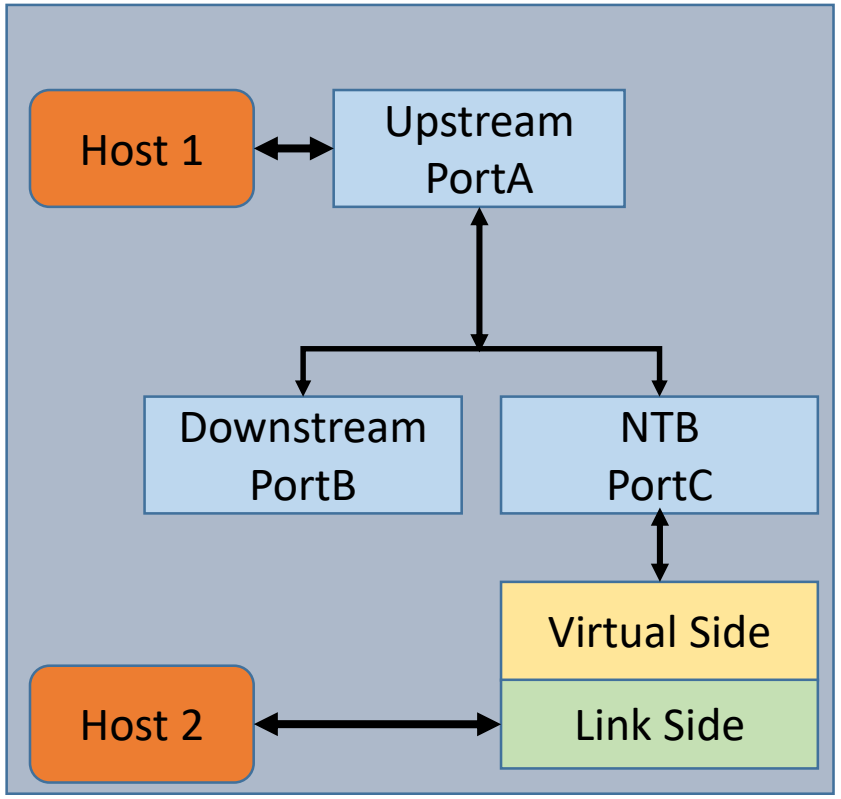
REDUNDANT CPU ON MTCA SYSTEM WITH PCI EXPRESS NON TRANSPARENT BRIDGE

L.Petrosyan (DESY)

PCI Express NTB

- 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



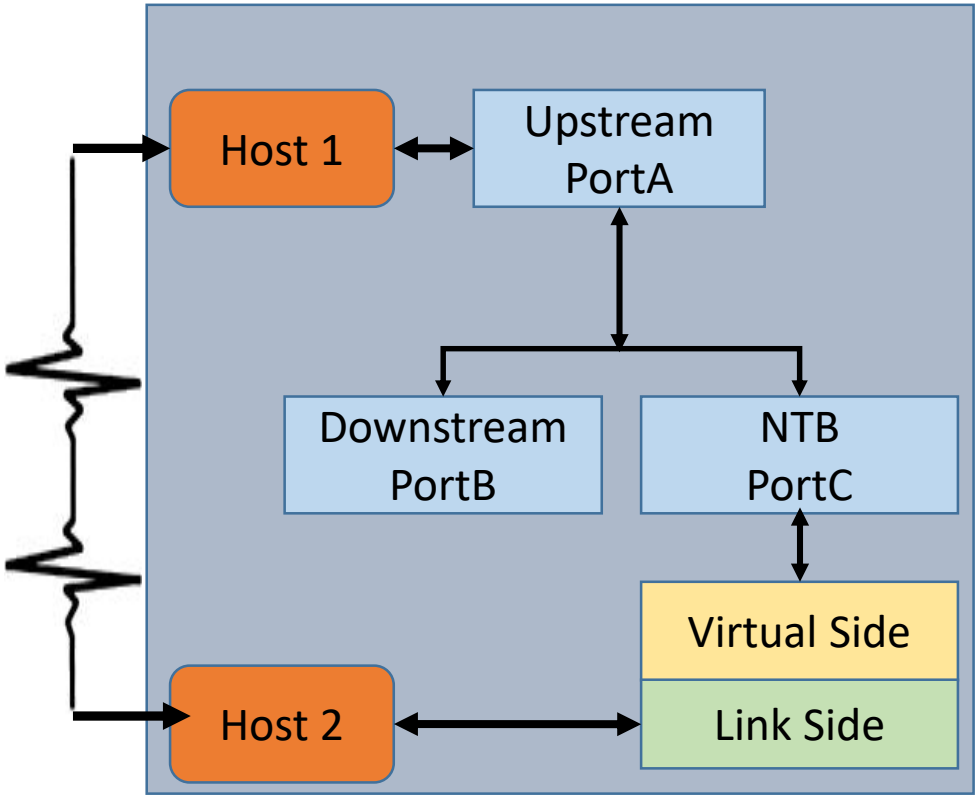
NTB Provides:

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

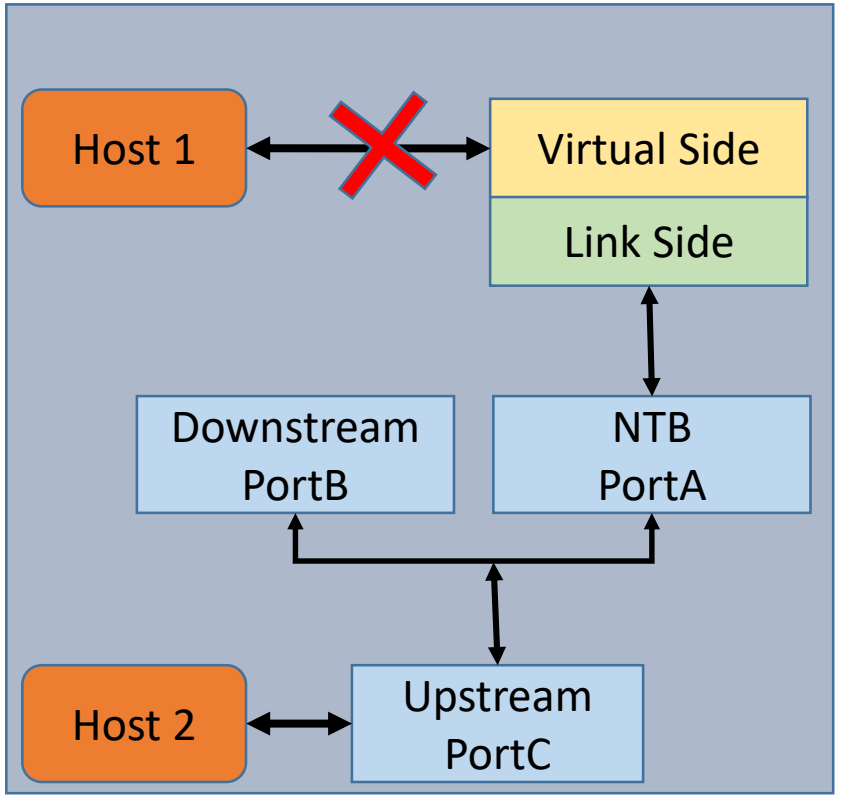
NTB FAILOVER

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 Hos1t 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.



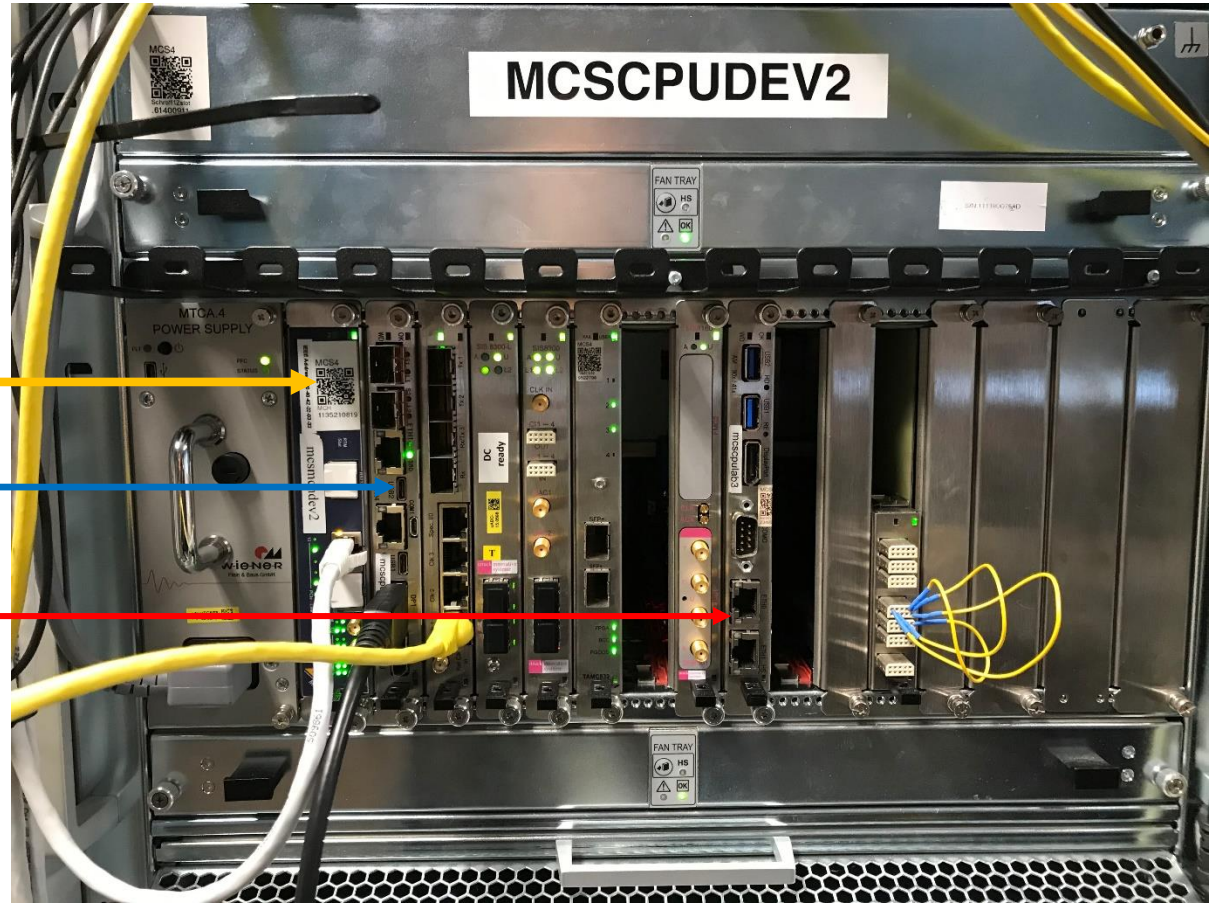
Redundant CPU Test

NAT MCH with PEX8748 PCIE Switch

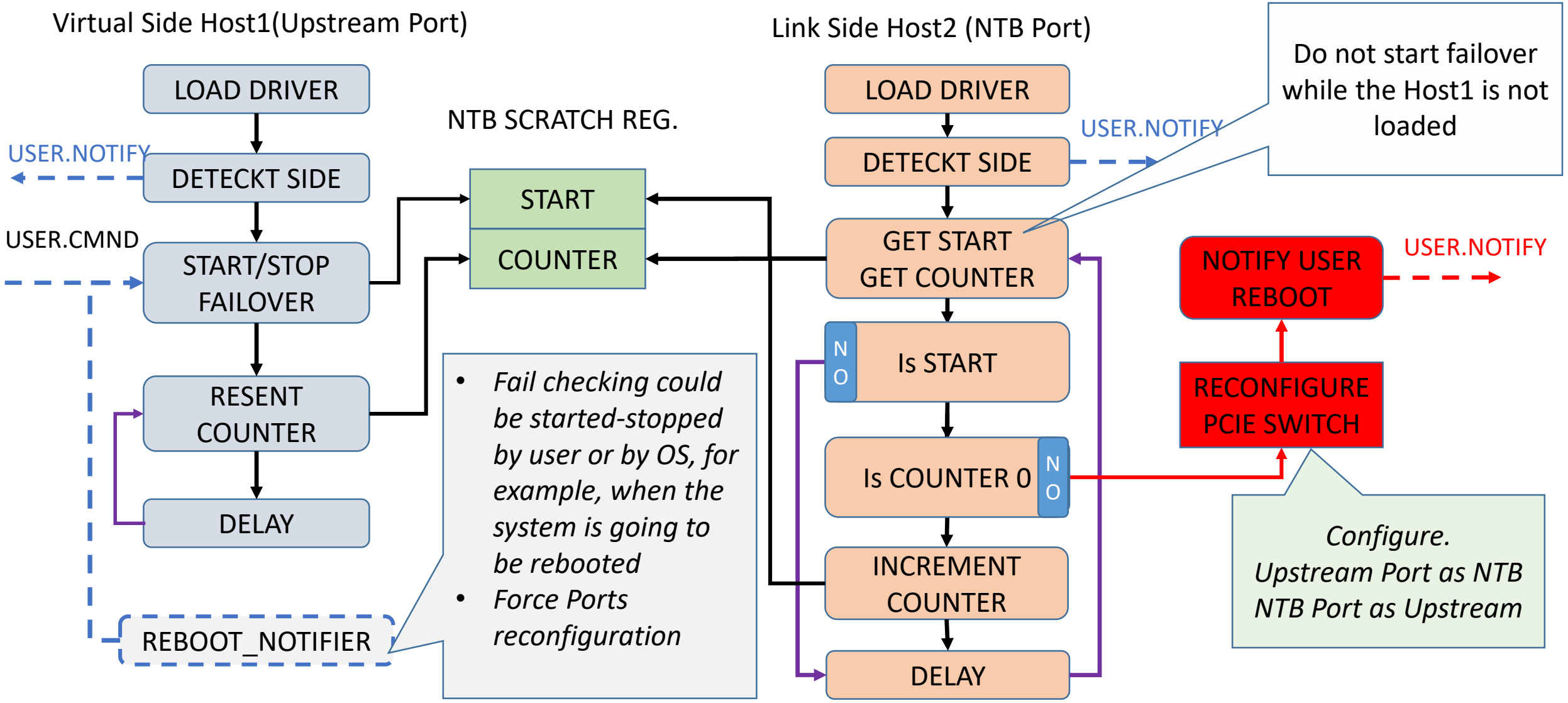
Host1 in Slot 1 (PCIE Switch Port19)

Host2 in Slot 8 (PCIE Switch Port16)

PCIENTBV – PCIE NTB device driver is running on both CPUs

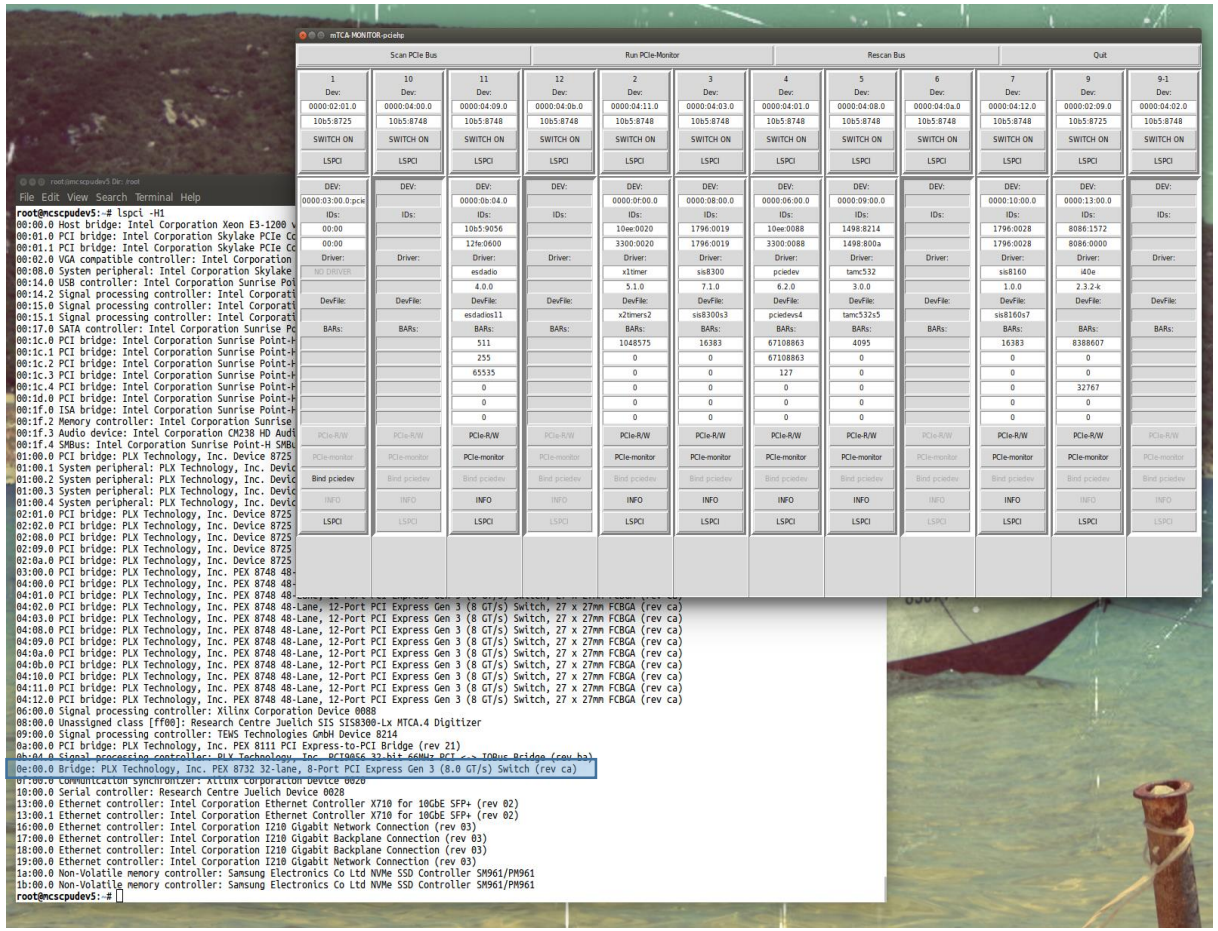


Redundant CPU Test

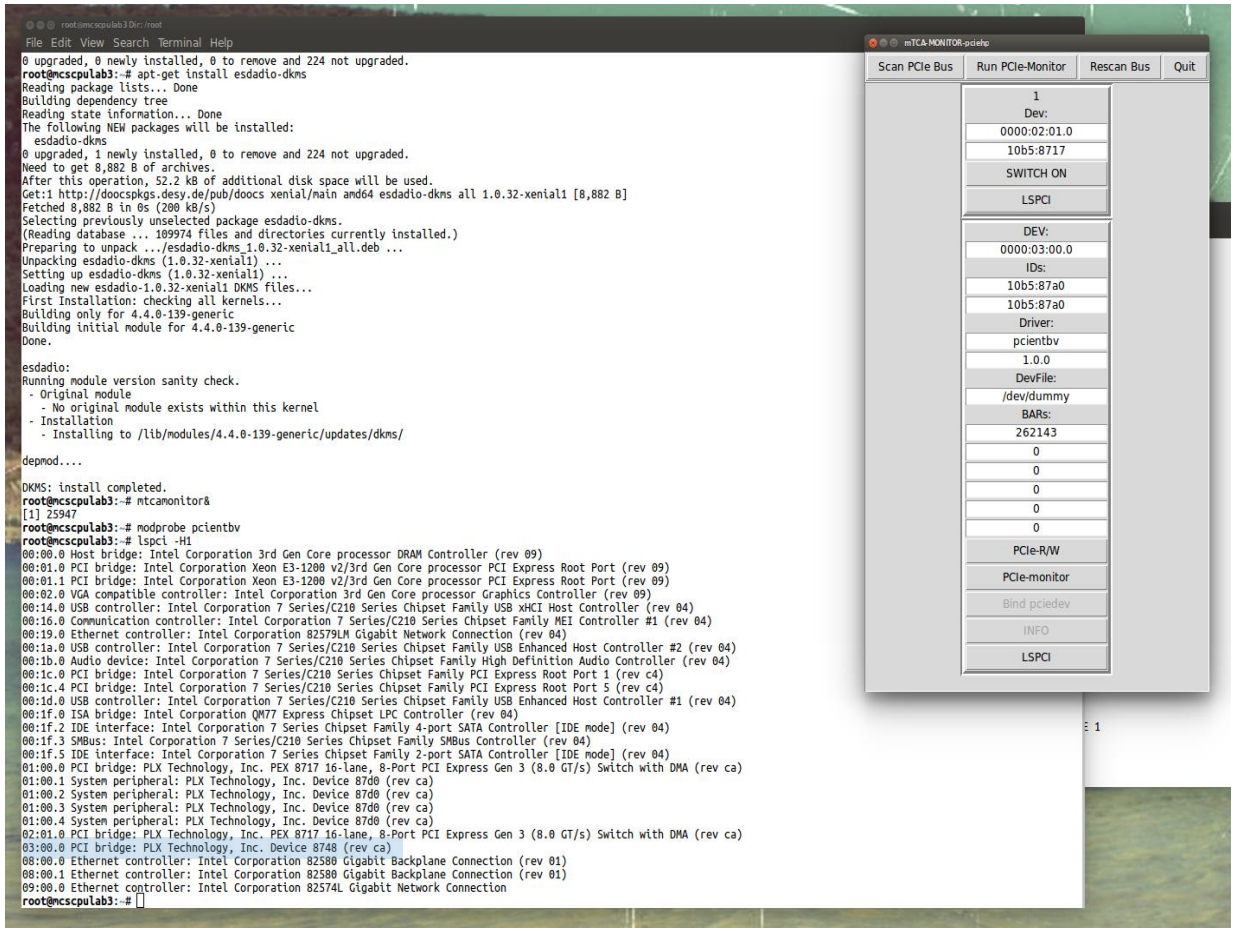


Redundant CPU Test

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



Redundant CPU Test

Heartbeats messages on the Link Side Host2 NTB devise driver

Host1 is OK

Host1 Fail

```
File  Edit  View  Search  Terminal  Help
Oct 24 14:52:54 mcscpulab3 kernel: [184590.463035] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG 0
Oct 24 14:52:56 mcscpulab3 kernel: [184592.438573] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341040497 DATA2
Oct 24 14:52:56 mcscpulab3 kernel: [184592.446820] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341040499 DATA 2 LINK SIDE 1
Oct 24 14:52:56 mcscpulab3 kernel: [184592.456084] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:52:56 mcscpulab3 kernel: [184592.463006] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG 1
Oct 24 14:52:58 mcscpulab3 kernel: [184594.438577] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341040997 DATA2
Oct 24 14:52:58 mcscpulab3 kernel: [184594.446582] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341040999 DATA 2 LINK SIDE 1
Oct 24 14:52:58 mcscpulab3 kernel: [184594.455633] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:52:58 mcscpulab3 kernel: [184594.462378] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG 1
Oct 24 14:53:00 mcscpulab3 kernel: [184596.438568] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341041497 DATA2
Oct 24 14:53:00 mcscpulab3 kernel: [184596.447089] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341041498 DATA 2 LINK SIDE 1
Oct 24 14:53:00 mcscpulab3 kernel: [184596.456667] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:53:00 mcscpulab3 kernel: [184596.463982] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG 1
Oct 24 14:53:02 mcscpulab3 kernel: [184598.438527] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341041997 DATA2
Oct 24 14:53:02 mcscpulab3 kernel: [184598.447215] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341041998 DATA 2 LINK SIDE 1
Oct 24 14:53:02 mcscpulab3 kernel: [184598.456944] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:53:02 mcscpulab3 kernel: [184598.464396] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG 1
Oct 24 14:53:04 mcscpulab3 kernel: [184600.438547] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341042497 DATA2
Oct 24 14:53:04 mcscpulab3 kernel: [184600.447399] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341042498 DATA 2 LINK SIDE 1
Oct 24 14:53:04 mcscpulab3 kernel: [184600.457294] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:53:04 mcscpulab3 kernel: [184600.464883] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG 1
Oct 24 14:53:04 mcscpulab3 kernel: [184600.473570] XXXXXXXXXXXXXXXXXPCIEDEV_TIMER:   VIRTUAL SIDE FAIL 4
Oct 24 14:53:04 mcscpulab3 kernel: [184600.528316] DMAR: DRHD: handling fault status reg 2
Oct 24 14:53:04 mcscpulab3 kernel: [184600.535535] DMAR: INTR-REMAP: Request device [[0f:00.0] fault index a0
Oct 24 14:53:04 mcscpulab3 kernel: [184600.535535] INTR-REMAP:[fault reason 34] Present field in the IRTE entry is clear
Oct 24 14:53:06 mcscpulab3 kernel: [184602.438540] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341042997 DATA2
Oct 24 14:53:06 mcscpulab3 kernel: [184602.447523] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341042998 DATA 2 LINK SIDE 1
Oct 24 14:53:06 mcscpulab3 kernel: [184602.457519] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:53:06 mcscpulab3 kernel: [184602.465181] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG -1
Oct 24 14:53:06 mcscpulab3 kernel: [184602.474033] XXXXXXXXXXXXXXXXXPCIEDEV_TIMER:   VIRTUAL SIDE FAIL 5
Oct 24 14:53:08 mcscpulab3 kernel: [184604.438526] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341043497 DATA2
Oct 24 14:53:08 mcscpulab3 kernel: [184604.447093] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341043498 DATA 2 LINK SIDE 1
Oct 24 14:53:08 mcscpulab3 kernel: [184604.456681] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:53:08 mcscpulab3 kernel: [184604.463948] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG -1
Oct 24 14:53:08 mcscpulab3 kernel: [184604.472440] XXXXXXXXXXXXXXXXXPCIEDEV_TIMER:   VIRTUAL SIDE FAIL 6
Oct 24 14:53:10 mcscpulab3 kernel: [184606.438523] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341043997 DATA2
Oct 24 14:53:10 mcscpulab3 kernel: [184606.447723] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341043999 DATA 2 LINK SIDE 1
Oct 24 14:53:10 mcscpulab3 kernel: [184606.457951] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE 1
Oct 24 14:53:10 mcscpulab3 kernel: [184606.465873] LLLLLLLLLLLLLLLPCIEDEV_TIMER:   LINK SIDE  SCRATCH_REG -1
Oct 24 14:53:10 mcscpulab3 kernel: [184606.474962] XXXXXXXXXXXXXXXXXPCIEDEV_TIMER:   VIRTUAL SIDE FAIL 7
Oct 24 14:53:12 mcscpulab3 kernel: [184608.438517] $$$$$$$$$$$$$PCIEDEV_TIMER:   EXPIRED AT 4341044497 DATA2
```

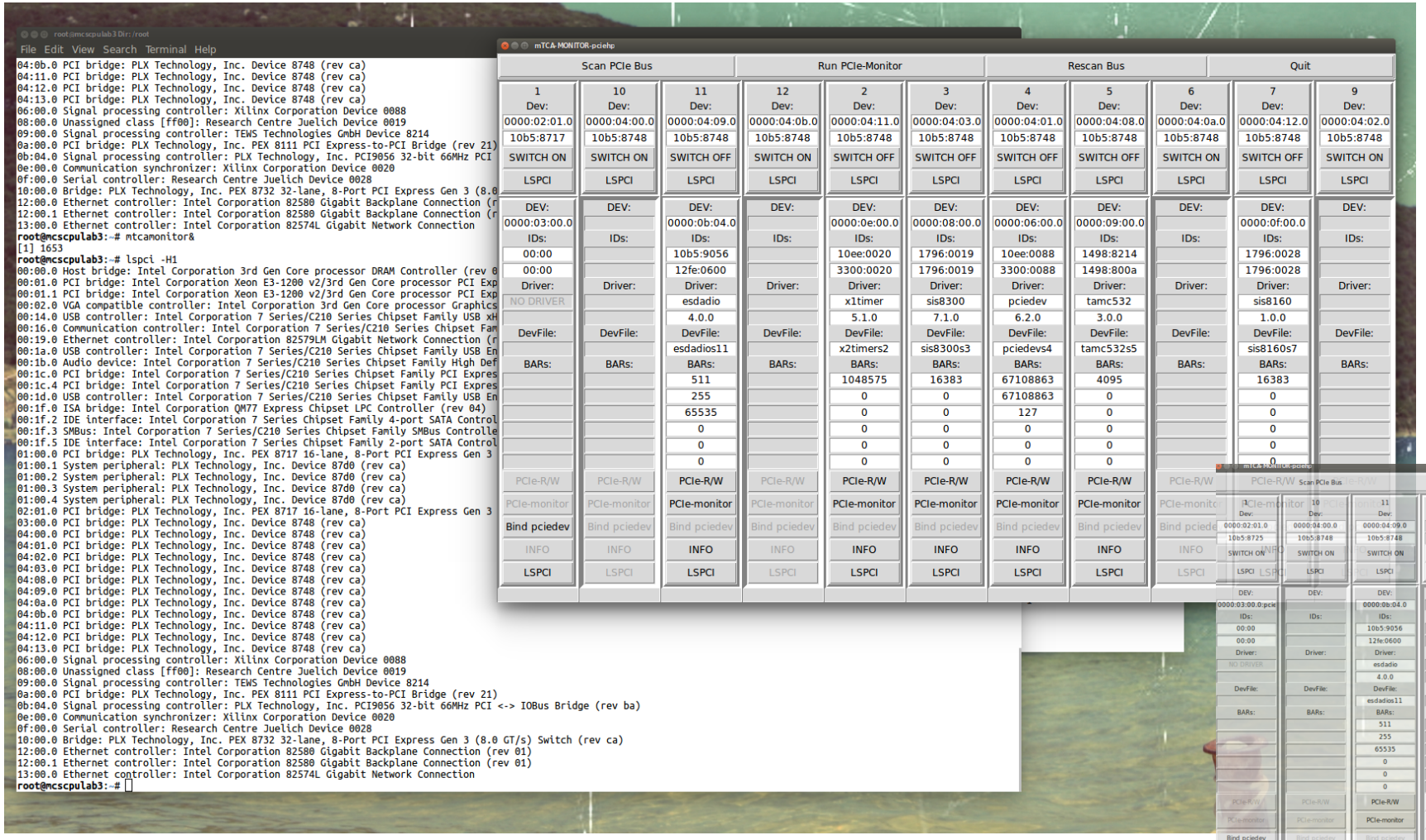
Using NAT MCH Diag tool, we can see the PCIE Switch Upstream Port Register is changed

```
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-c
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 'low
LSHM(0): FRU 9 sensor 79 LUN 0 '1V8' voltage 'lower non-critical go low' -deassertion

HUB REG 0x00000360 = 0x00003013
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):
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' -assertion
```

Port 19 Upstream, Port 16 NTB

Port 16 Upstream, Port 19 NTB

Redundant CPU Test



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

Info

- 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***

THANK YOU