x01
2025-04-25 20:58:55 Port80: 0xB000A691
2025-04-25 20:58:56 ApicId: 0x00000000
2025-04-25 20:58:56 CpuNum: 0
2025-04-25 20:58:56 RTC:    04/25/2025 20:58:55-00:00
2025-04-25 20:58:56 ErrorData: Degraded Device High-speed NIC1
2025-04-25 20:58:56 DeviceName:  High-speed NIC1
2025-04-25 20:58:56 Description: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act: Gen4, Width Exp: x16
   Act:x8.
2025-04-25 20:58:56 ERROR ENTRY - END
```

PCIe Lane Degrade – x1000c4s4b0n0

- Node dropped into the UEFI Interactive Shell instead of booting fully
- Look for frequency of the PCIe Lane Degrade



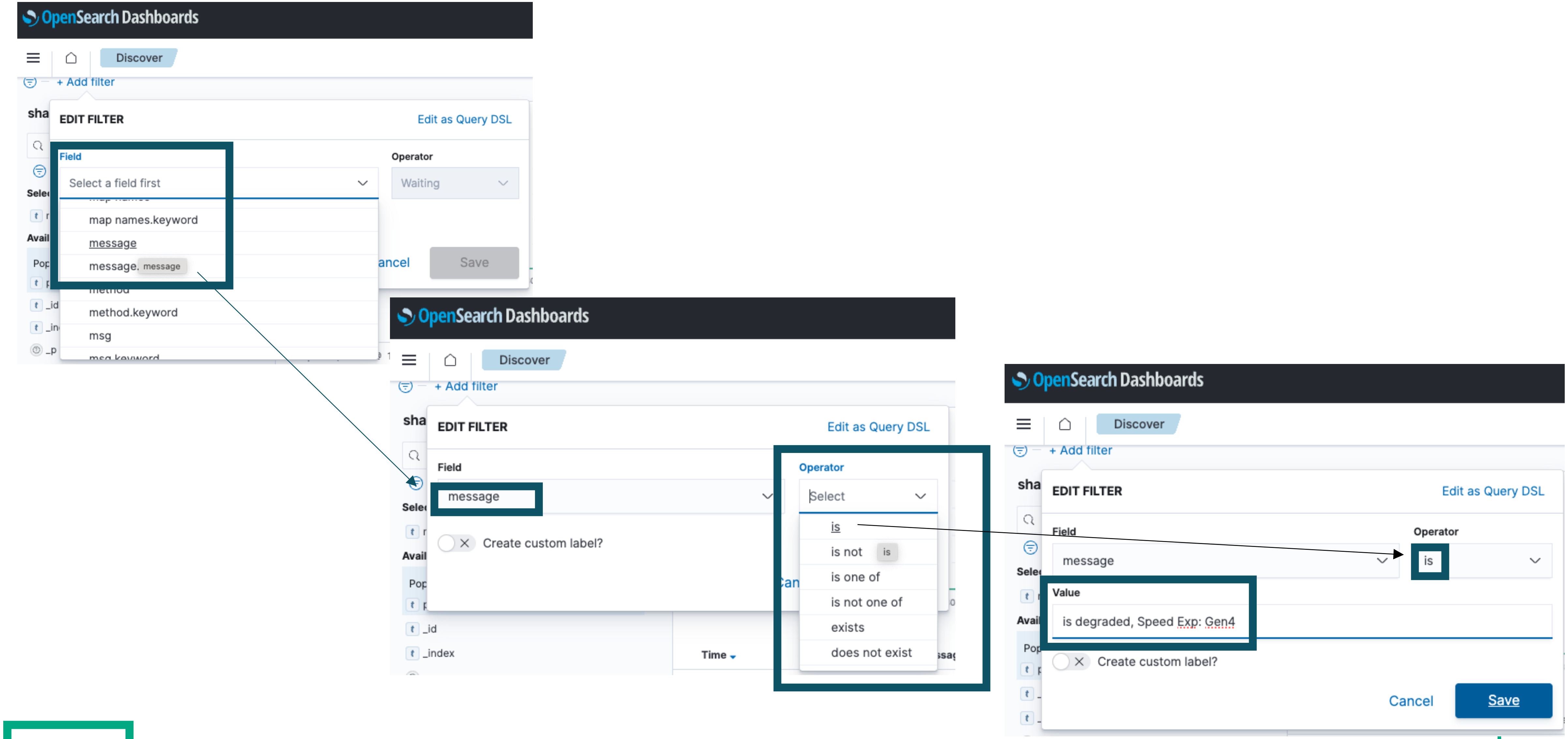
"console.hostname: x1000c4s4b0n0"

247,830 hits

Apr 25, 2025 @ 10:26:55.835 - Apr 25, 2025 @ 16:26:55.835

```
Time                              message
Apr 25, 2025 @ 16:26:21.577393910 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 Press ^[[1mESC^[[0m^[[37m^[[40m in 1 seconds to skip ^[[1m^[[33mstartup.nsh^[[0m^[[37m^[[40m or any other ke
                                  y to continue.
Apr 25, 2025 @ 16:26:21.577388620 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 ^[[1mmap^[[0m^[[37m^[[40m: No mapping found.
Apr 25, 2025 @ 16:26:21.577382839 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 UEFI v2.80 (American Megatrends, 0x00050016)
Apr 25, 2025 @ 16:26:21.577377479 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 EDK II
Apr 25, 2025 @ 16:26:21.577371518 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 ^[[!p^[[?3;4l^[[4l^[>^[[?1049h^[[22;0;0t^[[4l^[[?1h^[=^[[0m^[(B^[[1;24r^[[H^[[2J^[[H^[[2J^[[37m^[[40mUEFI In
                                  teractive Shell v2.2
Apr 25, 2025 @ 16:26:21.471115302 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 Press ^[[1mESC^[[0m^[[37m^[[40m in 1 seconds to skip ^[[1m^[[33mstartup.nsh^[[0m^[[37m^[[40m or any other ke
                                  y to continue.
Apr 25, 2025 @ 16:26:21.471109992 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 ^[[1mmap^[[0m^[[37m^[[40m: No mapping found.
Apr 25, 2025 @ 16:26:21.471104913 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 UEFI v2.80 (American Megatrends, 0x00050016)
Apr 25, 2025 @ 16:26:21.471099763 console.hostname: x1000c4s4b0n0 2025-04-25 21:26:20 EDK II
```

PCIe Lane Degrade – x1000c4s4b0n0 Filter



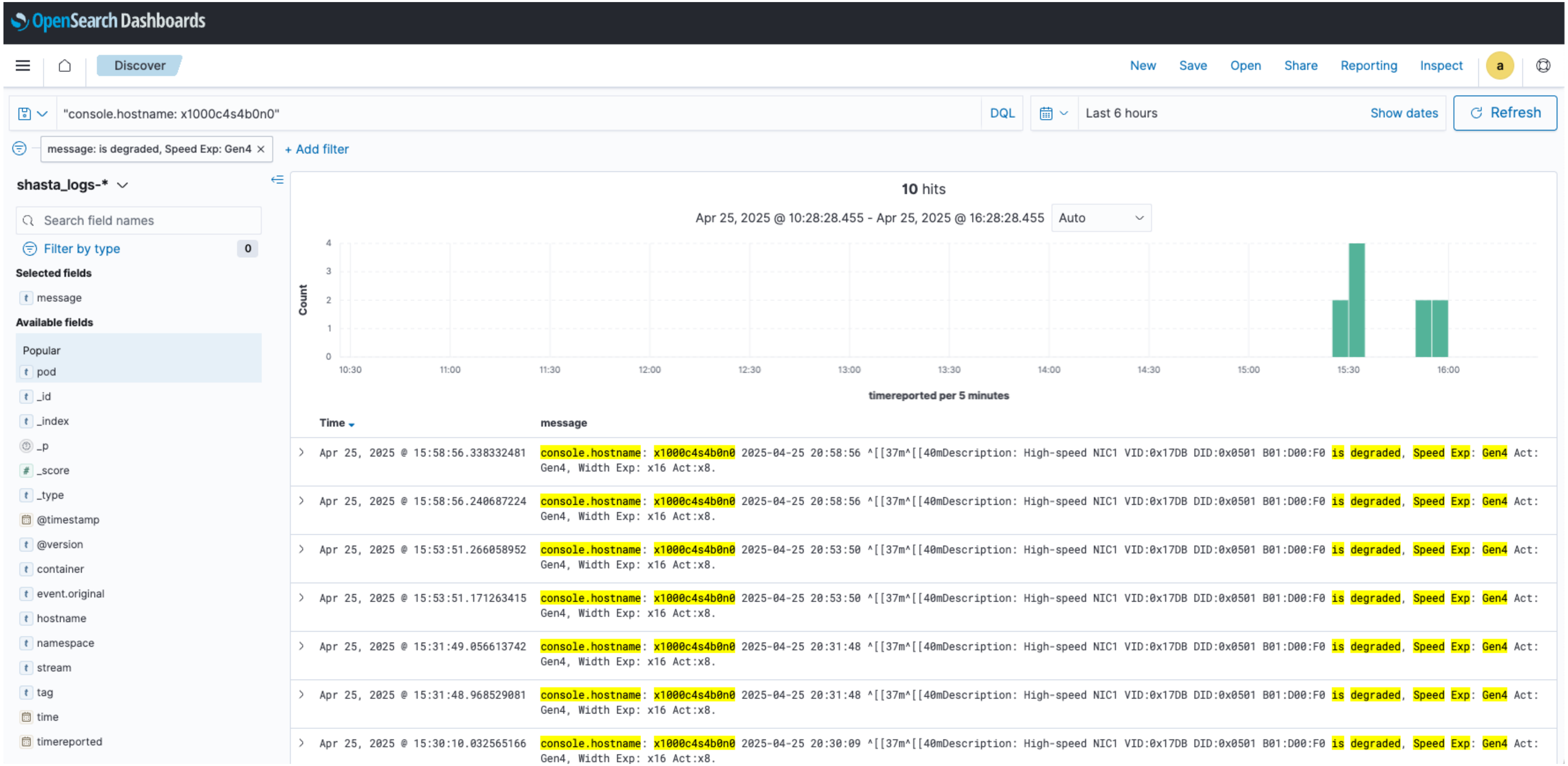
EDIT FILTER

Field: message

Operator: is

Value: is degraded, Speed Exp: Gen4

PCIe Lane Degrade – x1000c4s4b0n0 Frequency



"console.hostname: x1000c4s4b0n0"

message: is degraded, Speed Exp: Gen4 ×

10 hits

Apr 25, 2025 @ 10:28:28.455 - Apr 25, 2025 @ 16:28:28.455

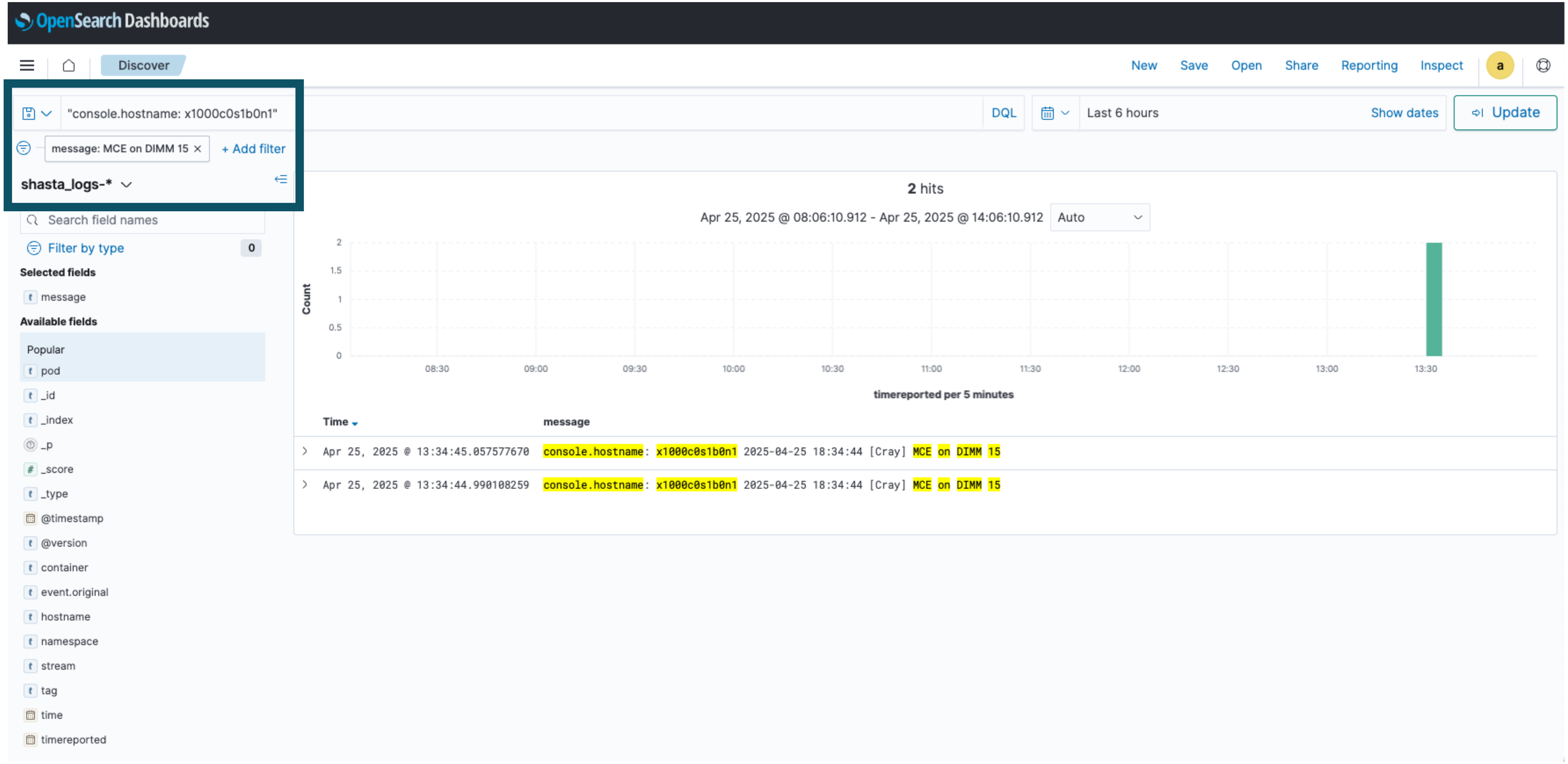
```
Apr 25, 2025 @ 15:58:56.338332481  console.hostname: x1000c4s4b0n0 2025-04-25 20:58:56 ^[[37m^[[40mDescription: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act:
                                   Gen4, Width Exp: x16 Act:x8.
Apr 25, 2025 @ 15:58:56.240687224  console.hostname: x1000c4s4b0n0 2025-04-25 20:58:56 ^[[37m^[[40mDescription: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act:
                                   Gen4, Width Exp: x16 Act:x8.
Apr 25, 2025 @ 15:53:51.266058952  console.hostname: x1000c4s4b0n0 2025-04-25 20:53:50 ^[[37m^[[40mDescription: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act:
                                   Gen4, Width Exp: x16 Act:x8.
Apr 25, 2025 @ 15:53:51.171263415  console.hostname: x1000c4s4b0n0 2025-04-25 20:53:50 ^[[37m^[[40mDescription: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act:
                                   Gen4, Width Exp: x16 Act:x8.
Apr 25, 2025 @ 15:31:49.056613742  console.hostname: x1000c4s4b0n0 2025-04-25 20:31:48 ^[[37m^[[40mDescription: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act:
                                   Gen4, Width Exp: x16 Act:x8.
Apr 25, 2025 @ 15:31:48.968529081  console.hostname: x1000c4s4b0n0 2025-04-25 20:31:48 ^[[37m^[[40mDescription: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act:
                                   Gen4, Width Exp: x16 Act:x8.
Apr 25, 2025 @ 15:30:10.032565166  console.hostname: x1000c4s4b0n0 2025-04-25 20:30:09 ^[[37m^[[40mDescription: High-speed NIC1 VID:0x17DB DID:0x0501 B01:D00:F0 is degraded, Speed Exp: Gen4 Act:
                                   Gen4, Width Exp: x16 Act:x8.
```

Machine Check Exception (MCE)

- Console log shows a correctable MCE for a DIMM on a node
- How often has this been happening on this node?
- Are there other nodes with a similar issue?

```
2025-04-25T00:18:11.06500 [546981.566380][T39869] mce: [Hardware Error]: Machine check events logged
2025-04-25T00:18:11.06501 [546981.570122][T39869] [Hardware Error]: Corrected error, no action required.
2025-04-25T00:18:11.06503 [546981.573638][T39869] [Hardware Error]: CPU:34 (17:31:0) MC17_STATUS[Over|CE|MiscV|AddrV|-|-|SyndV|CECC|-|-|Scrub]: 0xdc2041000000011b
2025-04-25T00:18:11.06504 [546981.579698][T39869] [Hardware Error]: Error Misc: 0x0000000000000000
2025-04-25T00:18:11.06505 [546981.582948][T39869] [Hardware Error]: Error Addr: 0x00000002335838c0
2025-04-25T00:18:11.06513 [546981.586198][T39869] [Hardware Error]: PPIN: 0x02b48c8ad411403f
2025-04-25T00:18:11.06514 [546981.589227][T39869] [Hardware Error]: IPID: 0x0000009600450f00, Syndrome: 0x348c08000a800400
2025-04-25T00:18:11.06515 [546981.593662][T39869] [Hardware Error]: Unified Memory Controller Ext. Error Code: 0
2025-04-25T00:18:11.06517 [546981.593693][T39869] EDAC MC1: 1 CE on dimm15 (csrow:0 channel:4 page:0x34b6b07 offset:0xc0 grain:64 syndrome:0x800)
2025-04-25T00:18:11.06518 [546981.602857][T39869] [Hardware Error]: cache level: L3/GEN, tx: GEN, mem-tx: RD
2025-04-25T18:34:47.55638 MCE to Software SMI error handler
2025-04-25T18:34:47.55642 CPU68 Local SMI Status = 0x00040000
2025-04-25T18:34:47.55644 [RAS]Local SMI Status: SmiSrcMca
2025-04-25T18:34:47.55645 Socket# 1, Ccd# 0, Ccx# 0, Core# 2, Thread# 0
2025-04-25T18:34:47.55646 MCA Bank Number : 17
2025-04-25T18:34:47.55647 MCA_STATUS : 0xDC2041000000011B
2025-04-25T18:34:47.55648 MCA_ADDR : 0x04000002335838C0
2025-04-25T18:34:47.55649 MCA_SYND : 0x348C08000A800400
2025-04-25T18:34:47.55650 MCA_MISC0 : 0xD01D0FFF01000000
2025-04-25T18:34:47.55651 MCA_MISC1 : 0xD01C0FF501000000
2025-04-25T18:34:47.55652 MCA_IPID : 0x0000009600450F00
2025-04-25T18:34:47.55654 [RAS]NormalizedAddr: 0x00000002335838C0
2025-04-25T18:34:47.55655 [RAS]NormalizedSocId: 0x1, DieId: 0x0, ChannelId: 0x4
2025-04-25T18:34:47.55656 translate_norm_to_dram_addr: noofbank = 1, noofrm = 0, noofrwolo = 6, noofrowhi = 0, noofcol = 5
2025-04-25T18:34:47.55657 translate_norm_to_dram_addr: RowLoBits = 16, RowHiBits = 0, ColBits = 10, CsBits = 0, BankBits = 4
2025-04-25T18:34:47.55659 ERROR ADDRESS : 0x00000002335838C0
2025-04-25T18:34:47.55661 ERROR ADDRESS LSB : 0x4
2025-04-25T18:34:47.55662 System Address Hi: 0x00000034
2025-04-25T18:34:47.55663 System Address Lo: 0xB6B070C0
2025-04-25T18:34:47.55664 DIMM Info (Chip Select): 0x0
2025-04-25T18:34:47.55665 DIMM Info (Bank): 0xB
2025-04-25T18:34:47.55666 DIMM Info (Row): 0x8CD6
2025-04-25T18:34:47.55667 DIMM Info (Column): 0x318
2025-04-25T18:34:47.55668 DIMM Info (rankmul): 0x0
2025-04-25T18:34:47.55669 MCA UMC Error:Type 0 Socket 1 Channel 4 DIMM 0
2025-04-25T18:34:47.55670 [Cray] MCE on DIMM 15
```

MCE on DIMM - x1000c0s1b0n1



"console.hostname: x1000c0s1b0n1"

message: MCE on DIMM 15

shasta_logs-*

Last 6 hours

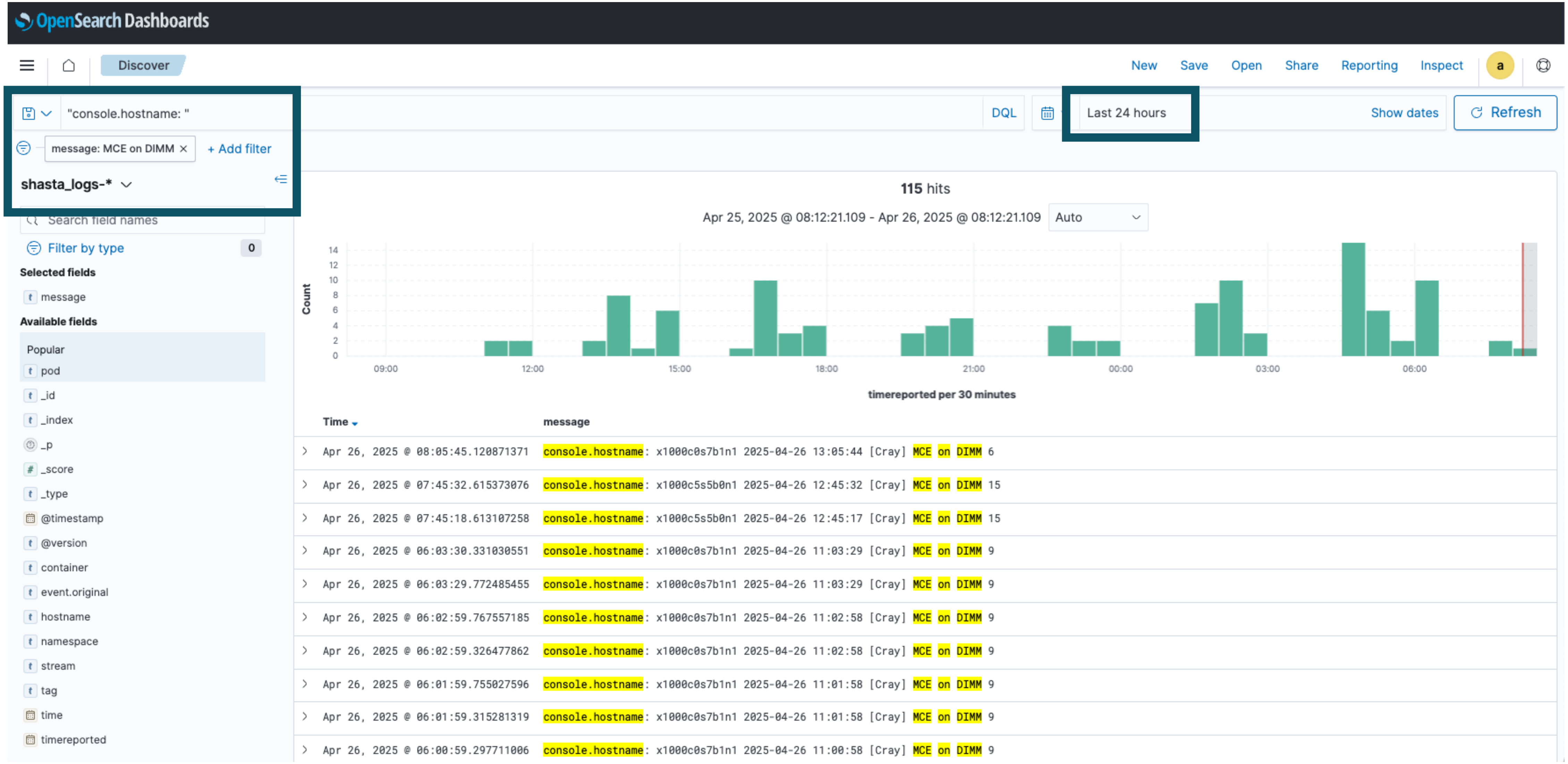
2 hits

Apr 25, 2025 @ 08:06:10.912 - Apr 25, 2025 @ 14:06:10.912

```
Time                                  message
Apr 25, 2025 @ 13:34:45.057577670     console.hostname: x1000c0s1b0n1 2025-04-25 18:34:44 [Cray] MCE on DIMM 15
Apr 25, 2025 @ 13:34:44.990108259     console.hostname: x1000c0s1b0n1 2025-04-25 18:34:44 [Cray] MCE on DIMM 15
```

https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

MCE on DIMM - All Nodes



"console.hostname: "

message: MCE on DIMM

shasta_logs-*

Last 24 hours

115 hits

Apr 25, 2025 @ 08:12:21.109 - Apr 26, 2025 @ 08:12:21.109 Auto

```
Time                                message
Apr 26, 2025 @ 08:05:45.120871371   console.hostname: x1000c0s7b1n1 2025-04-26 13:05:44 [Cray] MCE on DIMM 6
Apr 26, 2025 @ 07:45:32.615373076   console.hostname: x1000c5s5b0n1 2025-04-26 12:45:32 [Cray] MCE on DIMM 15
Apr 26, 2025 @ 07:45:18.613107258   console.hostname: x1000c5s5b0n1 2025-04-26 12:45:17 [Cray] MCE on DIMM 15
Apr 26, 2025 @ 06:03:30.331030551   console.hostname: x1000c0s7b1n1 2025-04-26 11:03:29 [Cray] MCE on DIMM 9
Apr 26, 2025 @ 06:03:29.772485455   console.hostname: x1000c0s7b1n1 2025-04-26 11:03:29 [Cray] MCE on DIMM 9
Apr 26, 2025 @ 06:02:59.767557185   console.hostname: x1000c0s7b1n1 2025-04-26 11:02:58 [Cray] MCE on DIMM 9
Apr 26, 2025 @ 06:02:59.326477862   console.hostname: x1000c0s7b1n1 2025-04-26 11:02:58 [Cray] MCE on DIMM 9
Apr 26, 2025 @ 06:01:59.755027596   console.hostname: x1000c0s7b1n1 2025-04-26 11:01:58 [Cray] MCE on DIMM 9
Apr 26, 2025 @ 06:01:59.315281319   console.hostname: x1000c0s7b1n1 2025-04-26 11:01:58 [Cray] MCE on DIMM 9
Apr 26, 2025 @ 06:00:59.297711006   console.hostname: x1000c0s7b1n1 2025-04-26 11:00:58 [Cray] MCE on DIMM 9
```

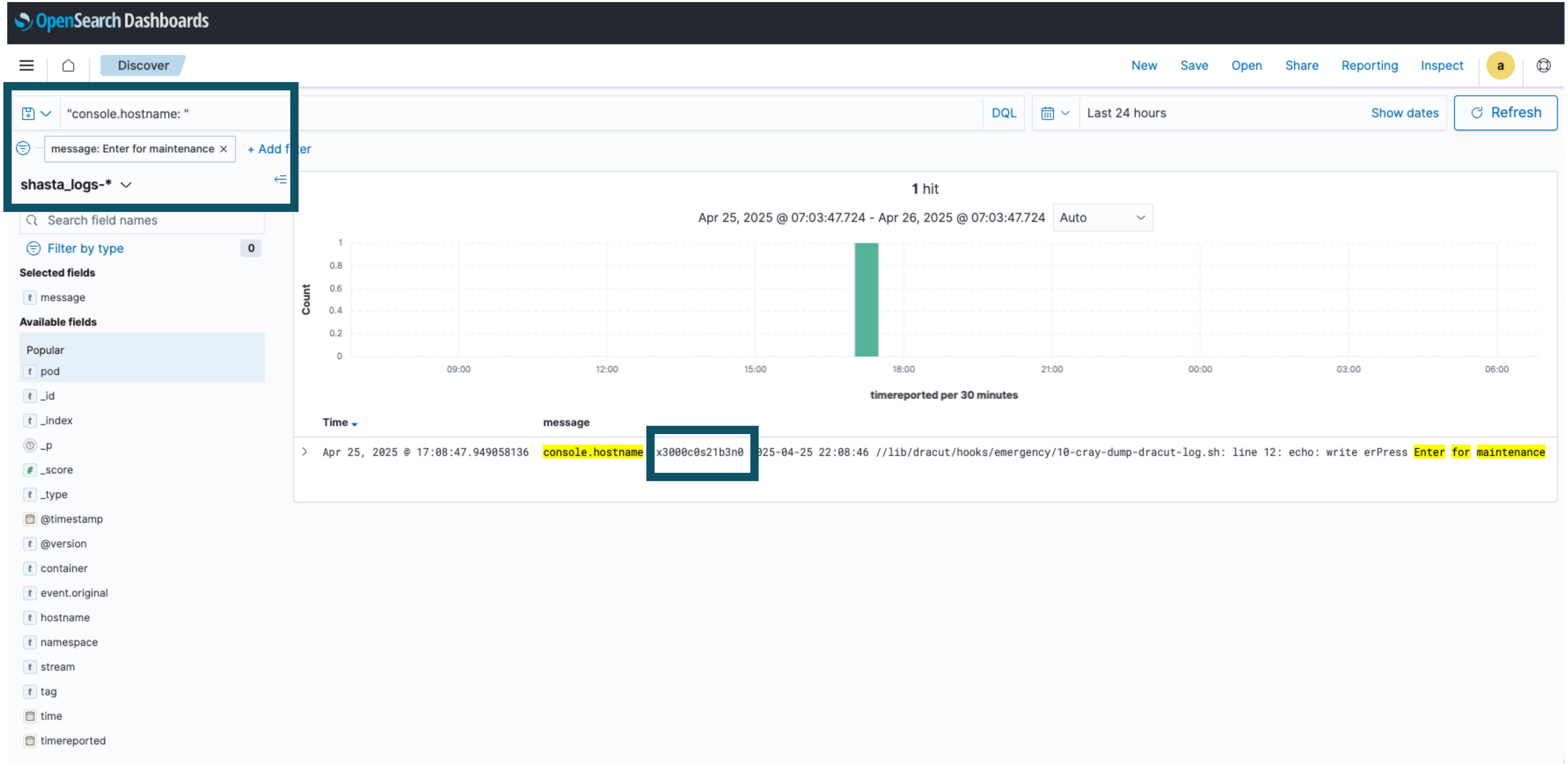
https://sma-dashboards.cmn.SYSTEM_DOMAIN_NAME

NO-CARRIER on HSN NIC

- Console log shows NO-CARRIER for an interface
  - tmp0 is renamed to hsn0
  - tmp1 is renamed to hsn1
  - tmp2 is renamed to hsn2
  - tmp3 is renamed to hsn3
- How often has this been happening on this node?
- Are there other nodes with a similar issue?

```
2025-04-25 22:06:43 [  164.072449] dracut-initqueue[5338]: INFO: Start cray-scripts-driver
2025-04-25 22:06:43 [  164.092085] dracut-initqueue[5338]: INFO: Start /lib/dracut/hooks/cray/links/15-run_cray_network_cfg_lldp.sh
2025-04-25 22:06:43 [  164.112445] dracut-initqueue[5339]: INFO: Start run_cray_network_cfg_lldp
2025-04-25 22:06:48 [  169.037306] dracut-initqueue[5339]: WARNING: Interface tmp0 is not UP (flags <NO-CARRIER,BROADCAST,MULTICAST,UP>)
2025-04-25 22:07:06 [  186.186896] dracut-pre-mount[5514]: iscsiadm: cannot make connection to 10.253.0.18: Network is unreachable
2025-04-25 22:07:06 [  187.187179] dracut-pre-mount[5514]: iscsiadm: cannot make connection to 10.253.0.18: Network is unreachable
2025-04-25 22:07:07 [  188.187485] dracut-pre-mount[5514]: iscsiadm: cannot make connection to 10.253.0.18: Network is unreachable
2025-04-25 22:07:15 [  196.308607] dracut-pre-mount[5514]: iscsiadm: cannot make connection to 10.253.0.18: Network is unreachable


<ConMan> Console [x3000c0s21b3n0] log opened at 2025-04-25 22:07:34 UTC.

<ConMan> Console [x3000c0s21b3n0] connected to <x3000c0s21b3>.
2025-04-25 22:07:43 [  223.732749] dracut-pre-mount[5524]: iscsiadm: cannot make connection to 10.253.0.23: Network is unreachable
2025-04-25 22:07:43 [  223.752095] dracut-pre-mount[5524]: iscsiadm: connection login retries (reopen_max) 5 exceeded
2025-04-25 22:07:43 [  223.772111] dracut-pre-mount[5524]: iscsiadm: Could not perform SendTargets discovery: iSCSI PDU timed out
2025-04-25 22:07:43 [  223.792689] dracut-pre-mount[5532]: iscsiadm: No active sessions.
2025-04-25 22:07:43 [  223.808557] dracut-pre-mount[5533]: iscsiadm: cannot make connection to 10.253.0.16: Network is unreachable
2025-04-25 22:07:44 [  224.770545] dracut-pre-mount[5533]: iscsiadm: cannot make connection to 10.253.0.16: Network is unreachable
2025-04-25 22:07:45 [  225.771245] dracut-pre-mount[5533]: iscsiadm: cannot make connection to 10.253.0.16: Network is unreachable
2025-04-25 22:07:46 [  226.771734] dracut-pre-mount[5533]: iscsiadm: cannot make connection to 10.253.0.16: Network is unreachable
2025-04-25 22:07:54 [  234.900884] dracut-pre-mount[5533]: iscsiadm: cannot make connection to 10.253.0.16: Network is unreachable
2025-04-25 22:08:02 [  243.052253] dracut-pre-mount[5533]: iscsiadm: connection login retries (reopen_max) 5 exceeded
2025-04-25 22:08:02 [  243.072244] dracut-pre-mount[5533]: iscsiadm: Could not perform SendTargets discovery: iSCSI PDU timed out
2025-04-25 22:08:02 [  243.092361] dracut-pre-mount[5543]: iscsiadm: No active sessions.
2025-04-25 22:08:09 [  250.196654] dracut-pre-mount[5544]: iscsiadm: cannot make connection to 10.253.0.8: Network is unreachable
2025-04-25 22:08:18 [  258.324895] dracut-pre-mount[5544]: iscsiadm: cannot make connection to 10.253.0.8: Network is unreachable
2025-04-25 22:08:26 [  266.452844] dracut-pre-mount[5544]: iscsiadm: cannot make connection to 10.253.0.8: Network is unreachable
2025-04-25 22:08:34 [  274.580782] dracut-pre-mount[5544]: iscsiadm: cannot make connection to 10.253.0.8: Network is unreachable
2025-04-25 22:08:42 [  282.708652] dracut-pre-mount[5544]: iscsiadm: cannot make connection to 10.253.0.8: Network is unreachable
2025-04-25 22:08:46 [  286.772550] dracut-pre-mount[5544]: iscsiadm: cannot make connection to 10.253.0.8: Network is unreachable
2025-04-25 22:08:46 [  286.796137] dracut-pre-mount[5544]: iscsiadm: connection login retries (reopen_max) 5 exceeded
2025-04-25 22:08:46 [  286.816084] dracut-pre-mount[5544]: iscsiadm: Could not perform SendTargets discovery: iSCSI PDU timed out
2025-04-25 22:08:46 [  286.836287] dracut-pre-mount[5500]: Warning: sbps-init.sh failed.
2025-04-25 22:08:46 [  286.836931] dracut-pre-mount[5495]: Warning: Unable to prepare squashfs file /tmp/cps/rootfs, dropping to debug.
2025-04-25 22:08:46 //lib/dracut/hooks/emergency/10-cray-dump-dracut-log.sh: line 12: echo: write erPress Enter for maintenance
2025-04-25 22:08:46 (or press Control-D to continue):
```

Enter for maintenance – All Nodes



Query: "console.hostname: "

Filter: message: Enter for maintenance ×

shasta_logs-*

Selected fields
- message

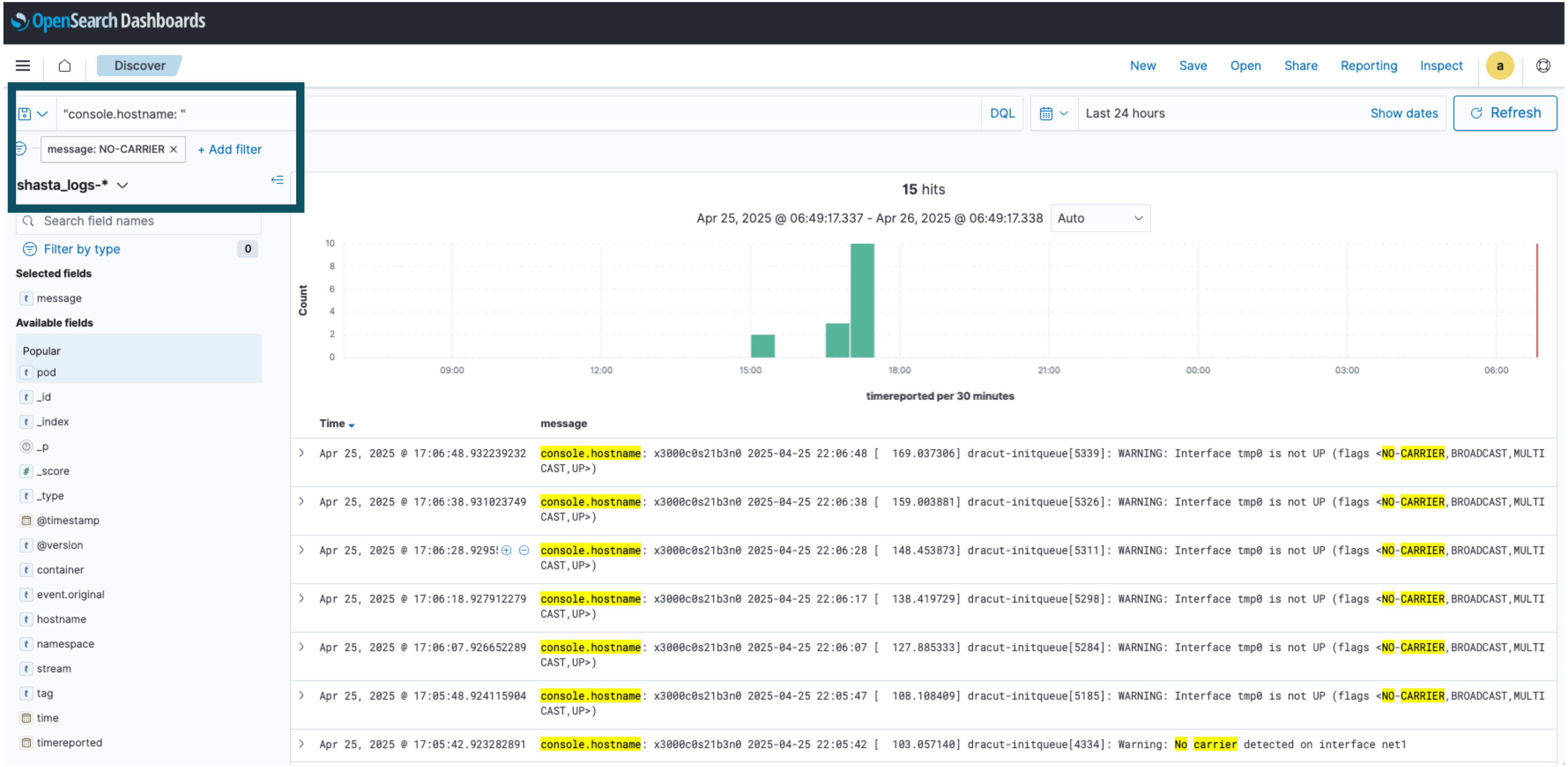
Available fields

Popular
- pod
- _id
- _index
- _p
- _score
- _type
- @timestamp
- @version
- container
- event.original
- hostname
- namespace
- stream
- tag
- time
- timereported

Apr 25, 2025 @ 07:03:47.724 - Apr 26, 2025 @ 07:03:47.724

Time	message
Apr 25, 2025 @ 17:08:47.949058136	console.hostname x3000c0s21b3n0 025-04-25 22:08:46 //lib/dracut/hooks/emergency/10-cray-dump-dracut-log.sh: line 12: echo: write erPress Enter for maintenance

NO-CARRIER on HSN NIC – All Nodes



"console.hostname: "

message: NO-CARRIER ×

shasta_logs-*

Selected fields

- message

Available fields

Popular

- pod
- _id
- _index
- _p
- _score
- _type
- @timestamp
- @version
- container
- event.original
- hostname
- namespace
- stream
- tag
- time
- timereported

15 hits

Apr 25, 2025 @ 06:49:17.337 - Apr 26, 2025 @ 06:49:17.338

```
Time                               message
Apr 25, 2025 @ 17:06:48.932239232  console.hostname: x3000c0s21b3n0 2025-04-25 22:06:48 [  169.037306] dracut-initqueue[5339]: WARNING: Interface tmp0 is not UP (flags <NO-CARRIER,BROADCAST,MULTI
                                   CAST,UP>)
Apr 25, 2025 @ 17:06:38.931023749  console.hostname: x3000c0s21b3n0 2025-04-25 22:06:38 [  159.003881] dracut-initqueue[5326]: WARNING: Interface tmp0 is not UP (flags <NO-CARRIER,BROADCAST,MULTI
                                   CAST,UP>)
Apr 25, 2025 @ 17:06:28.9295!      console.hostname: x3000c0s21b3n0 2025-04-25 22:06:28 [  148.453873] dracut-initqueue[5311]: WARNING: Interface tmp0 is not UP (flags <NO-CARRIER,BROADCAST,MULTI
                                   CAST,UP>)
Apr 25, 2025 @ 17:06:18.927912279  console.hostname: x3000c0s21b3n0 2025-04-25 22:06:17 [  138.419729] dracut-initqueue[5298]: WARNING: Interface tmp0 is not UP (flags <NO-CARRIER,BROADCAST,MULTI
                                   CAST,UP>)
Apr 25, 2025 @ 17:06:07.926652289  console.hostname: x3000c0s21b3n0 2025-04-25 22:06:07 [  127.885333] dracut-initqueue[5284]: WARNING: Interface tmp0 is not UP (flags <NO-CARRIER,BROADCAST,MULTI
                                   CAST,UP>)
Apr 25, 2025 @ 17:05:48.924115904  console.hostname: x3000c0s21b3n0 2025-04-25 22:05:47 [  108.108409] dracut-initqueue[5185]: WARNING: Interface tmp0 is not UP (flags <NO-CARRIER,BROADCAST,MULTI
                                   CAST,UP>)
Apr 25, 2025 @ 17:05:42.923282891  console.hostname: x3000c0s21b3n0 2025-04-25 22:05:42 [  103.057140] dracut-initqueue[4334]: Warning: No carrier detected on interface net1
```

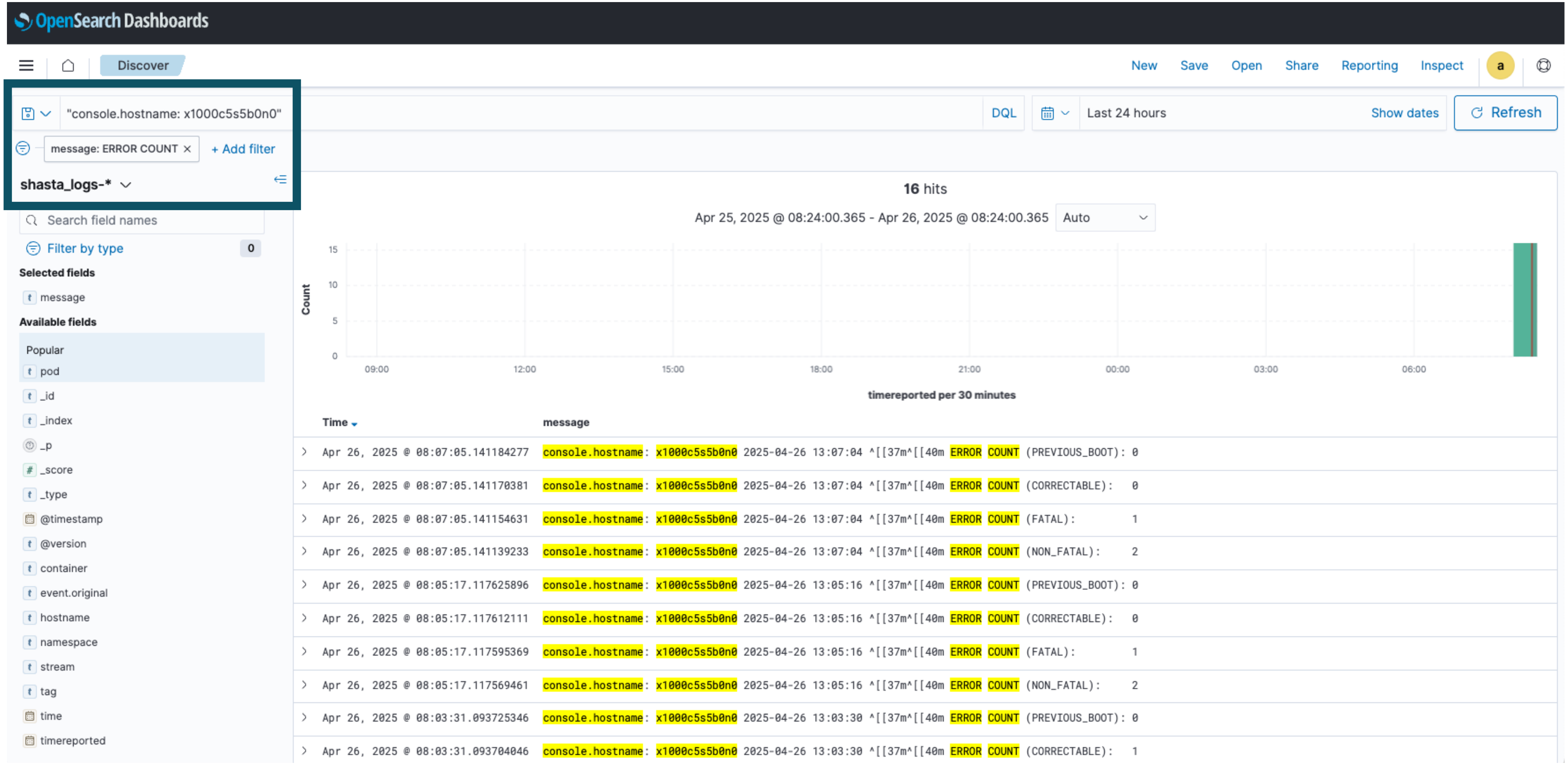
Node in boot loop

- Node in boot loop showing BIOS messages, but never starts PXE boot
- How often has this been happening on this node?
- Are there other nodes with a similar issue?

```
2025-04-26T13:03:24.41199 ERROR ENTRY - FATAL
2025-04-26T13:03:24.41200 Type:   8
2025-04-26T13:03:24.41201 Flags:  0x02
2025-04-26T13:03:24.41202 Port80: 0xB000A992
2025-04-26T13:03:24.41203 ApicId: 0x00000000
2025-04-26T13:03:24.41204 CpuNum: 0
2025-04-26T13:03:24.41210 RTC:    04/26/2025 13:03:23-00:00
2025-04-26T13:03:24.41211 ErrorData: Boot Failure No valid Boot devices found.
2025-04-26T13:03:24.41212 Description: No valid Boot devices found.
2025-04-26T13:03:24.41213 ERROR ENTRY - END
```

```
2025-04-26T13:03:34.11704 CrayContinueOnErrorCheck: Entry
2025-04-26T13:03:34.11705  ERROR COUNT (NON_FATAL):     2
2025-04-26T13:03:34.11706  ERROR COUNT (FATAL):         1
2025-04-26T13:03:34.11707  ERROR COUNT (CORRECTABLE):   1
2025-04-26T13:03:34.11708  ERROR COUNT (PREVIOUS_BOOT): 0
2025-04-26T13:03:34.11709 ContinueOnError DISABLED
2025-04-26T13:03:34.11710   Error(s) detected
2025-04-26T13:03:34.11711   Force Shell Boot
```

ERROR COUNT – x1000c5s5b0n0



"console.hostname: x1000c5s5b0n0"

message: ERROR COUNT ×

shasta_logs-*

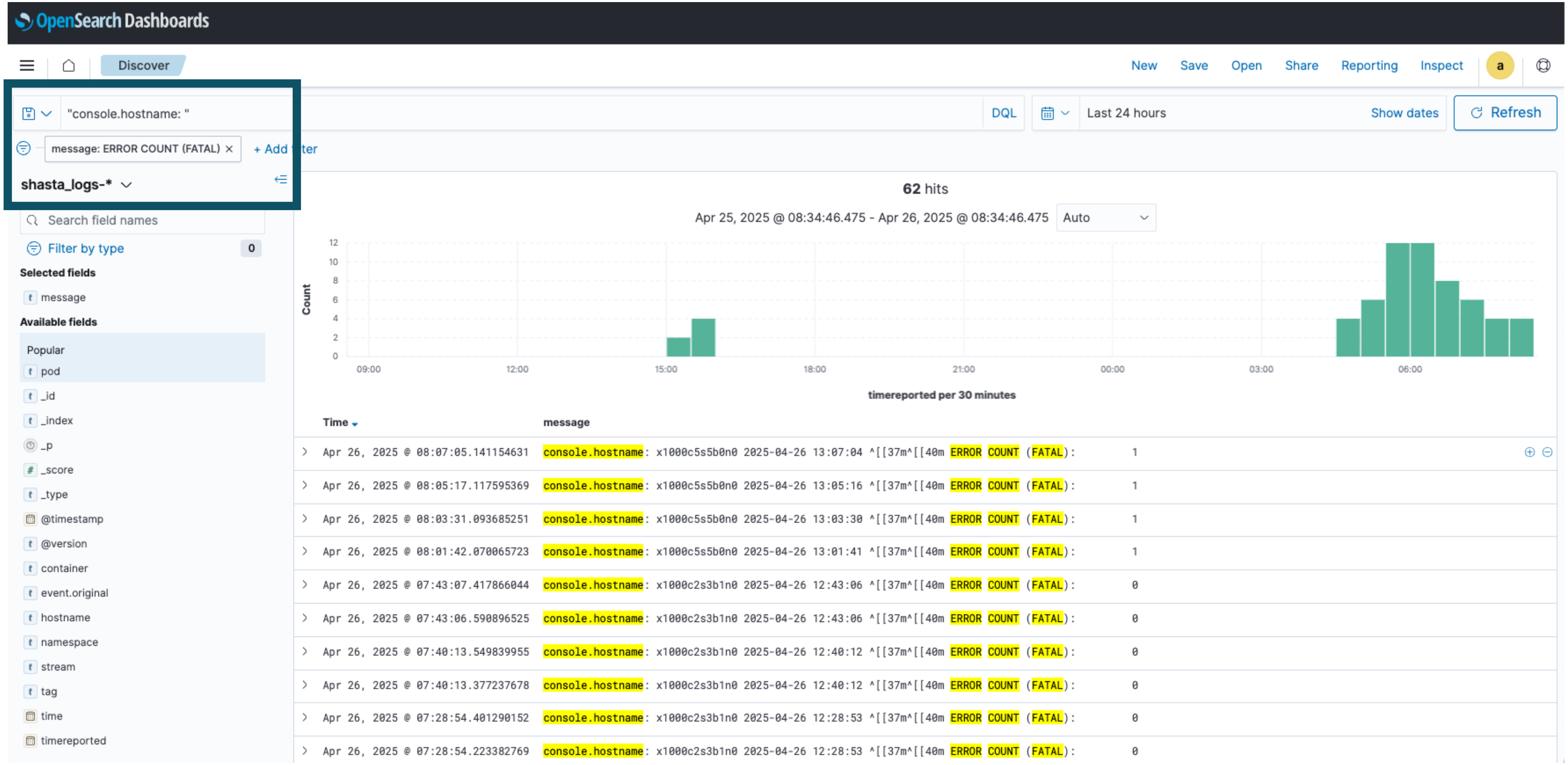
Last 24 hours

16 hits

Apr 25, 2025 @ 08:24:00.365 - Apr 26, 2025 @ 08:24:00.365 Auto

```
Time                                 message
Apr 26, 2025 @ 08:07:05.141184277    console.hostname: x1000c5s5b0n0 2025-04-26 13:07:04 ^[[37m^[[40m ERROR COUNT (PREVIOUS_BOOT): 0
Apr 26, 2025 @ 08:07:05.141170381    console.hostname: x1000c5s5b0n0 2025-04-26 13:07:04 ^[[37m^[[40m ERROR COUNT (CORRECTABLE):   0
Apr 26, 2025 @ 08:07:05.141154631    console.hostname: x1000c5s5b0n0 2025-04-26 13:07:04 ^[[37m^[[40m ERROR COUNT (FATAL):         1
Apr 26, 2025 @ 08:07:05.141139233    console.hostname: x1000c5s5b0n0 2025-04-26 13:07:04 ^[[37m^[[40m ERROR COUNT (NON_FATAL):     2
Apr 26, 2025 @ 08:05:17.117625896    console.hostname: x1000c5s5b0n0 2025-04-26 13:05:16 ^[[37m^[[40m ERROR COUNT (PREVIOUS_BOOT): 0
Apr 26, 2025 @ 08:05:17.117612111    console.hostname: x1000c5s5b0n0 2025-04-26 13:05:16 ^[[37m^[[40m ERROR COUNT (CORRECTABLE):   0
Apr 26, 2025 @ 08:05:17.117595369    console.hostname: x1000c5s5b0n0 2025-04-26 13:05:16 ^[[37m^[[40m ERROR COUNT (FATAL):         1
Apr 26, 2025 @ 08:05:17.117569461    console.hostname: x1000c5s5b0n0 2025-04-26 13:05:16 ^[[37m^[[40m ERROR COUNT (NON_FATAL):     2
Apr 26, 2025 @ 08:03:31.093725346    console.hostname: x1000c5s5b0n0 2025-04-26 13:03:30 ^[[37m^[[40m ERROR COUNT (PREVIOUS_BOOT): 0
Apr 26, 2025 @ 08:03:31.093704046    console.hostname: x1000c5s5b0n0 2025-04-26 13:03:30 ^[[37m^[[40m ERROR COUNT (CORRECTABLE):   1
```

ERROR COUNT (FATAL) – All nodes



"console.hostname: "

message: ERROR COUNT (FATAL)

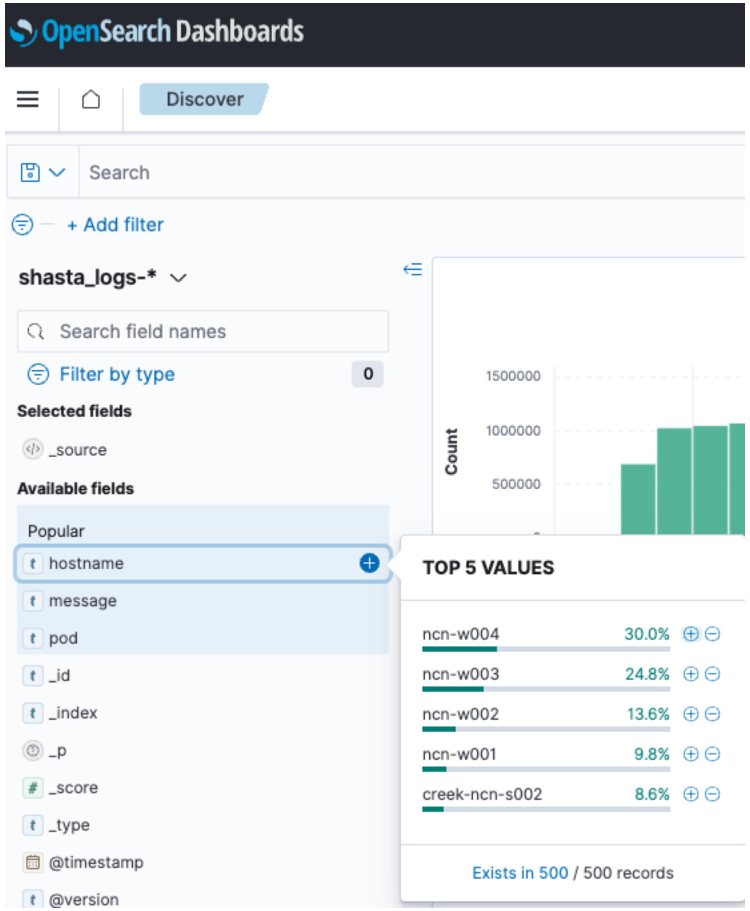
shasta_logs-*

62 hits

Apr 25, 2025 @ 08:34:46.475 - Apr 26, 2025 @ 08:34:46.475

Time	message
Apr 26, 2025 @ 08:07:05.141154631	console.hostname: x1000c5s5b0n0 2025-04-26 13:07:04 ^[[37m^[[40m ERROR COUNT (FATAL): 1
Apr 26, 2025 @ 08:05:17.117595369	console.hostname: x1000c5s5b0n0 2025-04-26 13:05:16 ^[[37m^[[40m ERROR COUNT (FATAL): 1
Apr 26, 2025 @ 08:03:31.093685251	console.hostname: x1000c5s5b0n0 2025-04-26 13:03:30 ^[[37m^[[40m ERROR COUNT (FATAL): 1
Apr 26, 2025 @ 08:01:42.070065723	console.hostname: x1000c5s5b0n0 2025-04-26 13:01:41 ^[[37m^[[40m ERROR COUNT (FATAL): 1
Apr 26, 2025 @ 07:43:07.417866044	console.hostname: x1000c2s3b1n0 2025-04-26 12:43:06 ^[[37m^[[40m ERROR COUNT (FATAL): 0
Apr 26, 2025 @ 07:43:06.590896525	console.hostname: x1000c2s3b1n0 2025-04-26 12:43:06 ^[[37m^[[40m ERROR COUNT (FATAL): 0
Apr 26, 2025 @ 07:40:13.549839955	console.hostname: x1000c2s3b1n0 2025-04-26 12:40:12 ^[[37m^[[40m ERROR COUNT (FATAL): 0
Apr 26, 2025 @ 07:40:13.377237678	console.hostname: x1000c2s3b1n0 2025-04-26 12:40:12 ^[[37m^[[40m ERROR COUNT (FATAL): 0
Apr 26, 2025 @ 07:28:54.401290152	console.hostname: x1000c2s3b1n0 2025-04-26 12:28:53 ^[[37m^[[40m ERROR COUNT (FATAL): 0
Apr 26, 2025 @ 07:28:54.223382769	console.hostname: x1000c2s3b1n0 2025-04-26 12:28:53 ^[[37m^[[40m ERROR COUNT (FATAL): 0

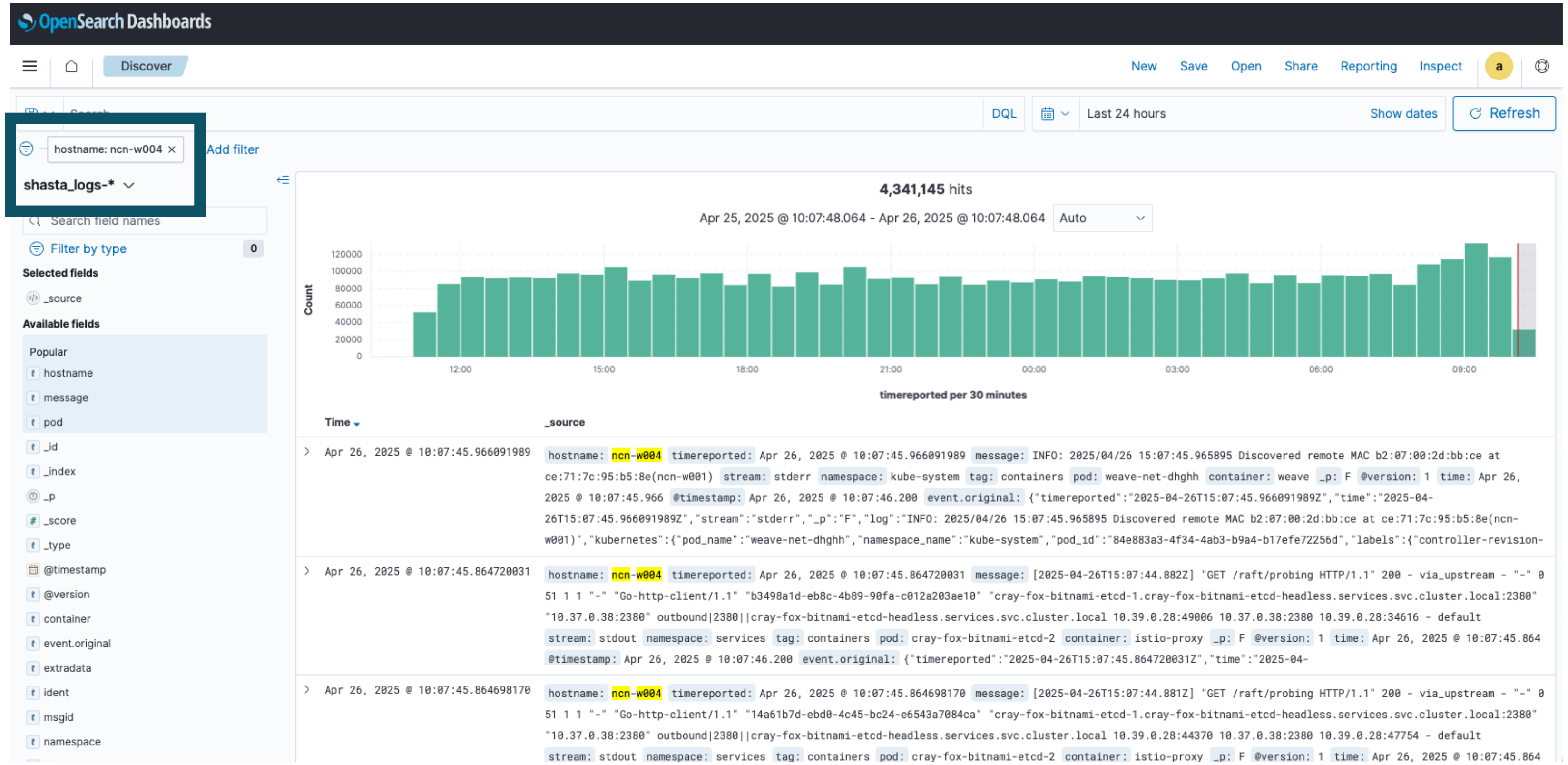
Searching Logs – Hostname ncn-w004



TOP 5 VALUES

ncn-w004	30.0%
ncn-w003	24.8%
ncn-w002	13.6%
ncn-w001	9.8%
creek-ncn-s002	8.6%

Exists in 500 / 500 records

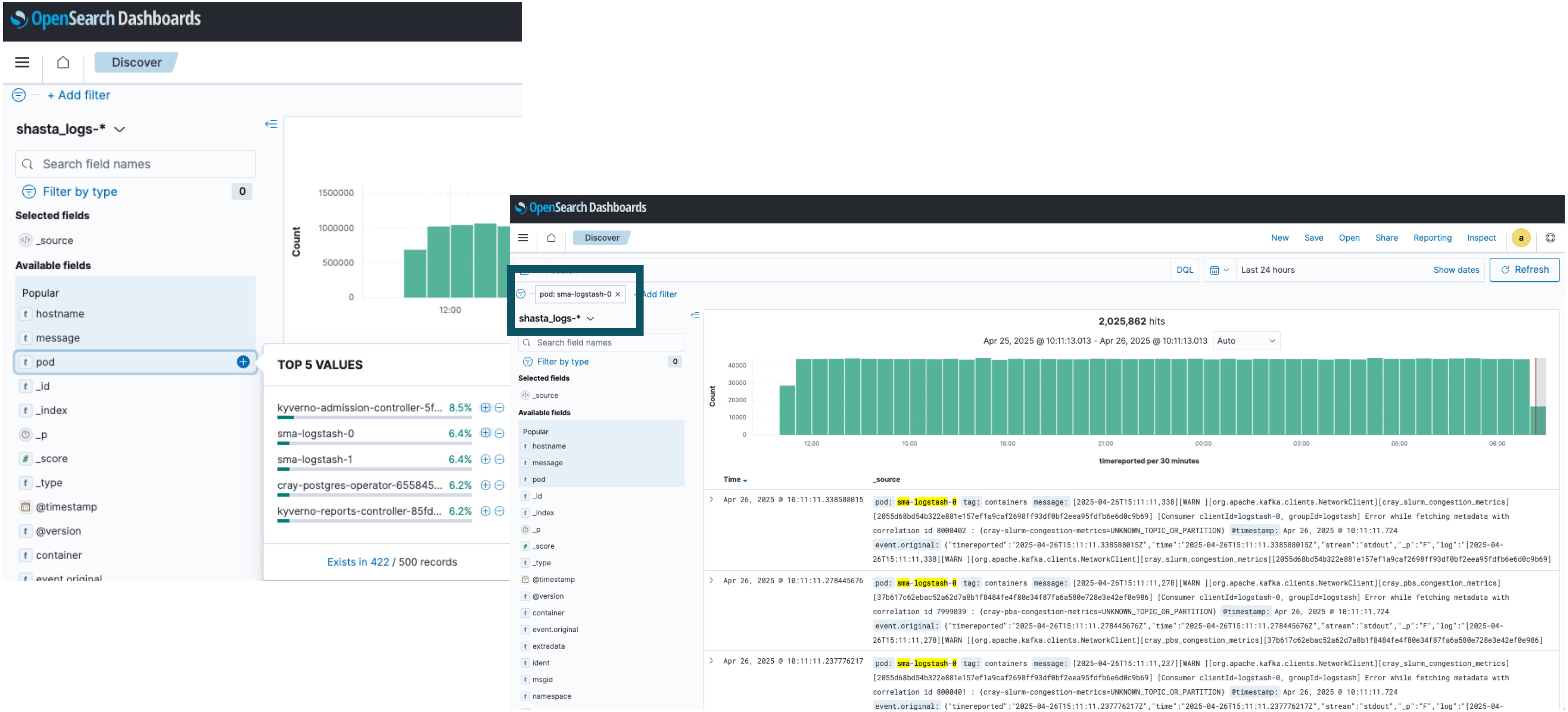


4,341,145 hits

Apr 25, 2025 @ 10:07:48.064 - Apr 26, 2025 @ 10:07:48.064 Auto

```
Time                              _source
> Apr 26, 2025 @ 10:07:45.966091989  hostname: ncn-w004 timereported: Apr 26, 2025 @ 10:07:45.966091989 message: INFO: 2025/04/26 15:07:45.965895 Discovered remote MAC b2:07:00:2d:bb:ce at
                                   ce:71:7c:95:b5:8e(ncn-w001) stream: stderr namespace: kube-system tag: containers pod: weave-net-dhghh container: weave _p: F @version: 1 time: Apr 26,
                                   2025 @ 10:07:45.966 @timestamp: Apr 26, 2025 @ 10:07:46.200 event.original: {"timereported":"2025-04-26T15:07:45.966091989Z","time":"2025-04-
                                   26T15:07:45.966091989Z","stream":"stderr","_p":"F","log":"INFO: 2025/04/26 15:07:45.965895 Discovered remote MAC b2:07:00:2d:bb:ce at ce:71:7c:95:b5:8e(ncn-
                                   w001)","kubernetes":{"pod_name":"weave-net-dhghh","namespace_name":"kube-system","pod_id":"84e883a3-4f34-4ab3-b9a4-b17efe72256d","labels":{"controller-revision-
> Apr 26, 2025 @ 10:07:45.864720031  hostname: ncn-w004 timereported: Apr 26, 2025 @ 10:07:45.864720031 message: [2025-04-26T15:07:44.882Z] "GET /raft/probing HTTP/1.1" 200 - via_upstream - "-" 0
                                   51 1 1 "-" "Go-http-client/1.1" "b3498a1d-eb8c-4b89-90fa-c012a203ae10" "cray-fox-bitnami-etcd-1.cray-fox-bitnami-etcd-headless.services.svc.cluster.local:2380"
                                   "10.37.0.38:2380" outbound|2380||cray-fox-bitnami-etcd-headless.services.svc.cluster.local 10.39.0.28:49006 10.37.0.38:2380 10.39.0.28:34616 - default
                                   stream: stdout namespace: services tag: containers pod: cray-fox-bitnami-etcd-2 container: istio-proxy _p: F @version: 1 time: Apr 26, 2025 @ 10:07:45.864
                                   @timestamp: Apr 26, 2025 @ 10:07:46.200 event.original: {"timereported":"2025-04-26T15:07:45.864720031Z","time":"2025-04-
> Apr 26, 2025 @ 10:07:45.864698170  hostname: ncn-w004 timereported: Apr 26, 2025 @ 10:07:45.864698170 message: [2025-04-26T15:07:44.881Z] "GET /raft/probing HTTP/1.1" 200 - via_upstream - "-" 0
                                   51 1 1 "-" "Go-http-client/1.1" "14a61b7d-ebd0-4c45-bc24-e6543a7084ca" "cray-fox-bitnami-etcd-1.cray-fox-bitnami-etcd-headless.services.svc.cluster.local:2380"
                                   "10.37.0.38:2380" outbound|2380||cray-fox-bitnami-etcd-headless.services.svc.cluster.local 10.39.0.28:44370 10.37.0.38:2380 10.39.0.28:47754 - default
                                   stream: stdout namespace: services tag: containers pod: cray-fox-bitnami-etcd-2 container: istio-proxy _p: F @version: 1 time: Apr 26, 2025 @ 10:07:45.864
```

Searching Logs – Pod sma-log-stash-0



TOP 5 VALUES

kyverno-admission-controller-5f...	8.5%
sma-logstash-0	6.4%
sma-logstash-1	6.4%
cray-postgres-operator-655845...	6.2%
kyverno-reports-controller-85fd...	6.2%

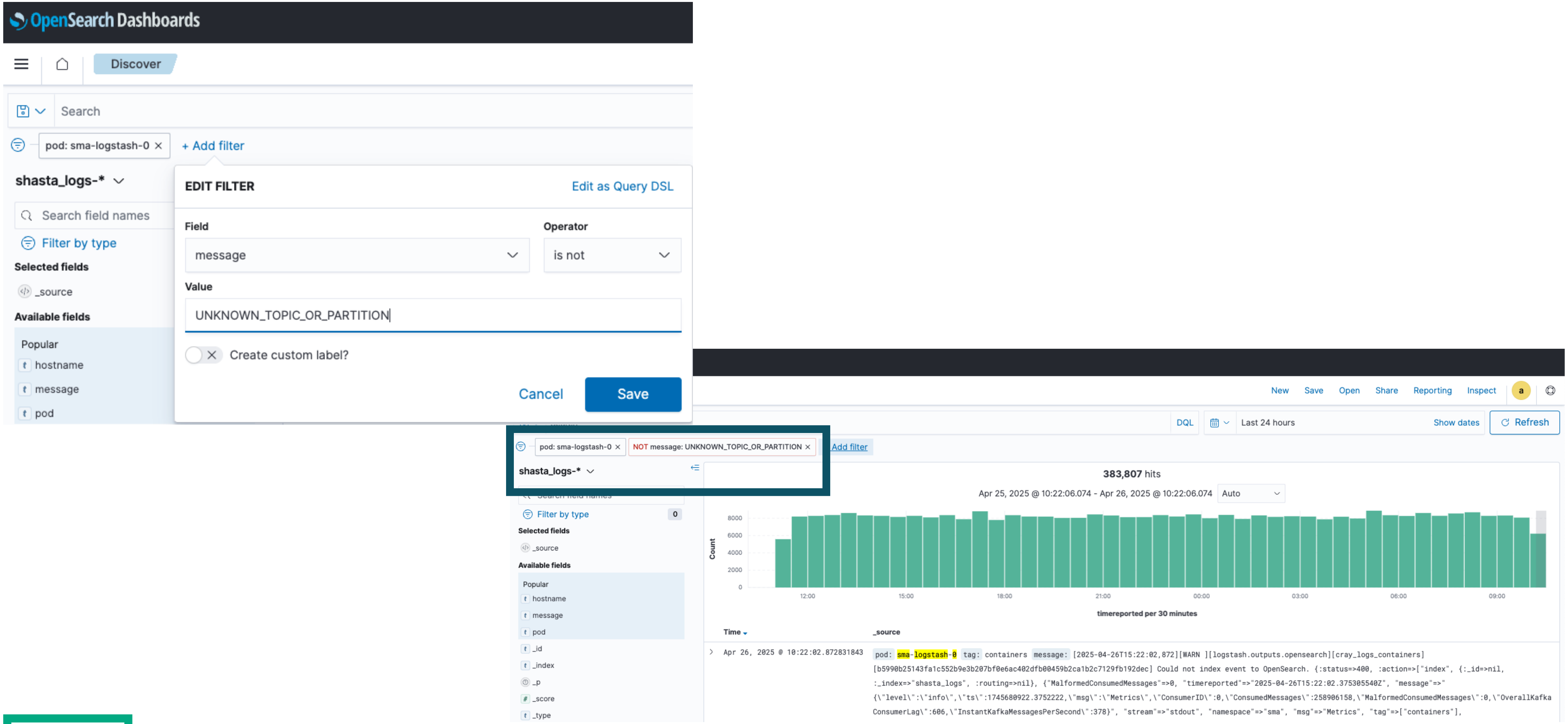
Exists in 422 / 500 records

2,025,862 hits

Apr 25, 2025 @ 10:11:13.013 - Apr 26, 2025 @ 10:11:13.013 Auto

```
Time                              _source
> Apr 26, 2025 @ 10:11:11.338588015  pod: sma-logstash-0 tag: containers message: [2025-04-26T15:11:11,338][WARN ][org.apache.kafka.clients.NetworkClient][cray_slurm_congestion_metrics]
                                  [2055d68bd54b322e881e157ef1a9caf2698ff93df0bf2eea95fdfb6e6d0c9b69] [Consumer clientId=logstash-0, groupId=logstash] Error while fetching metadata with
                                  correlation id 8000402 : {cray-slurm-congestion-metrics=UNKNOWN_TOPIC_OR_PARTITION} @timestamp: Apr 26, 2025 @ 10:11:11.724
                                  event.original: {"timereported":"2025-04-26T15:11:11.338588015Z","time":"2025-04-26T15:11:11.338588015Z","stream":"stdout","_p":"F","log":"[2025-04-
                                  26T15:11:11,338][WARN ][org.apache.kafka.clients.NetworkClient][cray_slurm_congestion_metrics][2055d68bd54b322e881e157ef1a9caf2698ff93df0bf2eea95fdfb6e6d0c9b69]
> Apr 26, 2025 @ 10:11:11.278445676  pod: sma-logstash-0 tag: containers message: [2025-04-26T15:11:11,278][WARN ][org.apache.kafka.clients.NetworkClient][cray_pbs_congestion_metrics]
                                  [37b617c62ebac52a62d7a8b1f8484fe4f80e34f87fa6a580e728e3e42ef0e986] [Consumer clientId=logstash-0, groupId=logstash] Error while fetching metadata with
                                  correlation id 7999039 : {cray-pbs-congestion-metrics=UNKNOWN_TOPIC_OR_PARTITION} @timestamp: Apr 26, 2025 @ 10:11:11.724
                                  event.original: {"timereported":"2025-04-26T15:11:11.278445676Z","time":"2025-04-26T15:11:11.278445676Z","stream":"stdout","_p":"F","log":"[2025-04-
                                  26T15:11:11,278][WARN ][org.apache.kafka.clients.NetworkClient][cray_pbs_congestion_metrics][37b617c62ebac52a62d7a8b1f8484fe4f80e34f87fa6a580e728e3e42ef0e986]
> Apr 26, 2025 @ 10:11:11.237776217  pod: sma-logstash-0 tag: containers message: [2025-04-26T15:11:11,237][WARN ][org.apache.kafka.clients.NetworkClient][cray_slurm_congestion_metrics]
                                  [2055d68bd54b322e881e157ef1a9caf2698ff93df0bf2eea95fdfb6e6d0c9b69] [Consumer clientId=logstash-0, groupId=logstash] Error while fetching metadata with
                                  correlation id 8000401 : {cray-slurm-congestion-metrics=UNKNOWN_TOPIC_OR_PARTITION} @timestamp: Apr 26, 2025 @ 10:11:11.724
                                  event.original: {"timereported":"2025-04-26T15:11:11.237776217Z","time":"2025-04-26T15:11:11.237776217Z","stream":"stdout","_p":"F","log":"[2025-04-
```

Searching Logs – Pod sma-log-stash-0 - no UNKNOWN_TOPIC_OR_PARTITION



pod: sma-logstash-0 ×

EDIT FILTER — Edit as Query DSL

Field: message

Operator: is not

Value: UNKNOWN_TOPIC_OR_PARTITION

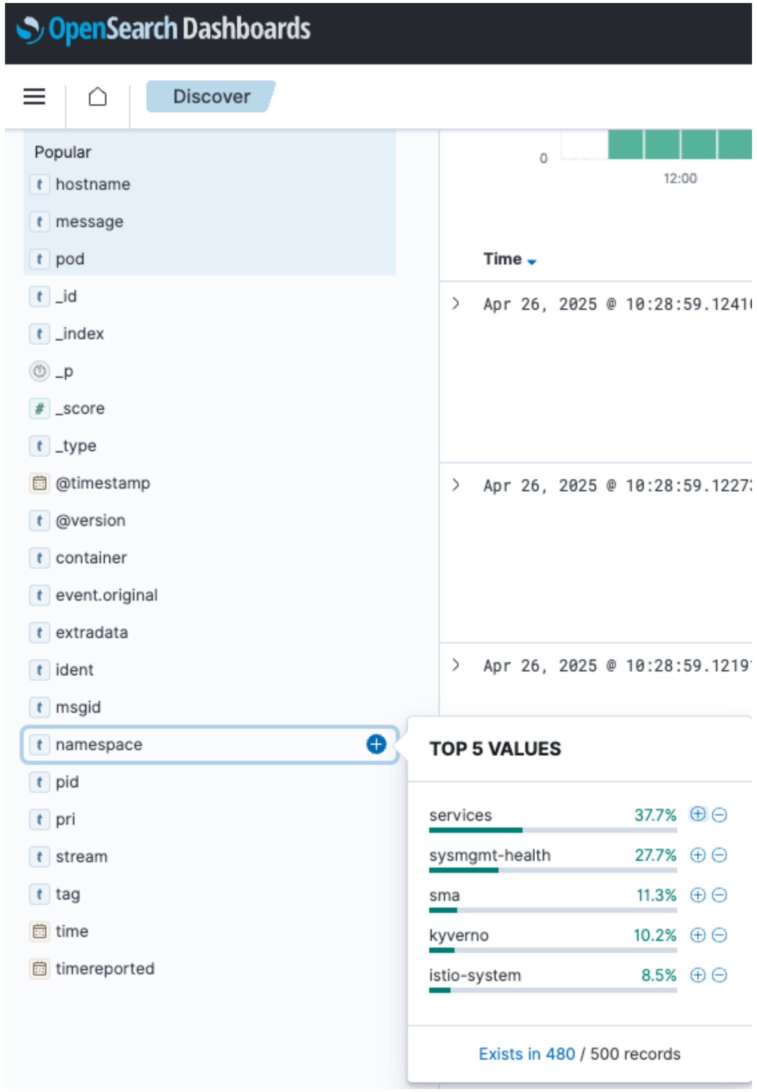
pod: sma-logstash-0 × NOT message: UNKNOWN_TOPIC_OR_PARTITION ×

383,807 hits

Apr 25, 2025 @ 10:22:06.074 - Apr 26, 2025 @ 10:22:06.074

```
Time          _source
Apr 26, 2025 @ 10:22:02.872831843   pod: sma-logstash-0 tag: containers message: [2025-04-26T15:22:02,872][WARN ][logstash.outputs.opensearch][cray_logs_containers]
[b5990b25143fa1c552b9e3b207bf0e6ac402dfb00459b2ca1b2c7129fb192dec] Could not index event to OpenSearch. {:status=>400, :action=>["index", {:_id=>nil,
:_index=>"shasta_logs", :routing=>nil}, {"MalformedConsumedMessages"=>0, "timereported"=>"2025-04-26T15:22:02.375305540Z", "message"=>"
{\"level\":\"info\",\"ts\":1745680922.3752222,\"msg\":\"Metrics\",\"ConsumerID\":0,\"ConsumedMessages\":258906158,\"MalformedConsumedMessages\":0,\"OverallKafka
ConsumerLag\":606,\"InstantKafkaMessagesPerSecond\":378}", "stream"=>"stdout", "namespace"=>"sma", "msg"=>"Metrics", "tag"=>["containers"],
```

Searching Logs - Namespace

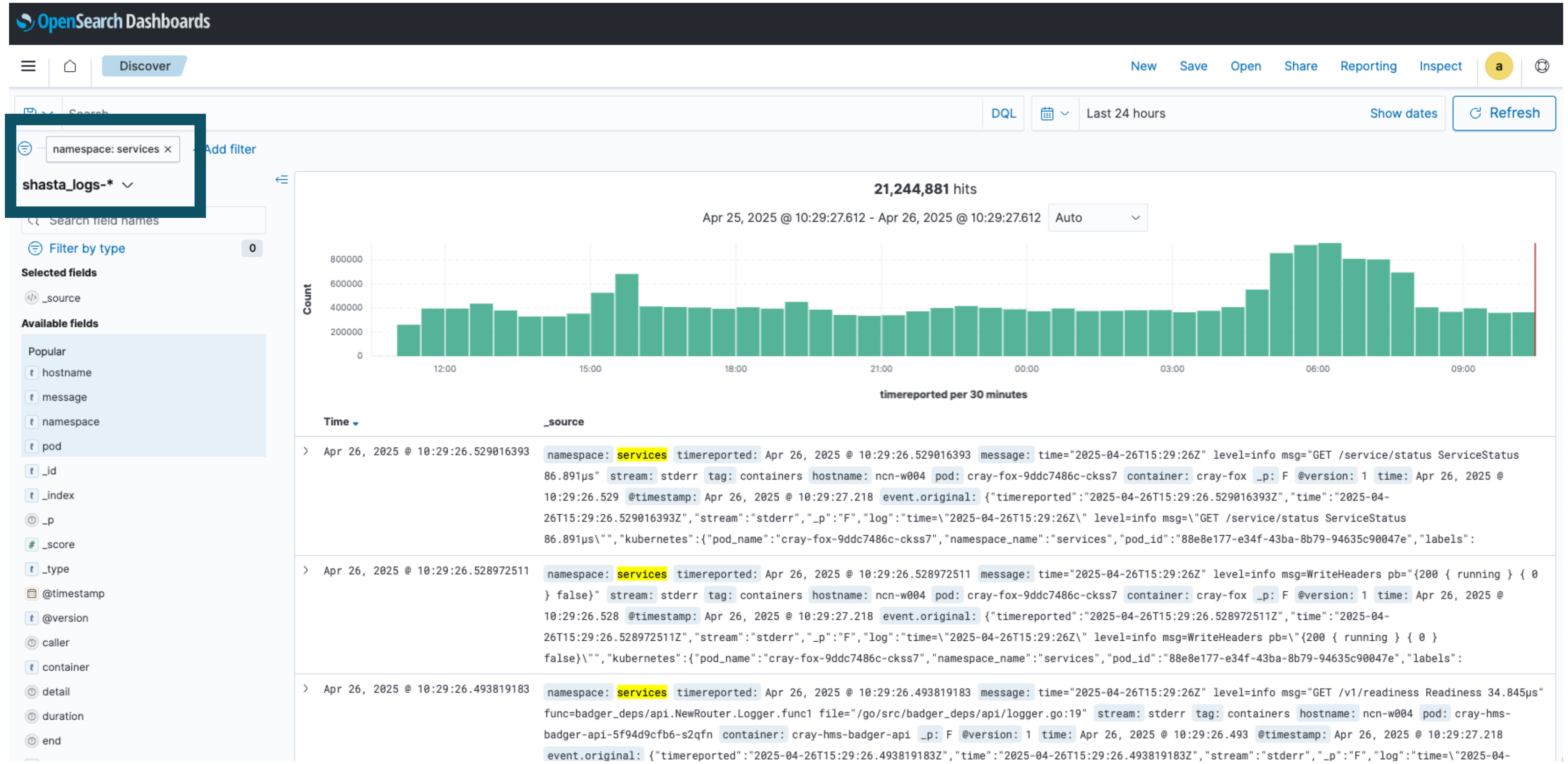


```
Apr 26, 2025 @ 10:28:59.1241
Apr 26, 2025 @ 10:28:59.1227
Apr 26, 2025 @ 10:28:59.1219
```

TOP 5 VALUES

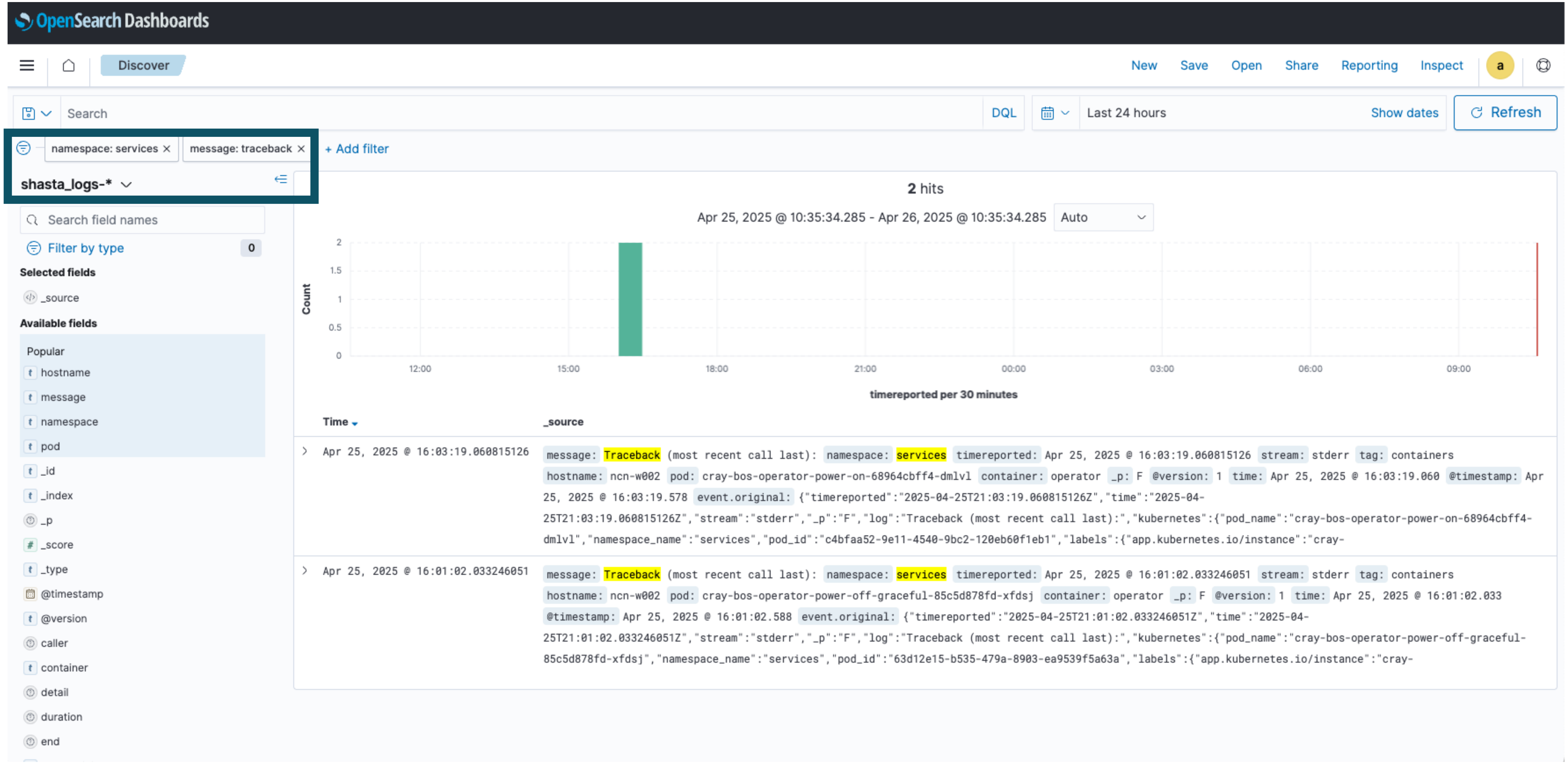
services	37.7%
sysmgmt-health	27.7%
sma	11.3%
kyverno	10.2%
istio-system	8.5%

Exists in 480 / 500 records



```
Apr 26, 2025 @ 10:29:26.529016393   namespace: services timereported: Apr 26, 2025 @ 10:29:26.529016393 message: time="2025-04-26T15:29:26Z" level=info msg="GET /service/status ServiceStatus
                                    86.891µs" stream: stderr tag: containers hostname: ncn-w004 pod: cray-fox-9ddc7486c-ckss7 container: cray-fox _p: F @version: 1 time: Apr 26, 2025 @
                                    10:29:26.529 @timestamp: Apr 26, 2025 @ 10:29:27.218 event.original: {"timereported":"2025-04-26T15:29:26.529016393Z","time":"2025-04-
                                    26T15:29:26.529016393Z","stream":"stderr","_p":"F","log":"time=\"2025-04-26T15:29:26Z\" level=info msg=\"GET /service/status ServiceStatus
                                    86.891µs\"","kubernetes":{"pod_name":"cray-fox-9ddc7486c-ckss7","namespace_name":"services","pod_id":"88e8e177-e34f-43ba-8b79-94635c90047e","labels":

Apr 26, 2025 @ 10:29:26.528972511   namespace: services timereported: Apr 26, 2025 @ 10:29:26.528972511 message: time="2025-04-26T15:29:26Z" level=info msg=WriteHeaders pb="{200 { running } { 0
                                    } false}" stream: stderr tag: containers hostname: ncn-w004 pod: cray-fox-9ddc7486c-ckss7 container: cray-fox _p: F @version: 1 time: Apr 26, 2025 @
                                    10:29:26.528 @timestamp: Apr 26, 2025 @ 10:29:27.218 event.original: {"timereported":"2025-04-26T15:29:26.528972511Z","time":"2025-04-
                                    26T15:29:26.528972511Z","stream":"stderr","_p":"F","log":"time=\"2025-04-26T15:29:26Z\" level=info msg=WriteHeaders pb=\"{200 { running } { 0 }
                                    false}\"","kubernetes":{"pod_name":"cray-fox-9ddc7486c-ckss7","namespace_name":"services","pod_id":"88e8e177-e34f-43ba-8b79-94635c90047e","labels":

Apr 26, 2025 @ 10:29:26.493819183   namespace: services timereported: Apr 26, 2025 @ 10:29:26.493819183 message: time="2025-04-26T15:29:26Z" level=info msg="GET /v1/readiness Readiness 34.845µs"
                                    func=badger_deps/api.NewRouter.Logger.func1 file="/go/src/badger_deps/api/logger.go:19" stream: stderr tag: containers hostname: ncn-w004 pod: cray-hms-
                                    badger-api-5f94d9cfb6-s2qfn container: cray-hms-badger-api _p: F @version: 1 time: Apr 26, 2025 @ 10:29:26.493 @timestamp: Apr 26, 2025 @ 10:29:27.218
                                    event.original: {"timereported":"2025-04-26T15:29:26.493819183Z","time":"2025-04-26T15:29:26.493819183Z","stream":"stderr","_p":"F","log":"time=\"2025-04-
```

Searching Logs – Namespace Services – Message traceback



namespace: services × message: traceback ×

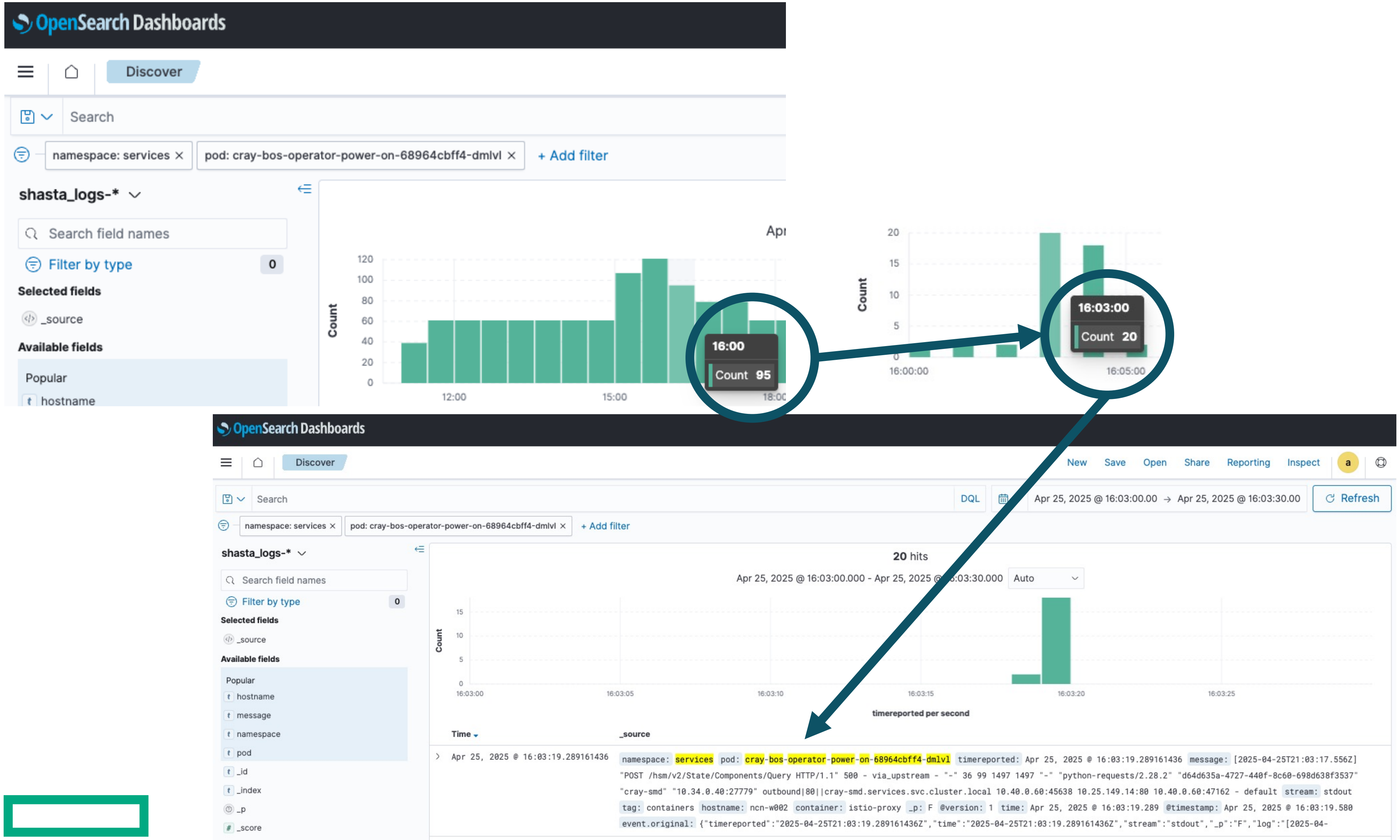
shasta_logs-*

2 hits

Apr 25, 2025 @ 10:35:34.285 - Apr 26, 2025 @ 10:35:34.285

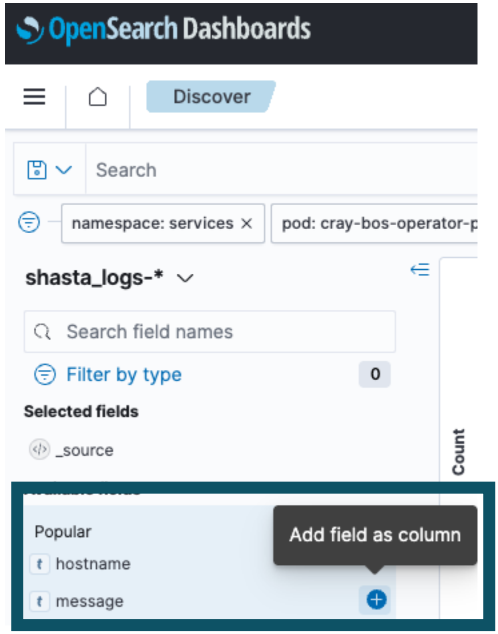
Time	_source
Apr 25, 2025 @ 16:03:19.060815126	message: Traceback (most recent call last): namespace: services timereported: Apr 25, 2025 @ 16:03:19.060815126 stream: stderr tag: containers hostname: ncn-w002 pod: cray-bos-operator-power-on-68964cbff4-dmlvl container: operator _p: F @version: 1 time: Apr 25, 2025 @ 16:03:19.060 @timestamp: Apr 25, 2025 @ 16:03:19.578 event.original: {"timereported":"2025-04-25T21:03:19.060815126Z","time":"2025-04-25T21:03:19.060815126Z","stream":"stderr","_p":"F","log":"Traceback (most recent call last):","kubernetes":{"pod_name":"cray-bos-operator-power-on-68964cbff4-dmlvl","namespace_name":"services","pod_id":"c4bfaa52-9e11-4540-9bc2-120eb60f1eb1","labels":{"app.kubernetes.io/instance":"cray-
Apr 25, 2025 @ 16:01:02.033246051	message: Traceback (most recent call last): namespace: services timereported: Apr 25, 2025 @ 16:01:02.033246051 stream: stderr tag: containers hostname: ncn-w002 pod: cray-bos-operator-power-off-graceful-85c5d878fd-xfdsj container: operator _p: F @version: 1 time: Apr 25, 2025 @ 16:01:02.033 @timestamp: Apr 25, 2025 @ 16:01:02.588 event.original: {"timereported":"2025-04-25T21:01:02.033246051Z","time":"2025-04-25T21:01:02.033246051Z","stream":"stderr","_p":"F","log":"Traceback (most recent call last):","kubernetes":{"pod_name":"cray-bos-operator-power-off-graceful-85c5d878fd-xfdsj","namespace_name":"services","pod_id":"63d12e15-b535-479a-8903-ea9539f5a63a","labels":{"app.kubernetes.io/instance":"cray-

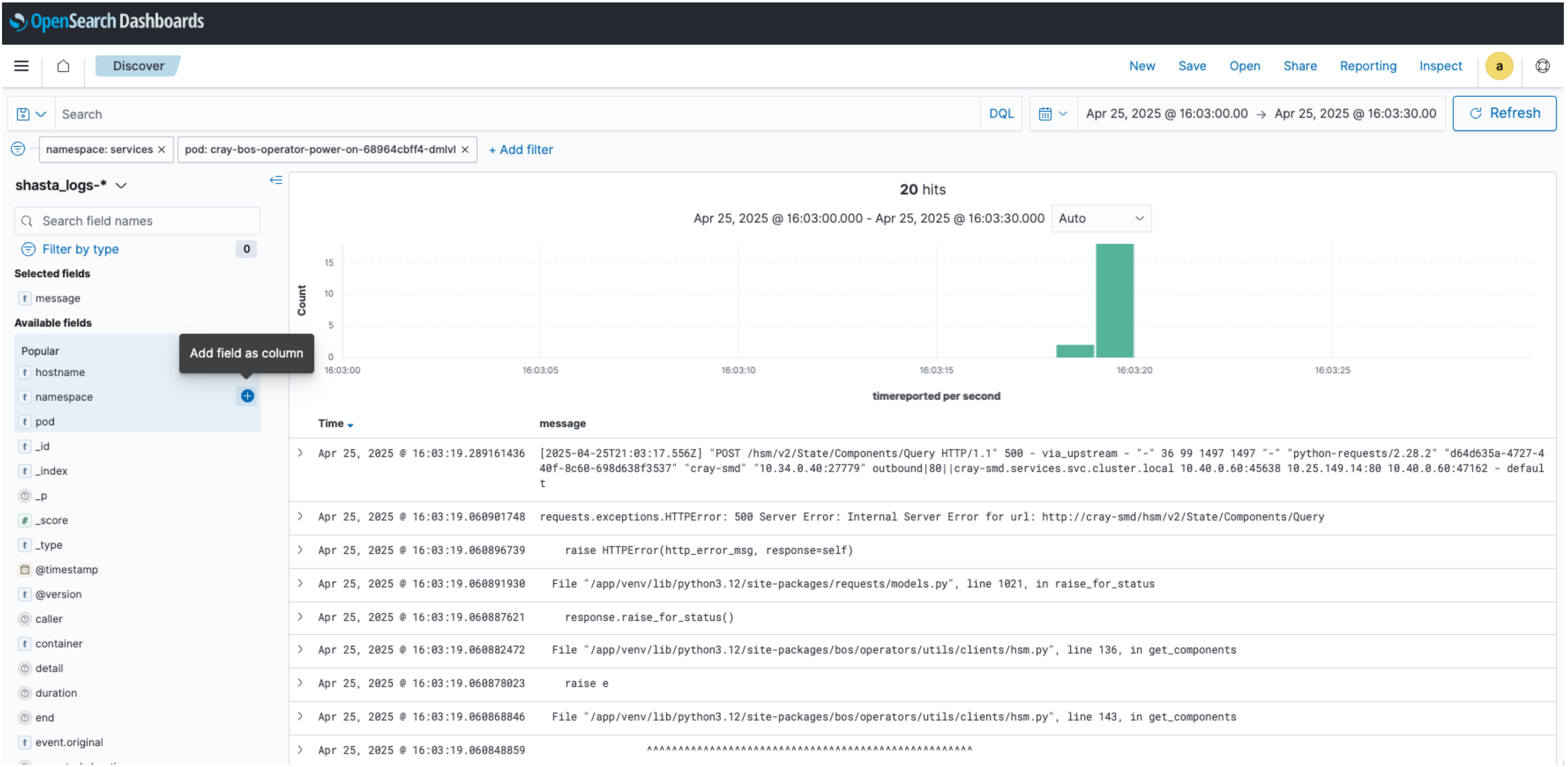
Searching Logs – Namespace Services – Message traceback – Time Zoom



```
Apr 25, 2025 @ 16:03:19.289161436   namespace: services pod: cray-bos-operator-power-on-68964cbff4-dmlvl timereported: Apr 25, 2025 @ 16:03:19.289161436 message: [2025-04-25T21:03:17.556Z]
"POST /hsm/v2/State/Components/Query HTTP/1.1" 500 - via_upstream - "-" 36 99 1497 1497 "-" "python-requests/2.28.2" "d64d635a-4727-440f-8c60-698d638f3537"
"cray-smd" "10.34.0.40:27779" outbound|80||cray-smd.services.svc.cluster.local 10.40.0.60:45638 10.25.149.14:80 10.40.0.60:47162 - default stream: stdout
tag: containers hostname: ncn-w002 container: istio-proxy _p: F @version: 1 time: Apr 25, 2025 @ 16:03:19.289 @timestamp: Apr 25, 2025 @ 16:03:19.580
event.original: {"timereported":"2025-04-25T21:03:19.289161436Z","time":"2025-04-25T21:03:19.289161436Z","stream":"stdout","_p":"F","log":"[2025-04-
```

Searching Logs – Namespace Services – Message traceback – Message field





namespace: services

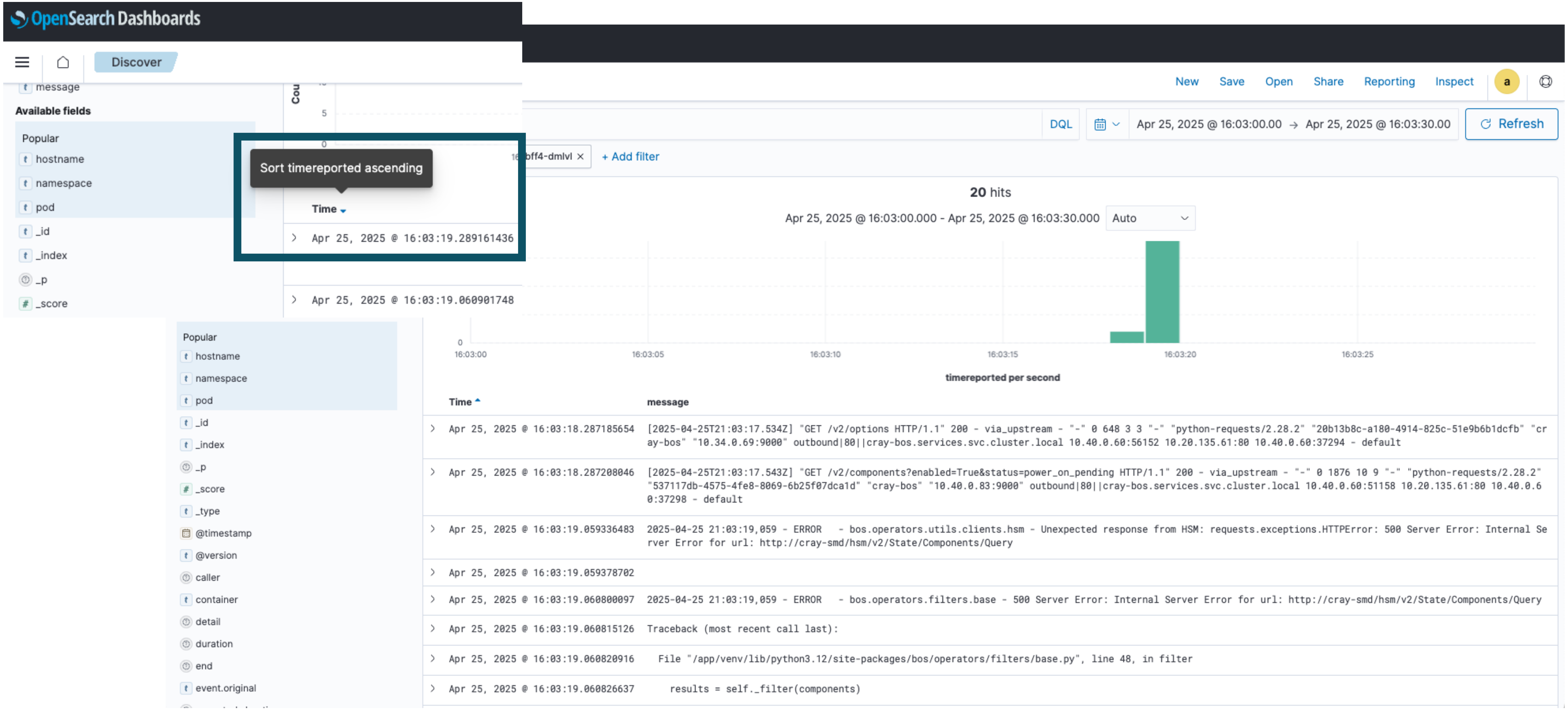
pod: cray-bos-operator-power-on-68964cbff4-dmlvl

20 hits

Apr 25, 2025 @ 16:03:00.000 - Apr 25, 2025 @ 16:03:30.000

```
Time ▾                            message
Apr 25, 2025 @ 16:03:19.289161436 [2025-04-25T21:03:17.556Z] "POST /hsm/v2/State/Components/Query HTTP/1.1" 500 - via_upstream - "-" 36 99 1497 1497 "-" "python-requests/2.28.2" "d64d635a-4727-4
                                  40f-8c60-698d638f3537" "cray-smd" "10.34.0.40:27779" outbound|80||cray-smd.services.svc.cluster.local 10.40.0.60:45638 10.25.149.14:80 10.40.0.60:47162 - defaul
                                  t
Apr 25, 2025 @ 16:03:19.060901748 requests.exceptions.HTTPError: 500 Server Error: Internal Server Error for url: http://cray-smd/hsm/v2/State/Components/Query
Apr 25, 2025 @ 16:03:19.060896739     raise HTTPError(http_error_msg, response=self)
Apr 25, 2025 @ 16:03:19.060891930   File "/app/venv/lib/python3.12/site-packages/requests/models.py", line 1021, in raise_for_status
Apr 25, 2025 @ 16:03:19.060887621     response.raise_for_status()
Apr 25, 2025 @ 16:03:19.060882472   File "/app/venv/lib/python3.12/site-packages/bos/operators/utils/clients/hsm.py", line 136, in get_components
Apr 25, 2025 @ 16:03:19.060878023     raise e
Apr 25, 2025 @ 16:03:19.060868846   File "/app/venv/lib/python3.12/site-packages/bos/operators/utils/clients/hsm.py", line 143, in get_components
Apr 25, 2025 @ 16:03:19.060848859                 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

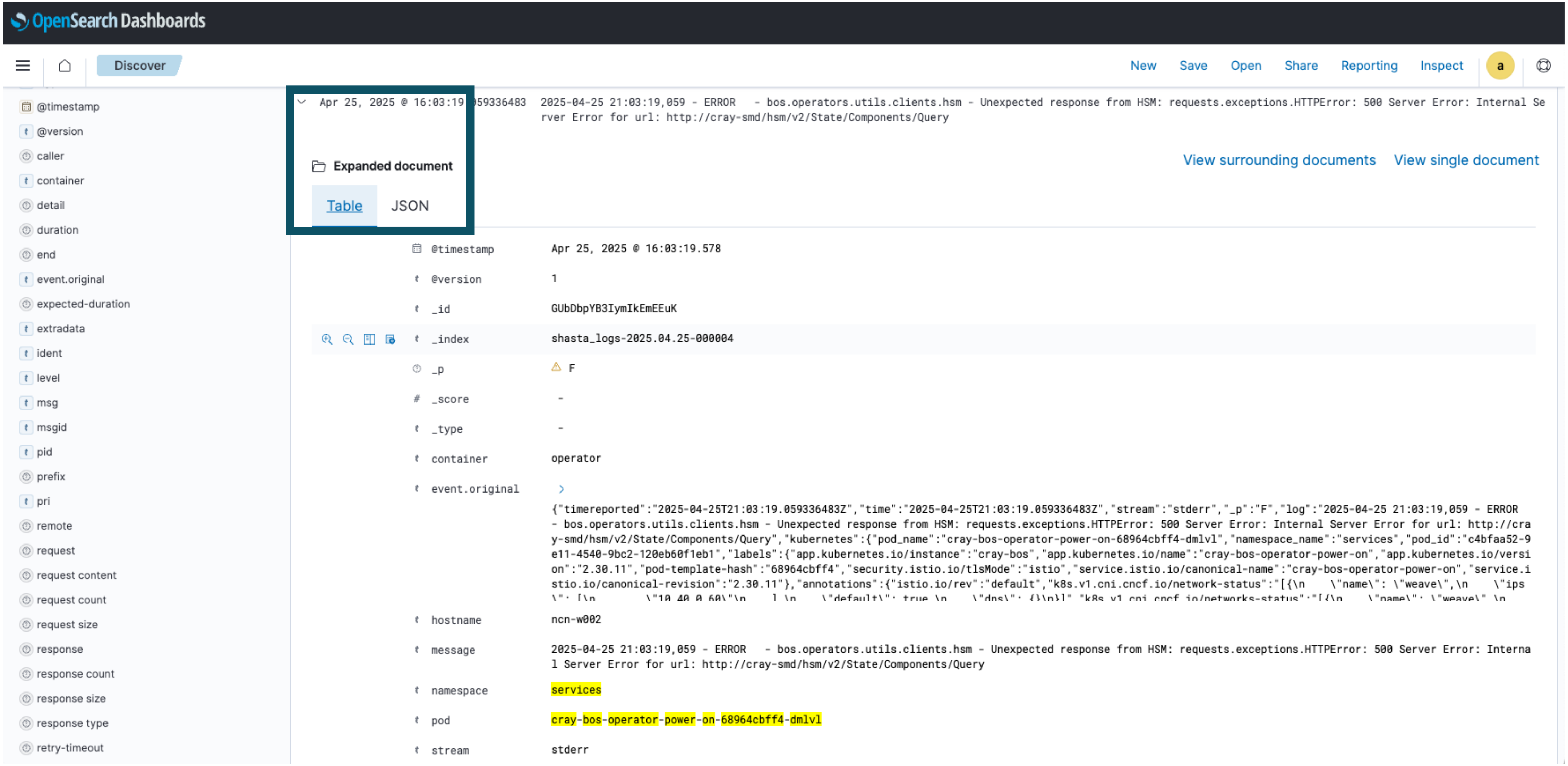
Searching Logs – Namespace Services – Message traceback – Time Sort



```
Apr 25, 2025 @ 16:03:19.289161436

Apr 25, 2025 @ 16:03:19.060901748
```

```
Time ▲                               message
Apr 25, 2025 @ 16:03:18.287185654  [2025-04-25T21:03:17.534Z] "GET /v2/options HTTP/1.1" 200 - via_upstream - "-" 0 648 3 3 "-" "python-requests/2.28.2" "20b13b8c-a180-4914-825c-51e9b6b1dcfb" "cr
                                   ay-bos" "10.34.0.69:9000" outbound|80||cray-bos.services.svc.cluster.local 10.40.0.60:56152 10.20.135.61:80 10.40.0.60:37294 - default
Apr 25, 2025 @ 16:03:18.287208046  [2025-04-25T21:03:17.543Z] "GET /v2/components?enabled=True&status=power_on_pending HTTP/1.1" 200 - via_upstream - "-" 0 1876 10 9 "-" "python-requests/2.28.2"
                                   "537117db-4575-4fe8-8069-6b25f07dca1d" "cray-bos" "10.40.0.83:9000" outbound|80||cray-bos.services.svc.cluster.local 10.40.0.60:51158 10.20.135.61:80 10.40.0.6
                                   0:37298 - default
Apr 25, 2025 @ 16:03:19.059336483  2025-04-25 21:03:19,059 - ERROR   - bos.operators.utils.clients.hsm - Unexpected response from HSM: requests.exceptions.HTTPError: 500 Server Error: Internal Se
                                   rver Error for url: http://cray-smd/hsm/v2/State/Components/Query
Apr 25, 2025 @ 16:03:19.059378702
Apr 25, 2025 @ 16:03:19.060800097  2025-04-25 21:03:19,059 - ERROR   - bos.operators.filters.base - 500 Server Error: Internal Server Error for url: http://cray-smd/hsm/v2/State/Components/Query
Apr 25, 2025 @ 16:03:19.060815126  Traceback (most recent call last):
Apr 25, 2025 @ 16:03:19.060820916    File "/app/venv/lib/python3.12/site-packages/bos/operators/filters/base.py", line 48, in filter
Apr 25, 2025 @ 16:03:19.060826637      results = self._filter(components)
```

Searching Logs – Namespace Services – Message traceback – Expanded



```
Apr 25, 2025 @ 16:03:19.059336483  2025-04-25 21:03:19,059 - ERROR   - bos.operators.utils.clients.hsm - Unexpected response from HSM: requests.exceptions.HTTPError: 500 Server Error: Internal Server Error for url: http://cray-smd/hsm/v2/State/Components/Query
```

@timestamp	Apr 25, 2025 @ 16:03:19.578
@version	1
_id	GUbDbpYB3IymIkEmEEuK
_index	shasta_logs-2025.04.25-000004
_p	⚠ F
_score	-
_type	-
container	operator
event.original	{"timereported":"2025-04-25T21:03:19.059336483Z","time":"2025-04-25T21:03:19.059336483Z","stream":"stderr","_p":"F","log":"2025-04-25 21:03:19,059 - ERROR - bos.operators.utils.clients.hsm - Unexpected response from HSM: requests.exceptions.HTTPError: 500 Server Error: Internal Server Error for url: http://cray-smd/hsm/v2/State/Components/Query","kubernetes":{"pod_name":"cray-bos-operator-power-on-68964cbff4-dmlvl","namespace_name":"services","pod_id":"c4bfaa52-9e11-4540-9bc2-120eb60f1eb1","labels":{"app.kubernetes.io/instance":"cray-bos","app.kubernetes.io/name":"cray-bos-operator-power-on","app.kubernetes.io/version":"2.30.11","pod-template-hash":"68964cbff4","security.istio.io/tlsMode":"istio","service.istio.io/canonical-name":"cray-bos-operator-power-on","service.istio.io/canonical-revision":"2.30.11"},"annotations":{"istio.io/rev":"default","k8s.v1.cni.cncf.io/network-status":"[{\n \"name\": \"weave\",\n \"ips\": [\n \"10.40.0.60\"\n],\n \"default\": true,\n \"dns\": {}\n}]","k8s.v1.cni.cncf.io/networks-status":"[{\n \"name\": \"weave\",\n
hostname	ncn-w002
message	2025-04-25 21:03:19,059 - ERROR - bos.operators.utils.clients.hsm - Unexpected response from HSM: requests.exceptions.HTTPError: 500 Server Error: Internal Server Error for url: http://cray-smd/hsm/v2/State/Components/Query
namespace	services
pod	cray-bos-operator-power-on-68964cbff4-dmlvl
stream	stderr

Telemetry and Grafana

- Grafana with CSM
- Refine existing dashboards
- Create new dashboards
- Grafana navigation
- Identify data sources for a dashboard

Grafana with CSM

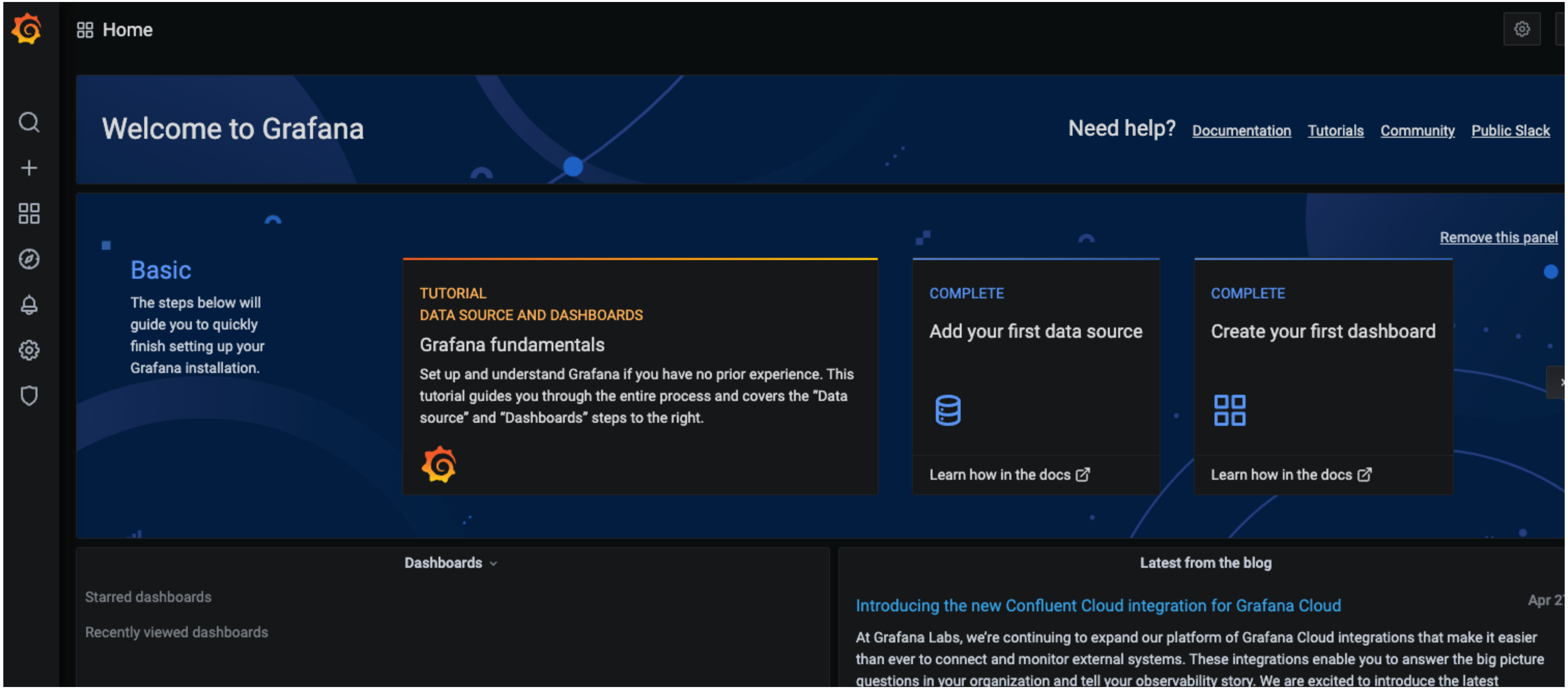
- Two Grafana instances are present on a CSM system
  - System Management Health Grafana
    - Includes numerous dashboards for visualizing metrics from prometheus and prometheus-istio

      https://grafana.cmn.SYSTEM_DOMAIN_NAME/
  - SMA Grafana
    - Includes system metric monitoring from these sources
      - Lightweight Data Monitoring Service (LDMS) statistics
      - HSN fabric performance, errors, congestion, and other statistics
      - Power, temperature and other sensor data from node, cabinet, and switch controllers

      https://sma-grafana.cmn.SYSTEM_DOMAIN_NAME/
  - Determine the external domain name by running the following command on any Kubernetes node:

```
ncn-m# kubectl get secret site-init -n loftsman \
-o jsonpath='{.data.customizations\.yaml}' | base64 -d | grep "external:"
external: SYSTEM_DOMAIN_NAME
```

System Management Health Grafana



Home

Welcome to Grafana

Need help? Documentation Tutorials Community Public Slack

Remove this panel

Basic

The steps below will guide you to quickly finish setting up your Grafana installation.

TUTORIAL
DATA SOURCE AND DASHBOARDS

Grafana fundamentals

Set up and understand Grafana if you have no prior experience. This tutorial guides you through the entire process and covers the “Data source” and “Dashboards” steps to the right.

COMPLETE

Add your first data source

Learn how in the docs

COMPLETE

Create your first dashboard

Learn how in the docs

Dashboards

Starred dashboards

Recently viewed dashboards

Latest from the blog

Introducing the new Confluent Cloud integration for Grafana Cloud

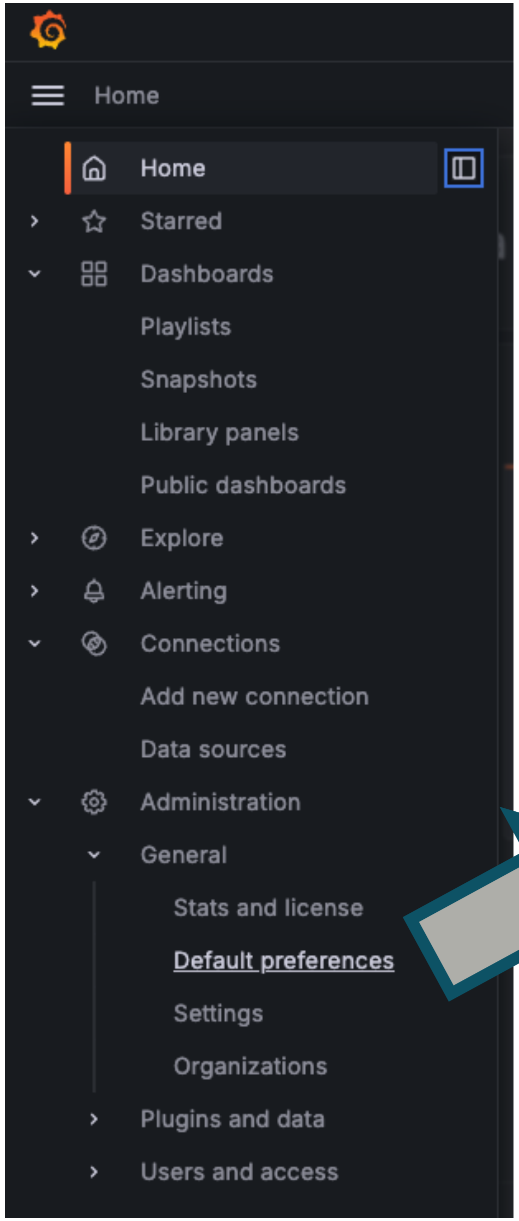
Apr 2

At Grafana Labs, we’re continuing to expand our platform of Grafana Cloud integrations that make it easier than ever to connect and monitor external systems. These integrations enable you to answer the big picture questions in your organization and tell your observability story. We are excited to introduce the latest

Choose default background

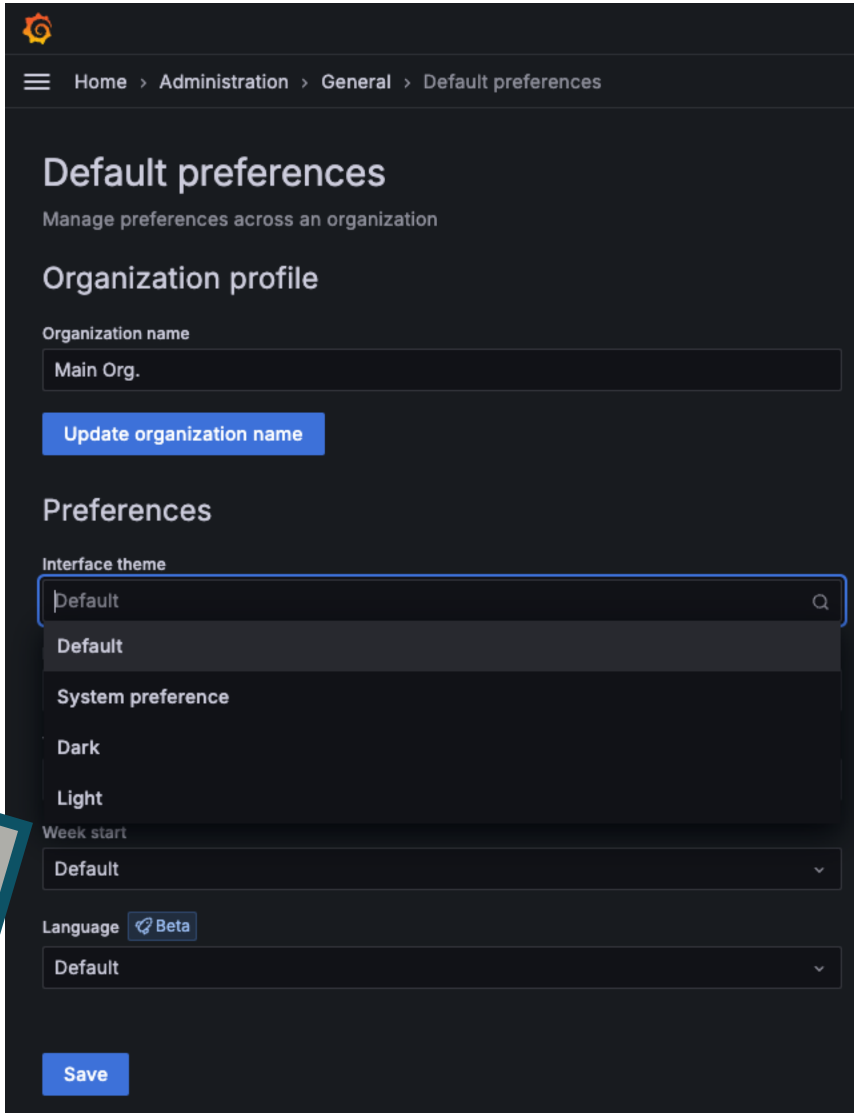
- Home menu has many Settings settings to adjust
- A different background color can be set
  - General menu
    - Default preferences
      - Select to get new window
      - Interface theme
        - Choose Default, System preference, Dark, or Light

- Default preferences window - top line shows hierarchical navigation to get to other areas
  - Home
  - Home/Administration
  - Home/Administration/General



Home

- Home
- Starred
- Dashboards
  - Playlists
  - Snapshots
  - Library panels
  - Public dashboards
- Explore
- Alerting
- Connections
  - Add new connection
  - Data sources
- Administration
  - General
    - Stats and license
    - Default preferences
    - Settings
    - Organizations
  - Plugins and data
  - Users and access



Home › Administration › General › Default preferences

Default preferences

Manage preferences across an organization

Organization profile

Organization name: Main Org.

Update organization name

Preferences

Interface theme: Default

- Default
- System preference
- Dark
- Light

Week start: Default

Language Beta: Default

Save

Refine existing Grafana dashboards

- Time range of data
  - Upper right corner of screen shows chosen time range
  - Pull down to change time range
- Refresh frequency
  - Can set dashboard to not refresh or refresh at specific interval
- Button to immediately refresh dashboard
- Drilling into data
  - Upper left on some dashboards has datasource and fields that can adjust scope of data being displayed
    - Compute nodes or NCNs
    - Kubernetes namespaces and pods
    - Geolocation by cabinet ID or lower in the hierarchy of components
    - One device or multiple devices or devices in a cabinet ID
    - One or more ports on a switch
    - Etc.

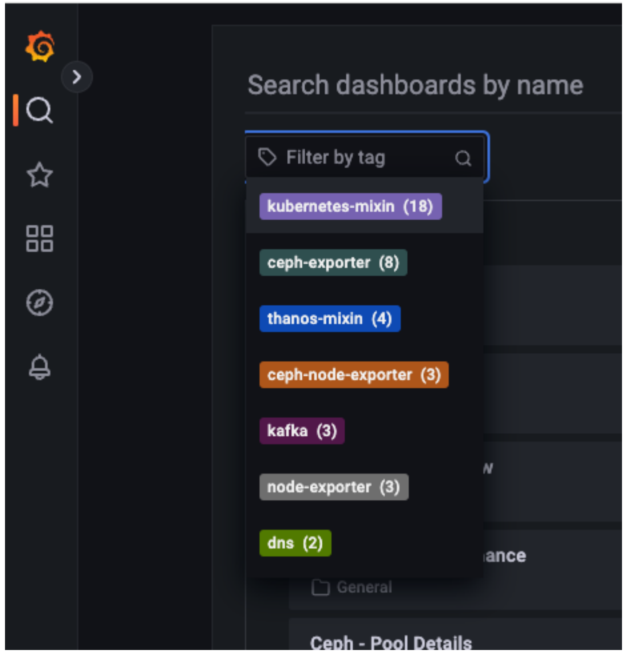
Create new Grafana dashboards

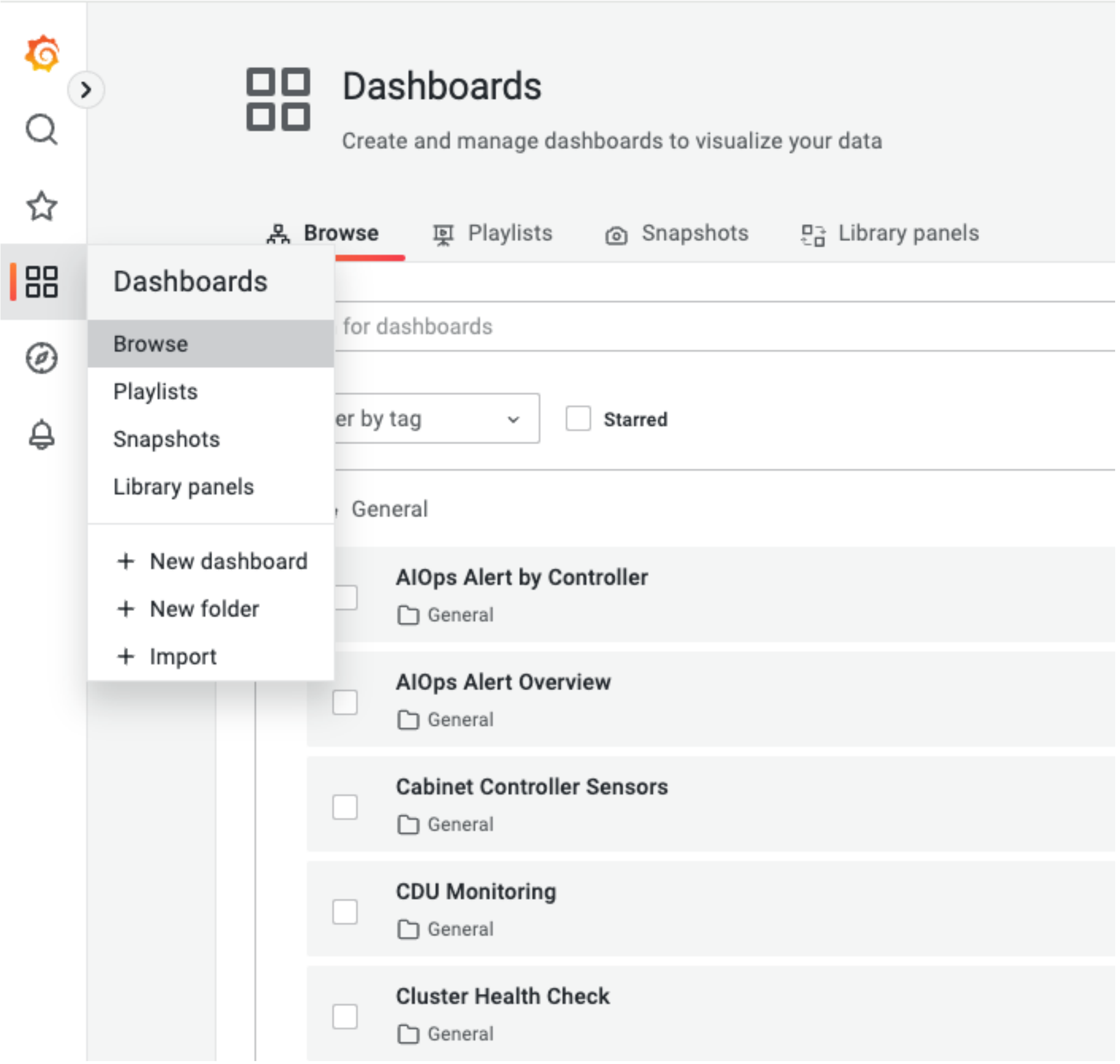
- Grafana's features and dashboard creation
  - https://grafana.com/docs/grafana/latest/
- Grafana panels and visualizations
  - The panel is the basic visualization building block in Grafana
  - Each panel has a query editor specific to the data source selected in the panel
  - The query editor allows you to build a query that returns the data you want to visualize.
  - https://grafana.com/docs/grafana/latest/features/panels/panels/
- Grafana dashboards
  - A dashboard is a set of one or more panels organized and arranged into one or more rows
  - https://grafana.com/docs/grafana/latest/features/dashboard/dashboards/

Grafana navigation

- Navigation bar on left side
  - Select 4 boxes icon to see list of dashboards
  - Choose browse

  - Select magnifying glass to search dashboards





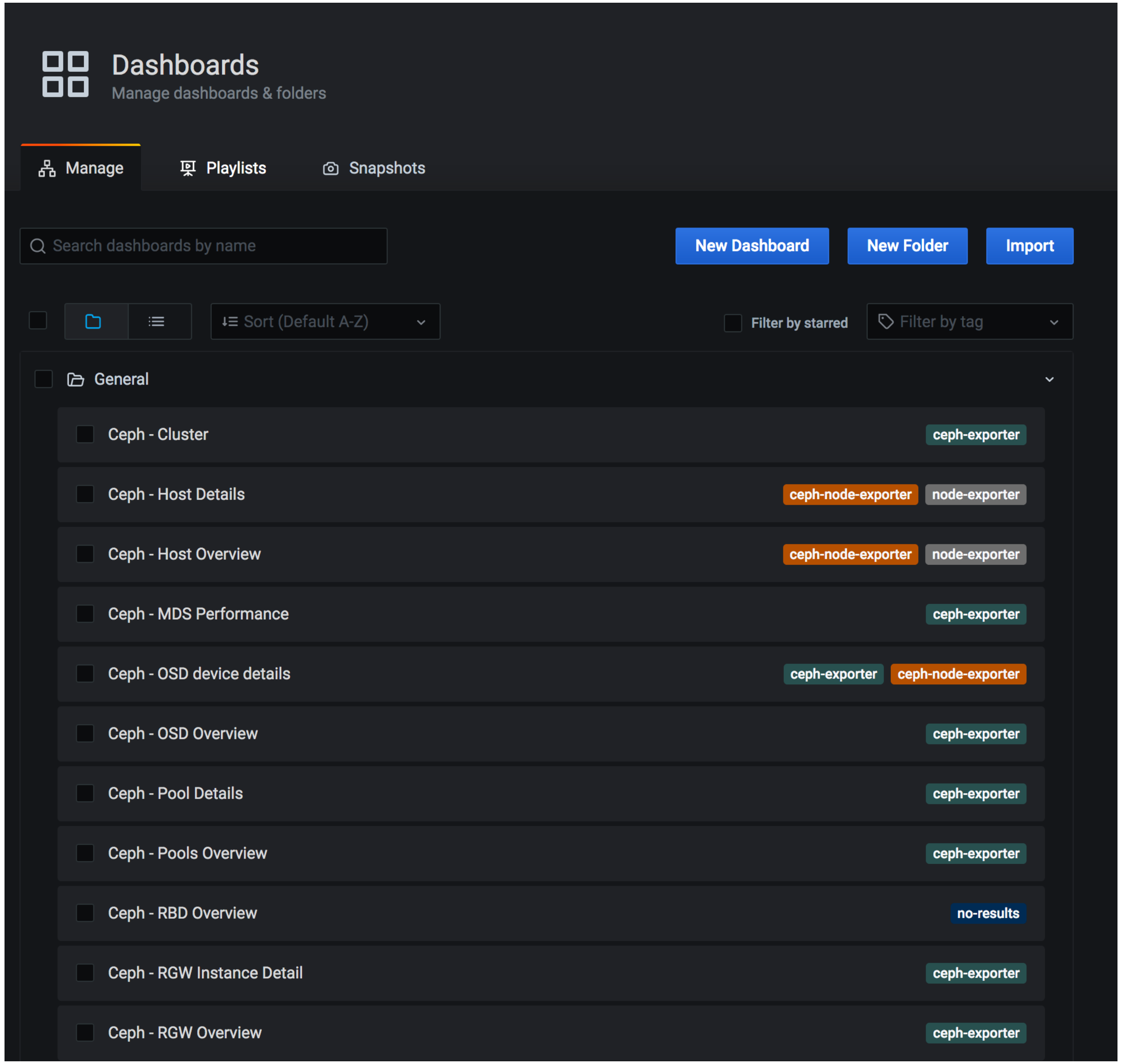
Data sources for Grafana dashboards

System Management Health Grafana dashboard sources

- amperage
- canu
- ceph-exporter
- ceph-node-exporter
- clusterview
- coredns
- dns
- fan speed
- goss
- hwmon
- IUF
- kafka
- Kea-dhcp
- kubernetes-mixin
- linux
- network
- no-results
- node
- node-exporter
- operator
- power
- Prometheus
- resolver
- smartmon
- temperature
- thanos-mixin
- unbound
- VictoriaMetrics
- vmagent
- vmalert
- voltage

SMA Grafana dashboard sources

- aiops
- AIops sub dashboard
- Alerta_dashboard
- crayFabricHealth
- DMTF
- fabric
- HMS
- HSN
- LDMS
- Slingshot
- slingshot_main_dashboard
- ss_network_status
- Ss_perf
- ss_quality_perf
- Sub dashboard



Dashboards
Manage dashboards & folders

Manage | Playlists | Snapshots

Search dashboards by name

New Dashboard | New Folder | Import

Sort (Default A-Z) | □ Filter by starred | Filter by tag

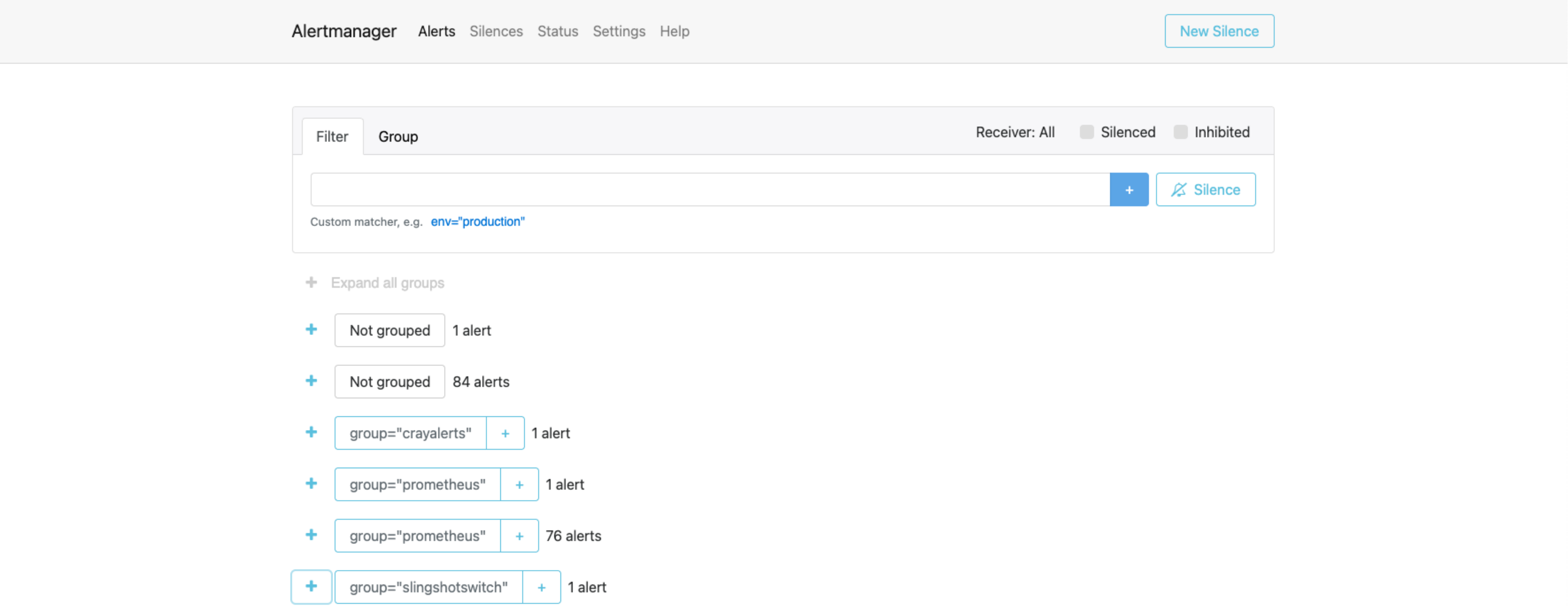
□ General

□ Ceph - Cluster	ceph-exporter
□ Ceph - Host Details	ceph-node-exporter node-exporter
□ Ceph - Host Overview	ceph-node-exporter node-exporter
□ Ceph - MDS Performance	ceph-exporter
□ Ceph - OSD device details	ceph-exporter ceph-node-exporter
□ Ceph - OSD Overview	ceph-exporter
□ Ceph - Pool Details	ceph-exporter
□ Ceph - Pools Overview	ceph-exporter
□ Ceph - RBD Overview	no-results
□ Ceph - RGW Instance Detail	ceph-exporter
□ Ceph - RGW Overview	ceph-exporter

Alerts and Alertmanager and cm health alertman

- Alertmanager
- cm health alertman

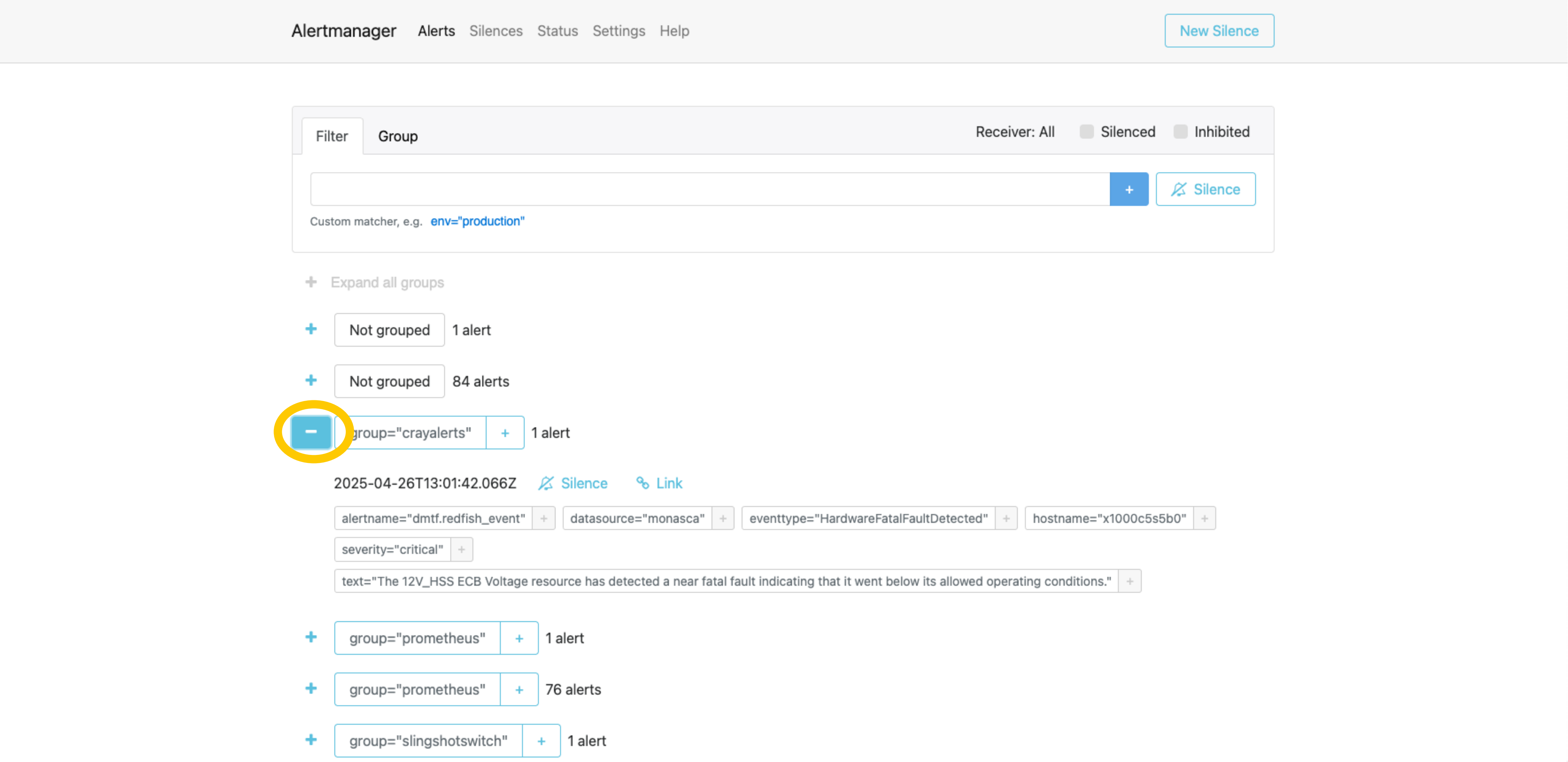
Alertmanager



Not grouped	1 alert
Not grouped	84 alerts
group="crayalerts"	1 alert
group="prometheus"	1 alert
group="prometheus"	76 alerts
group="slingshotswitch"	1 alert

https://alertmanager.cmn.SYSTEM_DOMAIN_NAME/

Alertmanager – DMTF.redfish_event



Not grouped 1 alert

Not grouped 84 alerts

group="crayalerts" 1 alert

2025-04-26T13:01:42.066Z Silence Link

alertname="dmtf.redfish_event" datasource="monasca" eventtype="HardwareFatalFaultDetected" hostname="x1000c5s5b0"

severity="critical"

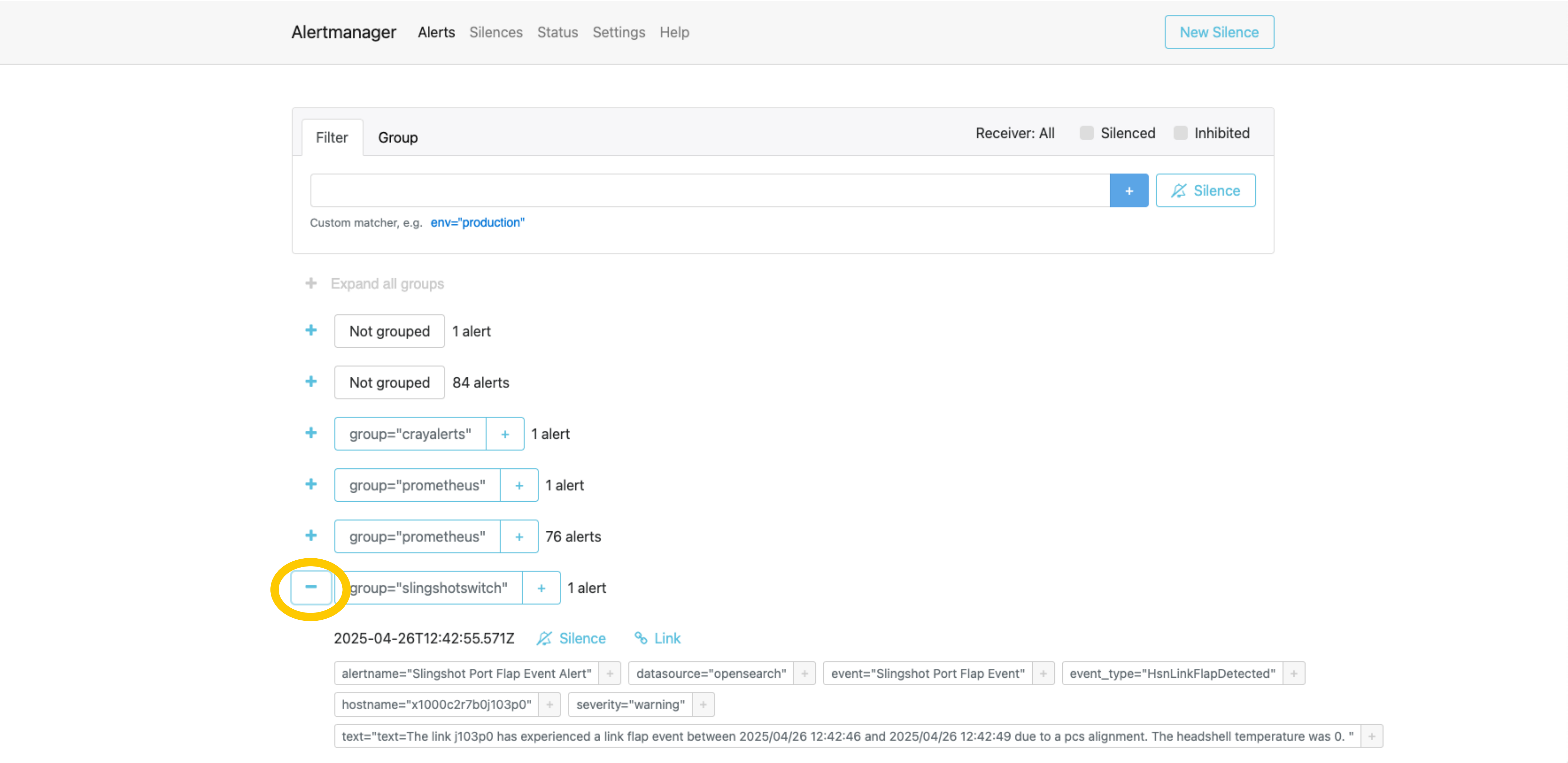
text="The 12V_HSS ECB Voltage resource has detected a near fatal fault indicating that it went below its allowed operating conditions."

group="prometheus" 1 alert

group="prometheus" 76 alerts

group="slingshotswitch" 1 alert

Alertmanager – Slingshot Port Flap Event



Not grouped 1 alert

Not grouped 84 alerts

group="crayalerts" 1 alert

group="prometheus" 1 alert

group="prometheus" 76 alerts

group="slingshotswitch" 1 alert

2025-04-26T12:42:55.571Z Silence Link

alertname="Slingshot Port Flap Event Alert"

datasource="opensearch"

event="Slingshot Port Flap Event"

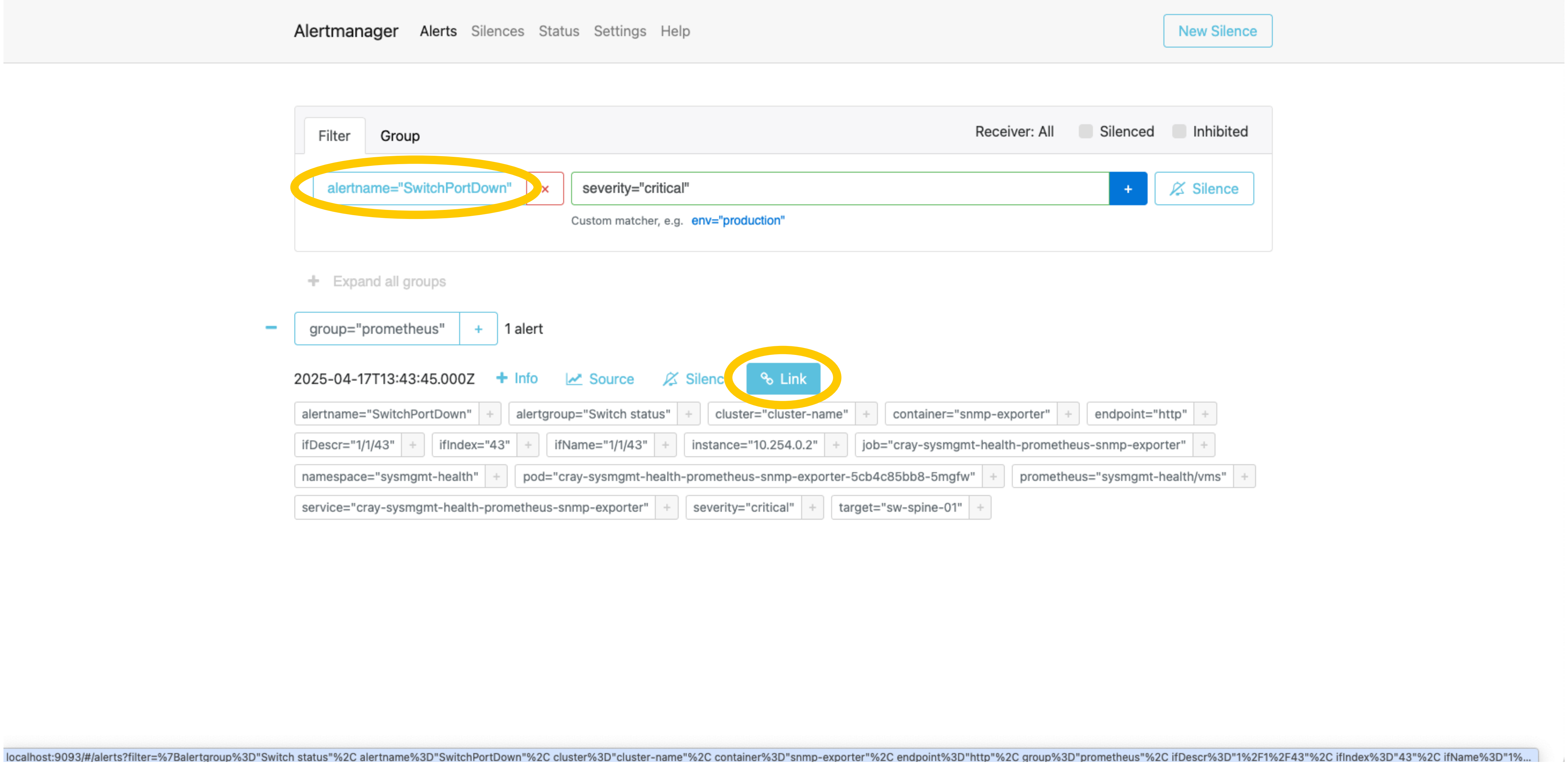
event_type="HsnLinkFlapDetected"

hostname="x1000c2r7b0j103p0"

severity="warning"

text="text=The link j103p0 has experienced a link flap event between 2025/04/26 12:42:46 and 2025/04/26 12:42:49 due to a pcs alignment. The headshell temperature was 0. "

Alertmanager – SwitchPortDown – sw-spine-001 1/1/43



alertname="SwitchPortDown"

severity="critical"

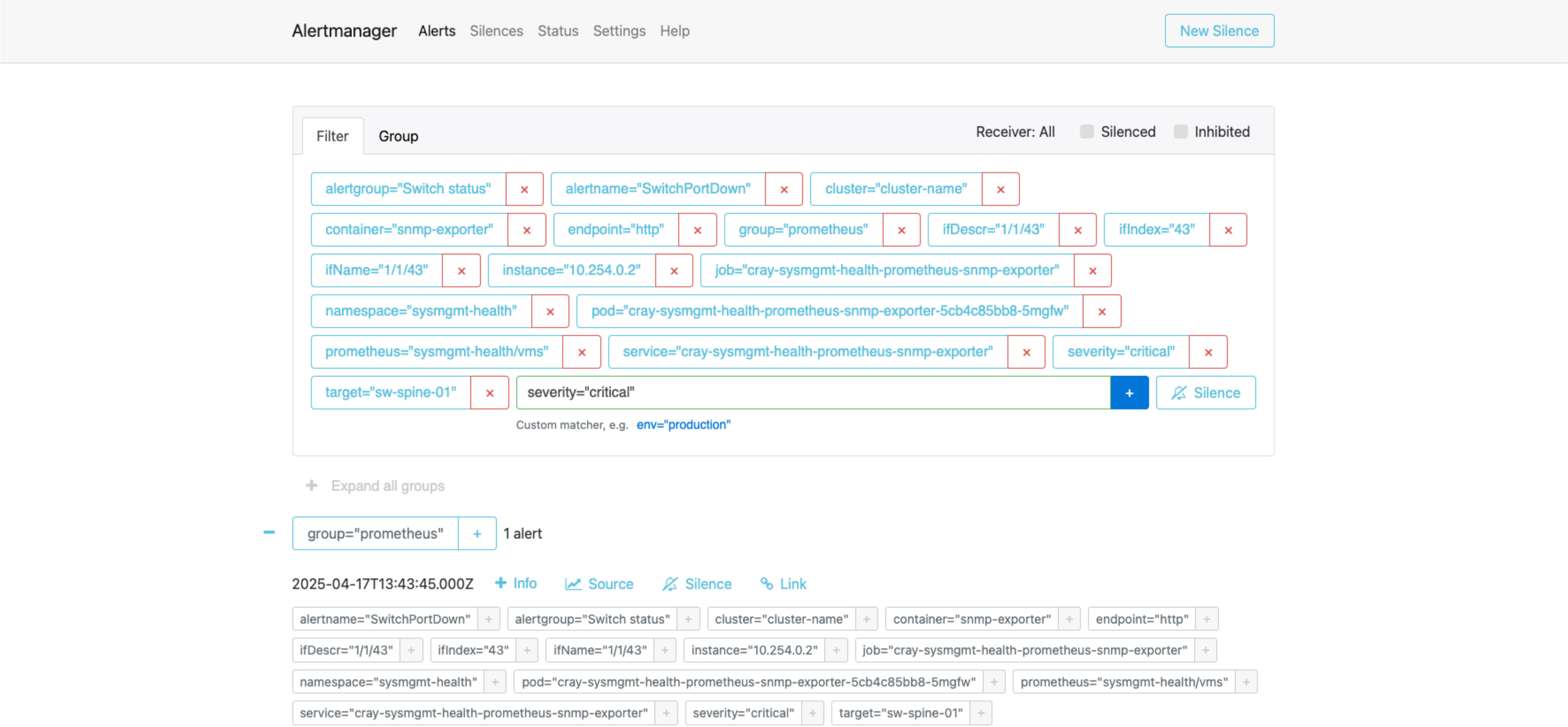
group="prometheus" 1 alert

2025-04-17T13:43:45.000Z

alertname="SwitchPortDown" alertgroup="Switch status" cluster="cluster-name" container="snmp-exporter" endpoint="http"
ifDescr="1/1/43" ifIndex="43" ifName="1/1/43" instance="10.254.0.2" job="cray-sysmgmt-health-prometheus-snmp-exporter"
namespace="sysmgmt-health" pod="cray-sysmgmt-health-prometheus-snmp-exporter-5cb4c85bb8-5mgfw" prometheus="sysmgmt-health/vms"
service="cray-sysmgmt-health-prometheus-snmp-exporter" severity="critical" target="sw-spine-01"

```
localhost:9093/#/alerts?filter=%7Balertgroup%3D"Switch status"%2C alertname%3D"SwitchPortDown"%2C cluster%3D"cluster-name"%2C container%3D"snmp-exporter"%2C endpoint%3D"http"%2C group%3D"prometheus"%2C ifDescr%3D"1%2F1%2F43"%2C ifIndex%3D"43"%2C ifName%3D"1%...
```

Alertmanager – SwitchPortDown – More Filters



Alertmanager   Alerts   Silences   Status   Settings   Help   New Silence

Filter   Group   Receiver: All   □ Silenced   □ Inhibited

alertgroup="Switch status" × alertname="SwitchPortDown" × cluster="cluster-name" ×

container="snmp-exporter" × endpoint="http" × group="prometheus" × ifDescr="1/1/43" × ifIndex="43" ×

ifName="1/1/43" × instance="10.254.0.2" × job="cray-sysmgmt-health-prometheus-snmp-exporter" ×

namespace="sysmgmt-health" × pod="cray-sysmgmt-health-prometheus-snmp-exporter-5cb4c85bb8-5mgfw" ×

prometheus="sysmgmt-health/vms" × service="cray-sysmgmt-health-prometheus-snmp-exporter" × severity="critical" ×

target="sw-spine-01" × severity="critical" + Silence

Custom matcher, e.g. env="production"

+ Expand all groups

− group="prometheus" + 1 alert

2025-04-17T13:43:45.000Z + Info Source Silence Link

alertname="SwitchPortDown" + alertgroup="Switch status" + cluster="cluster-name" + container="snmp-exporter" + endpoint="http" +

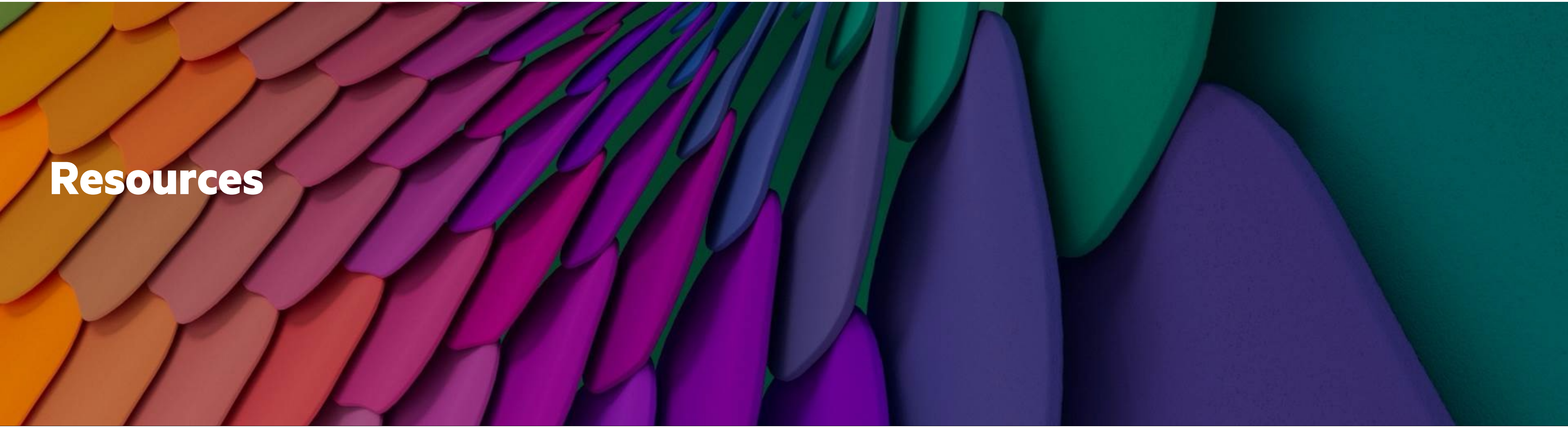
ifDescr="1/1/43" + ifIndex="43" + ifName="1/1/43" + instance="10.254.0.2" + job="cray-sysmgmt-health-prometheus-snmp-exporter" +

namespace="sysmgmt-health" + pod="cray-sysmgmt-health-prometheus-snmp-exporter-5cb4c85bb8-5mgfw" + prometheus="sysmgmt-health/vms" +

service="cray-sysmgmt-health-prometheus-snmp-exporter" + severity="critical" + target="sw-spine-01" +

cm health alertman

```
ncn# cm health alertman -h
usage: cm health alertman [-s|--status]
       [-g|--group]
       [-c|--cmd]
       [-h|--help]
       alert-group
       alert-group -h|--help
       alertman-command
       alertman-command -h|--help
positional arguments:
  alert-group         Display alerts associated with a
  specific alert-group, use
                      'cm health alertman -g' to see
  the supported alert-groups.
  alertman-command    Run a specific alertmanager
  command, use 'cm health
                      alertman -c' to see the supported
  alertman commands.
optional arguments:
  -h , --help         show this help message and exit.
  -s , --status       Display overall alerts status of
  the cluster.
  -g , --group        Display all supported alert-
  groups.
  -c , --cmd          Display all supported
  alertmanager commands.
---------------------------------------------
Examples:
 $ cm health alertman -s
 $ cm health alertman fabric
 $ cm health alertman fabric -h
 $ cm health alertman fabric -d n0
 $ cm health alertman query
 $ cm health alertman query -h
 $ cm health alertman silence –h
 ncn# cm health alertman -g
     compute
     fabric
     slingshothsn
     slingshotswitch
     prometheus
     aiops
     crayalerts
     cooldev
```



Resources

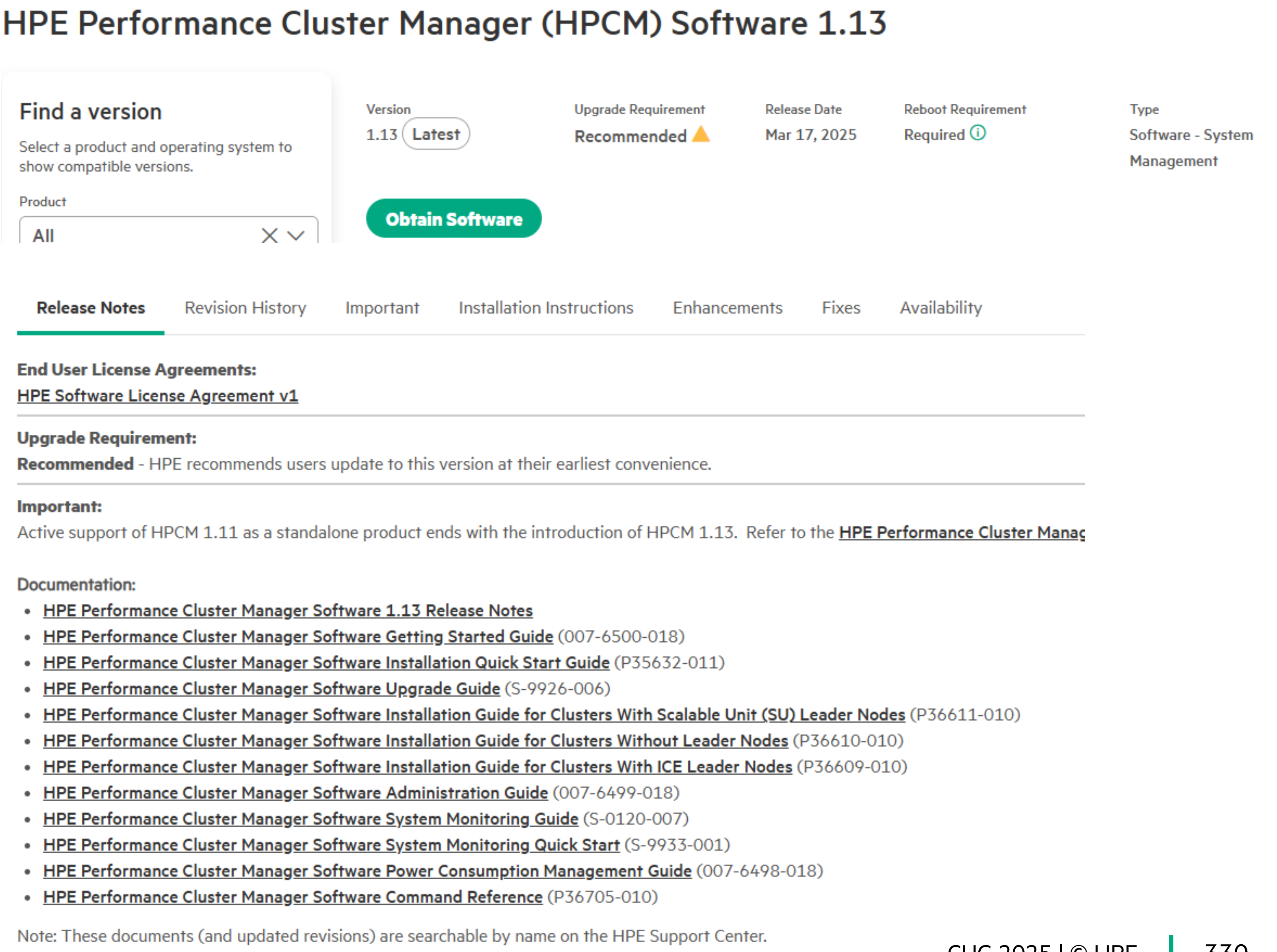
Resources

- Documentation
- Open-Source Software
- Training
- Related Presentations

Comprehensive documentation

Comprehensive documentation on **support.hpe.com** for administration, installation, monitoring, power management

*Keyword search **HPE Performance Cluster Manager** in Drivers and Software.*



HPE Performance Cluster Manager (HPCM) Software 1.13

Find a version
Select a product and operating system to show compatible versions.
Product
All

Version	Upgrade Requirement	Release Date	Reboot Requirement	Type
1.13 Latest	Recommended	Mar 17, 2025	Required	Software - System Management

End User License Agreements:
HPE Software License Agreement v1

Upgrade Requirement:
Recommended - HPE recommends users update to this version at their earliest convenience.

Important:
Active support of HPCM 1.11 as a standalone product ends with the introduction of HPCM 1.13. Refer to the HPE Performance Cluster Manag

Documentation:

- HPE Performance Cluster Manager Software 1.13 Release Notes
- HPE Performance Cluster Manager Software Getting Started Guide (007-6500-018)
- HPE Performance Cluster Manager Software Installation Quick Start Guide (P35632-011)
- HPE Performance Cluster Manager Software Upgrade Guide (S-9926-006)
- HPE Performance Cluster Manager Software Installation Guide for Clusters With Scalable Unit (SU) Leader Nodes (P36611-010)
- HPE Performance Cluster Manager Software Installation Guide for Clusters Without Leader Nodes (P36610-010)
- HPE Performance Cluster Manager Software Installation Guide for Clusters With ICE Leader Nodes (P36609-010)
- HPE Performance Cluster Manager Software Administration Guide (007-6499-018)
- HPE Performance Cluster Manager Software System Monitoring Guide (S-0120-007)
- HPE Performance Cluster Manager Software System Monitoring Quick Start (S-9933-001)
- HPE Performance Cluster Manager Software Power Consumption Management Guide (007-6498-018)
- HPE Performance Cluster Manager Software Command Reference (P36705-010)

Note: These documents (and updated revisions) are searchable by name on the HPE Support Center.

CSM Documentation - Installation

- HPE Cray EX System Software Getting Started Guide S-8000
- HPE Cray System Management (CSM) Markdown
  - https://github.com/Cray-HPE/docs-csm/tree/release/1.6
- HPE Cray System Management (CSM) HTML
  - https://cray-hpe.github.io/docs-csm/en-16/
- HPE Cray EX System CSM Diagnostics Installation and Configuration Guide
- HPE Cray EX System Diagnostic Utility (SDU) Installation Guide
- HPE Cray EX System HPC Firmware Pack Installation Guide S-8037
- HPE Cray EX System Monitoring Application Installation Guide S-8030
- HPE Cray Programming Environment Installation Guide: CSM on HPE Cray EX S-8003
- HPE Cray Supercomputing User Services Software Administration Guide: CSM on HPE Cray EX Systems (S-8063)
- HPE Slingshot Host Software Installation and Configuration Guide
- HPE Slingshot Release Notes
- HPE Slingshot Installation Guide for CSM
- HPE SUSE Linux Enterprise Operating System Installation Guide S-8028

CSM Documentation - Administration

- HPE Cray System Management (CSM) Markdown
  - https://github.com/Cray-HPE/docs-csm/tree/release/1.6
  - https://github.com/Cray-HPE/docs-csm/blob/release/1.6/operations/kubernetes/Kubernetes.md
  - https://github.com/Cray-HPE/docs-csm/blob/release/1.6/glossary.md
- HPE Cray System Management (CSM) HTML
  - https://cray-hpe.github.io/docs-csm/en-16/
- HPE Cray EX System CSM Diagnostics Administration Guide
- HPE Cray EX System Diagnostic Utility (SDU) Administration Guide
- HPE Cray EX System Monitoring Application Administration Guide S-8029
- HPE Cray Programming Environment User Guide: CSM on HPE Cray EX S-8005
- HPE Cray Supercomputing User Services Software Administration Guide: CSM on HPE Cray EX Systems (S-8063)
- HPE Cray User Access Node Software Administration Guide S-8033
- HPE Slingshot Host Software Administration Guide
- HPE Slingshot Host Software Troubleshooting Guide
- HPE Slingshot Administration Guide
- HPE Slingshot Fabric Command Reference Guide
- HPE Slingshot Troubleshooting
- HPE Slingshot Hardware Guide

Documentation – open-source tools

- CSM
  - MIT License
  - Github Hosted
    - https://github.com/Cray-HPE
  - Community Governance
    - https://github.com/Cray-HPE/community
  - Primary repository for the sat CLI written in Python: https://github.com/Cray-HPE/sat
  - Podman wrapper script written in Bash: https://github.com/Cray-HPE/sat-podman
  - An important library used by sat CLI: https://github.com/Cray-HPE/python-csm-api-client
  - Documentation starting point:
    - https://github.com/Cray-HPE/sat/blob/integration/CONTRIBUTING.md
    - https://github.com/Cray-HPE/sat/blob/integration/docs/developer/README.md

- 3rd party open-source
  - https://kubernetes.io/docs/home/
  - https://kubernetes.io/docs/reference/kubectl/cheatsheet/
  - https://lmgtfy.com/?q=kubernetes+troubleshooting
  - https://www.elastic.co/guide/en/kibana/current/index.html
  - https://grafana.com/docs/
  - https://github.com/aelsabbahy/goss
  - http://docs.ansible.com/
  - https://kubernetes.io/docs/reference/kubectl/jsonpath/
  - https://stedolan.github.io/jq/manual/
  - http://www.compciv.org/recipes/cli/jq-for-parsing-json/
  - https://osinside.github.io/kiwi/
  - https://ara.recordsansible.org/

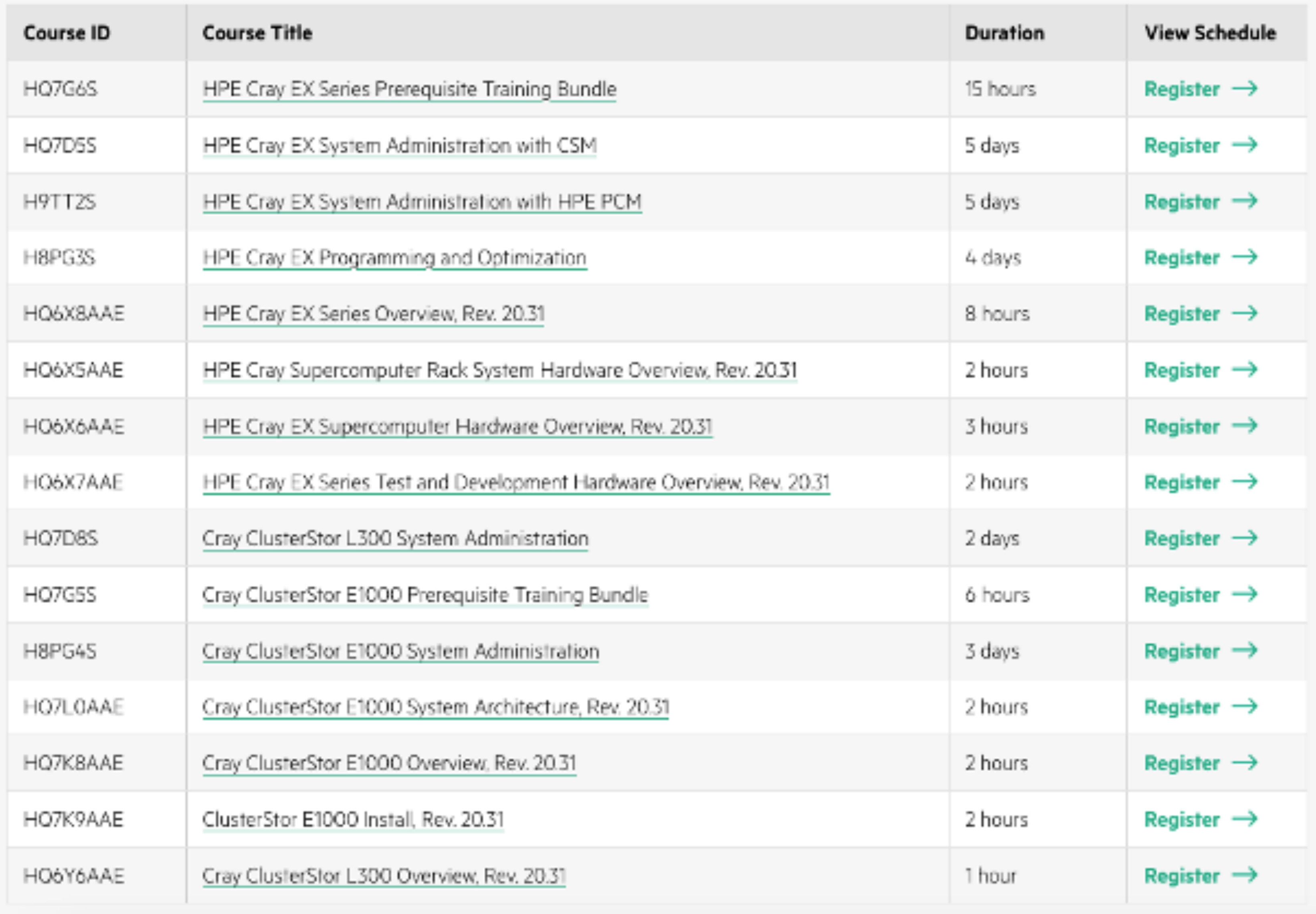
SUPERCOMPUTING: HPE CRAY EX Training

Where to start?

From HPE Edu
http://www.hpe.com/ww/training

- Select HPE Cray EX Series and ClusterStor Storage

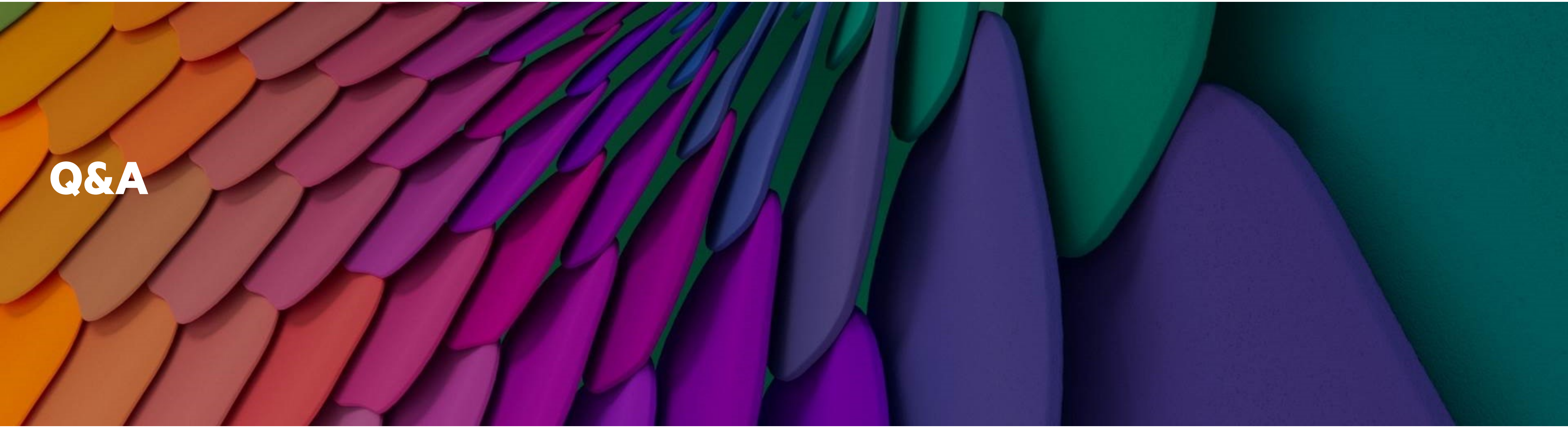
https://education.hpe.com/ww/en/training/portfolio/servers.html#ServersLearningPathsIntro



Course ID	Course Title	Duration	View Schedule
HQ7G6S	HPE Cray EX Series Prerequisite Training Bundle	15 hours	Register →
HQ7D5S	HPE Cray EX System Administration with CSM	5 days	Register →
H9TT2S	HPE Cray EX System Administration with HPE PCM	5 days	Register →
H8PG3S	HPE Cray EX Programming and Optimization	4 days	Register →
HQ6X8AAE	HPE Cray EX Series Overview, Rev. 20.31	8 hours	Register →
HQ6X5AAE	HPE Cray Supercomputer Rack System Hardware Overview, Rev. 20.31	2 hours	Register →
HQ6X6AAE	HPE Cray EX Supercomputer Hardware Overview, Rev. 20.31	3 hours	Register →
HQ6X7AAE	HPE Cray EX Series Test and Development Hardware Overview, Rev. 20.31	2 hours	Register →
HQ7D8S	Cray ClusterStor L300 System Administration	2 days	Register →
HQ7G5S	Cray ClusterStor E1000 Prerequisite Training Bundle	6 hours	Register →
H8PG4S	Cray ClusterStor E1000 System Administration	3 days	Register →
HQ7L0AAE	Cray ClusterStor E1000 System Architecture, Rev. 20.31	2 hours	Register →
HQ7K8AAE	Cray ClusterStor E1000 Overview, Rev. 20.31	2 hours	Register →
HQ7K9AAE	ClusterStor E1000 Install, Rev. 20.31	2 hours	Register →
HQ6Y6AAE	Cray ClusterStor L300 Overview, Rev. 20.31	1 hour	Register →

Related presentations and papers

- CUG 2025
  - BOF: CSM Updates, iSCSI boot content projection, and other CSM topics
  - BOF: CUG SIG System Monitoring Working Group
  - A Brief Summary of the HPCM Evolution Over Recent Releases
  - System Visualization Using Rackmap
  - HPE Slingshot Monitoring Software: Actionable Insights for HPC and AI Systems
  - Proactive Health Monitoring and Maintenance of High-Speed Slingshot Fabrics in HPC Environments
  - Hardware Triage Tool: Enhancements and Extensions
  - Detecting operating system noise with detect-detour
- CUG 2024
  - From Frontier to Framework: Enhancing Hardware Triage for Exascale Machines
  - AIOPS Empowered: Failure Prediction in System Management Software Tools
  - HPE Cray EX Power Monitoring Counters
  - Unification of Alerting Engines for Monitoring in System Management
- CUG 2022
  - Dealing with Metrics Data – Where is it, How to get it, What to do with it?



Q&A

Thank you

Sue Miller, susan.miller@hpe.com
Harold Longley, harold.longley@hpe.com
Pete Guyan, pguyan@hpe.com
Raghul Vasudevan, raghul-vasudevan@hpe.com