

Datacenter network switch down S513-C-IP163 | OTG0149898

Daniele POMPONI (IT-CS)

ASDF - July 18th, 2024

Agenda

- Summary and impact of the incident
- Trigger and root cause
- Detection
- Steps taken to diagnose, assess, and resolve
- Timeline and communication
- What went wrong and where we got lucky
- Follow-up actions

Summary and impact of the incident

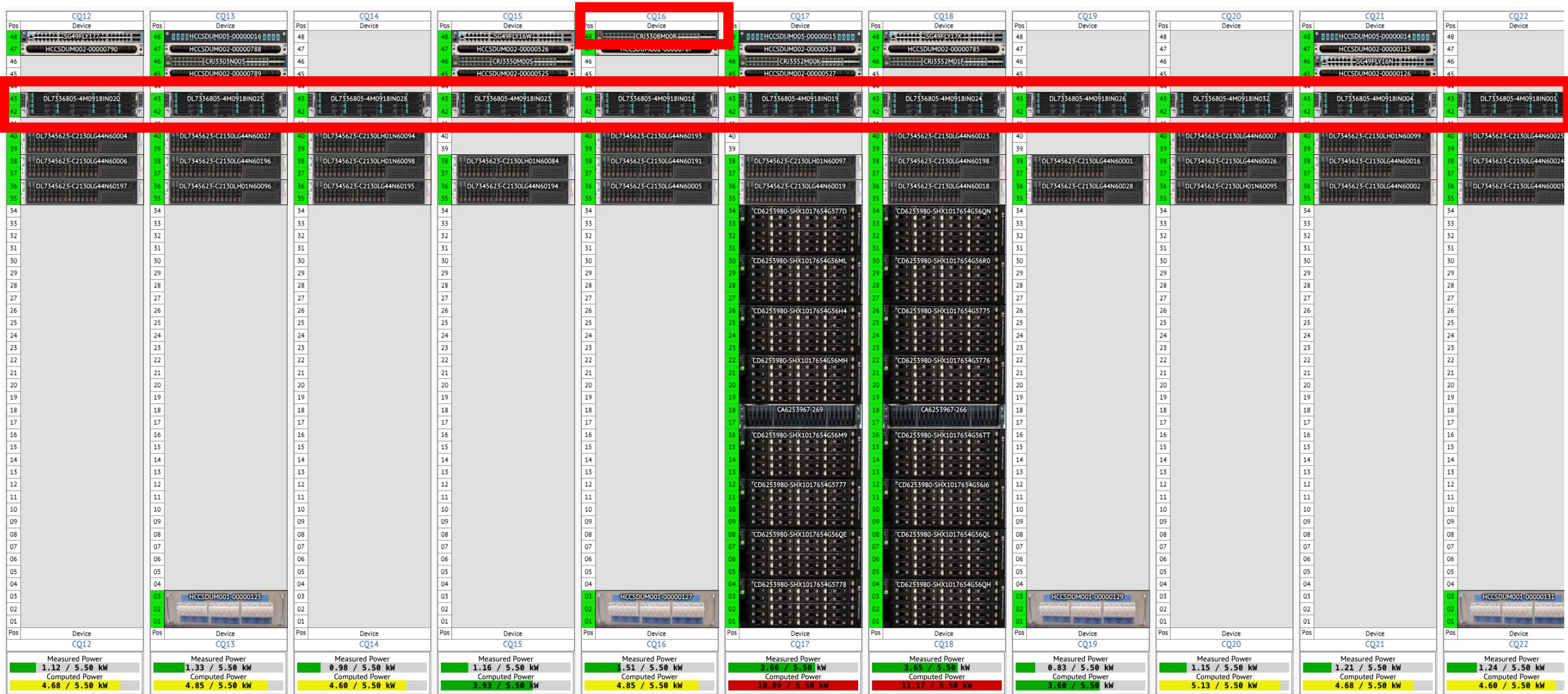
The datacenter network switch [N513-C-IP163-LBR7T-10](#) providing the service [S513-C-IP163](#) went down. The service was restored by replacing the faulty switch.

IMPACT:

- 11 quads (44 hypervisors) connected to this switch were offline during the outage.
- The Windows Terminal Servers infrastructure was degraded ([OTG0149894](#)), while 9 DBOD instances were down ([OTG0150013](#)).
- Due to the DBOD instances down, other services were impacted such as the SSO ([OTG0149903](#)) and CERN Library Catalogue web site ([OTG0149897](#)).

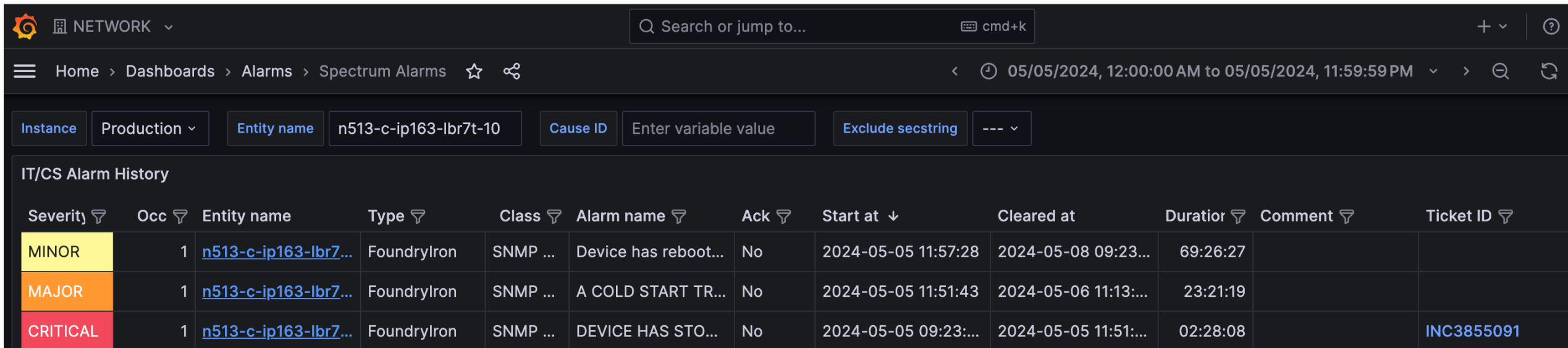
Trigger and root cause

Faulty network switch hosted in the Mainroom rack CQ16



Detection

- ✓ The network monitoring infrastructure (CA Spectrum) generated a CRITICAL alarm on the console.
- ✓ A GNI ticket (INC3855091) was generated and automatically routed towards the TI operators FE.



The screenshot shows a network monitoring interface with a search bar and navigation menu. The main content area displays a table of IT/CS Alarm History. The table has columns for Severity, Occ, Entity name, Type, Class, Alarm name, Ack, Start at, Cleared at, Duration, Comment, and Ticket ID. Three rows are visible, with the last row being CRITICAL and having a Ticket ID of INC3855091.

| Severity | Occ | Entity name | Type | Class | Alarm name | Ack | Start at | Cleared at | Duration | Comment | Ticket ID |
|----------|-----|--------------------------------------|-------------|----------|----------------------|-----|----------------------|----------------------|----------|---------|------------|
| MINOR | 1 | n513-c-ip163-lbr7... | FoundryIron | SNMP ... | Device has reboot... | No | 2024-05-05 11:57:28 | 2024-05-08 09:23... | 69:26:27 | | |
| MAJOR | 1 | n513-c-ip163-lbr7... | FoundryIron | SNMP ... | A COLD START TR... | No | 2024-05-05 11:51:43 | 2024-05-06 11:13:... | 23:21:19 | | |
| CRITICAL | 1 | n513-c-ip163-lbr7... | FoundryIron | SNMP ... | DEVICE HAS STO... | No | 2024-05-05 09:23:... | 2024-05-05 11:51:... | 02:28:08 | | INC3855091 |

Steps taken to diagnose, assess, and resolve

- ✓ The operator applied the corresponding procedure ([EDMS2955129](#)), and he escalated the ticket to the network piquet ([KB0007785](#)).
- ✓ The technician took in progress the incident within the SLA, and he reached the site.
- ✓ The technician first attempted to recover the service by doing a soft/hard reset.
- ✓ This didn't help; therefore, he applied the procedure to replace the switch.
- ✓ The service was restored following the replacement of the faulty switch.

Timeline and communication

- **09h23** | A critical alarm appeared on the Spectrum Console
- **09h33** | INC3855091 was generated
- **09h41** | The incident was taken in progress by the TI operator
- **09h48** | The incident was escalated to the network piquet
- **09h55** | The incident was taken in progress by the network piquet
- **10h43** | The incident was updated by the network technician to inform the persons in the loop that the soft reboot was not enough to re-establish the service
- **10h54** | OTG0149898 was published by the TI operator
- **10h56** | The incident was updated by the network technician to inform the persons in the loop that the hard reset didn't solve the problem either.
- **12:01** | The incident was updated (and closed) by the network technician to inform the persons in the loop that following the replacement of the faulty switch, the service was back online.

What went wrong

- ❖ From the network infrastructure point of view, nothing went wrong except the unpredictable hardware failure.
- ❖ The OTG should have been published earlier.

Where we got lucky

- ✓ The incident was resolved 1 hour before the SLA.
- ✓ This switch went down because faulty and not because of an electrical problem in the rack.

Reminder about today's SLA:

- Data switches always have 2 PSUs but **not** always connected to two different PDUs.
- Network technician piquet is available 24/7 (SLA 4 hours).
- **Since we stopped the CC operator service, broken PDUs/FUSE are not replaced outside working hours.**

A network switch down because of a power issue in the rack may remain down until the next working day.

We strongly recommend that the service managers take today's SLA into account also for the service's dependencies.

Followup actions

- ❖ **[COMPLETED]** Start the RMA process to get a replacement unit from the vendor.
- ❖ **[IN PROGRESS]** RQF2665813 – Create a new SNOW record producer dedicated for network incident to be used by TI to publish OTGs. Once this is completed, update the KBs accordingly.
 - ❖ The new RP is in production (thanks IT-TD-SM). The documentation is in the process of being updated.



home.cern