HTT – Hardware Triage Tool Updates

- Isa Wazirzada, Abhishek Mehta, Bhuvan Meda Rajesh, Vinanti Phadke

HTT – Hardware Triage Tool Overview

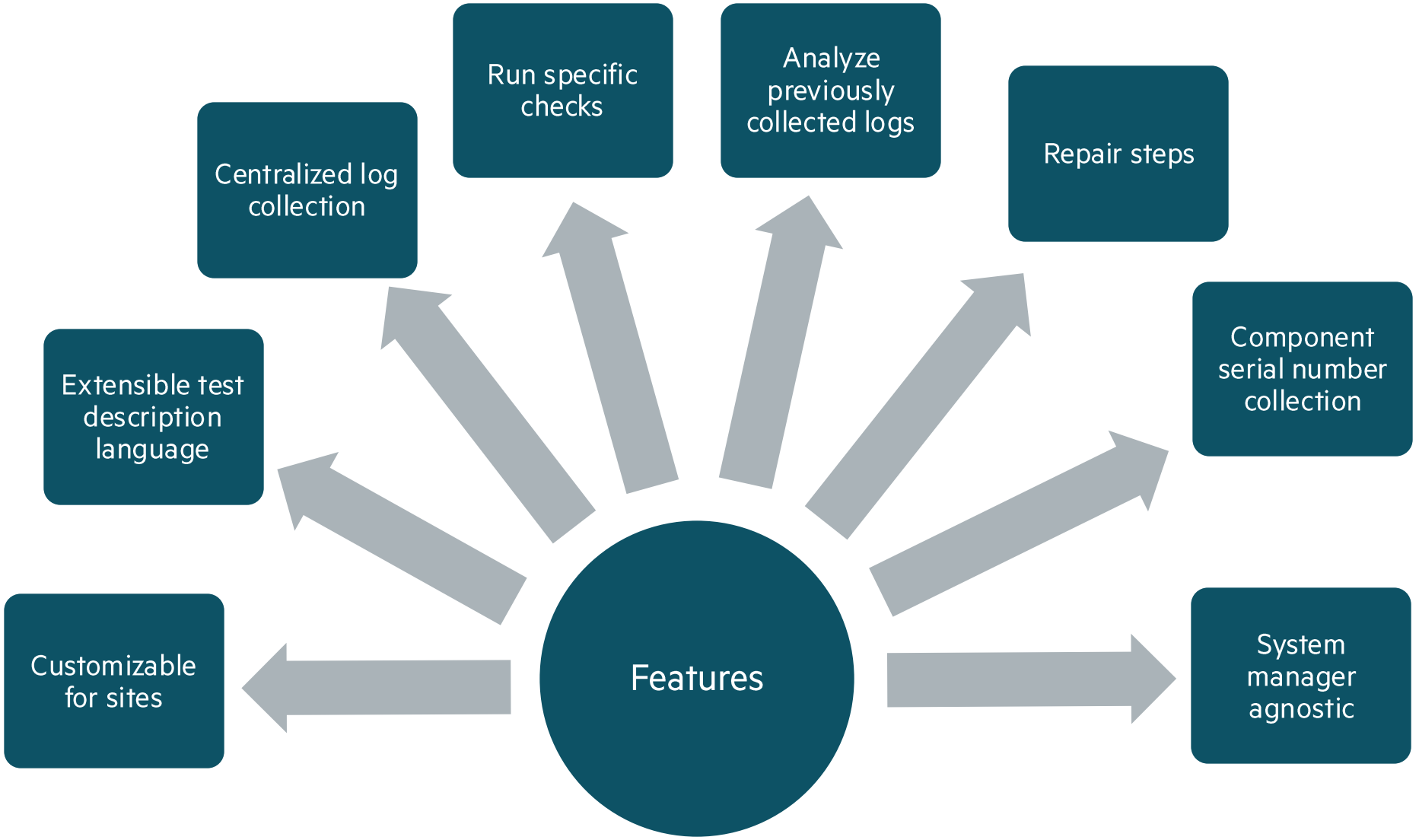
- Incorporates lessons learned from deployments across the world
- The hardware triage tool:
  - Checks for different failure signatures
  - Provides hardware actions and RMA codes (If applicable)
  - Builds a detailed support bundle even if it can't provide a diagnosis
- Current State
  - Can diagnose problems on several hardware programs
    - Cray EX: EX235a, EX255a, EX254n, EX4252, EX425, and the EX235n blades
    - Cray XD: XD220v, XD225v, XD295v, XD224, XD670 **– New**
  - Utilized in multiple geographies Europe Middle East and Africa (EMEA), Asia Pacific (APAC), and the Americas
  - Product level solution and System Manager agnostic
  - Tests for new hardware programs are being developed early in the product lifecycle

Hardware Triage Tool – What's in a name?

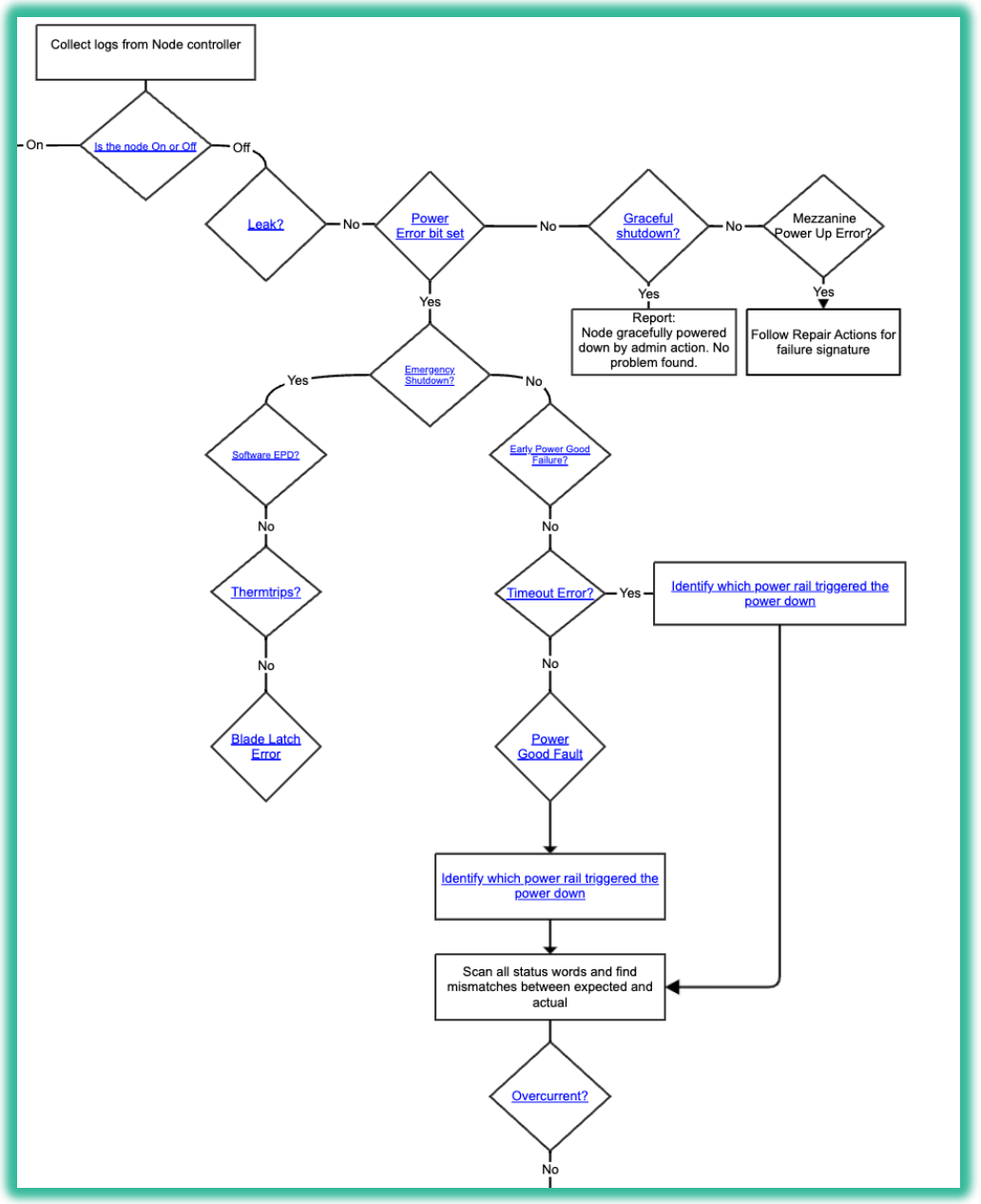
- What it is and it is not
- Administrators can run HTT to diagnose failures, for example if a compute node:

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
```

Test Description Language Enhancements

```
PrintBIOSVersion:
  exec_statement: dmidecode -s bios-version
    exec_statement_value_yes: "0"
    exec_statement_args: node
    yes condition:
      action: Successfully fetched the bios-version from dmidecode
    no condition:
      action: ERROR: Could not fetch the bios-version from dmidecode
```

```
Dracut _shell:
  search_pattern: Starting Dracut Emergency Shell
  search_file: current # One of the log files that's collected
  yes_condition:
    go_to: Check_interface_tmp0
  no_condition:
    go_to: EFI_Shell
```

Enhancement: Adding your own checks

- All test workflows reside in: **`/opt/clmgr/hardware-triage-tool/workflows_{ex,xd}`**
- Within an HPE provided test workflow additional checks can be added
- GPU Presence check example for the EX254n accelerator blade

```
MachineCheckEventsNC:
  custom_script: check_machine_events_nc.py
  custom_script_value_yes: "0"
  custom_script_args: log_path
  yes_condition:
    action: Contact HPE Support
  no_condition: None
...
---
GPU0_Presence_Check:
  exec_statement: nvidia-smi --id 0
  exec_statement_args: node
  exec_statement_value_yes: "0"
  exec_statement_value_no: "1"
  yes_condition:
    go_to: GPU1_Presence_Check
    action: "Test passed, no action required"
  no_condition:
    action: "Error: GPU0 not found"
```

Hardware Triage Tool – Command Line Arguments

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

```
Help page
usage: hwtriage [-h] [-r] [-n NODE_NAME] [-u USERNAME] [-p PASSWORD]
                [-l LOGPATH] [-ns {on,off}]
                [-hw {ex235a,ex255a,ex254n,ex4252,ex425,ex235n,xd220v,xd225v,xd295v,xd224,xd670}]
                [-ls] [-bs BEGIN_STAGE] [-rs RUN_STAGE] [-f INPUT_YAML]
                [-hy HARDWARE_YAML] [-sn] [-sno] [-k SSH_KEY]
                [-nck NC_SSH_KEY] [-t TIMEOUT] [-v] [-cpath CUSTOM_LOG_PATH]
                [-d]

optional arguments:
 -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,xd220v,xd225v,xd295v,xd224,xd670},
 --hardware {ex235a,ex255a,ex254n,ex4252,ex425,ex235n,xd220v,xd225v,xd295v,xd224,xd670}
                      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. Note: This currently doesn't work for XD224
                      hardware
 -sno, --serial-number-only
                      Collect the serial numbers into a file without
                      triaging. Note: This currently doesn't work for XD224
                      hardware
 -k SSH_KEY, --ssh-key SSH_KEY
                      Ssh key to enable passwordless ssh for node
 -nck NC_SSH_KEY, --nc-ssh-key NC_SSH_KEY
                      Ssh key to enable passwordless ssh for node controller
 -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
 -d, --disable-loader  To disable loader
```

Defining Hardware Configuration

- The hardware.yml file defines all supported hardware platforms
- Located at **`/opt/clmgr/hardware-triage-tool/hardware_config`**
- New feature allows creation of accurate hardware configuration files from nodes that are at the correct configuration levels
  - **`/opt/clmgr/hardware-triage-tool/generate_hardware_yaml.py`**

```
hardware:
  BIOSREV: None
  BIOSVER: 2.3.5
  CPUONLINE: 0-287
  DIMM_count: '4'
  DIMM_sizes: '1'
  DIMM_speed: '1'
  esm_link_speed: Disabled
  firmware_version: 1.5.53
  link_width: None
  mem_manufacture: '1'
  nic_speed: BS_200G
  number_of_nics: '4'
  pci_speed: 16.0 GT/s PCIe
  pci_width: '16'
  workflow_off: workflows_ex/workflow_ex254n_off.yml
  workflow_on: workflows_ex/workflow_ex254n_on.yml
```

Examples

Hardware Triage Tool Usage

- Installation on the admin node (HPCM) or ncn-m001 (CSM)
- HTT is invoked via the **`hwtriage`** command
  - Note the loader animation can now be disabled via the -d flag
  - Credentials are now automatically obtained from the System Manager (HPCM or CSM)
  - **`hwtriage -n x9000c3s5b0n1`**

```
Log collection completed
logging path : /var/log/hardware-triage-tool/x9000c3s5b0n1/20240419_1323
EX4252 Hardware is supported!
Triaging :x9000c3s5b0n1 ⣾
Node is in Off state
Analysis file : /var/log/hardware-triage-tool/x9000c3s5b0n1/20240419_1323/triage_output.json
Serial Numbers information :  /var/log/hardware-triage-tool/x9000c3s5b0n1/20240419_1323/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 the problem resurfaces replace the latch mechanism.
```

SIVOC Temperature Fault Example

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


logging path : /var/log/hardware-triage-tool/x1102c5s3b0n0/20240416_1608
ex254n Hardware is supported!
Triaging :x1102c5s3b0n0  ⁞

Node is in Off state
Triaging :x1102c5s3b0n0  ⁞Analysis file : /var/log/hardware-triage-tool/x1102c5s3b0n0/20240416_1608/triage_output.json
Serial Numbers information :/var/log/hardware-triage-tool/x1102c5s3b0n0/20240416_1608/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
logging path :  /var/log/hardware-triage-tool/x9000c1s0b0n1/20240502_1427
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/triage_output.json
Serial Numbers information :  /var/log/hardware-triage-tool/x9000c1s0b0n1/20240502_1427/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/check_node_health for more
details.
Triaging :x9000c1s0b0n1    ⁞Stage analysis : Unexpected_Booted Detected!
Triaging :x9000c1s0b0n1    ⁞DIMM 13 is missing
```

Enhanced Machine Check Error Reporting

```
Triaging: x9000c1s0b0n0    ⣿MCE Error Occured
MCE Error count: 80
Displaying the first 10 lines of the error records:

[25891.045323] mce: [Hardware Error]: Machine check events logged
    [25891.049052] [Hardware Error]: Corrected error, no action required.
    [25891.052566] [Hardware Error]: CPU:65 (19:1:1) MC18_STATUS[Over|CE|MiscV|AddrV|-|-|SyndV|CECC|-|-|Scrub]: 0xdc2041000000011b
    [25891.058580] [Hardware Error]: Error Misc: 0x0000000000000000
    [25891.061821] [Hardware Error]: Error Addr: 0x0000000303452800
    [25891.065070] [Hardware Error]: PPIN: 0x02b574e2cd790008
    [25891.068098] [Hardware Error]: IPID: 0x0000009600350f00, Syndrome: 0x397400400a800e00
    [25891.072519] [Hardware Error]: Unified Memory Controller Ext. Error Code: 0
    [25891.081708] [Hardware Error]: cache level: L3/GEN, tx: GEN, mem-tx: RD
    [53284.031154] mce: [Hardware Error]: Machine check events logged
Triaging: x9000c1s0b0n0    ⣿Stage analysis: MachineCheckEvents Detected!
```

Log Collection Example – New logging directory structure

```
ncn-m001:/var/log/hardware-triage-tool/x1002c0s3b1n0/20250425_1001 # ls -ltr
total 4424
-rw-r--r-- 1 root root   69180 Apr 18 14:14 ssif-current
-rw-r--r-- 1 root root 2922258 Apr 24 15:31 current
-rw-r--r-- 1 root root    2514 Apr 25 10:08 serial_numbers.txt
-rw-r--r-- 1 root root  166769 Apr 25 10:08 messages
-rw-r--r-- 1 root root    4096 Apr 25 10:08 power_down_req
-rw-r--r-- 1 root root    3917 Apr 25 10:08 powerfault_epd.Node0.nfpga.json
-rw-r--r-- 1 root root    2753 Apr 25 10:08 powerfault_epd.Node0.json
-rw-r--r-- 1 root root  222710 Apr 25 10:08 tlsproxy-current
-rw-r--r-- 1 root root       0 Apr 25 10:02 check_node_health.log
-rw-r--r-- 1 root root 1116009 Apr 25 10:02 dmesg_output.txt
-rw-r--r-- 1 root root    1916 Apr 25 10:02 triage_output.log
-rw-r--r-- 1 root root    1148 Apr 25 10:02 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
```

HPE Cray XD Support

HPE Cray XD Support

- Hardware Triage Tool support extended to – XD220v, XD225v, XD295v, XD224, XD670

- Cray XD line - HPE Redfish Crawler is the primary knowledge source
  - The HPE Redfish Crawler walks the Redfish API tree and collects **every** @odata.id reference
  - Collects all available information from a Redfish endpoint in one command
    - Logs
    - Serial numbers
    - Component health and state information
  - Available independently on support.hpe.com

- Hardware Triage tool test workflows check for:
  - Faulty power supplies
  - Missing components like memory DIMMs, accelerators, NICs, etc
  - Component health problems exposed via Redfish
  - Power draw and thermal load related problems
  - Node level consistency checks

Questions

Thank you!