

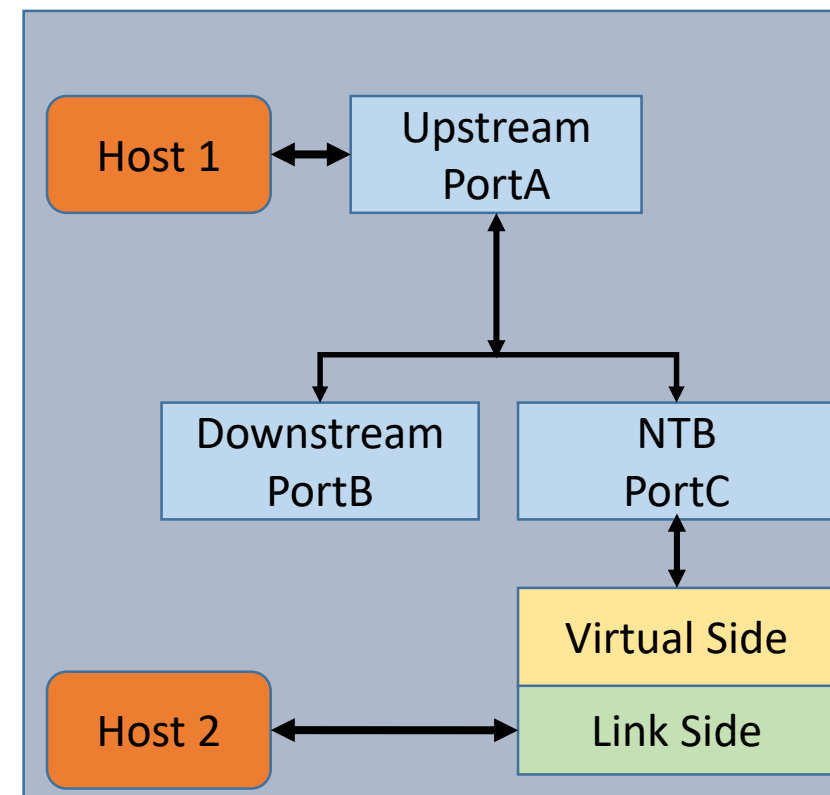
REDUNDANT CPU ON MTCA SYSTEM WITH PCI EXPRESS NON TRANSPARENT BRIDGE

L.Petrosyan (DESY)

- NTB used to connect two independent address/Host domains
- A Non-Transparent bridge consist of two back-to-back PCIe endpoints, a Virtual and Link side endpoints.
- NTB isolates Address spaces of different Hosts by appearing as an endpoint to each side

NTB Provides:

1. Allow to have second CPU on MTCA system (7th MicroTCA Workshop)
2. Allow to have redundant CPU on MTCA system (now)

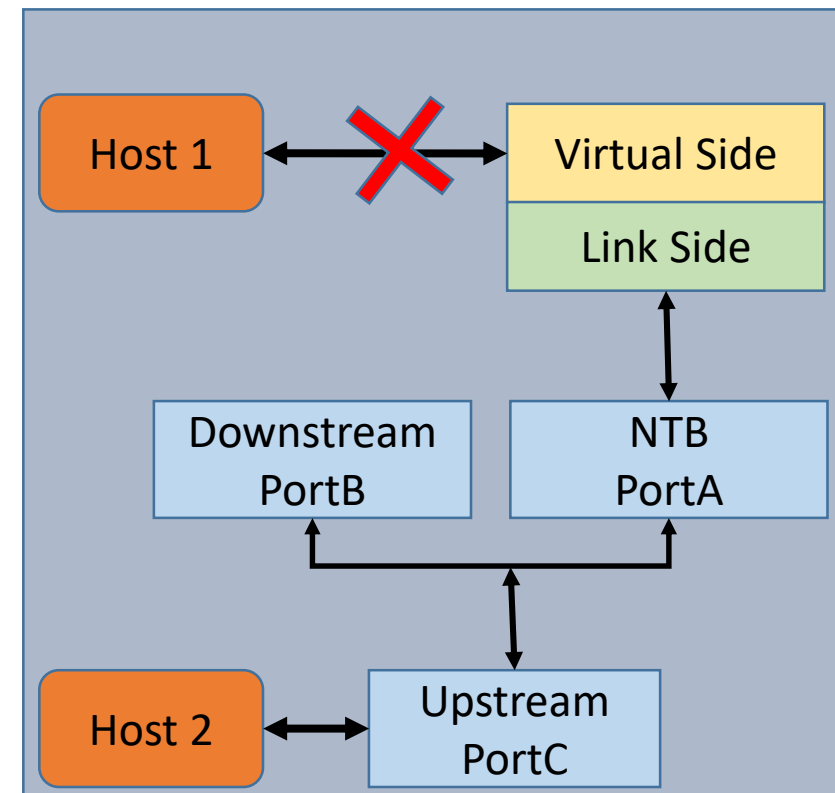
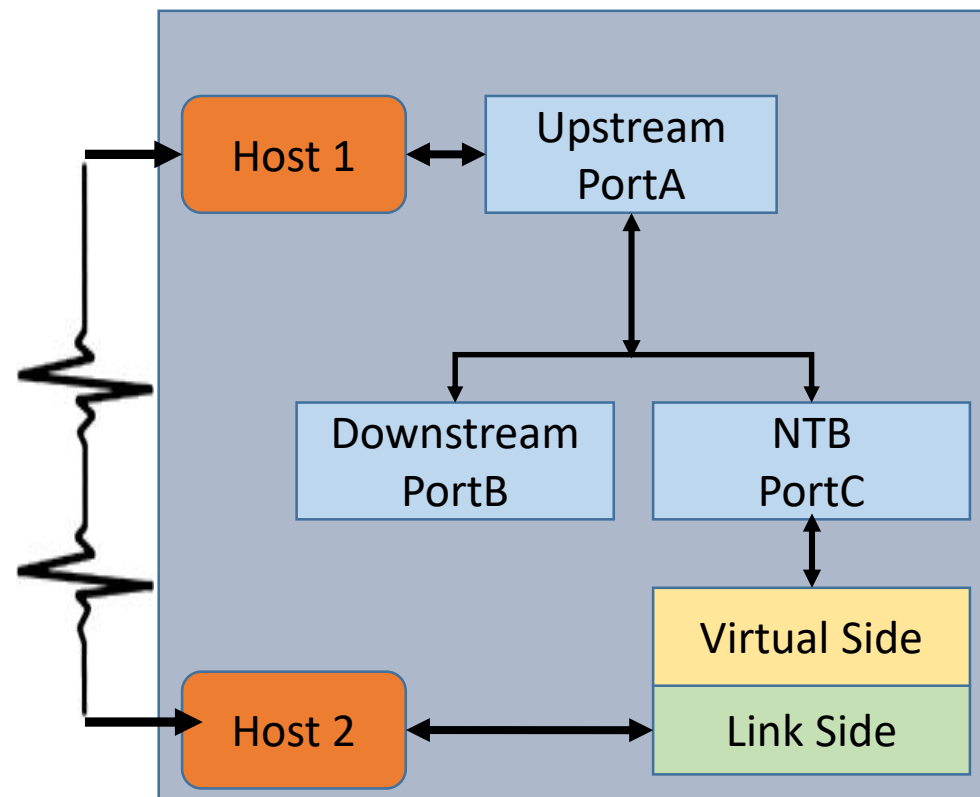


To protect against a failing Host taking the entire system down, a backup Host can be in place, ready to take over.

The passive Host2 is waiting for the active Host1 to fail.

When the Virtual side Host1 fail, passive Host2 initiates the sequence:

1. Configure PortA to be the NT Port and PortC to be the Upstream Port
2. Reset Upstream Port Secondary Bus
3. Re-enumerate the hierarchy and start operation.

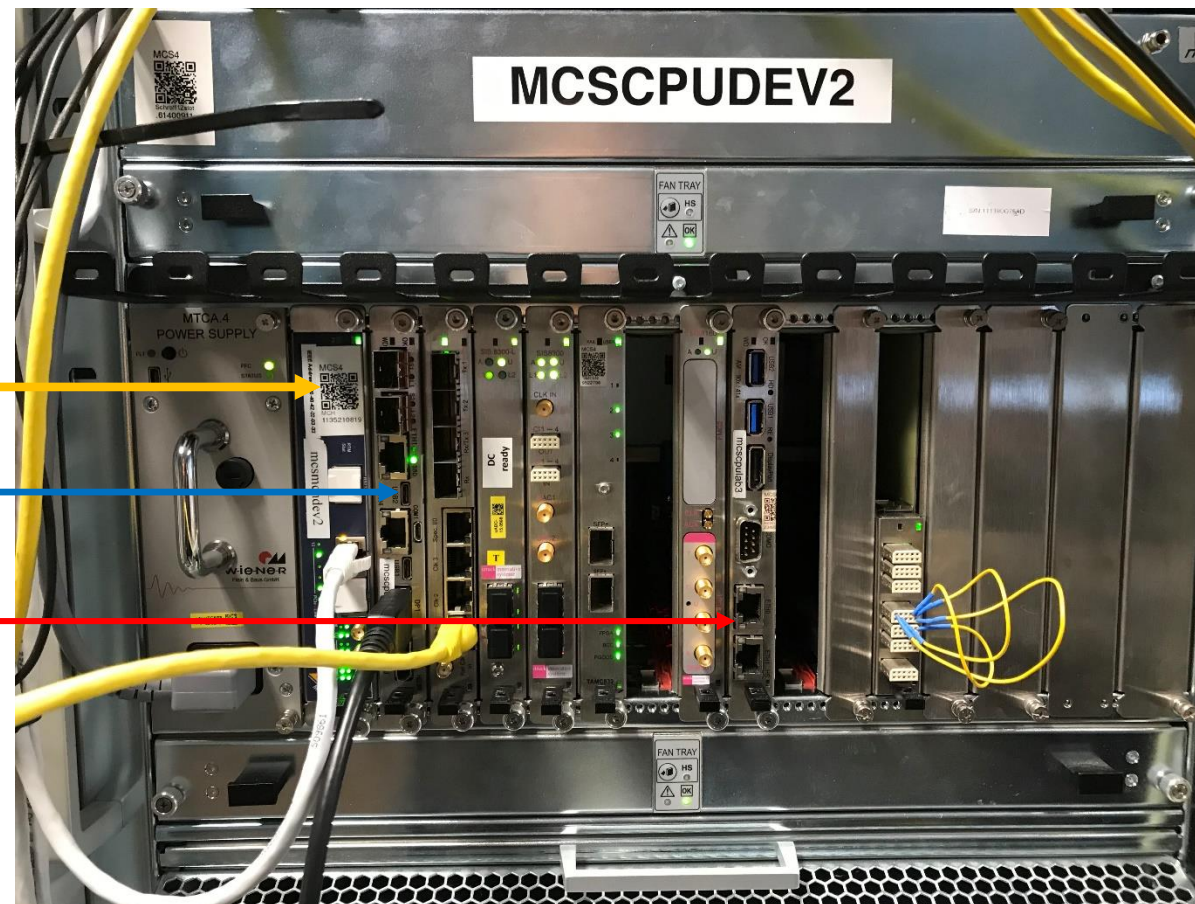


NAT MCH with PEX8748 PCIE Switch

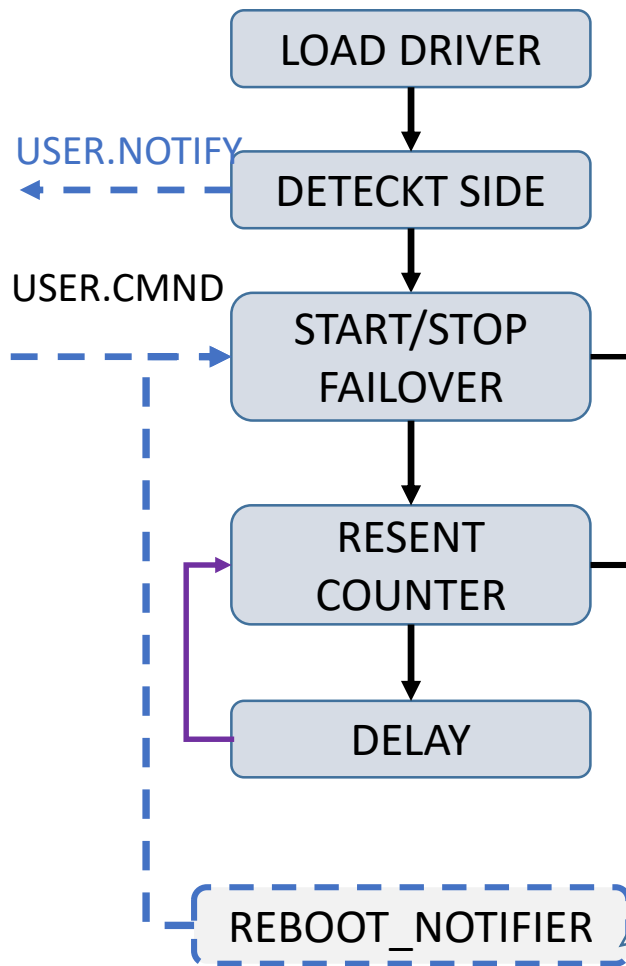
Host1 in Slot 1 (PCIE Switch Port19)

Host2 in Slot 8 (PCIE Switch Port16)

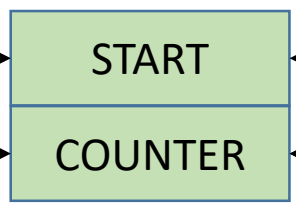
PCIENTBV – PCIE NTB device driver is running on both CPUs



Virtual Side Host1(Upstream Port)

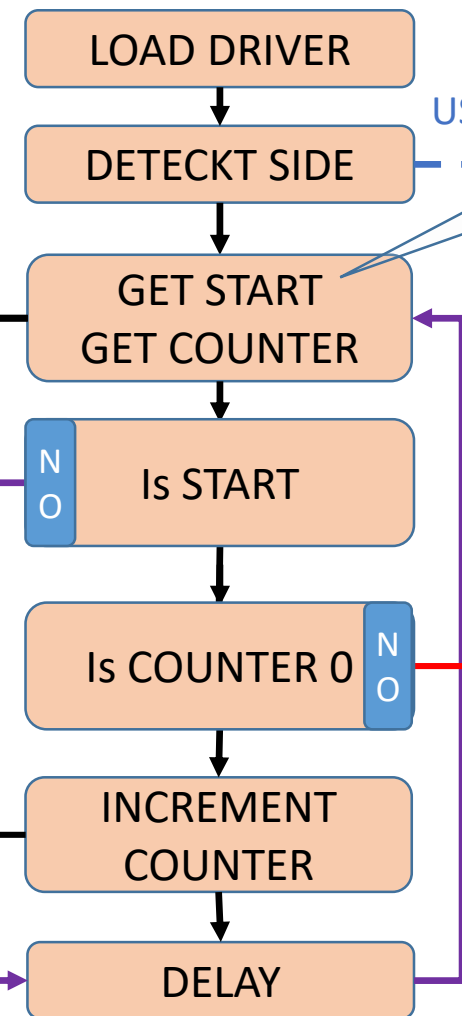


NTB SCRATCH REG.

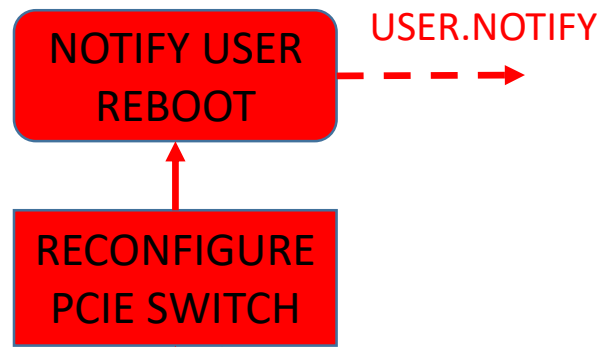


- Fail checking could be started-stopped by user or by OS, for example, when the system is going to be rebooted
- Force Ports reconfiguration

Link Side Host2 (NTB Port)



Do not start failover while the Host1 is not loaded



Configure.
Upstream Port as NTB
NTB Port as Upstream

Virtual Side Host 1.

All AMC cards are visible from this side
NTB Endpoint in slot 8 (Port 16, 10:00.0)

Link Side Host 2.

Only NTB Endpoint is visible

Slot	Dev	IDs	Driver	BARs
1	0000:02:01.0	10b5:8725	SWITCH ON	LSPCI
10	0000:04:00.0	10b5:8748	SWITCH ON	LSPCI
11	0000:04:00.0	10b5:8748	SWITCH ON	LSPCI
12	0000:04:0b.0	10b5:8748	SWITCH ON	LSPCI
2	0000:04:11.0	10b5:8748	SWITCH ON	LSPCI
3	0000:04:03.0	10b5:8748	SWITCH ON	LSPCI
4	0000:04:01.0	10b5:8748	SWITCH ON	LSPCI
5	0000:04:08.0	10b5:8748	SWITCH ON	LSPCI
6	0000:04:0a.0	10b5:8748	SWITCH ON	LSPCI
7	0000:04:12.0	10b5:8748	SWITCH ON	LSPCI
9	0000:02:09.0	10b5:8725	SWITCH ON	LSPCI
9.1	0000:04:02.0	10b5:8748	SWITCH ON	LSPCI

```

root@ncscpulab3:~# apt-get install esdadio-dkms
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
  esdadio-dkms
0 upgraded, 1 newly installed, 0 to remove and 224 not upgraded.
Need to get 8,882 B of archives.
After this operation, 52.2 kB of additional disk space will be used.
Get:1 http://doocspkgs.desy.de/pub/docs xenial/main amd64 esdadio-dkms all 1.0.32-xenial1 [8,882 B]
Fetched 8,882 B in 0s (200 kB/s)
Selecting previously unselected package esdadio-dkms.
(Reading database ... 189974 files and directories currently installed.)
Preparing to unpack .../esdadio-dkms_1.0.32-xenial1_all.deb ...
Unpacking esdadio-dkms (1.0.32-xenial1) ...
Setting up esdadio-dkms (1.0.32-xenial1) ...
Loading new esdadio-1.0.32-xenial1 DKMS files...
First Installation; checking all kernels...
Building only for 4.4.0-139-generic
Building initial module for 4.4.0-139-generic
Done.

esdadio:
Running module version sanity check.
 - Original module
 - No original module exists within this kernel
 - Installation
 - Installing to /lib/modules/4.4.0-139-generic/updates/dkms/

depmod....

DKMS: install completed.
root@ncscpulab3:~# mtcamonitora
[1] 25947
root@ncscpulab3:~# modprobe pcientbv
root@ncscpulab3:~# lspci -H
00:00.0 Host bridge: Intel Corporation 3rd Gen Core processor DRAM Controller (rev 09)
00:01.0 PCI bridge: Intel Corporation Xeon E3-1200 v2/3rd Gen Core processor PCI Express Root Port (rev 09)
00:01.1 PCI bridge: Intel Corporation Xeon E3-1200 v2/3rd Gen Core processor PCI Express Root Port (rev 09)
00:02.0 VGA compatible controller: Intel Corporation 3rd Gen Core processor Graphics Controller (rev 09)
00:14.0 USB controller: Intel Corporation 7 Series/C210 Series Chipset Family USB xHCI Host Controller (rev 04)
00:16.0 Communication controller: Intel Corporation 7 Series/C210 Series Chipset Family MEI Controller #1 (rev 04)
00:19.0 Ethernet controller: Intel Corporation 82579LM Gigabit Network Connection (rev 04)
00:1a.0 USB controller: Intel Corporation 7 Series/C210 Series Chipset Family USB Enhanced Host Controller #2 (rev 04)
00:1b.0 Audio device: Intel Corporation 7 Series/C210 Series Chipset Family High Definition Audio Controller (rev 04)
00:1c.0 PCI bridge: Intel Corporation 7 Series/C210 Series Chipset Family PCI Express Root Port 1 (rev c4)
00:1c.4 PCI bridge: Intel Corporation 7 Series/C210 Series Chipset Family PCI Express Root Port 5 (rev c4)
00:1d.0 USB controller: Intel Corporation 7 Series/C210 Series Chipset Family USB Enhanced Host Controller #1 (rev 04)
00:1f.0 ISA bridge: Intel Corporation QM77 Express Chipset LPC Controller (rev 04)
00:1f.2 IDE interface: Intel Corporation 7 Series Chipset Family 4-port SATA Controller [IDE mode] (rev 04)
00:1f.3 SMBus: Intel Corporation 7 Series/C210 Series Chipset Family SMBus Controller (rev 04)
00:1f.5 IDE interface: Intel Corporation 7 Series Chipset Family 2-port SATA Controller [IDE mode] (rev 04)
01:00.0 PCI bridge: PLX Technology, Inc. PEX 8717 16-lane, 8-Port PCI Express Gen 3 (8.0 GT/s) Switch with DMA (rev ca)
01:00.1 System peripheral: PLX Technology, Inc. Device 87d0 (rev ca)
01:00.2 System peripheral: PLX Technology, Inc. Device 87d0 (rev ca)
01:00.3 System peripheral: PLX Technology, Inc. Device 87d0 (rev ca)
01:00.4 System peripheral: PLX Technology, Inc. Device 87d0 (rev ca)
02:01.0 PCI bridge: PLX Technology, Inc. PEX 8717 16-lane, 8-Port PCI Express Gen 3 (8.0 GT/s) Switch with DMA (rev ca)
03:00.0 PCI bridge: PLX Technology, Inc. Device 8748 (rev ca)
08:00.0 Ethernet controller: Intel Corporation 82580 Gigabit Backplane Connection (rev 01)
08:00.1 Ethernet controller: Intel Corporation 82580 Gigabit Backplane Connection (rev 01)
09:00.0 Ethernet controller: Intel Corporation 82574L Gigabit Network Connection
root@ncscpulab3:~#

```

Heartbeats messages on the Link Side Host2 NTB device driver

Host1 is OK

Host1 Fail

```

File Edit View Search Terminal Help
5Oct 24 14:52:5 mcscpulab3 kernel: [184590.463035] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG 0
6Oct 24 14:52:5 mcscpulab3 kernel: [184592.438573] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341040497 DATA2
7Oct 24 14:52:5 mcscpulab3 kernel: [184592.446820] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341040499 DATA 2 LINK SIDE 1
8Oct 24 14:52:5 mcscpulab3 kernel: [184592.456084] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
9Oct 24 14:52:5 mcscpulab3 kernel: [184592.463006] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG 1
eOct 24 14:52:5 mcscpulab3 kernel: [184594.438577] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341040997 DATA2
eOct 24 14:52:58 mcscpulab3 kernel: [184594.446582] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341040999 DATA 2 LINK SIDE 1
5Oct 24 14:52:58 mcscpulab3 kernel: [184594.455633] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
eOct 24 14:52:58 mcscpulab3 kernel: [184594.462378] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG 1
tOct 24 14:53:00 mcscpulab3 kernel: [184596.438568] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341041497 DATA2
iOct 24 14:53:00 mcscpulab3 kernel: [184596.447089] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341041498 DATA 2 LINK SIDE 1
tOct 24 14:53:00 mcscpulab3 kernel: [184596.456667] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
eOct 24 14:53:00 mcscpulab3 kernel: [184596.463982] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG 1
7Oct 24 14:53:02 mcscpulab3 kernel: [184598.438527] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341041997 DATA2
7Oct 24 14:53:02 mcscpulab3 kernel: [184598.447215] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341041998 DATA 2 LINK SIDE 1
7Oct 24 14:53:02 mcscpulab3 kernel: [184598.456944] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
7Oct 24 14:53:02 mcscpulab3 kernel: [184598.464396] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG 1
eOct 24 14:53:04 mcscpulab3 kernel: [184600.438547] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341042497 DATA2
vOct 24 14:53:04 mcscpulab3 kernel: [184600.447399] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341042498 DATA 2 LINK SIDE 1
aOct 24 14:53:04 mcscpulab3 kernel: [184600.457294] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
gOct 24 14:53:04 mcscpulab3 kernel: [184600.464883] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG 1
gOct 24 14:53:04 mcscpulab3 kernel: [184600.473570] XXXXXXXXXXXXXXXXXXXXPCIEDEV_TIMER: VIRTUAL SIDE FAIL 4
Oct 24 14:53:04 mcscpulab3 kernel: [184600.528316] DMAR: DRHD: handling fault status reg 2
eOct 24 14:53:04 mcscpulab3 kernel: [184600.535535] DMAR: INTR-REMAP: Request device [[0f:00.0] fault index a0
dOct 24 14:53:04 mcscpulab3 kernel: [184600.535535] INTR-REMAP:[fault reason 34] Present field in the IRTE entry is clear
dOct 24 14:53:06 mcscpulab3 kernel: [184602.438540] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341042997 DATA2
Oct 24 14:53:06 mcscpulab3 kernel: [184602.447523] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341042998 DATA 2 LINK SIDE 1
5Oct 24 14:53:06 mcscpulab3 kernel: [184602.457519] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
Oct 24 14:53:06 mcscpulab3 kernel: [184602.465181] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG -1
iOct 24 14:53:06 mcscpulab3 kernel: [184602.474033] XXXXXXXXXXXXXXXXXXXXPCIEDEV_TIMER: VIRTUAL SIDE FAIL 5
5Oct 24 14:53:08 mcscpulab3 kernel: [184604.438526] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341043497 DATA2
rOct 24 14:53:08 mcscpulab3 kernel: [184604.447093] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341043498 DATA 2 LINK SIDE 1
eOct 24 14:53:08 mcscpulab3 kernel: [184604.456681] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
eOct 24 14:53:08 mcscpulab3 kernel: [184604.463948] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG -1
5Oct 24 14:53:08 mcscpulab3 kernel: [184604.472440] XXXXXXXXXXXXXXXXXXXXPCIEDEV_TIMER: VIRTUAL SIDE FAIL 6
eOct 24 14:53:10 mcscpulab3 kernel: [184606.438523] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341043997 DATA2
tOct 24 14:53:10 mcscpulab3 kernel: [184606.447723] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341043999 DATA 2 LINK SIDE 1
tOct 24 14:53:10 mcscpulab3 kernel: [184606.457951] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE 1
tOct 24 14:53:10 mcscpulab3 kernel: [184606.465873] LLLLLLLLLLLLLLLLLPCIEDEV_TIMER: LINK SIDE SCRATCH_REG -1
eOct 24 14:53:10 mcscpulab3 kernel: [184606.474962] XXXXXXXXXXXXXXXXXXXXPCIEDEV_TIMER: VIRTUAL SIDE FAIL 7
7Oct 24 14:53:12 mcscpulab3 kernel: [184608.438517] $$$$$$$$$$$$PCIEDEV_TIMER: EXPIRED AT 4341044497 DATA2

```

Using NAT MCH Diag tool, we can see the PCIE Switch Upstream Port Register is changed

```

PCIE (RET=0/0x0):
PCIE (RET=0/0x0): 8
select hub module (0=MCH1, 1=MCH2) (RET=0/0x0): LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion

Enter port (RET=0/0x0):
Enter address (RET=864/0x360):
Enter access mode (0=TP, 1=NT-L or 2=NT-V) (RET=0/0x0): LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion

HUB REG 0x00000360 = 0x00003013 ← Port 19 Upstream, Port 16 NTB
PCIE (RET=0/0x0):
PCIE (RET=0/0x0):
PCIE (RET=0/0x0):
PCIE (RET=0/0x0): LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -assertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion

PCIE (RET=0/0x0):
PCIE (RET=0/0x0): 8
select hub module (0=MCH1, 1=MCH2) (RET=0/0x0):
Enter port (RET=0/0x0):
Enter address (RET=864/0x360):
Enter access mode (0=TP, 1=NT-L or 2=NT-V) (RET=0/0x0): ← Port 16 Upstream, Port 19 NTB
HUB REG 0x00000360 = 0x00003310
PCIE (RET=0/0x0): LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -assertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -assertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -assertion

```

Port 19 Upstream, Port 16 NTB

Port 16 Upstream, Port 19 NTB

After the Host1 fail is detected, the NTB device driver has reconfigured the PCIE switch and rebooted the CPU
 Now Host 2 is Upstream host, and we can see all AMC modules from this host.
The same picture we had before on Host1

The terminal window shows the system boot process, including the detection of various hardware components like PCI bridges, signal processing controllers, and network controllers. The PCIe-Monitor application displays a detailed view of the PCIe bus configuration, organized into columns for each device (1-11). Each column contains information such as device ID, bus address, switch status, device type (e.g., LSPCI, PCIe-R/W), and driver details.

1	10	11	12	2	3	4	5	6	7	9
Dev: 0000:02:01.0	Dev: 0000:04:00.0	Dev: 0000:04:09.0	Dev: 0000:04:0b.0	Dev: 0000:04:11.0	Dev: 0000:04:03.0	Dev: 0000:04:01.0	Dev: 0000:04:08.0	Dev: 0000:04:0a.0	Dev: 0000:04:12.0	Dev: 0000:04:02.0
10b5:8717	10b5:8748	10b5:8748	10b5:8748	10b5:8748	10b5:8748	10b5:8748	10b5:8748	10b5:8748	10b5:8748	10b5:8748
SWITCH ON	SWITCH ON	SWITCH OFF	SWITCH ON	SWITCH OFF	SWITCH OFF	SWITCH OFF	SWITCH OFF	SWITCH ON	SWITCH OFF	SWITCH ON
LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI
DEV: 0000:03:00.0	DEV: 0000:0b:04.0	DEV: 0000:0e:00.0	DEV: 0000:08:00.0	DEV: 0000:06:00.0	DEV: 0000:09:00.0	DEV: 0000:0f:00.0	DEV: 0000:0e:00.0	DEV: 0000:0f:00.0	DEV: 0000:0e:00.0	DEV: 0000:13:00.0
IDs: 00:00	IDs: 10b5:9056	IDs: 10ee:0020	IDs: 1796:0019	IDs: 10ee:0088	IDs: 1498:8214	IDs: 1796:0028	IDs: 1498:800a	IDs: 1796:0028	IDs: 1796:0028	IDs: 8086:1572
00:00	12fe:0600	3300:0020	1796:0019	3300:0088	1498:800a	1796:0028	1498:800a	1796:0028	1796:0028	8086:0000
NO DRIVER	Driver: esdadio	Driver: x1timer	Driver: sis8300	Driver: pciedev	Driver: tamc532	Driver: sis8160	Driver: sis8160	Driver: sis8160	Driver: sis8160	Driver: sis8160
DevFile: 4.0	DevFile: 4.0	DevFile: 5.1.0	DevFile: 7.1.0	DevFile: 6.2.0	DevFile: 3.0.0	DevFile: 1.0.0	DevFile: 3.0.0	DevFile: 1.0.0	DevFile: 1.0.0	DevFile: 2.3.2-k
BARs: 511	BARs: esdadios11	BARs: x2timers2	BARs: sis8300s3	BARs: pciedevs4	BARs: tamc532s5	BARs: sis8160s7	BARs: tamc532s5	BARs: sis8160s7	BARs: sis8160s7	BARs: sis8160s7
255	65535	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W	PCIe-R/W
PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor	PCIe-monitor
Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev	Bind pciedev
INFO	INFO	INFO	INFO	INFO	INFO	INFO	INFO	INFO	INFO	INFO
LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI	LSPCI

- The source codes can be found on <https://github.com/MicroTCA>
- The information and Linux packages can be found on a DOOCS web page <http://doocs.desy.de>
- Mail [*doocs@desy.de*](mailto:doocs@desy.de)

THANK YOU