# Contents

# GB200/GB300 Shape Runbook: Partner Diagnostics Fail/Mod Codes Mapping

## Table of Contents

## Pre-Requisites:

- OCI CLI installed and working
  - Link to How to Install OCI CLI
- HOPS CLI installed and working
  - Link to How to Install HOPS CLI
- AllProxy (required by HOPS CLI)
  - Link to How to Install AllProxy

## Purpose

To outline actions for NVIDIA partner diagnostic fail/mod codes on GB200 shapes.

## When to Use

When triaging GB200 L10 and L11 Partner Diag test failures, with one or more MODS/DGX error codes in a fail log that need triaging.

The MODS/DGX fail codes in the FDT logs help identify the nature of the failure.

### Useful NVOnline IDs

- **PID 1116117** - Server RAS Guide
- **PID 1109712** - Debug and RAS Guide for NVIDIA Data Center Products
- **PID 1138824** - Multi Node NVLink Troubleshooting Guide

### Standard Log Collection

| Tray Type | Required Logs/Reports |
|---|---|
| **Compute Trays** | - NVIDIA bug report - FDT Logs (if available) - One-Click logs (can be retrieved with ILOM snapshots) |
| **NVSwitch Trays** | - NVOS Tech support logs from all 9 NVSwitches - Fabric Manager and NVLSM logs (from NVSwitch controller) |

**Important Note:**

**In the presence of more than one MODS/DGX error codes, it is important to note that one some may be the root cause and others a symptom of the failure.**

**Please follow the priority ordering noted below written with the goal of handling the codes that are likely the causes.**

## Brief Notices (Please Read)

### Q: What do I do if I am confused on what to do on the runbook?

**A:** If at anytime you are confused please reach out to CPV Tier 2 triage teams or CPV primary and secondary oncall for clarifications on repair steps. You can file a **Sev2/Sev3 ticket** in the **JIRA-SD HPC** queue.

For more urgent/time-sensitive issues, a slack channel you can reach out to is #cpv-triage.

Link to internal OCI slack channel: #cpv-triage

Link to JIRA-SD HPC ticket queue

We will attempt to address these issues as soon as we are able to.

### Q: What do I do if MODS code is not present in the runbook? ?

**A:** Please escalate to CPV Tier 2 triage team(s). You can escalate to Tier -2 by creating a linking ticket in GPUFM queue with CPV- GPU component item.

### Q: What do I do if the runbook does not resolve the issue and FDT fails the same way?

**A:** Please escalate to CPV Tier 2 triage team(s). In the spirit of saving time we would like to avoid hosts being stuck in CPV for long periods of time. If the runbooks don't work, or a runbook for your issue does not exist, please notify CPV tier 2 triage teams or someone under Gregory Siekas' team. You can create a JIRA-SD ticket, reach out to CPV tier 2 triage teams, primary and/or secondary CPV oncalls, or `#cpv-triage` slack channels. We would also be grateful if you are able to provide as many details as possible to help us troubleshoot and avoid doing the same triage/troubleshooting work you or others may have done previously.

### Q: Why are we not reseating GB200 trays anymore?

**A:** NVIDIA has advised OCI to avoid and IGNORE all reseating of GB200 compute trays and NVSwitch trays if it can be avoided. To quote one of the NVIDIA engineers, "reseating the trays can bend the pins on the tray" and Reseating multiple times is likely to cause damage to the NVLink connectivity for a compute tray. This is likely what contributed to hosts/racks in HSG being stuck in CPV for multiple weeks failing continually on FDT L10/L11 tests with NVLink failures.

### Q: Are FDT failures linked to RDMA network, should I cut a DO ticket?

**A:** Compute tray level and rack level fdt tests don't use RDMA links. Please don't cut any DO tickets to fix RDMA cabling for FDT failures. Cable validation , mlx check and wait for wpa tests will catch RDMA link failures.

## Steps to Triage

1. If there are any XID errors in the `dmesg` and/or kernel logs and the JIRA ticket and/or test logs indicate a timeout failure or the node has crashed, please cut a RHS ticket with these details. For reference, XID errors are reported to the host's operating system's kernel log or event log from the NVIDIA GPU driver. If you are curious to learn more, please refer to Nvidia XID Errors.
   - `dmesg` logs.
   - NVIDIA bug report (run `nvdebug` or `nvidia-bug-report.sh`).
   - Test logs from the failed test (i.e. FDT logs if you run FDT. . . ).
2. In the presence of more than one MODS/DGX error codes, it is important to note that some may be the cause and others a symptom of the failure.
3. Please follow the priority ordering starting at step 3 written with the goal of handling the codes that are likely the causes.

4. First, check if any MODS codes in seen in the list of MODS/DGX fail codes is present in NVLink-related fails. If FDT tests pass after following the runbook for the specified MODS errors, then you can end this runbook early.
5. Secondly, check General Non-RMA-eligible fails. If FDT tests pass after following the runbook for the specified MODS errors, then you can end this runbook early.
6. Thirdly, check Non-NVLink RMA-eligible fails. If FDT tests pass after following the runbook for the specified MODS errors, then you can end this runbook early.
7. If not, then check Other Non-NVLink fails.

### General Non-RMA-eligible fails:

**Action item:** On test fail, check the firmware version and retest with the latest test version. **If the fail is repeated, file an NVBug including dmesg, diagnostic error logs and bug report.**

| MODS Code | Error Message |
|---|---|
| xxxxxxxxx002 | Software error |
| xxxxxxxxx008 | Bad parameter passed to function |
| xxxxxxxxx021 | Script failed to execute |
| xxxxxxxxx077 | Timeout error |
| xxxxxxxxx144 | NVIDIA CUDA error |
| xxxxxxxxx240 | Unexpected result from hardware |
| xxxxxxxxx272 | Read parameter differs from expected |
| xxxxxxxxx318 | ECC detected a correctable error in L2 over threshold |
| xxxxxxxxx779 | Voltage value out of range |
| xxxxxxxxx818 | Mods detected an assertion failure |

### Non-NVLink RMA-eligible fails:

**Action item:** On test fail, check the firmware version and retest with the latest test version. **If the fail is repeated, start the RMA process.**

| MODS Code | Error Message |
|---|---|
| xxxxxxxxx083 | CRC/Checksum miscompare |
| xxxxxxxxx097 | Unexpected device interrupts |
| xxxxxxxxx194 | Bad memory |
| xxxxxxxxx276 | Hardware reports wrong status |
| xxxxxxxxx316 | ECC detected a correctable error over threshold |
| xxxxxxxxx319 | ECC detected an uncorrectable error in L2 |
| xxxxxxxxx320 | ECC detected a correctable error over threshold |
| xxxxxxxxx321 | ECC detected an uncorrectable error |
| xxxxxxxxx341 | Buffer mismatch |
| xxxxxxxxx363 | Row remapping failed |
| xxxxxxxxx539 | NVRM Generic Falcon Error |
| xxxxxxxxx541 | NVRM Detected memory error |
| xxxxxxxxx582 | GPU Stress Test found pixel miscompares |
| xxxxxxxxx612 | Invalid value for Tegra configuration variables |
| xxxxxxxxx614 | Extra golden code miscompare |
| xxxxxxxxx774 | Tegra test failed |

### Other Non-NVLink fails:

| MODS Code | Error Message | Suggested Action |
|---|---|---|
| xxxxxxxxx124 | InvalidInfoROM | Check firmware, run inforom recovery tool and retest. If still failing, then RMA. |
| xxxxxxxxx143 | PCI Express bus error | Ensure GPU and NVSwitch devices are detected on lspci and retest. If still failing, then RMA. |

| MODS Code | Error Message | Suggested Action |
|---|---|---|
| xxxxxxxxx167 | GFW boot reported a failure | Check firmware, reboot system, ensure GPU and NVSwitch devices are detected on lspci and retest. If still failing, then file NVbug. |

| MODS Code | Error Message | Suggested Action |
|---|---|---|
| xxxxxxxxx220 | PCIE device not found | Check firmware, ensure GPU and NVSwitch devices are detected on lspci, cold reboot and retest. If still failing, then RMA. |
| xxxxxxxxx317 | ECC detected an uncorrectable error in FB | Check firmware, power cycle host and retest. If still failing, then RMA. |

| MODS Code | Error Message | Suggested Action |
|---|---|---|
| xxxxxxxxx679 | NVRM invalid argument | Check firmware, run inforom recovery tool and retest. If still failing, then RMA. |

| MODS Code | Error Message | Suggested Action |
|---|---|---|
| xxxxxxxxx688 | NVRM invalid state or config | Inspect the compute tray fabric registration state with `nvidia-smi -q \| grep Fabric -A 9`. If fabric registration is not complete, then check the NVSwitch controller to see if it is running nmx-c with the correct NVSwitch configuration. You may need |

| MODS Code | Error Message | Suggested Action |
|---|---|---|
| xxxxxxxxx272 | Read parameter differs from expected | Retest. If the error is fatal, then power cycle, else ignore. |
| xxxxxxxxx936 | TLW Error | |

### NVLink-related FDT Test Failures /

If the following MODS error codes are seen, then this indicates a likely NVLink-related issue. If this is the case, then please follow the NVLink-related failure triage steps below:

| MODS Code | Error Message |
|---|---|
| xxxxxxxxx014 | Low bandwidth |
| xxxxxxxxx140 | NVLink bus error |

**Note on NVLink Related FDT Test Failures:** When a test was not run on a full rack of 18 hosts, there may have been interference from traffic initiated by the other hosts, that cause an NVLink-related fail.

### Checklist before troubleshooting NVLink-related CPV FDT test Failures:

**If any host on this rack is in a customer pool (not in `holding` pool), please reach out to SCE team before attempting any RMA or any CHS/DO action on the NVSwitches. You can view the rack SK id or host serial/rack serial in Compute Admin/C4PO: https://devops.oci.oraclecorp.com/compute-admin**

At any point when going through the below, please feel free to reach out to the CPV Tier 2 oncall primary and/or secondary. There may be a point where it will be more productive to escalate hosts/racks to CPV tier 2 triage teams to prevent hosts/racks from being stuck in CPV testing.

### NVLink-related failure triage steps:

1. Please review the JIRA ticket history of the host. Identify when and if the host was RMA'd. If the host has multiple failures between tests or has multiple repeated failures for the same FDT test, then please confirm and review the error history to the best of your ability. You may find the OCI CLI commands helpful to retrieve the debug logs for the host if the PAR links have expired. In particular, please review any HPC, GPUFM, CHS/DO, or RHS tickets that mention the host. Note any physical repairs that have been done on the host.

**If you are having trouble with some of the steps, please refer to CPV FDT NVLink Runbook Diagram**

2. Follow How to find rack firmware version to identify firmware version that is installed on the rack and post this in the HPC/GPUFM/RHS ticket(s). **Is there an action item for OCI here in the case of a mismatch? - Clarify with NVIDIA on how to deal with mismatched firmware versions and rack split between customer pools.**
3. Search the Compute Tray FDT Test Logs. Download the logs from the HPC/GPUFM ticket. If they are not there, then please follow directions in: How to Find Compute Tray FDT Logs.

You should see an error that shows some sort of MODS or DGX error.

```
MODS-000000000140 | NvlBwStressBg610Pulsy | nvlink | concurrent| NVLink | 0009:01:00.0_SN_1655124012808
```

OR

```
NvLink errors on 0018_01_00.0_SN_1655224089559 Link 4:
  Remote Connection                              : SharedQM3NvSwitch [ffff:0d:04.0], Link 2
```

4. If the error from step 2 is the first type, then use the slot indices of the compute tray (rack slot 25 in the example above) and the peer switch tray (rack slot 17 in the example above), along with the mapping table below to identify which specific compute tray(s) and switch tray(s) are involved in causing the failure. If the error from Step 2 is the second type, then use the NVSwitch Tray ASIC identified from the **"NVSwitch ASICs Contained"** column to identify the NVSwitch tray of interest.

| Chassis Phy Slot Number | COMPUTE or SWITCH Tray Index | Tray Name | NVSwitch ASICs Contained |
|---|---|---|---|
| 27 | 17 | Compute Tray 18 | N/A |
| 26 | 16 | Compute Tray 17 | N/A |
| 25 | 15 | Compute Tray 16 | N/A |
| 24 | 14 | Compute Tray 15 | N/A |
| 23 | 13 | Compute Tray 14 | N/A |
| 22 | 12 | Compute Tray 13 | N/A |
| 21 | 11 | Compute Tray 12 | N/A |
| 20 | 10 | Compute Tray 11 | N/A |
| 19 | 9 | Compute Tray 10 | N/A |
| 18 | 8 | Compute Tray 9 | N/A |
| 17 | 8 | Switch Tray 9 | SharedQM3NvSwitch [ffff:0d:10.0]<br>SharedQM3NvSwitch [ffff:0d:11.0] |
| 16 | 7 | Switch Tray 8 | SharedQM3NvSwitch [ffff:0d:0e.0]<br>SharedQM3NvSwitch [ffff:0d:0f.0] |
| 15 | 6 | Switch Tray 7 | SharedQM3NvSwitch [ffff:0d:0c.0]<br>SharedQM3NvSwitch [ffff:0d:0d.0] |
| 14 | 5 | Switch Tray 6 | SharedQM3NvSwitch [ffff:0d:0a.0]<br>SharedQM3NvSwitch [ffff:0d:0b.0] |
| 13 | 4 | Switch Tray 5 | SharedQM3NvSwitch [ffff:0d:08.0]<br>SharedQM3NvSwitch [ffff:0d:09.0] |
| 12 | 3 | Switch Tray 4 | SharedQM3NvSwitch [ffff:0d:06.0]<br>SharedQM3NvSwitch [ffff:0d:07.0] |
| 11 | 2 | Switch Tray 3 | SharedQM3NvSwitch [ffff:0d:04.0]<br>SharedQM3NvSwitch [ffff:0d:05.0] |
| 10 | 1 | Switch Tray 2 | SharedQM3NvSwitch [ffff:0d:02.0]<br>SharedQM3NvSwitch [ffff:0d:03.0] |
| 9 | 0 | Switch Tray 1 | SharedQM3NvSwitch [ffff:0d:00.0]<br>SharedQM3NvSwitch [ffff:0d:01.0] |
| 8 | 7 | Compute Tray 8 | N/A |
| 7 | 6 | Compute Tray 7 | N/A |

| Chassis Phy Slot Number | COMPUTE or SWITCH Tray Index | Tray Name | NVSwitch ASICs Contained |
|---|---|---|---|
| 6 | 5 | Compute Tray 6 | N/A |
| 5 | 4 | Compute Tray 5 | N/A |
| 4 | 3 | Compute Tray 4 | N/A |
| 3 | 2 | Compute Tray 3 | N/A |
| 2 | 1 | Compute Tray 2 | N/A |
| 1 | 0 | Compute Tray 1 | N/A |

5. If the whole rack (18 hosts) were part of this specific CPV FDT Test, then skip to step 13. Otherwise, we need to analyze the fabric manager logs in order to confirm that the failure was not caused as the side effects of a host on this rack that was not part of this test. For example, a host with a customer **OR** in CPV running a separate test. This is performed in steps 6, 7, and 8.

6. Obtain and/or access NVswitch controller Fabric Manager logs (you can download/make a copy if you want to) by following the directions in How to inspect NVSwitch Controller Fabric Manager Logs by following steps `A` through `H`.

7. Do a `grep` for `Fatal` errors in the Fabric Manager log (exclude `non-Fatal` errors), and then create a list of GPU GUIDs that had a `Fatal` error during or close to the timespan of the test. (Check the FDT logs for relevant start and end timestamps for the relevant FDT test ID).

```
$ grep -i "Fatal" fabricmanager.log* -A 11
--
[Oct 13 2025 18:14:38] [ERROR] [tid 202] Fabric Manager detected GPU NVL Fatal error on :
moduleId : 3
nodeId : 7
partitionId : 4137
gpuGuid : 0xED0A8727AFC3C6F9  # take note of the GPU GUIDs
portNum : 4
portStatus : 1
errorCode : 0x06
errorSubcode : 0x0E
portDownReasonCode : 0x0A
isErrorFirst : 0
errorStatus : 0x00000000
```

8. Now create a different list containing all of the GPU GUIDs that are on this rack but were not involved in this specific FDT test job ID. Note that some of these GPU GUIDs may potentially be in a customer pool or running a different CPV job.

This can also be found at How to get GPU Information for GB200 compute trays

```
# Identify GPU GUIDs for all compute tray hosts with `nvidia-smi -q`
# From the CPV bastion
# Use clush
$ cpv state overview-count -f RackNumber=<rack_number> | awk '{print $2}' # get compute tray IP address
$ clush -l ubuntu -w <compute_tray_ips> "nvidia-smi -q | grep -i guid " | awk '{print $5}'

# ssh into the compute tray
$ ssh ubuntu@<compute_tray_ip>
$ nvidia-smi -q | grep -i GUID
    GPU Fabric GUID                       : 0x593ac6c850427ec6
    GPU Fabric GUID                       : 0x4b8c9aed09601d7e
    GPU Fabric GUID                       : 0x2710d0c1defbc858
    GPU Fabric GUID                   : 0xf1fdf4af270643d7

# make a map for the GPU GUIDs, module IDs, serial numbers and GPU Fabric GUIDs for all GPUs on the rac
$ nvidia-smi -q | egrep "GPU 000|Module Id|Serial Number|GPU Fabric GUID"
GPU 00000008:01:00.0
    Serial Number                         : 1650325068949
        Chassis Serial Number             : 1820025180411
        Module Id                         : 2
        GPU Fabric GUID                   : 0x593ac6c850427ec6
GPU 00000009:01:00.0
    Serial Number                         : 1650325068949
        Chassis Serial Number             : 1820025180411
        Module Id                         : 1
        GPU Fabric GUID                   : 0x4b8c9aed09601d7e
GPU 00000018:01:00.0
    Serial Number                         : 1650225112368
        Chassis Serial Number             : 1820025180411
        Module Id                         : 4
        GPU Fabric GUID                   : 0x2710d0c1defbc858
GPU 00000019:01:00.0
    Serial Number                         : 1650225112368
        Chassis Serial Number             : 1820025180411
        Module Id                         : 3
        GPU Fabric GUID                   : 0xf1fdf4af270643d7
```

**Note: You may find it helpful to make a GPU GUID map containing all compute trays/GPUs on the rack that you are able to access.** To make the GPU GUID map, you can look into the NVLink mapping tool at Other Helpful Links and C4PO to identify which GPU belongs to which compute tray on the GB200 rack.

9. Find any entries in the `Fatal` error GPU GUID list from step 8 that contain a GPU GUID that is **NOT** present in the list of GPU GUIDs with Oracle/OCI from step 8. The goal here is to create a list of GPU GUIDs/compute tray hosts not part of this specific FDT CPV test job ID had `Fatal` errors that may have indirectly induced the test failure as observed in the CPV FDT test. Note all of this in the HPC/GPUFM/RHS ticket(s). If the list that was created is empty, then skip to step 13.

10. If the CPV FDT test involved fewer than 18 hosts on the rack, and the analysis from step 8 indicates that a compute tray or a switch tray other than the ones identified in step 4 had a fatal error occur over the duration of the test ID (i.e. `connectivity`, `nvlbwstress` etc…), then need to do an FDT run with all 18 hosts on the rack. This may involve engaging Strategic Customer Engagement (SCE) team to have the customer evacuate the rack if any part of the rack is with the customer.

11. Make sure all 18 hosts on the rack are in CPV and are not running any tests (make blocking HPC tickets if needed). Then proceed to run a rack-level FDT L11 test on the entire rack. If this re-run succeeds, then you can end this runbook early. If L11 was re-run and it fails, treat this as a fresh failure and then start back from Steps to Triage. You can also escalate to CPV tier 2 triage teams and/or CPV primary and secondary oncall for clarification and guidance on next steps.

12. If, instead, no hosts on this rack were with the customer, or the fabric manager logs indicate that no other nvswitch tray or compute tray other than the ones identified in step 4 had a failure occur, then continue on to the next step.

13. Initiate the GPUPR automated RMA process for the compute tray and re-run the FDT test. You can end the runbook here early if the retest succeeds. If the test fails again, then it is likely not a compute tray issue, and possibly an nvswitch tray or cable cartridge issue.

**STOP - MAKE SURE YOU READ STEP 15 BEFORE FOLLOWING STEPS 15-17**

14. **If there are any hosts in the customer pool, you must reach out to the SCE team and have the customer send the host(s) to CPV before proceeding to step 15 and onwards.**
15. Engage GPUPR team to start the process to RMA the suspect NVSwitch tray. Once the new NVSwitch is installed, please re-run the test. You can end the runbook here early if the retest succeeds. **See:** GPUPR Automated RMA runbook(s).
16. If the test from step 15 fails, cut a DO ticket to replace the cable cartridges. Once CHS/DO team is finished working, please re-run the test.
17. If the test from step 16 fails, escalate to CPV tier 2 making a Sev2 ticket. Our ticket queue is called **High Performance Computing** with Project key: **HPC**. See the link below for where to file the ticket.

- Link to HPC JIRA-SD ticket queue

## Other Helpful Links

| Link | Notes |
|---|---|
| NVLink Mapping Tool for Compute Trays and NVSwitches https://oracle.sharepoint.com/:x:/r/teams/AI2ComputeCollaboration/Shared%20Documents/Projects/GPU/NVIDIA%20GB2 | Please be careful if you have access to this information. It is from the strategic business partner website (NVOnline) and all documentation is highly restricted. If you do not have access and would like to request access to this link, please reach out to Gregory Siekas' team. It can be found on NVOnline if you have an account. |
| Substrate Credentials to access NVSwitches, BMCs, ILOMs, smartNIC/ARMs etc… | If you do not have access to the credentials then you may need to request access. Please reach out to a manager if you should be authorized to access these credentials. Please be very careful with these! The permissions are enforced via your SSO user for OCI and permissions portal. |
| Permissions Portal for Prod-Admin JIT activation | **This is needed to mark hosts as broken. Be careful with these permissions. Before marking a host as broken and/or terminating the host check to see if it has a customer instance. If it does, confirm with a manager that is oncall or TPM, AND Strategic Customer Engagement team (SCE) and ensure you have customer approval to do so. You will need a valid service ticket to attach to the action.** |
| GB200 Operations Swiss Army Knife | Documentation containing various runbooks and other information useful for GB200 operations. |
| GB200 NVSwitch Debugging Runbook | Contains some helpful commands and information if you need to investigate other parts of the NVSwitches. |
| HOPS Pinning Golden Set Config Source of Truth | Refer to this if you would like to identify version strings for firmware bundles for each GPU shape for Commercial. |
| HOPS Rack/Host Firmware Pinning Casper Config | Refer to this to determine which rack serial(s) or host serial(s) are pinned to which firmware bundle for Commercial. |

| Link | Notes |
|---|---|
| GPUPR Automated RMA runbook(s) | GPUPR runbooks for automated RMA for GB200 Compute trays and NVSwitch trays. |
| How to Identify hosts with Host-SmartNIC or SmartNIC-TOR link issues | This can be used if the host is unpingable and/or unsshable after attempting a powercycle. This runbook is owned by HPC/CPV team(s). If there are any issues with this runbook, please reach out to CPV team(s) at `#cpv-triage` internal OCI slack channel. The other SmartNIC runbooks below are owned by Card Management team. |
| Card Management Check SmartNIC to TOR Network Links | If you run into any issues with this runbook, please kindly reach out to Card Management DP team primary or secondary oncall. They are the SME's on SmartNIC/ARM cards. |
| Card Management BF3 SmartNIC Access Runbook ROT or SSH | If you run into any issues with this runbook, please kindly reach out to Card Management DP team primary or secondary oncall. They are the SME's on SmartNIC/ARM cards. |

## Appendix:

### How to Find Rack Firmware Version(s)

```
# Start hops-cli
$ cd hops-cli
$ source venv/bin/activate
$ hops-cli

# input yubikey pin when prompted

> host <host_serial>
> info | awk '{print $18}'
# the output is the rack serial.
> host <rack_serial>
> firmware_inventory

# note the output of the `firmware_inventory` command in the HPC/GPUFM/RHS ticket(s).
```

### How to Find Compute Tray FDT Logs

```
# 1. Access the Compute Tray host in CPV
$ ssh ubuntu@<compute_tray_ip>
$ cd ~/bio/jobs/results
$ cat execute.sh.log | grep -i "object put"

# Grab the MODS code from the run.log in /local/FDT/629...../dgx/logs-....../run.log
# For the failed test, look at the fieldiag_summary.log to get the details of the error.

#Output may look like below:

ubuntu@instance20251013181111:~/bio/jobs/results$ cat execute.sh.log | grep -i "object put"
+ /home/ubuntu/bin/oci os object put --namespace hpc --bucket-name Debug --file logs/logs-20251013-19363
+++ grep 'os object put'
+ /home/ubuntu/bin/oci os object put --namespace hpc --bucket-name JobResults --file /home/ubuntu/bio/j
```

```
# Identify the value in the "--name" field of any "oci os object put" commands
# We are looking for:
# fieldiag.log
# fdt_compute_tray_gb200_clustered/50ce059c-766c-456c-86bf-0af0e371bf69/logs-20251013-193630-2520XNG1XA
# usually the FDT logs will have this sort of naming convention: logs-{YYYYMMDD}-{HHMMSS}-{HOST_SERIAL}

# 2. The value for region may change depending on the region you are troubleshooting.
$ oci session authenticate --profile DEFAULT --tenancy-name bmc_operator_access --region ap-batam-1
# This may open up a browser to login. Please ensure your session uses bmc_operator_access

# 3. Download the FDT logs to your local machine:
$ oci os object get --profile DEFAULT --auth security_token -ns hpc -bn Debug --name fdt_compute_tray_gb

# Inspect the run.log in the file.
# Identify any tests that fail, then inspect the corresponding folder for any additional logs
# We are looking for any timestamps of when the failure occurred.
# This is needed to cross check with NVSwitch controller logs to identify if there was a conflicting wo
```

### How to inspect NVSwitch Controller Fabric Manager Logs

```
# A. start hops-cli.
$ cd hops-cli
$ source venv/bin/activate
$ hops-cli

# input yubikey PIN when prompted

> host <host_serial>
> info
# OR
> info | awk '{print $18}'
# identify rack serial

# B. Load rack in hops-cli.
> host <rack_serial>

# C. Filter down to NVSwitch model from "info" command.
> filter model=GPU_GB200-NVL72_Switch_S.01

# D. Find NVOS IP address for NVSwitch controller on elevation 19.
> nvos_presence
> info

# E. Create a new terminal.

# F. SSH to a substrate bastion host. HOPS VM or the one below can work.
$ ssh bastion-ad1.rb.ap-batam-1.oci.oracleiaas.com

# G. SSH to the NVSwitch controller
# password for the NVSwitch can be found in secret service using the link below. Adjust for your desire
# https://devops.oci.oraclecorp.com/secret-service/ap-batam-1/namespace/hops-gateway/secret/ad1-nvos
$ ssh admin@<nvswitch_controller_ip>
```

```
# H. Fabric manager log can be obtained/read in the path below.
$ cat /var/log/nmx/nmx-c/fabricmanager.log

# I. This is not required, but for example, if you are interested in inspecting the fatal errors, do th

$ grep -i "Fatal" fabricmanager.log* -A 11
--
[Oct 13 2025 18:14:38] [ERROR] [tid 202] Fabric Manager detected GPU NVL Fatal error on :
moduleId : 3
nodeId : 7
partitionId : 4137
gpuGuid : 0xED0A8727AFC3C6F9  # take note of the GPU GUIDs
portNum : 4
portStatus : 1
errorCode : 0x06
errorSubcode : 0x0E
portDownReasonCode : 0x0A
isErrorFirst : 0
errorStatus : 0x00000000

# J. Then consider posting your findings in the HPC/GPUFM/RHS ticket(s) if it can be of value.
```

### How to inspect NVSwitch Controller Subnet Manager logs

```
# 1. Identify GPU GUIDs for all compute tray hosts with `nvidia-smi -q`
# From the CPV bastion
# Use clush
$ cpv state overview-count -f RackNumber=<rack_number> | awk '{print $2}' # get compute tray IP address
$ clush -l ubuntu -w <compute_tray_ips> "nvidia-smi -q | grep -i guid " | awk '{print $5}'

# ssh into the compute tray
$ ssh ubuntu@<compute_tray_ip>
$ nvidia-smi -q | grep -i GUID
    GPU Fabric GUID                    : 0x593ac6c850427ec6
    GPU Fabric GUID                    : 0x4b8c9aed09601d7e
    GPU Fabric GUID                    : 0x2710d0c1defbc858
    GPU Fabric GUID                    : 0xf1fdf4af270643d7
# search GUID ids in the fabric manager logs and nv subnet manager logs (nvlsm.log)

# 2. Go back to the NVSwitch controller terminal and search Nvidia subnet manager log for the history o
$ grep -i "<gpu_guid>" nvlsm.log*
# OR
$ grep -in "<gpu_guid>" nvlsm.log*

# Manually inspect the other events happening around the time of the log events.
$ less +<line_number> nvlsm.log


$ echo <gpu_guid_1>,<gpu_guid_2>,... | xargs -n 1 -I {} grep -in {} nvlsm.log*

# 3 Parse and see if there are any notable errors, particularly link down errors, logs or other failure
```

#### Gather Logs for FDT failures

You will usually need at least the below: - FDT Test logs - NVDebug / Nvidia-bug-report.sh - NVSwitch Logs (if FDT test failure has NVLink-related error)

#### FDT Logs

Please see: [How to Find Compute Tray FDT Logs](#).

#### Gathering NVDebug logs (includes Nvidia-bug-report.sh)

```
# Identify Rack Number
# Connect to CPV bastion
# Get Rack Serial
$ cpv state read -t host -k <host_serial> | jq -r .[].rackNumber

# Connect to compute tray host

$ ssh ubuntu@<compute_tray_ip>
$ cd ~/nvdebug
$ sudo ./nvdebug -b "GB200 NVL" -t arm64 --local -v -o ./nvdebug-logs
$ sudo chmod -R 777 ./nvdebug/nvdebug-logs
$ hostname=$(cat ~/hostname.txt); cd ~/nvdebug/nvdebug-logs; for f in *.zip; do mv "\$f" "Rack<rack_num
# should rename nvdebug log file to RackXXXX_HOSTSERIAL_nvdebug-logs-{DATE}.zip or something.
# Upload logs to Object storage
$ oci os object put --namespace hpc --bucket-name Debug --file ~/nvdebug/nvdebug-logs/Rack5505_2518XNG0

# create PAR link for the new log(s)
$ oci os preauth-request create --namespace hpc --bucket-name Debug --profile DEFAULT --auth security_t

# Post this PAR Link to HPC/GPUFM/RHS Ticket(s) and anywhere else that is needed.

# For more details on usage, see below link:
# https://docs.oracle.com/en-us/iaas/tools/oci-cli/3.68.0/oci_cli_docs/cmdref/os/preauth-request.html
```

#### How to Retrieve ILOM Snapshots

From the ILOM using `connect_ws` from `hops-cli` or if you have `ssh` access to the ILOM:

```
-> start /SP/diag/shell
-> hwdiag gpu get_bmc_log hmc
-> hwdiag gpu get_bmc_log fpga
# exit

# configure to take full ILOM snapshot
-> set /SP/diag/snapshot dataset=full

# manually capture the snapshot and download to your local machine (avoid hops-cli timeout)
$ ssh -o UserKnownHostsFile=/dev/null -o StrictHostKeyChecking=no -o HashKnownHosts=no -l cpjohnso -p 2
```

Script to run from your developer workstation if you are on OCNA Network + VPN:

```
# #######################################################################################
# # ORIGINAL ILOM SNAPSHOT SCRIPT FROM ZACK BERKSHIRE
# # % ip=10.160.106.216
# # % bastion=bastion-ad1.rb.ap-batam-1.oci.oracleiaas.com
# # % user=$(whoami)
# # % pwd=$(ssh ${user}@operator-access-token.svc.ad1.ap-batam-1 'generate --mode password')
# # % sshpass -p "$pwd"  ssh -oProxyCommand="ssh -W %h:%p ${bastion}" -l${user} ${ip} set /SP/diag/snap
# # % sshpass -p "$pwd"  ssh -oProxyCommand="ssh -W %h:%p ${bastion}" -l${user} ${ip} set -script /SP/d
# ########################################################################################
```

You can copy the below code and run it as a bash script on your developer machine.

You will need to be on the OCNA VPN network and have permissions and access to the substrate bastion.

```
############### 3. ILOM snapshot collection (ALL Compute tray hosts)  ###############
#1. reload pkcs11
# $ pkill -9 ssh-agent;pkill -9 ssh-pkcs11-helper;ssh-add -s /usr/local/lib/opensc-pkcs11.so
# #put in yubikey PIN
# $ reload-ssh
# #put in yubikey PIN

#2. restart allproxy
# use whatever way you would like to start allproxy...
# e.g.
# $ source venv/bin/activate
# $ venv/bin/allproxy --use-ossh

#3. get ILOM IPs
# use either hops-cli -> host <rack_serial> -> info and parse ILOM IP's

# > host <rack_serial>
# get GB200 compute tray model string
# > info
# filter down to compute trays
# > filter model=GPU_GB200-NVL72_S.01

# obtain ILOM IP addresses for compute trays
# > network_info | awk '{print "\""$1 " " $7"\""}'
# copy and paste the output into ilom_ip_list variable as shown below.


# alternatively you can open the rack in C4PO -> go to HOST section below and add ILOM IP column

ilom_ip_list=(
 "2515XKG30G <ILOM_IP>"
 "2516XKG41V <ILOM_IP>"
 "2513XNG0EA <ILOM_IP>"
 "2513XNG0FH <ILOM_IP>"
 "2512XNG02D <ILOM_IP>"
 "2513XNG02H <ILOM_IP>"
 "2513XNG0DF <ILOM_IP>"
 "2513XNG0EJ <ILOM_IP>"
 "2513XNG0D9 <ILOM_IP>"
 "2513XNG0E5 <ILOM_IP>"
 "2513XNG0G4 <ILOM_IP>"
 "2513XNG0DB <ILOM_IP>"
 "2513XNG0D6 <ILOM_IP>"
 "2513XNG0C8 <ILOM_IP>"
 "2513XNG0DA <ILOM_IP>"
 "2513XNG0FV <ILOM_IP>"
 "2507XNG03U <ILOM_IP>"
 "2509XNG10X <ILOM_IP>"
)

#4 Run script:
getIlomSnapshot()
{
    local ilom_ip=$1
    local host_serial=$2
    local bastion=$3
    local rack_number=$4

    bastion=bastion-ad1.rb.ap-batam-1.oci.oracleiaas.com
    DATE=`date +%F-%H%M%S`

    if [[ -z "$ilom_ip" ]]; then
        echo "Error: ilom_ip parameter is empty or missing."
        return 1
    fi
    if [[ -z "$host_serial" ]]; then
        echo "Error: host_serial parameter is empty or missing."
        return 1
    fi
    if [[ -z "$bastion" ]]; then
        echo "Error: bastion parameter is empty or missing."
        return 1
    fi
    if [[ -z "$rack_number" ]]; then
        echo "Error: rack_number parameter is empty or missing."
        return 1
    fi

    user=$(whoami)
    pwd=$(ssh ${user}@operator-access-token.svc.ad1.ap-batam-1 'generate --mode password')

    echo "user: $user"
    echo "password: $pwd"

    ilom_snapshot_directory=${rack_number}/${rack_number}-ilom-snapshots
    mkdir -p ${ilom_snapshot_directory}

    echo "collecting ILOM snapshot for $host_serial at $ilom_ip"
    sshpass -p "$pwd" ssh -oStrictHostKeyChecking=no -oProxyCommand="ssh -W %h:%p ${bastion}" -l${user}
    sshpass -p "$pwd" ssh -oProxyCommand="ssh -W %h:%p ${bastion}" -l${user} ${ilom_ip} set -script /SP
    # set dataset back to a default value?
    # dataset : Possible values = normal, normal-logonly, fruid, fruid-logonly, full, full-logonly, ser
    # sshpass -p "$pwd" ssh -oProxyCommand="ssh -W %h:%p ${bastion}" -l${user} ${ip} set /SP/diag/snaps
}

# virt node or bastion that can access substrate/ILOM
bastion=bastion-ad1.rb.ap-batam-1.oci.oracleiaas.com
rack_number=

for entry in "${ilom_ip_list[@]}"; do
    serial=$(echo "$entry" | awk '{print $1}')
    ip=$(echo "$entry" | awk '{print $2}')
    echo "getting snapshot for $serial $ip..."
    getIlomSnapshot "$ip" "$serial" "$bastion" "$rack_number"
done
```

### How to Retrieve SOSReport:

Link to SOSReport Documentation on Confluence:

Sometimes this is needed by other teams (VTS, RHS, OHD, etc.) in order to help qualify a host for RMA.

SOSReport *should* be installed on the CPV OS Image for compute trays by default.

```
# From the CPV bastion:

# 1. ssh to the GPU compute tray host.
$ ssh ubuntu@<compute_tray_ip>

# 2. Run SOSReport. This may take a few minutes.
$ sudo sosreport

# 3. Upload the log to Object Storage and/or SCP back to your developer machine.
# from CPV bastion, scp the log from the compute tray back to the bastion.
# Note: the file is named slightly differently for each time you run.
$ scp ubuntu@<compute_tray_ip>:/var/tmp/sosreport-bur8r2403-s06u34-2022-04-13-zbvabpj.tar.xz .
```

### How to Update the Firmware of a GB200 Compute Tray(s) or NVswitch Tray(s):

**Note: Please engage CPV Tier 2 oncalls if a partial rack has split firmware versions. A partial rack is a rack that contains hosts that are in use by the customer and are in the customer pool(s), and some hosts in OCI/CPV. If you have a partial rack, we will likely need to engage and coordinate with SCE team to identify a plan to coordinate any firmware upgrades.**

1. Please identify the current firmware versions the host/rack have and cross check with the expected version(s) using the resources below.

- **How to Find Rack Firmware Version**
  - Follow these directions to identify the current firmware version that is installed the host/rack.
- **HOPS Rack/Host Firmware Pinning Casper Config**
  - Use the latest version of this Casper Config file to identify which firmware bundle the host/rack is pinned to.
- **HOPS Pinning Golden Set Config Source of Truth**
  - Use the latest version of this config file to identify which firmware bundle has which firmware version based on the firmware version strings you find.

2. If there is a mismatch or if you would like to update the firmware version a host/rack is pinned to, please create a BMP ticket to request HOPS team to pin the host/rack to the desired firmware version.

| Links | Notes |
|---|---|
| Link to BMP ticket queue | Example BMP ticket to request firmware pin |

3. Once the BMP ticket is closed and HOPS team has given confirmation that the pinning configuration has been updated and deployed, **please check the rack to make sure there are no customer instances and hosts are not in the customer pool(s).**

4. If there are, please engage Strategic Customer Engagement Team (SCE) to get approval from the customer and coordinate the recycle.

5. If the whole rack is with OCI and/or in CPV `holding` pool or you have SCE team and customer approval, please terminate all instances and GMCs on this rack. This will trigger the whole rack to be recycled and reprovisioned by going through HOPS to pick up the latest firmware version. This may take a few hours. Please reach out to CPV Tier 2 triage team(s) or CPV primary/secondary oncalls at `#cpv-triage` or create an HPC ticket. **It is important that all GMCs on the rack need to be terminated in order for the NVSwitch trays to be recycled correctly.** If there are any issues, please reach out to CPV team oncalls and/or GPUCP team oncalls if the issue is due to NVSwitch recycling not triggering.

### FDT Test IDS

| List of FDT Test IDs |
|---|
| Checkinforom |
| environmentcheck |
| Inventory |
| TegraCpu |
| TegraCpu4 |
| TegraCpu5 |
| TegraMemory |
| CpuMemorySweep |
| TegraClink |
| Gpustress |
| Gpumem |
| Pcie |
| C2C |
| CPUVDD_PowerStress |
| CpuGpuConstPower |
| Connectivity |
| NvlBwStress |
| NvlBwStressBg610 |
| NvlBwStressBg610Pulsy |
| DimmStress |
| CpuGpuSyncPulsePower |
| ThermalSteadyState |
| SyslogErrorCheck |
| KernLogErrorCheck |
| DmesgLogErrorCheck |
| SyslogAERCheck |
| KernLogAERCheck |
| DmesgLogAERCheck |

Grab the MODS code from step 2 for the failed test. Look at the `fieldiag_summary.log` or other `fieldiag` log files to get the details of the error. Most importantly, please identify the start and end timestamps of the failed test in the log breakdown.

`/local/FDT` will be generated by the CPV FDT test automations/scripts. This should remain after the FDT test has finished running. However, this path will disappear upon power cycling the host. If the last test CPV ran on the host is an FDT test (exclude pretest(s)) and the file path is not there, you can follow these steps:

Please refer to FDT Test IDs to view all of the available test IDs in the FDT test suite.

```
# find mountable SSD device.
# path may be /dev/md0 or have a different number appended to it.
$ ls /dev | grep -i md
```

```
md0

# mount the SSD device to filepath.
$ sudo mount /dev/md0 /local

$ ls /local
FDT

$ cd /local/FDT
# Then proceed to look for the FDT test results/test logs.
```

### FDT Test Log File Paths

You can use the below as a reference for where to find more detailed logs per FDT test.

Example file paths:

#### Single Node FDT Test Log File Paths

```
/local/FDT/629.../dgx/logs-<yyyymmdd>-<hhmmss>/<fdt_test_id>/<GPU_PCI_ADDR_GPU_SN>/fieldiag.log
```

```
/local/FDT/629.../dgx/logs-<yyyymmdd>-<hhmmss>/<fdt_test_id>/<GPU_PCI_ADDR_GPU_SN>/fieldiag.jso
```

An example of a full test log file path is:

```
/local/FDT/629-24975-0000-FLD-50447-rev2/dgx/logs-20251016-071032/Checkinforom/0008_01_00.0_SN_16502250
```

#### Multi Node FDT Test Log File Paths

```
/local/FDT/629.../dgx/logs-<yyyymmdd>-<hhmmss>/COMPUTE_NODE_X/connectivity/<GPU_PCI_ADDR_GPU_SN>/fieldia
```

```
/local/FDT/629.../dgx/logs-<yyyymmdd>-<hhmmss>/COMPUTE_NODE_X/connectivity/<GPU_PCI_ADDR_GPU_SN>/fieldia
```

Example full path:

```
/local/FDT/629-24972-4975-FLD-50448-rev3/dgx/logs-20251013-202824/COMPUTE_NODE_10/Connectivity/0008:01:
```

The above file path will contain start and end timestamps for the `connectivity` test shown on in the overall FDT summary from the `run.log`.

```
$ oci session authenticate --profile DEFAULT --tenancy-name bmc_operator_access --region r1

$ oci os object list -ns hpc -bn Debug --auth security_token --profile DEFAULT --prefix 2425XLG0BG

# Look for any logs you might want to look through. You can download them by running the below:

$ oci os object get -ns hpc -bn Debug --auth security_token --profile DEFAULT --name /name/of/object/st
```

Pulling FDT test logs from OCI CLI:

```
./logs-20251013-152625/<fdt_test_id>/*/fieldiag.log
```

```
./logs-20251013-152625/<fdt_test_id>/*/fieldiag.jso
```

```
/logs-20251013-152625/FDT/629.../dgx/logs-<yyyymmdd>-<hhmmss>/COMPUTE_NODE_X/connectivity/<GPU_PCI_ADDR
```

```
/logs-20251013-152625/FDT/629.../dgx/logs-<yyyymmdd>-<hhmmss>/COMPUTE_NODE_X/connectivity/<GPU_PCI_ADDR
```

### How to get GPU Information for GB200 Compute Trays

#### This includes: - GPU PCI ID - Serial Number - Chassis Serial Number - Module ID - GPU Fabric GUID

It may be good and comprehensive to run this command on each compute tray on the rack for rack-level troubleshooting.

```
$ nvidia-smi -q | egrep "GPU 000|Module Id|Serial Number|GPU Fabric GUID"
```

You can use `clush` or `xargs` to run this on all compute trays that are accessible from the CPV Burn-in orchestrator bastion host.

```
# 1. Identify the Rack Number you are working on/
# you can use:
# - C4PO
# - cpv state read -t host -k <host_serial> | jq -r .[].rackNumber
# hops-cli command -> load host -> run 'info' command

# 2. Obtain IP addresses of the hosts that are in CPV:
$ cpv state overview-count -f RackNumber=<rack_number> | grep -i "10." | awk '{print $2}' | grep -i "10

# 3. Use `clush` or `xargs` to iterate through all host IP addresses on the rack:
$ clush -l ubuntu -w <host1_ip>,<host2_ip> "nvidia-smi -q | egrep \"GPU 000|Module Id|Serial Number|GPU
GPU 00000008:01:00.0
    Serial Number                         : 1760125301973
        Chassis Serial Number             : 1820125181020
        Module Id                         : 2
        GPU Fabric GUID                   : 0x9cd49482d5262392
GPU 00000009:01:00.0
    Serial Number                         : 1760125301973
        Chassis Serial Number             : 1820125181020
        Module Id                         : 1
        GPU Fabric GUID                   : 0x8c2c19528bf876ac
GPU 00000018:01:00.0
    Serial Number                         : 1650325014536
        Chassis Serial Number             : 1820125181020
        Module Id                         : 4
        GPU Fabric GUID                   : 0xd54716fe66366f36
GPU 00000019:01:00.0
    Serial Number                         : 1650325014536
        Chassis Serial Number             : 1820125181020
        Module Id                         : 3
        GPU Fabric GUID                   : 0x6617182fe21c9f5c
```

**Additional Notes:**

- Module Id is the physical slot number for the GPU which corresponds to which cable cartridge - e.g. Module Id 2 is Cable Cartridge 2
- Serial Number will tell you which Bianca Module Serial Number - handy if it comes down to RMA Bianca Module

You can turn this into a table for each host containing the information and use it to reference for other directions/steps in this runbook.

FDT NVLink Runbook Diagram

If you get confused, please refer to the below NV Bug ID for how to approach this situation. You may need to reach out to someone with an NVOnline account and ask if they would be able to give you a summary of the NVBug and rundown if you do not currently have access to the NVOnline/partner website.

**NVBug: 5573659**

If you would like to improve and/or otherwise modify the content of the diagram, please reach out to CPV team for access to the Visio diagram.

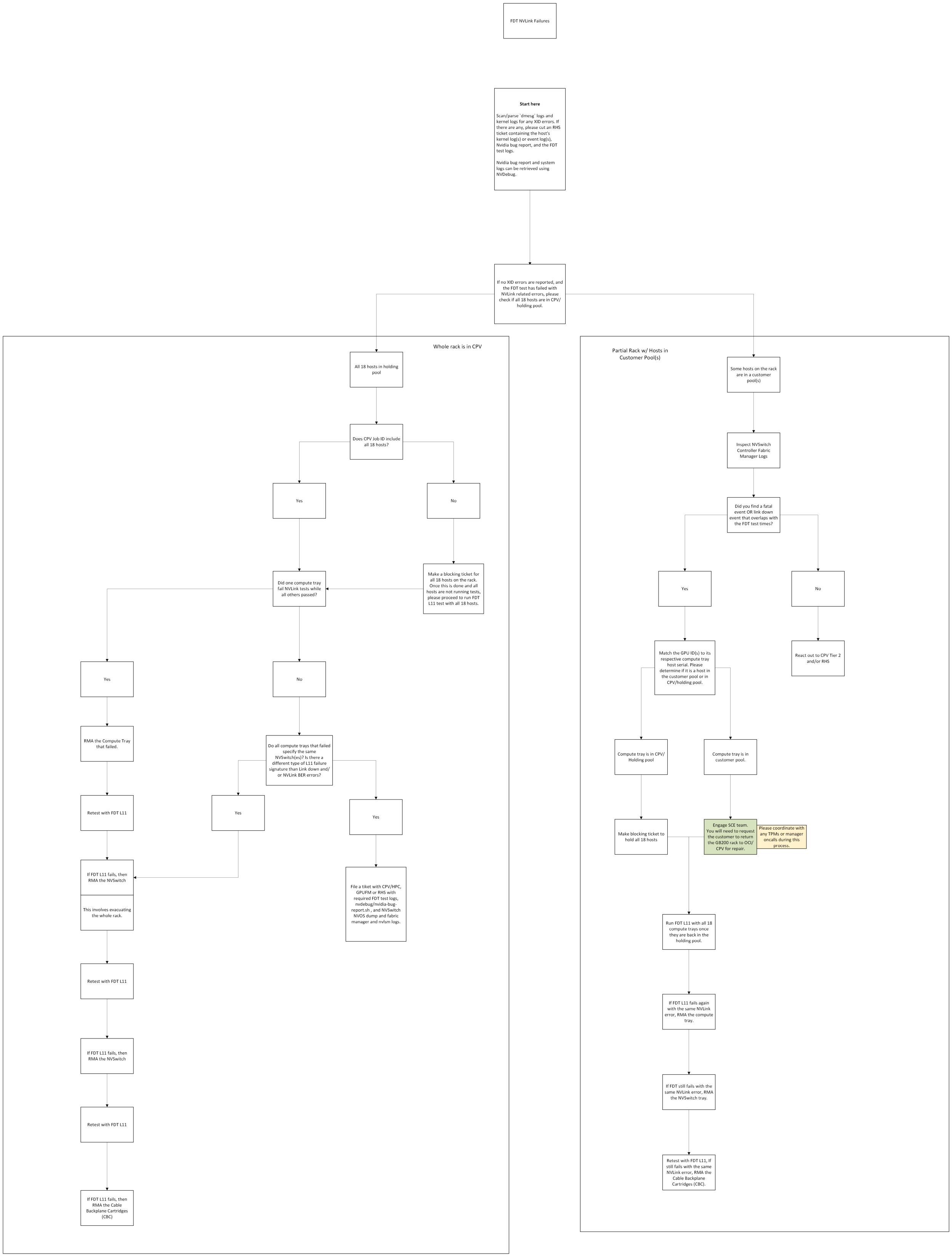
FDT NVLink Failures
Start here
Scan/parse `dmesg` logs and kernel logs for any XID errors. If there are any, please cut an RHS ticket containing the host's kernel log(s) or event log(s), Nvidia bug report, and the FDT test logs.
Nvidia bug report and system logs can be retrieved using NVDebug.
If no XID errors are reported, and the FDT test has failed with NVLink related errors, please check if all 18 hosts are in CPV/ holding pool.
Whole rack is in CPV
All 18 hosts in holding pool
Does CPV Job ID include all 18 hosts?
Yes
No
Make a blocking ticket for all 18 hosts on the rack. Once this is done and all hosts are not running tests, please proceed to run FDT L11 test with all 18 hosts.
Did one compute tray fail NVLink tests while all others passed?
Yes
No
RMA the Compute Tray that failed.
Do all compute trays that failed specify the same NVSwitch(es)? Is there a different type of L11 failure signature than Link down and/ or NVLink BER errors?
Retest with FDT L11
Yes
Yes
If FDT L11 fails, then RMA the NVSwitch
This involves evacuating the whole rack.
File a tiket with CPV/HPC, GPUFM or RHS with required FDT test logs, nvdebug/nvidia-bug-report.sh , and NVSwitch NVOS dump and fabric manager and nvlsm logs.
Retest with FDT L11
If FDT L11 fails, then RMA the NVSwitch
Retest with FDT L11
If FDT L11 fails, then RMA the Cable Backplane Cartridges (CBC)
Partial Rack w/ Hosts in Customer Pool(s)
Some hosts on the rack are in a customer pool(s)
Inspect NVSwitch Controller Fabric Manager Logs
Did you find a fatal event OR link down event that overlaps with the FDT test times?
Yes
No
Match the GPU ID(s) to its respective compute tray host serial. Please determine if it is a host in the customer pool or in CPV/holding pool.
React out to CPV Tier 2 and/or RHS
Compute tray is in CPV/ Holding pool
Compute tray is in customer pool.
Make blocking ticket to hold all 18 hosts
Engage SCE team. You will need to request the customer to return the GB200 rack to OCI/ CPV for repair.
Please coordinate with any TPMs or manager oncalls during this process.
Run FDT L11 with all 18 compute trays once they are back in the holding pool.
If FDT L11 fails again with the same NVLink error, RMA the compute tray.
If FDT still fails with the same NVLink error, RMA the NVSwitch tray.
Retest with FDT L11, If still fails with the same NVLink error, RMA the Cable Backplane Cartridges (CBC).

Figure 1: CPV FDT NVLink Runbook Diagram