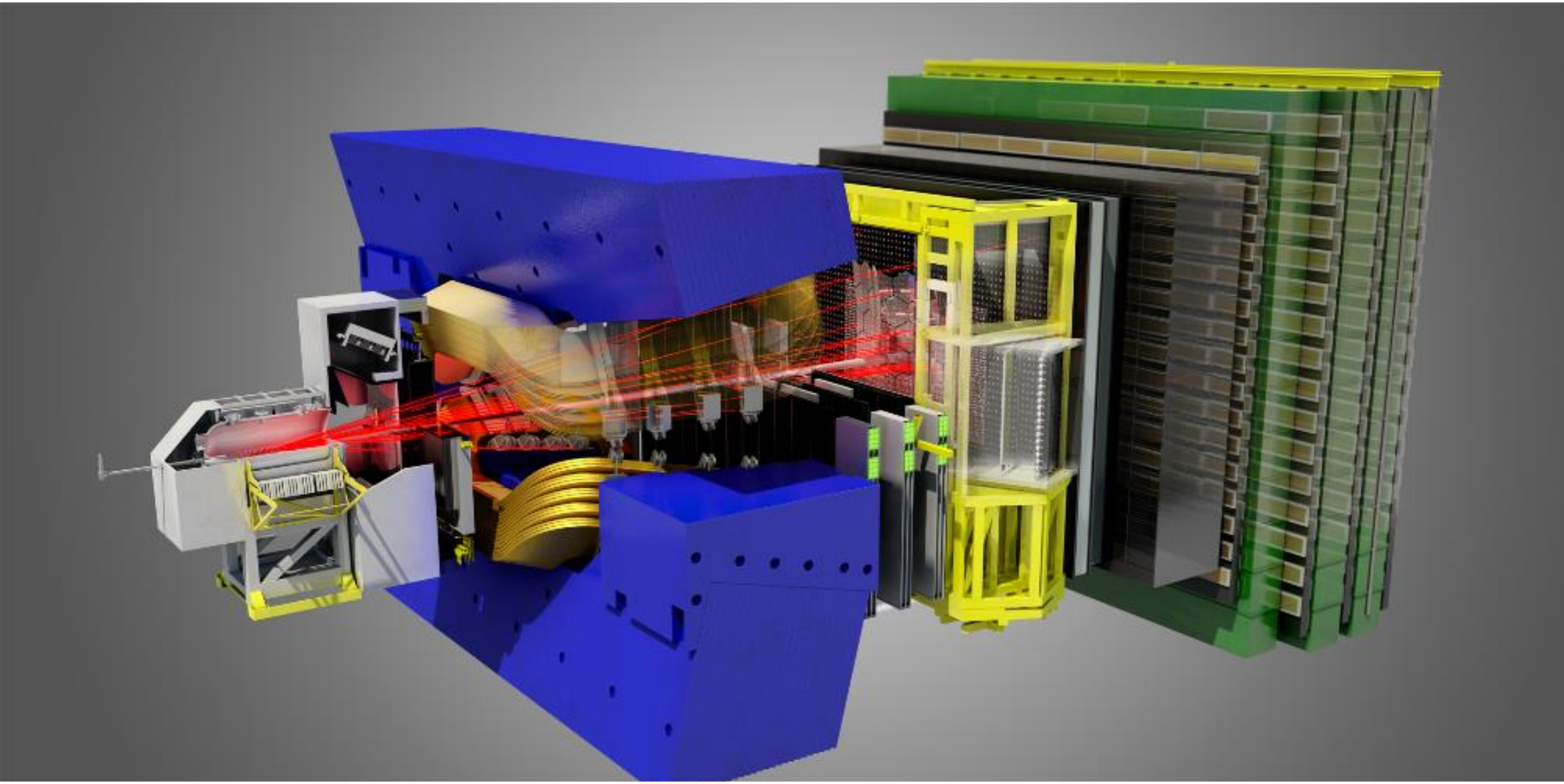


LHCb: status and plans for the online system upgrade

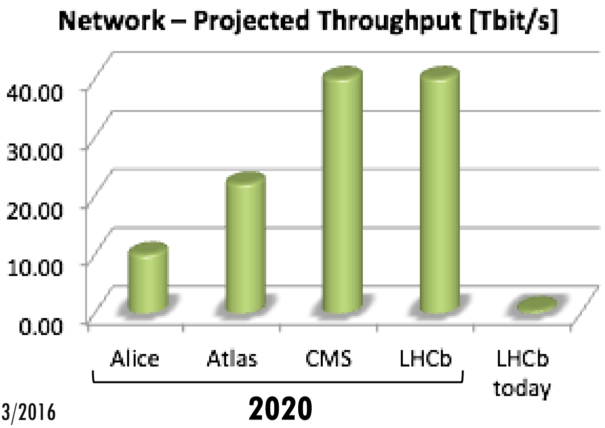
Paolo Durante
paolo.durante@cern.ch
on behalf of the LHCb collaboration

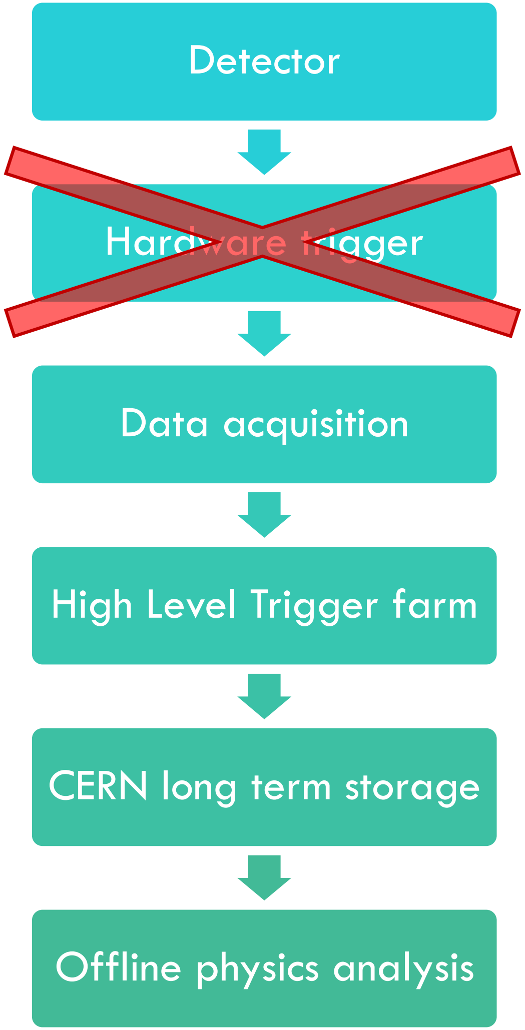
Outline

- New readout system
  - Slow & fast control
  - Optical links
  - Readout board
  - Event building
  - Data centre
- Online mile-stones

Run3 upgrade

- Higher luminosity from LHC
- Sub-detectors will upgrade
- Removal of hardware trigger
  - Inefficient at high luminosity
  - New readout electronics
- Filter farm will need to handle:
  - Larger **event size** (~50KB to ~100KB)
  - Larger **event rate** (~1MHz to ~40MHz)
- New challenges for DAQ & High-Level Trigger





Detector

~~Hardware trigger~~

Data acquisition

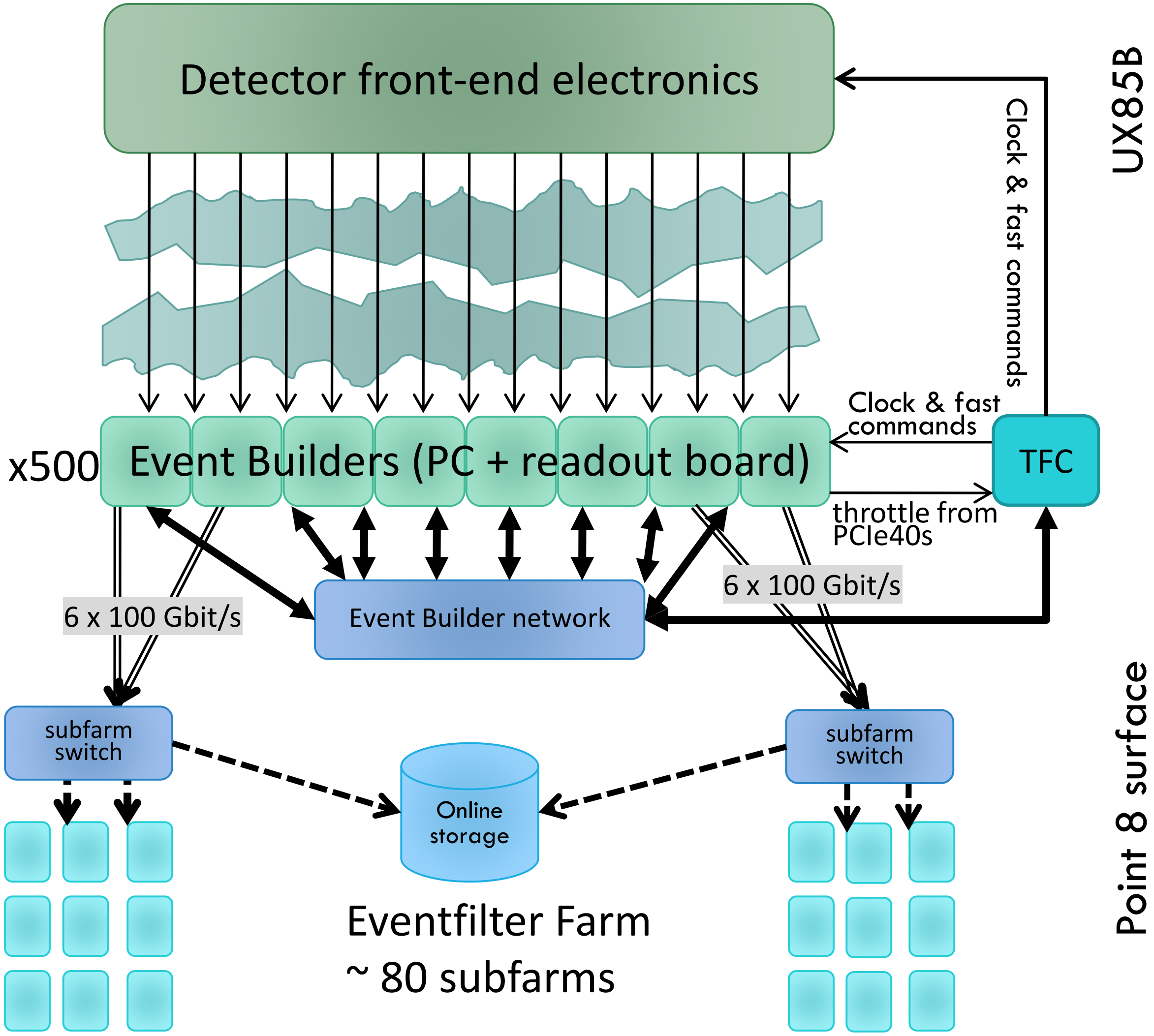
High Level Trigger farm

CERN long term storage

Offline physics analysis

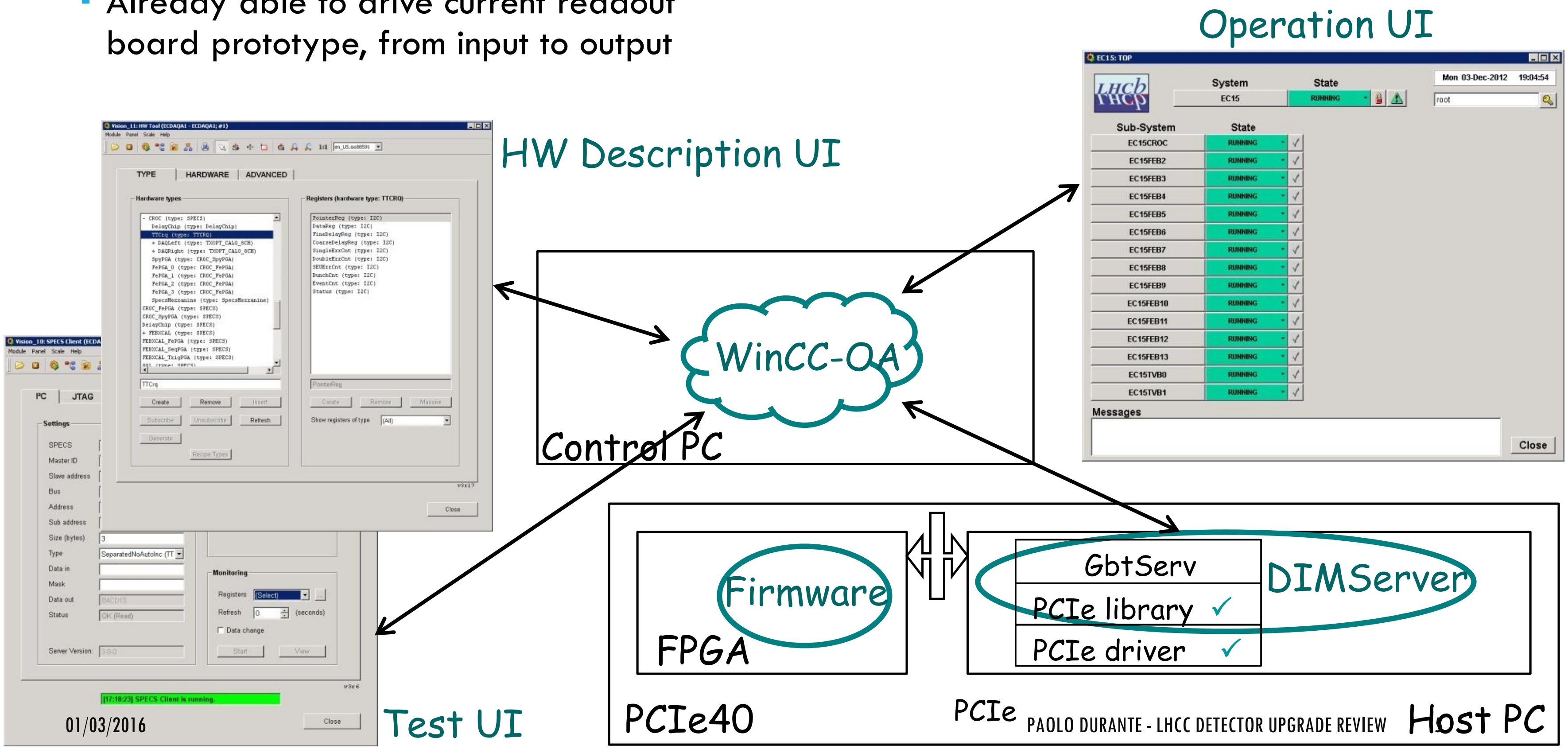
Run3 Online System

- Dimensioning the system:
  - **~10 000** optical links
    - detector to surface **(~350 m)**
    - up to **4.8 Gbps/link**
  - **~500** readout nodes (up to ~24 links/node)
  - **~40 MHz** event rate
  - **~100 KB** event size
- High bisection bandwidth in event builder network
  - ~32 Tb/s aggregate bandwidth
  - Leverage emerging 100G technologies
- Global configuration and control via ECS subsystem
- Global synchronization via TFC subsystem



ECS (Experiment Control System)

- Controls and monitors <u>all subsystems</u>
  - DAQ, TFC, HLT, farm…
- Continuity from current implementation
  - JCOP / DIM / WinCCOA / SMI++ / Recipes
- Already able to drive current readout board prototype, from input to output
- Frontends rely on GBT-SCA hardware by EP-ESE
- Low-level components are being implemented



EC15: TOP

Mon 03-Dec-2012 19:04:54

root

System	State
EC15	RUNNING

Sub-System	State
EC15CROC	RUNNING
EC15FEB2	RUNNING
EC15FEB3	RUNNING
EC15FEB4	RUNNING
EC15FEB5	RUNNING
EC15FEB6	RUNNING
EC15FEB7	RUNNING
EC15FEB8	RUNNING
EC15FEB9	RUNNING
EC15FEB10	RUNNING
EC15FEB11	RUNNING
EC15FEB12	RUNNING
EC15FEB13	RUNNING
EC15TVB0	RUNNING
EC15TVB1	RUNNING

Messages

Vision_11: HW Tool (ECDAQA1 - ECDAQA1; #1)

TYPE | HARDWARE | ADVANCED

Hardware types

```
- CROC (type: SPECS)
   DelayChip (type: DelayChip)
   TTCrq (type: TTCRQ)
   + DAQLeft (type: TXOPT_CALO_8CH)
   + DAQRight (type: TXOPT_CALO_8CH)
   SpyPGA (type: CROC_SpyPGA)
   FePGA_0 (type: CROC_FePGA)
   FePGA_1 (type: CROC_FePGA)
   FePGA_2 (type: CROC_FePGA)
   FePGA_3 (type: CROC_FePGA)
   SpecsMezzanine (type: SpecsMezzanine)
CROC_FePGA (type: SPECS)
CROC_SpyPGA (type: SPECS)
DelayChip (type: SPECS)
+ FEBXCAL (type: SPECS)
FEBXCAL_FePGA (type: SPECS)
FEBXCAL_SeqPGA (type: SPECS)
FEBXCAL_TrigPGA (type: SPECS)
```

TTCrq

Registers (hardware type: TTCRQ)

```
PointerReg (type: I2C)
DataReg (type: I2C)
FineDelayReg (type: I2C)
CoarseDelayReg (type: I2C)
SingleErrCnt (type: I2C)
DoubleErrCnt (type: I2C)
SEUErrCnt (type: I2C)
BunchCnt (type: I2C)
EventCnt (type: I2C)
Status (type: I2C)
```

PointerReg

Show registers of type: (All)

Vision_10: SPECS Client (ECDA

I²C | JTAG

Settings

Size (bytes): 3

Type: SeparatedNoAutoInc (TT

Monitoring

Registers: (Select)

Refresh: 0 (seconds)

☐ Data change

[17:18:23] SPECS Client is running.

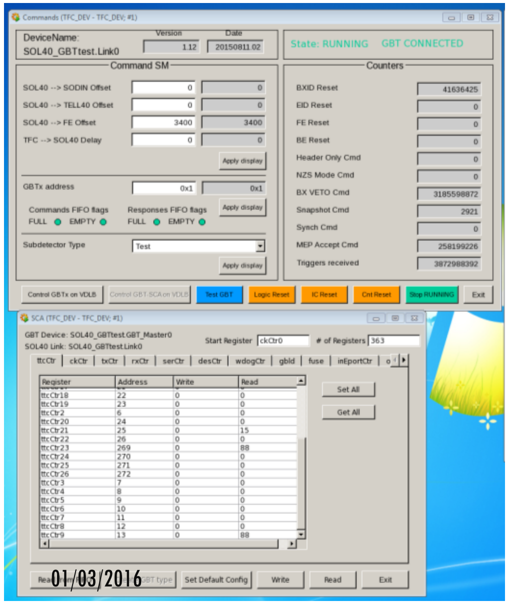
TFC (Timing & Fast Control)

Current status

- Already integrated in firmware
- Uses same readout board hardware as the DAQ (PCIe40)
- Can send fast commands to frontends
  - SciFi, UT, Muon ASICS already being tested
- Programmable internal throttle for bandwidth regulation

Ongoing work

- PON (Passive Optical Network) technology integration (with EP-ESE)
- Clock phase tests on readout board (with CPPM)
- Continue feedback and compliance testing with frontend experts



Commands (TFC_DEV - TFC_DEV; #1)

DeviceName: SOL40_GBTtest.Link0 — Version 1.12 — Date 20150811.02

State: RUNNING GBT CONNECTED

Command SM

SOL40 --> SODIN Offset	0	0
SOL40 --> TELL40 Offset	0	0
SOL40 --> FE Offset	3400	3400
TFC --> SOL40 Delay	0	0
GBTx address	0x1	0x1

Commands FIFO flags FULL ● EMPTY ● — Responses FIFO flags FULL ● EMPTY ●

Subdetector Type: Test

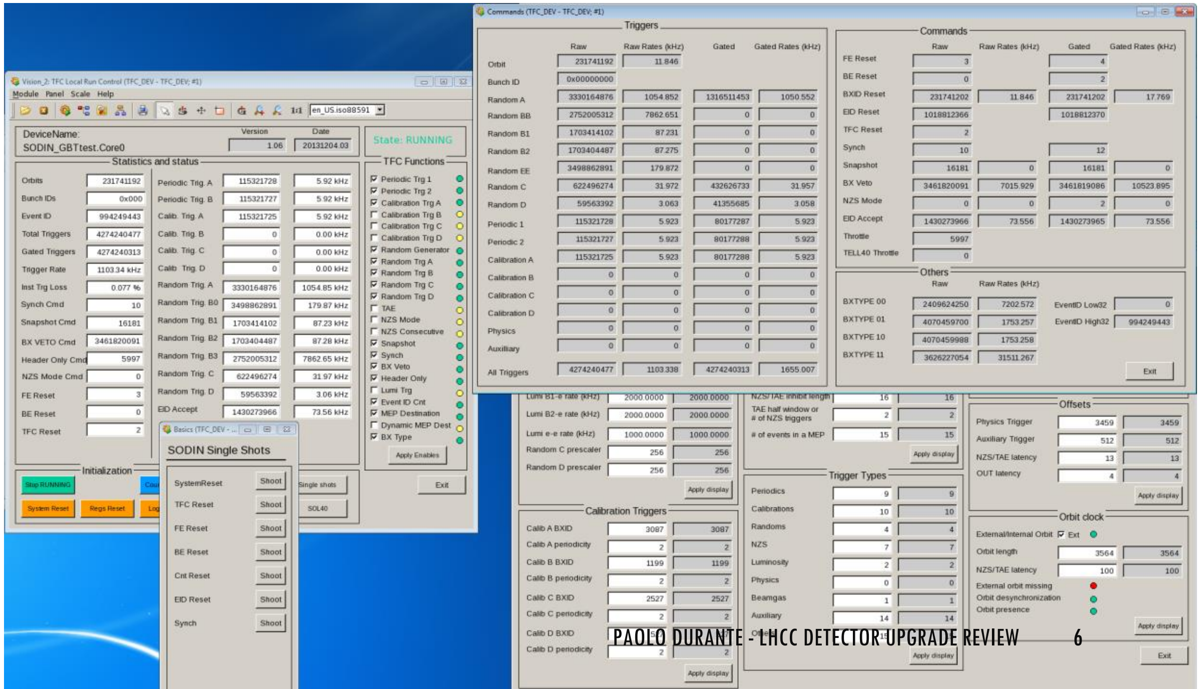
Counters

BXID Reset	41636425
EID Reset	0
FE Reset	0
BE Reset	0
Header Only Cmd	0
NZS Mode Cmd	0
BX VETO Cmd	3185598872
Snapshot Cmd	2921
Synch Cmd	0
MEP Accept Cmd	258199226
Triggers received	3872988392

SCA (TFC_DEV - TFC_DEV; #1)

GBT Device: SOL40_GBTtest.GBT_Master0
SOL40 Link: SOL40_GBTtest.Link0
Start Register: ckCtr0 — # of Registers: 363

Register	Address	Write	Read
ttcCtr18	22	0	0
ttcCtr19	23	0	0
ttcCtr2	6	0	0
ttcCtr20	24	0	0
ttcCtr21	25	0	15
ttcCtr22	26	0	0
ttcCtr23	269	0	88
ttcCtr24	270	0	0
ttcCtr25	271	0	0
ttcCtr26	272	0	0
ttcCtr3	7	0	0
ttcCtr4	8	0	0
ttcCtr5	9	0	0
ttcCtr6	10	0	0
ttcCtr7	11	0	0
ttcCtr8	12	0	0
ttcCtr9	13	0	88



Vision_2: TFC Local Run Control (TFC_DEV - TFC_DEV; #1)

DeviceName: SODIN_GBTtest.Core0 — Version 1.06 — Date 20131204.03

State: RUNNING

Statistics and status

Orbits	231741192	Periodic Trig. A	115321728	5.92 kHz
Bunch IDs	0x000	Periodic Trig. B	115321727	5.92 kHz
Event ID	994249443	Calib. Trig. A	115321725	5.92 kHz
Total Triggers	4274240477	Calib. Trig. B	0	0.00 kHz
Gated Triggers	4274240313	Calib. Trig. C	0	0.00 kHz
Trigger Rate	1103.34 kHz	Calib. Trig. D	0	0.00 kHz
Inst Trg Loss	0.077 %	Random Trig. A	3330164876	1054.85 kHz
Synch Cmd	10	Random Trig. B0	3498862891	179.87 kHz
Snapshot Cmd	16181	Random Trig. B1	1703414102	87.23 kHz
BX VETO Cmd	3461820091	Random Trig. B2	1703404487	87.28 kHz
Header Only Cmd	5997	Random Trig. B3	2752005312	7862.65 kHz
NZS Mode Cmd	0	Random Trig. C	622496274	31.97 kHz
FE Reset	3	Random Trig. D	59563392	3.06 kHz
BE Reset	0	EID Accept	1430273966	73.56 kHz
TFC Reset	2			

TFC Functions

- ☑ Periodic Trg 1 ●
- ☑ Periodic Trg 2 ●
- ☑ Calibration Trg A ●
- ☐ Calibration Trg B ○
- ☐ Calibration Trg C ○
- ☐ Calibration Trg D ○
- ☑ Random Generator ●
- ☑ Random Trg A ●
- ☑ Random Trg B ●
- ☑ Random Trg C ●
- ☑ Random Trg D ●
- ☐ TAE ○
- ☐ NZS Mode ○
- ☐ NZS Consecutive ○
- ☑ Snapshot ●
- ☑ Synch ●
- ☑ BX Veto ●
- ☑ Header Only ●
- ☐ Lumi Trg ○
- ☑ Event ID Cnt ●
- ☑ MEP Destination ●
- ☐ Dynamic MEP Dest ○
- ☑ BX Type ●

Basics (TFC_DEV - ...)

SODIN Single Shots

SystemReset	Shoot
TFC Reset	Shoot
FE Reset	Shoot
BE Reset	Shoot
Cnt Reset	Shoot
EID Reset	Shoot
Synch	Shoot

Commands (TFC_DEV - TFC_DEV; #1)

Triggers

	Raw	Raw Rates (kHz)	Gated	Gated Rates (kHz)
Orbit	231741192	11.846		
Bunch ID	0x00000000			
Random A	3330164876	1054.852	1316511453	1050.552
Random BB	2752005312	7862.651	0	0
Random B1	1703414102	87.231	0	0
Random B2	1703404487	87.275	0	0
Random EE	3498862891	179.872	0	0
Random C	622496274	31.972	432626733	31.957
Random D	59563392	3.063	41355685	3.058
Periodic 1	115321728	5.923	80177287	5.923
Periodic 2	115321727	5.923	80177288	5.923
Calibration A	115321725	5.923	80177288	5.923
Calibration B	0	0	0	0
Calibration C	0	0	0	0
Calibration D	0	0	0	0
Physics	0	0	0	0
Auxiliary	0	0	0	0
All Triggers	4274240477	1103.338	4274240313	1655.007

Commands

	Raw	Raw Rates (kHz)	Gated	Gated Rates (kHz)
FE Reset	3		4	
BE Reset	0		2	
BXID Reset	231741202	11.846	231741202	17.769
EID Reset	1018812366		1018812370	
TFC Reset	2			
Synch	10		12	
Snapshot	16181	0	16181	0
BX Veto	3461820091	7015.929	3461819086	10523.895
NZS Mode	0	0	2	0
EID Accept	1430273966	73.556	1430273965	73.556
Throttle	5997			
TELL40 Throttle	0			

Others

	Raw	Raw Rates (kHz)		
BXTYPE 00	2409624250	7202.572	EventID Low32	0
BXTYPE 01	4070459700	1753.257	EventID High32	994249443
BXTYPE 10	4070459988	1753.258		
BXTYPE 11	3626227054	31511.267		

Lumi B1-e rate (kHz)	2000.0000	2000.0000
Lumi B2-e rate (kHz)	2000.0000	2000.0000
Lumi e-e rate (kHz)	1000.0000	1000.0000
Random C prescaler	256	256
Random D prescaler	256	256

NZS/TAE inhibit length	16	16
TAE half window or # of NZS triggers	2	2
# of events in a MEP	15	15

Calibration Triggers

Calib A BXID	3087	3087
Calib A periodicity	2	2
Calib B BXID	1199	1199
Calib B periodicity	2	2
Calib C BXID	2527	2527
Calib C periodicity	2	2
Calib D BXID	[illegible]	[illegible]
Calib D periodicity	2	2

Trigger Types

Periodics	9	9
Calibrations	10	10
Randoms	4	4
NZS	7	7
Luminosity	2	2
Physics	0	0
Beamgas	1	1
Auxiliary	14	14
Others	15	[illegible]

Offsets

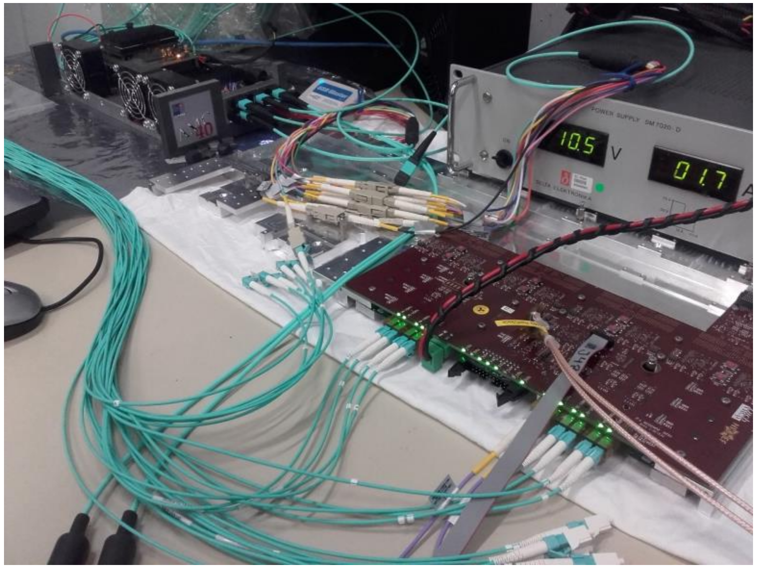
Physics Trigger	3459	3459
Auxiliary Trigger	512	512
NZS/TAE latency	13	13
OUT latency	4	4

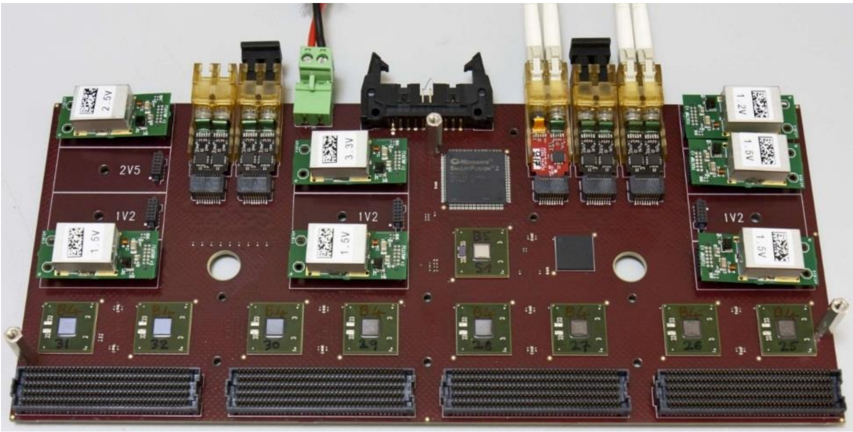
Orbit clock

External/Internal Orbit	☑ Ext ●	
Orbit length	3564	3564
NZS/TAE latency	100	100
External orbit missing	●	
Orbit desynchronization	●	
Orbit presence	●	

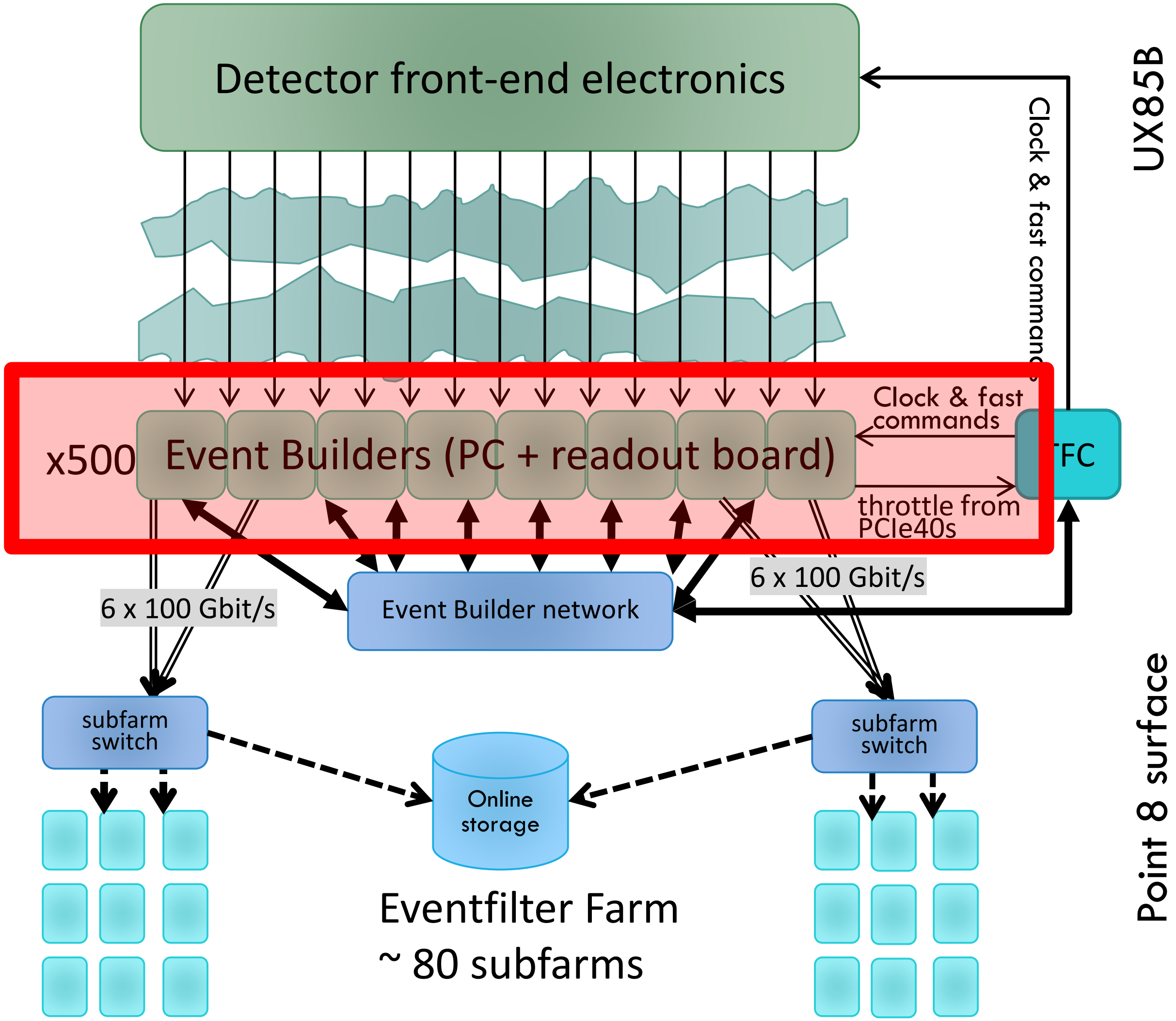
Long-distance optics

- Counting room on surface
  - Power, cooling, space constraints in underground area
  - ~350 meter distance
- Based on EP-ESE technology
  - Rad-hard Versatile Link on frontends
  - Initially qualified for ~100m
- Fiber infrastructure by EN-EL
  - Pilot installation at end 2014
- Loopback tests in 2015
  - ~9 months, ~700 meters
  - Avago MiniPOD transceivers
  - Bit Error Rate < 10^{-18}
  - Full system equivalent: < 5 errors/day
  - ✓ LHCC milestone
- Continued tests in 2016
  - Versatile TX on frontend prototype
  - MiniPOD RX on readout board prototype



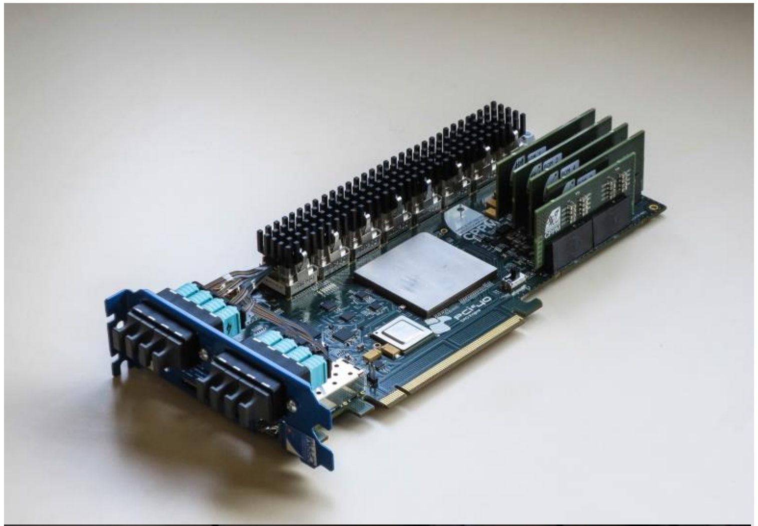


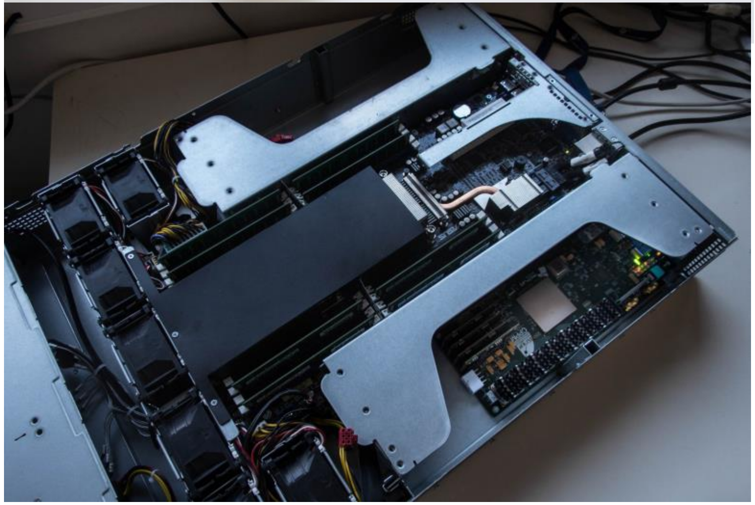
Readout boards / Event builders



Readout board hardware (PCIe40)

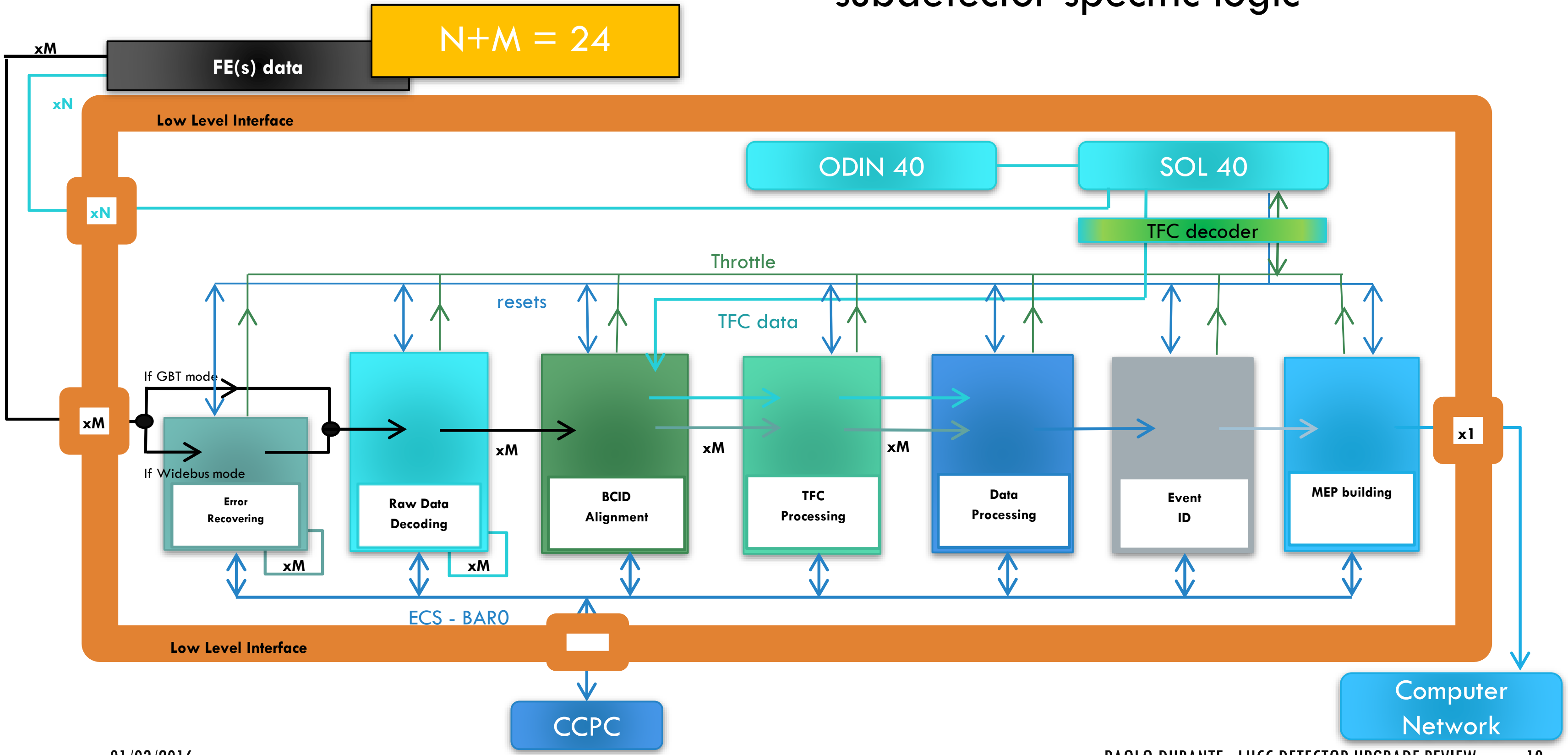
- PCI Express add-in card
  - Altera Arria10 FPGA
  - 100Gbps DMA engine to event-builder memory
- High-density optical IO
  - Up to 48 transceivers (Avago MiniPODs)
  - Reuse same HW for timing distribution system
- Decouple FPGA from network
  - Maximum flexibility in network technology
- Exploit commercial technologies
  - PCI Express Gen3 interconnect
  - COTS servers designed for GPU acceleration
- 2^nd generation readout board
  - Developed at CPP Marseille
- Pre-production launched
  - Ready end of Q2
- Market survey completed
  - Tender in H2 2016





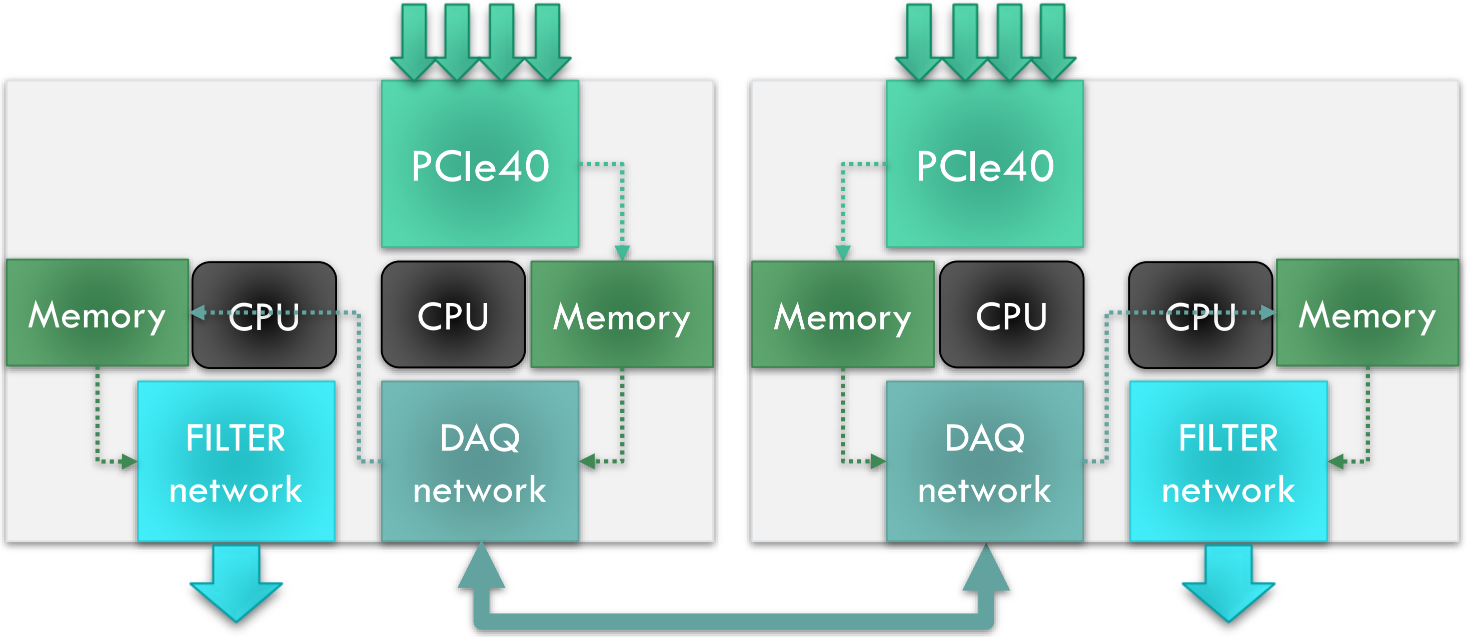
Readout board firmware

- Common architecture for all subdetectors
- Joint effort of LAPP, CERN, CPPM
- Highly configurable through frontend-specific parameters
- Handles ECS, TFC & DAQ, subdetector-specific logic



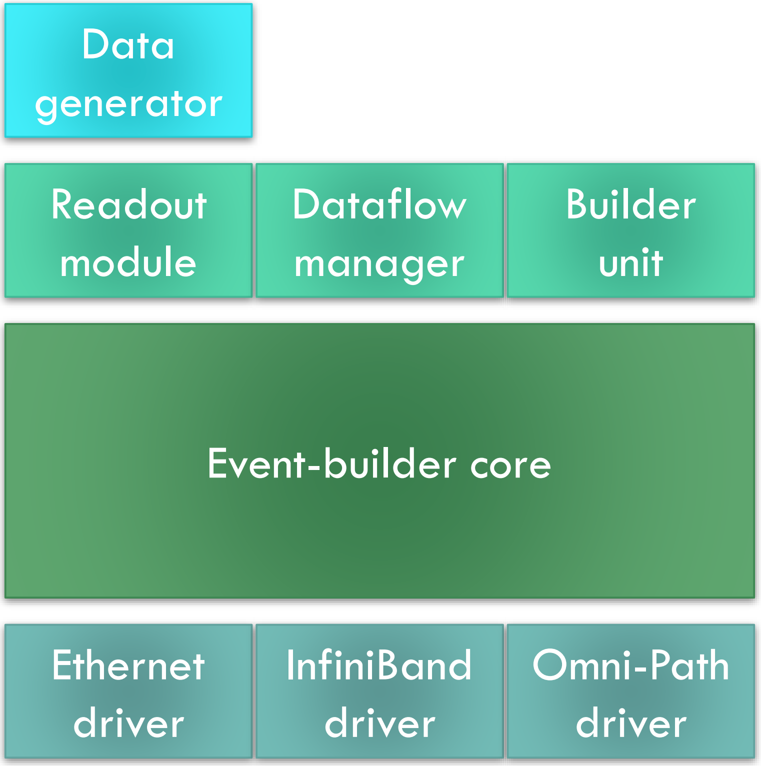
Readout unit dataflow

- A single Readout unit must sustain ~400Gbps IO bandwidth
- Optimize memory bandwidth
  - Design for zero-copy operations and RDMA over the network
  - Organize dataflow according to topology and IO resources
  - Exploit full network bandwidth



Event-building software (DAQPIPE)

- Recreate distributed event-building dataflow of LHCb Run3
- Modular architecture, “drivers” for each network technology under evaluation
- Leverage existing HPC sites to assess scalability
- Close collaboration with the industry through CERN OpenLab
- Already achieving ~86Gbps with Infiniband EDR
  - Meets our target
  - Reduced scale setup
- Off-site tests for Infiniband EDR and Intel Omni-Path currently ongoing…



Data generator

Readout module

Dataflow manager

Builder unit

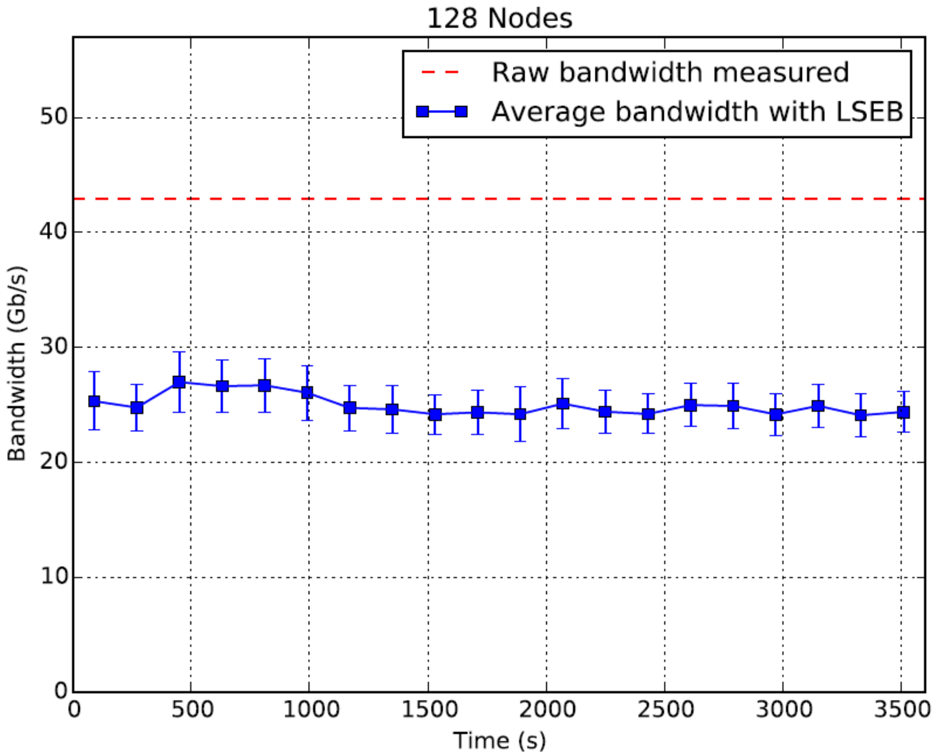
Event-builder core

Ethernet driver

InfiniBand driver

Omni-Path driver

128node InfiniBand scalability (CINECA)



- Cores allocated: 8/16
- Size of buffer: ~128KB
- Average bw: 24.98 Gb/s

Performance limited by non-exclusive utilization of cluster

Matteo Manzali - INFN CNAF - Università degli Studi di Ferrara

Future data centre at Point 8

	Turnkey commercial solution (Requires minimal support from CERN engineering groups)	**Leverage existing infrastructure at Point 8**
Building	Buy pre-fabricated containers from a commercial supplier	Accommodate the farm in an existing building (SX8 hall)
Cooling	Cooling solution depends on the vendor (e. g. free air cooling)	Passive rear-door heat exchangers using primary water from existing cooling towers ▪ Compatible with DCLC for hot spots ▪ Test setup at the pit to evaluate performance with on site “warm” cooling water

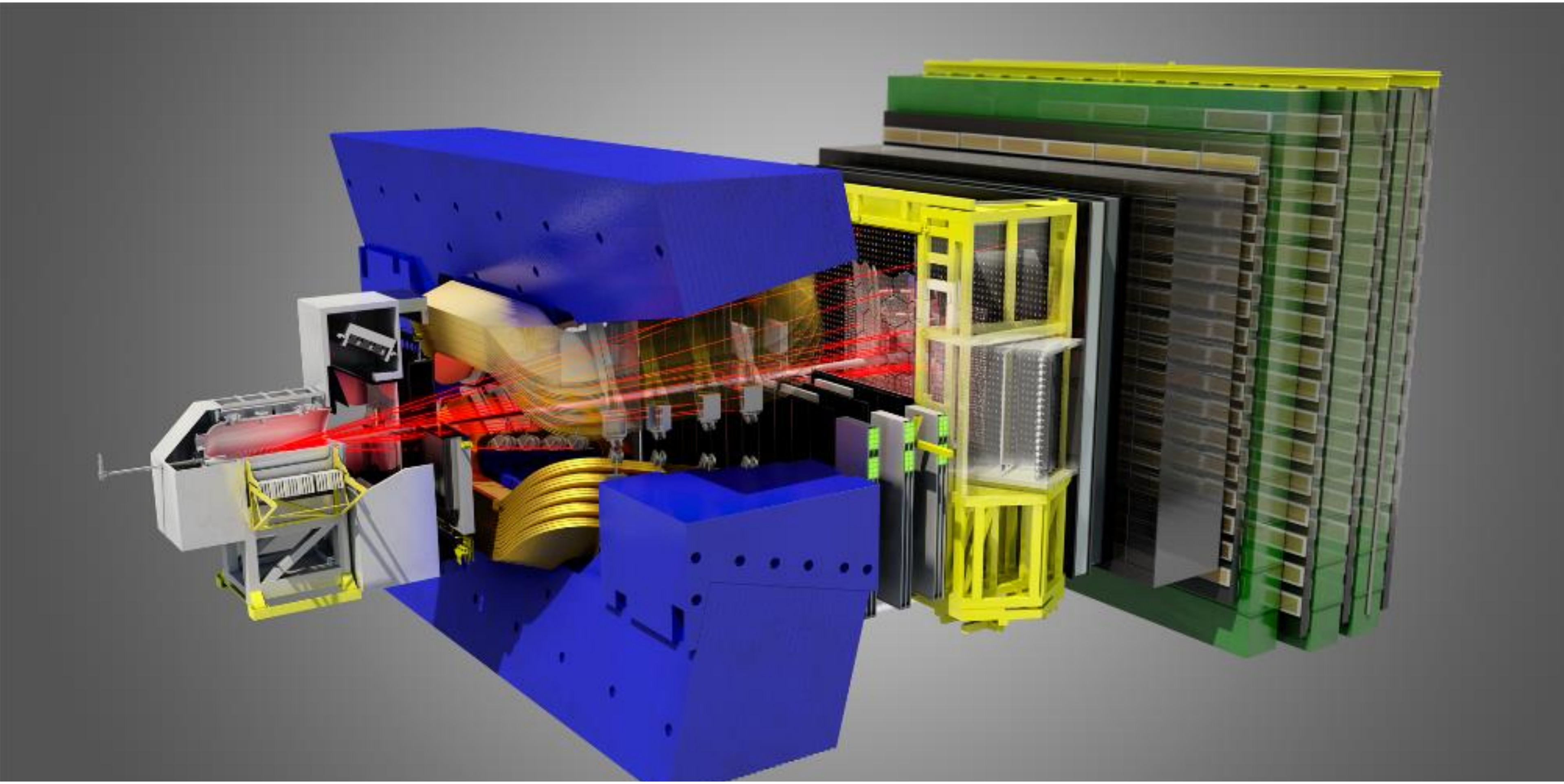
A review to decide the most cost-effective solution will take place in April.

Online mile-stones

- ✓ Q1 2016: long-distance fibres validation (successful)
- ✓ Q2 2016: event-building with 32-port switch (successful)
  - Thanks to the Bologna, CNAF and Ferrara teams

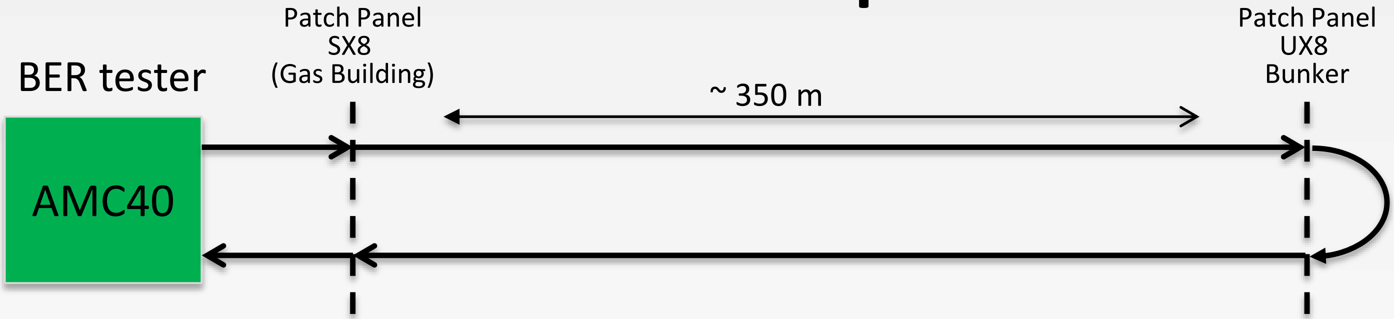
  → would be good to repeat this in a lab-setup (better tuning etc…)
- □ Q2 2016: event-builder scalability to 600 nodes

  ✓Achieved with 128 nodes, larger scale test still this year
  need to get a slot from external HPC sites (France, US…)
- □ Q2 2017: data-centre solution defined
- □ Q4 2017: full DAQ network test
- □ Q3 2018: network technology decision
- □ Q3 2019: ready for commissioning



Thank you for your time

Setup



- Review of TDR suggested a test installation at P8 to ensure the viability of the long distance read-out
- Installed 3 Trunk cables with 144 fibers each from different providers at P8
  - 1 x 144 Fibers from Fibernet, pre-connectorized
  - 1 x 144 Fibers from CERN, pre-connectorized
  - 1 x 144 Fibers from CERN, spliced
- All 432 Fibers have been tested and have expected Attenuation
- Currently Measuring the spliced fibers
  - Worst case scenario due to additional attenuation of splice point
- Testing on loop with 2 x 350m
  - Higher attenuation than in final setup
  - AMC40 transmitter has more optical power though
  - Longer fiber just about compensates for the stronger transmitter

Result so far

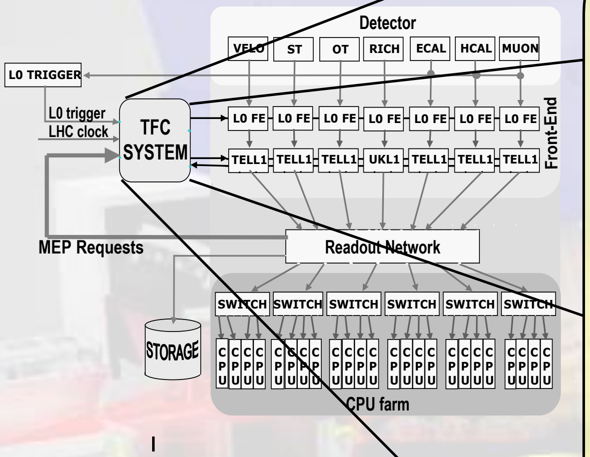
- Setup has been running for over 9 months now
- 10^{18} bits tested so far
- 0 Errors
- BER < 10^{-18} with 95% confidence

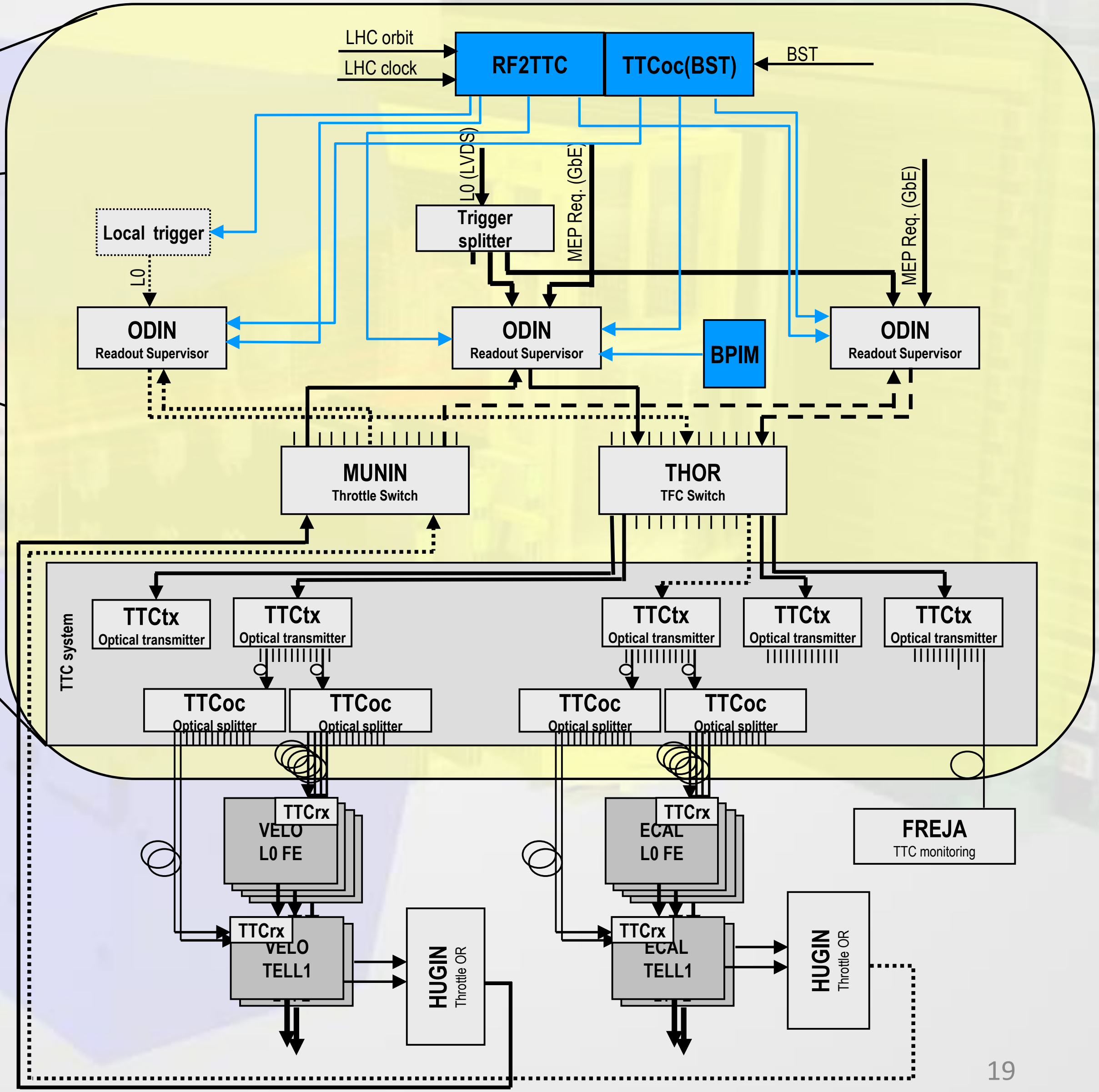
$$CL = 1 - e^{-N \times BER_s} \times \sum_{k=0}^{E} \frac{(N \times BER_S)^k}{k!}$$

CL = Confidence Level
N = Number of bits tested
E = Errors
BER_s = Specific Error Rate

Confidence	63%	74%	95%
Aggregated BER	< 3.1×10^{-18}	< 4×10^{-18}	< 9×10^{-18}
Error rate for 12k fibers	< 16 / day	< 20 / day	< 45 / day

Timing distribution at LHCb

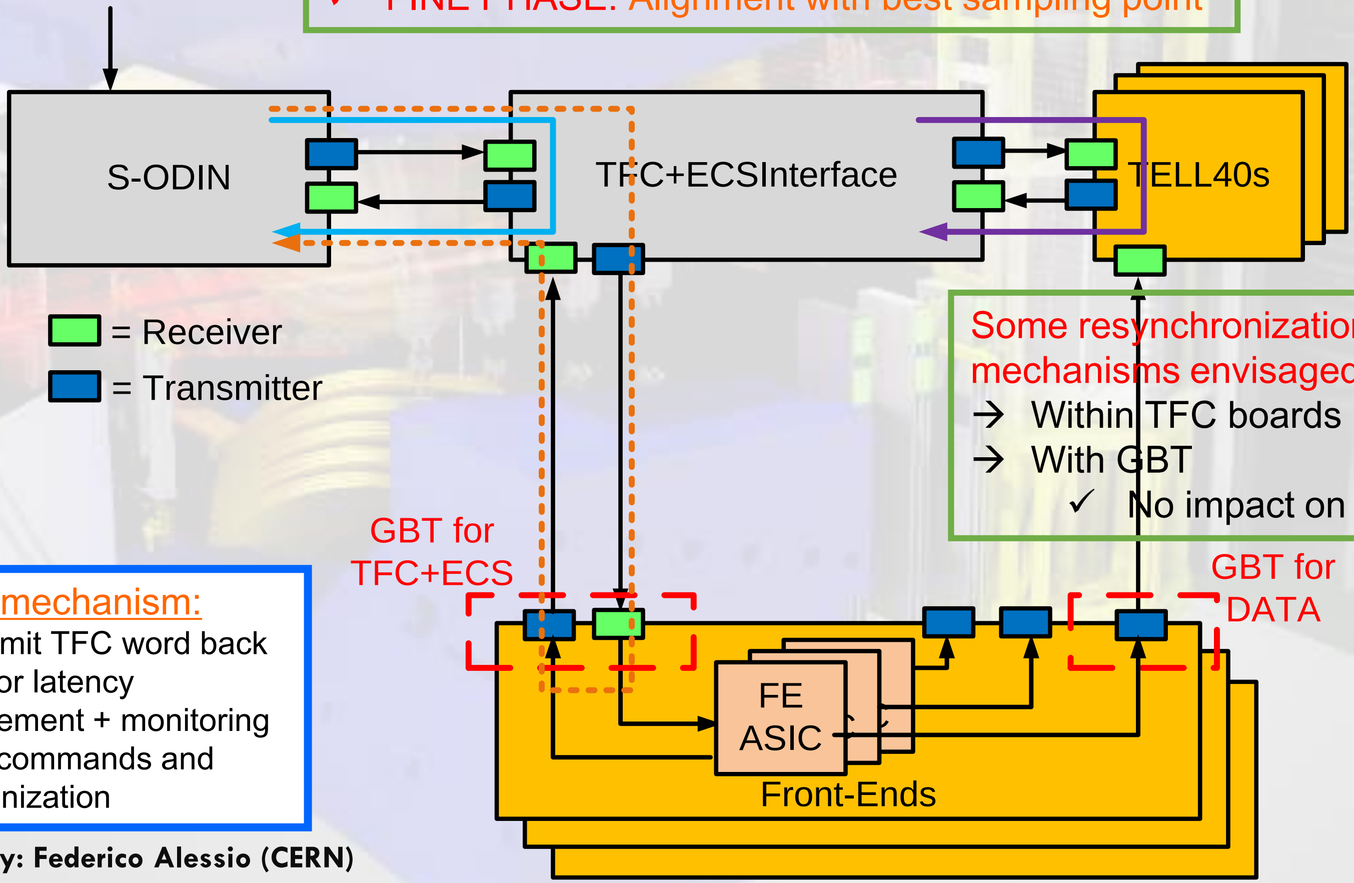




- Unidirectional
- 1:32 max split ratio
- Low bandwidth
- Obsolete components
- Very low payload for Bunch synchronous data

Upgrade proposal 1:TTC over GBT

From TFC point of view, we ensure constant:
- ✓ LATENCY: Alignment with BXID
- ✓ FINE PHASE: Alignment with best sampling point



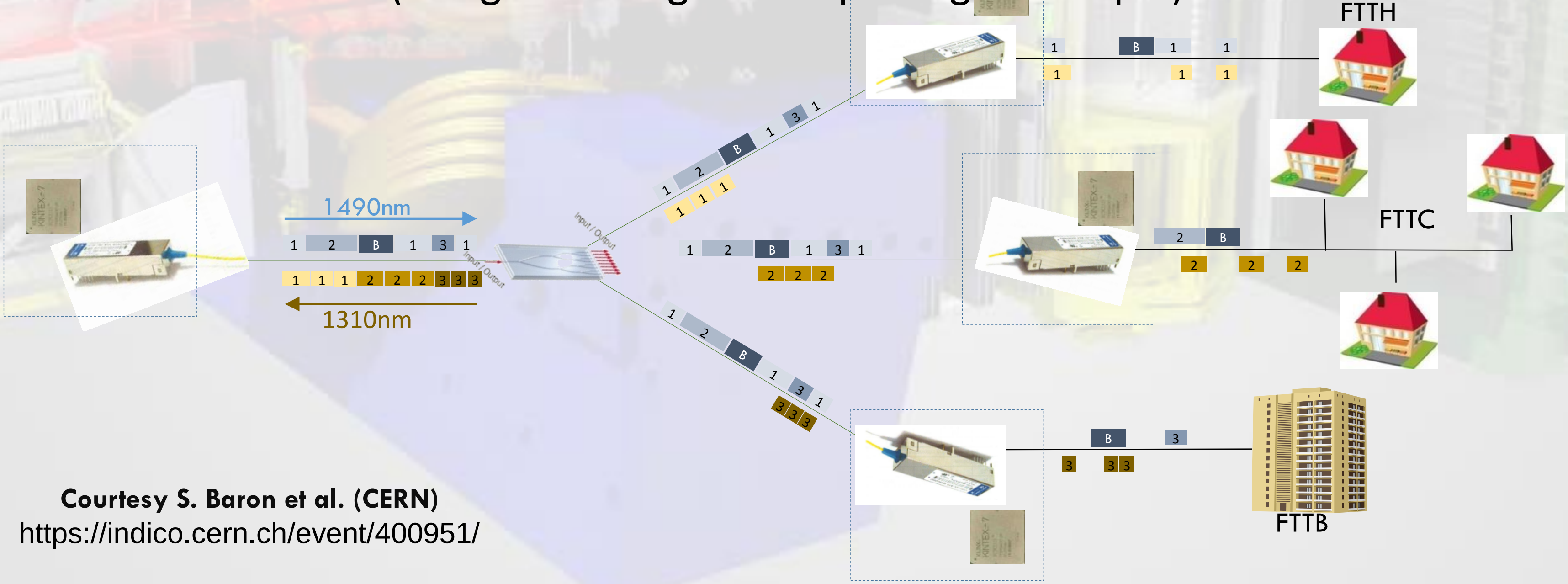
Some resynchronization mechanisms envisaged:
- → Within TFC boards
- → With GBT
  - ✓ No impact on FE itself

Loopback mechanism:
- → re-transmit TFC word back
- → allows for latency measurement + monitoring of TFC commands and synchronization

Courtesy: Federico Alessio (CERN)

Upgrade proposal 2:TTC over PON

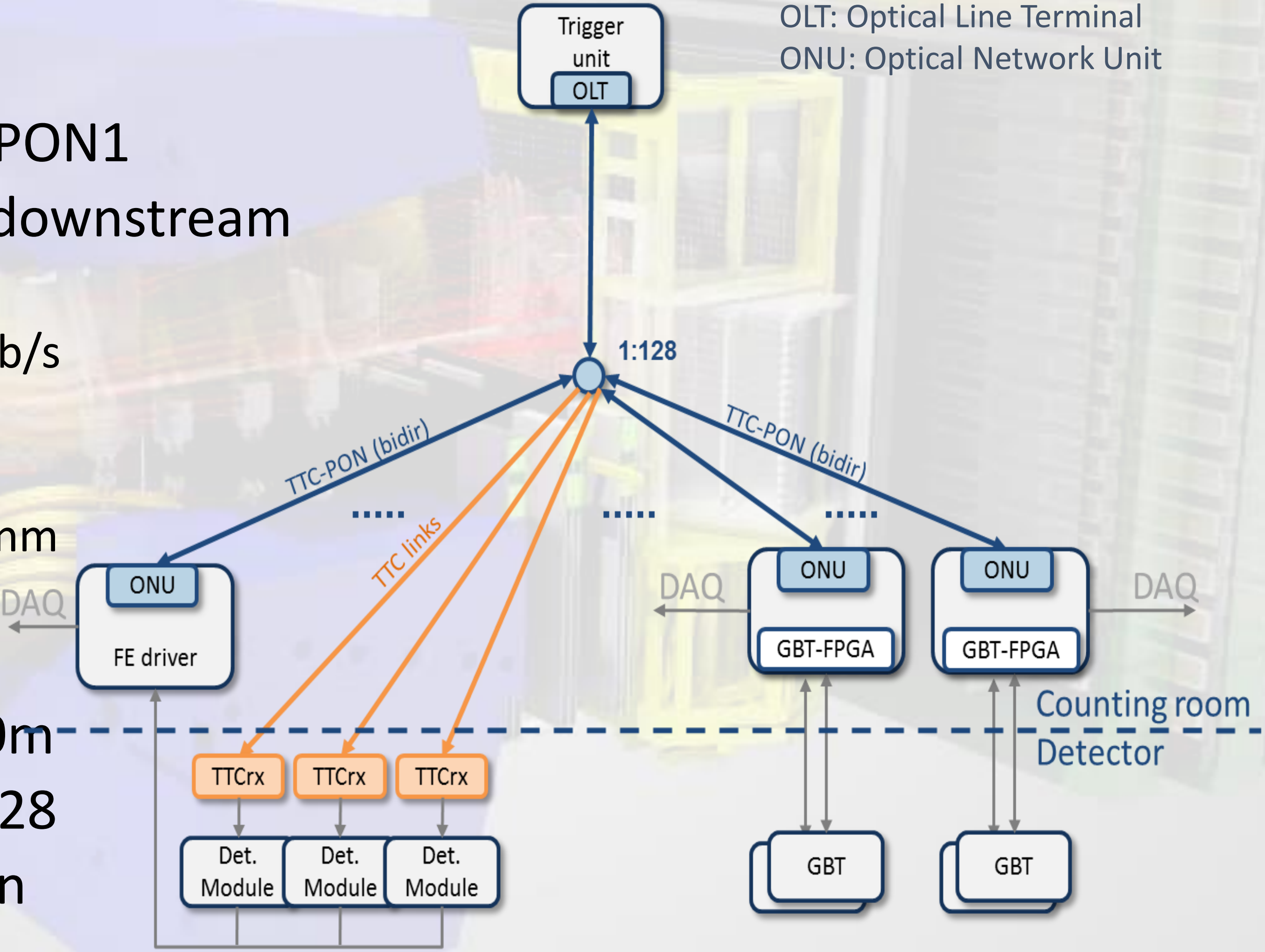
- Major topology of growing Access Network Market known as:
    - Fiber To The X (FTTx) (http://en.wikipedia.org/wiki/Fiber_to_the_x)
- Point-to-MultiPoint (P2M)
- One single fibre in charge of both downstream and upstream transmissions (using wavelength multiplexing technique)



Courtesy S. Baron et al. (CERN)
https://indico.cern.ch/event/400951/

TTC-PON

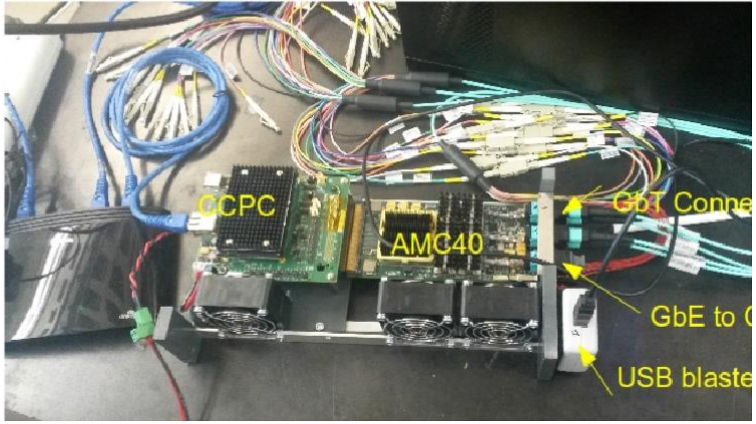
- Standard: XG-PON1
- Fixed latency downstream
- Line rate
  - Down: 9.6 Gb/s
  - Up: 2.4 Gb/s
- Wavelength:
  - Down: 1577nm
  - Up: 1270nm
- Bidirectional
- Range: 1-1000m
- Split ratio: 1:128
- Error detection



OLT: Optical Line Terminal
ONU: Optical Network Unit

AMC40 readout board

- Philosophy: one common board for DAQ and timing/control
- 1st generation readout board
  - Developed at CPP Marseille
- Mezzanine for ATCA crate
- Altera Stratix V FPGA
- ≤36 bidirectional optical links
  - Avago MiniPODs
  - GBT / GWT / 10GbE protocols
- Produced and made available to the collaboration (MiniDAQ1)
- Firmware designed for migration to 2nd generation readout board



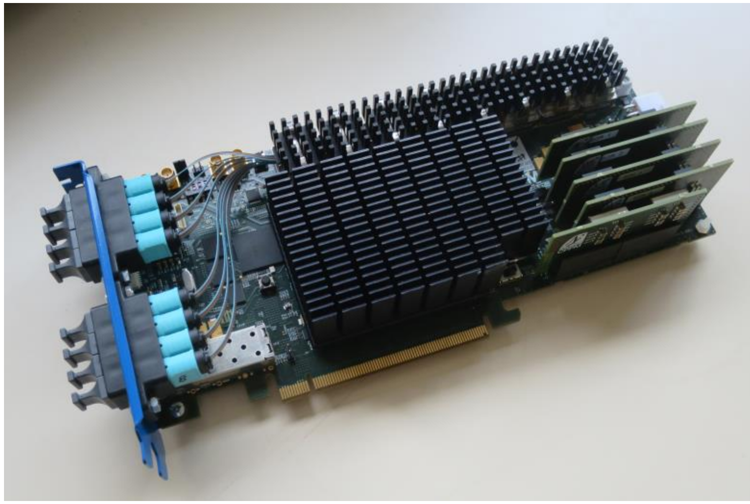
PCIe40 status

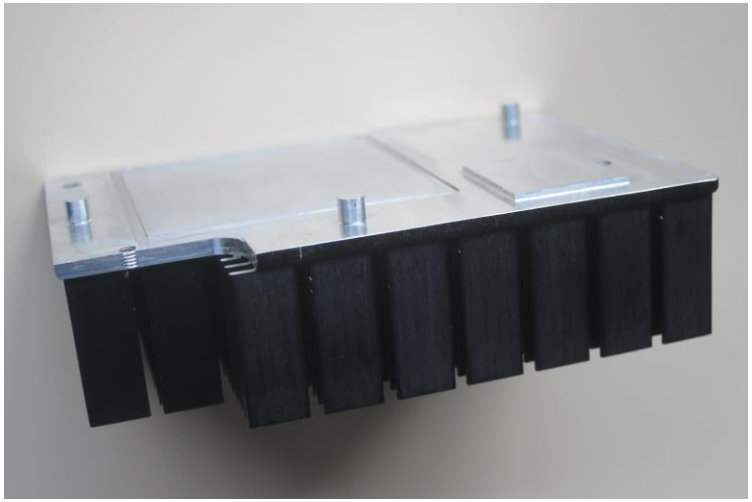
Current status

- Low-Level Interface available
  - Slow control through PCIe
  - 100Gbps DMA
  - On-board communication (I2C, SPI)
  - Filtered PLL clock
  - CvP (programming via PCIe)
  - Temperature and current monitoring
- New custom heat-sink design
- Procurement ongoing after tender

Next plans

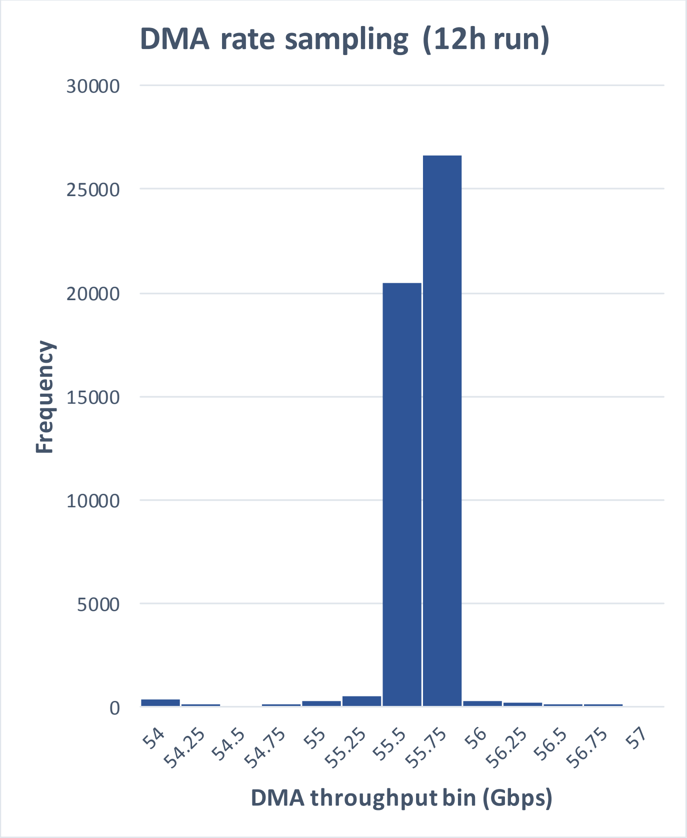
- x24 GBT integration
- PON tests for timing distribution
- Optical links characterization
  - Eye aperture
  - BER
- Cooling characterization
  - *f*(firmware, airflow, heat-sink)
- Monitoring via IPMI



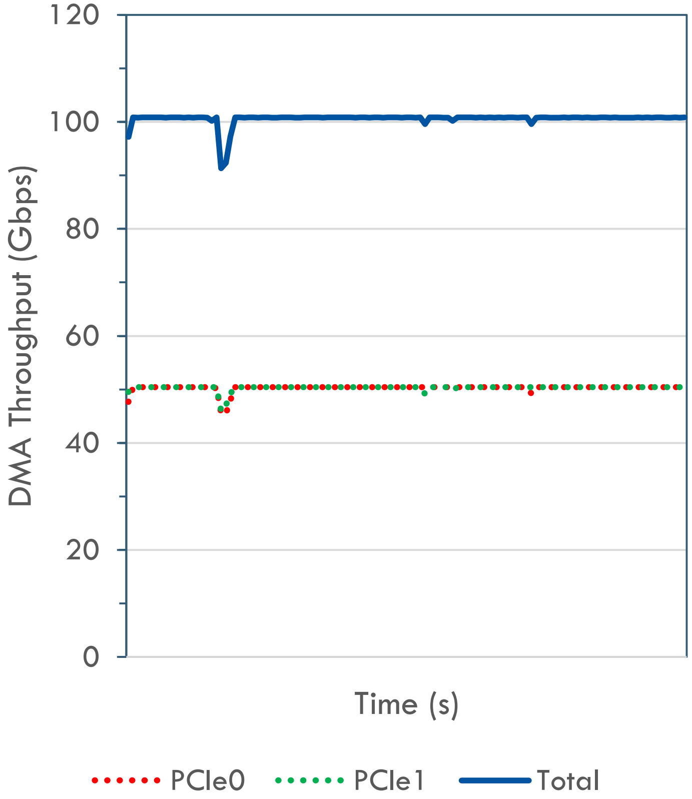


PCIe Gen3 DMA performance

1 interface (optimized)



2 interfaces (unoptimized)



Firmware status

AMC40

- ✓Readout over GBT / WideBus
- ✓Fixed-size frontend protocol
- ✓6 optical links
- ✓Front-end configuration via GBT-SCA
- ✓Data monitoring inside FPGA
- ✓UDP output protocol over 10GBASE-R
- ✓Embedded FE data generator
- ✓Data verification outside FPGA
- ✓Configurable data rate throttle
- □Variable-size FE protocol

PCIe40

… all of the previous, plus:

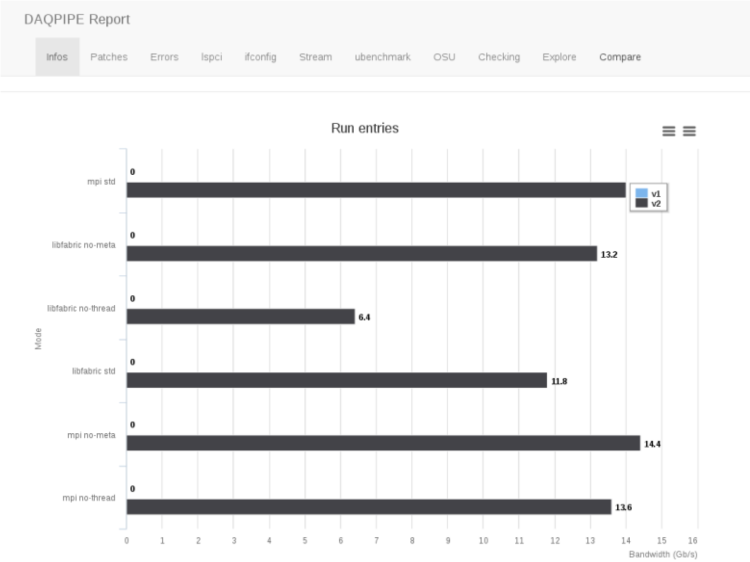
- ✓HW-specific low level interface
- ✓PCI Express Gen3 DMA interface & driver

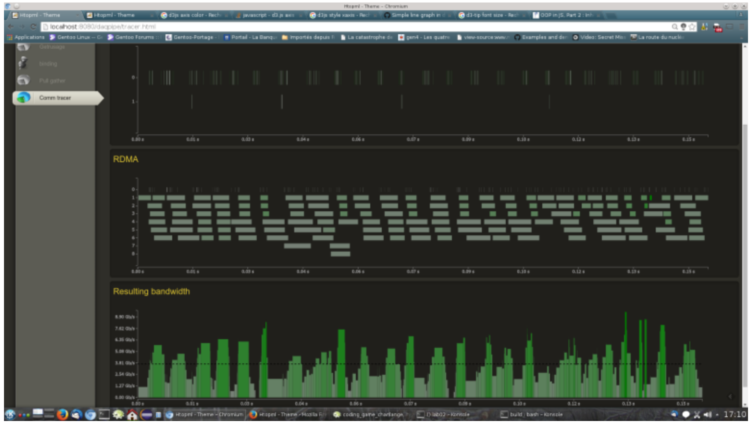
- □Increase number of optical links
- □Integrate LLI & DMA into firmware
- □Implement remaining frontend protocols

Event-building software performance

- DAQPIPE now supports following APIs:
- - **Libfabric**
- - **MPI**
- - **TCP**
- Implemented automatic benchmarking and reporting
- (with web-based interface)
- On-the-fly monitoring (htopml)
- Benchmarks of
- - Intel **Omni-Path**
- - Infiniband **EDR**

  currently in progress!
- Already observing **86 Gb/s** on **EDR**


DAQPIPE Report
Infos
Patches
Errors
lspci
ifconfig
Stream
ubenchmark
OSU
Checking
Explore
Compare
Run entries
v1
v2
Mode
mpi std
libfabric no-meta
libfabric no-thread
libfabric std
mpi no-meta
mpi no-thread
Bandwidth (Gb/s)


Comm tracer
RDMA
Resulting bandwidth
17:10