Hewlett Packard
Enterprise

AIOPS EMPOWERED FAILURE PREDICTION IN SYSTEM MANAGEMENT SOFTWARE TOOLS

Deepak Nanjundaiah
Subrahmanya Joshi
Sergey Serebryakov

May 09, 2024

CUG2024
Diverse Universe
PERTH AUSTRALIA

AGENDA

- Why AI and ML for HPC systems?
- AIOps Machine Learning and Failure Prediction Framework
- AIOps Dashboards
- Current Status and Future Work

WHY AI / ML FOR HPC DATA CENTERS?

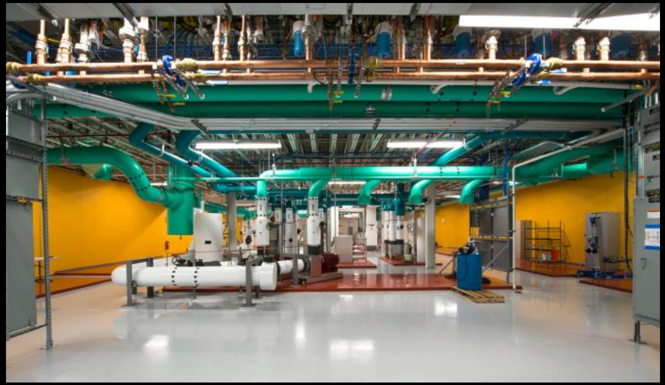
Large number of metrics (thousands)

Real-time data streams at NREL datacenter: $\mathcal{O}$(10^3) facility data points per minute and $\mathcal{O}$(10^6) Eagle data points per minute.

- Do not know where to look.
- Data is coming real time at high rates
- Threshold-based methods (setpoints) are used, but produce many false positives and thus not scalable
- Some anomalies can only be identified in high-dimensional spaces (multiple metrics).
- Setpoints and Dashboards are not always sufficient to identify anomalies

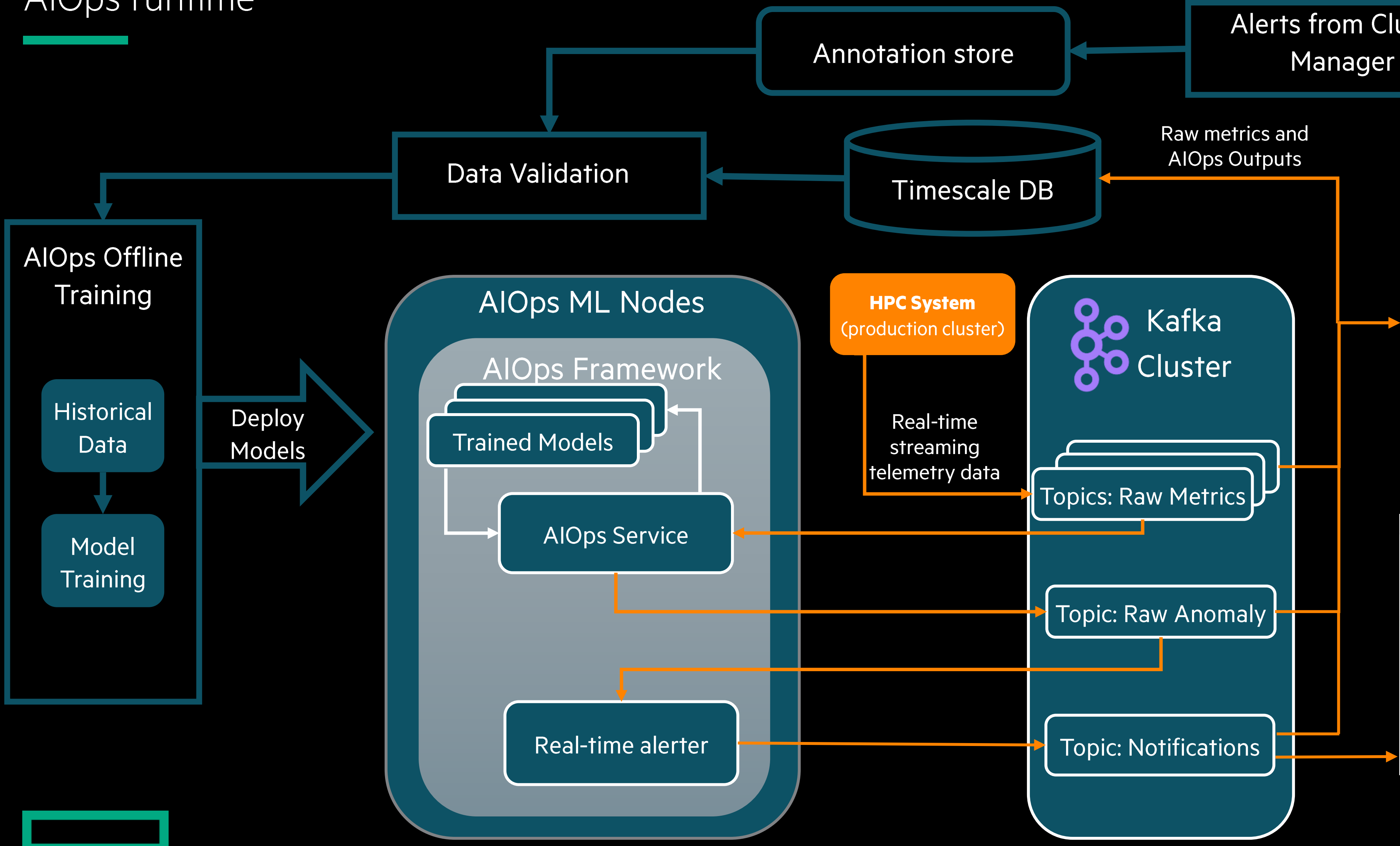
Broad range of problems (not only anomaly detection)

- Anomaly Detection (single metric and multi-metric models)
- Forecasting
- Preventative maintenance
- Reducing carbon footprint by optimizing datacenter ops (Digital Twins)



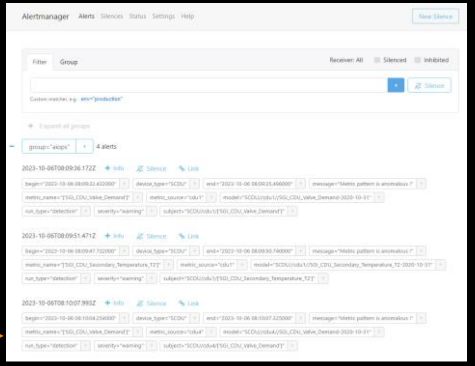
AIOPS MACHINE LEARNING FRAMEWORK

AIOps runtime



Analysis Dashboards

Notification Dashboards



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

group="aiops" 4 alerts

2023-10-06T08:09:36.172Z

begin="2023-10-06 08:09:32.432000" device_type="SCDU" end="2023-10-06 08:09:35.496000" message="Metric pattern is anomalous !"
metric_name="['SGI_CDU_Valve_Demand']" metric_source="cdu1" model="SCDU//cdu1//SGI_CDU_Valve_Demand-2020-10-31"
run_type="detection" severity="warning" subject="SCDU/cdu1/['SGI_CDU_Valve_Demand']"

2023-10-06T08:09:51.471Z

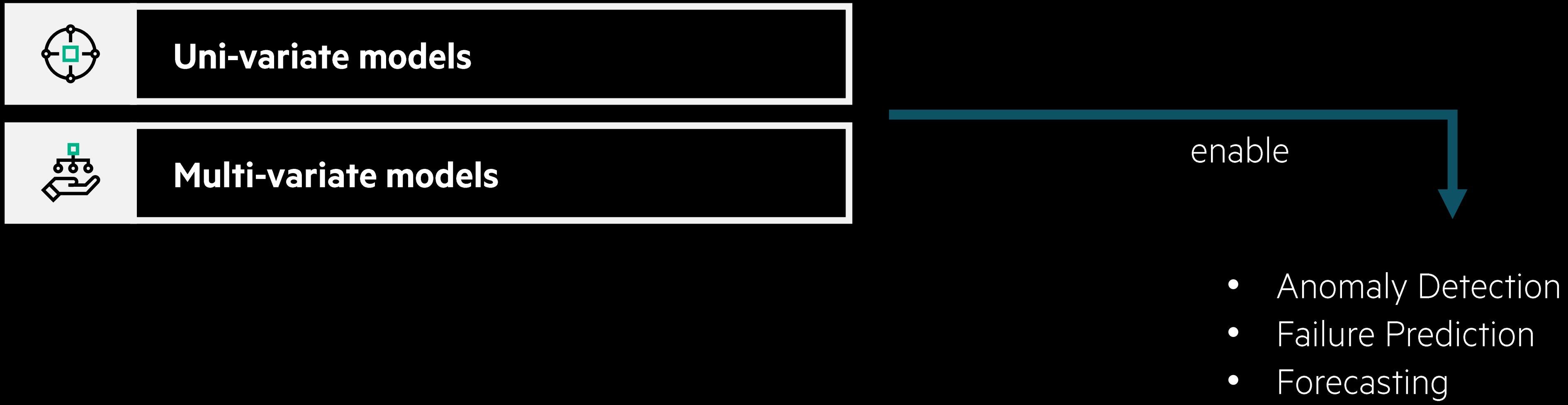
begin="2023-10-06 08:09:47.722000" device_type="SCDU" end="2023-10-06 08:09:50.740000" message="Metric pattern is anomalous !"
metric_name="['SGI_CDU_Secondary_Temperature_T2']" metric_source="cdu1" model="SCDU//cdu1//SGI_CDU_Secondary_Temperature_T2-2020-10-31"
run_type="detection" severity="warning" subject="SCDU/cdu1/['SGI_CDU_Secondary_Temperature_T2']"

2023-10-06T08:10:07.993Z

begin="2023-10-06 08:10:04.254000" device_type="SCDU" end="2023-10-06 08:10:07.325000" message="Metric pattern is anomalous !"
metric_name="['SGI_CDU_Valve_Demand']" metric_source="cdu4" model="SCDU//cdu4//SGI_CDU_Valve_Demand-2020-10-31"
run_type="detection" severity="warning" subject="SCDU/cdu4/['SGI_CDU_Valve_Demand']"

AIOPS MACHINE LEARNING FRAMEWORK

From statistical to machine learning models for time series data



Uni-variate models

Multi-variate models

enable

- Anomaly Detection
- Failure Prediction
- Forecasting

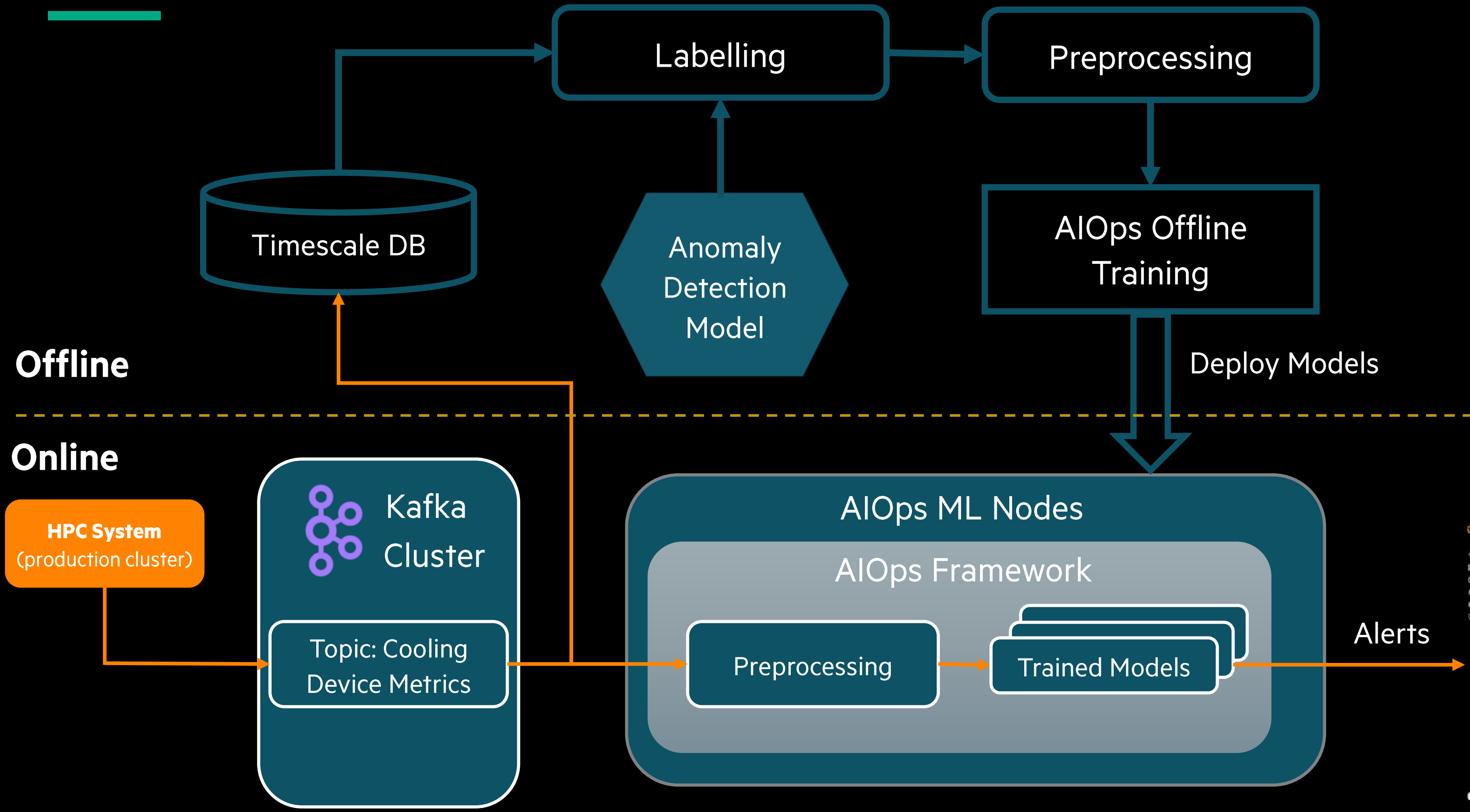
AIOPS MACHINE LEARNING FRAMEWORK

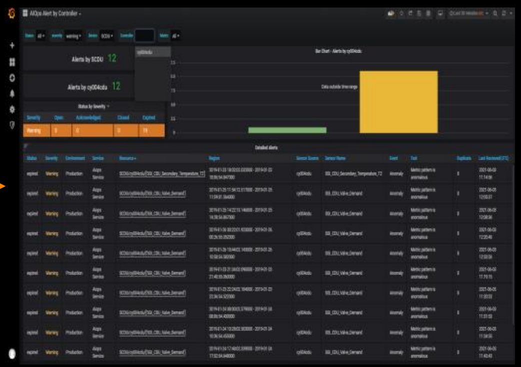
AIOps Failure Prediction Framework: Overview

- Some failures are easily predictable, while others are extremely hard to predict.
- 2 types of failures. Some failures occur suddenly, while others are preceded by some specific set of events.
- These occurrences could be anomalies or, at times, simply a normal sequence of events.
- We can train a model to understand these events, and during inference, could potentially predict them even before they happen.
- To accomplish this, we require historical data with known and accurately labeled failures.

AIOPS MACHINE LEARNING FRAMEWORK

AIOps Failure Prediction Framework: Online Inference and Offline training architecture

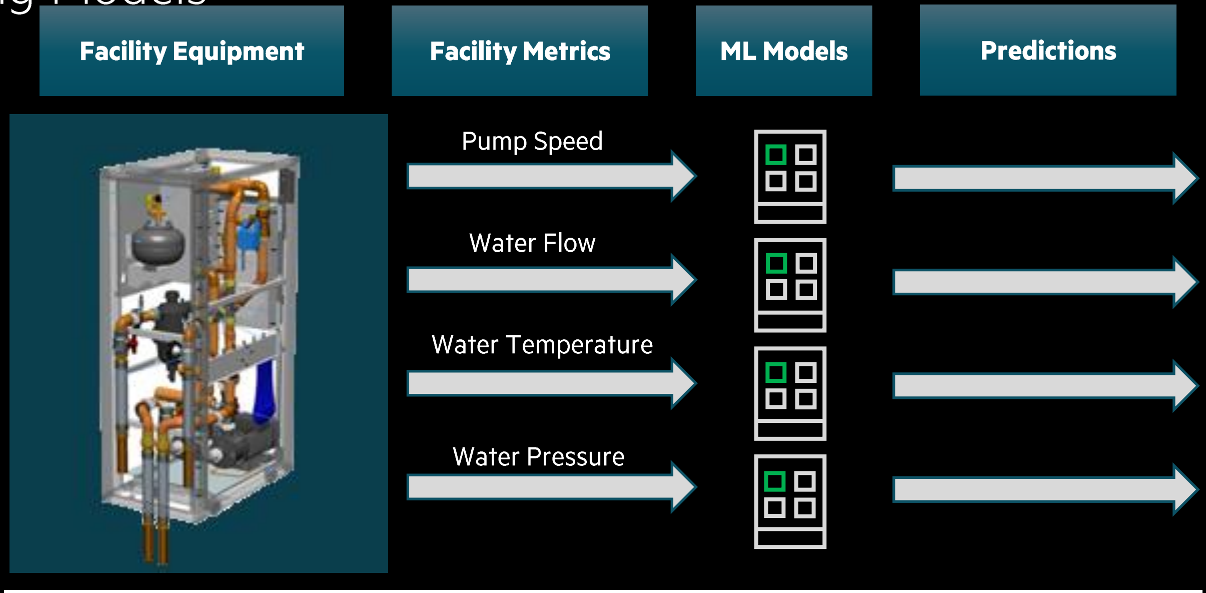


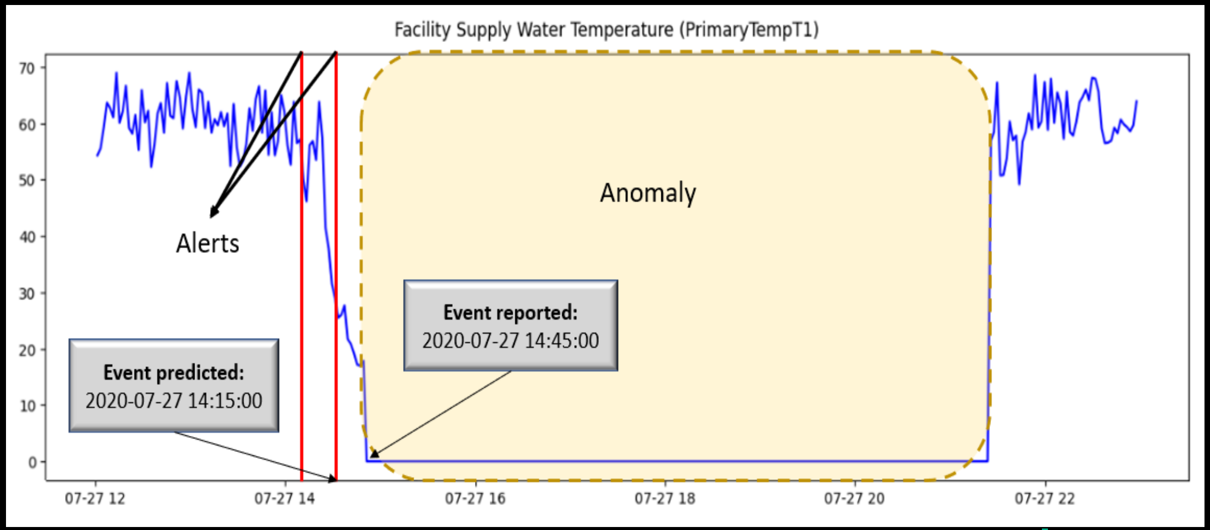


AIOPS MACHINE LEARNING FRAMEWORK

AIOps Failure Prediction Framework: Machine Learning Models

- We use LSTM and fully connected neural networks for failure prediction.
- LSTMs very effectively capture and utilize long term dependencies in sequential data.
- Anomaly detection models, which are part of AIOps itself label the data.
- Models are uni-variate, with each being trained and utilized for inference on a single metric.
- Models employ a semi-supervised approach, wherein labeled data is automatically generated from existing detection models.





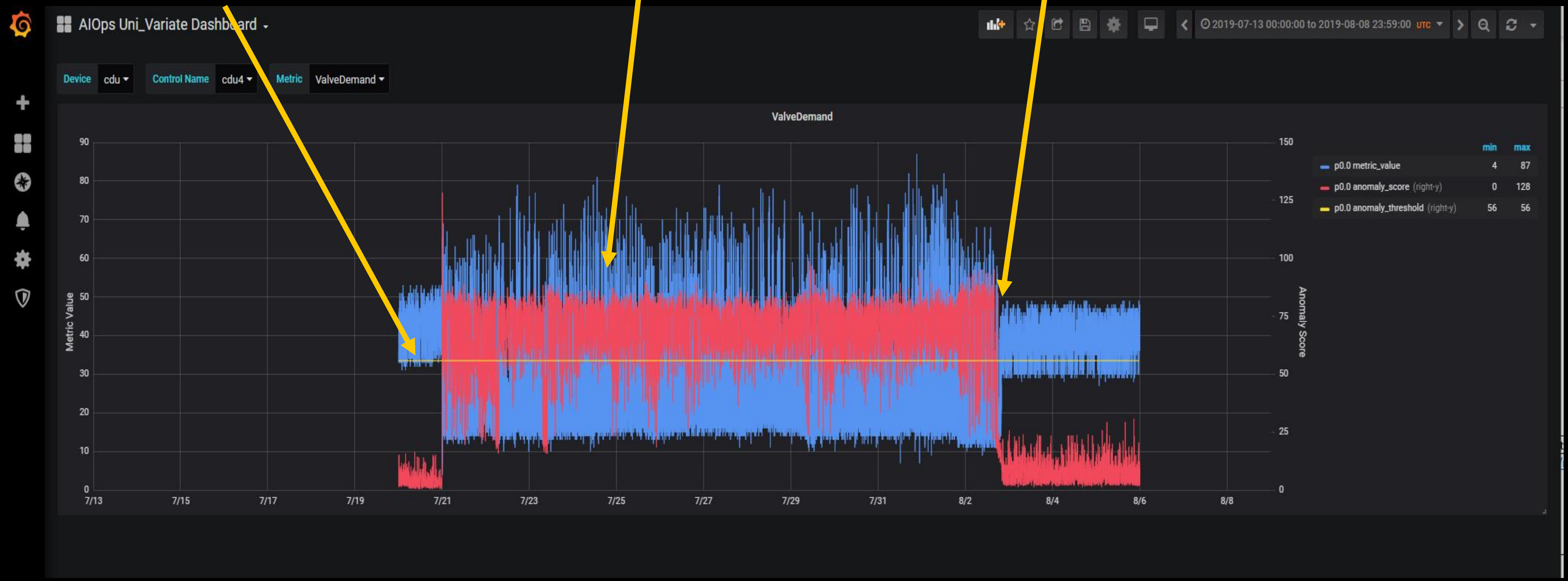
AIOPS DASHBOARDS

Univariate anomaly detection dashboard for Coolant Distribution Units (CDUs)

Raw metric (blue)

Anomaly threshold (yellow line)

Anomaly score (red)

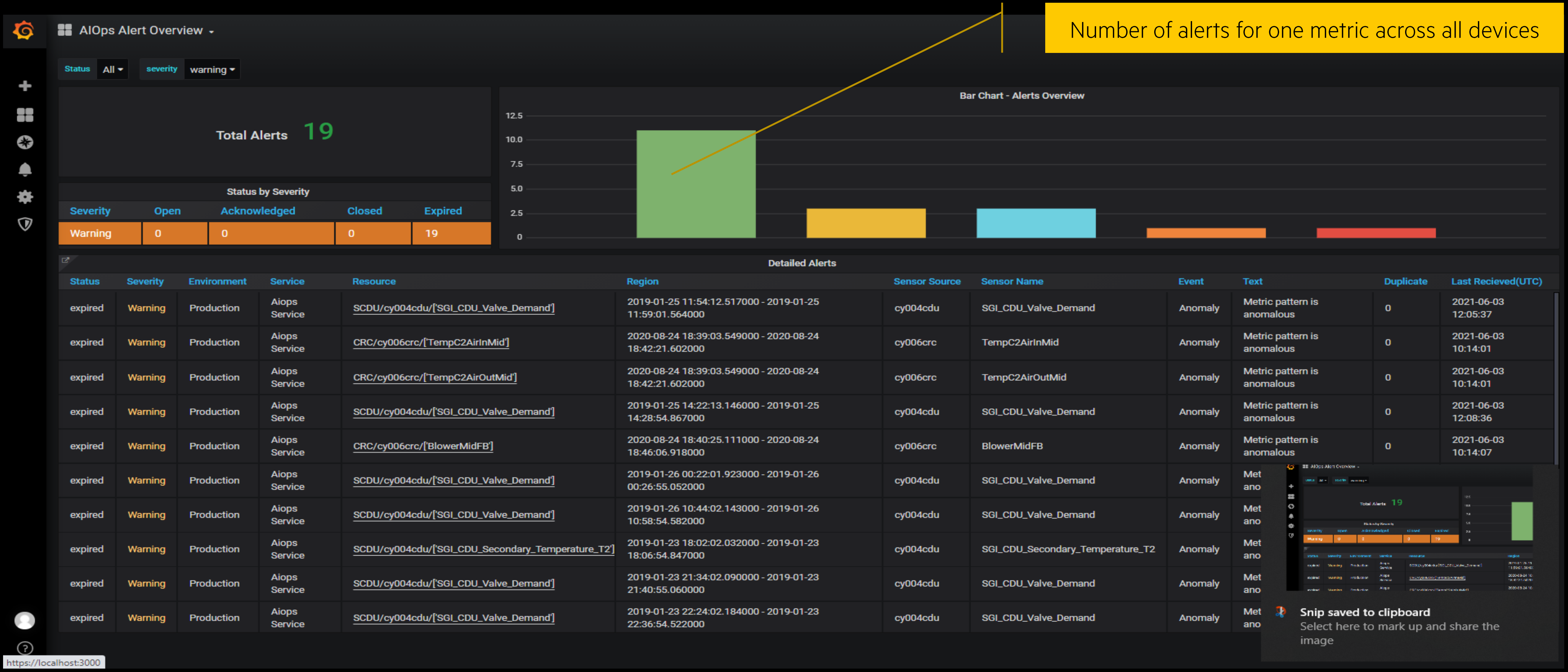


Device: cdu
Control Name: cdu4
Metric: ValveDemand

	min	max
🟦 p0.0 metric_value	4	87
🟥 p0.0 anomaly_score (right-y)	0	128
🟨 p0.0 anomaly_threshold (right-y)	56	56

AIOPS DASHBOARDS

AIOps alert overview



Number of alerts for one metric across all devices

Total Alerts 19

Status by Severity

Severity	Open	Acknowledged	Closed	Expired
Warning	0	0	0	19

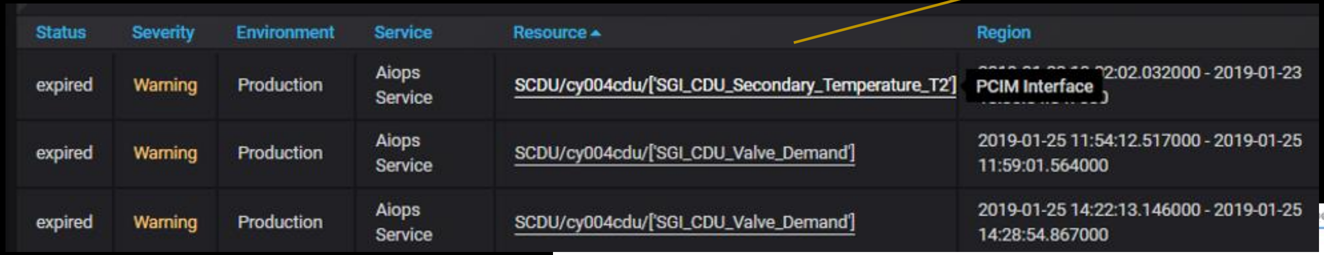
Detailed Alerts

Status	Severity	Environment	Service	Resource	Region	Sensor Source	Sensor Name	Event	Text	Duplicate	Last Recieved(UTC)
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-25 11:54:12.517000 - 2019-01-25 11:59:01.564000	cy004cdu	SGI_CDU_Valve_Demand	Anomaly	Metric pattern is anomalous	0	2021-06-03 12:05:37
expired	Warning	Production	Aiops Service	CRC/cy006crc/['TempC2AirInMid']	2020-08-24 18:39:03.549000 - 2020-08-24 18:42:21.602000	cy006crc	TempC2AirInMid	Anomaly	Metric pattern is anomalous	0	2021-06-03 10:14:01
expired	Warning	Production	Aiops Service	CRC/cy006crc/['TempC2AirOutMid']	2020-08-24 18:39:03.549000 - 2020-08-24 18:42:21.602000	cy006crc	TempC2AirOutMid	Anomaly	Metric pattern is anomalous	0	2021-06-03 10:14:01
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-25 14:22:13.146000 - 2019-01-25 14:28:54.867000	cy004cdu	SGI_CDU_Valve_Demand	Anomaly	Metric pattern is anomalous	0	2021-06-03 12:08:36
expired	Warning	Production	Aiops Service	CRC/cy006crc/['BlowerMidFB']	2020-08-24 18:40:25.111000 - 2020-08-24 18:46:06.918000	cy006crc	BlowerMidFB	Anomaly	Metric pattern is anomalous	0	2021-06-03 10:14:07
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-26 00:22:01.923000 - 2019-01-26 00:26:55.052000	cy004cdu	SGI_CDU_Valve_Demand	Anomaly	Met ano		
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-26 10:44:02.143000 - 2019-01-26 10:58:54.582000	cy004cdu	SGI_CDU_Valve_Demand	Anomaly	Met ano		
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Secondary_Temperature_T2']	2019-01-23 18:02:02.032000 - 2019-01-23 18:06:54.847000	cy004cdu	SGI_CDU_Secondary_Temperature_T2	Anomaly	Met ano		
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-23 21:34:02.090000 - 2019-01-23 21:40:55.060000	cy004cdu	SGI_CDU_Valve_Demand	Anomaly	Met ano		
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-23 22:24:02.184000 - 2019-01-23 22:36:54.522000	cy004cdu	SGI_CDU_Valve_Demand	Anomaly	Met ano		

Snip saved to clipboard
Select here to mark up and share the image

AIOPS DASHBOARDS

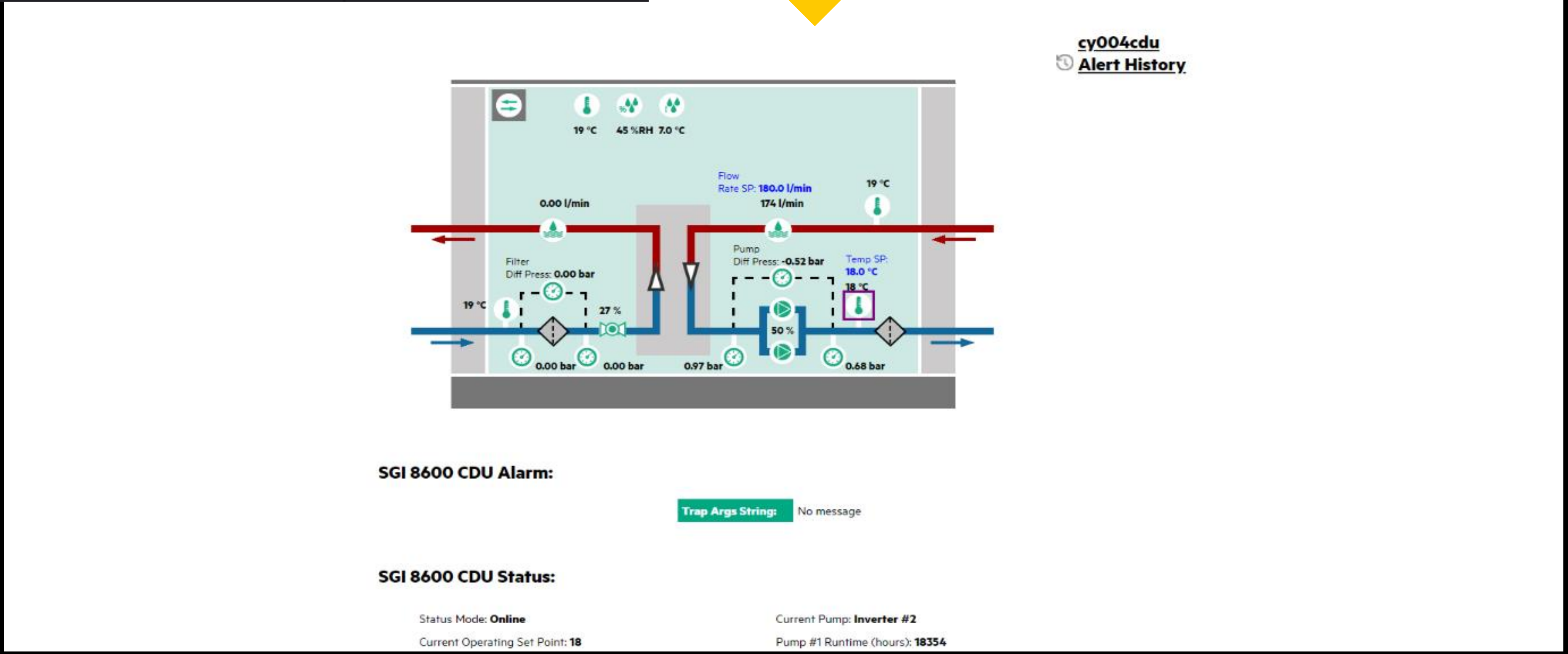
Connection with HPCM PCIM visualization



Status	Severity	Environment	Service	Resource ▲	Region
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Secondary_Temperature_T2']	[illegible]2:02.032000 - 2019-01-23 [illegible]
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-25 11:54:12.517000 - 2019-01-25 11:59:01.564000
expired	Warning	Production	Aiops Service	SCDU/cy004cdu/['SGI_CDU_Valve_Demand']	2019-01-25 14:22:13.146000 - 2019-01-25 14:28:54.867000

Click metric source to open PCIM interface for this device. This metric will be shown inside a pink rectangle.

PCIM – Power and Cooling Infrastructure Manager



SGI 8600 CDU Alarm:

Trap Args String: No message

SGI 8600 CDU Status:

Status Mode: **Online**

Current Operating Set Point: **18**

Current Pump: **Inverter #2**

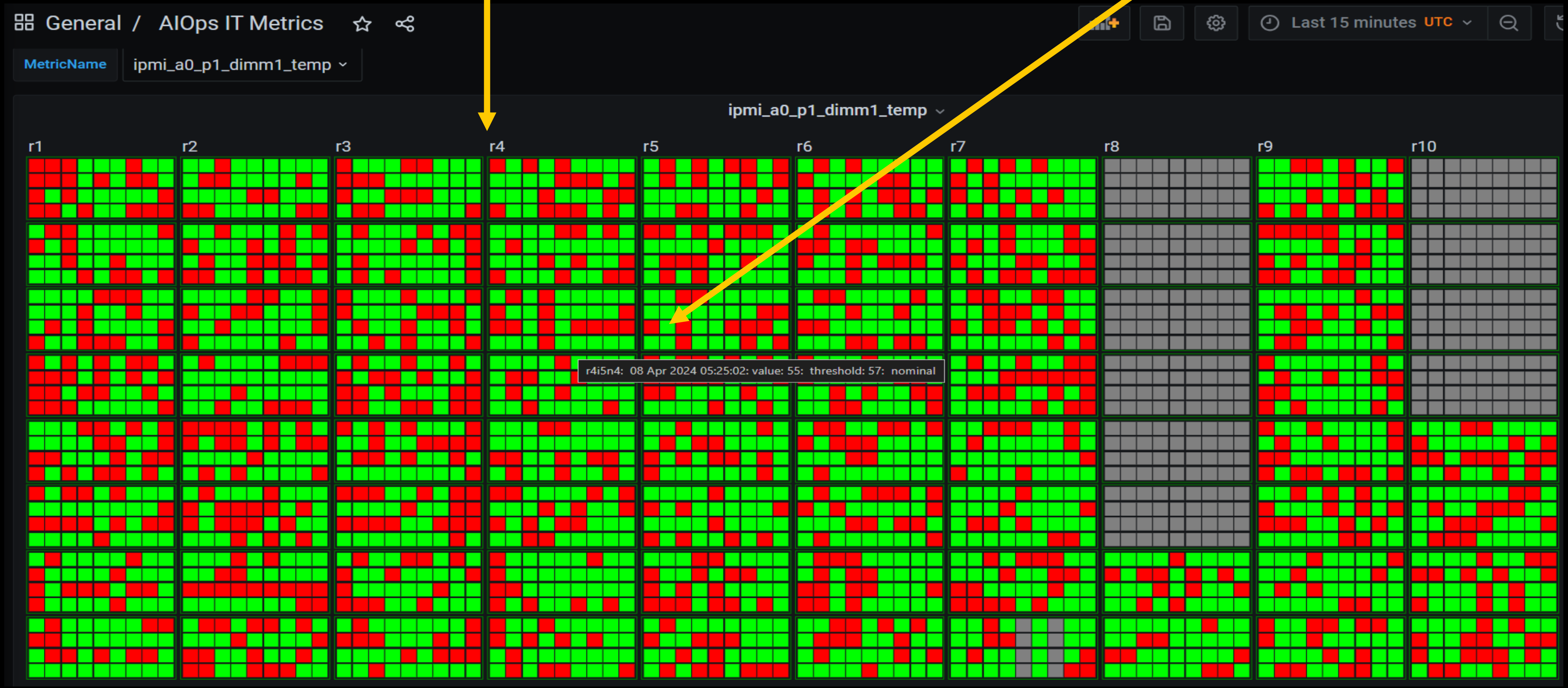
Pump #1 Runtime (hours): **18354**

AIOPS DASHBOARDS

Visualizing IT metrics

compute nodes are displayed in the rack orientation

Green is nominal , red represent anomalous



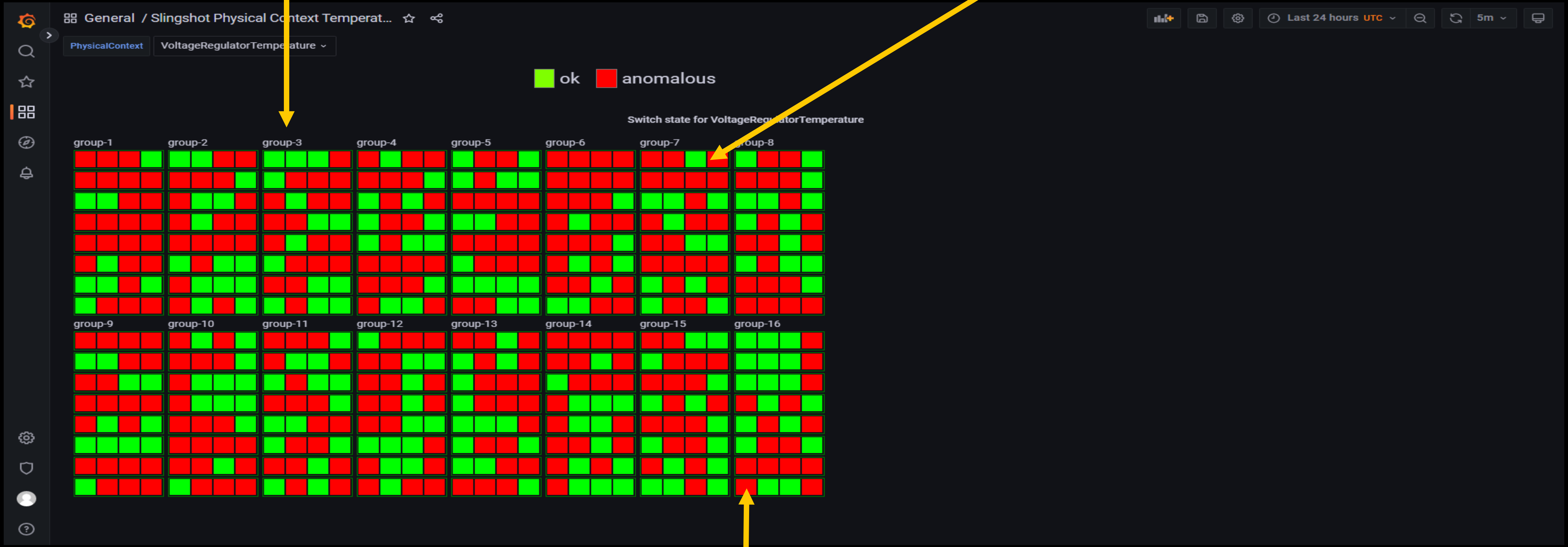
r4i5n4: 08 Apr 2024 05:25:02: value: 55: threshold: 57: nominal

AIOPS DASHBOARDS

Visualizing Slingshot metrics

Switches are displayed in the group orientation

Green is nominal , red represent anomalous



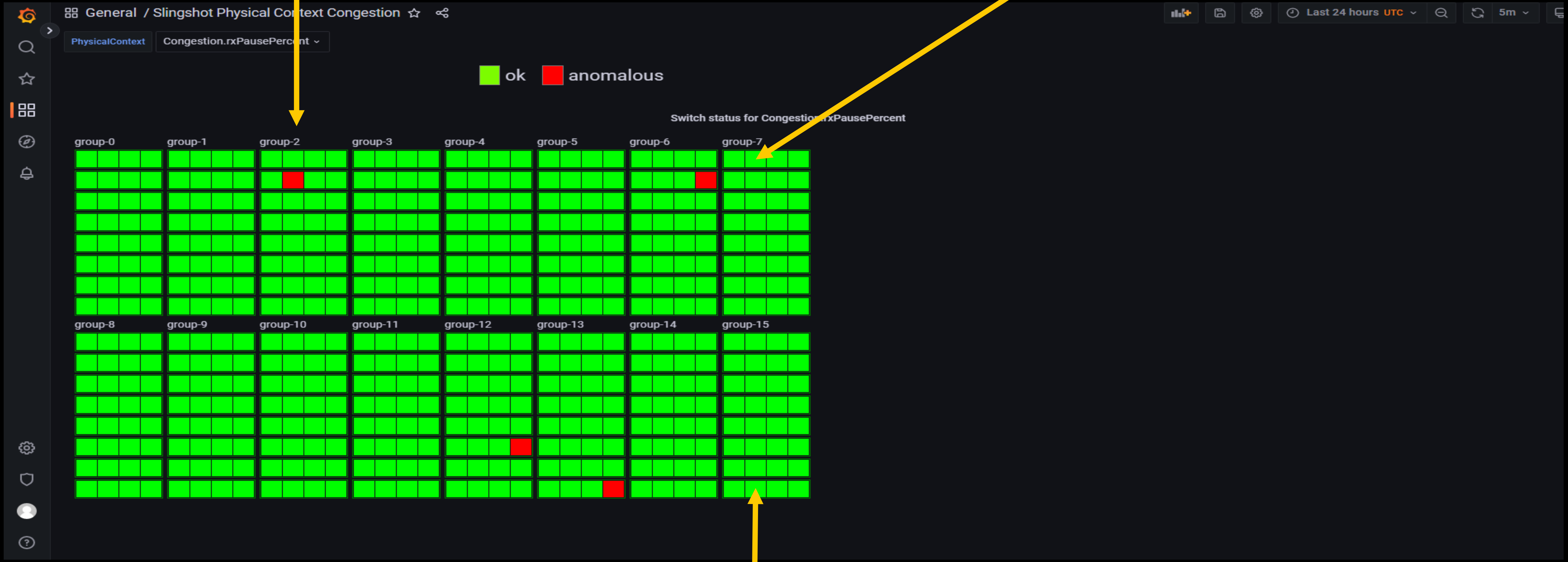
Click on each switch ,display detail dashboard

AIOPS DASHBOARDS

Visualizing Slingshot metrics

Switches are displayed in the group orientation

Green is nominal , red represent anomalous



General / Slingshot Physical Context Congestion

PhysicalContext | Congestion.rxPausePercent

Last 24 hours UTC | 5m

🟩 ok 🟥 anomalous

Switch status for Congestion.rxPausePercent

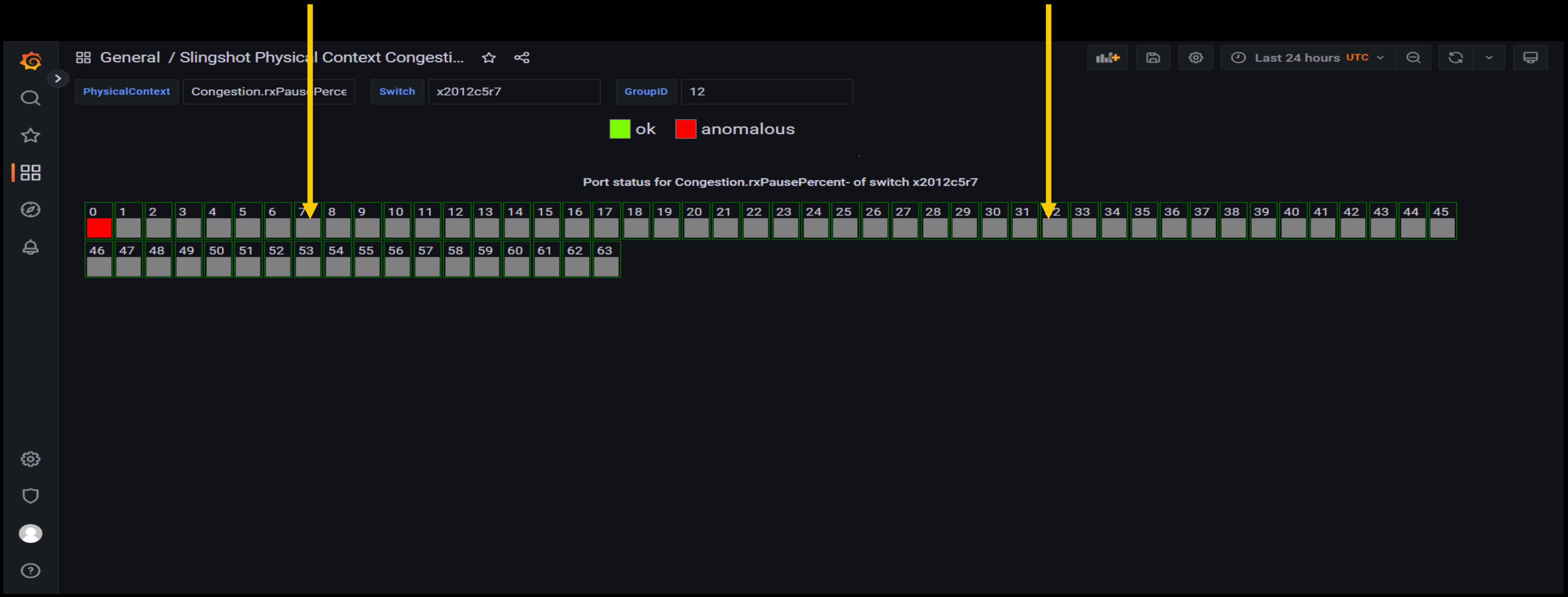
Click on each switch ,display detail dashboard

AIOPS DASHBOARDS

Visualizing Slingshot metrics

details of all ports inside the switch.

Green is nominal , red represent anomalous



General / Slingshot Physical Context Congesti...

PhysicalContext: Congestion.rxPausePerce

Switch: x2012c5r7

GroupID: 12

🟩 ok 🟥 anomalous

Port status for Congestion.rxPausePercent- of switch x2012c5r7

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

STATUS AND FUTURE WORK

CURRENT STATUS

- AIOps has been deployed in NREL's production environment since June 2020.
- AIOps has detected 50% more hardware-related anomalies (using historical data).
- AIOps has shown that 40% more high priority incidents that turned into critical events could have been prevented by demonstrated early anomaly detection.
- AIOps supports MLOps, Automation of Model training/creation and reload, model re-training and performance monitoring.
- AIOps supports anomaly detection for
  - Facility metrics CDU (cooling distribution unit), CRC (cooling rack controller), ARCS (Adaptive Rack Cooling System)
  - IT telemetry metrics , with new Grafana visualization panel
  - Slingshot telemetry metrics , with new Grafana visualization panel
- AIOps supports metric forecast and failure prediction for CDU metrics

- AIOps is integrated with HPCM 1.11 and CSM 1.4
- AIOps Integration with GreenLake : A proof of concept (POC) has already been conducted in collaboration with the Green Lake COM team for the purpose of anomaly detection and the reduction of carbon emissions by optimizing the data center ops.

FUTURE WORK – 2024

- Multi-variate failure prediction
- Platform independent solution that can be easily leveraged across HPE infrastructure portfolio (GreenLake).
- Analyzing and predicting job failures.
- Log analytics : Anomaly Detection and Diagnosis from System Logs through Deep Learning.

THANK YOU

Deepak Nanjundaiah, HPC System Software, nanjundaiah@hpe.com
Sergey Serebryakov, HPE Labs, sergey.serebryakov@hpe.com
Subrahmanya Joshi, HPC System Software, subrahmanya-vinayak-joshi@hpe.com
Amarnath Chilumukuru, HPC System Software, amarnath.c@hpe.com
Piyali Hazra, HPC System Software, piyali.hazra@hpe.com