eicSys GmbH

Embedded Integrated Control Systems

Saving Costs and Increasing Data Throughput in MicroTCA.4 by Hardware Concept Extension and Optimization of Data Transfer Performance

ONE TECHNOLOGY MULTIPLE SOLUTIONS

EicSys	Jalmuzna Wojciech	wojciech.jalmuzna@eicsys.eu
N.A.T.	Vollrath Dirksen	vollrath@nateurope.com
		mtca-helpdesk@desy.de
		support@mtca.eu

2013 ICALEPCS

Saving Costs & Increasing Data Throughput

Agenda

- Motivation
- Hardware Concept Extension
- Optimization of Data Transfer Performance
- Description of Life Demo
- Summary

eicSys GmbH
Embedded Integrated Control Systems

Motivation
Saving Costs and Increasing Data Throughput

- Desy uses MTCA.4, Core-i7, IO boards with FPGA and PCIexpress
- EicSys analyse such system for optimization and better reuse in other projects

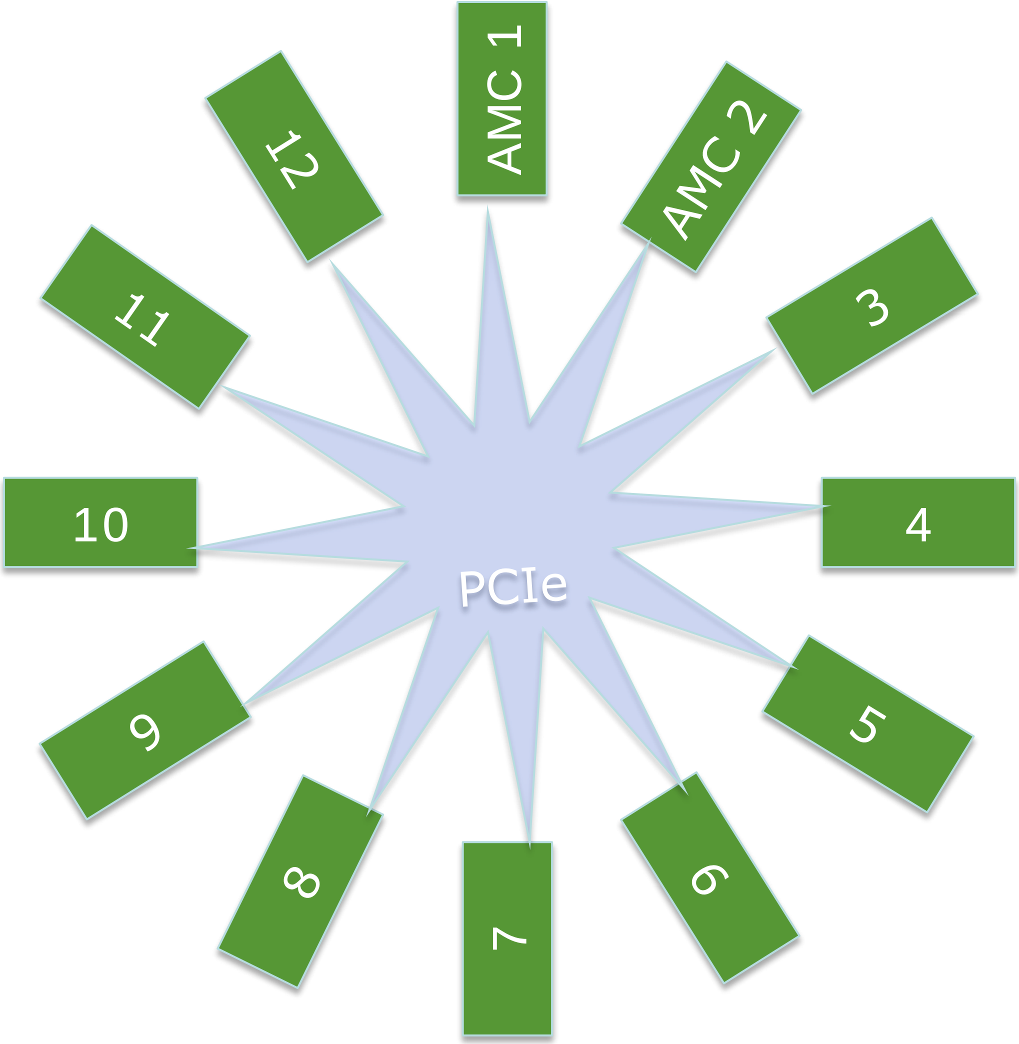
- Goal is to demonstrate how to
  - maximize the number of IO slots in MicroTCA.4 systems
  - save development time normally caused by different IO drivers
  - use different PCIexpress data xfr speeds in one system
  - optimize PCIexpress IO data bandwidth to CPU
  - save CPU time for data processing normally waste for IO reading
  - demonstrate data bottlenecks and solutions to overcome them

Saving Costs & Increasing Data Throughput

Agenda

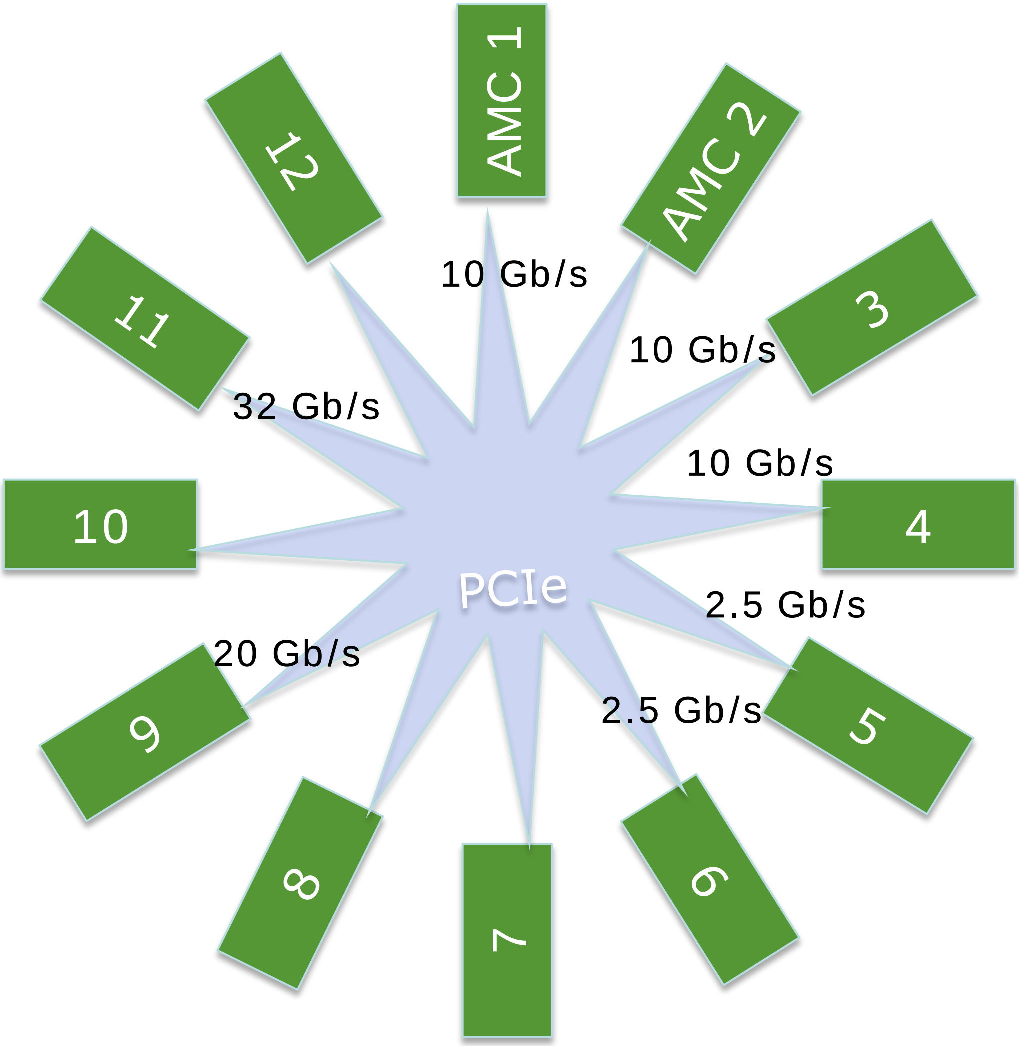
- Motivation
- Hardware Concept Extension
- Optimization of Data Transfer Performance
- Description of Life Demo
- Summary

Star Topology
IPMI, PCIe (SRIO, XAUI) and GbE



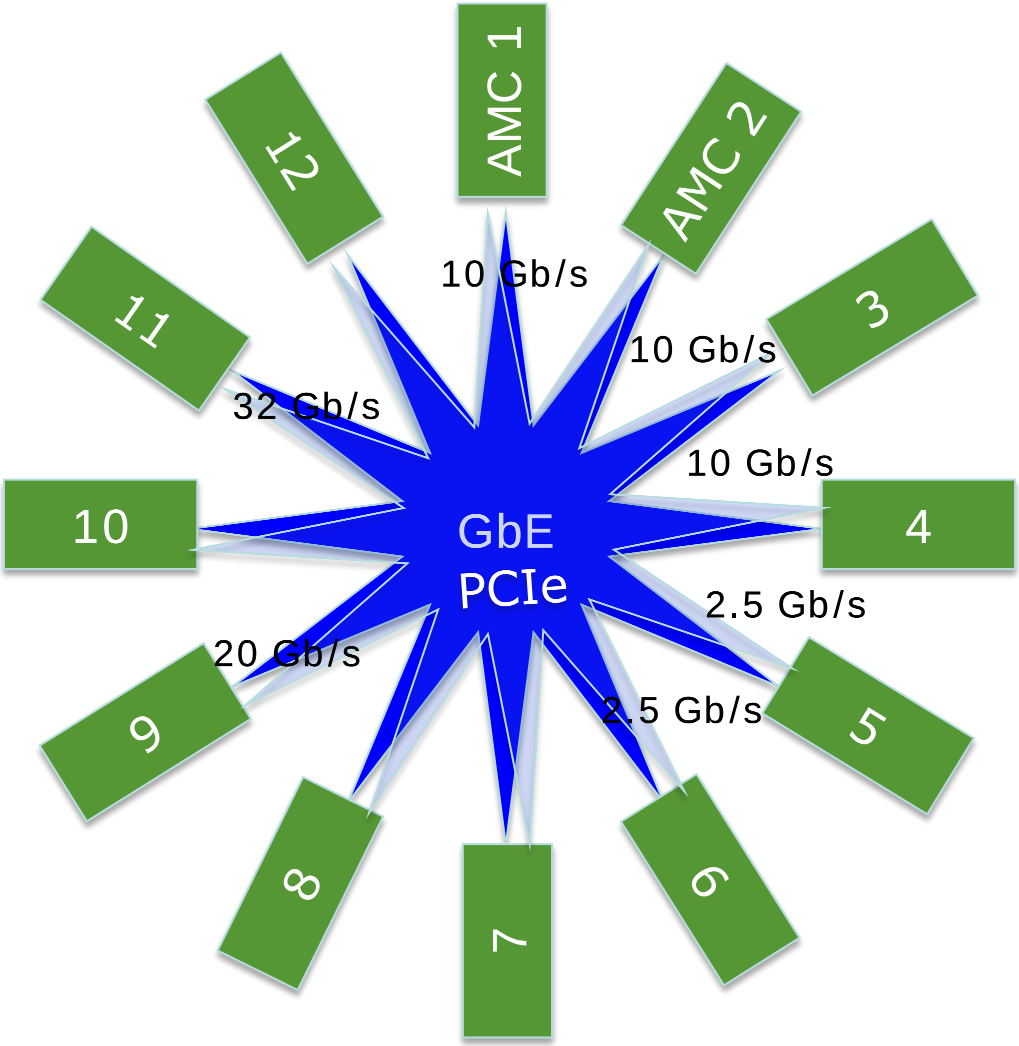
Star Topology

IPMI, PCIe (SRIO, XAUI) and GbE



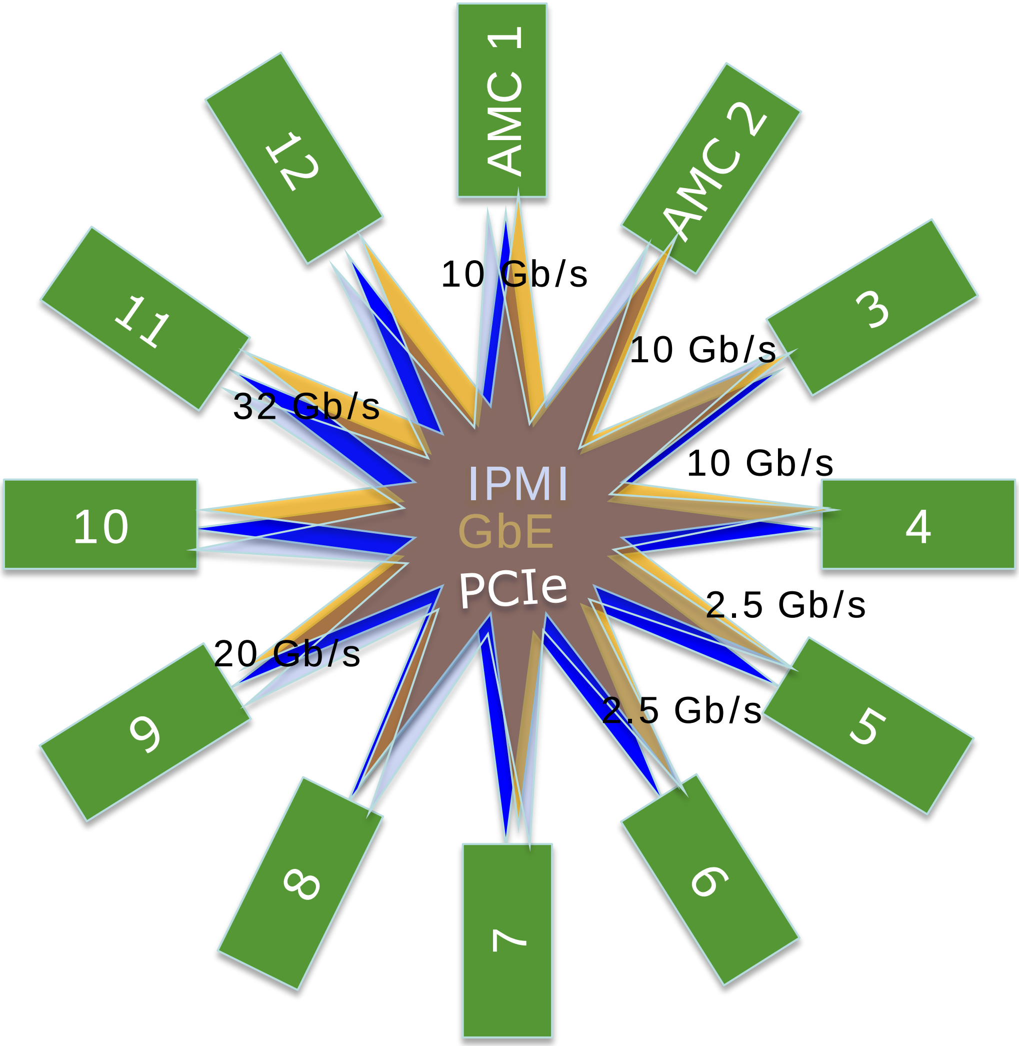
Star Topology

IPMI, PCIe (SRIO, XAUI) and GbE



Star Topology

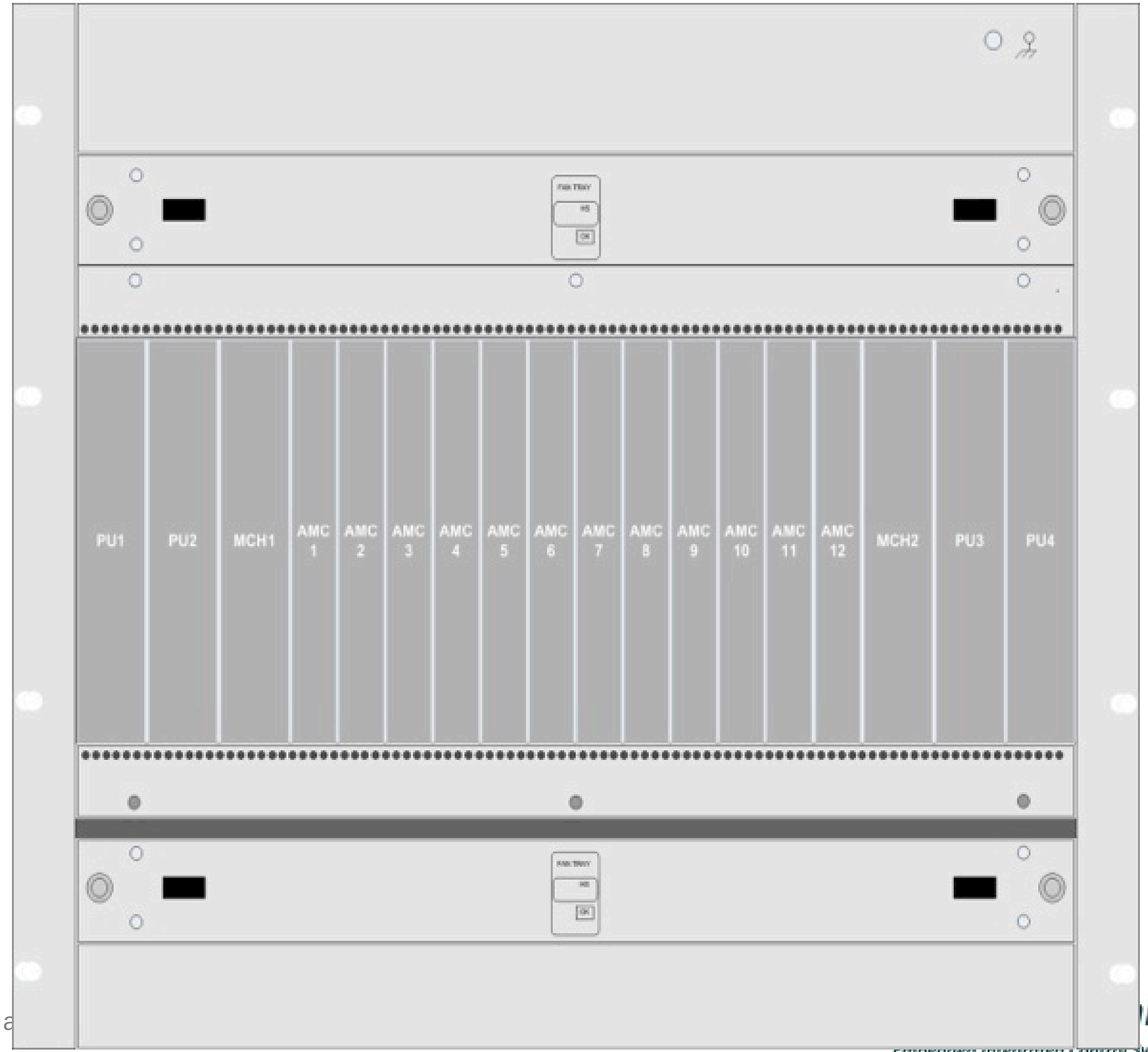
IPMI, PCIe (SRIO, XAUI) and GbE



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

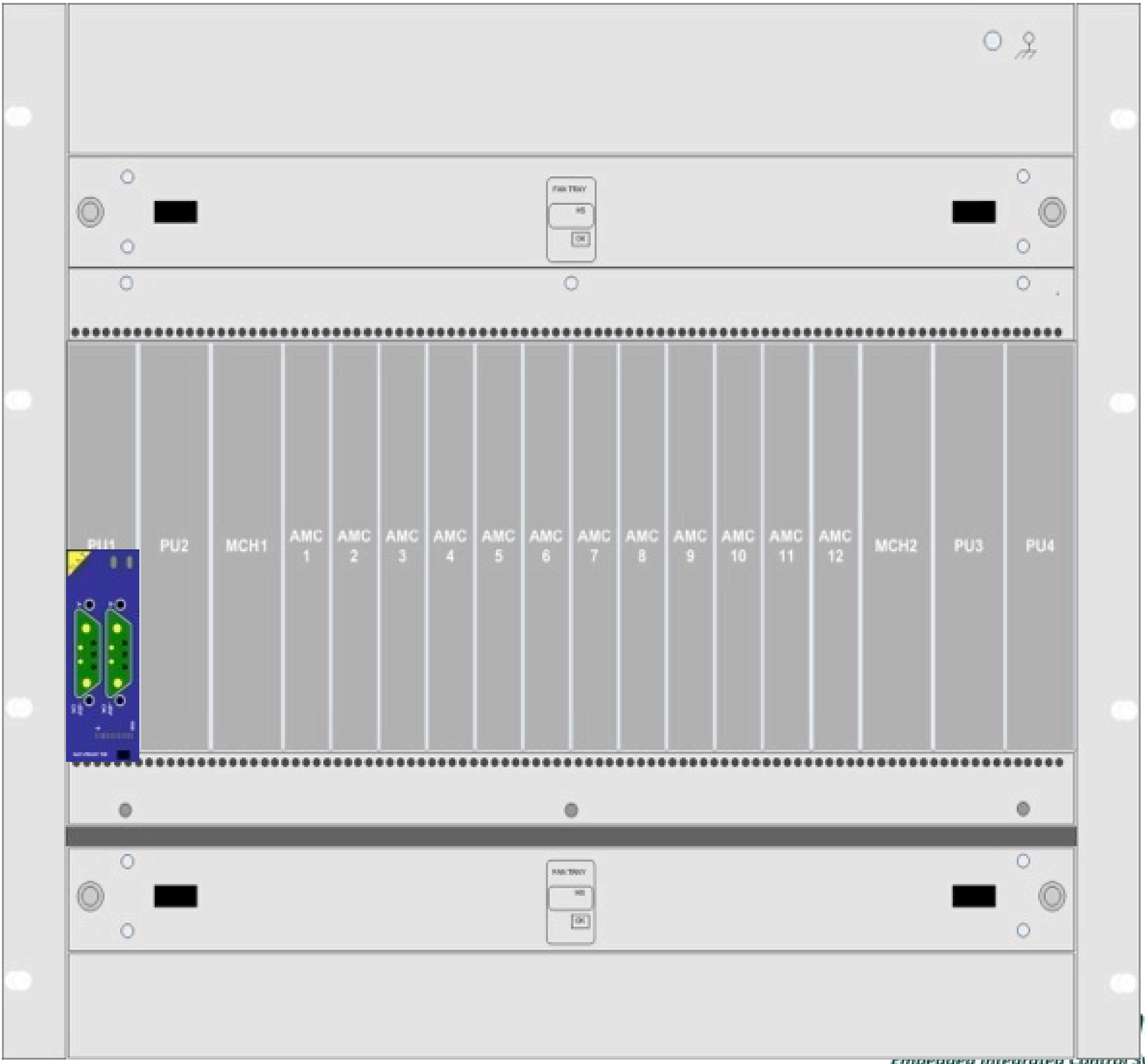
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,…

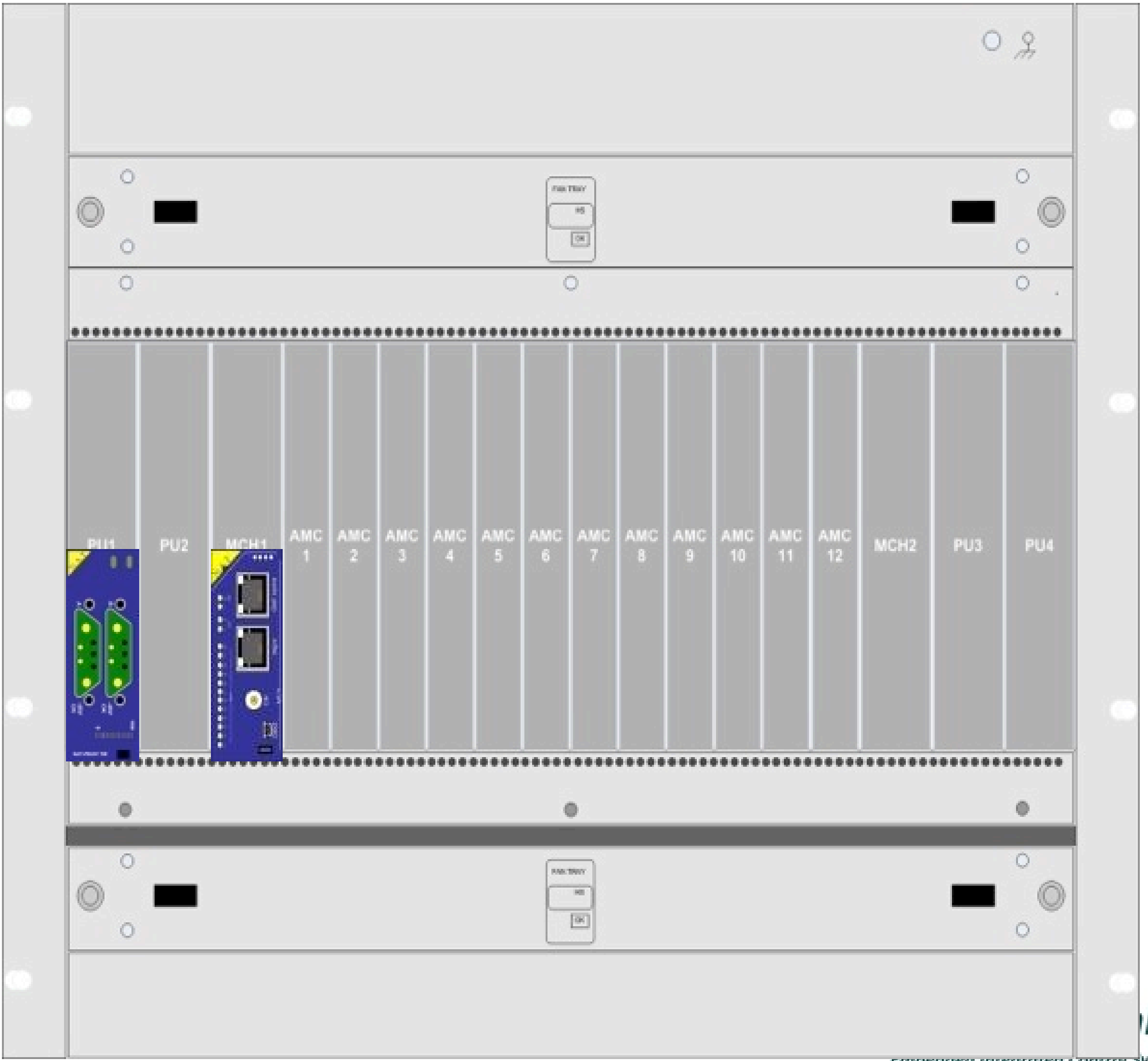
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

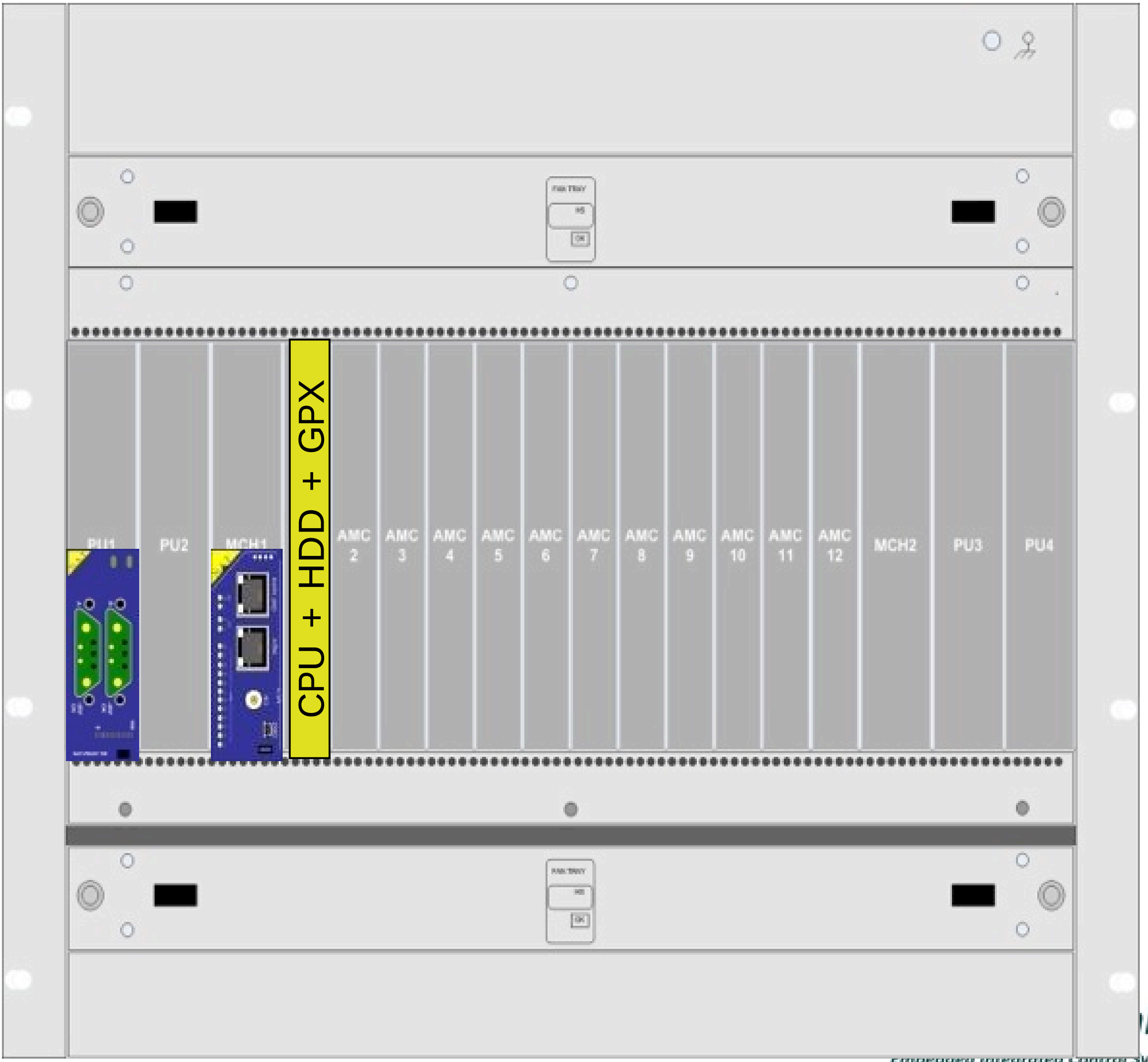
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

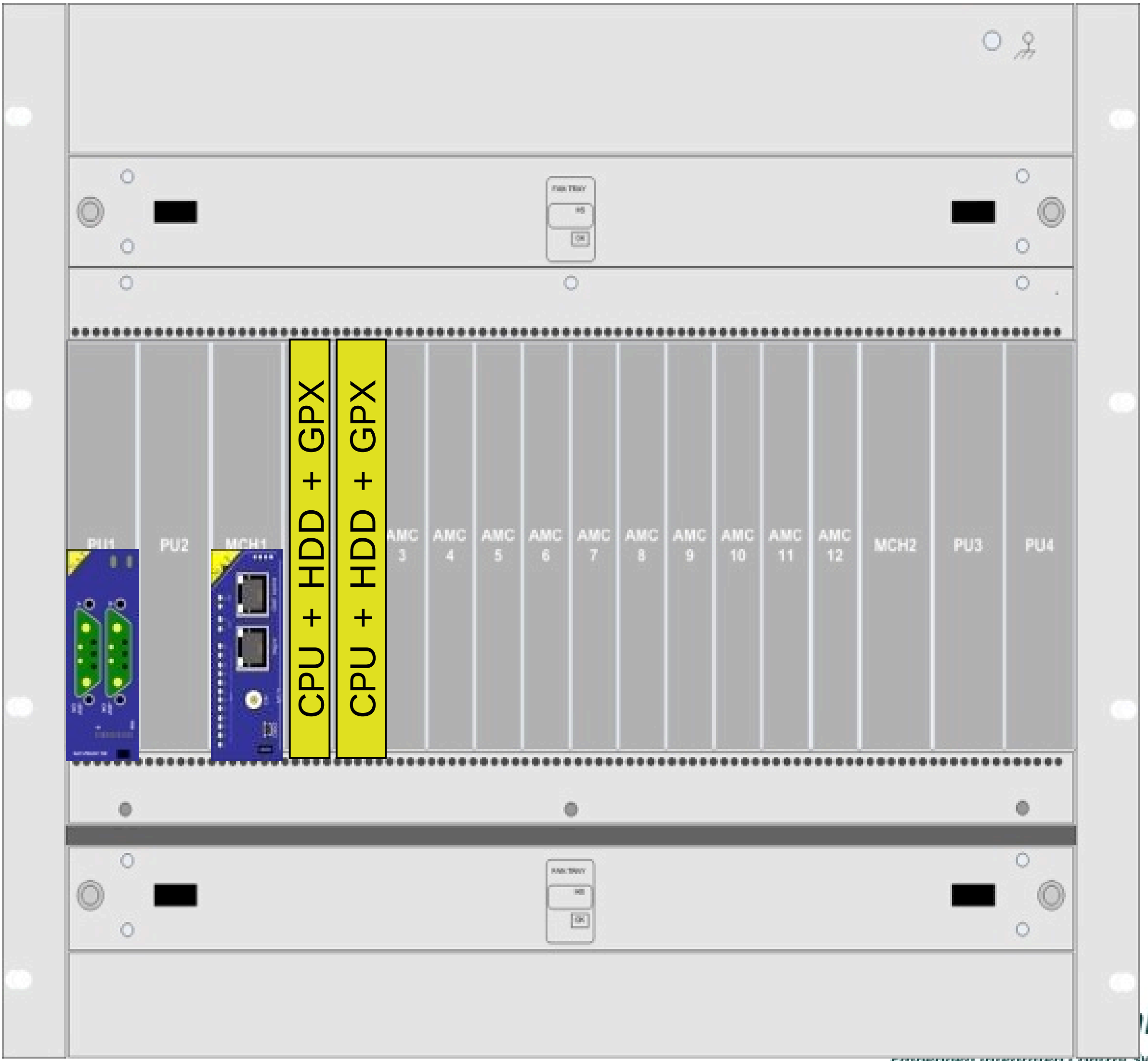
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

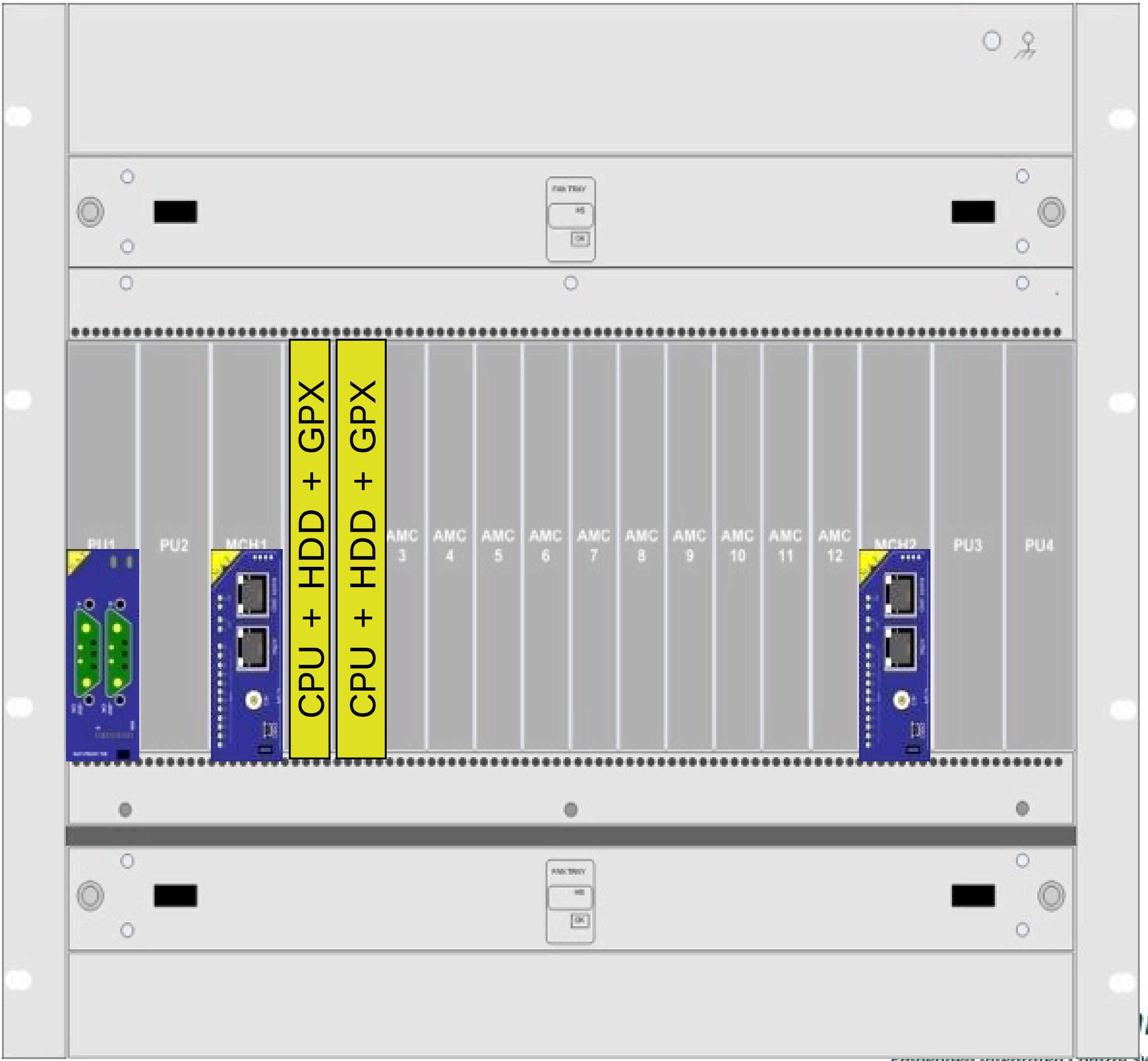
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

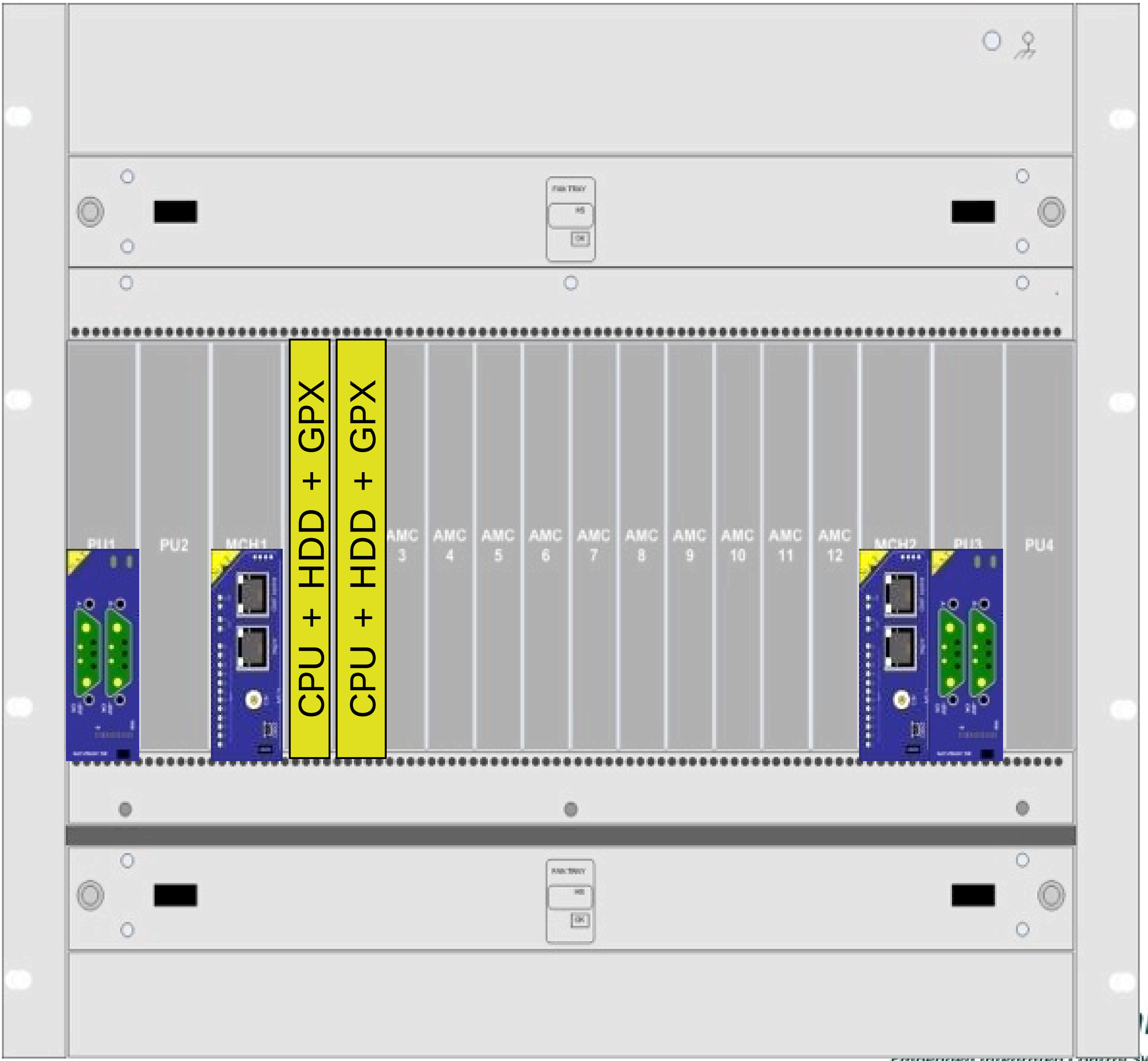
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

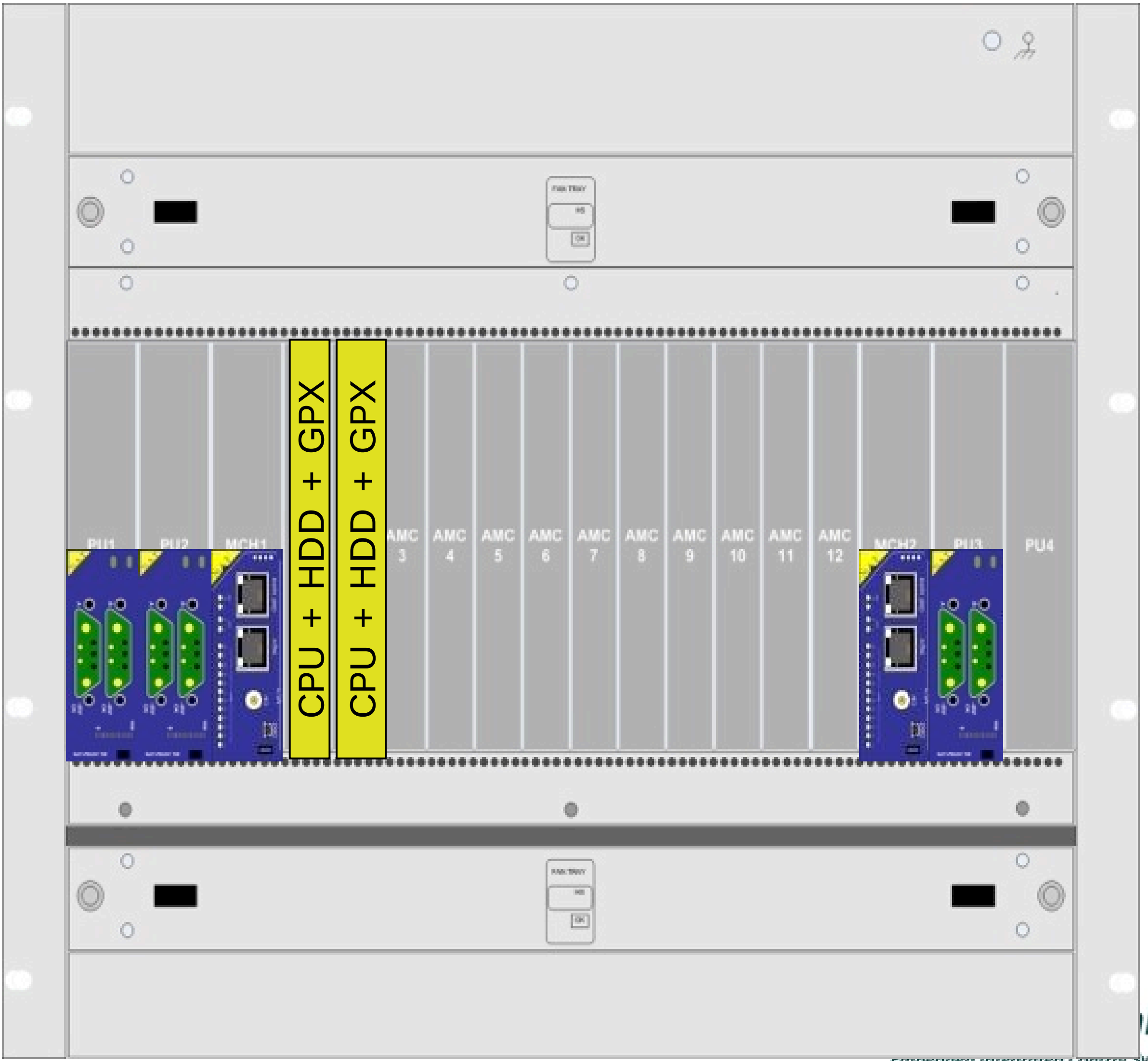
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

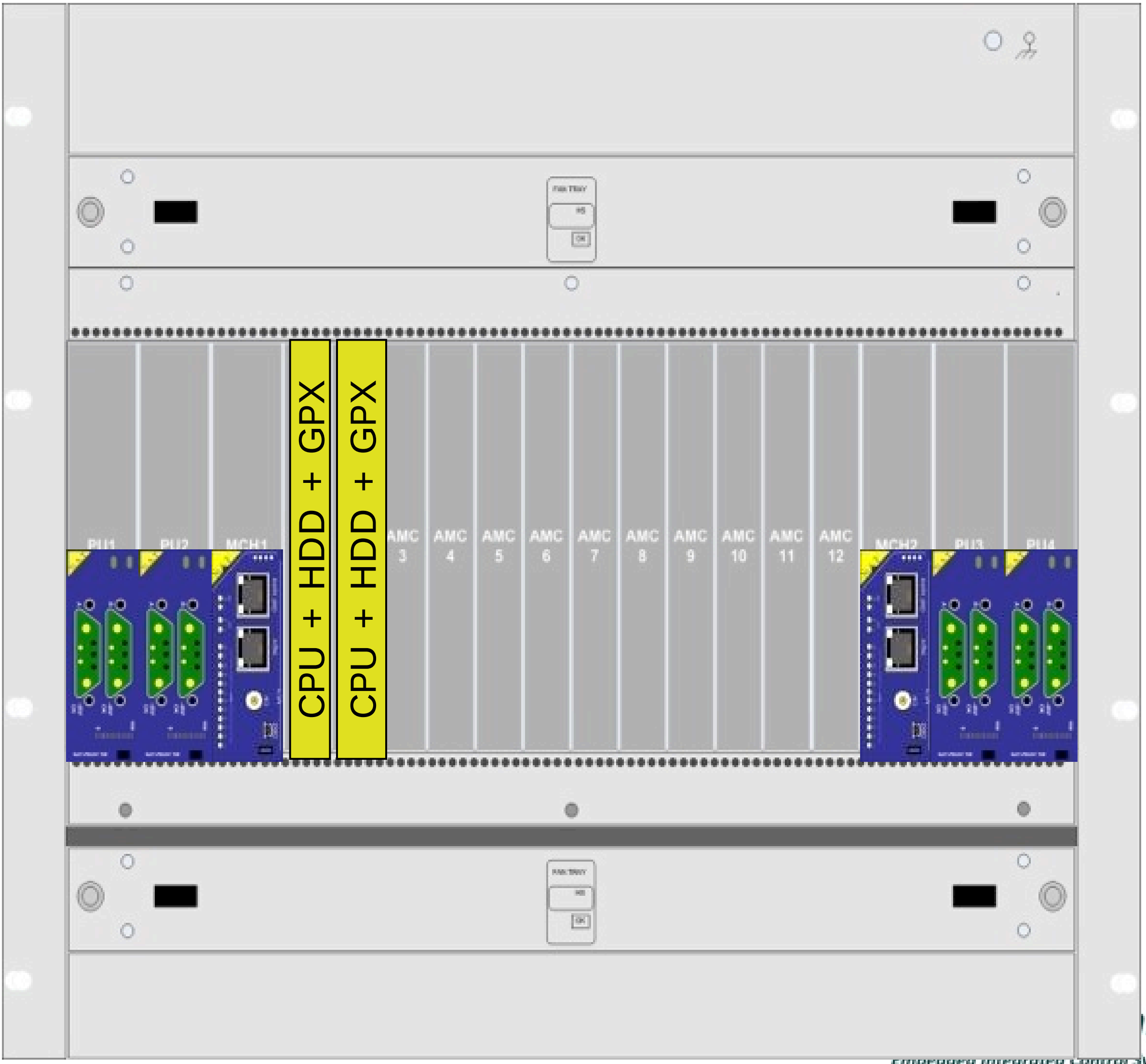
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

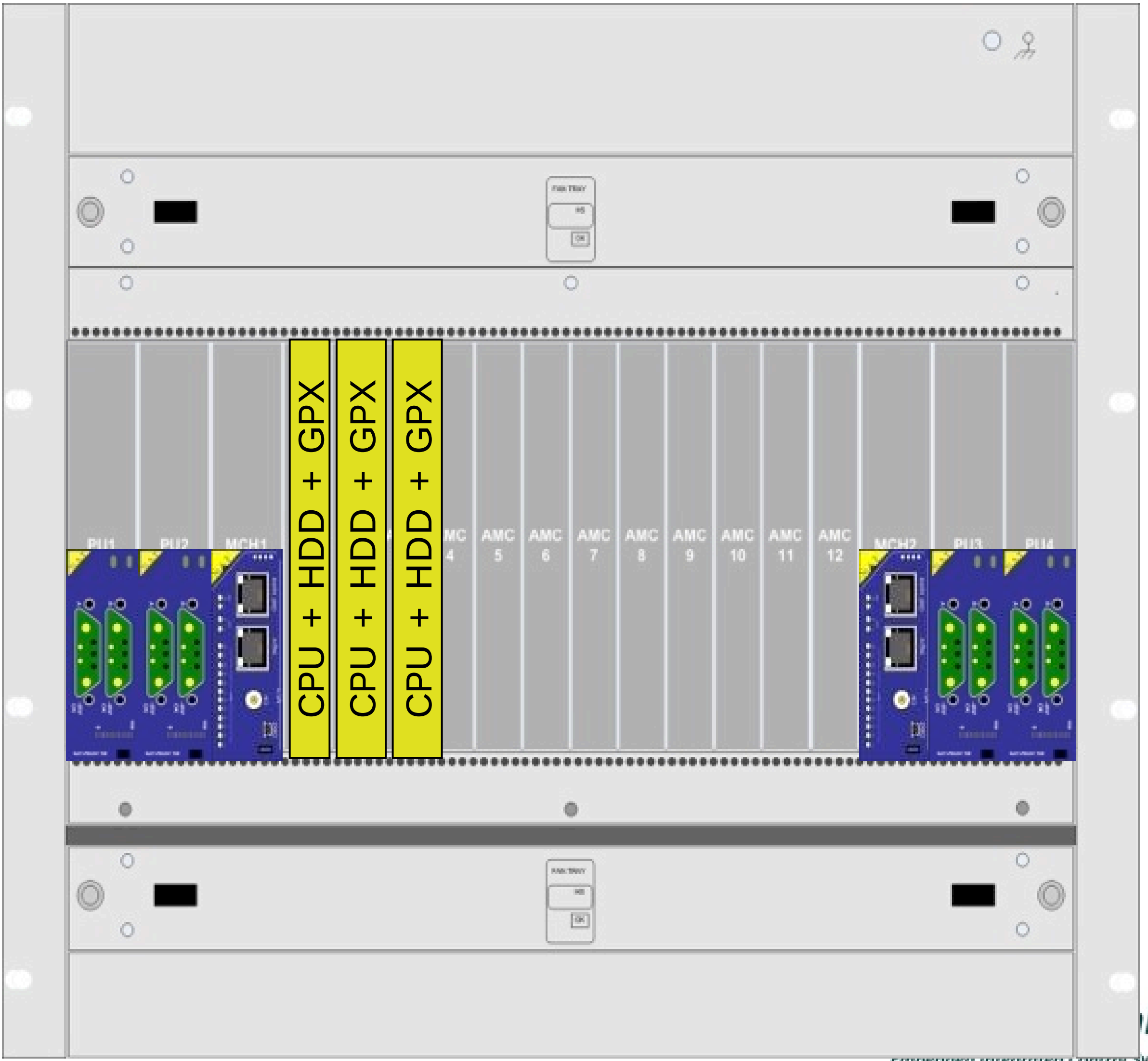
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

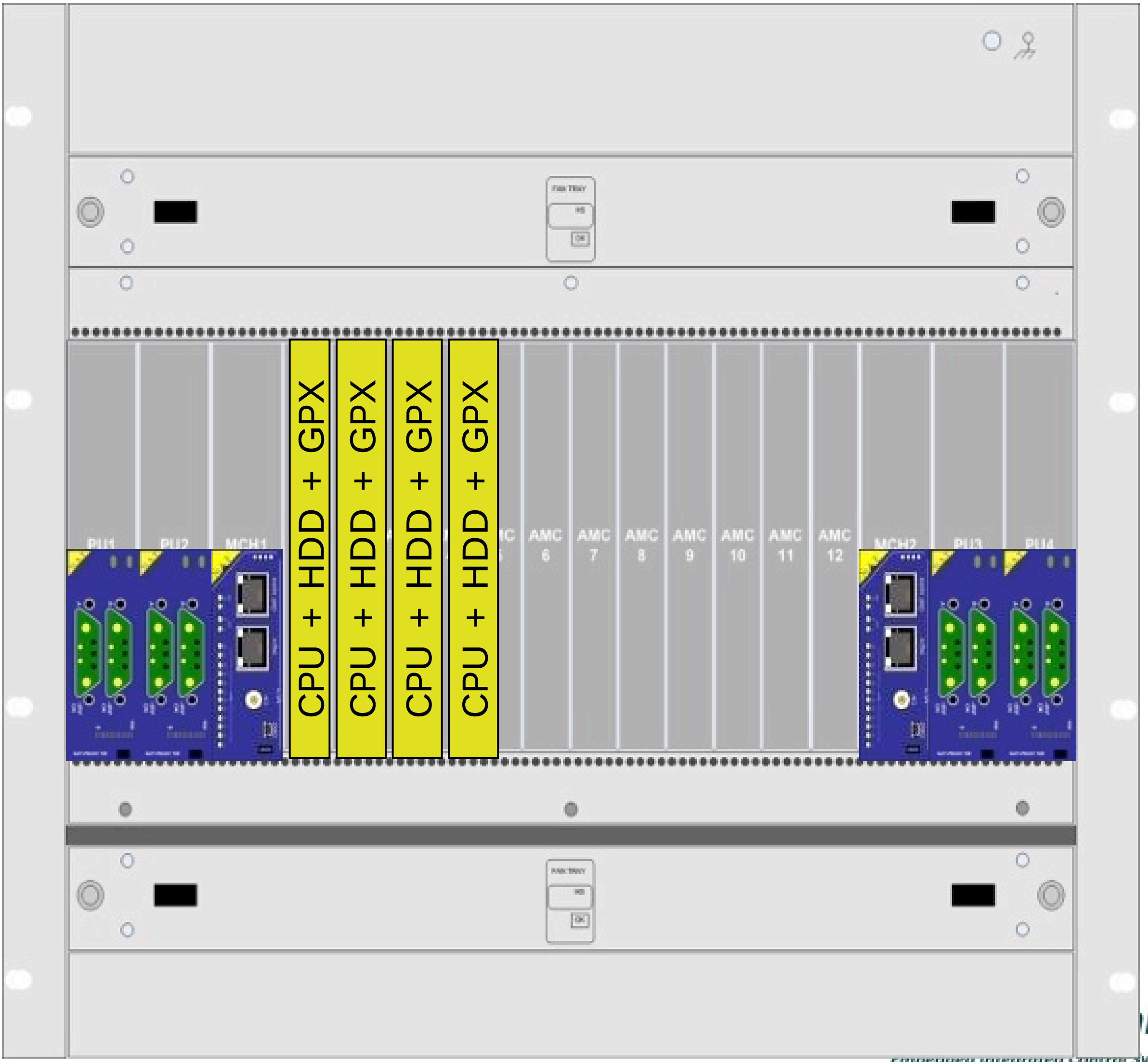
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

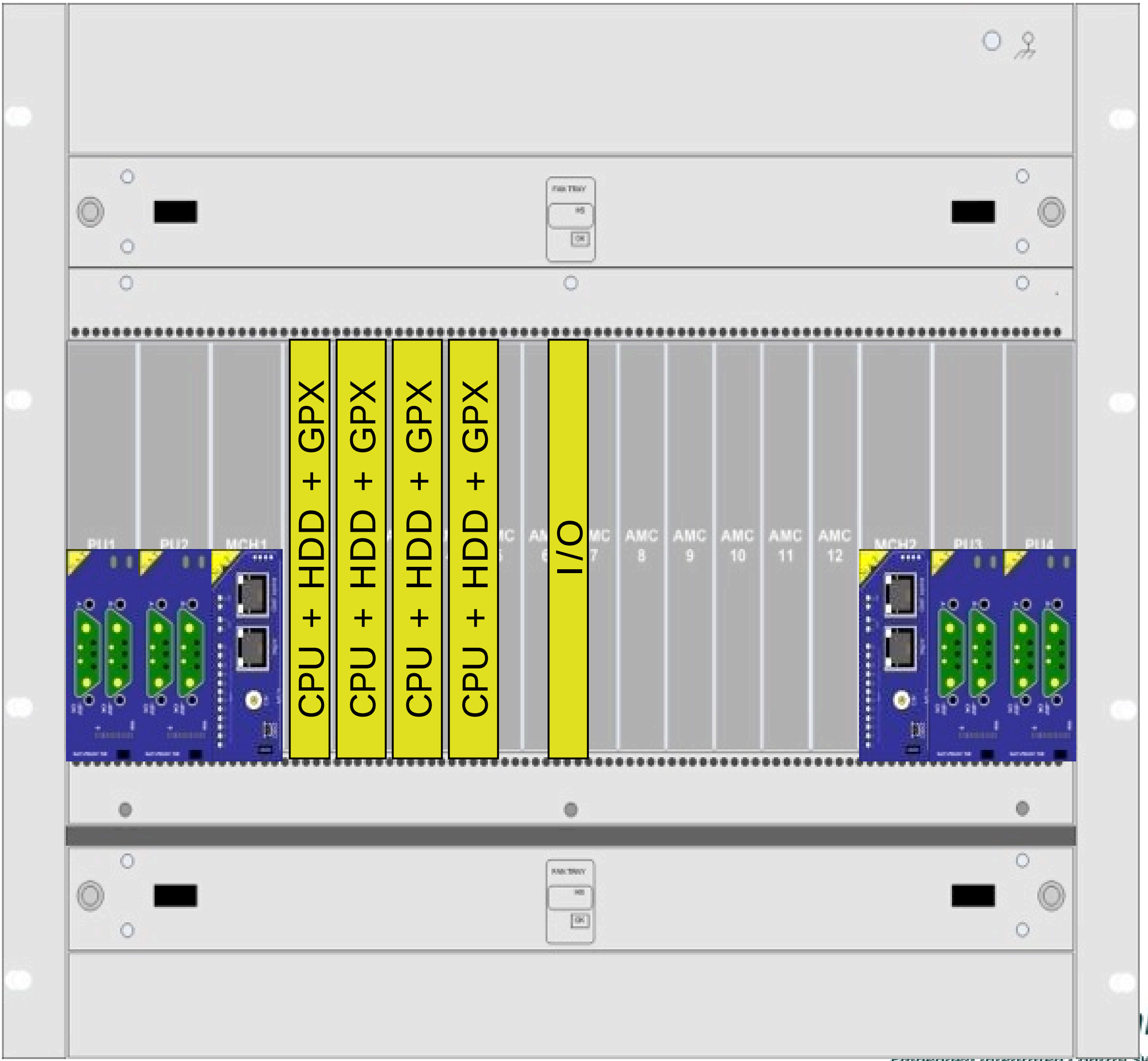
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

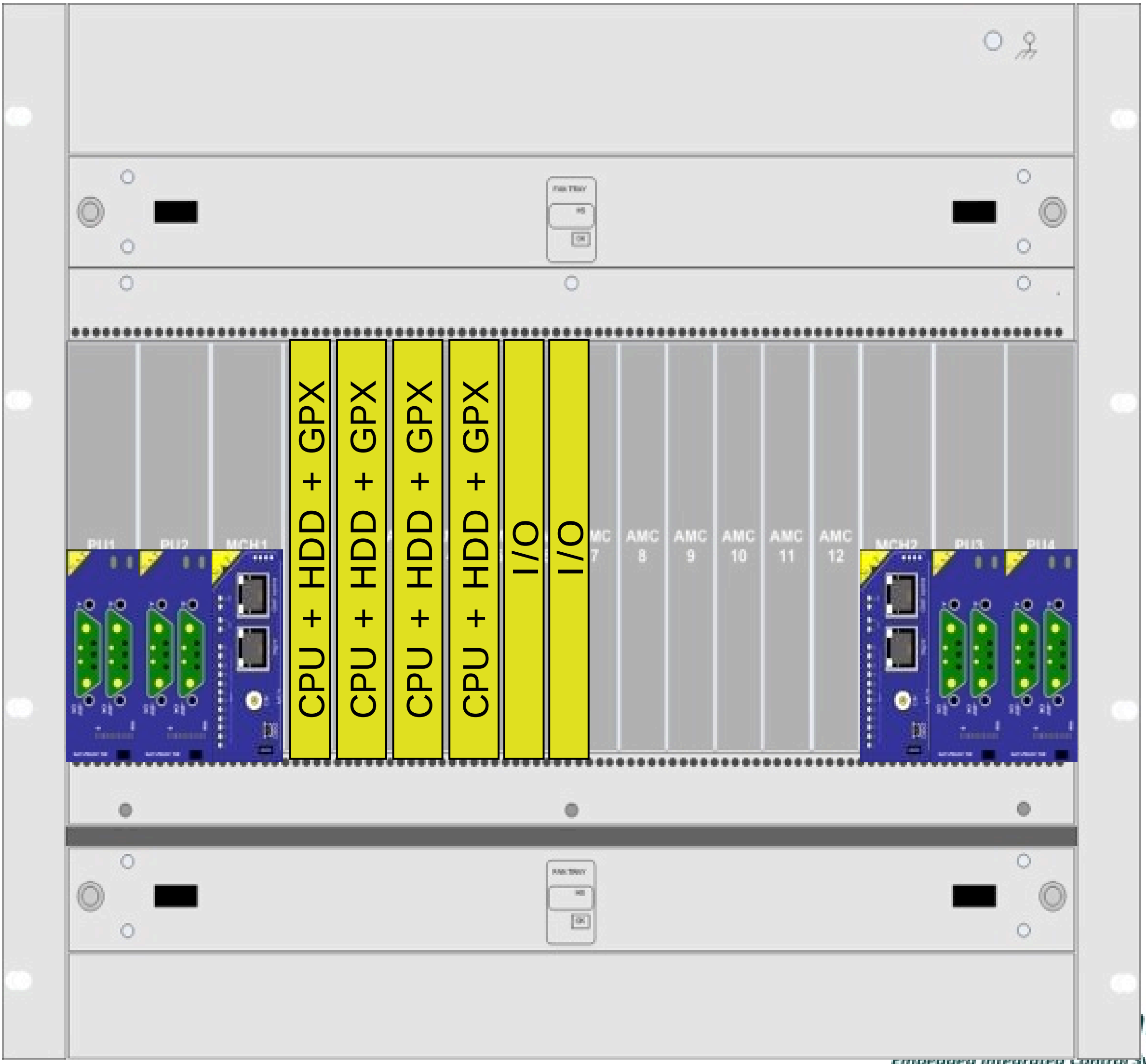
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

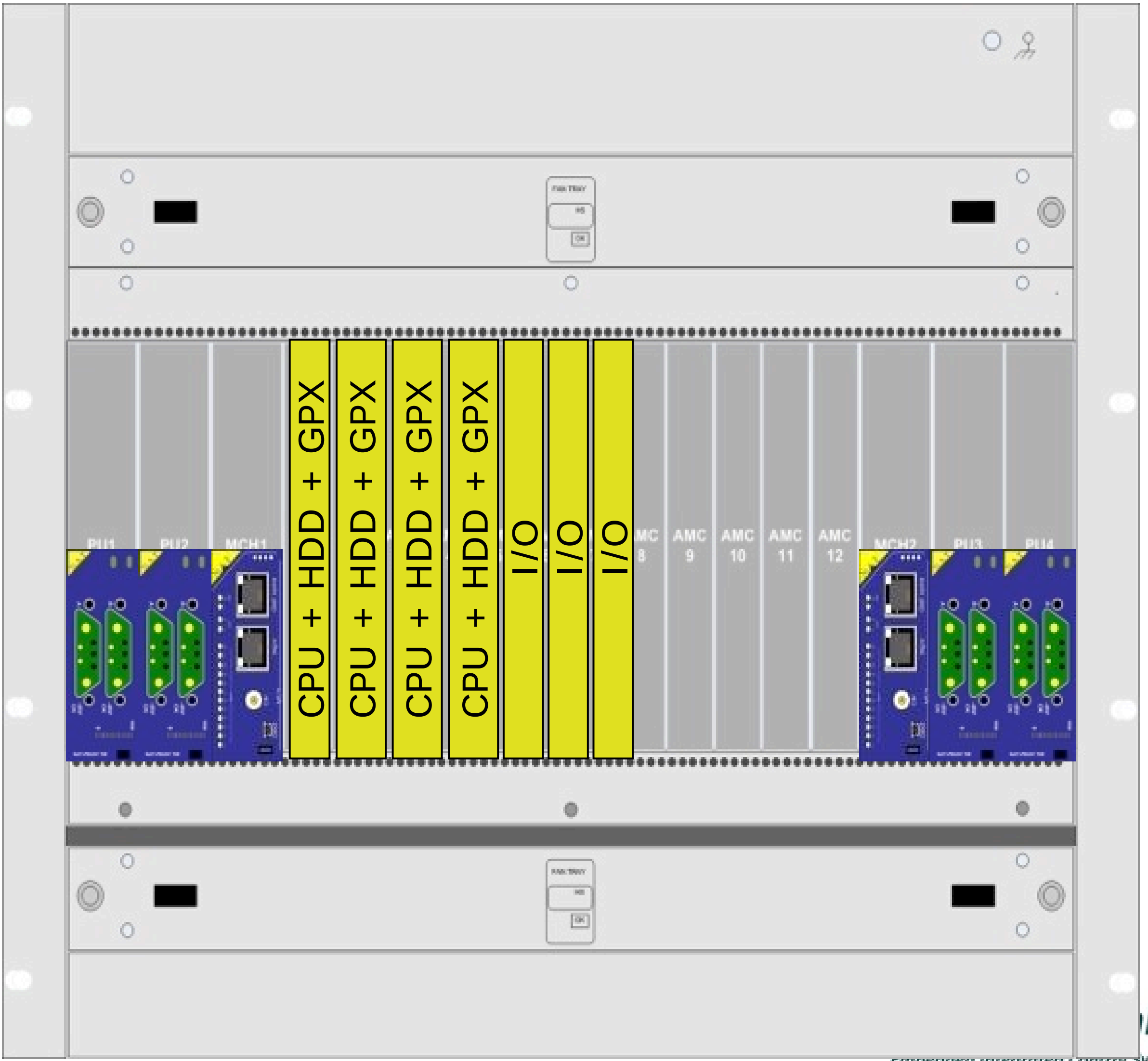
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

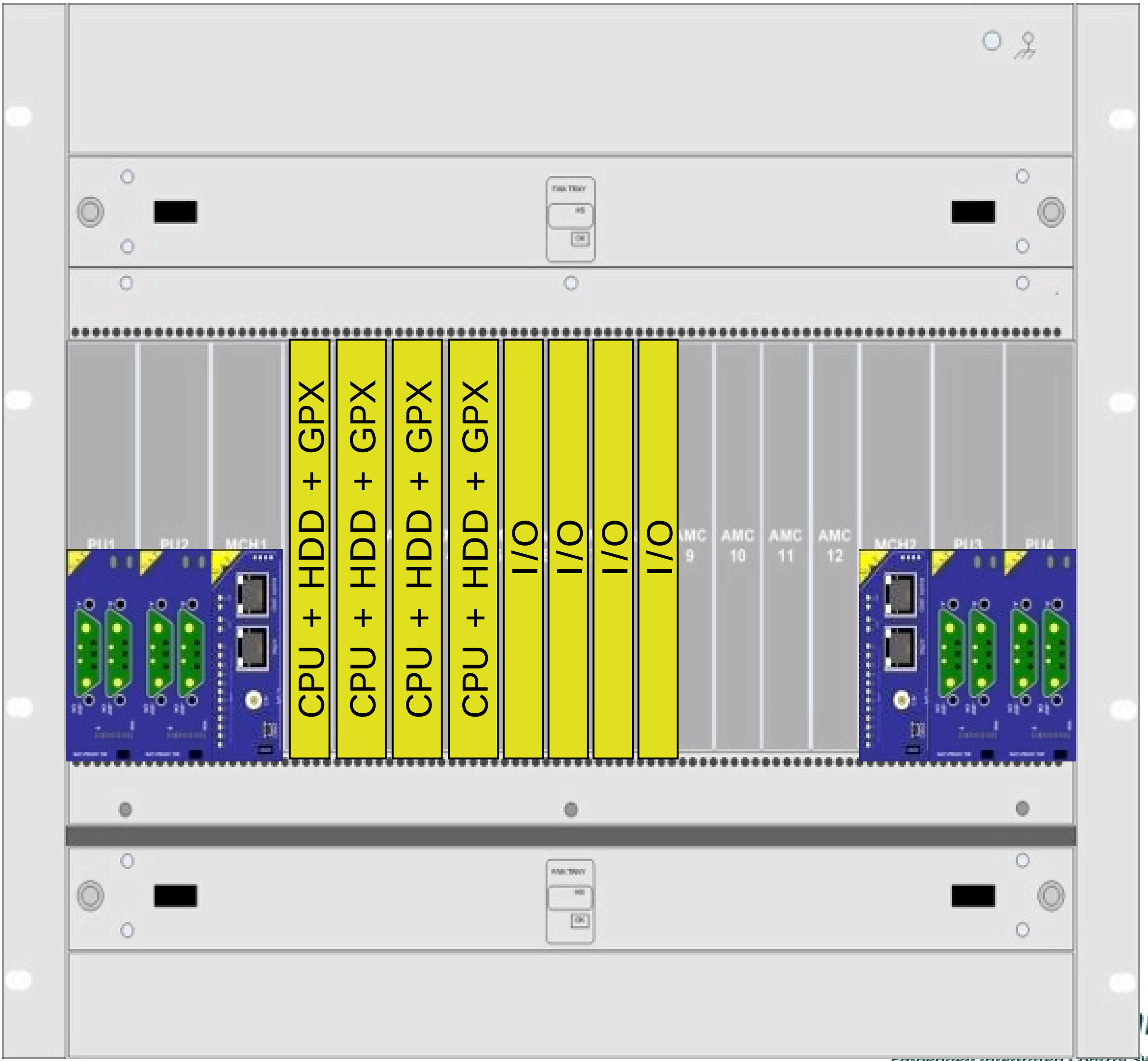
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

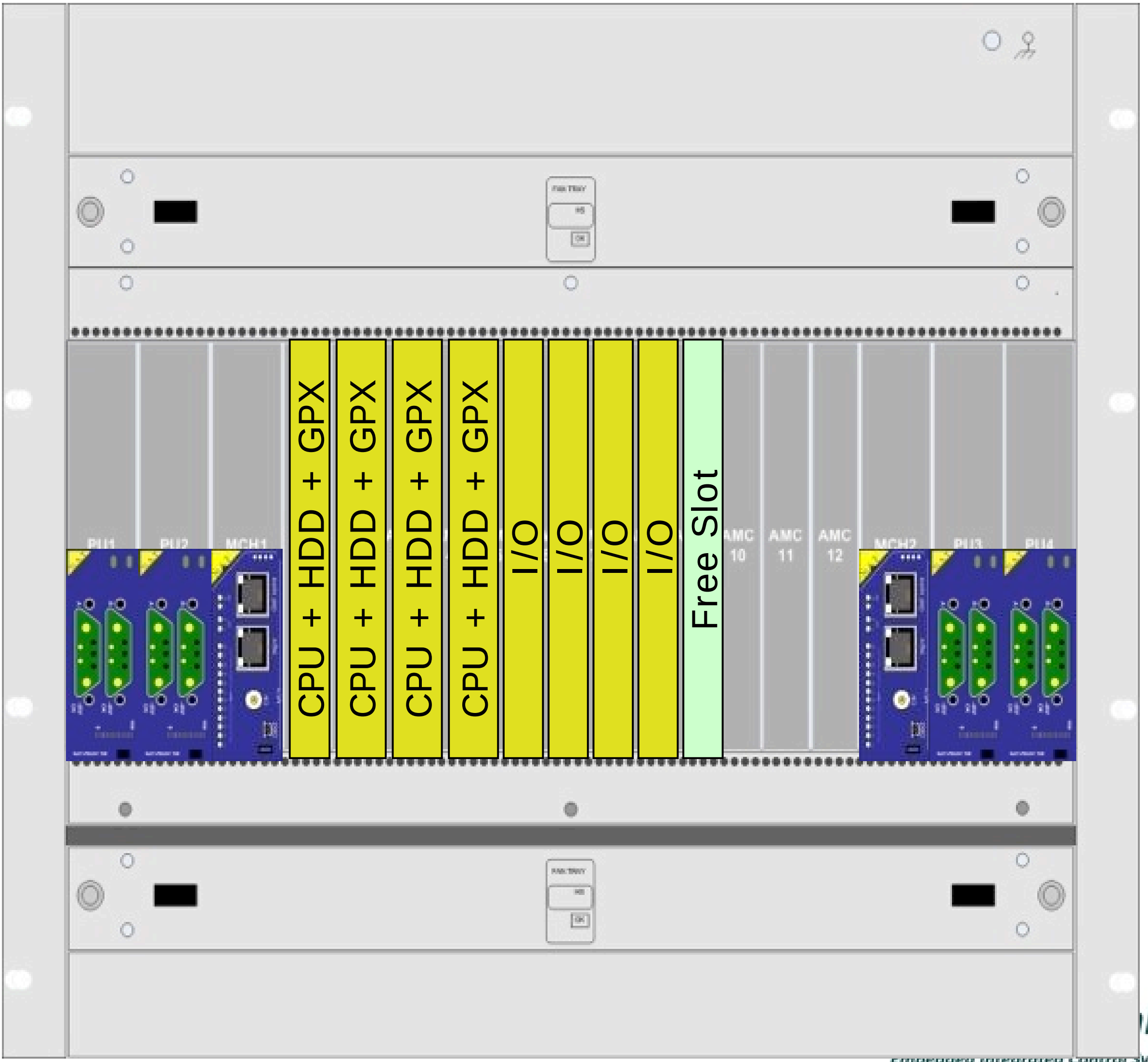
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

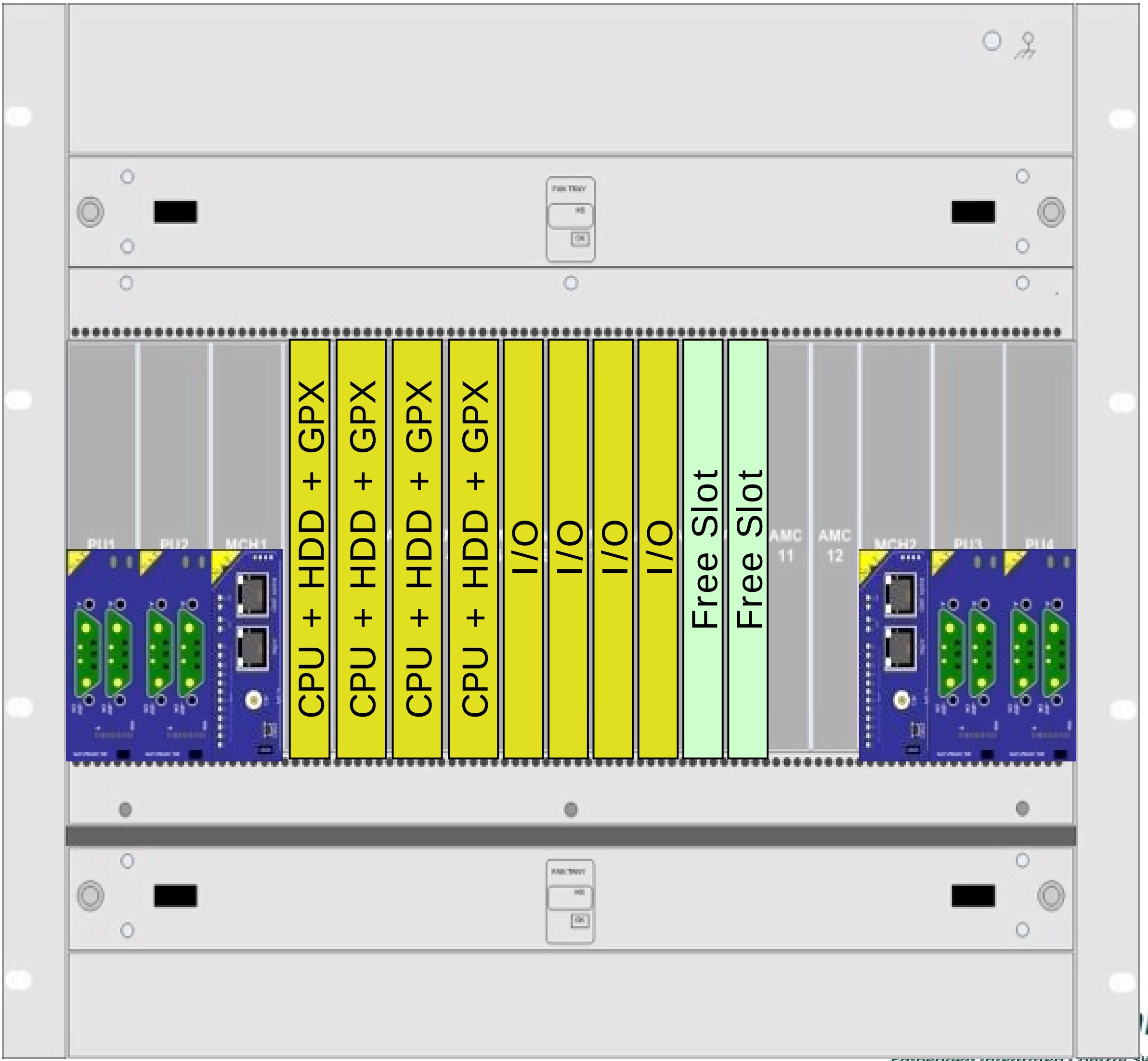
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

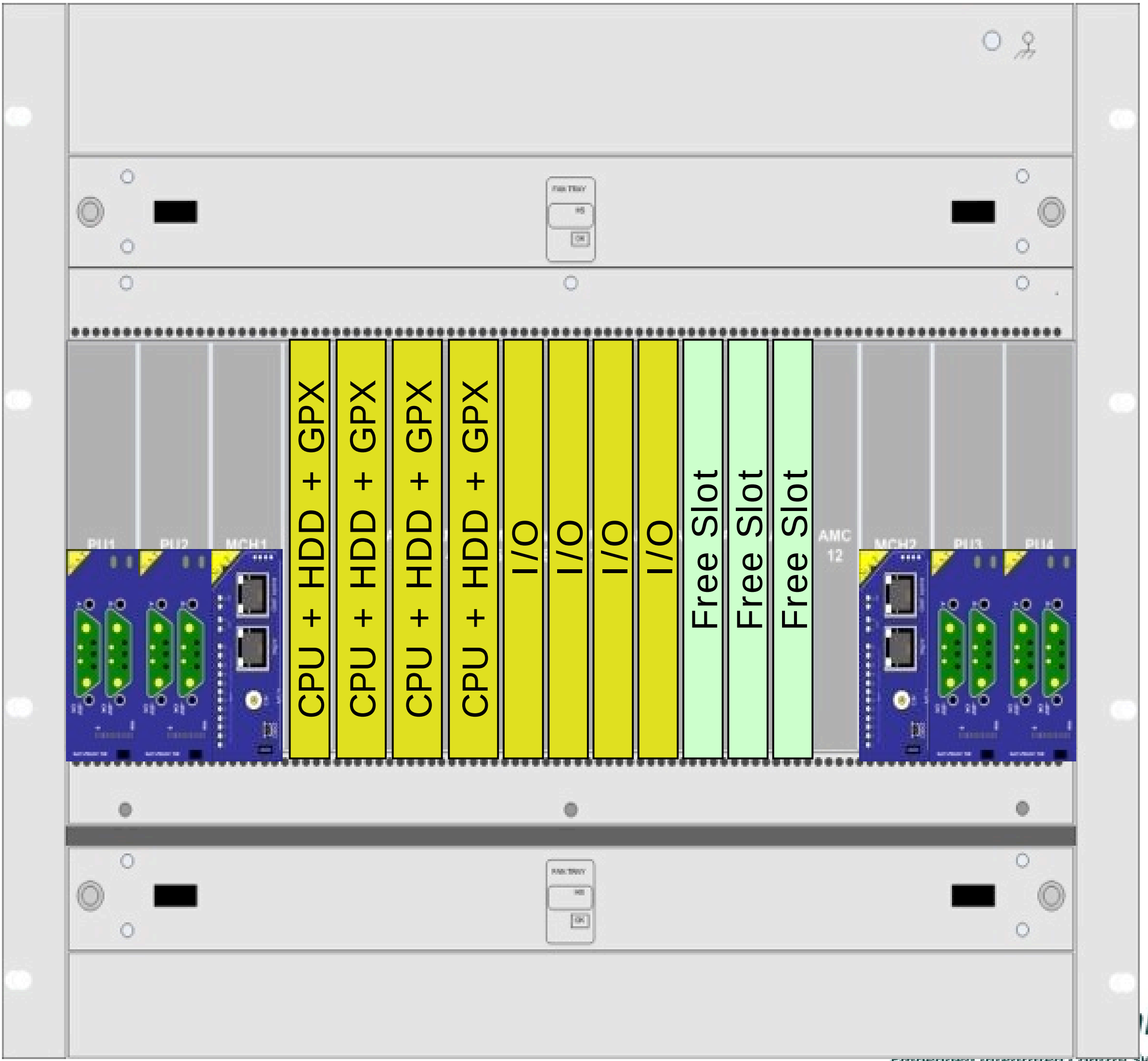
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

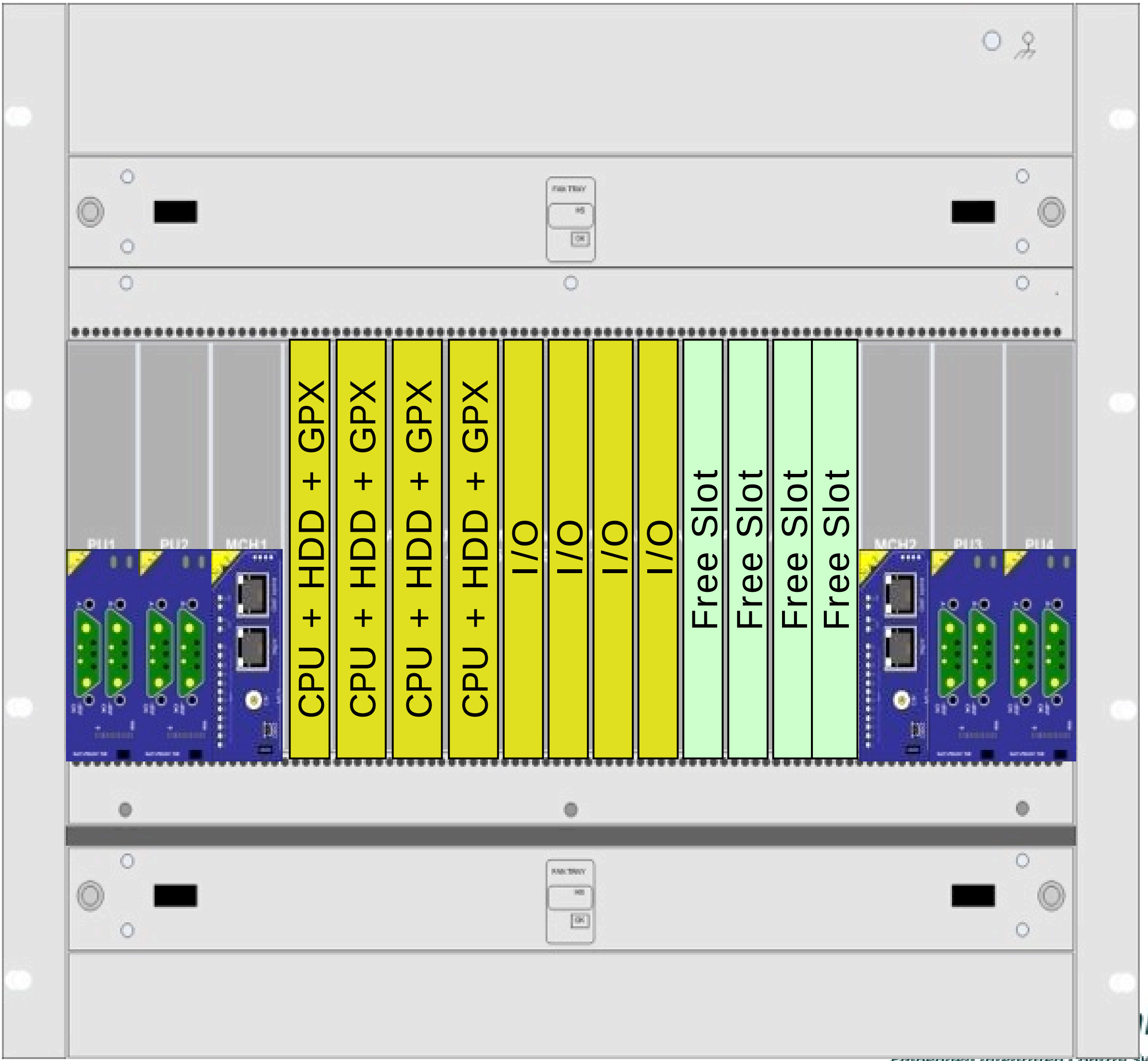
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

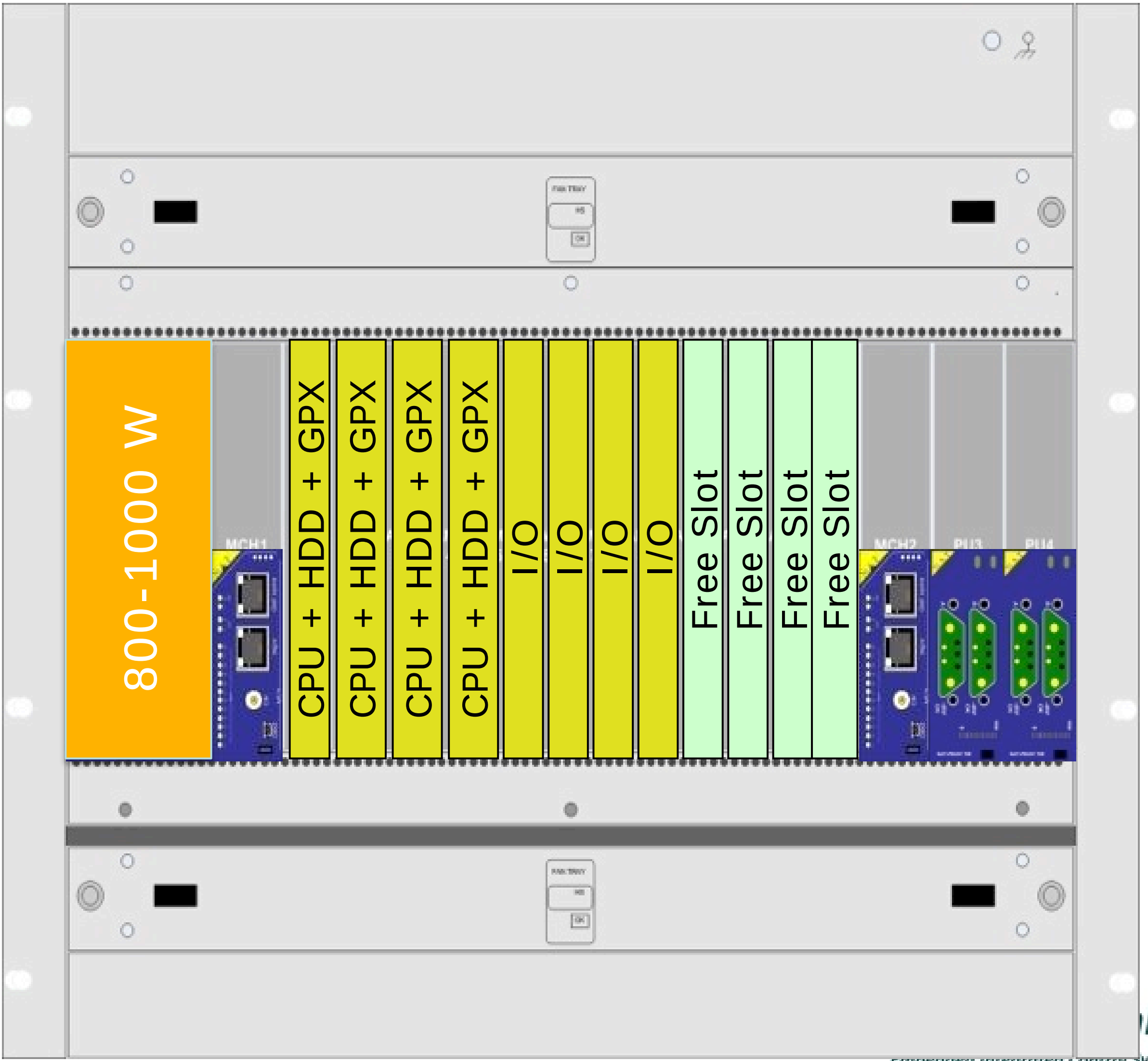
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,…

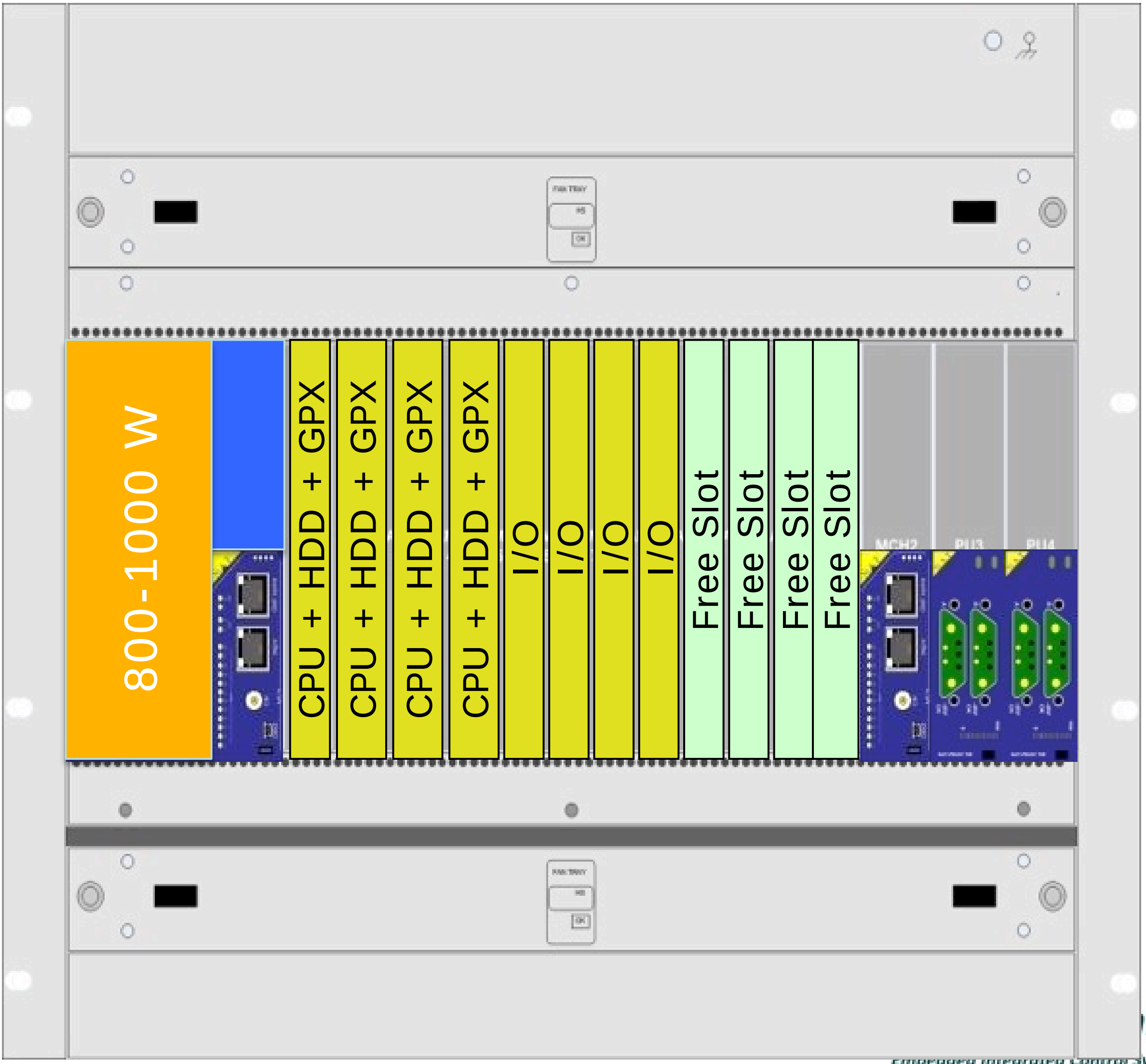
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

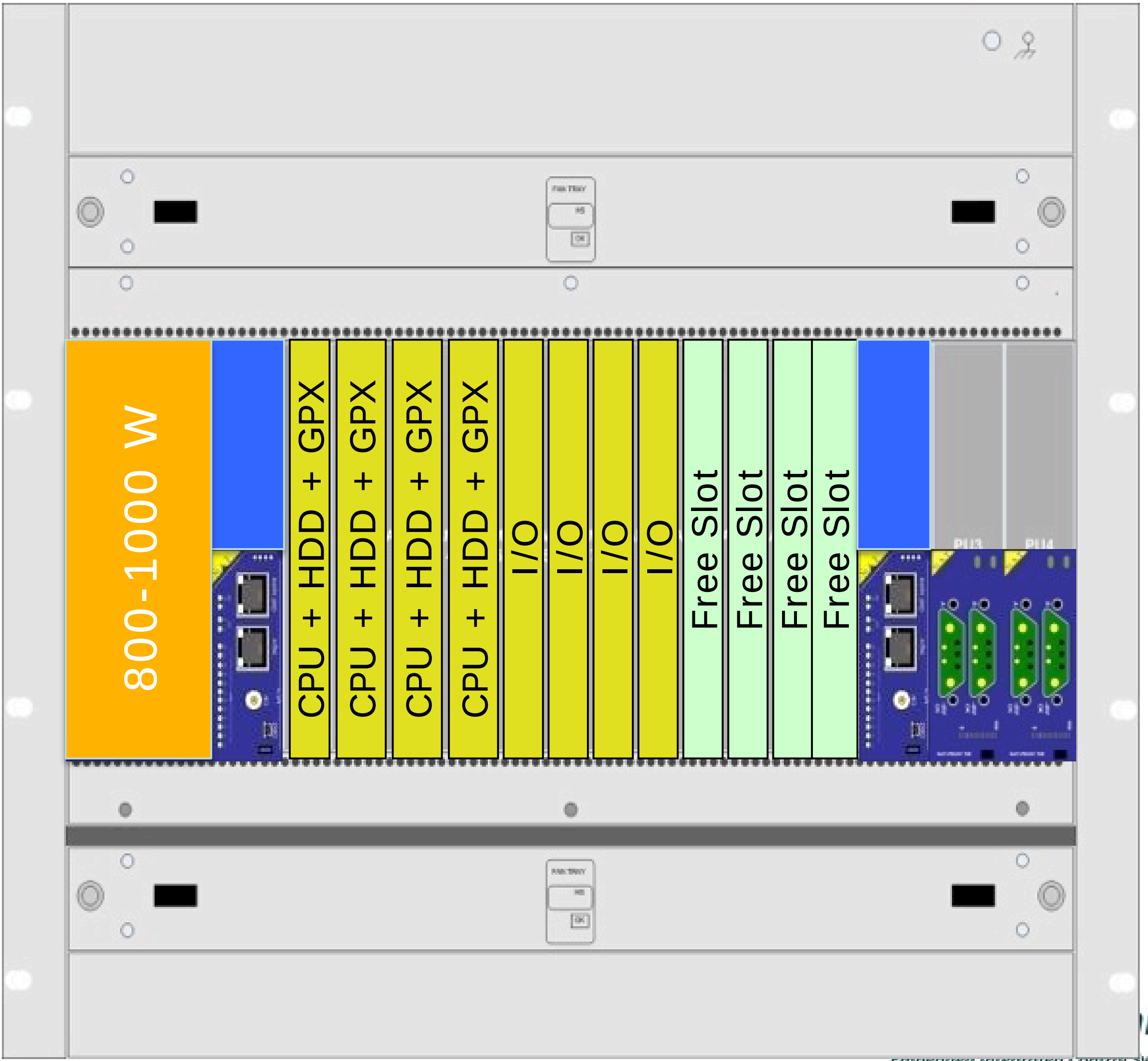
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

µRTM, multiple PrAMCs, multiple Transfers,...

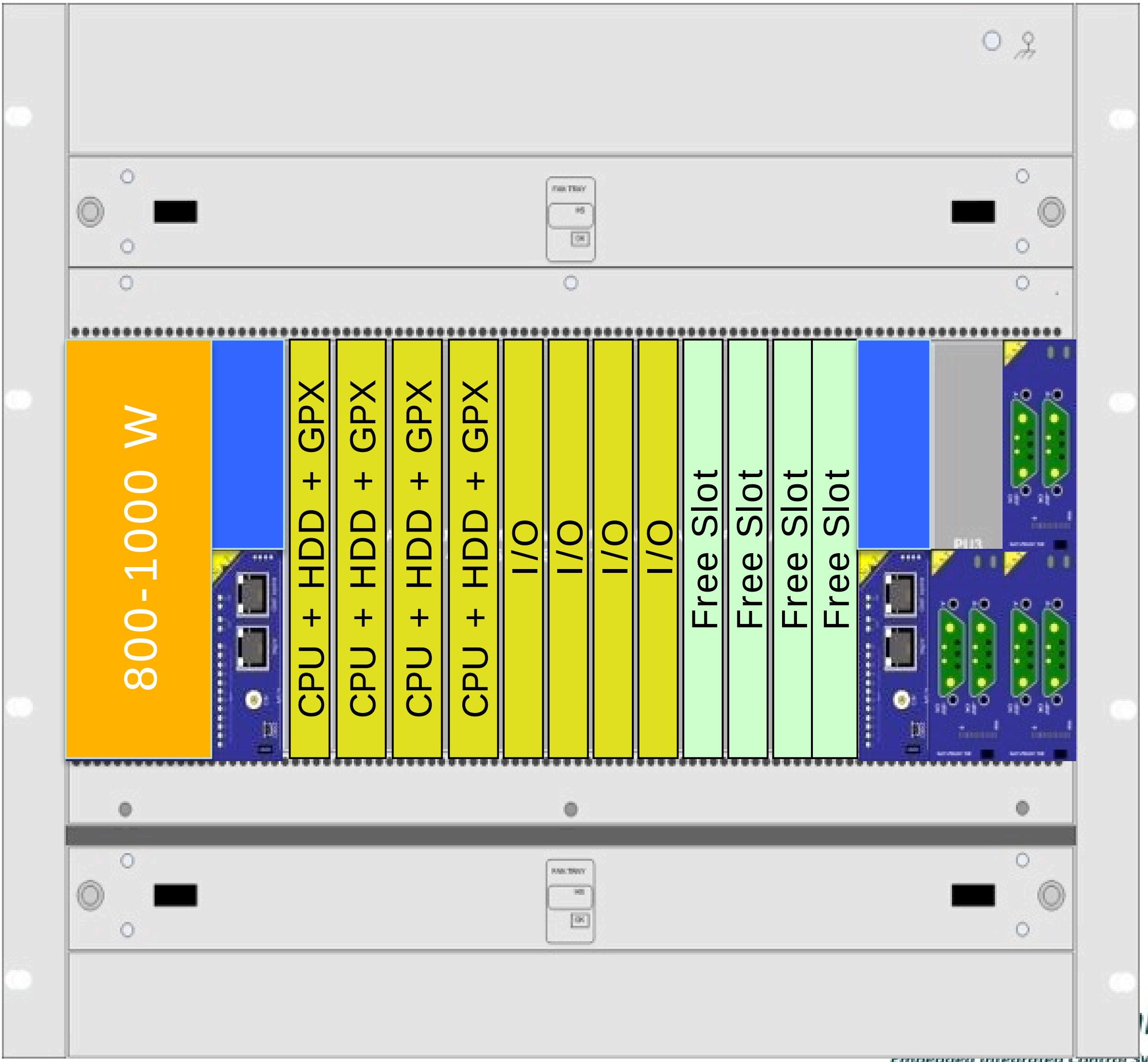
- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



Get more out of MTCA.4

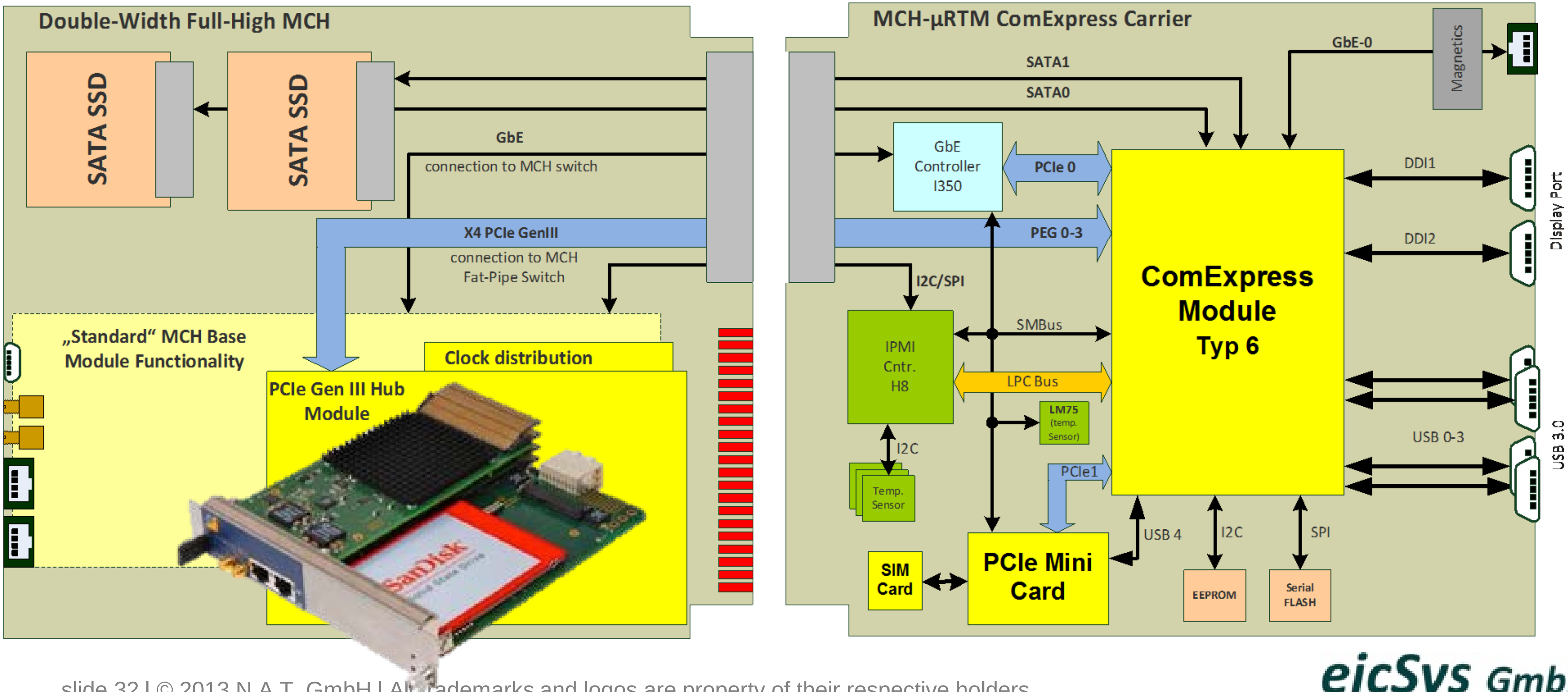
µRTM, multiple PrAMCs, multiple Transfers,...

- Which one is the Root Complex?
- Unused space above MCH
- Unused space behind MCH



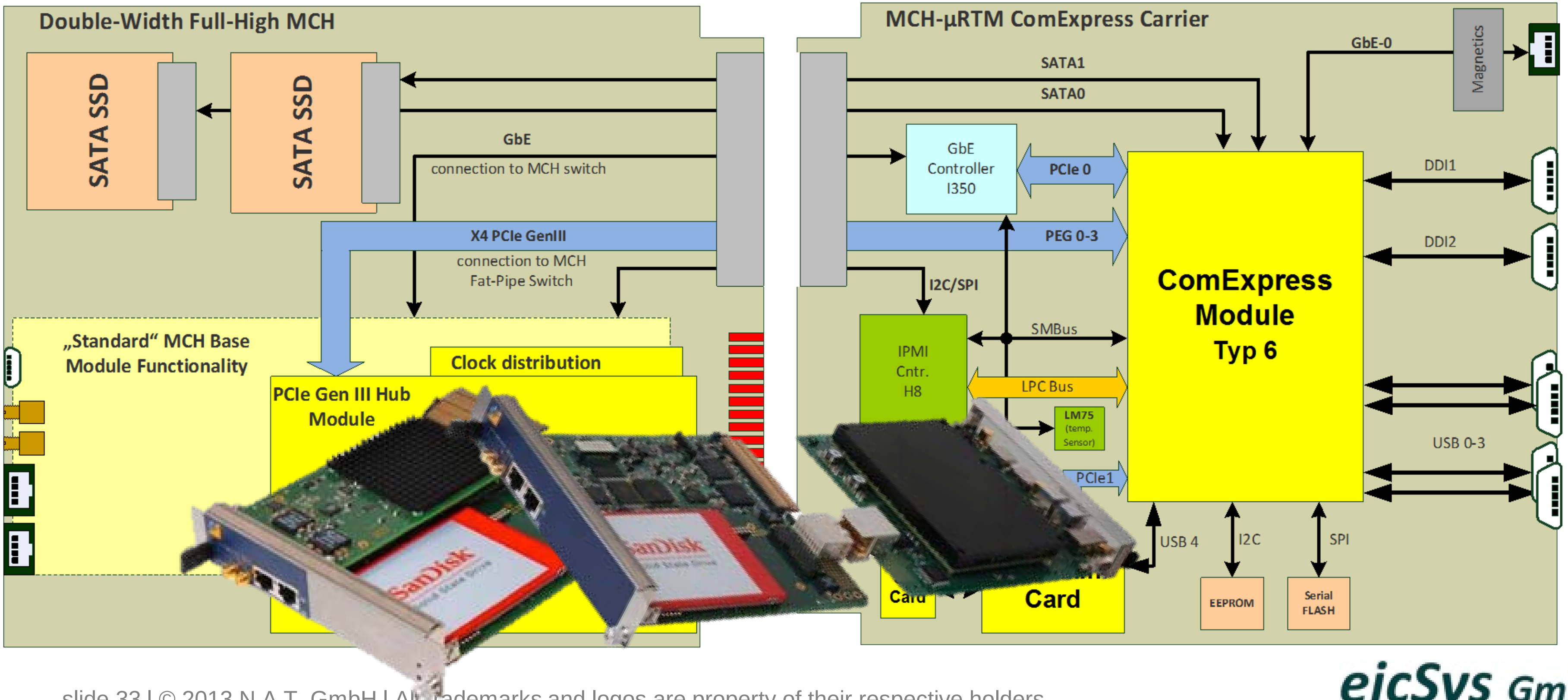
NAT-MCH, double-width
with µRTM ComExpress Carrier

- Saves 1 slot for CPU and minimum 1 slot for storage
- 7 Slot MTCA.4-Chassis: full feature MCH, Core-i7, 512 GB and 7 IO slots
- 12 Slot MTCA.4 chassis: 2 full feature MCH, 2 Core-i7, 1 TB, 12 IO slots



NAT-MCH, double-width with µRTM ComExpress Carrier

- Saves 1 slot for CPU and minimum 1 slot for storage
- 7 Slot MTCA.4-Chassis: full feature MCH, Core-i7, 512 GB and 7 IO slots
- 12 Slot MTCA.4 chassis: 2 full feature MCH, 2 Core-i7, 1 TB, 12 IO slots



ComExpress allow future upgrade

Options of interest

- ComExpress Type-6
- Available Processors
  - Core-i7 (Ivy-Bridge)
    - – 4 Core-i7 35W(SV)
      – 2 Core-i7 25W(LV)
      – 2 Core-i7 17W (ULV)
  - PowerPC (QorIQ-P1022)
  - VIA (Eden, Nano)
- Memory
  - Up to 8 Gbyte DDR3 incl ECC
- Display-Ports
- SATA
- USB3.0

MCH demanded by SLAC and Desy

NAT-MCH-PHYS

STAT
Primary/Secondary LED

RS232

GbE1

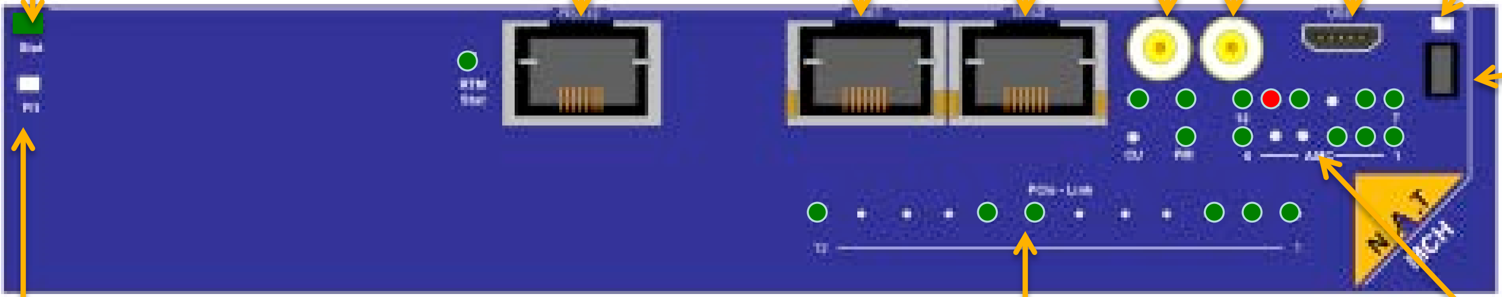
GbE2

CLK1&2

USB

Blue LED

Hot-Swap Hdl



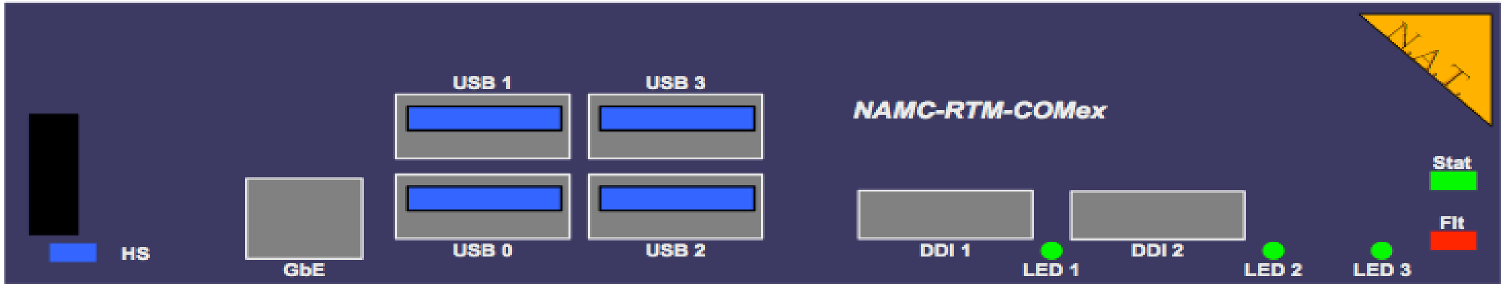
FLT
Red LED =Fault

PCIexpress Status and Speed LEDs:

off	no PCIe link active
on	PCIe-Gen3 link active
fast flashing	PCIe-Gen2 link active
slow flashing	PCIe-Gen1 link active

FRU Status LEDs:

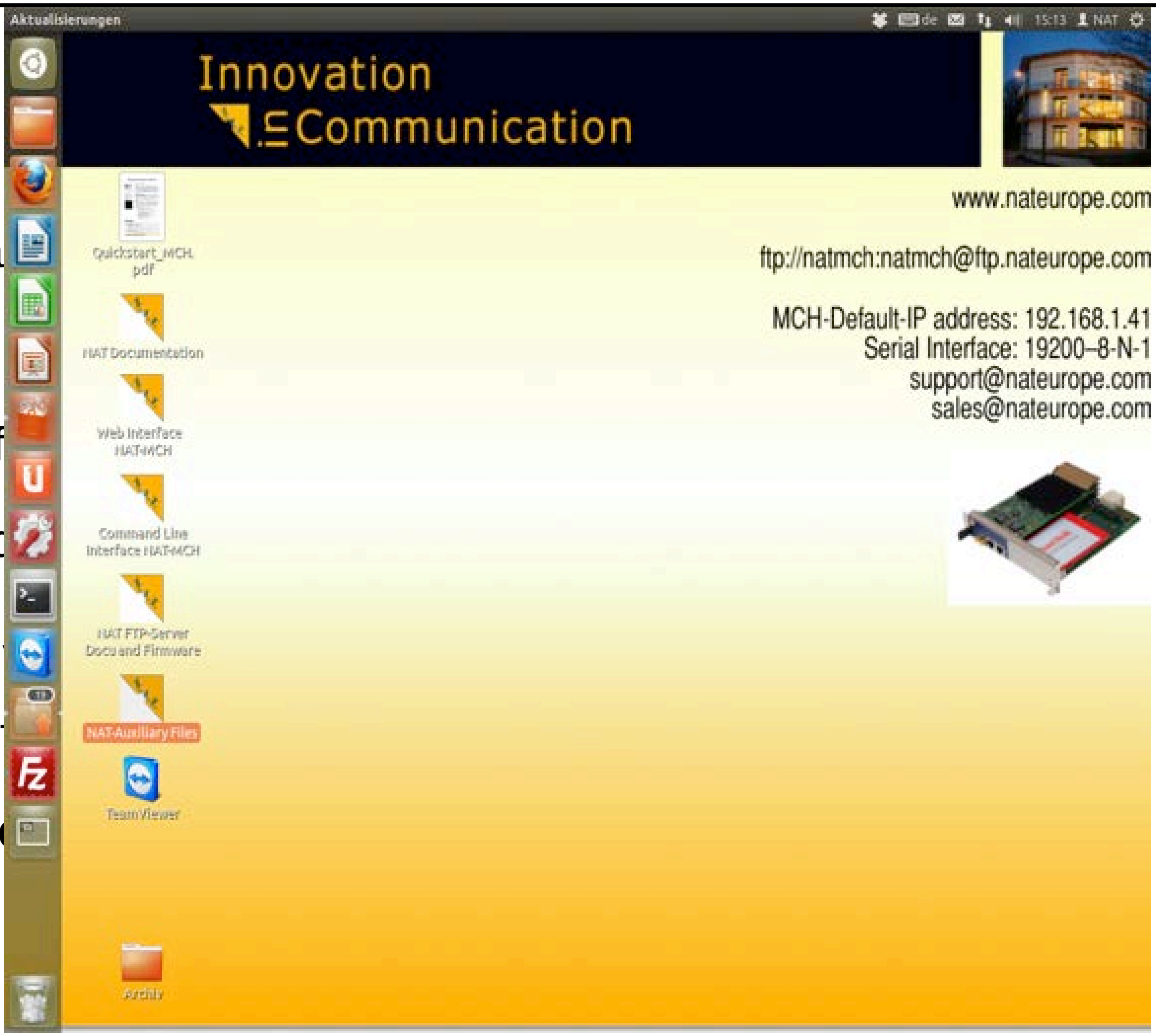
AMC 1-12
CU 1, 2
PM 1, 2



Out of the Box experience

Save Desy Support Resources

- Quick Sta
- NAT docu
- Webinterf
- Command
- NAT-FTP-
- Auxiliary
- NAT Rem



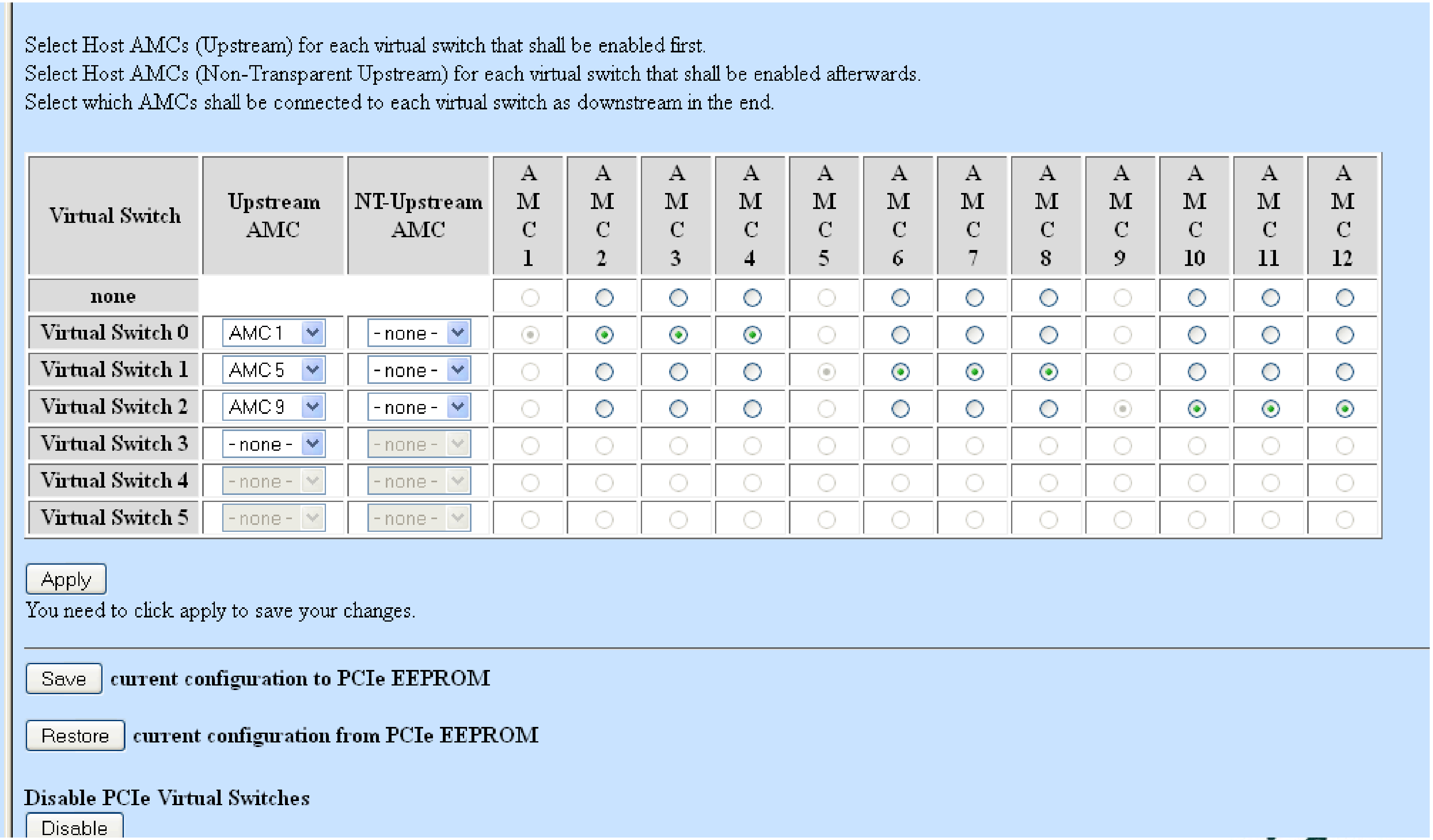
www.nateurope.com

ftp://natmch:natmch@ftp.nateurope.com

MCH-Default-IP address: 192.168.1.41
Serial Interface: 19200–8-N-1
support@nateurope.com
sales@nateurope.com

Up to 6 PCIexpress-Cluster

Configuring the Data Plane



Select Host AMCs (Upstream) for each virtual switch that shall be enabled first.
Select Host AMCs (Non-Transparent Upstream) for each virtual switch that shall be enabled afterwards.
Select which AMCs shall be connected to each virtual switch as downstream in the end.

Virtual Switch	Upstream AMC	NT-Upstream AMC	AMC 1	AMC 2	AMC 3	AMC 4	AMC 5	AMC 6	AMC 7	AMC 8	AMC 9	AMC 10	AMC 11	AMC 12
none			○	○	○	○	○	○	○	○	○	○	○	○
Virtual Switch 0	AMC 1	- none -	●	●	●	●	○	○	○	○	○	○	○	○
Virtual Switch 1	AMC 5	- none -	○	○	○	○	●	●	●	●	○	○	○	○
Virtual Switch 2	AMC 9	- none -	○	○	○	○	○	○	○	○	●	●	●	●
Virtual Switch 3	- none -	- none -	○	○	○	○	○	○	○	○	○	○	○	○
Virtual Switch 4	- none -	- none -	○	○	○	○	○	○	○	○	○	○	○	○
Virtual Switch 5	- none -	- none -	○	○	○	○	○	○	○	○	○	○	○	○

You need to click apply to save your changes.

current configuration to PCIe EEPROM

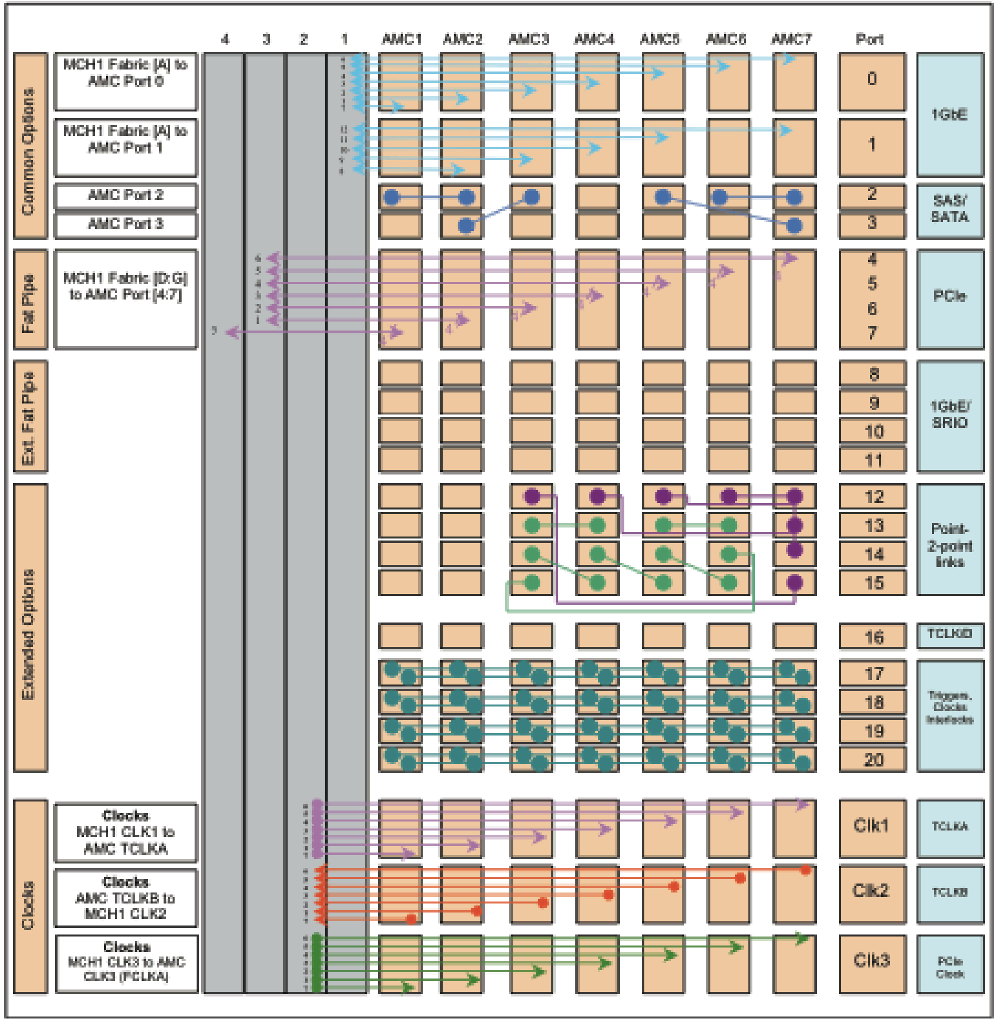
current configuration from PCIe EEPROM

Disable PCIe Virtual Switches

Backplane Topology MTCA.4

Multiple differential Connections & Transfers

Backplane Topology 7-slot Physics, 23005-466 [28.09.2011]



4 3 2 1 AMC1 AMC2 AMC3 AMC4 AMC5 AMC6 AMC7 Port

Common Options: MCH1 Fabric [A] to AMC Port 0 | MCH1 Fabric [A] to AMC Port 1 | AMC Port 2 | AMC Port 3

Fat Pipe: MCH1 Fabric [D:G] to AMC Port [4:7]

Ext. Fat Pipe

Extended Options

Clocks: Clocks MCH1 CLK1 to AMC TCLKA | Clocks AMC TCLKB to MCH1 CLK2 | Clocks MCH1 CLK3 to AMC CLK3 (PCLKA)

Port	
0	1GbE
1	
2	SAS/ SATA
3	
4	PCIe
5	
6	
7	
8	1GbE/ SRIO
9	
10	
11	
12	Point-2-point links
13	
14	
15	
16	TCLKD
17	Triggers, Clocks, Interlocks
18	
19	
20	
Clk1	TCLKA
Clk2	TCLKB
Clk3	PCIe Clock

Saving Costs & Increasing Data Throughput

Agenda

- Motivation
- Hardware Concept Extension
- Optimization of Data Transfer Performance
- Description of Life Demo
- Summary

DMA transfer

Optimization of Data Transfer Performance

- Direct memory access (DMA) – allows subsystem to get access to the memory independently of the CPU
- Most efficient way to access the memory
- In case of PCIe max performance can be achieved due do large frame sizes

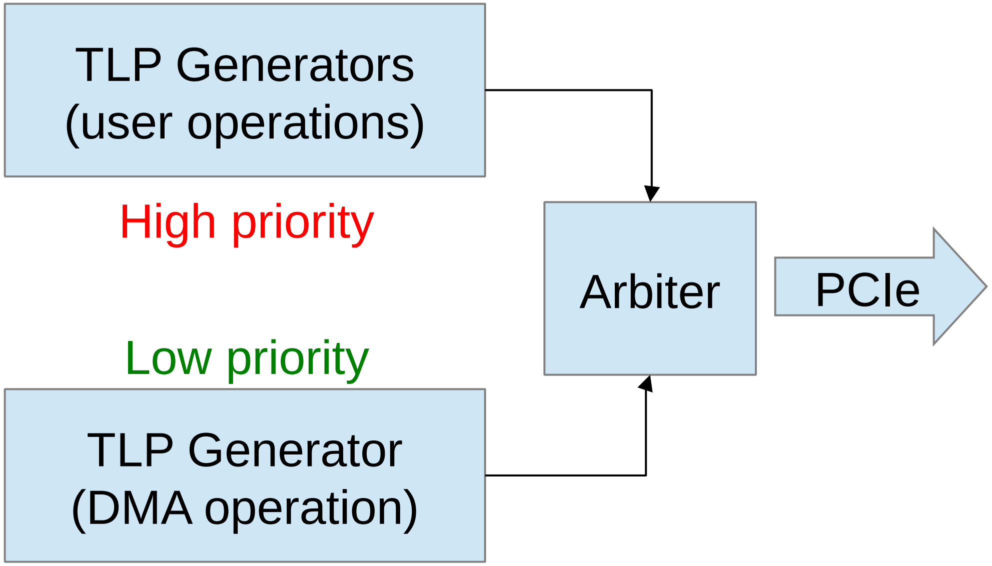
Implementation must be done on all levels of application (firmware, driver, software)

Flexibility ! Many ways to implement it, we just show one of possibilities.

Firmware layer

Optimization of Data Transfer Performance

- Virtex 5 based PCIe endpoint
- IRQ based notifications at the end of transfer
- On-board DDR2 memory access
- Single DMA transfer
- Scatter list implementation (1024 entries)



On firmware level, DMA transfer is relatively simple

Driver layer

Optimization of Data Transfer Performance

- Most of the work is done here, there is a lot of OS limitations:
  - Limitations of a single DMA transfer
  - Buffers on the kernel side or user side
  - It is even possible to reserve buffers in physical memory during system start-up
  - IRQ handling

- To get max performance some improvements to the firmware are needed
  - for example double buffering of scatter lists
  - circular buffer for scatter lists, etc.).
  - It is always balance between performance and required resources.

What has been chosen for live demo ?

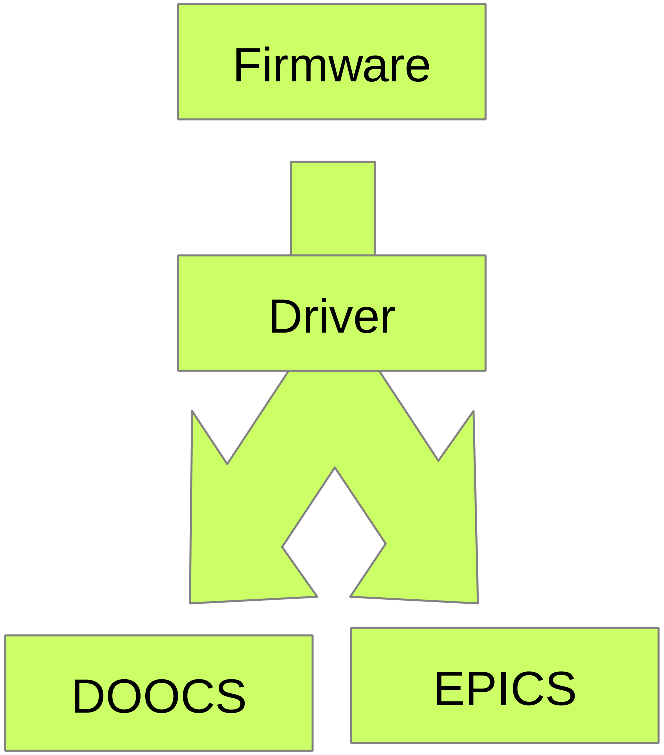
Software layer

Optimization of Data Transfer Performance

- Control over the firmware using IOCTL functions provided by the driver

- Buffers for data are allocated on application side

- Application (doocs or epics) receives SIGUSR1 when acquisition buffers are ready
- For performance evaluation, dedicated app has been implemented – it transfers large blocks of data from several boards.



Saving Costs & Increasing Data Throughput

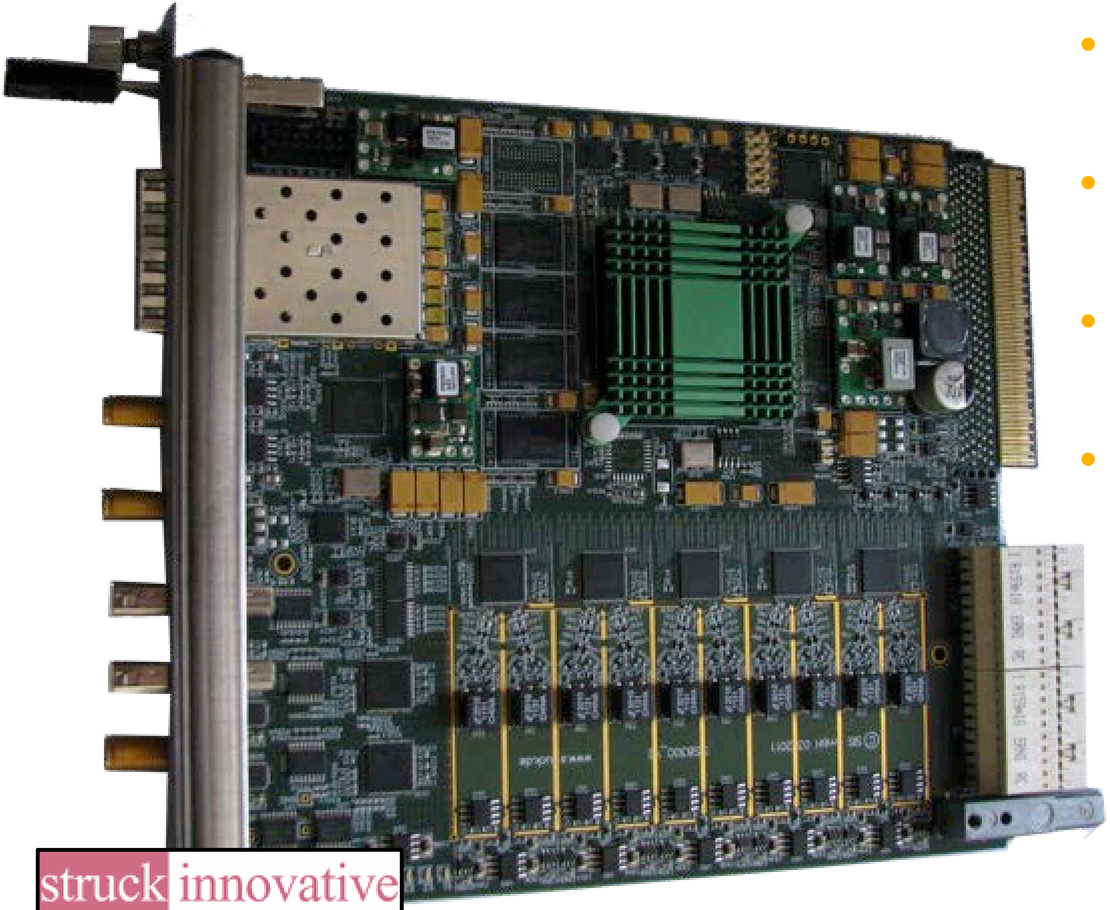
Agenda

- Motivation
- Hardware Concept Extension
- Optimization of Data Transfer Performance
- Description of Life Demo
- Summary

eicSys GmbH
Embedded Integrated Control Systems

Description of Live Demo

SIS8300 – digitizer board



struck innovative systeme

- MTCA.4 digitizer board
- 10 analog channels
- Virtex 5 FPGA
- PCIe x4 GEN1 capability

Boxards are running
eicSys **uni_daq_firmware**

eicSys GmbH
Embedded Integrated Control Systems

Description of Live Demo

MTCA.4 Chassis, 4 AMCs, MCH + COMex

- All software tools showed in the presentation
- EPICS demo (data readout and visualization)
- DOOCS demo (data readout and visualization)
- Demonstration of basic uni_daq_firmware framework functions

Please visit us at the exhibition

MTCA.4 Debugging
Inventory

- show_fru

```
 FRU  Device  State  Name
=========================================
 0   MCH      M4    NMCH-CM
 3   mcmc1    M4    NAT-MCH-MCMC
 5   AMC1     M4    SIS8300
 6   AMC2     M4    SIS8300
 7   AMC3     M4    SIS8300
 8   AMC4     M4    TAMC900-10
40   CU1      M4    Cooling  Unit
50   PM1      M4    PDM
60   Clk1     M4    MCH-Clock
61   Hub1     M4    MCH-PCIe
64   RTM1     M4    MCH-RTM-ComEx
```

MTCA.4 Debugging

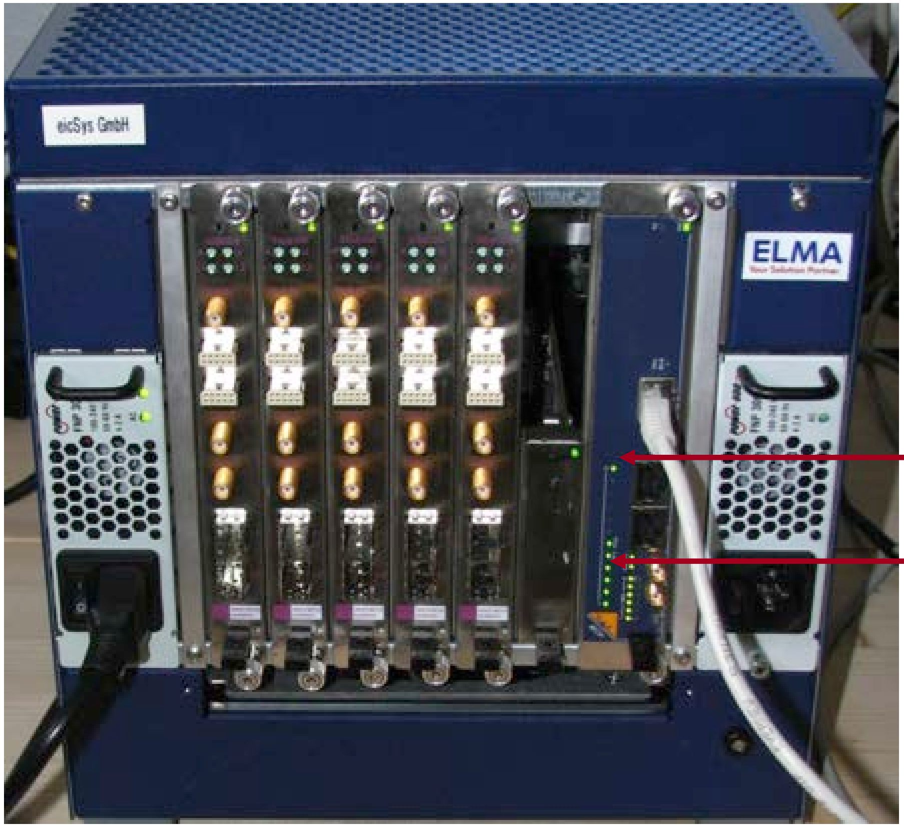
Power Budget

- show_pm

```
PM1: - online, primary(fru 50)    : budget 50.0 A (alloc 25.3 A avail 24.7 A)
|---------------------------------------------------------|
|chan  FRU FruId  primPM  secPM  PS1  POn  ENA   MP  PP  Amps |
|---------------------------------------------------------|
  1   MCH1     3   1       -      y    y    y    y   y   3.8
  2   MCH2     4   -       -
  3    CU1    40   1       -      y    -    y    y   y   2.5
  4    CU2    41   -       -
  5   AMC1     5   1       -      y    -    y    y   y   5.0
  6   AMC2     6   1       -      y    -    y    y   y   5.0
  7   AMC3     7   1       -      y    -    y    y   y   5.0
  8   AMC4     8   1       -      y    -    y    y   y   4.0
  9   AMC5     9   1       -      -    -    -    -   -
 10   AMC6    10   1       -      -    -    -    -   -
 11   AMC7    11   -       -
 12   AMC8    12   -       -
 13   AMC9    13   -       -
 14  AMC10    14   -       -
 15  AMC11    15   -       -
 16  AMC12    16   -       -
|---------------------------------------------------------|
```

Description of Live Demo

PCIe Gen1 and Gen3 configurations



PCIEx2/PCIEx4 **Gen-3**
MCH-RTM-COMex-i7

PCIEx4 **Gen-1**
AMC1,2,3,4

Acquisition on all boards
is synchronized by one trigger

MTCA.4 Debugging

E-Keying

- show_ekey

```
EKeying information - activated Links:
--------------------------------------
 AMC FRU State Channel Type Port
======================================================
AMC1   5   M4    0     PCIe   4 <-> MCH1 Fabric D downstream Gen 1, no SSC
                               5 <-> MCH1 Fabric E downstream Gen 1, no SSC
                               6 <-> MCH1 Fabric F downstream Gen 1, no SSC
                               7 <-> MCH1 Fabric G downstream Gen 1, no SSC

AMC2   6   M4    0     PCIe   4 <-> MCH1 Fabric D downstream Gen 1, no SSC
                               5 <-> MCH1 Fabric E downstream Gen 1, no SSC
                               6 <-> MCH1 Fabric F downstream Gen 1, no SSC
                               7 <-> MCH1 Fabric G downstream Gen 1, no SSC

AMC3   7   M4    0     PCIe   4 <-> MCH1 Fabric D downstream Gen 1, no SSC
                               5 <-> MCH1 Fabric E downstream Gen 1, no SSC
                               6 <-> MCH1 Fabric F downstream Gen 1, no SSC
                               7 <-> MCH1 Fabric G downstream Gen 1, no SSC


...........
```

MTCA.4 Debugging

Result of PCIexpress Training

- show_link_state

```
AMC  1 Port  4 is PCIe - x4 - 2,5 GT/s
AMC  1 Port  5 is PCIe - x4 - 2,5 GT/s
AMC  1 Port  6 is PCIe - x4 - 2,5 GT/s
AMC  1 Port  7 is PCIe - x4 - 2,5 GT/s
AMC  2 Port  4 is PCIe - x4 - 2,5 GT/s
AMC  2 Port  5 is PCIe - x4 - 2,5 GT/s
AMC  2 Port  6 is PCIe - x4 - 2,5 GT/s
AMC  2 Port  7 is PCIe - x4 - 2,5 GT/s
AMC  3 Port  4 is PCIe - x4 - 2,5 GT/s
AMC  3 Port  5 is PCIe - x4 - 2,5 GT/s
AMC  3 Port  6 is PCIe - x4 - 2,5 GT/s
AMC  3 Port  7 is PCIe - x4 - 2,5 GT/s
AMC  4 Port  4 is PCIe - x4 - 2,5 GT/s
AMC  4 Port  5 is PCIe - x4 - 2,5 GT/s
AMC  4 Port  6 is PCIe - x4 - 2,5 GT/s
AMC  4 Port  7 is PCIe - x4 - 2,5 GT/s
local RTM link status:
   Ethernet - 1000Base-BX
   PCIe - x4 – 8 GT/s
```

Description of Live Demo

Benchmark results

CPU PCIe Conf	1 [MB/s]	2 [MB/s]	3 [MB/s]	4 [MB/s]	bandwidth [MB/s]	max [MB/s]	link usage [%]
PCIe x4 GEN1	829.75	432.09 431.19	287.44 287.16 287.45	215.60 215.55 215.64 215.62	862.41	914.28 1000	94.32
PCIe x2 GEN3	828.82	813.24 812.68	556.20 553.08 556.24	417.46 415.48 417.46 417.47	1667.87	1801.14 1970	92.6
PCIe x4 GEN3	?	?	?	?		3602.28 3940	
PCIe x16 GEN3	?	?	?	?			

For more details please visit our live demo on the exhibition

eicSys GmbH
Embedded Integrated Control Systems

Saving Costs & Increasing Data Throughput

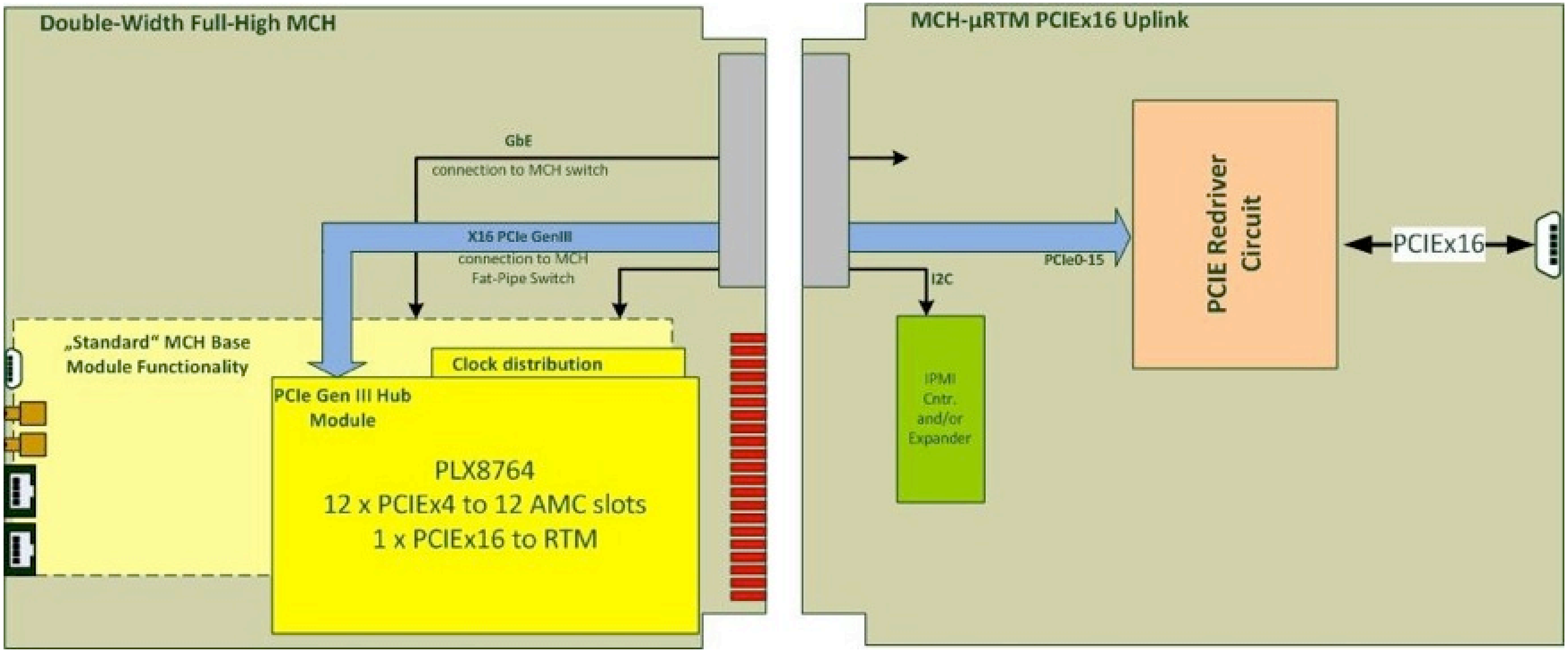
Agenda

- Motivation
- Hardware Concept Extension
- Optimization of Data Transfer Performance
- Description of Life Demo
- Summary

Bottleneck Analysis

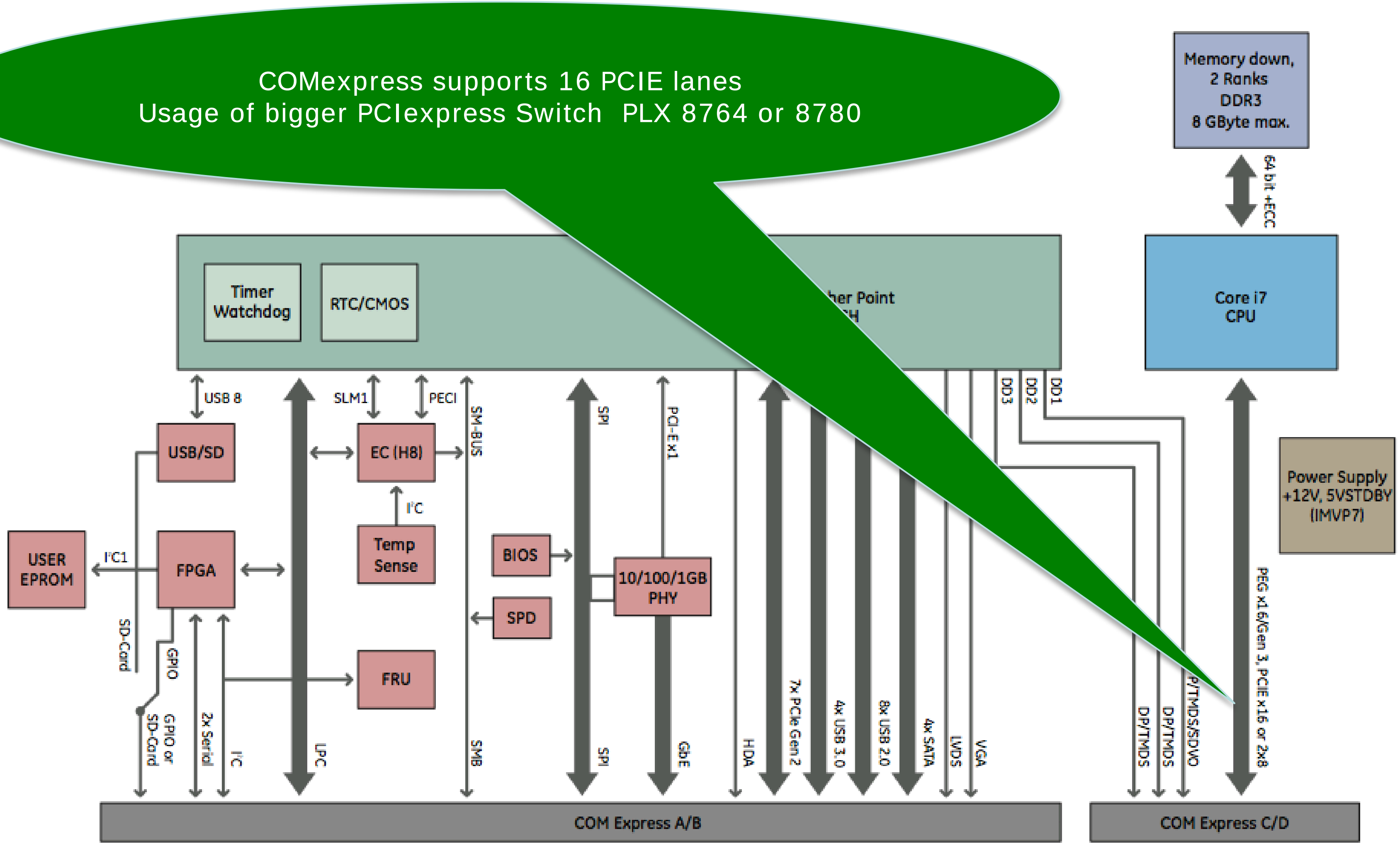
- Processing Power
  - Increase number of CPU cores => power consumption
  - Additional AMC-CPU boards => PCIexpress Clustering, NT-Mode
  - Alternative: Freescale QorIQ-P4080 (8 cores), P5040 (64-Bit)
- Data Path
  - PCIexpress Gen1 (4 *2.5 Gbaud= 10 Gbaud, 10b/8b coding, 20% overhead)
  - PCIexpress Gen2 (4 * 5 Gbaud = 20 Gbaud, 10b/8b coding)
  - PCIexpress Gen 3 (4 * 8 Gbaud = 32 Gbaud, 128b/130b coding, 1,54% overhead)
    - PCIe 3.0's 8 GT/s bit rate effectively delivers 985 MB/s per lane, double speed of PCEe 2.0
  - Increase number of lanes to Central CPU => 16 lanes

Solution 1:
External Processing Power



Solution 2: NAT-MCH-RTM-COMex-i7
with16 PCIexpress lanes

COMexpress supports 16 PCIE lanes
Usage of bigger PCIexpress Switch PLX 8764 or 8780



eicSys GmbH
Embedded Integrated Control Systems

Thank you very much!

Questions?

Jalmuzna Wojciech
wojciech.jalmuzna@eicsys.eu

Vollrath Dirksen
vollrath@nateurope.com
mtca-helpdesk@desy.de
support@mtca.eu

www.eicsys.eu
www.nateurope.com

MTCA.4 Training:

www.mtca.eu

2014: Training Level-2 and -3

eicSys GmbH
Embedded Integrated Control Systems