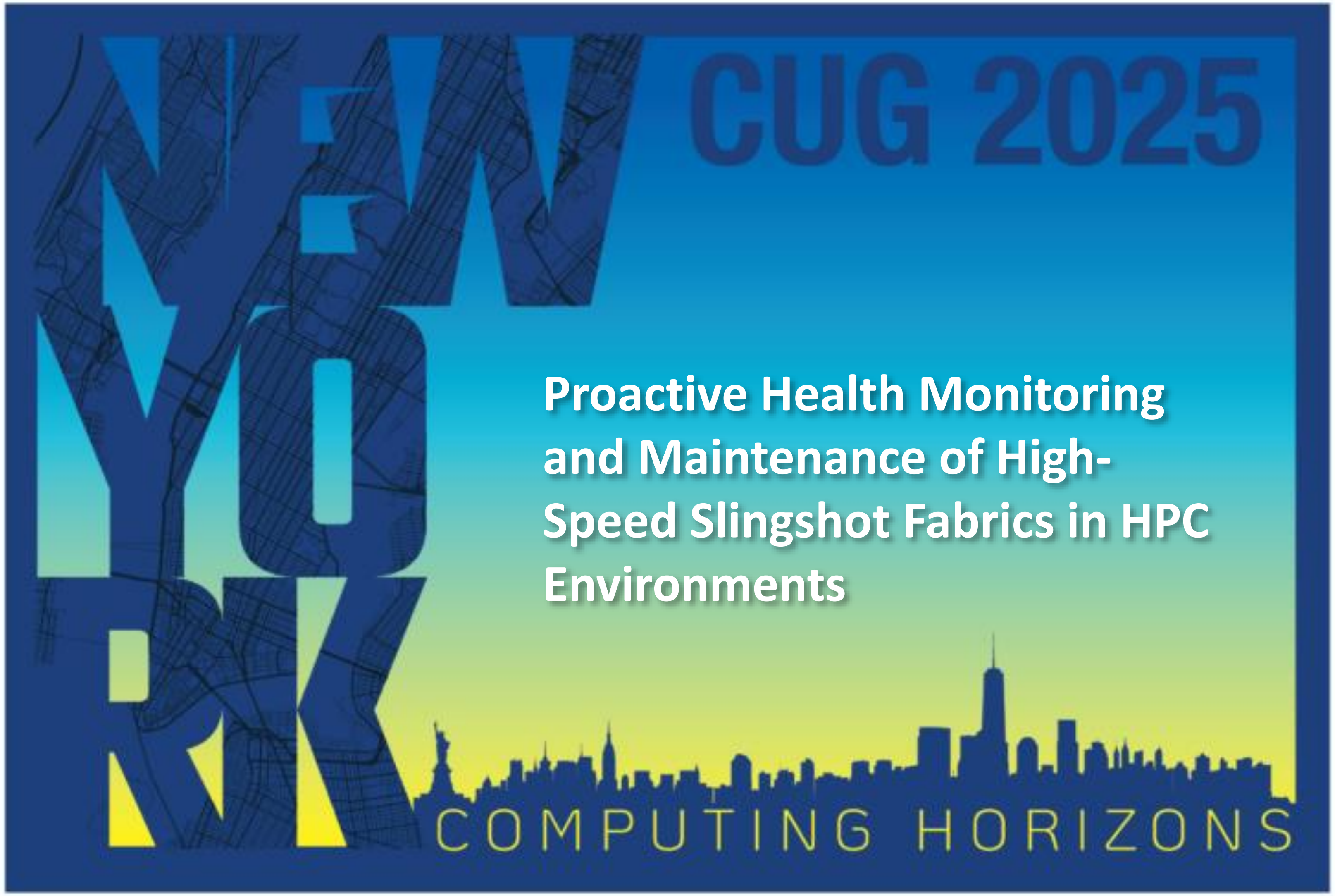


Proactive Health Monitoring and Maintenance of High-Speed Slingshot Fabrics in HPC Environments

Slingshot Case #1

Issue Summary

Jobs are failing and seeing CXI Errors being reported by nodes

Description

When running MPI jobs, end users are reporting job failures. Also seeing the following error messages being reported by the nodes during job execution:

```
[Tue Feb 18 20:48:41 2025] cxi_ss1 0000:21:00.0: cxi0[hsn0]: C_EC_CRIT: C1_HNI error: pfc_fifo_oflw
  (46) (was first error at 1739908121:067337283)

[Tue Feb 18 20:48:41 2025] cxi_ss1 0000:21:00.0: cxi0[hsn0]: C_EC_CRIT: pfc_fifo_oflw_cntr: 213

[Tue Feb 18 20:48:41 2025] cxi_ss1 0000:21:00.0: cxi0[hsn0]: C_EC_DEGRD_NS: C_IXE error: pbuf_rd_err
  (48) (was first error at 1739908121:067509999)

[Tue Feb 18 20:48:41 2025] cxi_ss1 0000:21:00.0: cxi0[hsn0]: C_EC_DEGRD_NS: pbuf_rd_errors: 14
```

Slingshot Case #2

Issue Summary
A New Issue was found with Slingshot Host Software v12.0.0, USS v1.3 on Blanca Peak blades

Description
New settings introduced in USS v1.3 for IOMMU.passthrough can cause nodes running Slingshot Host Software v12.0.0 to crash when running LNET.

Slingshot Case #3

Issue Summary

After upgrading to Slingshot v2.3.0 jobs are starting to fail when they used to work fine

Description

Recently upgraded the system to Slingshot v2.3.0. Everything appears to have come up normally but now when we try to run jobs, we are seeing a number of packet drops being reported and jobs are failing.

Approach

- **This talk will focus on Slingshot, but the concepts here apply to any technology**
- **There are (2) paths to address problems**
  - **Reactive**
    - Typically, these are unknown problems (never experienced before)
    - Can be known issues, but not proactively looking for them
    - Zero-day software/hardware/environment issues
    - Can be a long-drawn-out process getting to root cause
  - **Proactive**
    - Known, there is an established fingerprint
- **Our goal**:
  - Learn from reactive (all) problems to facilitate proactive checks
  - Integrate learnings into on-system-provided tools
  - Build proactive checks into HPE backend to help customers identify issues
    - Compliments Customer Advisories and the use of local tooling to identify issues

How we get there

Both reactive and proactive solutions start with system state and telemetry data being available to HPE

- Slingshot built-in tooling, used by the site admins
  - Ongoing monitoring of the output from tools such as `fmn-show-status`, Redfish alerts, telemetry
  - Reactive debug checking of tools such as `linkdbg`, `show-flaps` and `fmn-check-fabric`
  - Leveraging SMS and/or AIOps to provide real-time alerting and reporting of fabric and system status
- Automate the capture of this data capture over time
  - Capture stateful data from Slingshot built-in tooling
  - Capture logs before they roll off
  - Data Securely uploaded to HPE to get this feedback loop going
- Reactive learning across customers
  - Identify fingerprints of known issues based upon the uploaded data and resolution of cases
  - Data feeds identification of new proactive checks, new customer advisories, tooling, and knowledge articles
  - Scope and scale provides valuable feedback to R&D for future product enhancements

How we get there: Secure Upload

- **Deploy HPE’s System Diagnostic Utility (SDU) Toolkit**
  - Embedded in CSM/HPCM/Slingshot and ClusterStor/Neo 7.x
  - Includes **SDU CLI**
    - Collects system state and telemetry data using extensible plugins
    - Forms a support bundle of files referred to as a collection
  - Includes **HPE RDA** for secure transport
    - HPE's de facto solution for secure upload
  - Makes use of **Metis**
    - Backend system for receiving SDU data from RDA
    - Provides hooks for SFDC integration
    - Allows for proactive checks, reporting and data analysis
- Check out the Knowledge Article: Getting Started with HPC Cray System Diagnostic Utility (SDU)

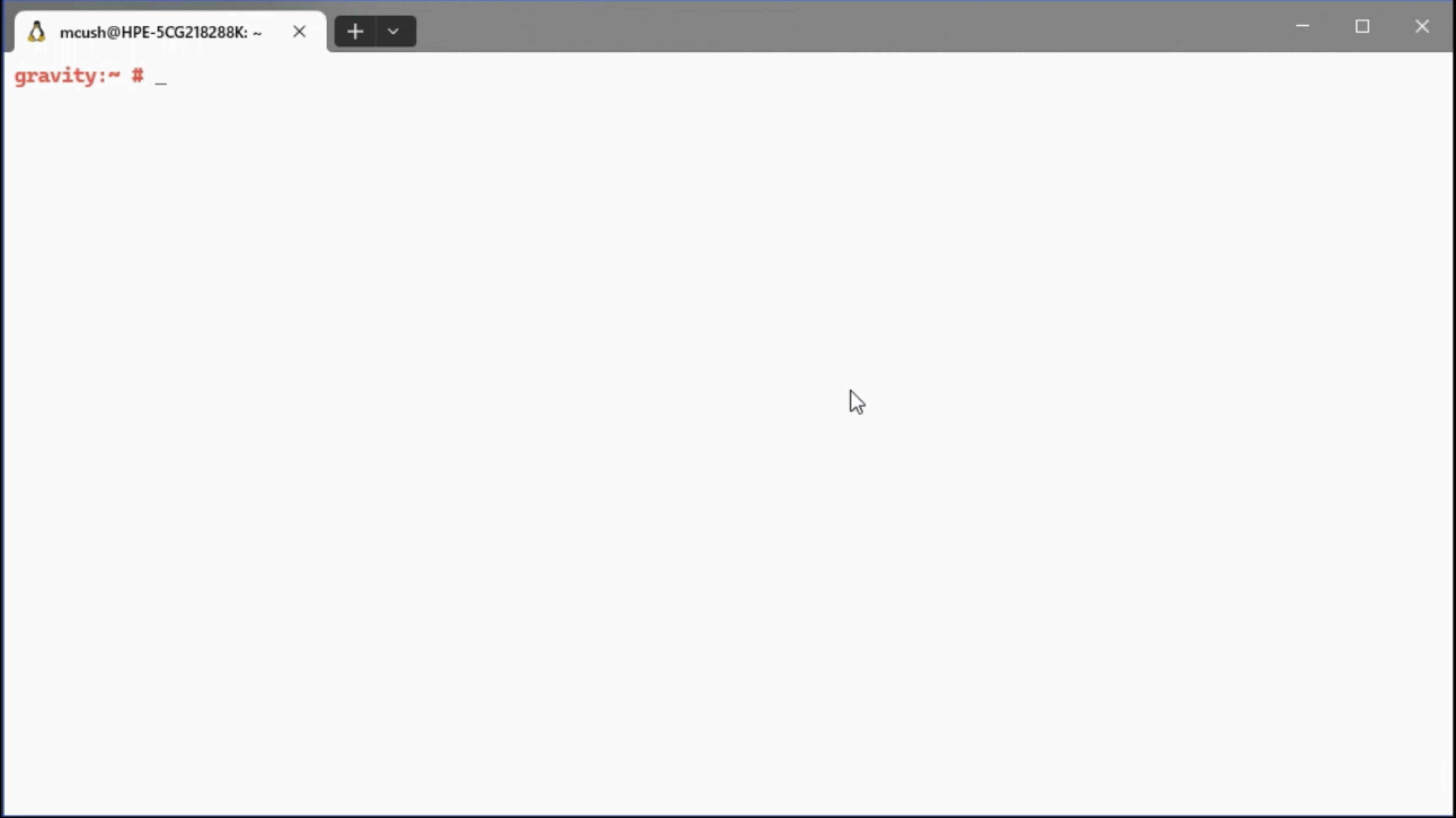
Metis: In ancient Greek religion and mythology, was the pre-Olympian goddess of **wisdom, counsel and deep thought**, and a member of the Oceanids.

Additional Efficiencies with Secure Upload

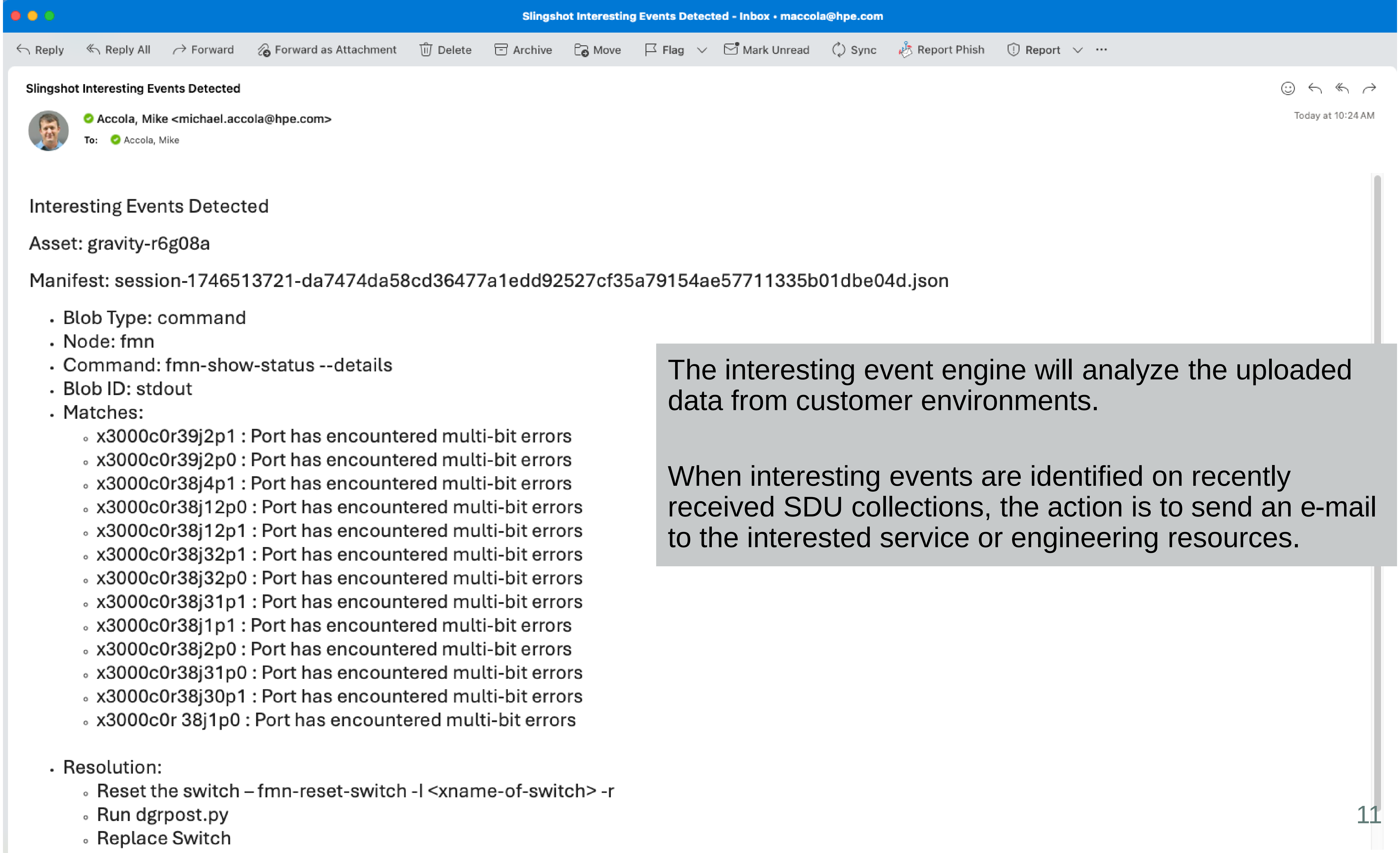
- **Proactive efficiency – Phase I: Daily Scenario**
  - Sends Asset Facts to service agents – attached to the asset in HPE's SFDC (our view of your assets in HPESC/DCE)
  - Internal cron will automatically send
    - Health information
    - Firmware versions
    - Software versions
- **Reactive efficiency**
  - **Triage scenario**
    - Standard SDU method to capture most of the data needed by HPE service engineers and R&D
  - **Adhoc scenario**
    - Allows users to very easily upload any requested service telemetry
      - Copy or sym link objects of interest (command output, files, etc.) to the adhoc directory
      - `# sdu scenario adhoc --ref=sfdc:123456789`
      - Avoid the pain and time of shuffling support data around and uploading to SFDC or HPRC manually
      - HPE service agent will be informed when the data has been successfully uploaded
  - **Case creation** (**coming soon – currently in testing**)
    - This will create an SFDC/HPESC case and inform the case team
    - `# sdu scenario adhoc --ref=sfdc:create`

System Diagnostic Utility (SDU) Overview

- System Diagnostic Utility (SDU) included in CSM 1.2+, HPCM 1.7+ and ClusterStor Neo 7+
- Collects fabric data via fabric manager
- Securely uploads to HPE via Remote Device Access (RDA)
  - RDA is the most secure transport in the industry
  - Utilized globally on hundreds of thousands of devices
  - See HPE Remote Device Access Security Technical Paper for details
- Standard data collection scenario or support-team directed queries
- HPE Supported Proactive Monitoring
  - Regular automated collection enable later proactive diagnosis as rules are built



```
gravity:~ # _
```



Slingshot Interesting Events Detected

Accola, Mike <michael.accola@hpe.com>

To: Accola, Mike

Today at 10:24 AM

Interesting Events Detected

Asset: gravity-r6g08a

Manifest: session-1746513721-da7474da58cd36477a1edd92527cf35a79154ae57711335b01dbe04d.json

- Blob Type: command
- Node: fmn
- Command: fmn-show-status --details
- Blob ID: stdout
- Matches:
  - x3000c0r39j2p1 : Port has encountered multi-bit errors
  - x3000c0r39j2p0 : Port has encountered multi-bit errors
  - x3000c0r38j4p1 : Port has encountered multi-bit errors
  - x3000c0r38j12p0 : Port has encountered multi-bit errors
  - x3000c0r38j12p1 : Port has encountered multi-bit errors
  - x3000c0r38j32p1 : Port has encountered multi-bit errors
  - x3000c0r38j32p0 : Port has encountered multi-bit errors
  - x3000c0r38j31p1 : Port has encountered multi-bit errors
  - x3000c0r38j1p1 : Port has encountered multi-bit errors
  - x3000c0r38j2p0 : Port has encountered multi-bit errors
  - x3000c0r38j31p0 : Port has encountered multi-bit errors
  - x3000c0r38j30p1 : Port has encountered multi-bit errors
  - x3000c0r 38j1p0 : Port has encountered multi-bit errors
- Resolution:
  - Reset the switch – fmn-reset-switch -l <xname-of-switch> -r
  - Run dgrpost.py
  - Replace Switch

The interesting event engine will analyze the uploaded data from customer environments.

When interesting events are identified on recently received SDU collections, the action is to send an e-mail to the interested service or engineering resources.

How we get there: Interesting Event Rules Engine

- Critical issue “fingerprint” identified, and an interesting event rule is created
- Incoming collections are scanned for a fingerprint match
- Interesting events engine informs engineers which customers/assets have a fingerprint match
- This allows engineers to proactively inform site teams/customers to avoid or quickly mitigate issue before it becomes critical
- Types of scans
  - Real time scan
    - Identify known issues proactively as new collections arrive
  - Historical scan
    - Research historical collections to learn about unique conditions that may be present to create proactive interesting events detection
    - Determine scale / scope of issues

Slingshot Case #1

Issue Summary
Jobs are failing and seeing CXI Errors being reported by nodes

How this would have been "Proactively" detected
As information is uploaded to the Metis database we parse through the output of the `fmn-show-status` command and see that the number of interfaces in the cassini-policy does not match the number of ports that are in the qos-ll_be_be_et-cassini-policy then we notify the system admins to validate that the proper set of port-policies are applied to the fabric.

Once these settings were applied properly, the errors on the system went away.

Slingshot Case #2

Issue Summary
A New Issue was found with Slingshot Host Software v12.0.0, USS v1.3 on Blanca Peak blades

How this would have been "Proactively" detected
Proactive parsing of the regularly collected data provides a list of all of the customers that currently have Blanca Peak blades installed in their system.

HPE Proactively reaches out to the customers potentially affected and warns them to stay on Slingshot Host Software v11.1.0 OR apply certain settings to avoid the issues that have been seen with the newer SHS release.

This also allows the Slingshot Host Software team to determine the impact if this issue based upon the number of customers with Blanca Peak vs other types of blades so the priority can be set appropriate to continue to release SHS v12.0.0 as it is or to stop the release of SHS v12.0.0 and wait for v12.0.1!

Having this data at hand also allows the Slingshot team to better know the adoption rates of different releases and focus on where testing and fixes may need to be ported for issues.

Slingshot Case #3

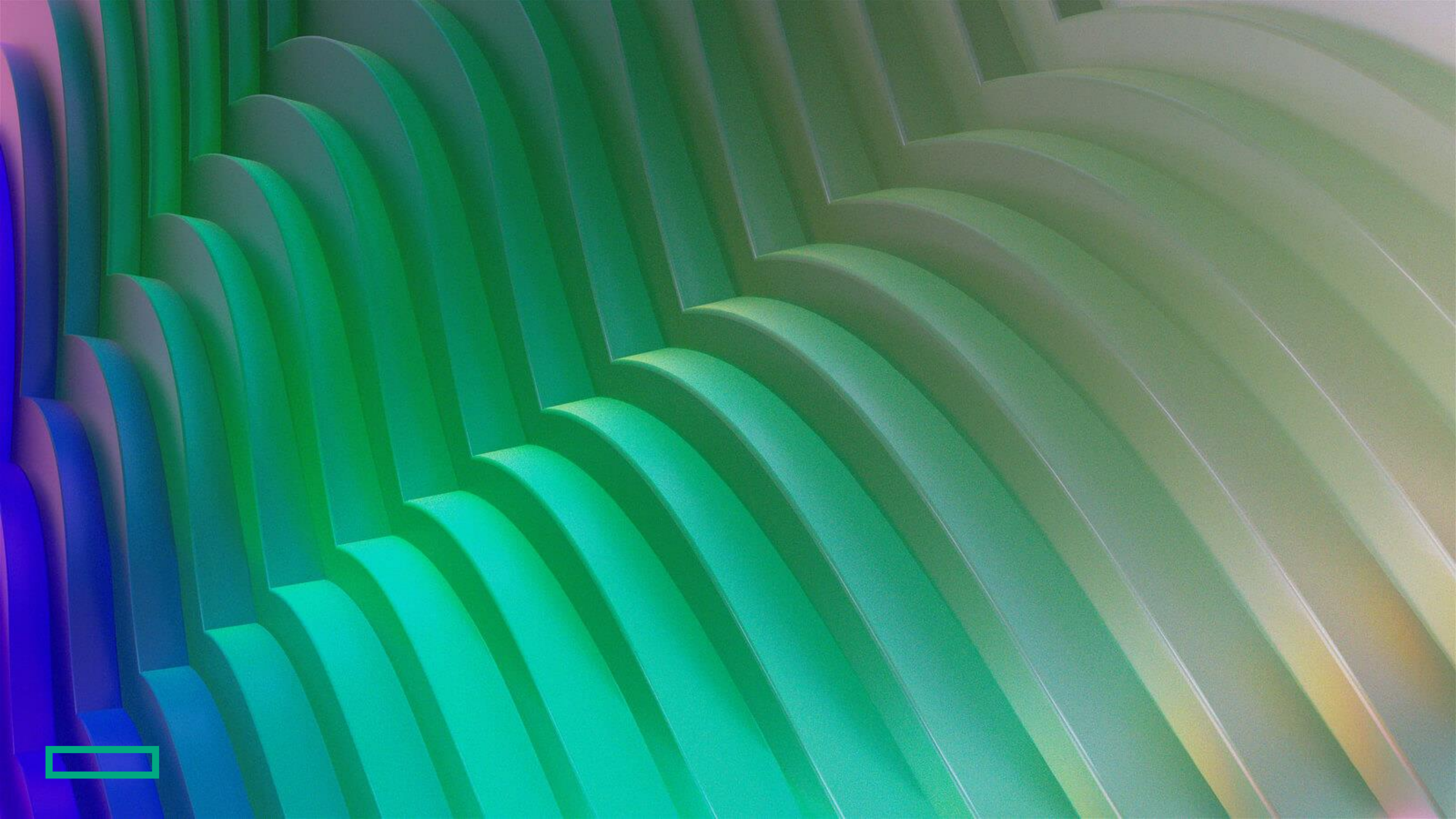
Issue Summary
After upgrading to Slingshot v2.3.0 jobs are starting to fail when they used to work fine

How this would have been "Proactively" detected
Regular updates are being captured and send to HPE via SDU. Part of those regular updates are a list of Switch ASIC errors that are being reported by the switches. This new issue was discovered in Slingshot v2.3.0 which uncovered a new type of Empty Route that can cause traffic to stall on the system after a switch has been reset under very specific conditions. The ASIC Error Output is parsed daily, and it is detected that this particular error is happening on a switch in the fabric. The Site Admin is contacted with this issue and a New Customer Advisory is referenced along with an updated `fmn-debug` RPM which has a tool to better detect this issue moving forward.

Once the updated fmn-debug rpm is installed and the script is in place, we can quickly find and recover from this issue until the next Slingshot Release resolves the issue.

Conclusion

- HPE needs your help to move from reactive to proactive support of Slingshot
  - More support telemetry from more systems enhances our view of Slingshot behavior across multiple environments
- Our L3 support and R&D teams are working together to build out interesting events rules to improve your support experience
- Part of the journey to proactive support is to improve the reactive support experience via the adhoc scenario, case reference on upload, and the coming case creation feature of SDU.
- Expected outcomes
  - Better (and Faster) case resolution
  - Introduction of new proactive capabilities to avoid cases altogether!
  - Reduce the chance of multi-customer failures
  - Driving prioritization of testing and fixes by knowing what our customers are running
  - Help systems be more stable and productive by catching issues BEFORE they cause problems



Security

Client

- The SDU Toolkit runs in an OCI container, which isolates it from the host
- SDU CLI is open, written in Python, easily reviewed
- The upload of support telemetry is 100% controlled by the customer
- The collections are written locally and can be inspected prior to upload
- The collection objects are hashed and verified in the collection manifest
- The standard scenarios do not collect PII or customer data

Transport

- RDA is the most secure transport in the industry
- Utilized globally on hundreds of thousands of devices
- See HPE Remote Device Access Security Technical Paper for details

Backend

- Each asset’s support telemetry is stored in its own container
- Only HPC service and R&D team members have access to the uploaded data

SDU One Slide

System Diagnostic Utility (SDU)

Method to Move Data Easily to HPE

- Helps facilitate the capture and secure movement of system diagnostic data to HPE
- Easy to use method to upload multiple data packages simultaneously
- Provides complete control to the customer allowing to Opt-In/Out of any and all data collection

Built-In to Every System

- SDU ships with both CSM and HPCM Software
- Easy to enable interface with complete access controls to allow/deny access

Fully Integrated into HPE Support and Systems

- Only HPC service and R&D team members have access to the uploaded data
- Data is used to create/update support cases
- Data is also used to update census data so HPE understands our customer's environments and can proactively prioritize and provide support moving forward.