Hewlett Packard
Enterprise

HTT – Hardware Triage Tool

Isa Wazirzada, Abhishek Mehta, Vinanti Phadke

HTT – Hardware Triage Tool Overview

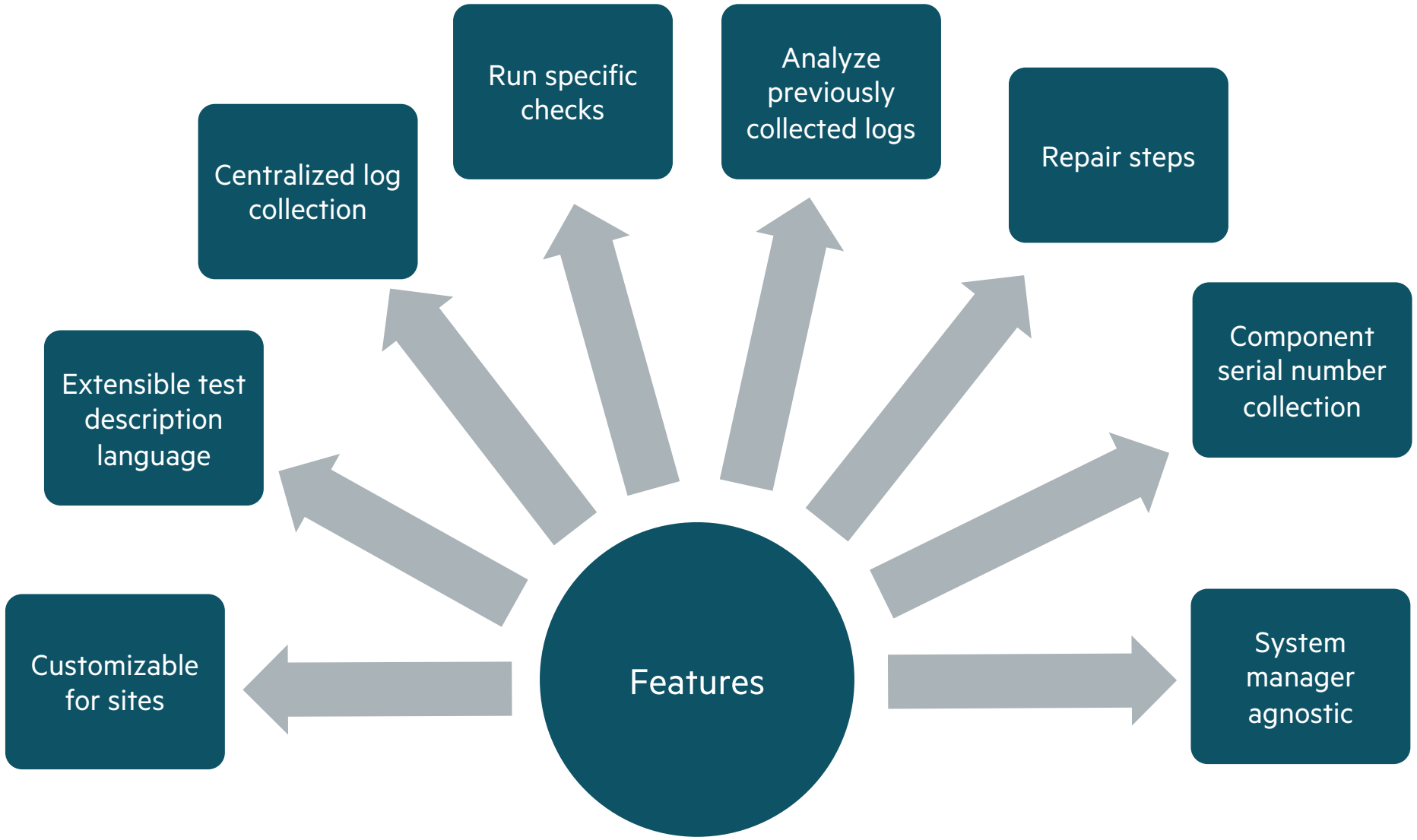
- Incorporates lessons learned from deployments across the world
- The hardware triage tool:
  - Checks for different failure signatures
  - Provides hardware actions and RMA codes (If applicable)
  - Builds a detailed support bundle even if it can't provide a diagnosis
- Current State
  - Can diagnose problems on several hardware programs
    - EX235a, EX255a, EX254n, EX4252, EX425, and the EX235n blades
  - Being utilized in EMEA, the Americas, and APAC
  - Product level solution
  - Tests for new hardware programs are being developed early in the product lifecycle

Hardware Triage Tool – What's in a name?

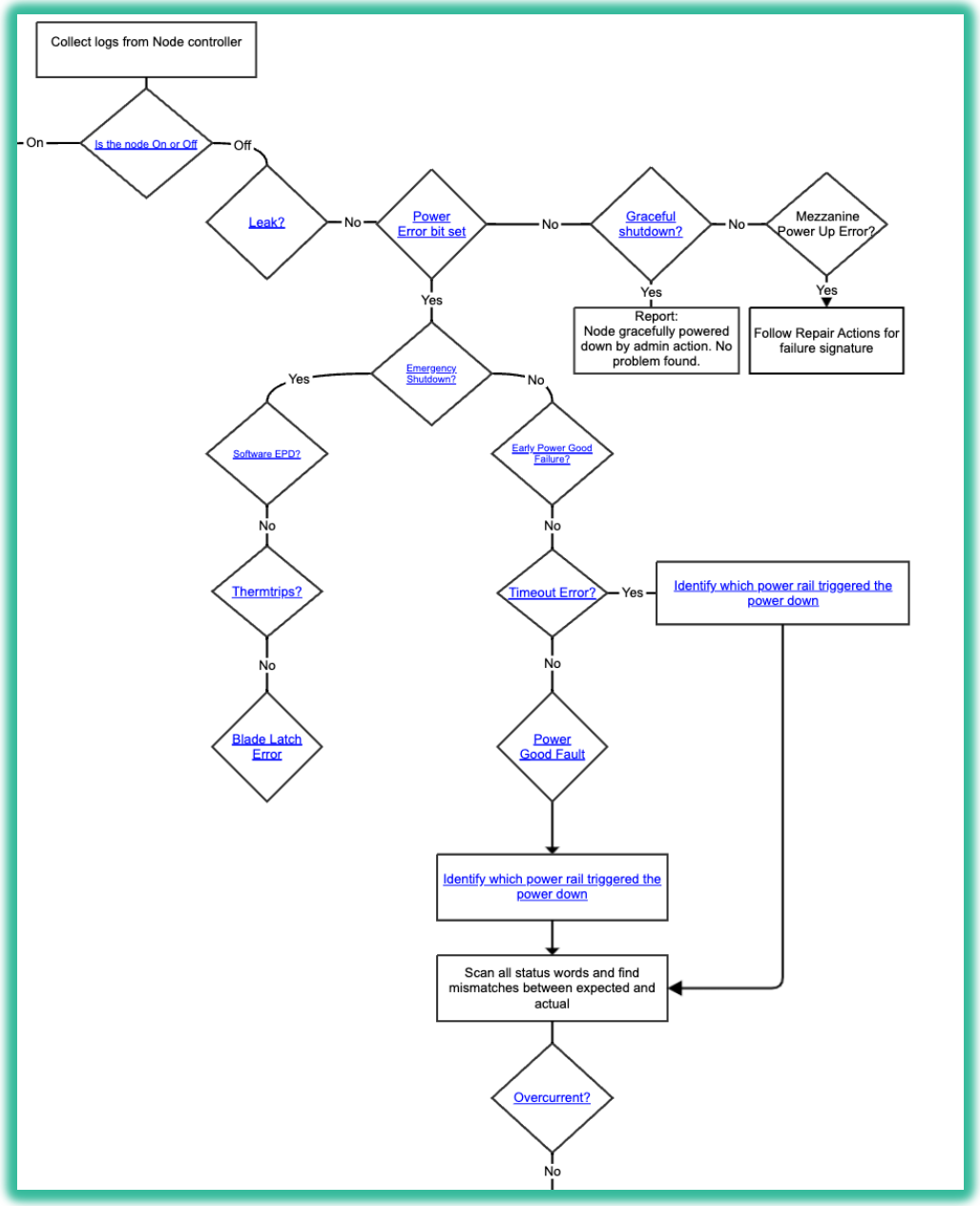
- What it is and it is not
- Administrators can run HTT to diagnose failures, for example if a compute node:

Fails to power on	Powers on but fails to boot (stuck at UEFI shell)	Boots but fails health checks
Unexpectedly reboots	Unexpectedly powers down (Emergency Power Down)	Becomes unresponsive during a job but didn't power off

Hardware Triage Tool Features



Defining Test Workflows



```
MezzError:
  custom_script: check_mezz_error.py
  custom_script_value_yes: "0"
  custom_script_args: log_path
  yes_condition:
    action: |
      - swap NMC0-NMC1
      - swap risers to see if it follows the mezz
  no_condition:
    action: None
    go_to: Dracut_shell
```

A Closer Look at the Test Description Language

```
Graceful Power Down:
  exec_statement: 'cat $logpath/power_down_req '
  exec_statement_value_yes: "1"
  yes_condition:
    action: Graceful shutdown happened
  no_condition:
    go_to: MezzError
```

```
Power CAP Check:
  input_json: powerfault_epd.Node$n.nfpga.json
  key1:
    input: '["Registers","R_NFPGA_GPNC_N$n_PWR_CSR_CAP","Val"]'
    value:
      value_no: "0x0"
  yes_condition:
    go_to: PowerError
  no_condition:
    go_to: Emergency Shutdown
```

```
Power CAP Check:
  input_json: powerfault_epd.Node$n.nfpga.json
  key1:
    input: '["Registers","R_NFPGA_BLPNC_PWR_CSR_CAP","Val"]'
    value:
      value_no: "0x0"
  yes_condition:
    go_to: PowerError
  no_condition:
    go_to: Emergency Shutdown
```

A Closer Look at the Test Description Language

```
cat powerfault_epd.Node0.nfpga.json | jq .Registers.R_NFPGA_BLPNC_PWR_CSR_CAP.Val
```

Defining Test Workflows

```
Power CAP Check:
  input_json: powerfault_epd.Node$n.nfpga.json
  key1:
    input: '["Registers","R_NFPGA_GPNC_N$n_PWR_CSR_CAP","Val"]'
    value:
      value_no: "0x0"
  yes_condition:
    go_to: PowerError
  no_condition:
    go_to: Emergency Shutdown
```

```
Emergency Shutdown:
  input_json: powerfault_epd.Node$n.nfpga.json
  key1:
    input: '["Registers","R_NFPGA_GPNC_N$n_PWR_CSR_CAP","Bits","emergency_shutdown"],["Registers","R_NFPGA_GPNC_N$n_PWR_CSR","Bits","emergency_shutdown"]'
    value:
      value_yes: "0x1"
  yes_condition:
    go_to: SoftwareEPD
  no_condition:
    go_to: EarlyPowerGoodFailure
```

Hardware Triage Tool – Command Line Arguments

- `/opt/clmgr/hardware-triage-tool/hwtriage -h`

```
-h, --help            show this help message and exit
-r, --revision        Show the revision and exit.
-n NODE_NAME, --node-name NODE_NAME
                      Enter the node name to perform the checks
-u USERNAME, --username USERNAME
                      Username to access node controller and the redfish
                      calls
-p PASSWORD, --password PASSWORD
                      Password to access node controller and redfish calls
-l LOGPATH, --logpath LOGPATH
                      Provide the full log path to perform the checks
-ns {On,Off}, --node-state {On,Off}
                      Provide the node power state
-hw {ex235a,ex255a,ex254n,ex4252,ex425,ex235n}, --hardware {ex235a,ex255a,ex254n,ex4252,ex425,ex235n}
                      Provide the node hardware type
-ls, --list-stages    To list stages in a yaml file
-bs BEGIN_STAGE, --begin-stage BEGIN_STAGE
                      Enter the stage name from where the check will start
-rs RUN_STAGE, --run-stage RUN_STAGE
                      To run only one stage from yaml file
-f INPUT_YAML, --input-yaml INPUT_YAML
                      To pass an input config yml file as input
-hy HARDWARE_YAML, --hardware-yaml HARDWARE_YAML
                      To pass a hardware config yml file as input
-sn, --show-serial-number
                      To display the serial number info with the triage
                      result
-sno, --serial-number-only
                      Collect the serial numbers into a file without
                      triaging
-k SSH_KEY, --ssh-key SSH_KEY
                      Ssh key to enable passwordless ssh
-t TIMEOUT, --timeout TIMEOUT
                      Timeout duration for collecting logs in seconds,
                      default=120
-v, --verbose         To have a verbose output
-cpath CUSTOM_LOG_PATH, --custom-log-path CUSTOM_LOG_PATH
                      Provide the custom log path to store the triage logs
                      in the case to override the default log path
```

Defining Hardware Configuration

- The hardware.yml file defines all supported hardware platforms
- Located at **`/opt/clmgr/hardware-triage-tool/hardware.yml`**

```
hardware_family:
  - name: "ex235a"
    attributes:
      hardware:
          workflow_on: "workflows/workflow_EX235a_on.yml"
          workflow_off: "workflows/workflow_EX235a_off.yml"
          number_of_nics: "4"
          firmware_version: "1.5.41-ESM"
          nic_speed: "BS_200G"
          pci_speed: "16.0 GT/s PCIe"
          pci_width: "16"
          link_speed: "16.0 GT/s PCIe"
          link width: "16"
          esm_link_speed: "25.0 GT/s"
          BIOSVER: "1.6.2"
          BIOSREV: "5.21"
          CPUONLINE: "0-127"
          mem_manufacture: "1"
          DIMM_sizes: "1"
          DIMM_speed: "1"
          DIMM_count: "8"
```

Examples

Hardware Triage Tool Usage

- Installation on the admin node (HPCM) or ncn-m001 (CSM)
- HTT is invoked via the `hwtriage` command
  - **`hwtriage -u root -p [redacted] -n x9000c3s5b0n1`**

```
Log collection completed


logging path :  /opt/clmgr/hardware-triage-
tool/logs/x9000c3s5b0n1.2023.09.19.13.23/x9000c3s5b0
 EX4252   Hardware is supported!
Triaging :x9000c3s5b0n1

Node is in Off state
Analysis file :  /opt/clmgr/hardware-triage-
tool/logs/x9000c3s5b0n1.2023.09.19.13.23/x9000c3s5b0/triage_output.json
Serial Numbers information :  /opt/clmgr/hardware-triage-
tool/logs/x9000c3s5b0n1.2023.09.19.13.23/x9000c3s5b0/serial_numbers.txt
Triaging :x9000c3s5b0n1 Stage analysis : PowerError Detected!
Stage analysis : Emergency Shutdown Detected!
Stage analysis : BladeLatch Detected!
Recommended  action :  First check to make sure latch is closed, if it isn't then close the latch, If
problem resurfaces over time, replace the latch
Triage completed!
```

SIVOC Temperature Fault Example

```
./hwtriage -n x1102c5s3b0n0 -u root -p REDACTED
Info: Nodename not in xname format, mapping xname
Log collection completed


logging path : /var/log/hardware-triage-tool/x1102c5s3b0n0_20240416_1608/x1102c5s3b0
ex254n Hardware is supported!
Triaging :x1102c5s3b0n0  ⁞

Node is in Off state
Triaging :x1102c5s3b0n0  ⁞Analysis file : /var/log/hardware-triage-tool/x1102c5s3b0n0_20240416_1608/x1102c5s3b0/triage_output.json
Serial Numbers information :/var/log/hardware-triage-tool/x1102c5s3b0n0_20240416_1608/x1102c5s3b0/serial_numbers.txt
Triaging :x1102c5s3b0n0  ⁞Stage analysis : Power CAP Check Detected!
Stage analysis : PowerError Detected!
Triaging :x1102c5s3b0n0  ⁞Error: Temperature fault/warning fault detected for SIVOC
Repair Actions:
 1. Check for coolant leaks.
  a. If leak found, replace cooling loop.
 2. Check that SIVOC cables are populated and properly mated at both ends of the cable:
  a. Control cable
  b. Input Power cable
  c. SIVOC Radsok seating with node card
 3. Swap the SIVOC with another functional node card (usually the partner node card):
  a. If the problem follows the SIVOC, replace the SIVOC.
  b. If the problem stays with the node card, swap the SIVOC control cable with the other node card.
  c. If the problem follows the SIVOC control cable, replace the control cable.
  d. If the problem stays with the node card, swap the SIVOC power cable with the other node card.
  e. If the problem follows the SIVOC power cable, replace the power cable.
  f. If the problem stays with the node card, replace the node card.
```

Missing DIMM Example

```
./hwtriage -n x9000c1s0b0n1 -u root -p REDACTED
logging path :  /var/log/hardware-triage-tool/x9000c1s0b0n1_20240502_1427/x9000c1s0b0
ex425 Hardware is supported!
Triaging :x9000c1s0b0n1    ⣿
Node is in On state
Node is booted
Triaging :x9000c1s0b0n1    ⣿Analysis file :  /var/log/hardware-triage-tool/x9000c1s0b0n1_20240502_1427/x9000c1s0b0/triage_output.json
Serial Numbers information :  /var/log/hardware-triage-tool/x9000c1s0b0n1_20240502_1427/x9000c1s0b0/serial_numbers.txt
=== Checking for hardware mismatch
Triaging :x9000c1s0b0n1    ⣿ DIAG_ERROR: BIOS version does not match expected version - Got '1.7.2' but expected one of [1.6.3]
Triaging :x9000c1s0b0n1    ⣿=== Checked ===
=== Checking for port and link failure on the node ===
Triaging :x9000c1s0b0n1    ⣿=== Checked ===
=== Checking for board calibration error ===
=== Checked ===
=== Performing amdgpu checks ===
=== Checked ===
Triaging :x9000c1s0b0n1    ⣿Stage analysis : CheckNodeHealth_Failure Detected!
Recommended  action :  ChecknodeHealth Failure Detected. Look into /var/log/hardware-triage-tool/x9000c1s0b0n1_20240502_1427/x9000c1s0b0/check_node_health for more details.
Triaging :x9000c1s0b0n1    ⣿Stage analysis : Unexpected_Booted Detected!
Triaging :x9000c1s0b0n1    ⣿DIMM 13 is missing
```

Log Collection Example

```
ncn-m001:/var/log/hardware-triage-tool/x1003c7s6b1n1_20240404_1357/x1003c7s6b1 # ls -ltr
total 1668
-rw-r--r-- 1 root root  173557 Apr  4 13:57 messages
-rw-r--r-- 1 root root    6766 Apr  4 13:57 powerfault_epd.Node1.nfpga.json
-rw-r--r-- 1 root root    7723 Apr  4 13:58 powerfault_epd.Node1.json
-rw-r--r-- 1 root root  221967 Apr  4 13:58 tlsproxy-current
-rw-r--r-- 1 root root 1266880 Apr  4 13:58 current
-rw-r--r-- 1 root root    5930 Apr  4 13:58 nodeeventd-current
-rw-r--r-- 1 root root    4096 Apr  4 13:58 power_down_req
-rw-r--r-- 1 root root    2583 Apr  4 13:58 serial_numbers.txt
-rw-r--r-- 1 root root    1073 Apr  4 13:58 triage_output.json
```

Serial Number Collection Example

```
Chassis Components information
Part: Enclosure, Part Number: 101920703.D, Serial Number: HA19310270
NMC - Part: Mezz0, Serial Number: EP22110746
NMC - Part: Mezz1, Serial Number: EP22110775

Node information
Node: Node1, Part Number: 101920703.D, Serial Number: HA19310270

Node1's Dimms information
Node: Node1, DIMM: DIMM 0 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE2AD
Node: Node1, DIMM: DIMM 1 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE04C
Node: Node1, DIMM: DIMM 2 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE062
Node: Node1, DIMM: DIMM 3 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE1E6
Node: Node1, DIMM: DIMM 4 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE2D8
Node: Node1, DIMM: DIMM 5 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE1C9
Node: Node1, DIMM: DIMM 6 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE208
Node: Node1, DIMM: DIMM 7 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE2ED
Node: Node1, DIMM: DIMM 8 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE044
Node: Node1, DIMM: DIMM 9 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE0A6
Node: Node1, DIMM: DIMM 10 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE291
Node: Node1, DIMM: DIMM 11 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE026
Node: Node1, DIMM: DIMM 12 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE23A
Node: Node1, DIMM: DIMM 13 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE01B
Node: Node1, DIMM: DIMM 14 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE1CF
Node: Node1, DIMM: DIMM 15 Part Number: M393A2K43DB3-CWE, Serial Number: 037EE269

Node1's Processors information
Node: Node1, Processor: CPU0 Part Number: N/A, Serial Number: 9HP3547S90103
Node: Node1, Processor: CPU1 Part Number: N/A, Serial Number: 9HP3547S90072

Node1's SIVOC information
Node: Node1, Type: PowerSupply, Serial Number: 19CS2300164

Node1's Firmware Information
Node: Node1, Component: Node1.BIOS, Version: ex425.bios-1.7.2
Node: Node1, Id: Node1.HPCNet0, Component: SS11 200Gb 2P NIC Mezz Firmware, Version: 1.5.41
Node: Node1, Id: Node1.HPCNet1, Component: SS11 200Gb 2P NIC Mezz Firmware, Version: 1.5.41

FPGAs Firmware Information
Name: BMC, Component: Baseboard Management Controller, Version: nc.1.8.4-17-shasta-release.arm.2023-09-01T22:05:43+00:00.278b9e1
Name: nFPGA0, Component: Cray nFPGA-WNC Logic Device - Hardware Management, Version: 5.04
Name: mFPGA0, Component: Cray mFPGA-SAW0 Logic Device - Network Mezzanine HSN, Version: 2.02
Name: mFPGA1, Component: Cray mFPGA-SAW1 Logic Device - Network Mezzanine HSN, Version: 2.02
```

Questions

Thank you!