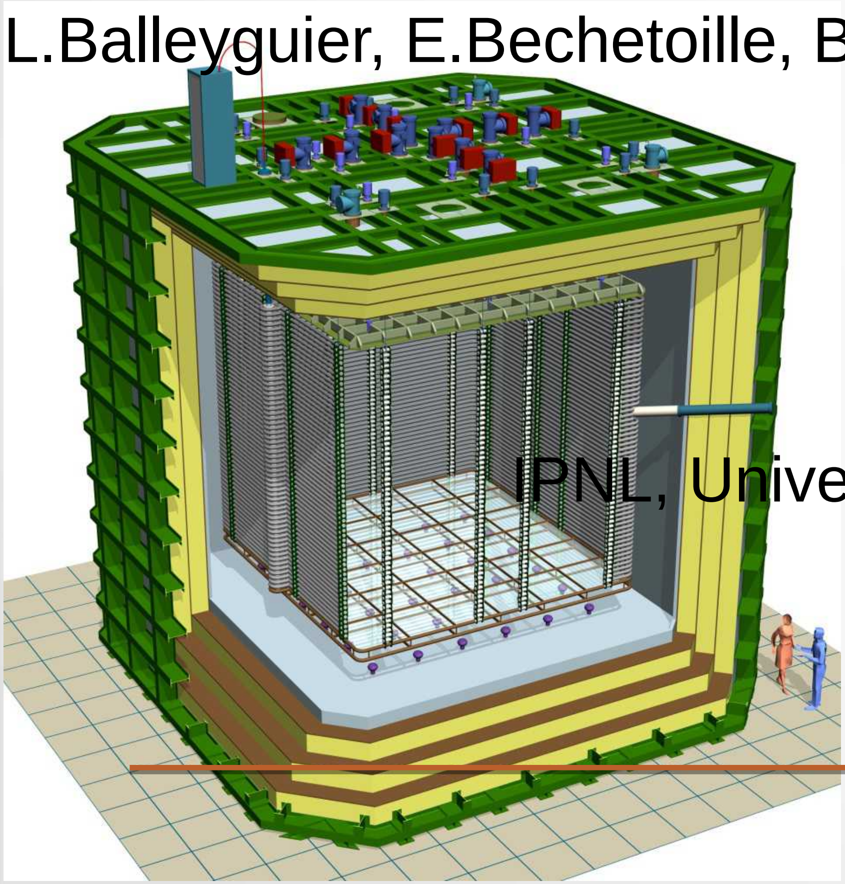
WA105 global DAQ architecture

D.Autiero, D.Caiulo, S.Galymov, J.Marteau, E.Pennacchio

L.Balleyguier, E.Bechetoille, B.Carlus, C.Girerd, H.Mathez, D.Pugnère



IPNL, Université de Lyon, CNRS-IN2P3, UMR 5822

SBN-DUNE Meeting

November 20, 2015

Outline

- **WA105 DAQ general features**
- **µTCA standard and implementation : crates & AMC/ADC boards**
- **FPGA back-end processing boards & OpenCL**
- **White Rabbit time distribution standard**
- **Local storage design**

WA105 @CERN : R/O features

- **F/E-inside** : ASIC in the cold → feedthrough (power supply / signal)

- **F/E-outside** : **µTCA** (charge/PM) + **White Rabbit** (time/CLK/trigger)
  - ✓ µTCA crate on the cryostat directly coupled to the feedthroughs
  - ✓ 12 slots for charge R/O (10/12 occupied : ~ 16 % contingency)
  - ✓ **Single MCH + WR slave board**
    - ✓ Time/CLK/trigger distribution to the backplane
    - ✓ 1 x 10 Gbe uplink

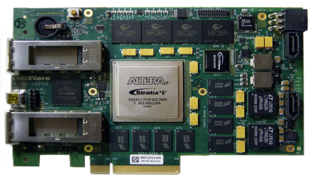
- **B/E** : **FPGA** processing boards inside standard PC :
  - ✓ 1 PC for 6 µTCA crates (6/8 occupied : ~ 25 % contingency)
  - ✓ **OpenCL** framework for developers

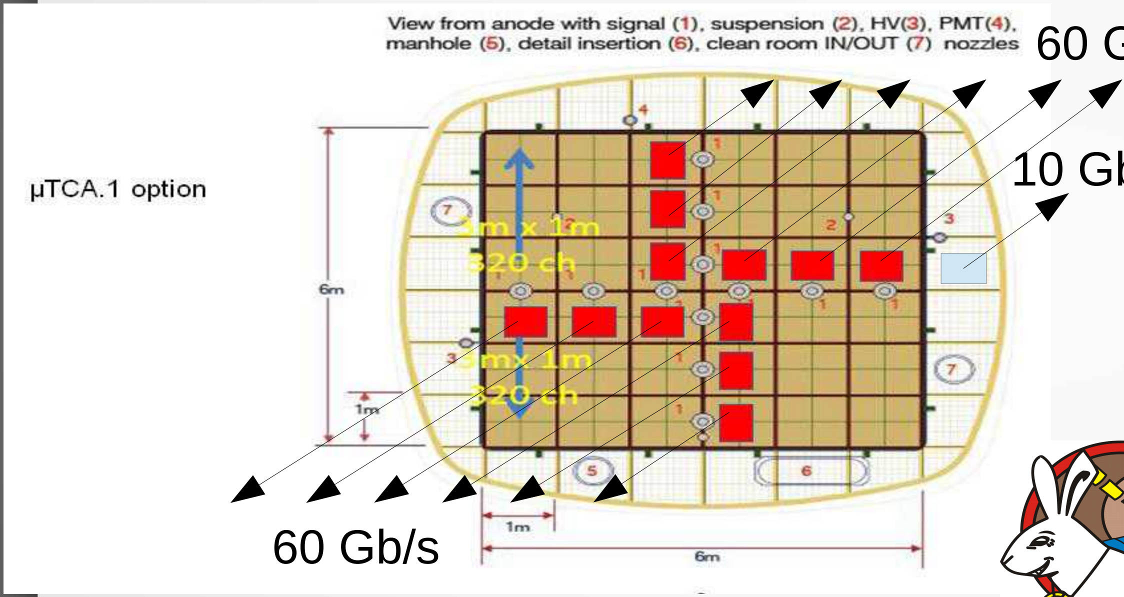
- **Local buffer and online processing** :
  - ✓ **DELL** servers farm (15 OSS, 2 MDS, 1 CS) → 1 PB (1 – 10 days)
  - ✓ Online processing (384 cores) → shared with LxBatch system ?
  - ✓ Connexion to CERN : 20 Gbps, EOS then CASTOR, 2.5 PB final

µTCA-DAQ architecture

2 x 40 Gb/s

E.B.1



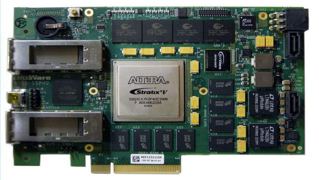

View from anode with signal (1), suspension (2), HV(3), PMT(4),
manhole (5), detail insertion (6), clean room IN/OUT (7) nozzles
3m x 1m
320 ch
3m x 1m
320 ch
6m
1m
1m
6m
1
2
3
4
5
6
7
7

µTCA.1 option

60 Gb/s

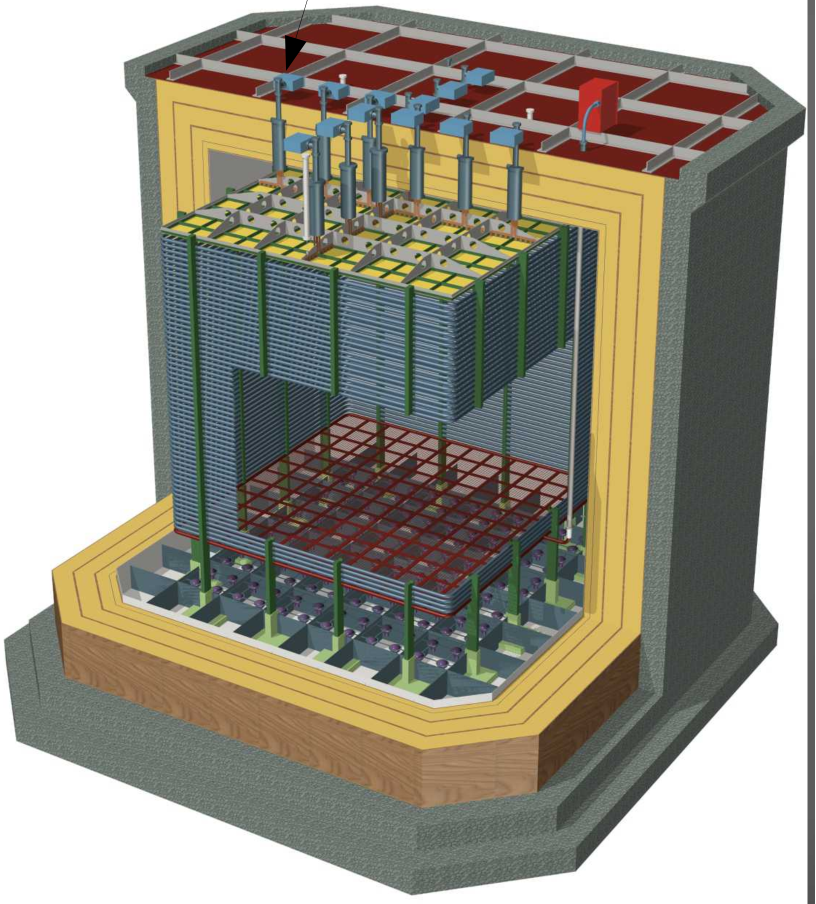
10 Gb/s

60 Gb/s

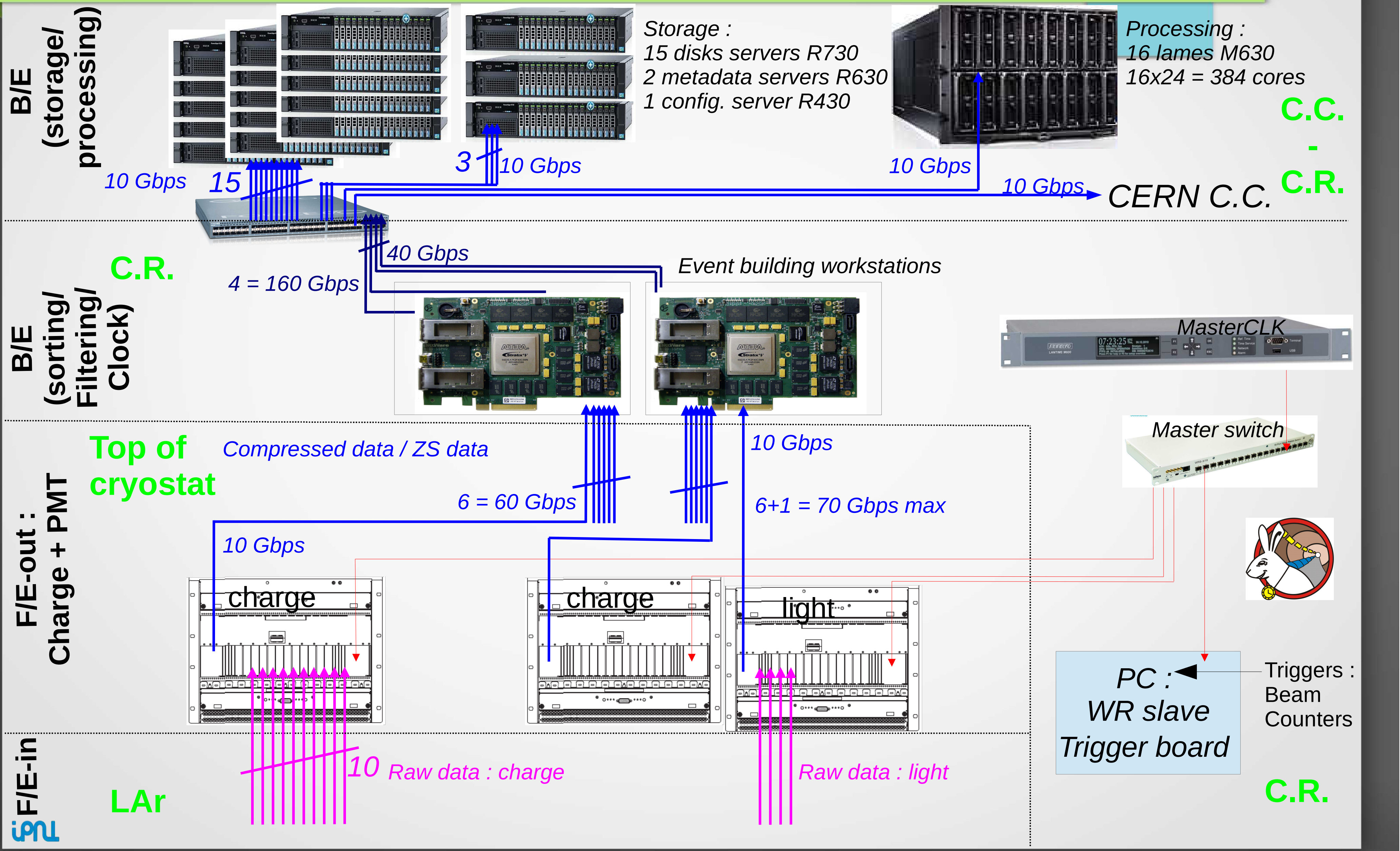


E.B.2

2 x 40 Gb/s



WA105 data network



B/E (storage/ processing)

Storage :
15 disks servers R730
2 metadata servers R630
1 config. server R430

Processing :
16 lames M630
16x24 = 384 cores

C.C. - C.R.

10 Gbps 15

3 10 Gbps

10 Gbps

10 Gbps → CERN C.C.

B/E (sorting/ Filtering/ Clock)

C.R.

40 Gbps

4 = 160 Gbps

Event building workstations

MasterCLK

Master switch

F/E-out : Charge + PMT

Top of cryostat

Compressed data / ZS data

10 Gbps

6 = 60 Gbps

6+1 = 70 Gbps max

10 Gbps

charge

charge

light

PC :
WR slave
Trigger board

Triggers :
Beam
Counters

C.R.

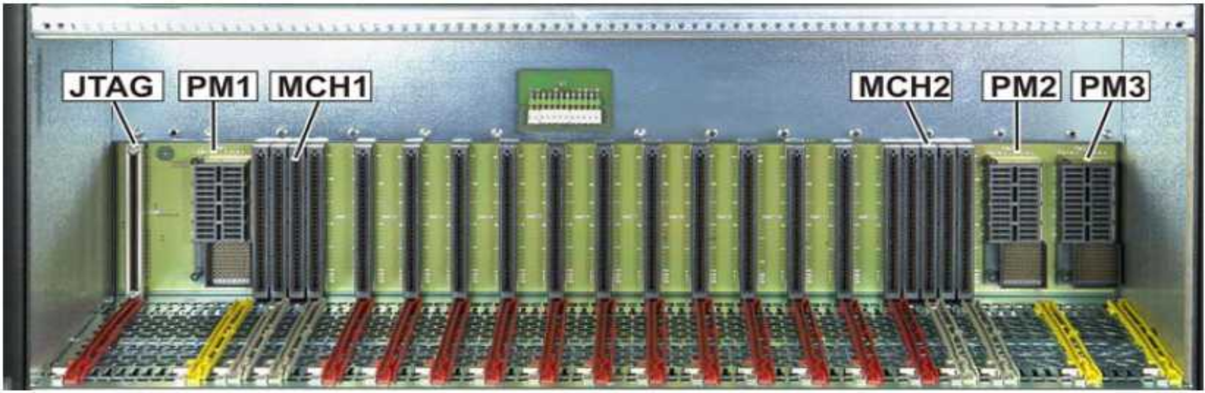
F/E-in

LAr

10 Raw data : charge

Raw data : light

µTCA crate requirements



Crate backplane

640 ch. per crate = 10x64 ch. AMC cards
– 1 crate per chimney
– 12 crates in total = 7680 ch.
– 16 bits per sample, 400ns sampling rate
– Max. drift : 4000µs => 10k samples per ch.

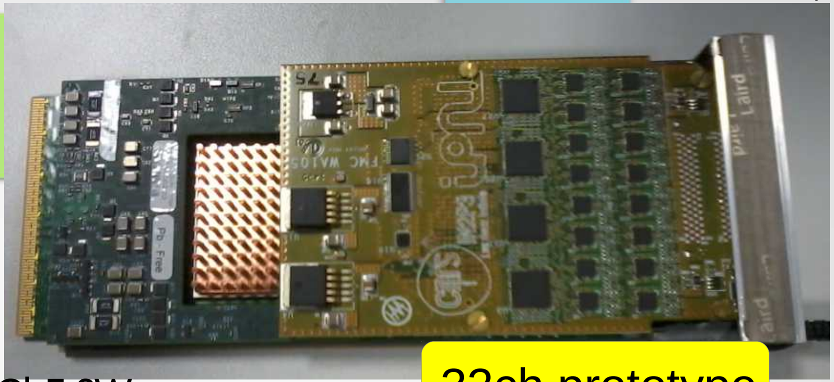
Data rates :
– Max. rate per crate :
10 Gbits / (16b.x640ch.x10k) ~ 100 Hz (nominal beam)
– Zero-suppression : 6µs shaping =>
~15 samples for a single hit => max. red. factor ~600
– Huffman compression : factor 10

Schroff reference : 11850-015, 8 U µTCA Shelf,

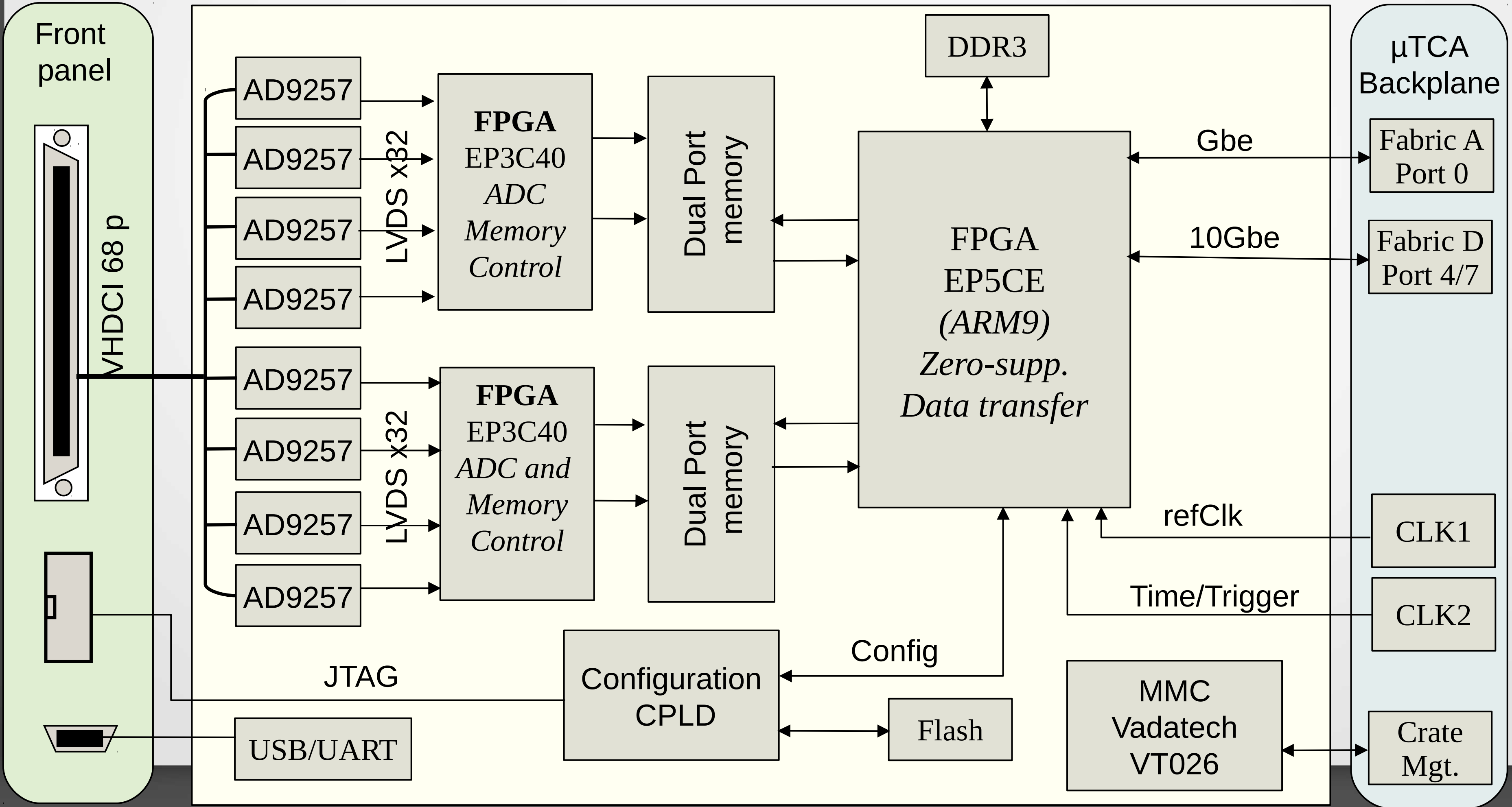
- 12+2+3+1 slot for AMC double mid-size modules
- 19” rack mountable
- 12 AMC Double mid-size slots
- 2 redundant MicroTCA Carrier Hub (MCH) slots
- 2 Power Module (PM) slots (6 HP Double), right side
- 1 Power Module (PM) slots (12 (9) HP dble), left side
- 1 slot for a JTAG module (dble compact)
- 5 splitting kits to install single module in a double slot

AMC digitization card

- µTCA standard (double width , full height)
- 64 input channels (2V / 14 bits / 2.5 Msps to 20 Msps)
- Control through MCH 1GbE backplane link port 0/1
- Data transmission through 10GbE backplane and MCH 10GbE SW
- Local buffer dual port memory (512K + 256K 18 bits) 12288 samples/ch



32ch prototype



WA105 B/E board

Bittware S5-PCIe-HQ board : S5PHQ-D8

- FPGA

Altera® Stratix® V GX/GS FPGA
20 full-duplex, high-performance, multi-gigabit SerDes transceivers @ up to 14.1 GHz
Up to 952,000 logic elements (LEs) available
Up to 62 Mb of embedded memory
1.4 Gbps LVDS performance
Up to 3,926 18×18 variable-precision multipliers
Embedded HardCopy Blocks

- Memory

Two banks of up to 8 GBytes DDR3 SDRAM (x64)
Four banks of up to 18 MBytes QDRII+ (x18)
128 MBytes of Flash memory for booting FPGA

- PCIe Interface

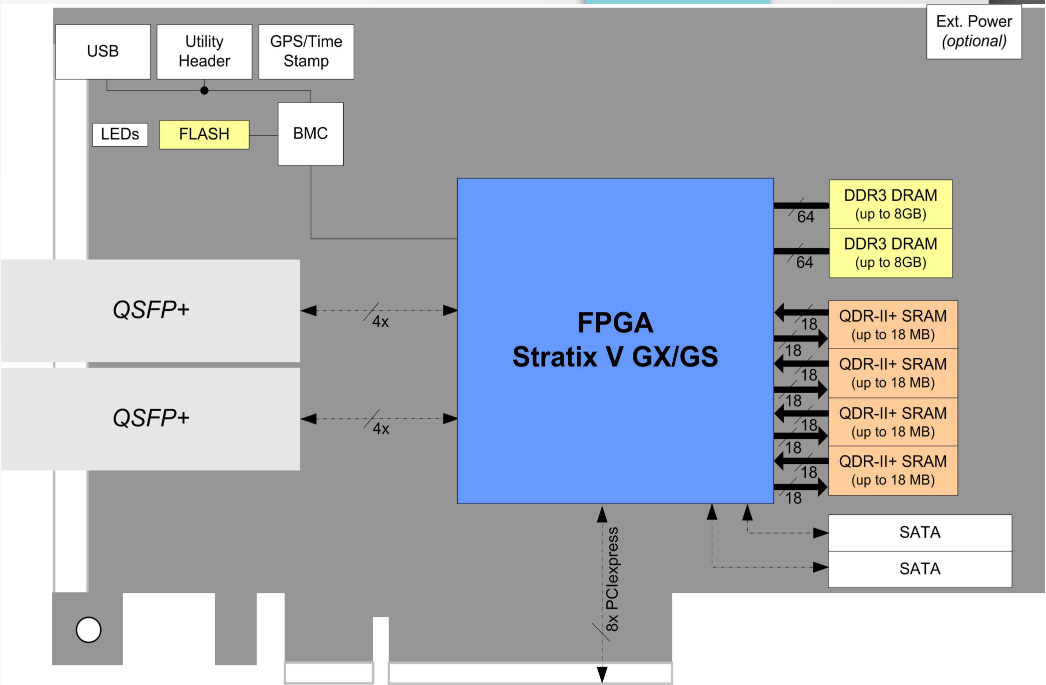
x8 Gen1, Gen2, Gen3 direct to FPGA

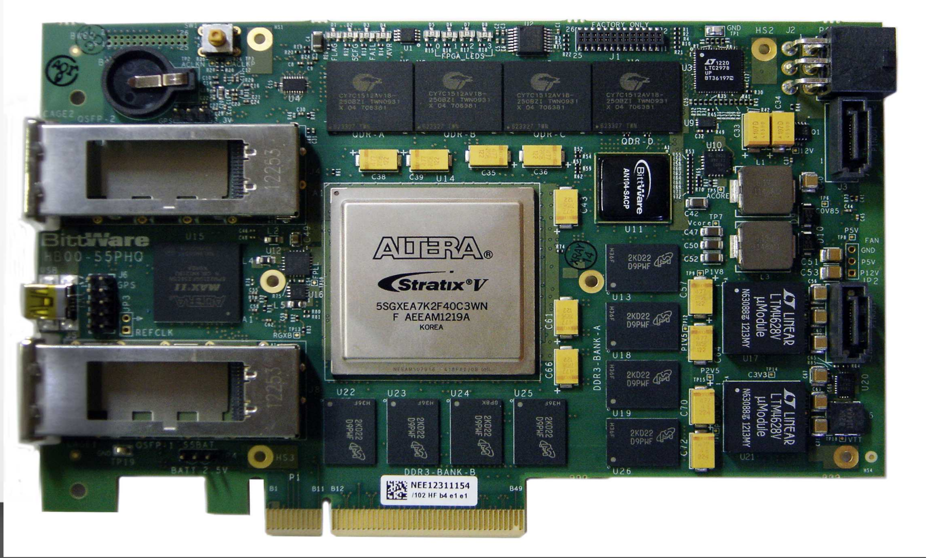
- Debug Utility Header

RS-232 port to Stratix V
JTAG debug interface to Stratix V

- QSFP+ Cages

2 QSFP+ cages on front panel connected directly to FPGA via 8 SerDes (no external PHY)
Each supports 40 GigE or four 10 GigE interfaces
Can be optionally adapted for use as SFP+





Future FPGA-based B/E system

Terabox

- 24 TeraFLOPS processing: 16x Altera Arria 10 or Stratix V FPGAs

Up to 18 million logic elements (Arria 10 GX)
Up to 62,000 multipliers (Stratix V GS)
Dual card wrt WA105 cards
New FPGA-board with Stratix X available by the time of DUNE

- 1.28 Terabits/sec I/O

- 128x 10GigE, 32x 40GigE, 32x 100 GigE,
  or 32x QDR Infiniband

- 6.5 Terabits/sec memory bandwidth

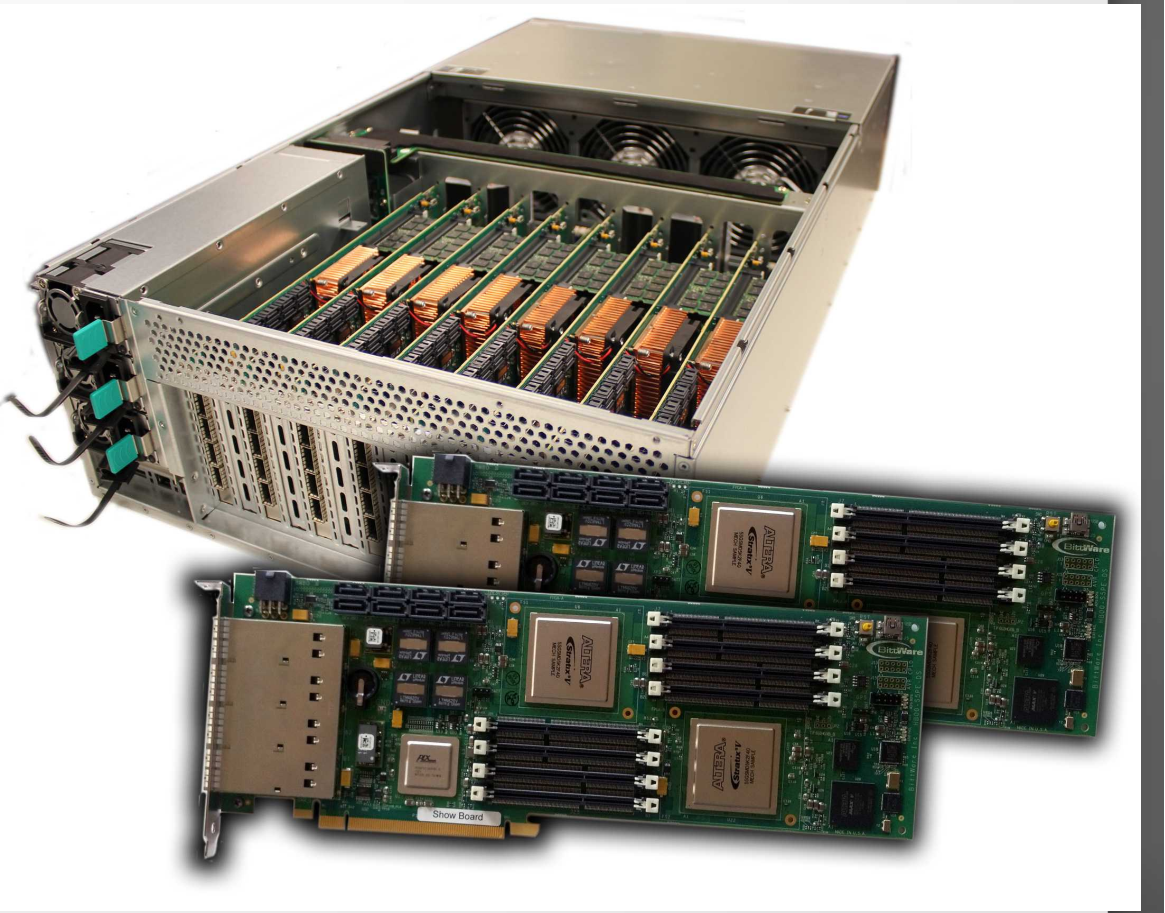
Up to 64 banks DDR3-1600 (512 GBytes)
DDR4, QDR-IV, QDRII+, and RLDRAM3 memory options

- 4U or 5U Rackmount PCIe system
  (server, industrial, or expansion)

Dual socket Intel Ivy Bridge with up to 12 cores
Up to 768 GBytes of system memory
8 Gen3 x16 PCIe slots

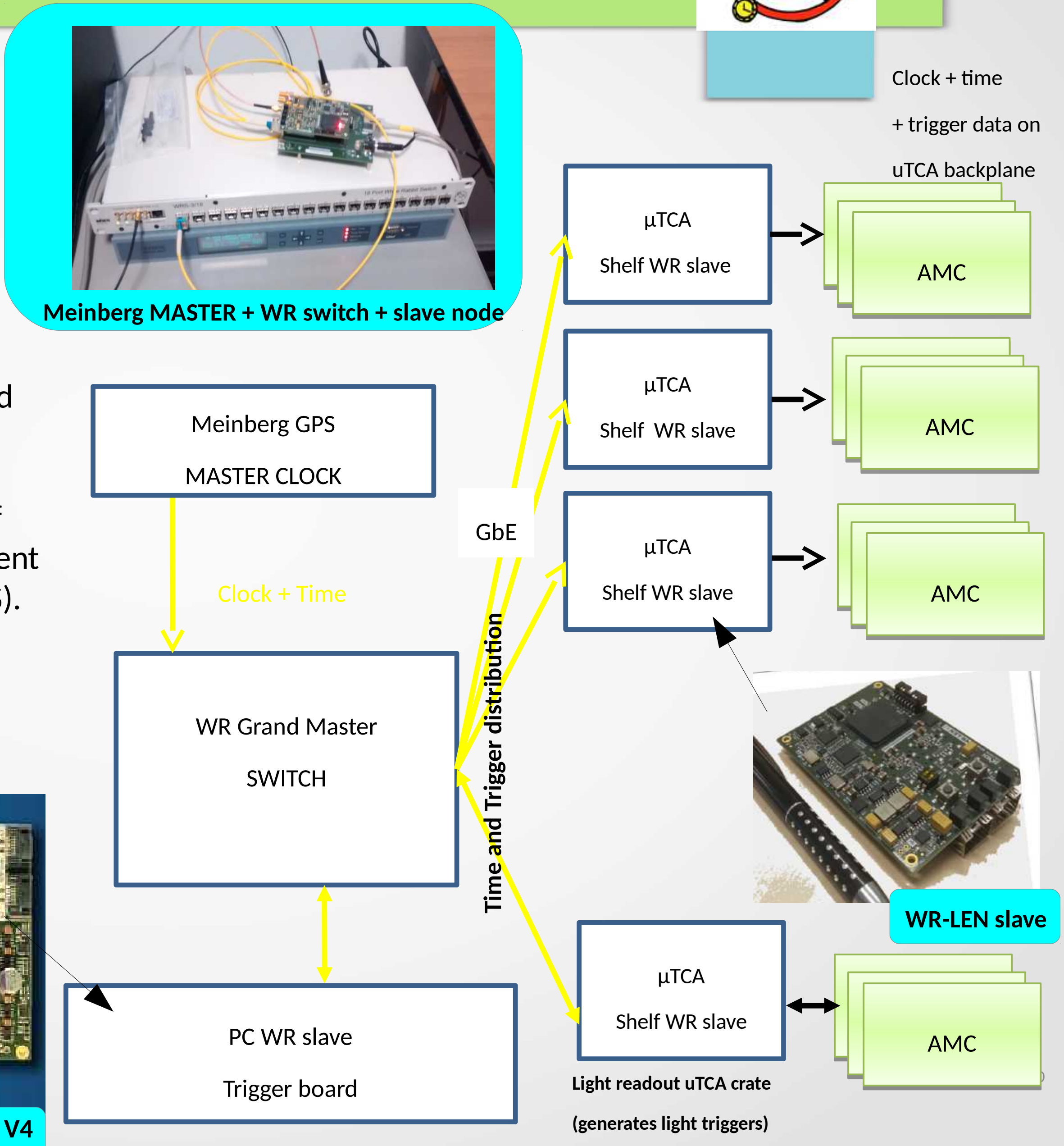
- Complete software support

Windows and Linux 64 drivers, interface libraries, and
hardware management
FPGA development kit for Arria 10 and Stratix V

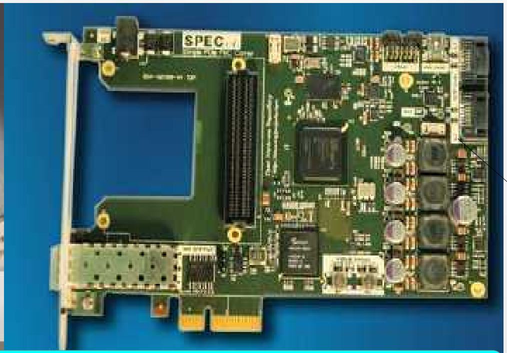


White Rabbit scheme

- WR is an evolution of the synchronization scheme based on **synchronous Ethernet + PTP** which was previously developed at IPNL in 2008: http://arxiv.org/abs/0906.2325
- WR is accurate at sub-ns level, enough to align the 400ns samples
- At the level of the charge readout DAQ is distributed the beam trigger timestamp.
- Trigger time info starts and closes the acquisition of the samples belonging to the drift window of an event in each AMC (important when operating without ZS).
- The beam trigger can be time-stamped on the PC trigger board and be broadcasted to the microTCA crates via the WR time distribution network



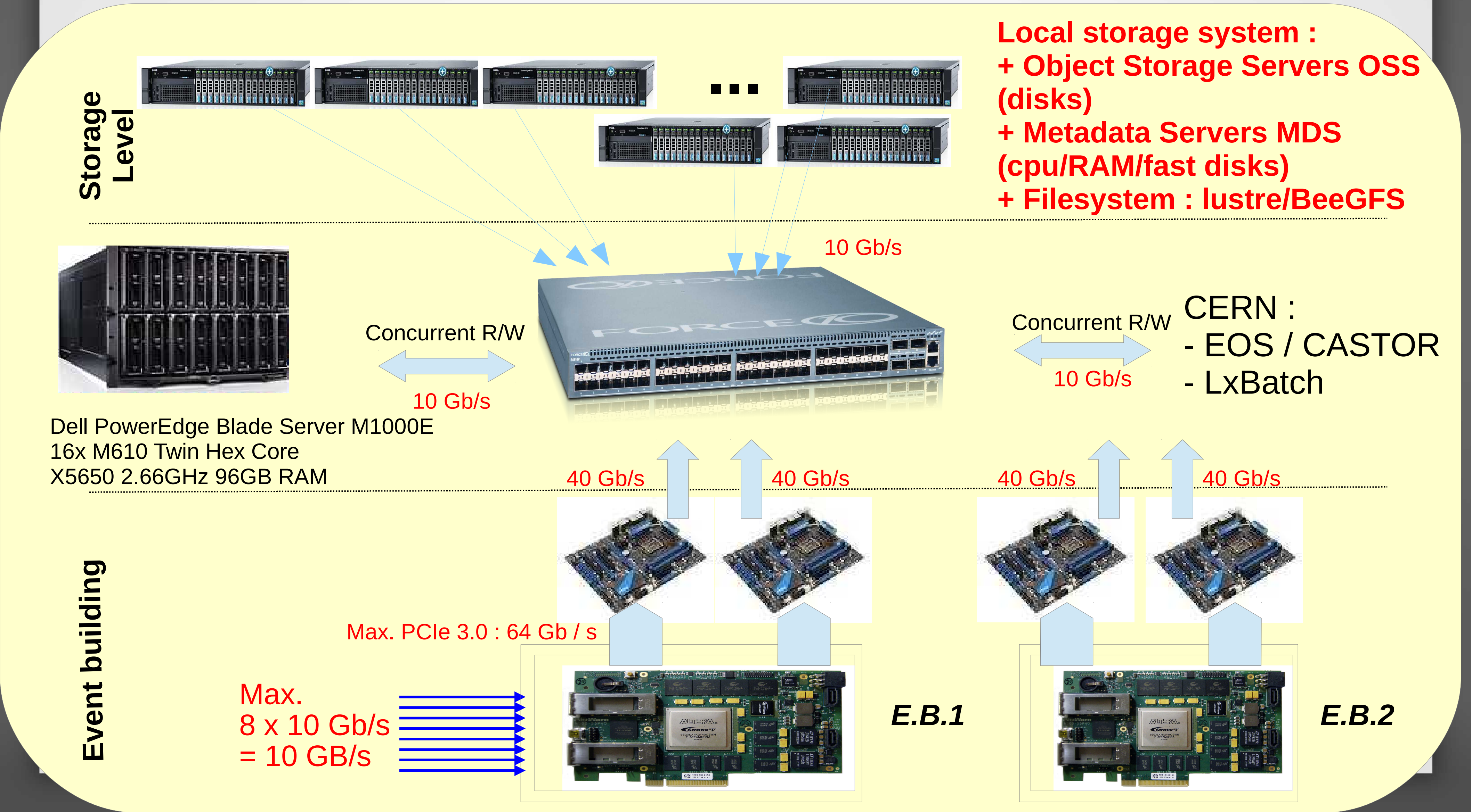
FMC Fine Delay 1 ns 4 channels



SPEC FMC PCIe carrier V4

Distributed DELL-based solution

CERN requirements : ~3 days autonomous data storage for each experiment : ~1PB
WA105 ~ LHC-experiment requirements



Storage Level

Local storage system :
+ Object Storage Servers OSS (disks)
+ Metadata Servers MDS (cpu/RAM/fast disks)
+ Filesystem : lustre/BeeGFS

...

10 Gb/s

Concurrent R/W

10 Gb/s

Dell PowerEdge Blade Server M1000E
16x M610 Twin Hex Core
X5650 2.66GHz 96GB RAM

Concurrent R/W

10 Gb/s

CERN :
- EOS / CASTOR
- LxBatch

40 Gb/s

40 Gb/s

40 Gb/s

40 Gb/s

Event building

Max. PCIe 3.0 : 64 Gb / s

Max.
8 x 10 Gb/s
= 10 GB/s

E.B.1

E.B.2

DELL-based solution : configuration

storage servers :

* 15 R730XD (storage servers) including :
* 16 disks 6To
* 32Go RAM
* 2 disks system RAID 1, 300 Go 10k
* 1 network card Intel X540 double port 10 GB
* 4 years extended guarantee (D+1 intervention)
* 2 processors Intel Xeon E5-2609 v3
* raid H730P
* Rails with management arm
* double power supply

metadata servers (MDS) :

* 2 R630 (metadata servers), including :
* 2 disks 200 Go SSD SAS Mix Use MLC 12Gb/s
* 2 processors Intel Xeon E5-2630 v3
* 32Go DDR4
* RAID H730p
* network : Intel X540 2 ports 10 Gb
* 4 years extended guarantee (D+1 intervention)
* Rails with management arm
* double power supply

configuration server :

* 1 R430 (configuration server)
* 1 processor E5-2603 v3
* RAID H730
* 2 hard disks 500 Go Nearline SAS 6 Gbps 7,2k
* 16 Go DDR4
* Rails with management arm
* double power supply

Offline computing farm: 16*24 = 384 cores

* 1 blade center PowerEdge M1000e with 16 lames M630, each including :
* 128Go DDR4
* 2 processors Intel Xeon E5-2670 v3
* 4 years extended guarantee (D+1 intervention)
* 2 hard disks 500 Go SATA 7200 Tpm
* netwok Intel X540 10 Gb

Switch Force10, S4820T (see next slide) :

***** 48 x 10GbaseT ports
* 4 x 40G QSFP+ ports
* 1 x AC PSU
* 2 fans

10/40 switches specs.

FORCE10 / CISCO :

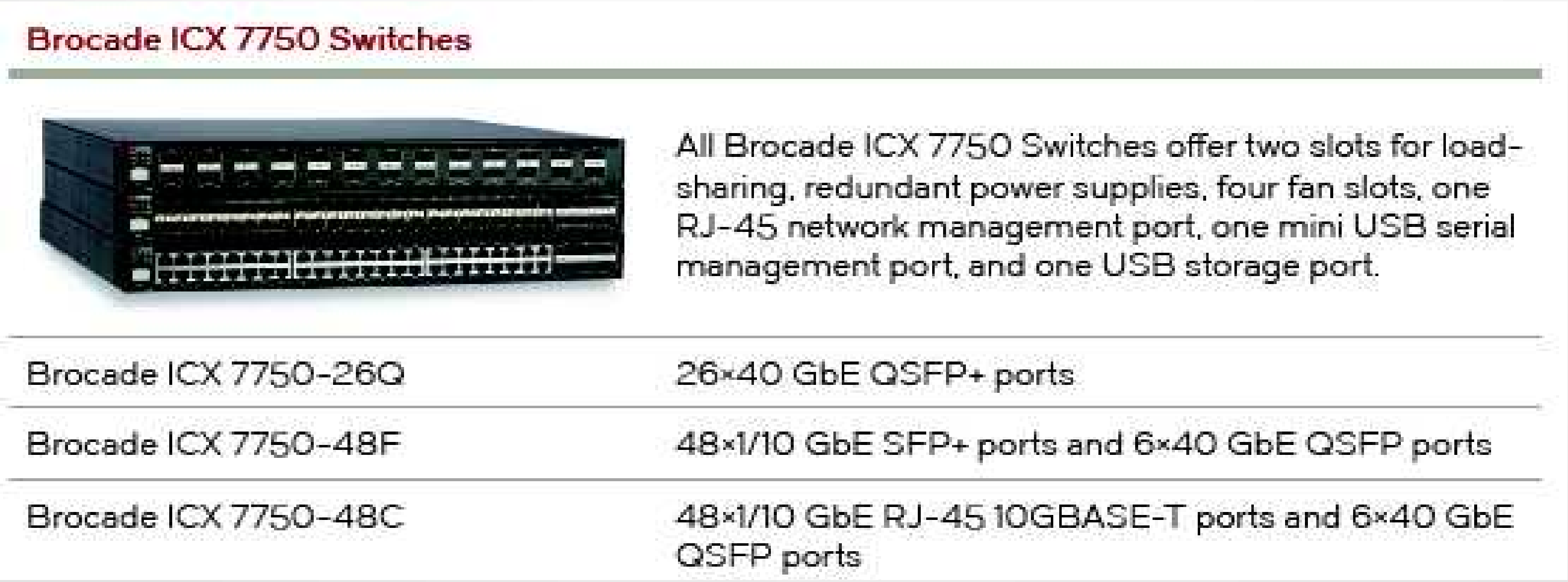
OPTION 1 : Nexus 3172TQ : 48 ports 1/10G baseT, 6 ports 40G QSFP+
N3K-C3172TQ-10GT Nexus 3172T 48 x 1/10GBase-T and 6 QSFP+ ports

OPTION 2 : Nexus 9372TX : 48 ports 1/10G baseT, 6 ports 40G QSFP+
N9K-C9372TX Nexus 9300 with 48p 1/10G-T and 6p 40G QSFP+

OPTION 3 : Nexus 9396TX : 48 ports 1/10G baseT, 12 ports 40G QSFP+
N9K-C9396TX Nexus 9300 48p 1/10GBASE-T & additional uplink module req.
N9K-M12PQ ACI Uplink Module for Nexus 9300 12p 40G QSFP

~10 to 15 k€

BROCADE :



Brocade ICX 7750 Switches

All Brocade ICX 7750 Switches offer two slots for load-sharing, redundant power supplies, four fan slots, one RJ-45 network management port, one mini USB serial management port, and one USB storage port.

Brocade ICX 7750-26Q	26×40 GbE QSFP+ ports
Brocade ICX 7750-48F	48×1/10 GbE SFP+ ports and 6×40 GbE QSFP ports
Brocade ICX 7750-48C	48×1/10 GbE RJ-45 10GBASE-T ports and 6×40 GbE QSFP ports

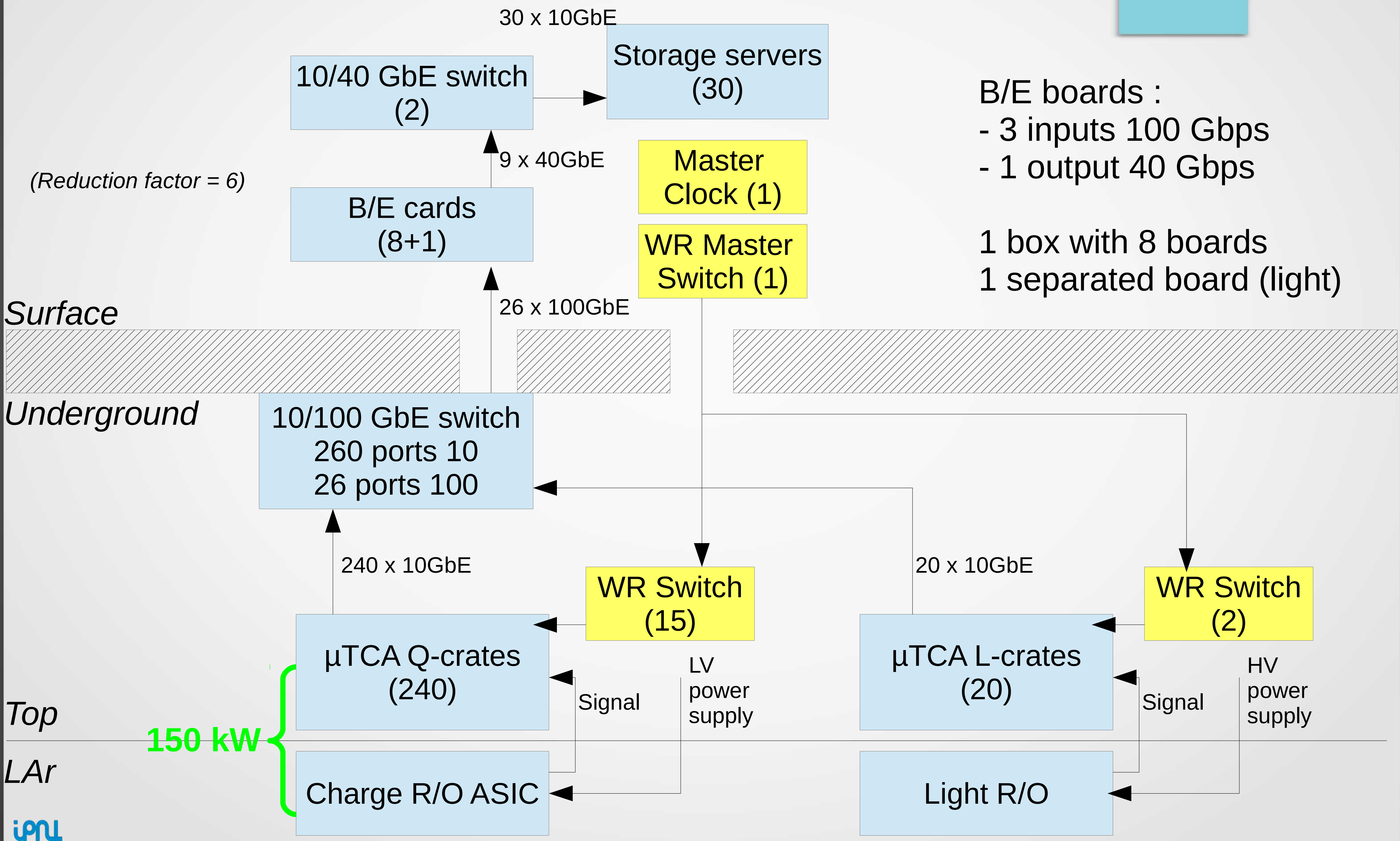
Conclusions

- WA105 DAQ system is developed using **µTCA** and **White Rabbit** standards.
- The online Event Building is performed on a reduced number of powerful **FPGA** processing boards, using OpenCL framework for software developments.
- Offline local storage is designed through a **10/40Gbe network**.
- **The scheme is scalable to DUNE (µTCA+WR, 10/100Gbe, B/E)**

3x1x1 prototype used as a testbench for WA105 readout & DAQ chains (**¼ of 6x6x6) :**

- 3 crates for charge readout (synchronized via WR)
- 1 B/E card + 1 Event Builder (EB)
- 1 local storage system (~100 TB distributed among 5 servers)
- 1 switch 10/40 to interconnect the crates, the EB, the local storage and CERN EOS

DUNE proposed scheme



B/E boards :
- 3 inputs 100 Gbps
- 1 output 40 Gbps

1 box with 8 boards
1 separated board (light)