



Nvidia OEM Network Product Update

2025 Q1

Agenda

- BFB Bundle Images and Update

- BF3 Portfolio Update – Cold aisle

- BF3 Hardware Update – Power supply

- BF3 NVQual Update

- BF3 Firmware Update – NC-SI PKG ID Change

- BF3 Host Management – Privilege and Host Management

- BF3 Reset Control

Agenda

- **BFB Bundle Images and Update**

- BF3 Portfolio Update – Cold aisle

- BF3 Hardware Update – Power supply

- BF3 NVQual Update

- BF3 Firmware Update – NC-SI PKG ID Change

- BF3 Host Management – Privilege and Host Management

- BF3 Reset Control

BFB – BFB Bundle Images

- BlueField Image Types - [Types and Methods of Updating BlueField Software Image - NVIDIA Docs](#)
- BFB – BlueField Bundle images – a proprietary format used to update and recover the BlueField device. There are two types of BFB images available:
 - BF-Bundle – Includes BlueField firmware components, Arm OS, and DOCA. This bundle contains everything needed to update all possible components of a BlueField device.
 - BF-FW-Bundle – Includes only the BlueField firmware components and excludes Arm OS and DOCA. This is typically used for day-2 operations or by customers working in NIC mode who do not require Arm OS or DOCA running on the device.
- ISO – Similar to BF-Bundle, this format includes BlueField firmware components, Arm OS, and DOCA packages in an ISO standard format. The BlueField ISO image is based on the standard Ubuntu ISO image for Arm64, but with an updated kernel and added DOCA packages. PXE booting the BlueField device with the ISO image results in the installation of the Arm OS, including DOCA, and the update of BlueField firmware components.
- PLDM – Similar to BF-FW-Bundle, this format includes only BlueField firmware components and does not include Arm OS or DOCA. The image is distributed per SKU to keep the image size small, as required by the limitations of some platform BMCs.
- Repository – An online repository used for updating with standard Linux tools. This method updates DOCA from NVIDIA's repository and Arm OS packages from Canonical's updates repository. BlueField firmware components are also updated. Post-installation steps are required to upgrade BlueField firmware components (i.e., ATF/UEFI, NIC firmware, and BMC components).

[NVIDIA DOCA 2.8.0 Downloads | NVIDIA Developer](#)

NVIDIA DOCA 2.8.0 Downloads

Select

Click on the green buttons that describe your target platform. Only supported platforms will be shown. By downloading and using the software, you agree to fully comply with the terms and conditions of the [DOCA EULA](#).

Host or BlueField	Host-Server	BlueField	
Deployment Package	BF-Bundle	BF-FW-Bundle	BF-BMC
Distribution	Ubuntu		
Version	22.04		
Installer Type	BFB	ISO	

[Software Installation and Upgrade](#)

Info

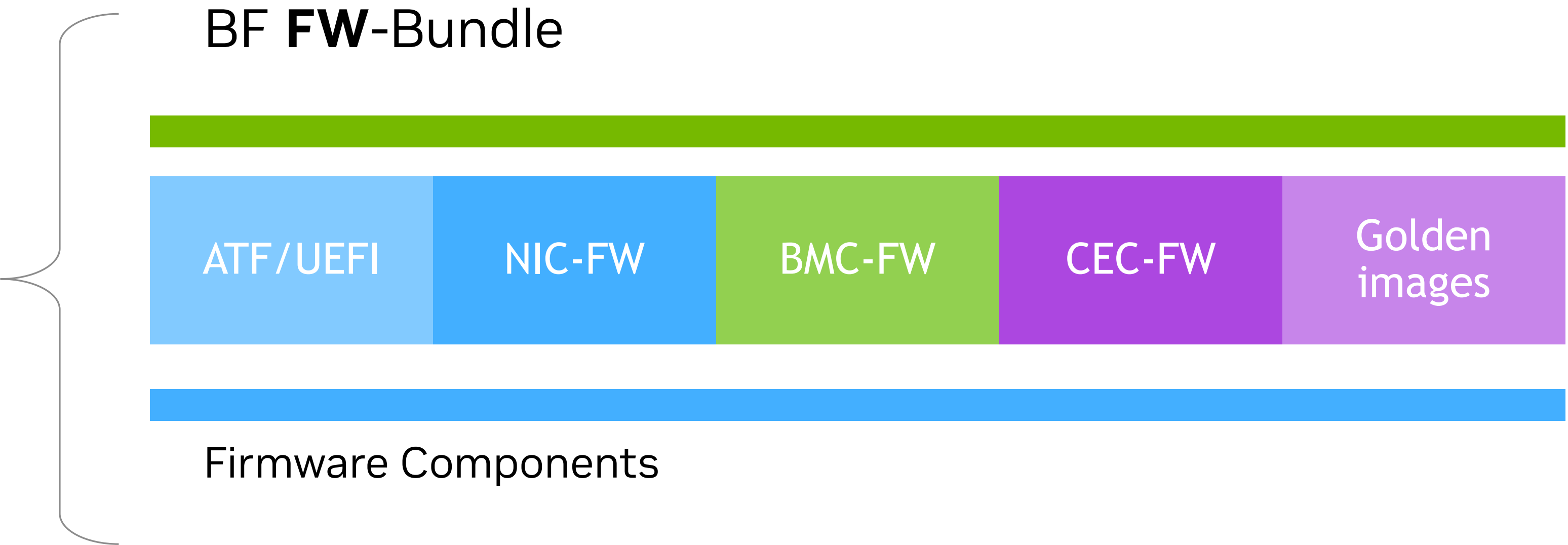
The minimum BMC Firmware version that supports this method of BMC upgrade from BFB image, is 23.07. If your BMC firmware version is lower, follow the [NVIDIA BlueField BMC Software](#) documentation to upgrade BMC firmware. The BMC version can be obtained by following instructions [here](#).

BF-Bundle Vs. BF-FW Bundle

***new* Firmware-only Bundle**

- “light” in size
- Used for day2 provisioning w/o impacting installed ARM-OS.

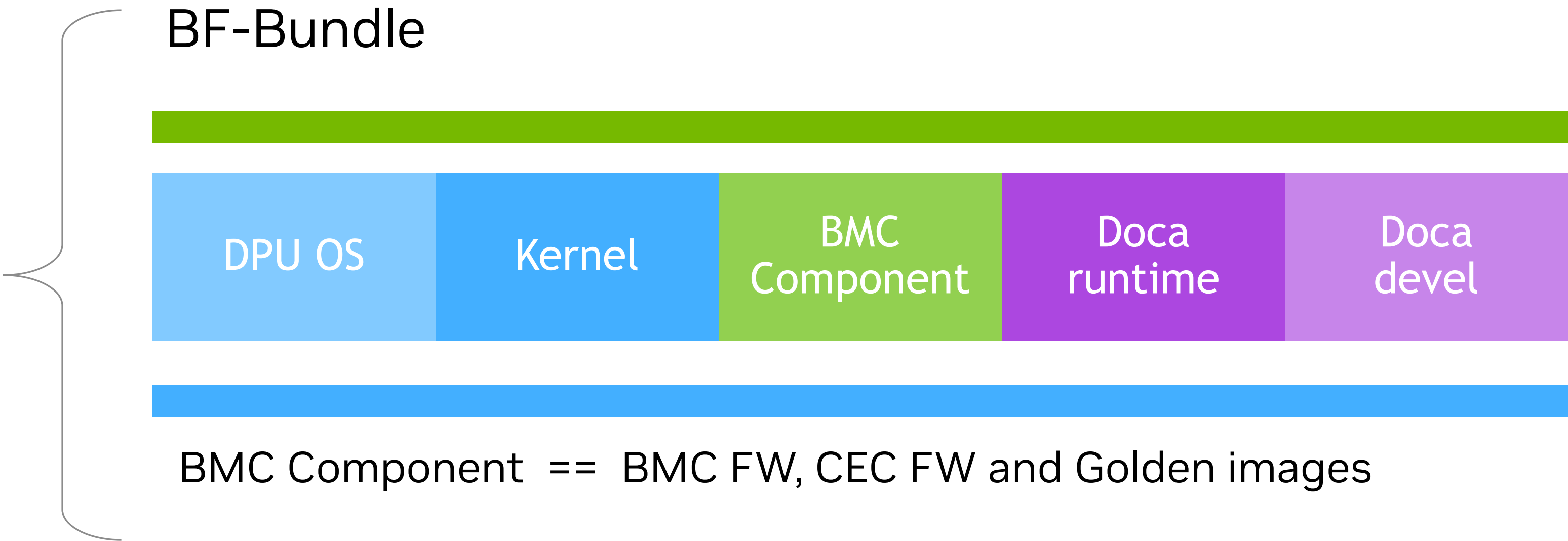
NOTE: requires customers to upgrade packages separately (e.g., via Linux standard tools (yam/apt upgrade)).



VS.

BF-Bundle (full BFB)

- Installs everything / (OS, Kernel, pre-built binaries).
- **Purpose:** Production Line / Recovery



BFB – PLDM Firmware Update

- Changes and New Features in 4.9.1
- Upgrading BlueField Software Components Using PLDM
- **Warning**
 - PLDM firmware update is possible only if the currently running version (i.e., the version to update from) is DOCA 2.9.0/BSP 4.9.0 or higher.

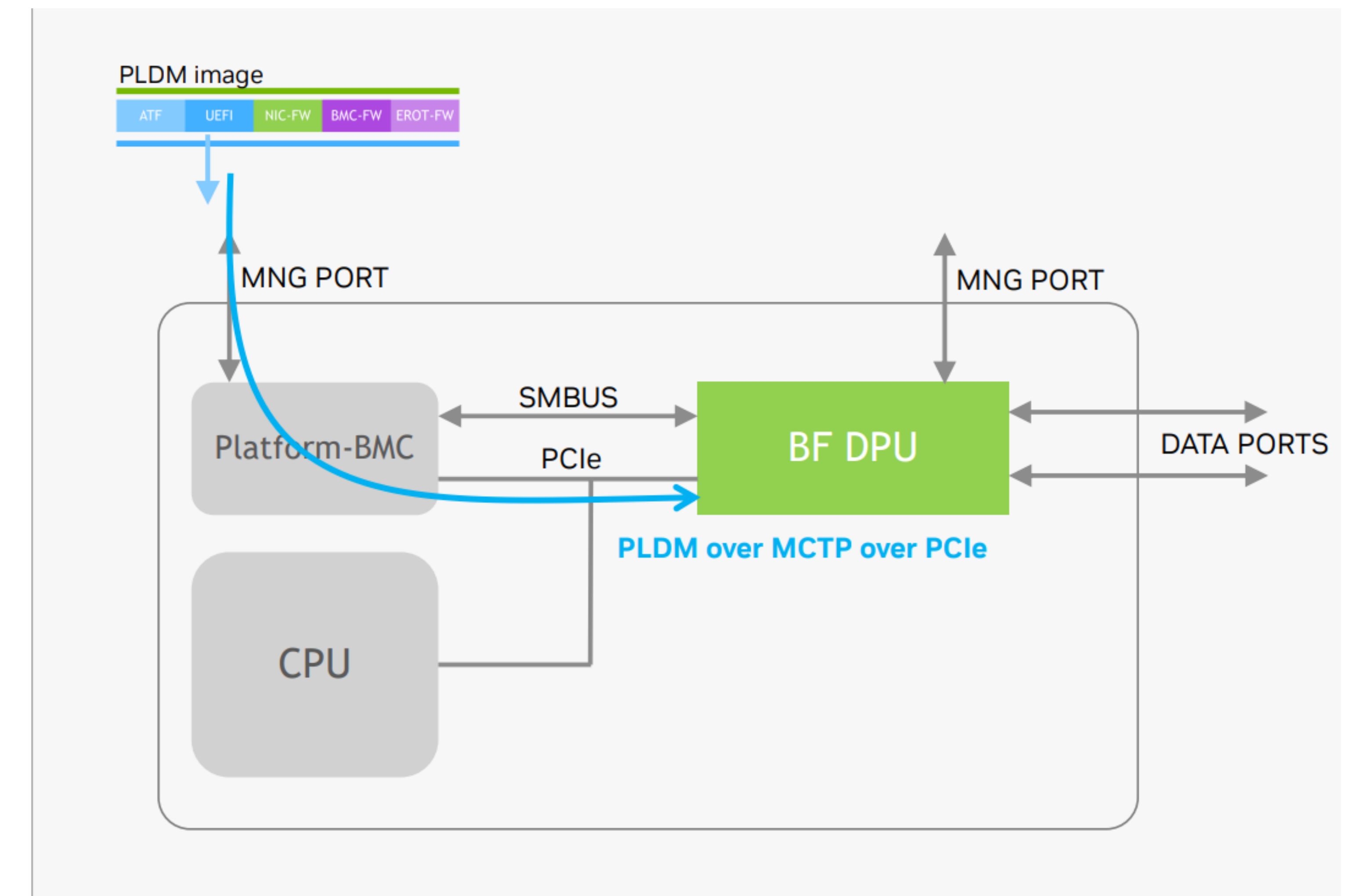
Changes and New Features in 4.9.1

- **PLDM firmware update support for BlueField-3 devices**

Added support for PLDM firmware update – BlueField-3 devices now support firmware updates via the PLDM type-5 protocol over MCTP over PCIe, allowing future updates to be performed through the platform's BMC. This implementation complies with PLDM specification version 1.0. Note that the firmware image size is approximately 100MB.

Known limitations:

 - With this release, firmware updates using the PLDM flow can only be applied with a full-system power cycle. Support for applying updates via server reboot will be added in a future release.
 - When operating in DPU mode, administrators must manually configure BMC credentials in a local Arm OS config file to enable BMC and CEC updates. Provisioning of credentials will be automated in a future release.
- Enhanced firmware reset flow for Sync1 utilizing community-accepted hot reset kernel flow
- Added logging of NVMe/eMMC wipe operations to RShim log
- Bug fixes



BFB – BF Bundle Upgrade Procedure

- mlx5_core Driver Prints Fatal Error During BFB Installation

- During BFB image installation, the mlx5_core driver prints fatal messages on the x86 host. T as shown below. This behavior started in 4.7.0/2.7.0 where the BFB image also updates NIC firmware the BMC software.

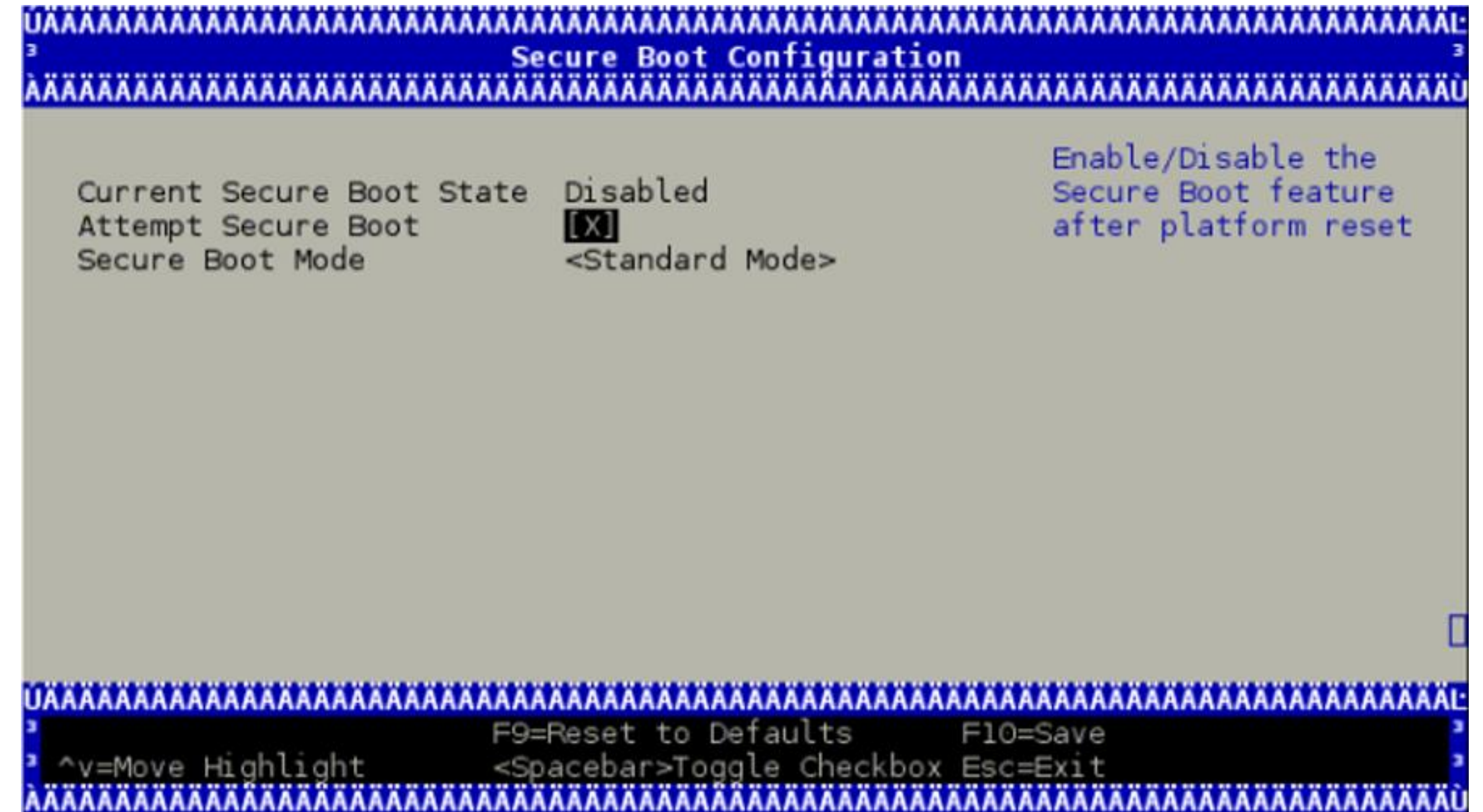
How to Proceed

- This behavior is expected. The installation should be followed by a power cycle to recover and activate the NIC firmware and the BMC software.

```
[Tue Apr 9 06:59:36 2024] mlx5_core 0000:03:00.0: poll_health:1037:(pid 0): Fatal error 3 detected
[Tue Apr 9 06:59:36 2024] mlx5_core 0000:03:00.1: poll_health:1037:(pid 0): Fatal error 3 detected
[Tue Apr 9 06:59:36 2024] mlx5_core 0000:03:00.1: mlx5_health_try_recover:339:(pid 2778): handling bad device here
[Tue Apr 9 06:59:36 2024] mlx5_core 0000:03:00.1: mlx5_handle_bad_state:290:(pid 2778): starting teardown
[Tue Apr 9 06:59:36 2024] mlx5_core 0000:03:00.1: mlx5_error_sw_reset:241:(pid 2778): start
[Tue Apr 9 06:59:36 2024] mlx5_core 0000:03:00.1: mlx5_error_sw_reset:274:(pid 2778): end
[Tue Apr 9 06:59:39 2024] mlx5_core 0000:03:00.1: E-Switch: Disable: mode(LEGACY), nvfs(0), necvfs(0), active vports(0)
[Tue Apr 9 06:59:39 2024] mlx5_core 0000:03:00.1: mlx5_wait_for_pages:916:(pid 2778): Skipping wait for vf pages stage
[Tue Apr 9 06:59:39 2024] mlx5_core 0000:03:00.1: mlx5_wait_for_pages:916:(pid 2778): Skipping wait for vf pages stage
[Tue Apr 9 06:59:47 2024] mlx5_core 0000:03:00.0: mlx5_health_try_recover:339:(pid 5577): handling bad device here
[Tue Apr 9 06:59:47 2024] mlx5_core 0000:03:00.0: mlx5_handle_bad_state:290:(pid 5577): starting teardown
[Tue Apr 9 06:59:47 2024] mlx5_core 0000:03:00.0: mlx5_error_sw_reset:241:(pid 5577): start
[Tue Apr 9 06:59:47 2024] mlx5_core 0000:03:00.0: mlx5_error_sw_reset:274:(pid 5577): end
[Tue Apr 9 06:59:47 2024] mlx5_core 0000:03:00.0: E-Switch: Disable: mode(LEGACY), nvfs(0), necvfs(0), active vports(0)
[Tue Apr 9 06:59:47 2024] mlx5_core 0000:03:00.0: mlx5_wait_for_pages:916:(pid 5577): Skipping wait for vf pages stage
[Tue Apr 9 06:59:47 2024] mlx5_core 0000:03:00.0: mlx5_wait_for_pages:916:(pid 5577): Skipping wait for vf pages stage
[Tue Apr 9 06:59:49 2024] mlx5_core 0000:03:00.1: mlx5_health_try_recover:345:(pid 2778): starting health recovery flow
[Tue Apr 9 06:59:49 2024] mlx5_core 0000:03:00.1: mlx5_pci_slot_reset Device state = 2 pci_status: 0. Enter
[Tue Apr 9 06:59:49 2024] mlx5_core 0000:03:00.1: wait vital counter value 0x21100 after 1 iterations
[Tue Apr 9 06:59:49 2024] mlx5_core 0000:03:00.1: mlx5_pci_slot_reset Device state = 2 pci_status: 1. Exit, err = 0, result = 5, recovered
```


BFB – Disable Secure Boot

- Disable Secure Boot
- [UEFI Secure Boot - NVIDIA Docs](#)
 - Method#1 – UEFI menu
 - Method#2 – Redfish API by on-board BMC



```
curl -k -u root:<password> -H "Content-Type: application/octet-stream" -X GET https://<BF-BMC-IP>/redfish/v1/Systems/Bluefield/SecureBoot
{
  "@odata.id": "/redfish/v1/Systems/Bluefield/SecureBoot",
  "@odata.type": "#SecureBoot.v1_1_0.SecureBoot",
  "Description": "The UEFI Secure Boot associated with this system.",
  "Id": "SecureBoot",
  "Name": "UEFI Secure Boot",
  "SecureBootCurrentBoot": "Enabled",
  "SecureBootEnable": true,
  "SecureBootMode": "SetupMode"
}
curl -k -u root:<BF-BMC-PASSWORD> -X PATCH https://<BF-BMC-IP>/redfish/v1/Systems/Bluefield/SecureBoot -H 'Content-Type: application/json' -d
'{"SecureBootEnable": false}'
```


BFB – Rshim Owner Ship

- [SoC Management Interface - NVIDIA Docs](#)
- The RShim interface may be owned by the BlueField BMC or the host (Windows or Linux). In situations where users do not have access to the host, they would want to transfer RShim ownership to the BMC.
- [Known Issues - NVIDIA Docs](#)

4129718	<p>Description: If the path of installation of the bfb image (i.e., <code>DPU_OS</code> in "Deploying BlueField Software Using BFB from BMC") is called and the RShim on the host is not connected, the BMC takes the RShim. If the RShim on the host is connected, calling this path returns an error.</p> <hr/> <p>Workaround: To reclaim RShim ownership to the host:</p> <p>a. Open the <code>rshim.conf</code> file:</p> <pre>vim /etc/rshim.conf</pre> <p>b. Uncomment the following line:</p> <pre>FORCE_MODE 1</pre> <p>c. Restart RShim:</p> <pre>systemctl restart rshim</pre> <hr/> <p>Discovered in version: 24.10</p>
---------	--

Agenda

- BFB Bundle Images and Update

- **BF3 Portfolio Update – Cold aisle**

- BF3 Hardware Update – Power supply

- BF3 NVQual Update

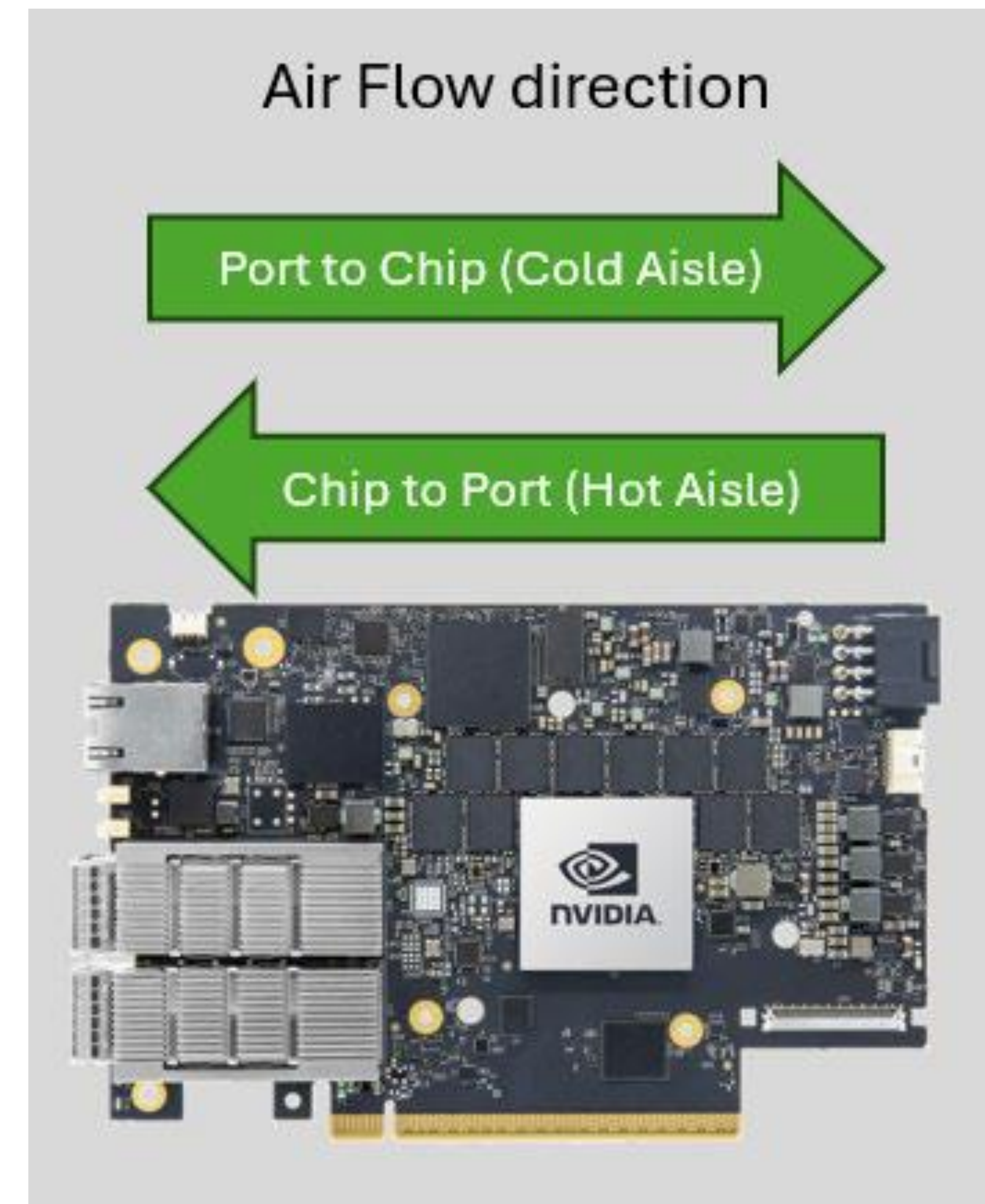
- BF3 Firmware Update – NC-SI PKG ID Change

- BF3 Host Management – Privilege and Host Management

- BF3 Reset Control

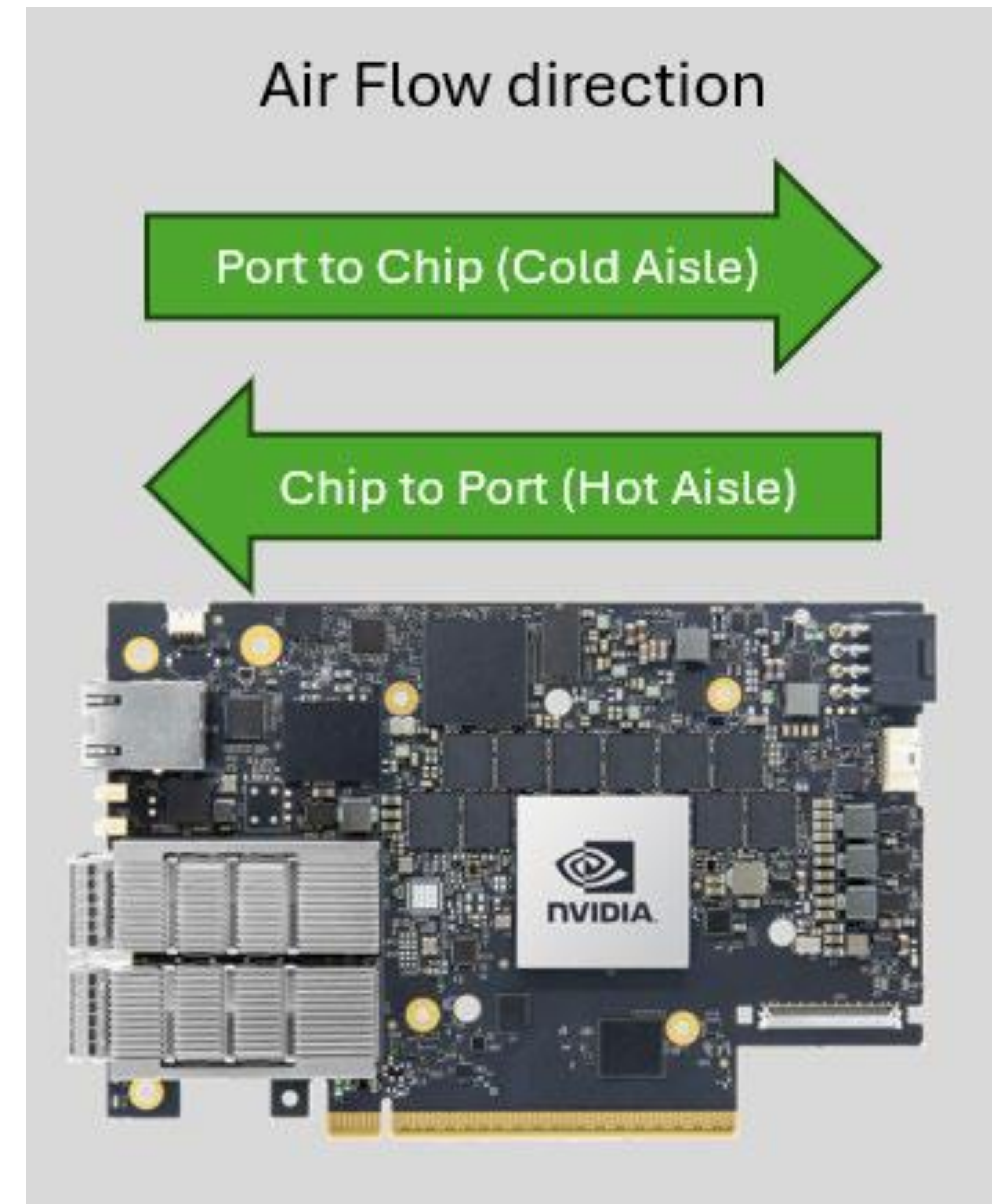
Cold Aisle Consideration

- Consult with NVIDIA Field Applications Engineering before starting Network NVQual thermal testing on BlueField-3 platforms in a cold aisle cooling environment.
- Some cases require additional measurement
- **1103274 NVIDIA BlueField-3 Networking Platform Product Specifications**
- B3240 Cold Aisle – OPN 900-9D3B6-00CN-PA0
Prototype now



Cold Aisle Consideration

- 确定BF3是否是在server的冷通道(Cold Aisle)位置, 如果是需要在BF3的电池上贴上热电偶, 测试电池的温度, 记录整个运行nvqual过程中的温度值, 并摘取最大值, 和nvqual log 和report时一起提交。
- BF3的电池温度不能超过60度



Agenda

- BFB Bundle Images and Update

- BF3 Portfolio Update – Cold aisle

- **BF3 Hardware Update – Power supply**

- BF3 NVQual Update

- BF3 Firmware Update – NC-SI PKG ID Change

- BF3 Host Management – Privilege and Host Management

- BF3 Reset Control

BFB – Prevent DPU from booting if ATX is missing

- [Release Notes Change Log History - NVIDIA Docs](#)
- [Changes and New Features in 4.8.0](#)
 - Introduced new behavior for 150W platforms where, to ensure proper operation of the BlueField, ATF will not boot BlueField-3 platforms if the ATX +12V is not connected.
- [Supported Interfaces - NVIDIA Docs](#)

External PCIe Power Supply Connector

✓ Note

Applies to following DPUs only. The external ATX power cable is not supplied with the DPU package; however, this is a standard cable usually available in servers.

- **B3220:** 900-9D3B6-00CV-AA0 and 900-9D3B6-00SV-AA0
- **B3240:** 900-9D3B6-00CN-AB0, 900-9D3B6-00SN-AB0 and 900-9D3B6-00CN-PA0
- **B3210:** 900-9D3B6-00CC-AA0 and 900-9D3B6-00SC-AA0
- **B3210E:** 900-9D3B6-00CC-EA0 and 900-9D3B6-00SC-EA0

To power up the above-mentioned DPUs, it is necessary to use a supplementary 8-pin ATX power cable. Since the power provided by the PCIe golden fingers is limited to **66W**, a total maximum of up to **150W** is enabled through the ATX 8-pin connector and the PCIe x16 golden fingers.

- In case customers in production with BF3 and they didn't plug in the ATX power
- Please make sure that these customers do not upgrade before they plug in the ATX power.
- How to detect presence of ATX cable w/o physical access to card
- Run from BF3 ARM
 - # modprobe mlxbf-ptm
 - # cat /sys/kernel/debug/mlxbf-ptm/monitors/status/atx_power_available

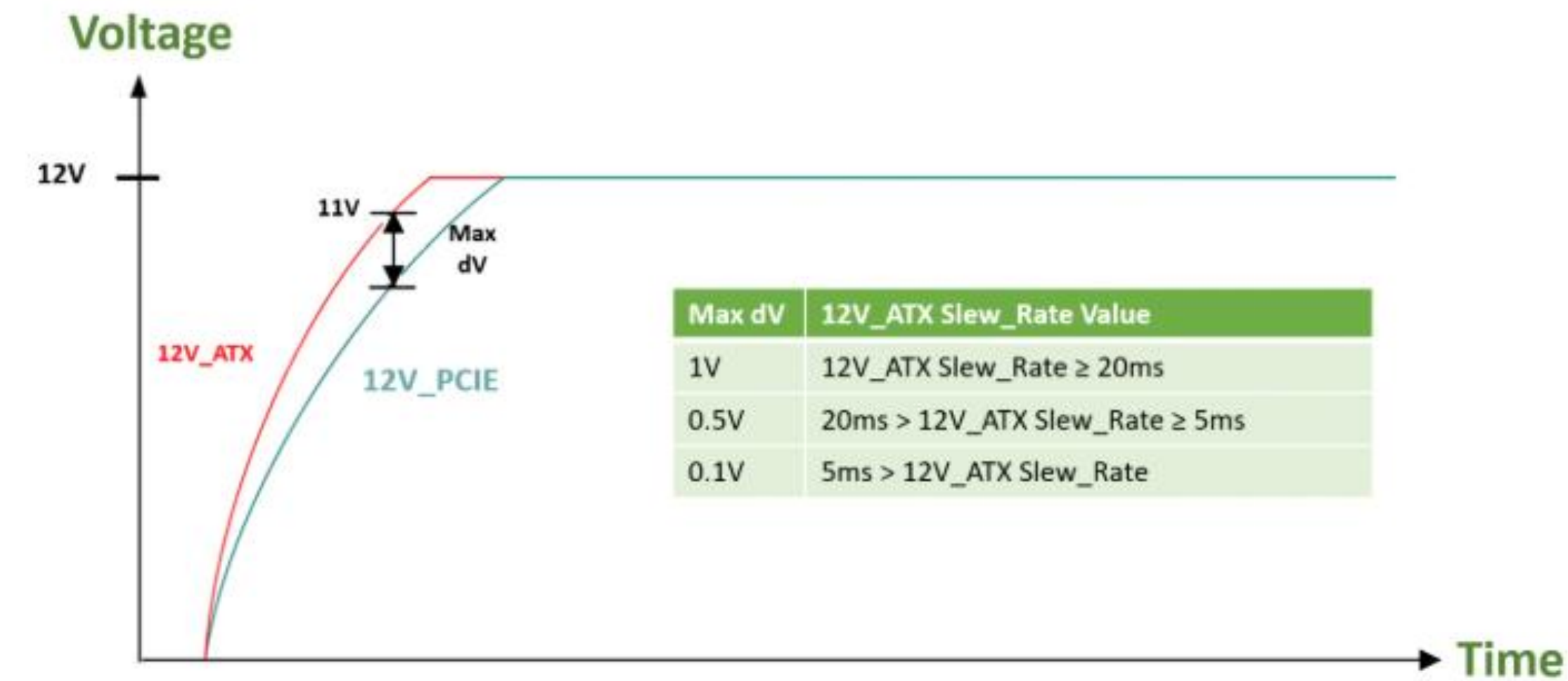
```
Nvidia BlueField-3 rev1 BL1 V1.0
NOTICE: Running as 9009D3B600SVAA system
NOTICE: BL2: v2.2(release):4.9.0-25-g0ce57e322
NOTICE: BL2: Built : 22:23:21, Oct 30 2024
NOTICE: BL2 built for hw (ver 2)
NOTICE: # Finished initializing DDR MSS0
NOTICE: # Finished initializing DDR MSS1
NOTICE: DDR POST passed.
NOTICE: BL31: v2.2(release):4.9.0-25-g0ce57e322
NOTICE: BL31: Built : 22:23:22, Oct 30 2024
NOTICE: BL31 built for hw (ver 2), lifecycle GA Secured
CRITICAL ERROR: ATX power not detected! Halting system!!
```


BFB – DPU Power Up and Power Down Sequence

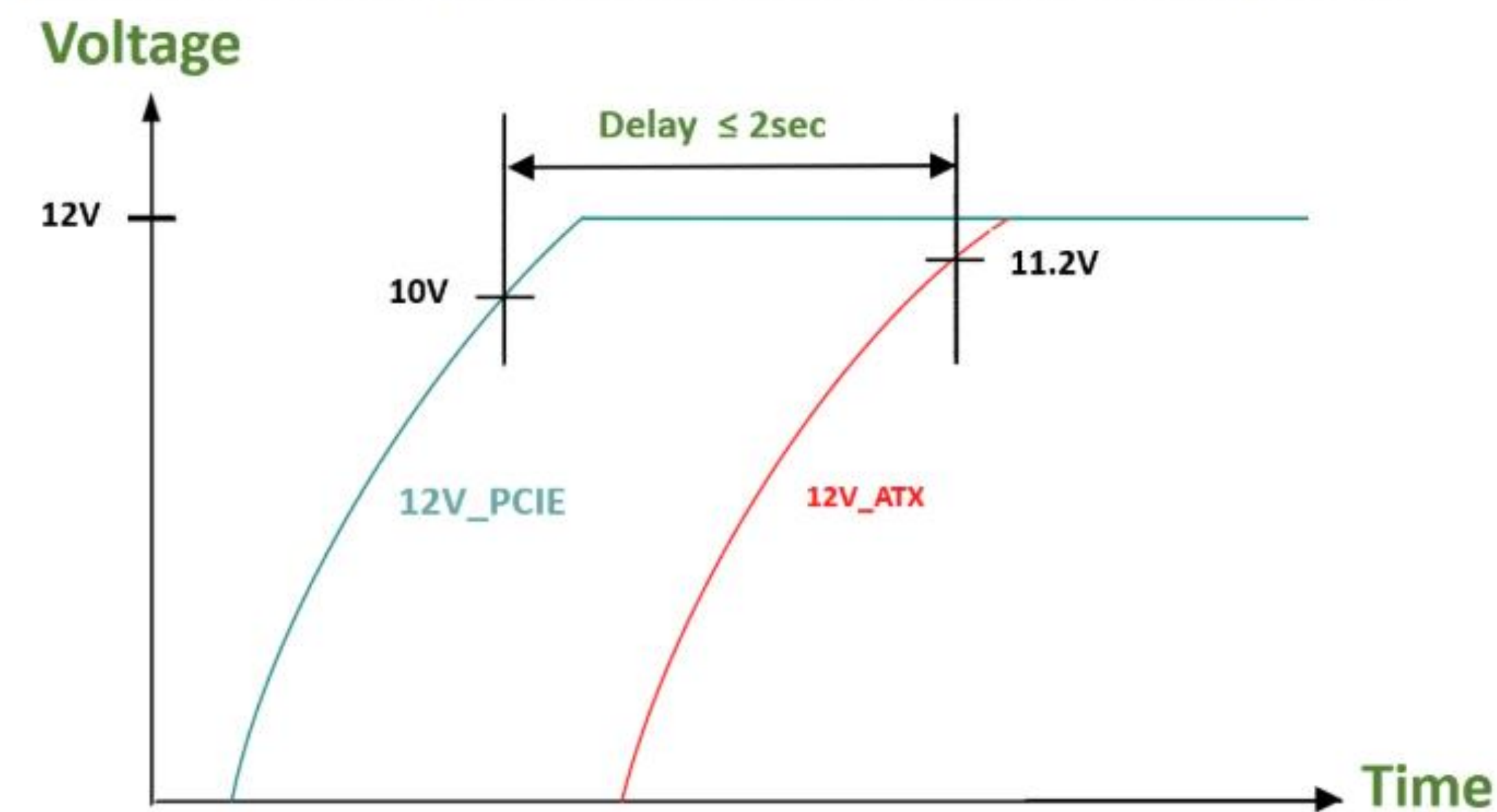
- DPU Power-Up and Power-Down Sequences

Power-Up Sequence

1. The 12V_ATX voltage can exceed the 12V_PCIE voltage by a maximum allowable voltage difference (dV) when the 12V_ATX reaches 11V. See below graph and table describing the dV between the 12V_ATX and 12V_PCIE voltages.



2. The 12V_ATX can be powered up after the 12V_PCIE, with a maximum delay of 2 seconds. The below graph illustrates the delay between the 12V_ATX and 12V_PCIE voltages at power-up.



DPU Power-Up and Power-Down Sequences

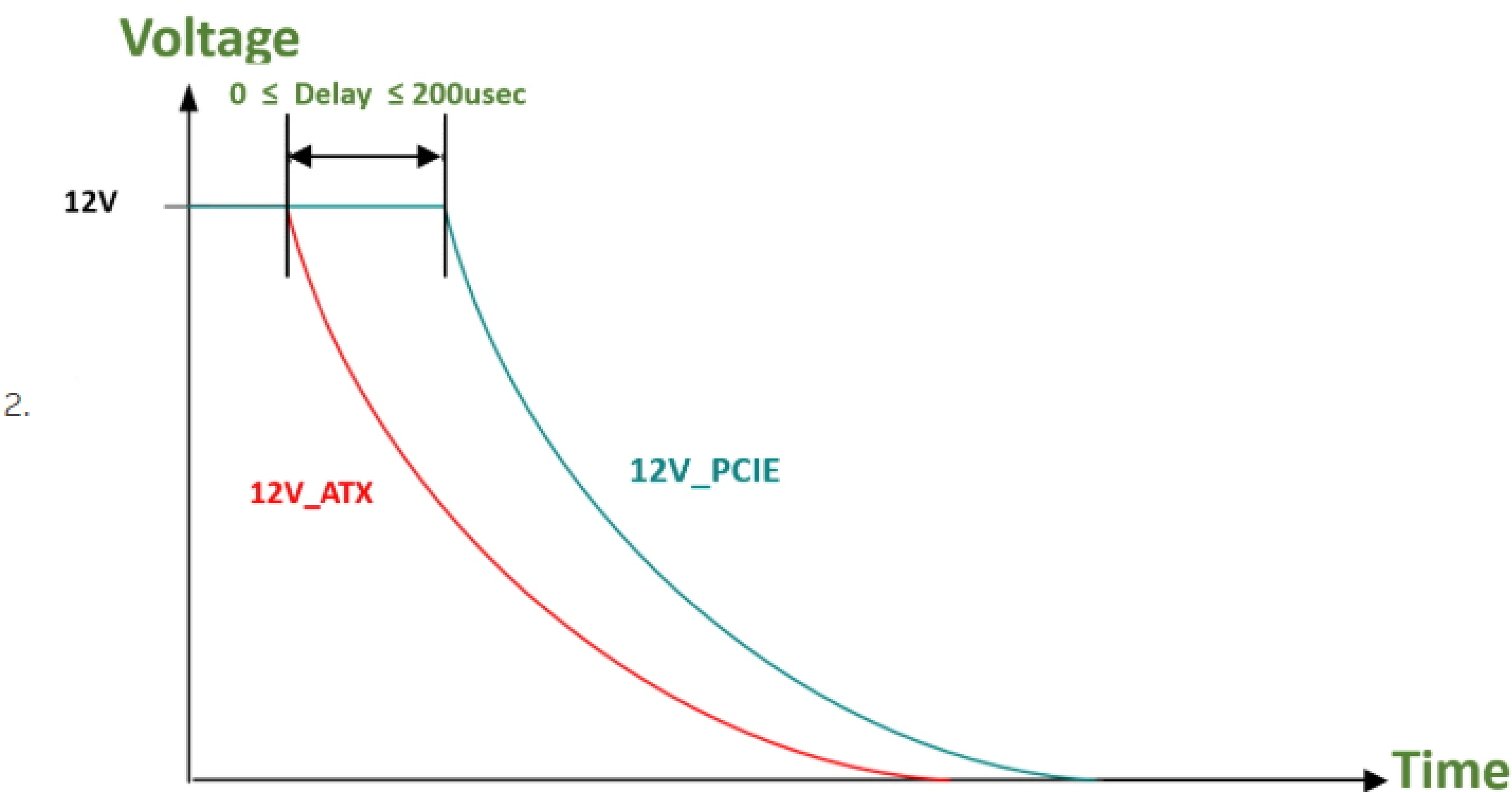
✓ Note

The power-up and power-down sequences listed below apply to DPUs with x16 PCIe extension option.

- B3220 Model:** 900-9D3B6-00CV-AA0 and 900-9D3B6-00SV-AA0
- B3240 Model:** 900-9D3B6-00CN-AB0 and 900-9D3B6-00SN-AB0
- B3210 Model:** 900-9D3B6-00CC-AA0 and 900-9D3B6-00SC-AA0
- B3210E Model:** 900-9D3B6-00CC-EA0 and 900-9D3B6-00SC-EA0

Power-Down Sequence

1. The 12V_PCIE voltage can be powered down simultaneously with the 12V_ATX voltage, or within a maximum delay of 200usec. The below graph illustrates the delay between the 12V_ATX and 12V_PCIE voltages at power-down.



3. The 12V_PCIE voltage must not be powered down while the 12V_ATX voltage is powered up.

DOCA 2.8 – PCN-001828

- The release addresses critical bugs and is therefore classified as a mandatory software upgrade for all BlueField-3 units.
- Details check PCN-001828



Agenda

- BFB Bundle Images and Update

- BF3 Portfolio Update – Cold aisle

- BF3 Hardware Update – Power supply

- **BF3 NVQual Update**

- BF3 Firmware Update – NC-SI PKG ID Change

- BF3 Host Management – Privilege and Host Management

- BF3 Reset Control

Network Adapter Leverage

- Recommended to run NVQual with the most power and thermal-intensive configuration, using the maximum number of network adapters the server is intended to support.
- **Leverage request**
 - Customer should create a new NVBug with leverage request.
 - Leverage request ticket should not include NVQual report, but a pointer to the parent/leveraged ticket.
 - Leverage ticket should include justification for getting approval
 - Example: (<https://nvbugspro.nvidia.com/bug/4850695>).
 - Leverage is approved under next conditions:
 - Configuration change: 8x B3140H to 8x CX7.
 - CX7's location (slot) are same as B3140's
 - No change in CPU and GPU device
 - No change in thermal design (fan, fan control, air baffles, ambient temperature)
- **PID: 1123578 NVIDIA Qualified Program Leveraging Guidelines**
- **Leverage**
 - B3140H <> B3220 <> B3240 (HA)
 - B3240 (CA) <?> B3220 > CX7

Nvqual 2.10 Update

- 使用最新的nvqual release

- [1086533 NVIDIA Network NVQual Kit](#)
- [1111407 NVIDIA Network NVQual x86 Host Docker Container Image](#)
- [1111408 NVIDIA Network NVQual Arm64 Host Docker Container Image](#)
- [1110878 NVIDIA Network NVQual BlueField-3 Docker Container Image](#)
- [1120109 Network NVQual User Guide for NVIDIA Networking Products.pdf](#)

- NVIDIA_Network_NVQual_Qualification_Template.xlsm → NVQual Test Results → Qualification Requirements

NVQual Test Item	NVQual Test Name	Qualification Requirements
Test #1	Combined Stress Test (DPU / ConnectX-7/ConnectX-8)	<div><p>Minimum systems: 1 (multiple systems recommended)</p><p>Test Condition: Max server supported configuration</p><p>Minimum Loops:</p><p><u>GPUs in the system:</u></p><ul style="list-style-type: none">- One loop of Network NVQual with GPU Thermal NVQual running concurrently- One loop with only Network NVQual.<p>Two loops total.</p><p><u>No GPUs in the system:</u> One loops of Network NVQual</p></div>

如何从nvqual log中判断结果

[16/12/2024 12:18:15] PLAYER_5 INFO

Accelerators	
Accelerator	Information
bf_ssd_stressor	<div>ESC[4mnvme0n1ESC[0m</div> <div>RUN STATUS (read): BW average: 1498MiB/s (1571MB/s) BW range: 1498MiB/s-1498MiB/s (1571MB/s-1571MB/s) data transferred: 5266GiB (5654GB) I/O runtime: 3600072-3600072msec device: nvme0n</div> <div>DISK STATISTICS (read/write): operations: 21569408/0 request merging: 0/0 runtime: 446987239/0 I/O request queue depth: 446987239 I/O utilization: 100.00% device: nvme0n</div> <div>PERFORMANCE TEST: ESC[32mPASSESC[0m</div>
dma	<div>Duration: 3600018726 micro seconds Enqueued jobs: 17443061 Dequeued jobs: 17443061 Throughput: 4845 Operations/s Ingress rate: 075.707 Gib/s Egress rate: 075.707 Gib/s</div> <div>PERFORMANCE TEST: ESC[32mPASSESC[0m</div>
stress_ng_intense	<div>CPU cores utilization threshold: 90%</div> <div>PERFORMANCE TEST: ESC[32mPASSESC[0m</div>

ib_traffic	<div>server ip: 127.0.1.1 client ip: 127.0.1.1 server_device: mlx5_6 client_device: mlx5_9 BW_average[Gb/sec]: 739.69 PERFORMANCE TEST: ESC[32mPASSESC[0m</div>
host_mlxlink_counter_0_0_0	<div>depth: 0, index: 0, node: 0, rdma\device: 0000:99:00.0 PCIe Operational (Enabled) Info ----- Depth, pcie index, node : 0, 0, 0 Link Speed Active (Enabled) : 32G-Gen 5 (32G-Gen 5) Link Width Active (Enabled) : 16X (16X) Management PCIe Performance Counters Info ----- RX Errors : 0 TX Errors : 0 CRC Error dllp : 0 CRC Error tlp : 0 Effective ber : 15E-255 PERFORMANCE TEST: ESC[32mPASSESC[0m</div>
host_pcie_eye_0_0_0	<div>depth: 0, index: 0, node: 0 device: 0000:99:00.0 connectd to: 0000:95:01.0 +-----+ PCIe Eye Last FOM Test +-----+ Lane Average Value Status Info +-----+ 15 124.4 ESC[32mPASSESC[0m +-----+ Notes: Last FOM fail threshold: 49 Last FOM pass threshold: 70 +-----+ PCIe Eye +-----+ Test Name Status Info +-----+ Link Speed ESC[32mPASSESC[0m Link Width ESC[32mPASSESC[0m </div>

如何从nvqual log中判断结果

Measurements					
Run authentication	Min measurement	Max measurement	Avg measurement	Performance Pass/Fail	Fail info
module temperature mlx5 6 [°C]	55.0	59.0	56.77	ESC[32mPASSESC[0m	State: steady state Time in SS: 450.283063 seconds Time from PENDING to SS: 899.406955 seconds Last measurements window: 300 seconds Max temperature delta allowed: 10 °C Required time to switch from PENDING to SS: 600 seconds Max PENDING time allowed: 3600 seconds Failed cycles: 0 out of 5 allowed
temperature [°C]	61.0	74.0	70.95		
bf_ssd_temperature_nvme0n1 [°C]	47.0	61.0	57.9	ESC[32mPASSESC[0m	
power [W]	33.0	58.0	54.4		
throttling_state [% CPU]	100	100	100.0		
arm_frequency [MHz]	disabled	disabled	disabled		
[16/12/2024 12:18:15] PLAYER_5 INFO					
Measurement Counters					
Run authentication	Initial count	Final count	Status		
power_throttling	0	0			
thermal_throttling	0	0	ESC[32mPASSESC[0m		

- 所有测试项有PASS和FAIL的结果显示，Power throttling 和 Thermal throttling 必须PASS
- CX7/BF3温度是从 thermal throttling count上看，必须为0
- Fail info 可以忽略
- 必须包括模块温度（使用光模块）

nvqual 常见问题- bf_ssd_stressor FAIL

Accelerators	
Accelerator	Information
bf_ssd_stressor	<div><div>ESC[4m/dev/nvme0n1:ESC[0m</div><div>RUN STATUS (read): BW average: 528MiB/s (554MB/s) BW range: 528MiB/s-528MiB/s (554MB/s-554MB/s) data transferred: 1856GiB (1993GB) I/O runtime: 3600192-3600192msec device: nvme0n1</div><div>DISK STATISTICS (read/write): operations: 7603559/92190 request merging: 0/4757 runtime: 445490552/14938258 I/O request queue depth: 460472676 I/O utilization: 100.00% device: nvme0n1</div><div>PERFORMANCE TEST: ESC[31mFAILESC[0m Average throughput 554.0 MBs is lower than minimal required throughput 600 MBs</div></div>

1. 检查DPU的root partition是否是nvme, 如果是重新烧录BFB, 指定root为emmc

[DPU] df -h

Filesystem	Size	Used	Avail	Use%	Mounted on
/dev/mmcblk0p2	38G	7.3G	29G	21%	/
/dev/mmcblk0p1	50M	8.6M	41M	18%	/boot/efi

cat bf.cfg

```
BMC_USER="root"
BMC_PASSWORD="OpenEuler12#$"
BMC_REBOOT="yes"
device="/dev/mmcblk0" # 指定烧入emmc
```

bfb-install --bfb bf-bundle-2.9.0-90_24.10_ubuntu-22.04_dev.bfb -r rshim0 -c bf.cfg

2. 如果root是emmc, 格式化nvme

[DPU] nvme format /dev/nvme0n1

3. 手动测试SSD性能

taskset -c 0 sudo time fio --rw=read --runtime=120 --time_based --ioengine=libaio --group_reporting --exitall --filename=/dev/nvme0n1 --name=/dev/nvme0n1 --bs=4096k --numjobs=16 --iodepth=16 --size=1G --loops=1 --invalidate=1 --randrepeat=1 --direct=1 --norandommap

nvqual 常见问题- ib_traffic FAIL

Accelerators	
Accelerator	Information
ib_traffic	server ip: 127.0.1.1
	client ip: 127.0.1.1
	server_device: mlx5_5
	client_device: mlx5_4
	BW_average[Gb/sec]: 179.38
	PERFORMANCE TEST: ESC [31m FAIL ESC [0m
	INFO: average BW 179.38 Gb/sec is lower than minimal required BW (200.0) on Gen_5.

```
16886: [run on local host]: numactl -C 49 ib_write_bw --ib-dev=mlx5_9 --ib-port=1 --port=18519 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1
16916: [run on local host]: numactl -C 49 ib_write_bw --ib-dev=mlx5_8 --ib-port=1 --port=18519 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1 127.0.1.1
16942: [run on local host]: numactl -C 49 ib_write_bw --ib-dev=mlx5_7 --ib-port=1 --port=18518 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1
16972: [run on local host]: numactl -C 49 ib_write_bw --ib-dev=mlx5_6 --ib-port=1 --port=18518 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1 127.0.1.1
16998: [run on local host]: numactl -C 1 ib_write_bw --ib-dev=mlx5_3 --ib-port=1 --port=18517 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1
17028: [run on local host]: numactl -C 1 ib_write_bw --ib-dev=mlx5_2 --ib-port=1 --port=18517 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1 127.0.1.1
17054: [run on local host]: numactl -C 1 ib_write_bw --ib-dev=mlx5_0 --ib-port=1 --port=18515 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1 127.0.1.1
17083: [run on local host]: numactl -C 1 ib_write_bw --ib-dev=mlx5_1 --ib-port=1 --port=18515 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1
17116: [run on local host]: numactl -C 1 ib_write_bw --ib-dev=mlx5_4 --ib-port=1 --port=18516 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1 127.0.1.1
17142: [run on local host]: numactl -C 1 ib_write_bw --ib-dev=mlx5_5 --ib-port=1 --port=18516 --size=1048576 --duration=120 --bidirectional --connection=RC --qp=1 --report_gbits --CPU-freq --post_list=1
```

- 多卡跑nvqual的ib_traffic是，nvqual将所有的ib_write_bw绑定到了固定的cpu core上，导致性能不足
- 联系相应SA
- Nvqual参数中禁止ib_traffic : --disabled_criteria ib_traffic
- 最新**Nvqual-2.10.02**版本已经解决次问题

nvqual 常见问题- running with container FAIL

```
[10/12/2024 06:06:59] CONTROLLER INFO  
bf_mlxlink_counter_3_1_0's process of player 0 completed.  
  
[10/12/2024 06:06:59] CONTROLLER INFO  
bf_pcie_eye_3_1_0's process of player 0 completed.  
  
[10/12/2024 06:11:59] CONTROLLER INFO  
NVQual rc_status has been changed from 0 to 6  
  
[10/12/2024 06:11:59] CONTROLLER ERROR  
module_temperature_mlx5_1's process of player 0 is still alive after join timeout.
```

- 运行 Nvqual 2.9.0 的container 模式，出现错误消息
- 解决方法是运行bf的container模式或者non-container模式
- 最新 **Nvqual 2.10.2** 已经解决

Agenda

- BFB Bundle Images and Update

- BF3 Portfolio Update – Cold aisle

- BF3 Hardware Update – Power supply

- BF3 NVQual Update

- **BF3 Firmware Update – NC-SI PKG ID Change**

- BF3 Host Management – Privilege and Host Management

- BF3 Reset Control

New BF3 FW(32.42.1000, 32.43.1014 and later) NC-SI pkg id changed

In the new BF3 FW, Pkg id is changed from 0x3 to 0x7 (some customized BF3 pkg id is changed from 0x0 to 0x4)

- host BMC should implement the pkg id discover mechanism as below. (some customers use hard code of pkg id)

Network Controller Sideband Interface (NC-SI) Specification DSP0222

The Channel ID field comprises two subfields, Package ID and Internal Channel ID, as described in Table 2.

Table 2 – Channel ID format

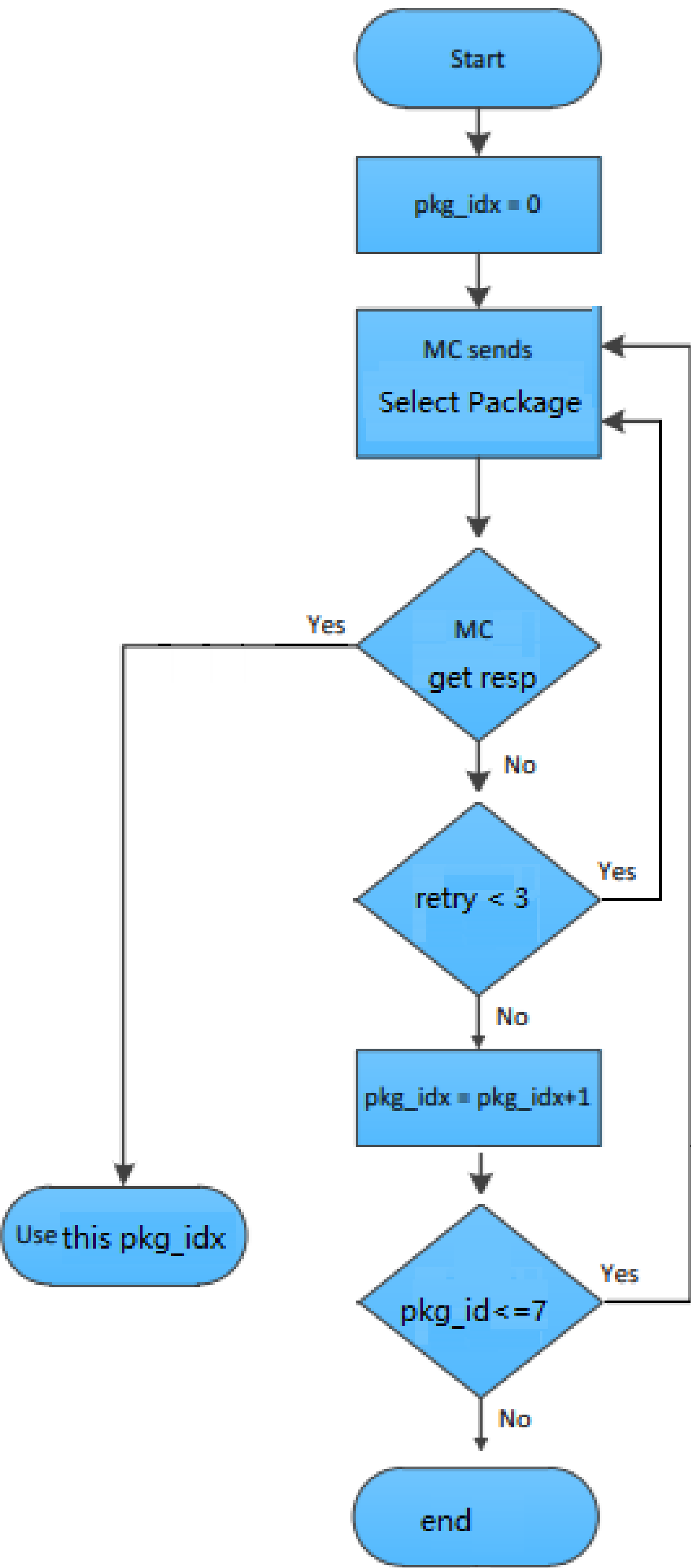
Bits	Field Name	Description
[7..5]	Package ID	The Package ID is required to be common across all channels within a single Network Controller that share a common NC-SI physical interconnect. The system integrator will typically configure the Package IDs starting from 0 and increasing sequentially for each physical Network Controller. The Network Controller shall allow the least significant two bits of this field to be configurable by the system integrator, with the most significant bit of this field = 0b. An implementation is allowed to have all 3 bits configurable.
[4..0]	Internal Channel ID	The Network Controller shall support Internal Channel IDs that are numbered starting from 0 and increasing sequentially for each Pass-through channel

Discover package

The Management Controller issues a Select Package command starting with the lowest Package ID (see 8.4.5 for more information). Because the Management Controller is assumed to have no prior knowledge of whether the Network Controller is enabled for hardware arbitration, the Select Package command is issued with the Hardware Arbitration parameter set to 'disable'.

If the Management Controller receives a response within the specified response time, it can record that it detected a package at that ID. If the Management Controller does not receive a response, it is recommended that the Management Controller retry sending the command. Three total tries is typical. (This same retry process should be used when sending all commands to the Network Controller and will be left out of the descriptions in the following steps.) If the retries fail, the Management Controller can assume that no Network Controller is at that Package ID and can immediately repeat this step 2) for the next Package ID in the sequence.

Network Controller Sideband Interface (NC-SI) Specification DSP0222



Agenda

- BFB Bundle Images and Update

- BF3 Portfolio Update – Cold aisle

- BF3 Hardware Update – Power supply

- BF3 NVQual Update

- BF3 Firmware Update – NC-SI PKG ID Change

- **BF3 Host Management – Privilege and Host Management**

- BF3 Reset Control

DPU Access Privilege (1/2)

- NVIDIA provides OEM NCSI command for host BMC to control the access privilege to DPU. The Purpose is to avoid unexpected changes to DPU by unauthorized or error operations.

Use cases:

- SuperNIC mode/DPU mode – a wrong operation of mlxconfig may reset DPU to default mode.
- Administrator may control the unexpected DPU FW upgrade or PCC FW upgrade
- Only through specific IPMI command to authorize the access to DPU settings.


配置项	生效机制	配置方式	Mellanox OEM command	功能描述
Set Host Access to Smart NIC CPU	Power Cycle	BMC NCSI	CMD ID: 0x12 Parameter ID: 0x19	打开/关闭Host侧通过rshim方式访问DPU (console, reboot, bfb 升级)
Set External Host Privileges	Power Cycle	BMC NCSI	CMD ID: 0x12 Parameter ID: 0x32	打开/关闭Host侧对DPU的各项访问权限 (NIC reset/FW update/PCC update)
Set Smart NIC Mode	Power Cycle	BMC NCSI	CMD ID: 0x12 Parameter ID: 0x33	配置DPU的工作模式 (SuperNIC/DPU)

- [PID: 1106035 NVIDIA Mellanox Specific NC-SI OEM Commands Application Note](#)

DPU Access Privilege (2/2)

- NCSI Command details:

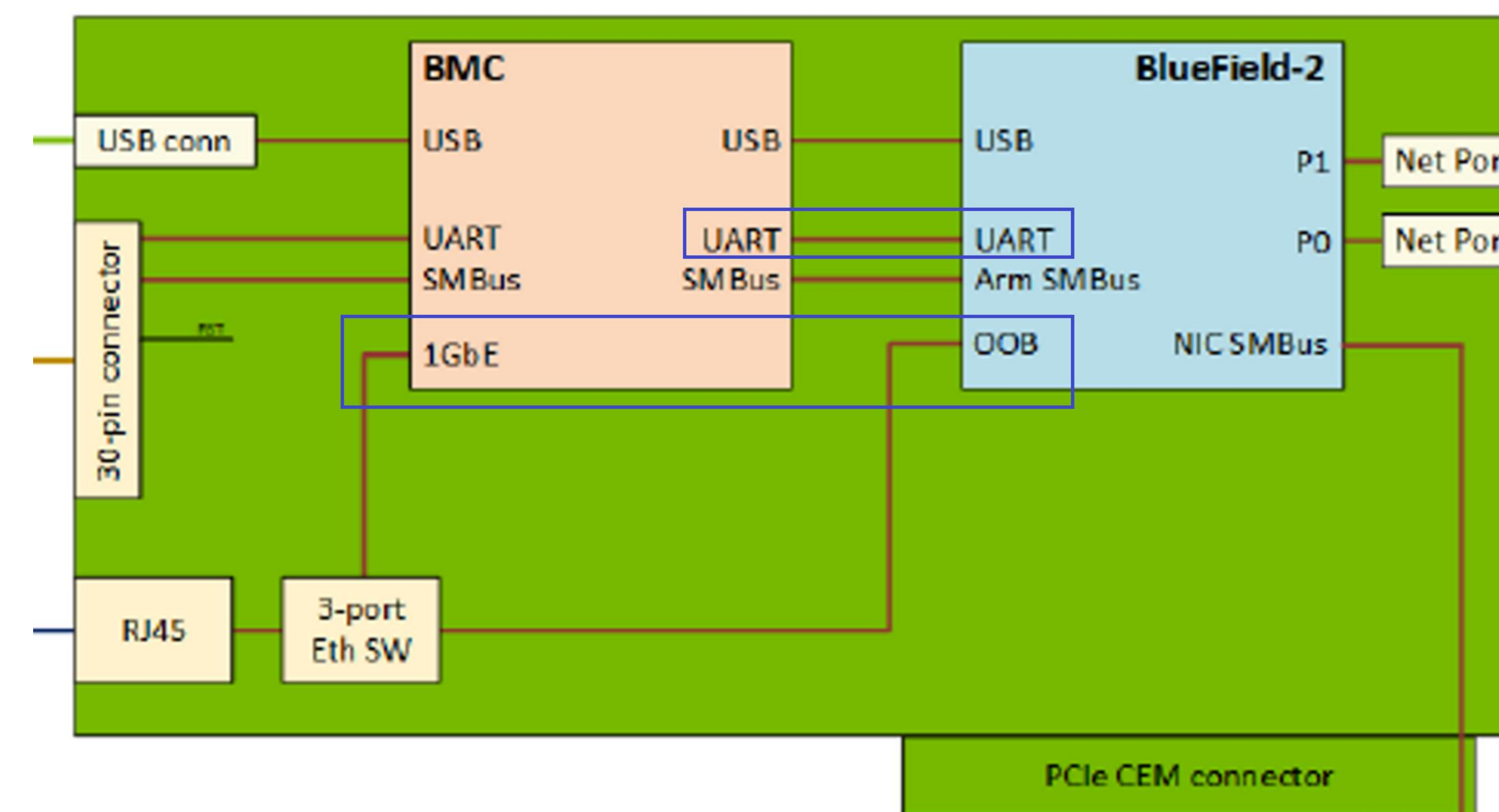
```
root@ :~# mlxconfig -d mlx5_2 q|grep -i priv
R0    HOST_PRIV_RSHIM                ENABLE(1)
R0    HOST_PRIV_FLASH_ACCESS        DEVICE_DEFAULT(0)
R0    HOST_PRIV_NV_HOST             DEVICE_DEFAULT(0)
R0    HOST_PRIV_NV_PORT             DEVICE_DEFAULT(0)
R0    HOST_PRIV_NV_GLOBAL            DEVICE_DEFAULT(0)
R0    HOST_PRIV_NV_INTERNAL_CPU      DISABLE(2)
R0    HOST_PRIV_PCC_UPDATE           DEVICE_DEFAULT(0)
R0    HOST_PRIV_FW_UPDATE            DEVICE_DEFAULT(0)
R0    HOST_PRIV_NIC_RESET            DEVICE_DEFAULT(0)
```



- Server BMC SW supports the settings and provide IPMI interface to access.

DPU ARM/BMC In-band management(1/4)

- Access DPU BMC without OOB connected.
 - Login DPU ARM from Host through IP/UART over rshim over PCIe
 - minicom -D /dev/rshim0/console
 - ssh root@dpu_tmfifo_net0
 - IPMI to set DPU BMC static IP
 - sudo ipmitool lan set 1 ipsrc static
 - sudo ipmitool lan set 1 ipaddr 11.11.11.2
 - sudo ipmitool lan set 1 defgw ipaddr 11.11.11.1
 - sudo ipmitool lan set 1 netmask 255.255.255.0
 - sudo ifconfig oob_net0 11.11.11.1/24 up
 - SSH BMC from DPU ARM



DPU ARM/BMC In-band management(2/4)

- Example - Access DPU BMC without OOB connected, upgrade BMC FW

- 1. 设置 tmfifo_net0 192.168.100.1/24
 - minicom -D /dev/rshim0/console 设置 DPU ARM tmfifo_net0 192.168.100.2/24
- 2. 设置 tmfifo_net1 192.168.101.1/24
 - minicom -D /dev/rshim1/console 设置 DPU ARM tmfifo_net0 192.168.101.2/24
- 3. scp *.fwpkg ubuntu@192.168.10X.2:/home/ubuntu
- 4. ssh ubuntu@192.168.10X.2 后, 设置BMC OOB ip, 设置ARM OOB IP

sudo ipmitool lan set 1 ipsrc static

sudo ipmitool lan set 1 ipaddr 11.11.11.2

sudo ipmitool lan set 1 defgw ipaddr 11.11.11.1

sudo ipmitool lan set 1 netmask 255.255.255.0

sudo ifconfig oob_net0 11.11.11.1/24 up

- 5. 升级BMC

Upgrade BMC CEC: //升级大约需要20s

- curl -k -u root:'Nvidia@123!' -H "Content-Type: application/octet-stream" -X POST -T /home/ubuntu/cec1736-ecfw-00.02.0152.0000-n02-rel-prod.fwpkg https://11.11.11.2/redfish/v1/UpdateService
- curl -k -u root:'Nvidia@123!' -X GET https://11.11.11.2/redfish/v1/TaskService/Tasks/0 | jq -r '.PercentComplete'

Upgrade BMC FW: // 升级大约需要12 分钟

- curl -k -u root:'Nvidia@123!' -H "Content-Type: application/octet-stream" -X POST -T /home/ubuntu/bf3-bmc-23.10-5_opn.fwpkg https://11.11.11.2/redfish/v1/UpdateService/update
- curl -k -u root:'Nvidia@123!' -X GET https://11.11.11.2/redfish/v1/TaskService/Tasks/1 | jq -r '.PercentComplete'

DPU ARM/BMC In-band management(3/4)

- In case DPU ARM is not able to access through Host rshim or OOB IP.
 - DPU BMC OOB DHCP enabled by default
 - Use DHCP Server (OpenDHCPServer) to connect DPU OOB and assign DPU BMC IP
 - IPMI login DPU BMC and SOL redirect to ARM console
 - `ipmitool -c 17 -I lanplus -H <DPU_BMC_IP> -U root -P <DPU_BMC_password> sol set set-in-progress set-complete 1`
 - `ipmitool -c 17 -I lanplus -H <DPU_BMC_IP> -U root -P <DPU_BMC_password> sol set enabled true 1`
 - `ipmitool -c 17 -I lanplus -H <DPU_BMC_IP> -U root -P <DPU_BMC_password> sol active`

← ↻ ⚠ Not secure | 192.168.10.11:6789

Open DHCP Server Version 2.00RC Windows Build 1060

Server: NV-6FMQDY3 32bit <http://dhcpserver.sourceforge.net/>

Active Leases			
Mac Address	IP	Lease Expiry	Hostname (first 20 chars)
5c:25:73:16:ad:29	192.168.10.26	22-Nov-24 20:11:49	dpu-bmc
5c:25:73:1f:5b:45	192.168.10.22	22-Nov-24 18:39:06	dpu-bmc
5c:25:73:1f:7a:a3	192.168.10.23	22-Nov-24 19:09:04	dpu-bmc
5c:25:73:3f:a9:c9	192.168.10.24	22-Nov-24 19:45:44	dpu-bmc
5c:25:73:42:d5:ff	192.168.10.25	22-Nov-24 20:31:35	dpu-bmc

```
root@l-csi-ar-0908h:~# ipmitool -C 17 -I lanplus -H 10.7.157.97 -U root -P Nvidia_12345! sol set set-in-progress set-complete 1
root@l-csi-ar-0908h:~# ipmitool -C 17 -I lanplus -H 10.7.157.97 -U root -P Nvidia_12345! sol set enabled true 1
root@l-csi-ar-0908h:~# ipmitool -C 17 -I lanplus -H 10.7.157.97 -U root -P Nvidia_12345! sol activate
[SOL Session operational. Use ~? for help]
```

```
l-csi-bf3-200g-03 login: ubuntu
Password:
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 5.15.0-1035-bluefield aarch64)
```

```
* Documentation: https://help.ubuntu.com
* Management:   https://landscape.canonical.com
* Support:       https://ubuntu.com/pro
```

```
System information as of Tue Apr 30 05:56:40 UTC 2024
```

```
System load:      1.1865234375
Usage of /:        6.7% of 116.76GB
Memory usage:     30%
Swap usage:       0%
Temperature:      84.9 C
```


DPU ARM/BMC In-band management(4/4)

- In case DPU ARM is not able to access through Host rshim or OOB IP.
 - DPU BMC OOB DHCP enabled by default
 - Use DHCP Server (OpenDHCPServer) to connect DPU OOB and assign DPU BMC IP
 - SSH to DPU BMC then login ARM through UART
 - `ssh root@<DPU_BMC_ip>`

```
root@bluesphere:~# obmc
obmc-console-client          obmc-mellanox-mac-syncd.sh
obmc-console-server         obmc-secure-copy-image
obmc-flash-bmc              obmcutil
root@bluesphere:~# obmc-console-client

CentOS Linux 7 (AltArch)
Kernel 4.19.161-bf.113.gadcd9e3 on an aarch64

l-csi-bf2-100g-51 login: root
```


Agenda

- BFB Bundle Images and Update

- BF3 Portfolio Update – Cold aisle

- BF3 Hardware Update – Power supply

- BF3 NVQual Update

- BF3 Firmware Update – NC-SI PKG ID Change

- BF3 Host Management – Privilege and Host Management

- **BF3 Reset Control**

DPU RESET CONTROL (BMC 23.07 and later)

Reset level	Command	comments
Hard reset of DPU arm cores and nic subsystem	<pre>curl -k -u root:'<password>' -H "Content-Type: application/json" -X POST https://<bmc_ip>/redfish/v1/Systems/Bluefield/Actions/ComputerSystem.Reset -d '{"ResetType": "PowerCycle"}'</pre> <pre>ipmitool chassis power cycle</pre>	A hard reset of BlueField is permitted only when all connected hosts assert the PERST signal. It is crucial for all host devices to assert the PERST signal when BlueField-3 is shared among multiple hosts to enable a hard reset.
Force Hard reset of DPU arm cores and nic subsystem	<pre>curl -k -u root:'<password>' -H "Content-Type: application/json" -X POST https://<bmc_ip>/redfish/v1/Systems/Bluefield/Oem/Nvidia/SOC.ForceReset</pre>	Force hard reset of the BlueField happens without waiting for All_STANDBY or PERST. Users must make sure the server is ready for the reset!

DPU RESET CONTROL (BMC 23.07 and later)

Reset level	Command	comments
Hard reset of DPU arm cores	<pre>curl -k -u root:'<password>' -H "Content-Type: application/json" -X POST https://<bmc_ip>/redfish/v1/Systems /Bluefield/Actions/ComputerSystem.R eset -d '{"ResetType": "ForceRestart"}'</pre> <pre>ipmitool chassis power reset</pre>	
Soft shutdown of DPU arm cores	<pre>curl -k -u root:'<password>' -H "Content-Type: application/json" -X POST https://<bmc_ip>/redfish/v1/Systems /Bluefield/Actions/ComputerSystem.R eset -d '{"ResetType": "GracefulShutdown"}'</pre> <pre>ipmitool power soft</pre>	This command is relevant only for BlueField-3 devices.

