Design, Operation and Future of the CMS DAQ system

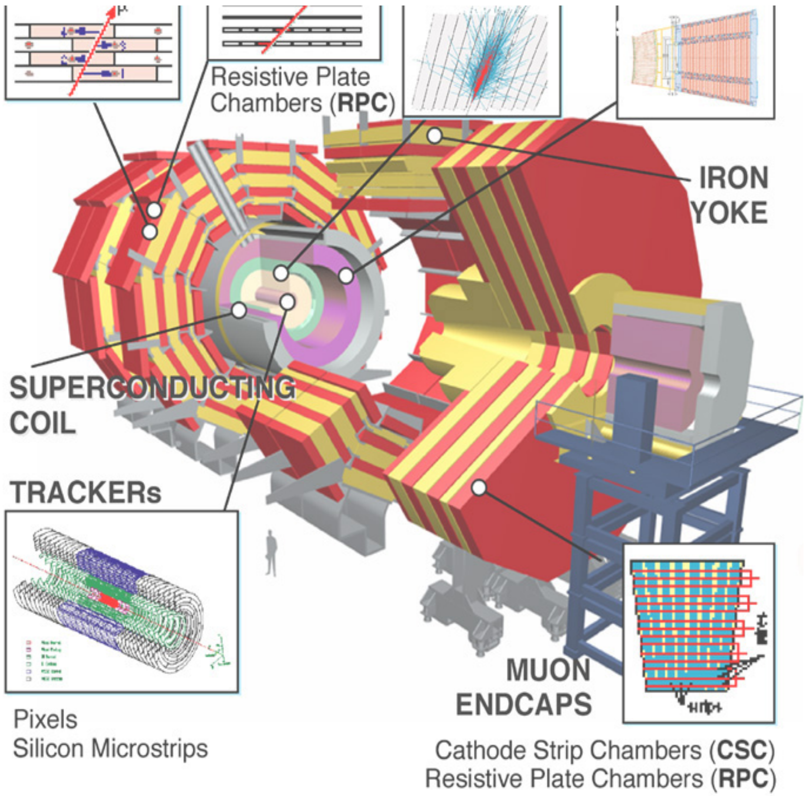
TIPP 2011 Saturday 11 June 2011

Frans Meijers (CERN-PH)

On behalf of the CMS DAQ group

CMS design parameters and DAQ requirements

Detectors

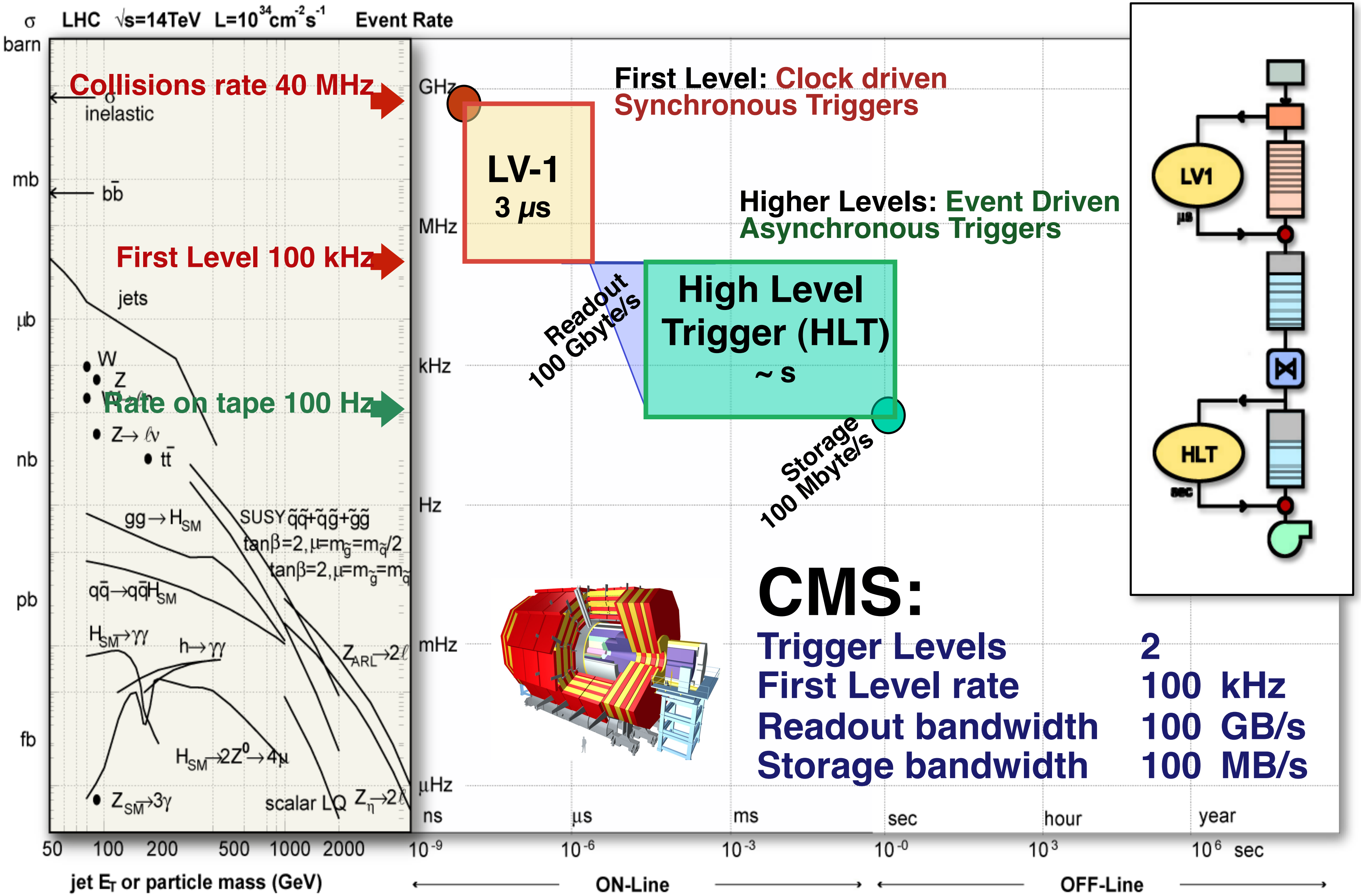


Total weight : 12,500 t	Overall length : 21.6 m
Overall diameter : 15 m	Magnetic field : 4 Tesla

Detector	Channels	Control	Ev. Data
Pixel	60000000	1 GB	50 (kB)
Tracker	10000000	1 GB	650
Preshower	145000	10 MB	50
ECAL	85000	10 MB	100
HCAL	14000	100 kB	50
Muon DT	200000	10 MB	10
Muon RPC	200000	10 MB	5
Muon CSC	400000	10 MB	90
Trigger		1 GB	16

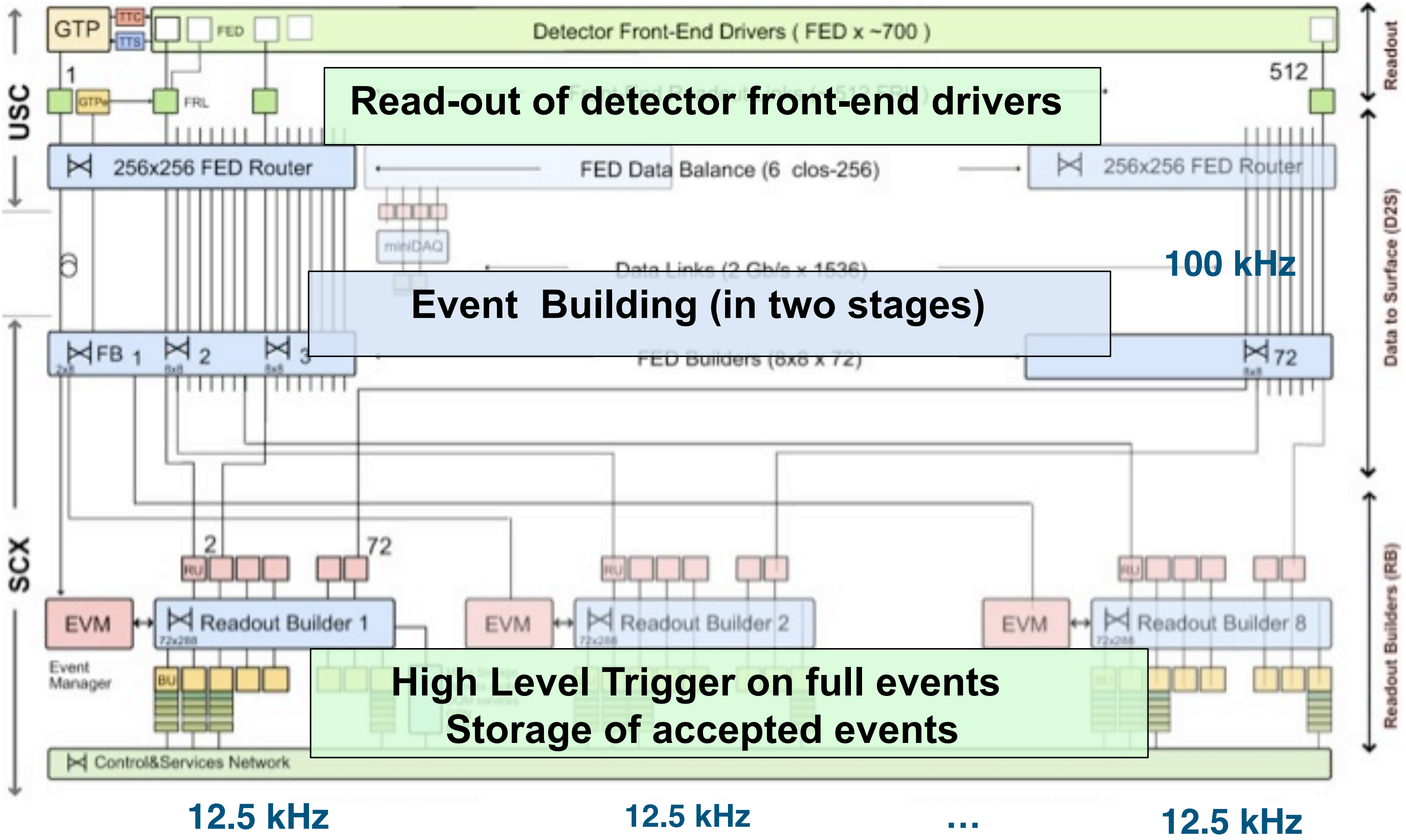
Average Event size	**1 Mbyte**
Max LV1 Trigger	**100 kHz**
Online rejection	**99.999%**
System dead time	**~ %**

Two Trigger levels



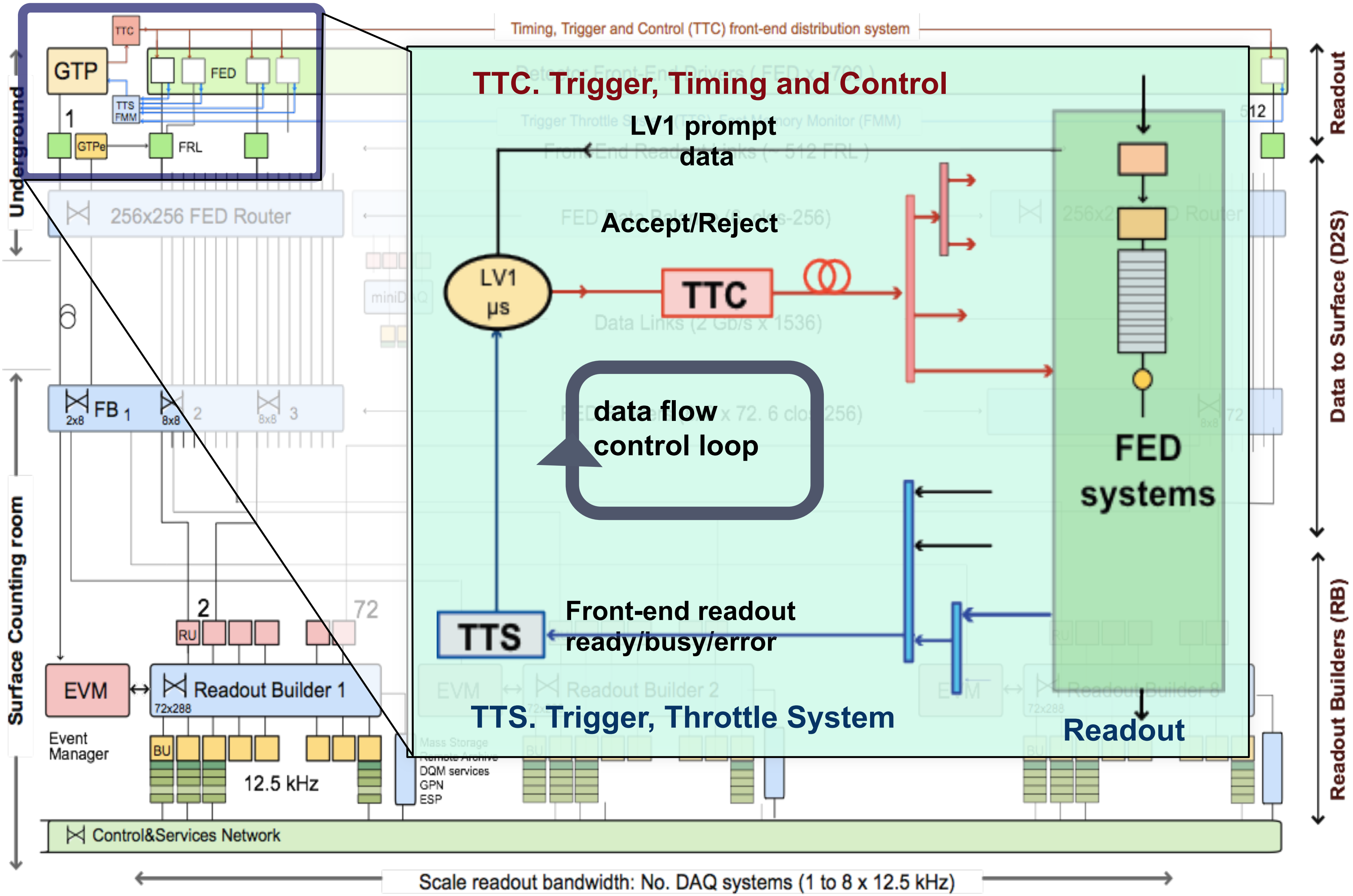
IMPLEMENTATION

CMS DAQ

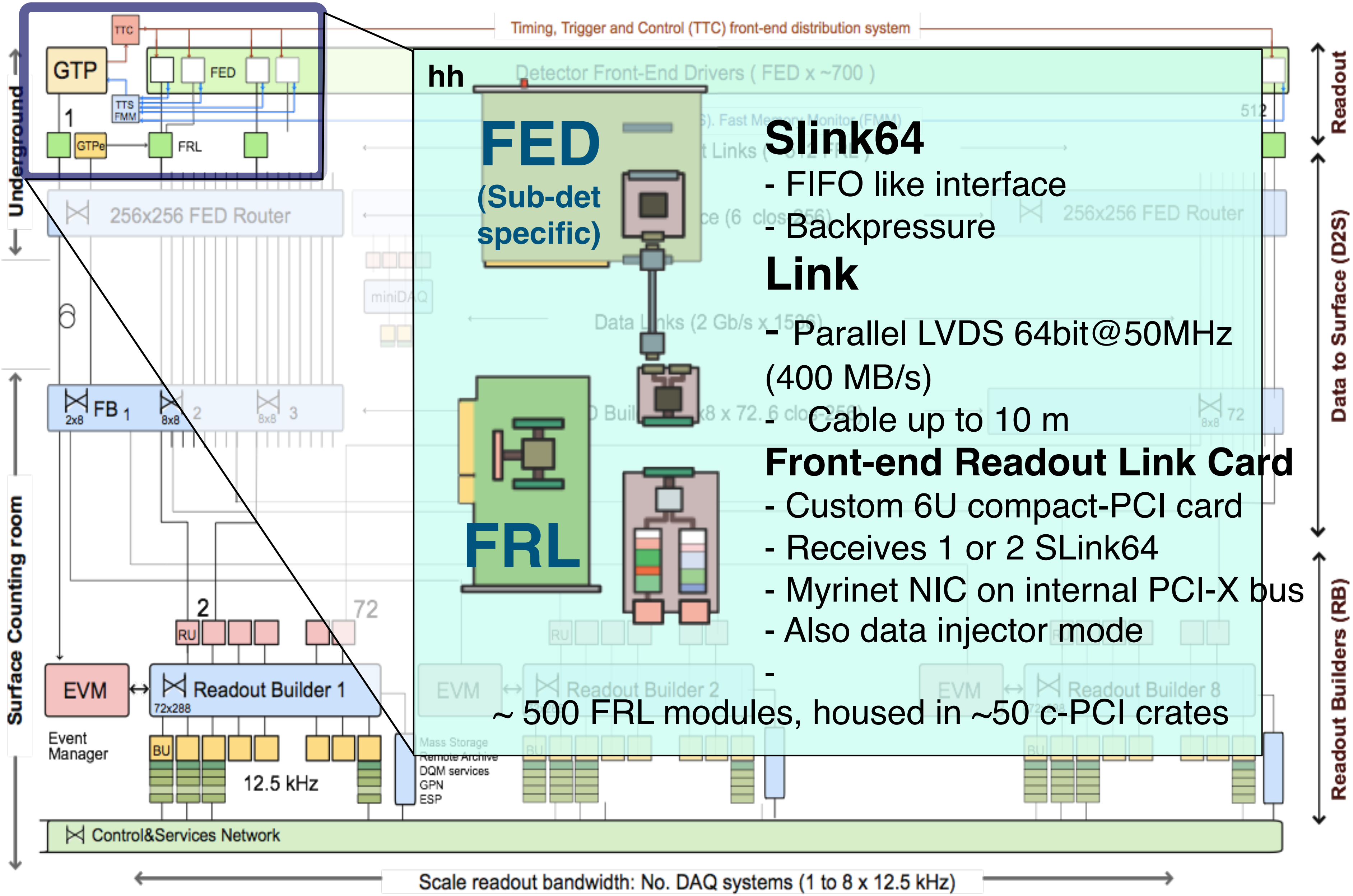


GTP
TTC
TTS
FED
Detector Front-End Drivers (FED x ~700)
1
512
GTPe
FRL
Read-out of detector front-end drivers
256x256 FED Router
FED Data Balance (6 clos-256)
256x256 FED Router
miniDAQ
100 kHz
Event Building (in two stages)
FB 1
2
3
72
2
72
EVM
Readout Builder 1
EVM
Readout Builder 2
EVM
Readout Builder 8
Event Manager
BU
High Level Trigger on full events
Storage of accepted events
Control&Services Network
USC
SCX
Readout
Data to Surface (D2S)
Readout Builders (RB)
12.5 kHz
12.5 kHz
…
12.5 kHz

Front-end model



Uniform interface - Readout



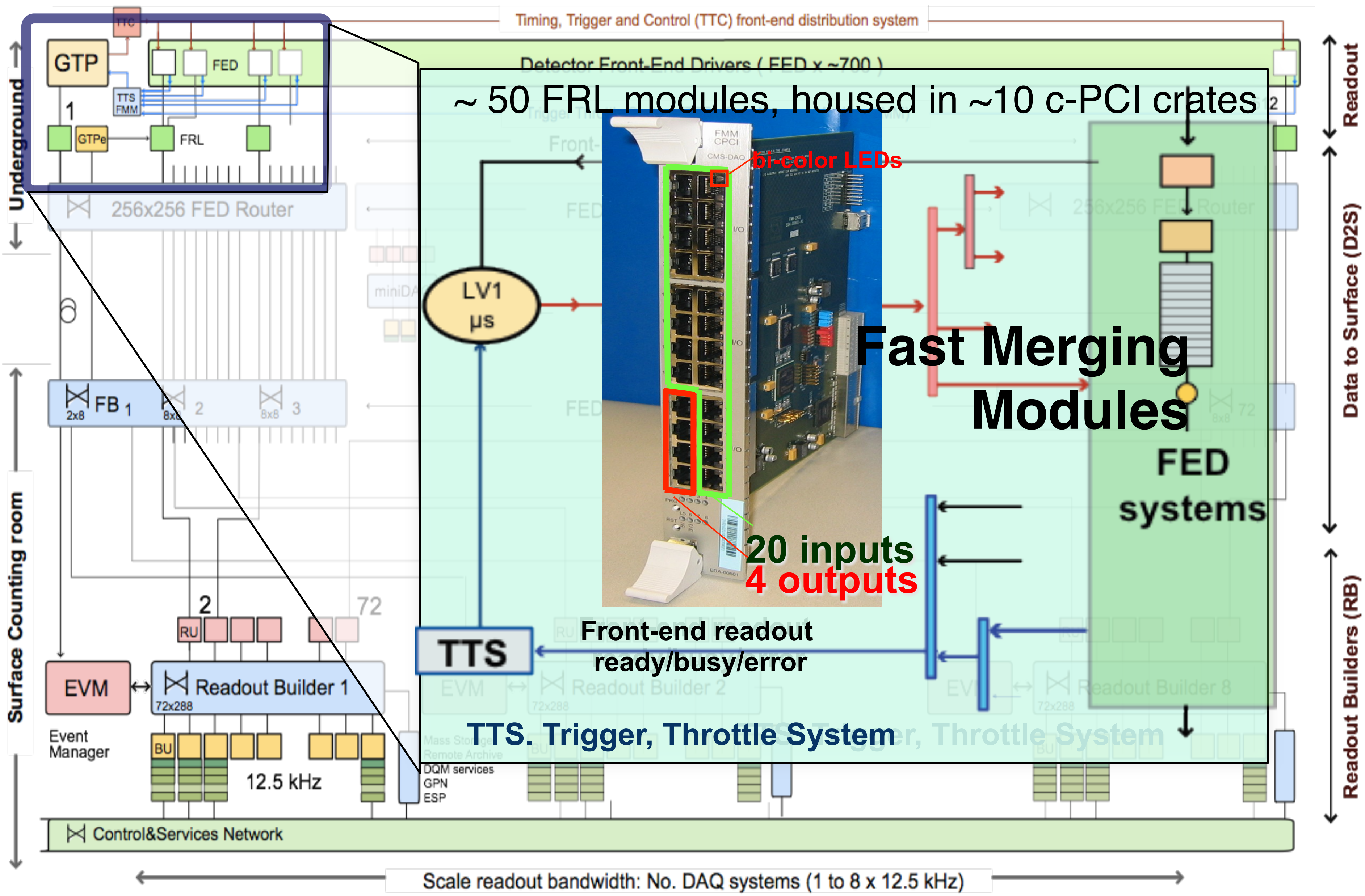
Slink64
- FIFO like interface
- Backpressure

Link
- Parallel LVDS 64bit@50MHz (400 MB/s)
- Cable up to 10 m

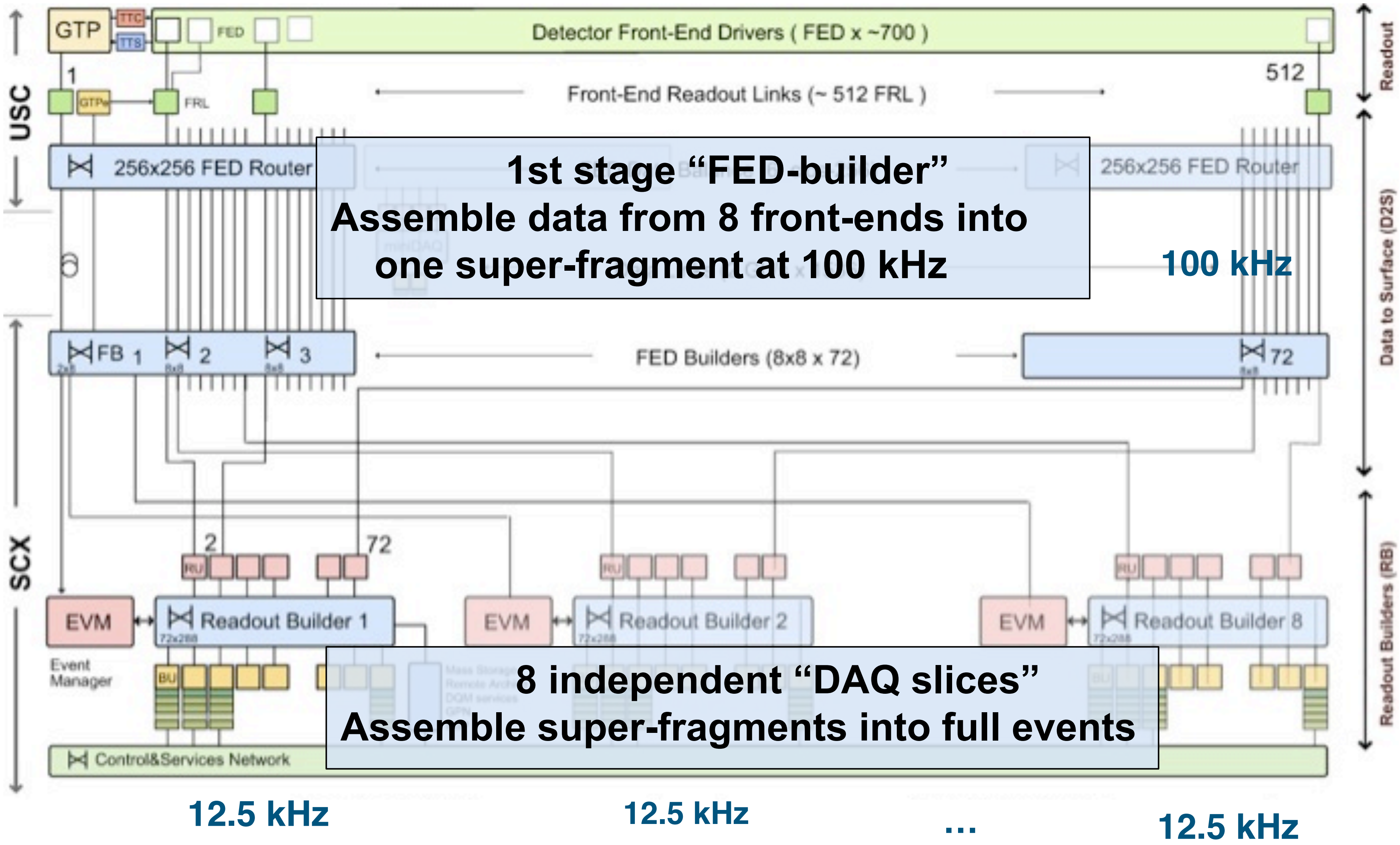
Front-end Readout Link Card
- Custom 6U compact-PCI card
- Receives 1 or 2 SLink64
- Myrinet NIC on internal PCI-X bus
- Also data injector mode
-

~ 500 FRL modules, housed in ~50 c-PCI crates

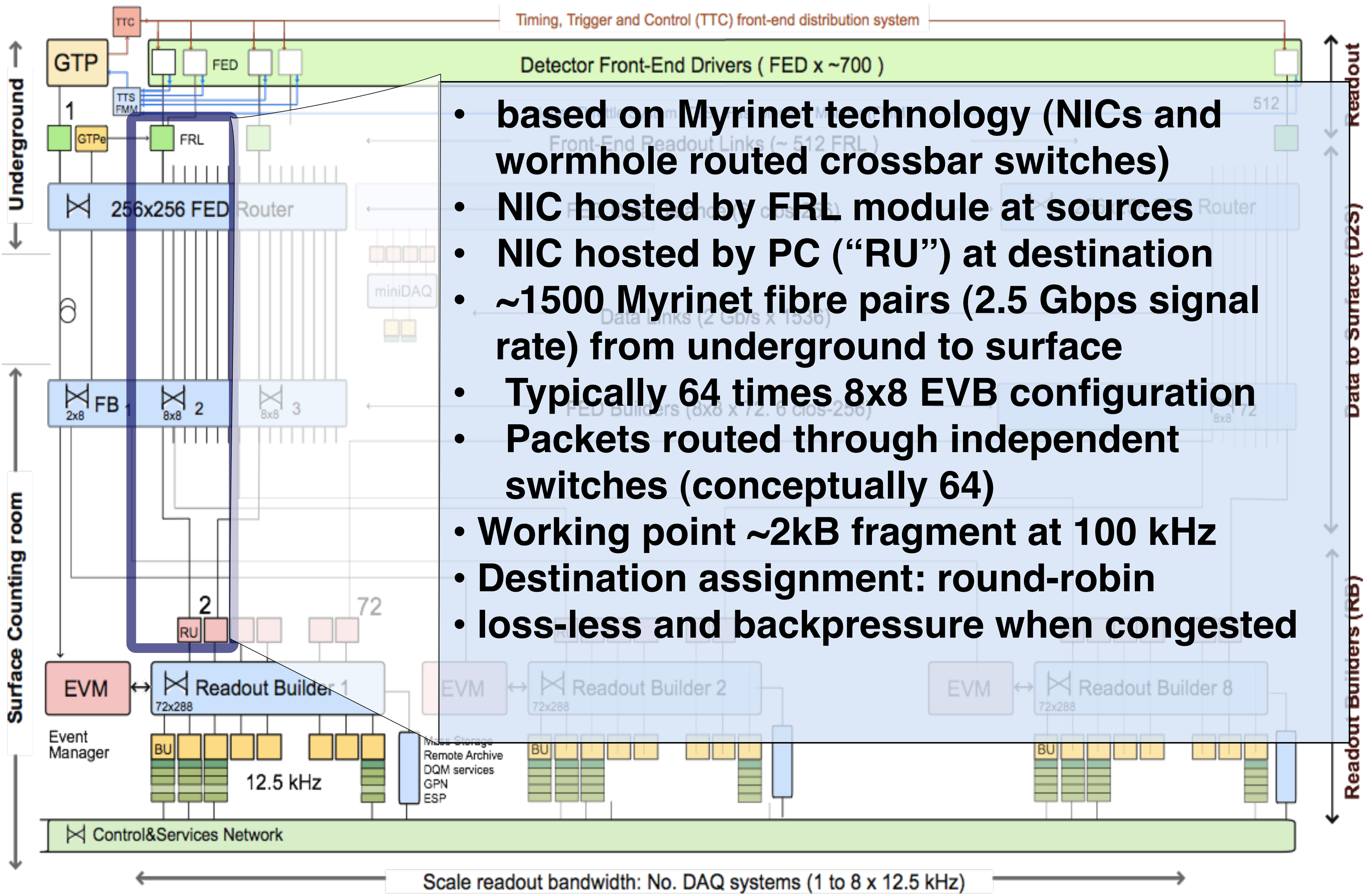
Uniform Interface – TTS



2-Stage Event Builder

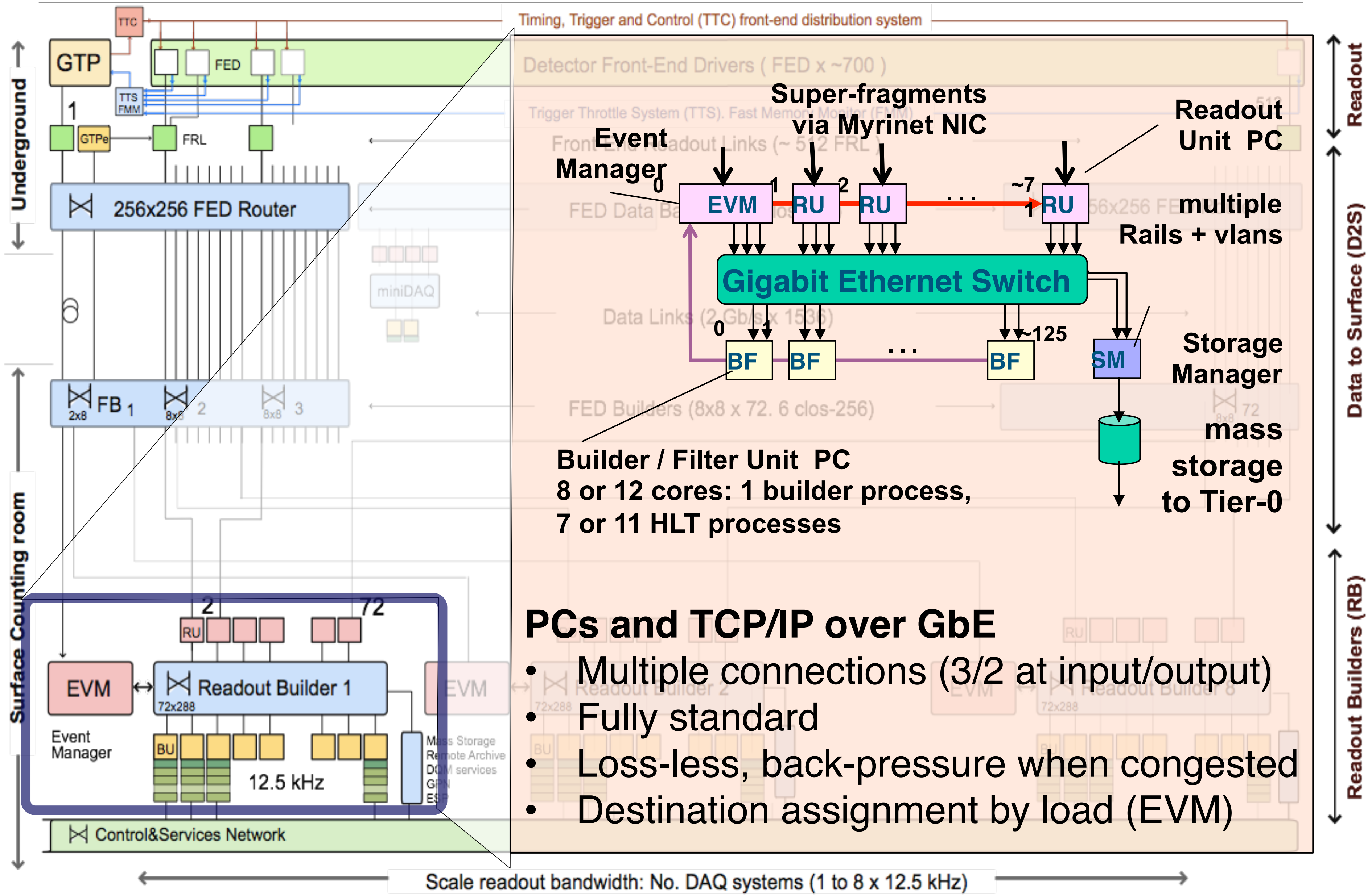


Super-Fragment Builder (1st stage)



- based on Myrinet technology (NICs and wormhole routed crossbar switches)
- NIC hosted by FRL module at sources
- NIC hosted by PC (“RU”) at destination
- ~1500 Myrinet fibre pairs (2.5 Gbps signal rate) from underground to surface
- Typically 64 times 8x8 EVB configuration
- Packets routed through independent switches (conceptually 64)
- Working point ~2kB fragment at 100 kHz
- Destination assignment: round-robin
- loss-less and backpressure when congested

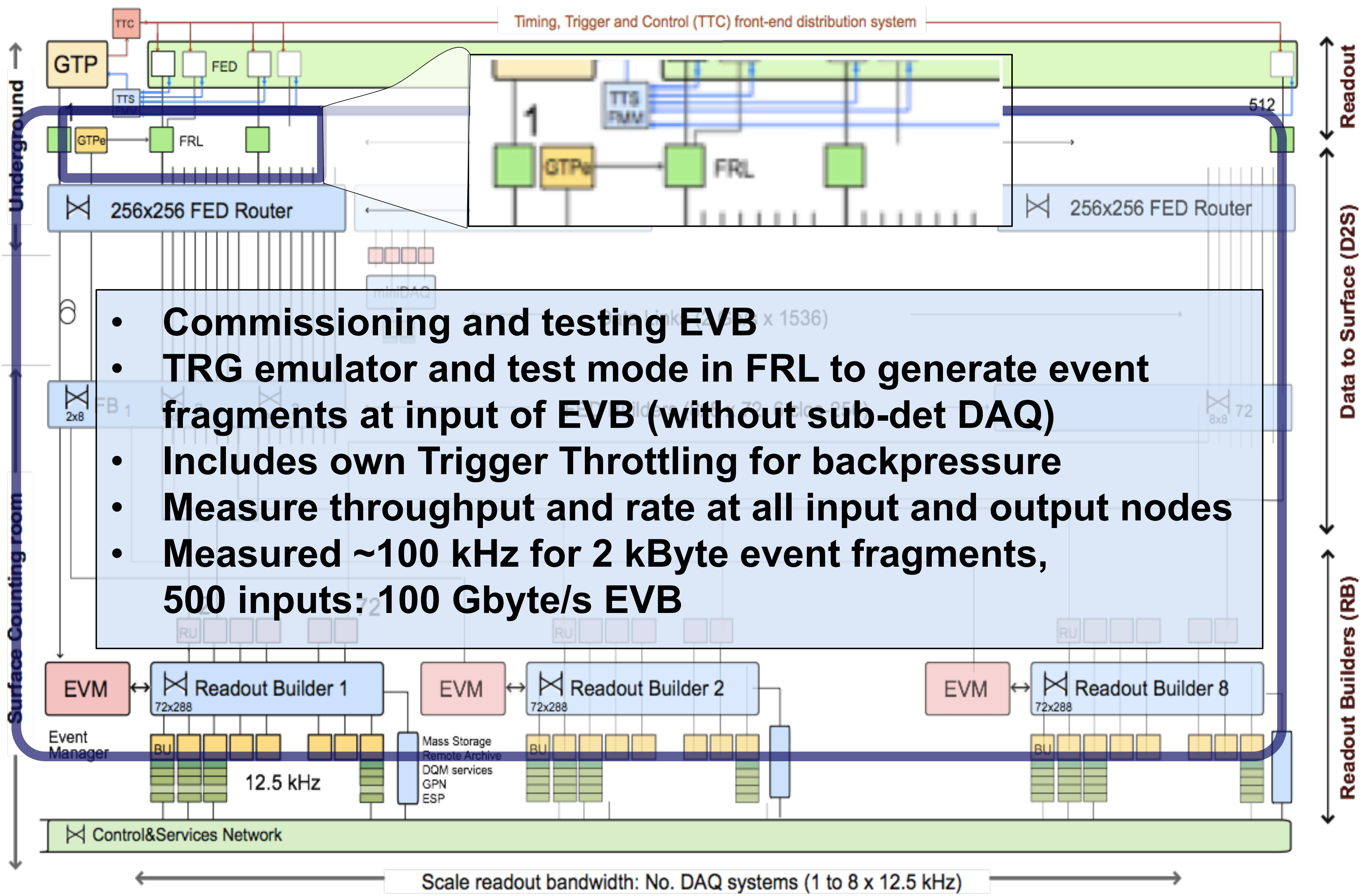
DAQ slice builder (2nd stage)



PCs and TCP/IP over GbE

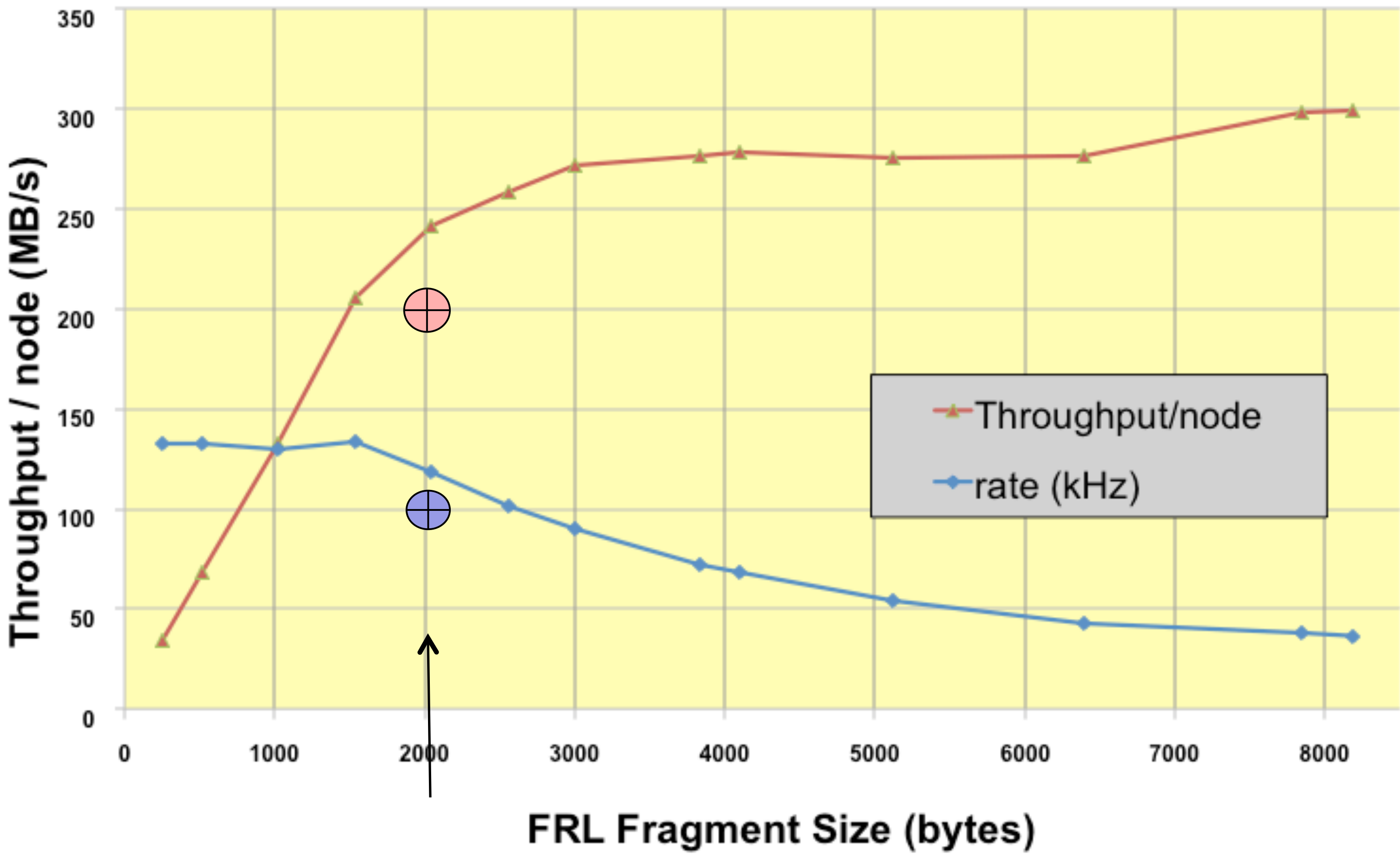
- Multiple connections (3/2 at input/output)
- Fully standard
- Loss-less, back-pressure when congested
- Destination assignment by load (EVM)

Full-EVB and emulator mode



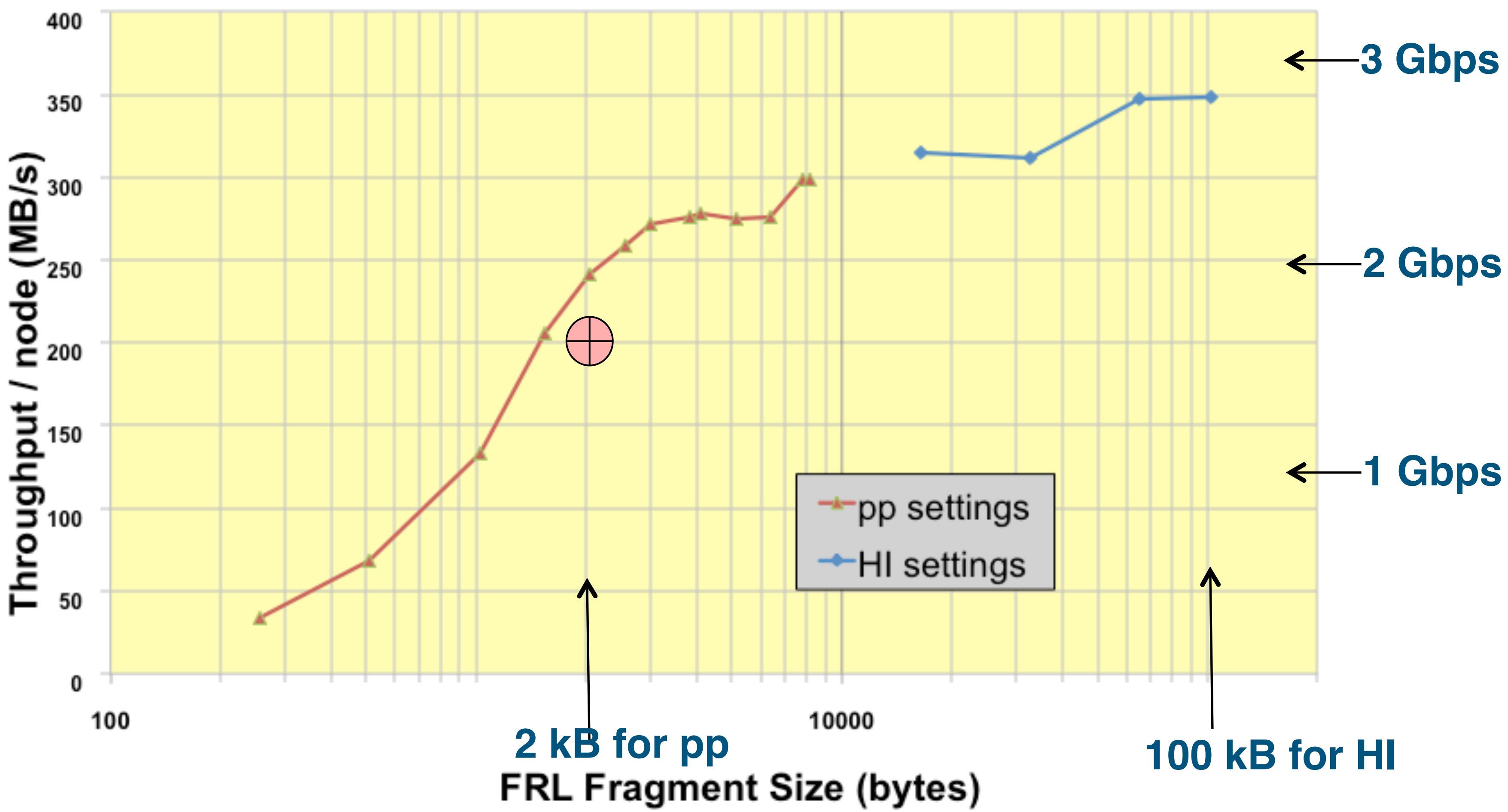
- Commissioning and testing EVB
- TRG emulator and test mode in FRL to generate event fragments at input of EVB (without sub-det DAQ)
- Includes own Trigger Throttling for backpressure
- Measure throughput and rate at all input and output nodes
- Measured ~100 kHz for 2 kByte event fragments, 500 inputs: 100 Gbyte/s EVB

Full EVB performance



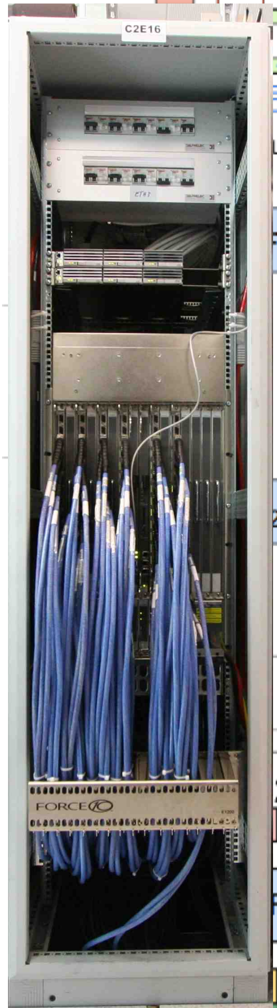
Working point is 2 kByte fragment
Rate ~125 kHz, Throughput ~250 MByte/s on each node
Aggregate EVB Throughput 125 GByte/s

Full EVB performance / pp and HI



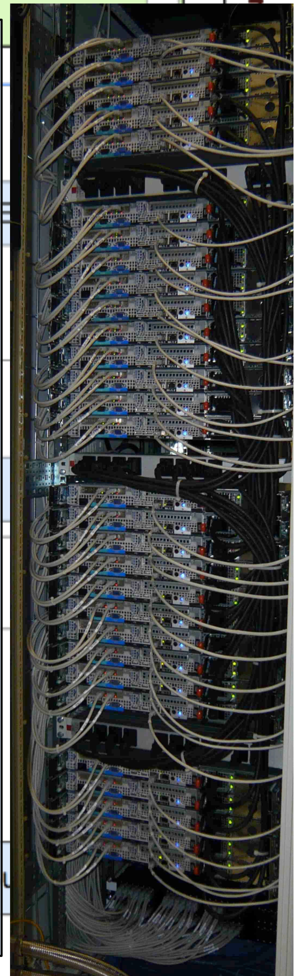
- Large fragment sizes: reach the 3x1Gbps of the RU-PC ethernet output
- Aggregate EVB throughput 175 Gbyte/s

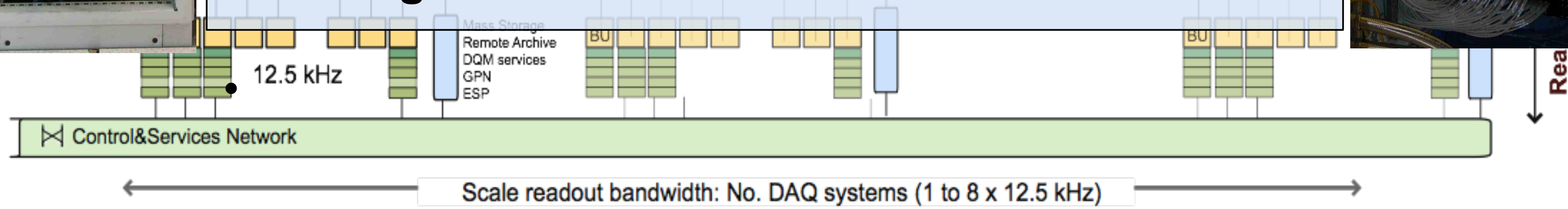
EVB – HLT installation



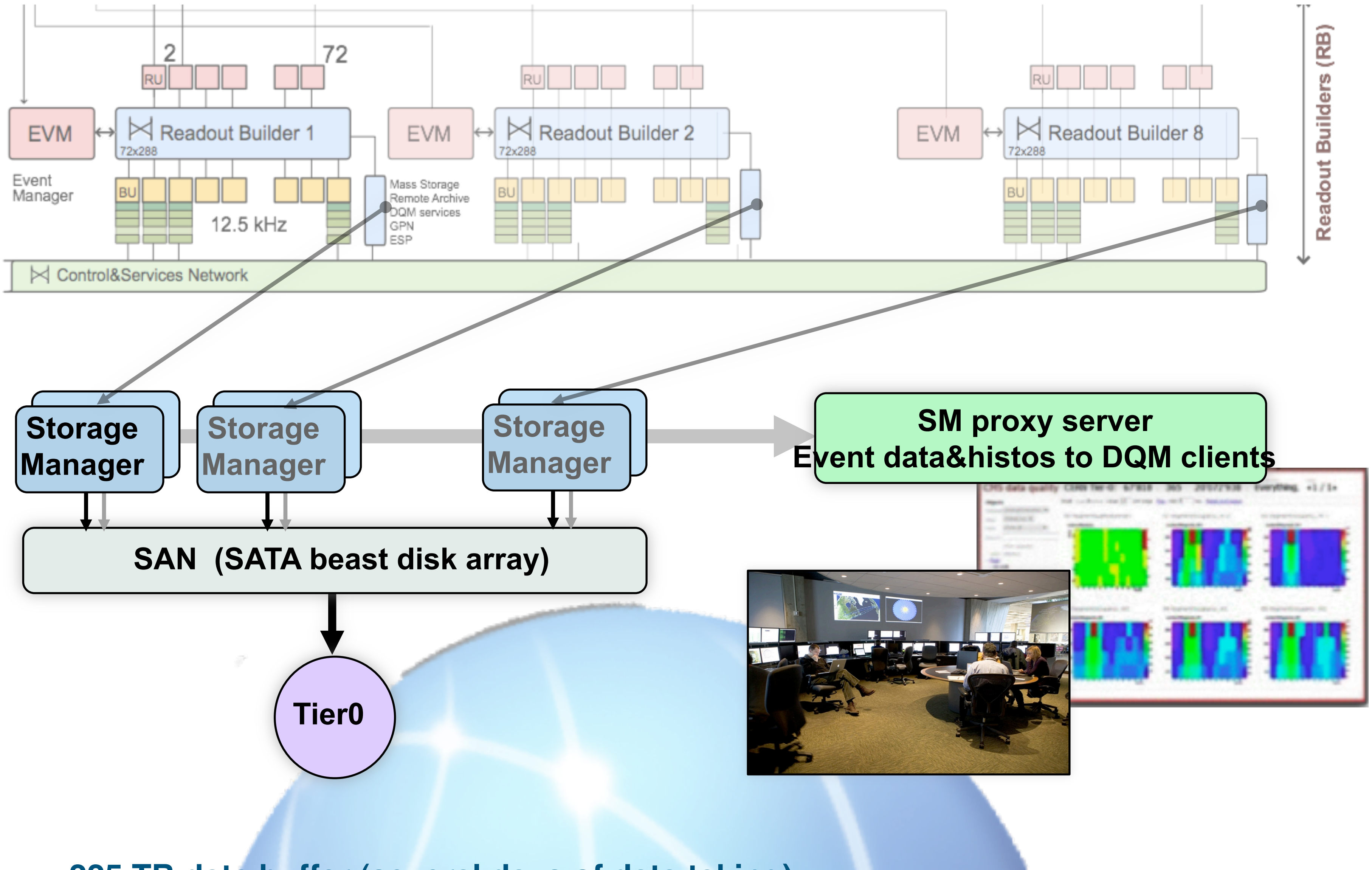
- **EVB – input "RU" PC nodes**
  - 640 times dual 2-core E5130 (2007)
  - Each node has 3 links to GbE switch
- **Switches**
  - 8 times F10 E1200 routers
  - In total ~4000 ports
- **EVB – output + HLT node ("BU-FU")**
  - 720 times dual 4-core E5430, 16 GB (2008)
  - 288 times dual 6-core X5650, 24 GB (2011)
  - Each node has 2 links to GbE switch

HLT Total: 1008 nodes, 9216 cores, 18 TB memory
@100 kHz: ~90 ms/event
Can be easily expanded by adding PC nodes and recabling EVB network



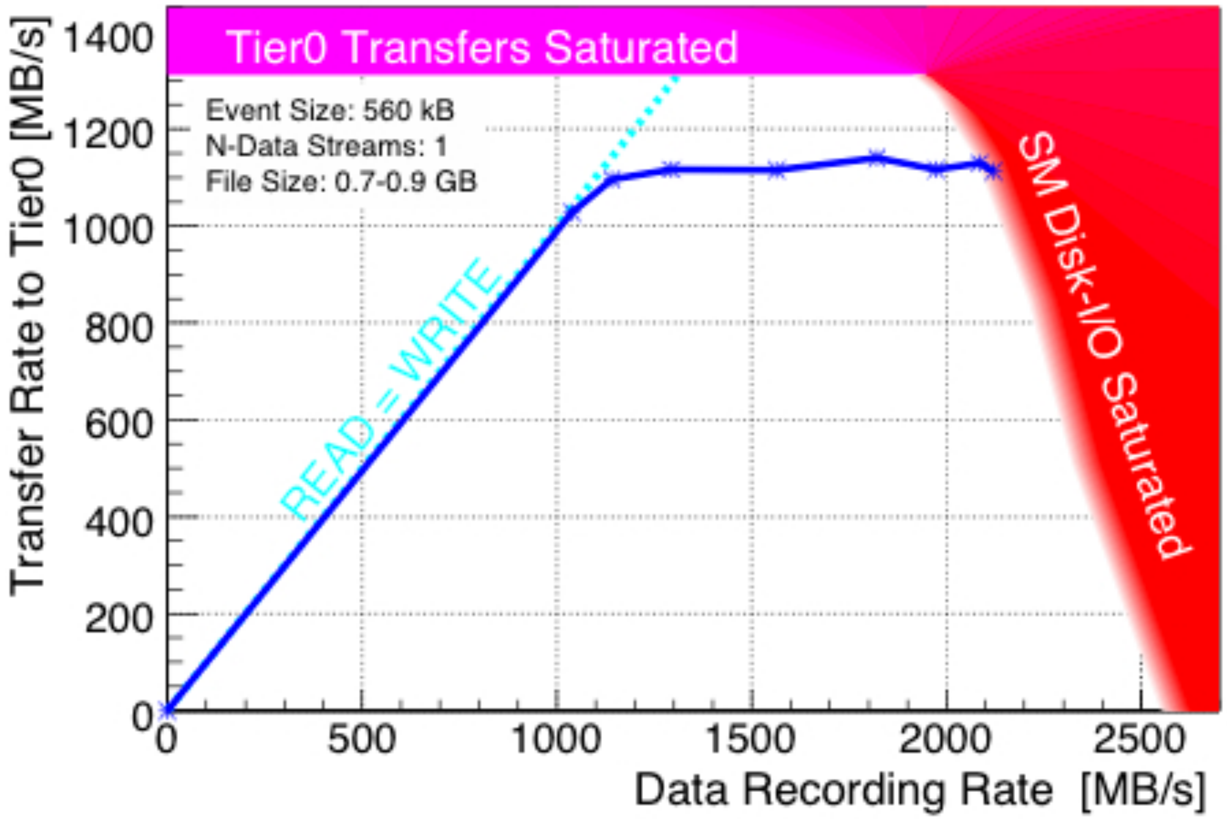


Storage, DQM, T0 transfer



~225 TB data buffer (several days of data taking)
DQM clients request event data (sampling) and histos via HTTP

Storage Manager Performance



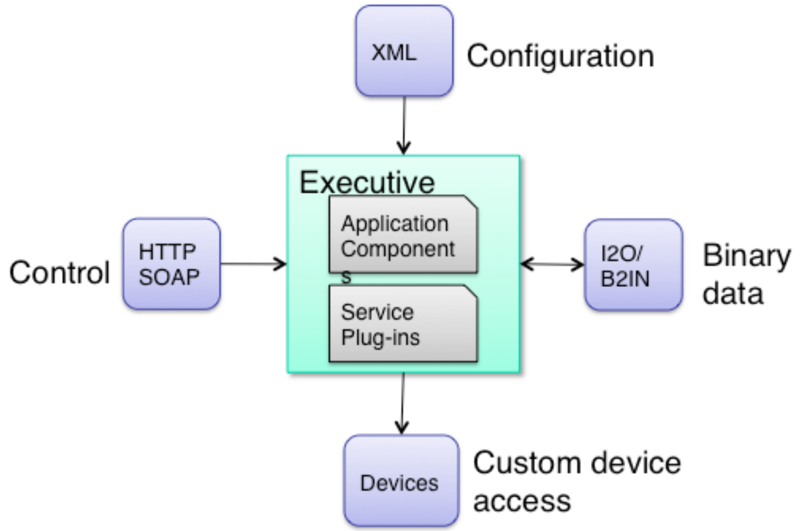
- Total capacity: 300 TB
- HLT compresses event data (root); reduction by factor ~2
- Event data to disk
  - pp; ~200 MB/s, design 600 MB/s
  - Heavy Ions: ~1.4 GB/s (up to 2.8 GB/s w/o transfer)

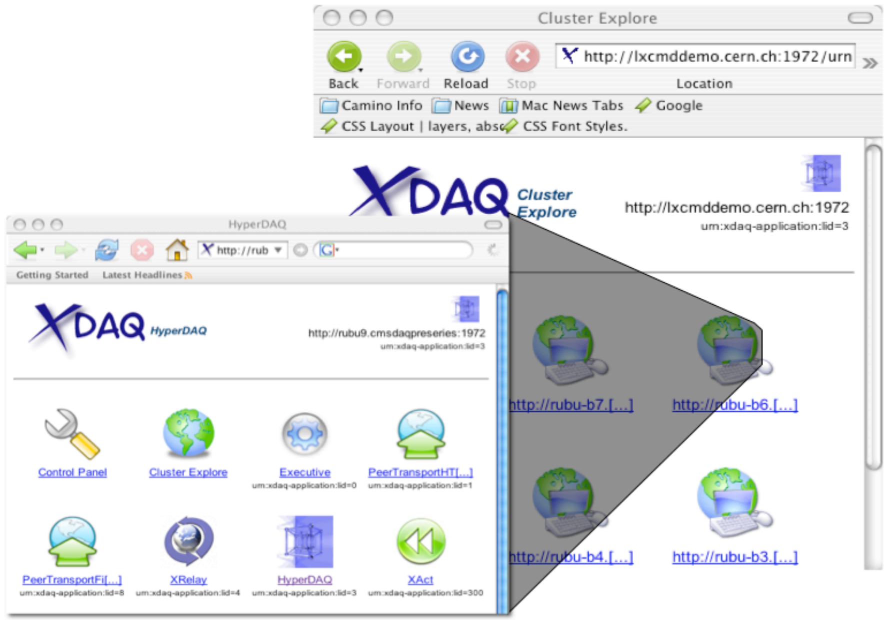
Online Software

- DAQ (sub-detector and “central” DAQ)
  - to transfer event data, built events, interface to custom hardware
  - to Control, configure and monitor the event flow
  - Layered approach with framework (XDAQ and run-control)
  - Applications using XDAQ, RC as foundation
- CMSSW offline software (C++)
  - The event reconstruction and selection used for High Level Trigger
- Detector Control System (“slow control”)
  - PVSS based + JCOP
- IT infrastructure
  - ~2500 linux nodes, ~100 Windows nodes, network

XDAQ framework and components

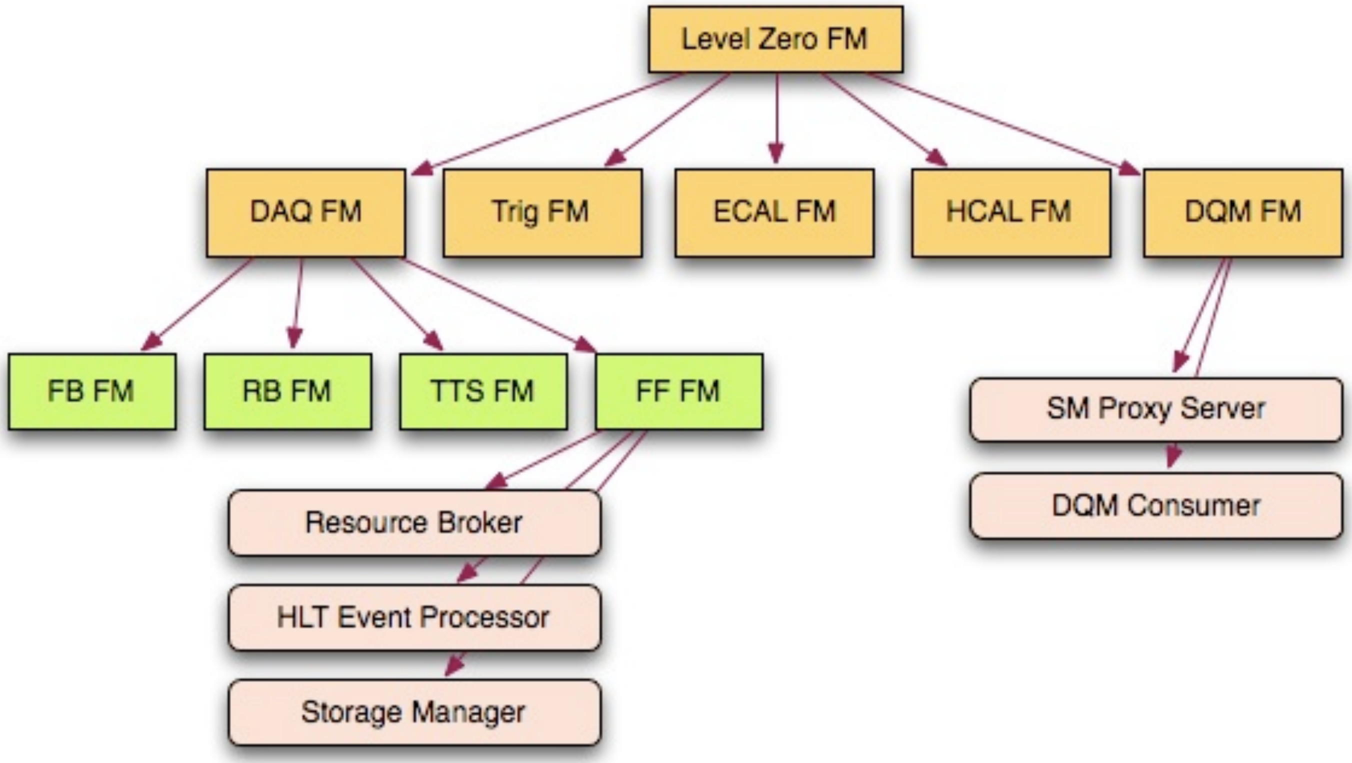
- C++ Framework and components
- Reusable building blocks for
  - Hardware access
  - Transport protocols
  - Services
- Dynamic configuration based on XML
- Controlled and browsable with http/soap





Run control

- Configures and controls all (~10 k) applications
- Hierarchy of finite-state machines (Function Manager)
- Uses Java / Web technologies

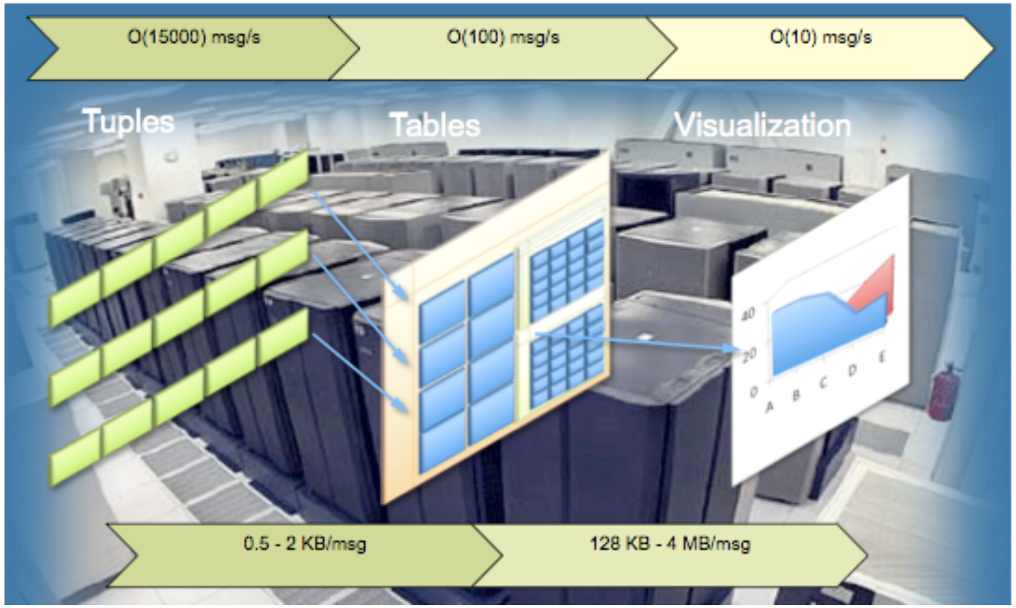


Top level control Web - GUI

- GUI is a web-page
- Top level is Global state machine, aware of LHC states, eg stable beams
- Trigger configuration and clock source (LHC/local)
- Automatic Sub-system configurations, eg level of zero suppression
- Cross-checks and warnings to help the DAQ shifter

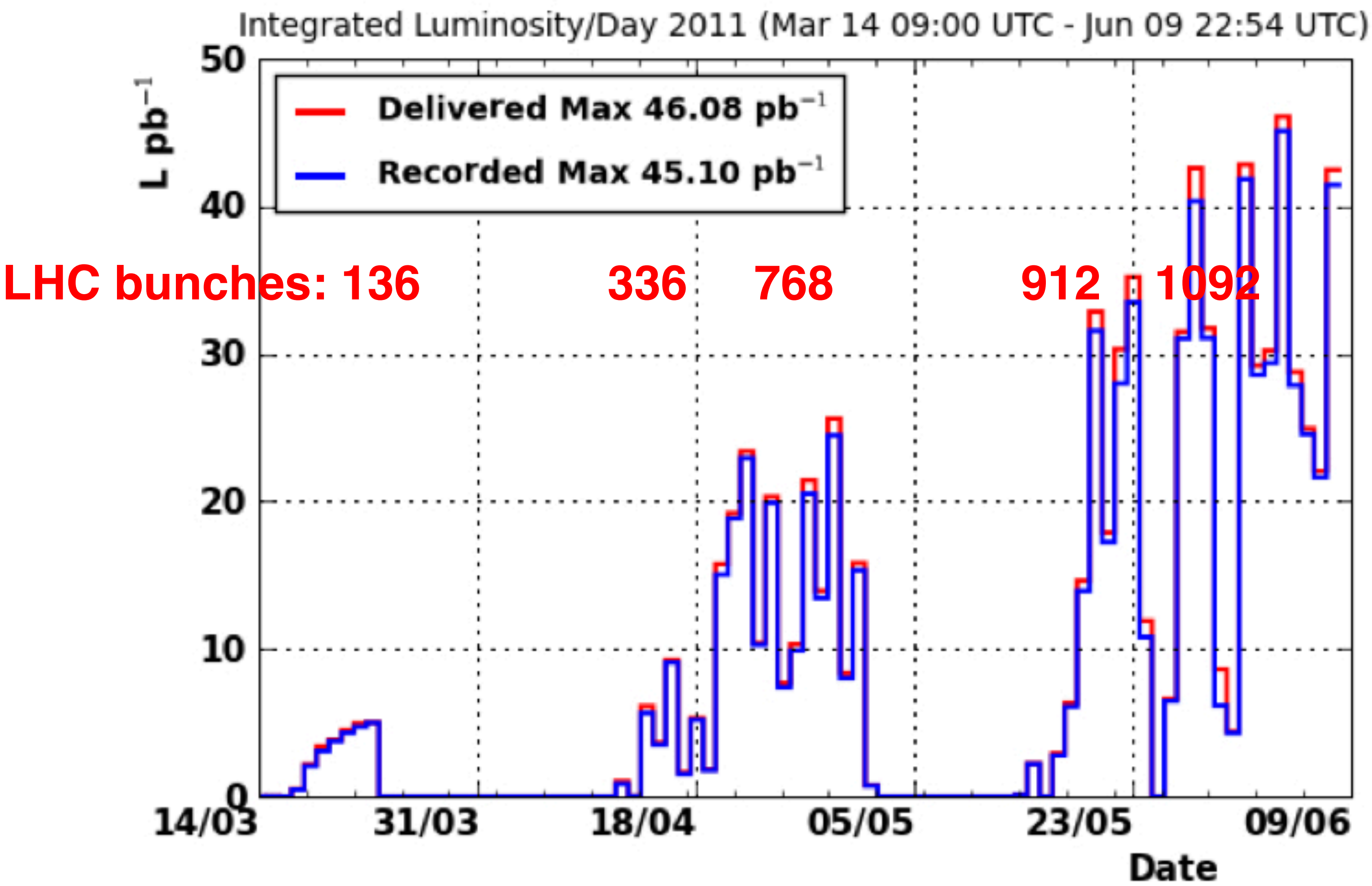
DAQ monitoring

- Monitoring of tuples and error messages
  - O(2k) PCs
  - O(20 k) applications
- Collect and aggregate
  - Hierarchy of collectors
  - Load balancing
- Access service for
  - Visualization applications
  - Error reporting GUIs
  - Expert system
- Persistent storage
  - in relational dBase
- Latency ~1 s



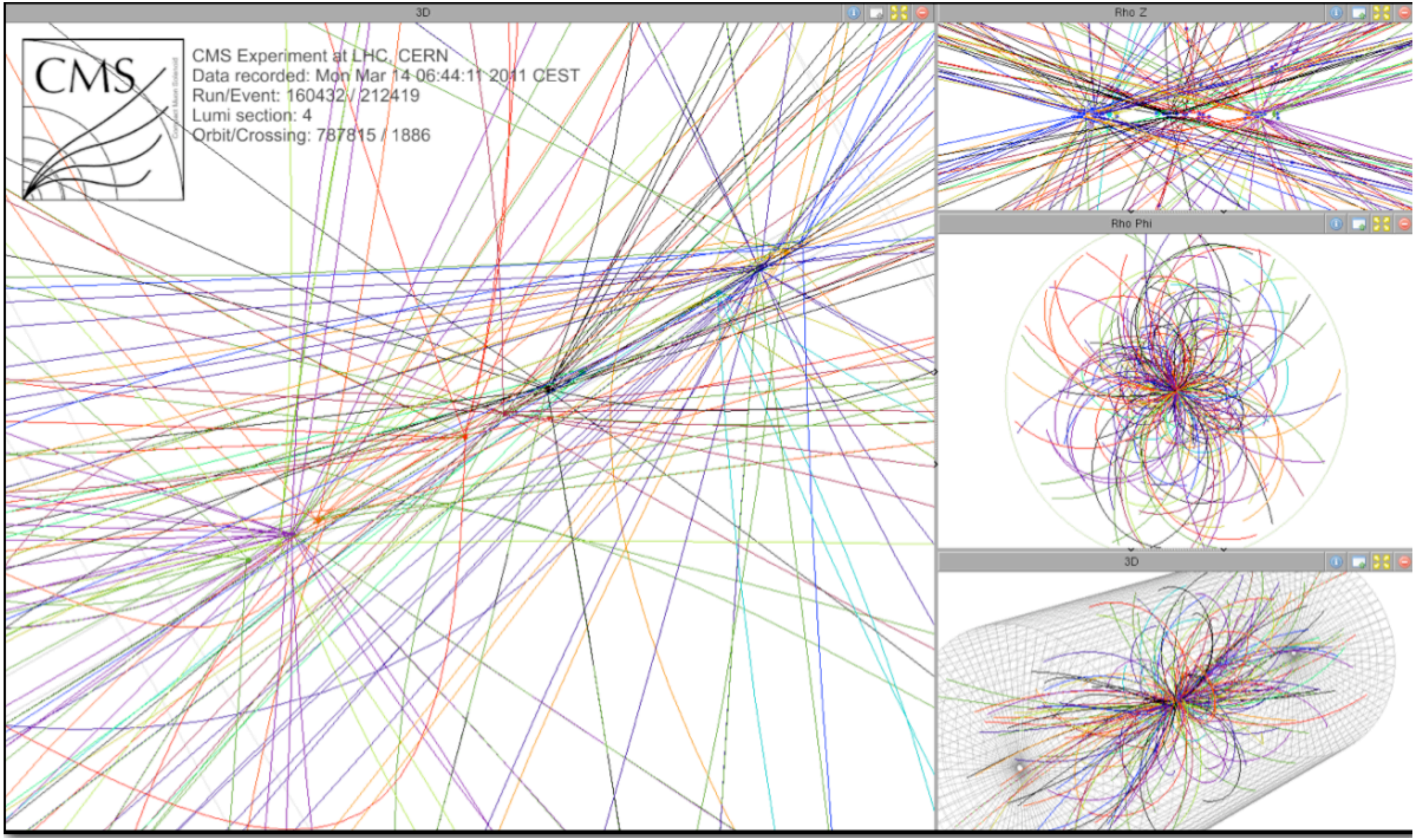
OPERATION

CMS data taking

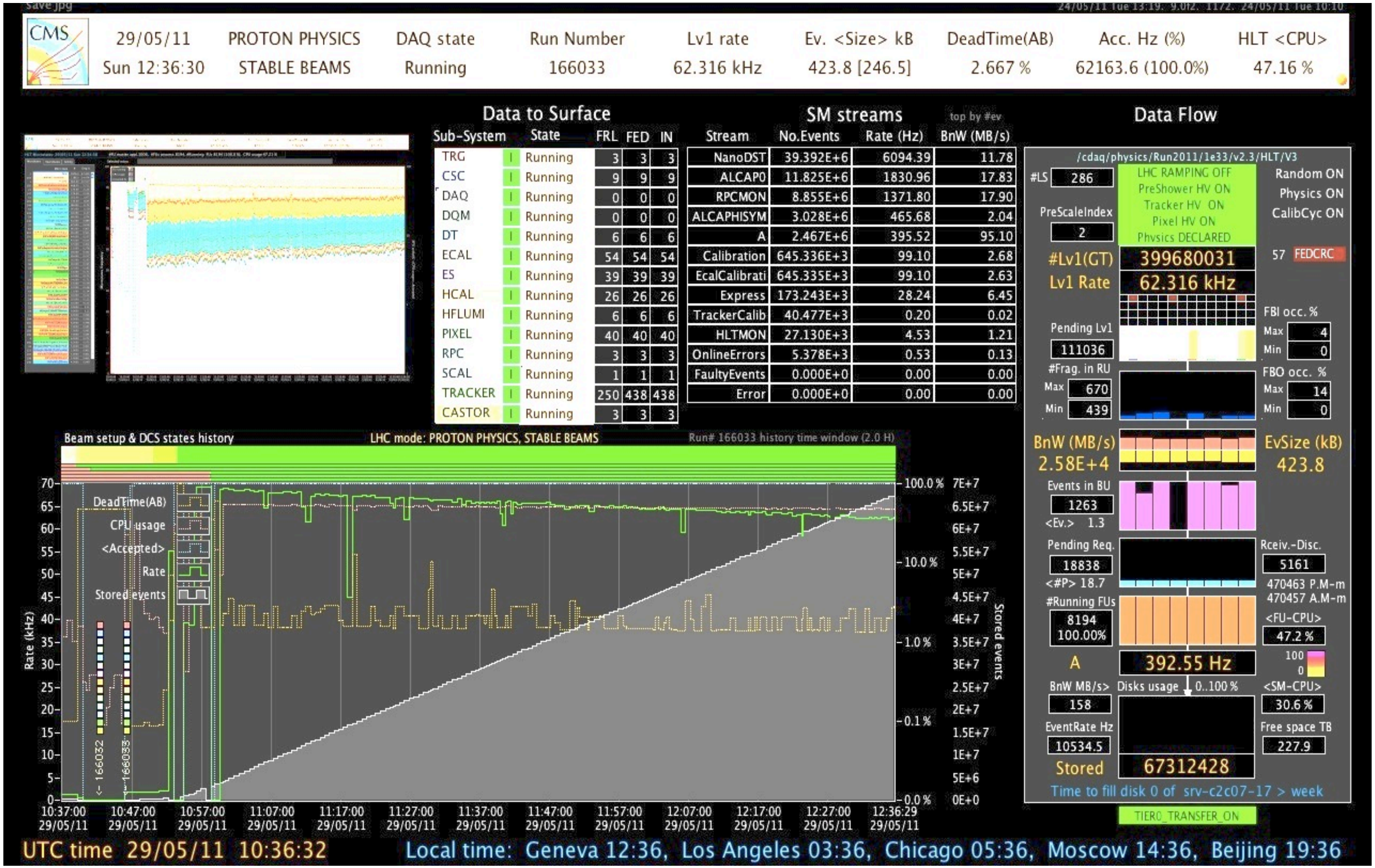


- 2010: Lumi pp delivered/collected 47/43 pb^{-1}; Pb-Pb 9.5/8.7 ub^{-1}
- 2011: Lumi pp (till 09.06) delivered/collected 831/763 pb^{-1}
- CMS Overall data taking efficiency: ~92%

pp collisions in CMS

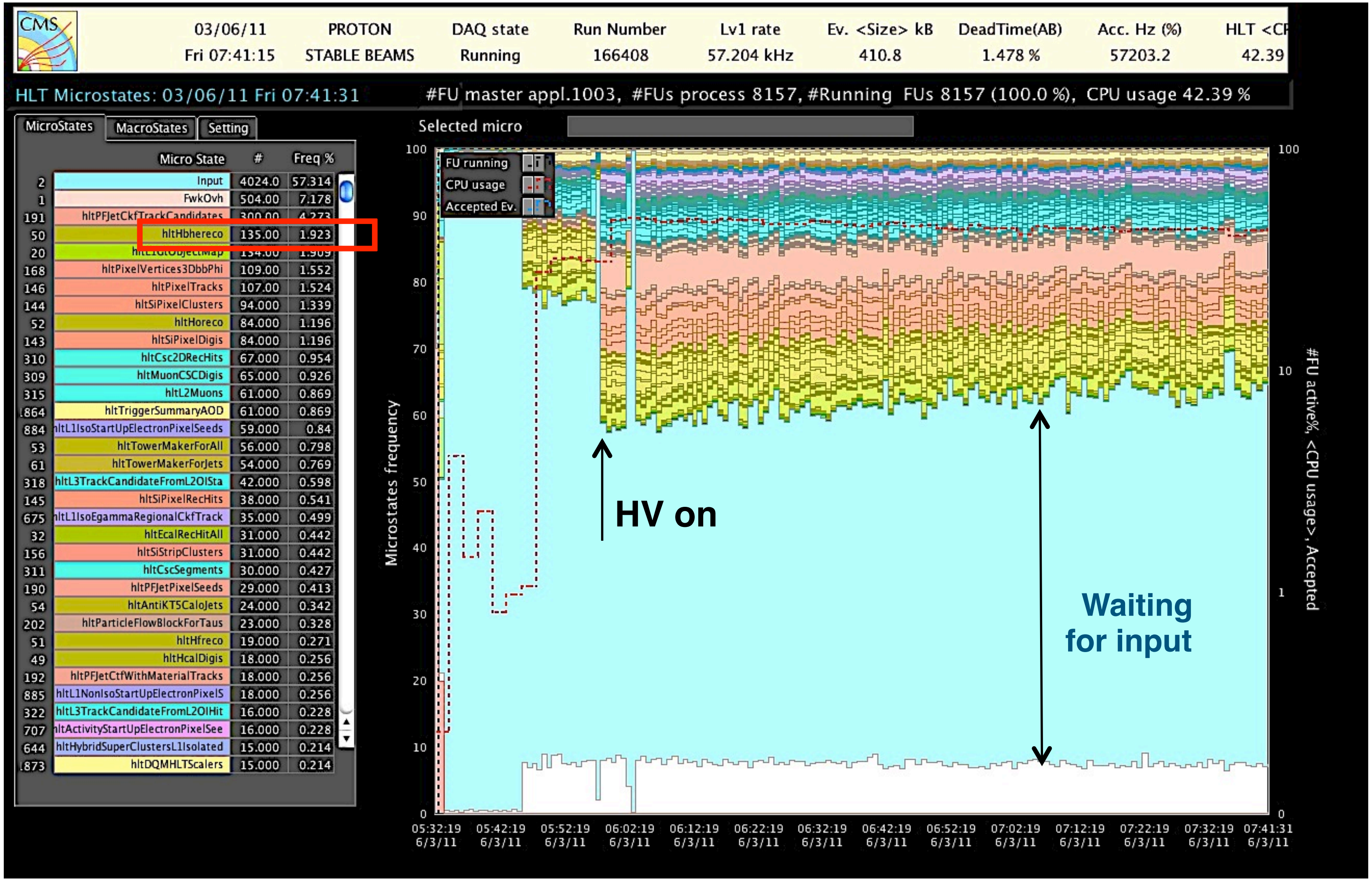


LHC 1092 bunches



LHC Fill 1815 1092 bunches, 1042 colliding, lumi ~1300 10^30, L1 70 kHz, Stream A ~400 Hz, HLT CPU ~50%

real-time HLT profiler



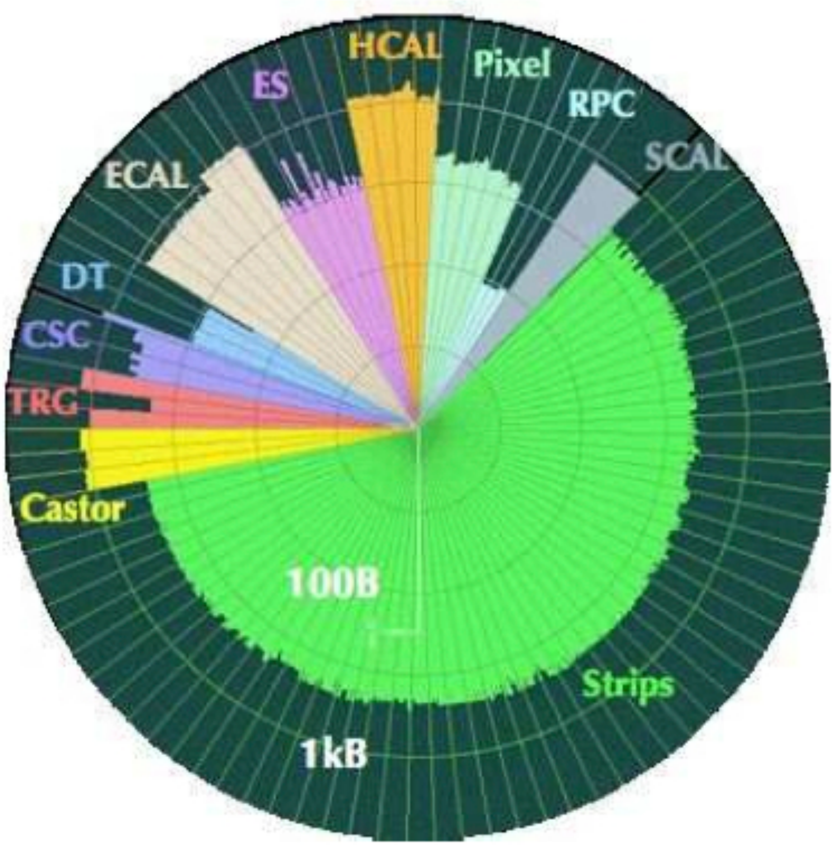
HLT streams

SM streams			top by #ev
Stream	No.Events	Rate (Hz)	BnW (MB/s)
NanoDST	39.392E+6	6094.39	11.78
ALCAP0	11.825E+6	1830.96	17.83
RPCMON	8.855E+6	1371.80	17.90
ALCAPHISYM	3.028E+6	465.68	2.04
A	2.467E+6	395.52	95.10
Calibration	645.336E+3	99.10	2.68
EcalCalibrati	645.335E+3	99.10	2.63
Express	173.243E+3	28.24	6.45
TrackerCalib	40.477E+3	0.20	0.02
HLTMON	27.130E+3	4.53	1.21
OnlineErrors	5.378E+3	0.53	0.13
FaultyEvents	0.000E+0	0.00	0.00
Error	0.000E+0	0.00	0.00

- High Level Trigger
  - Performs 2nd level trigger
  - Categorizes events in streams and PD (Physics Data) sets
- Stream A is “physics” stream with several PDs
- Challenge for physics groups to restrain to less than ~300 Hz total

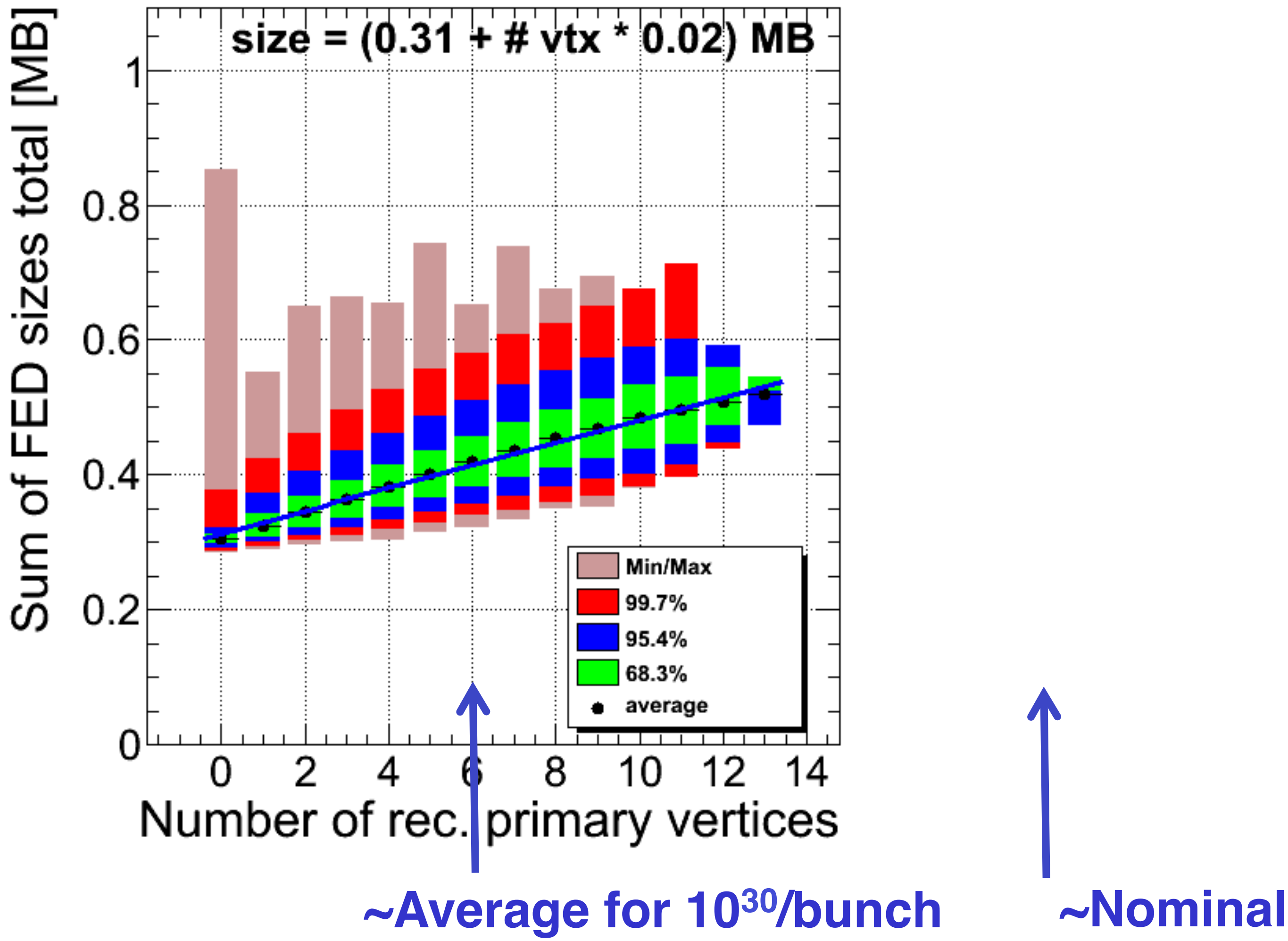
- (Can change L1 and HLT pre-scales “on-the-fly” at “lumi-section” boundaries)

Event Sizes (FED and EVB) in pp



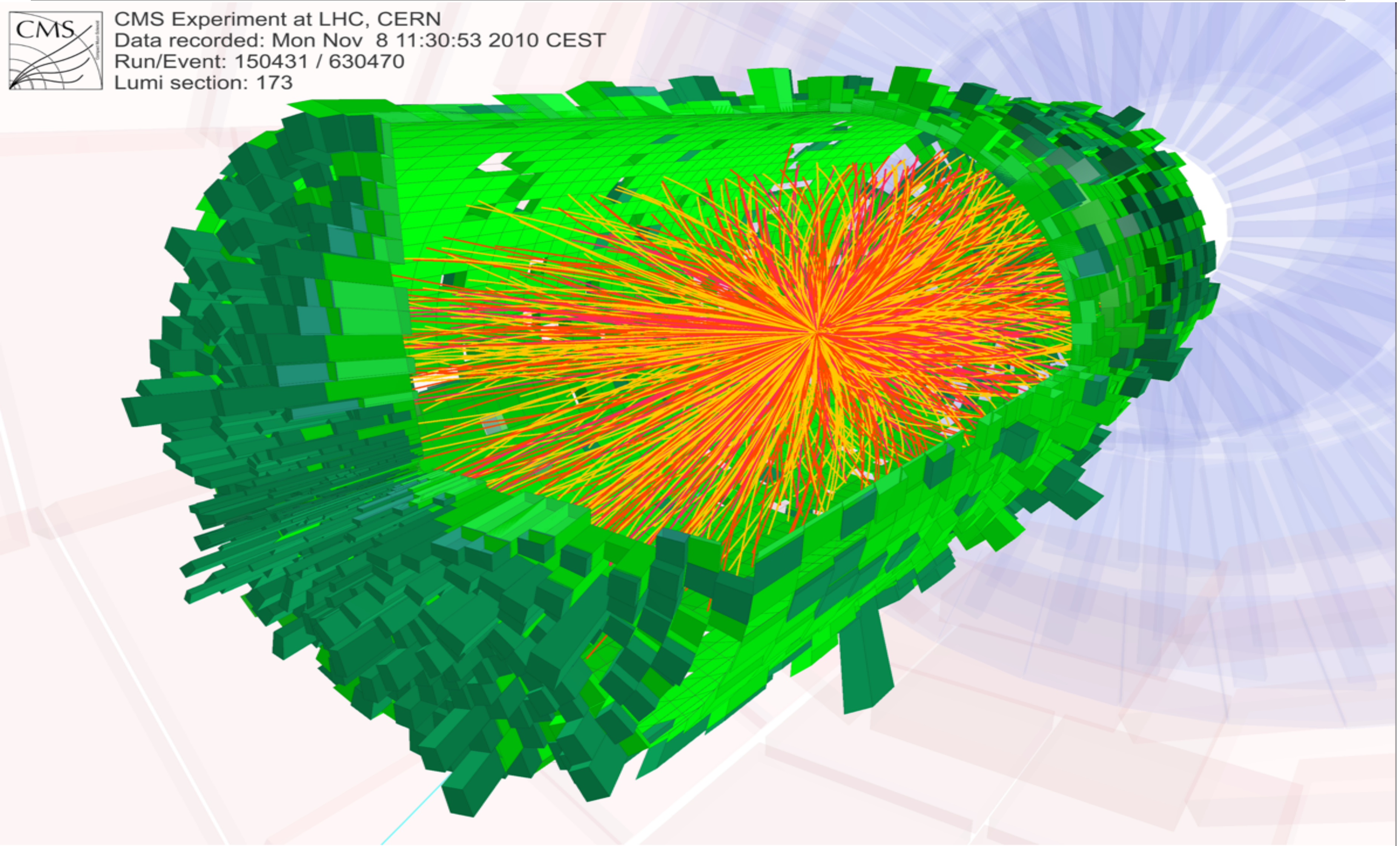
- Almost all sub-det FEDs apply zero-suppression (ZS)
- Data volume increasing with lumi/bunch
- For Tracker 2-1 FED to FRL merging
- Total size ~400 kB/evt, after (ROOT) compression in HLT ~200 kByte
- Nominal: 2 kByte per FRL (1 or 2 FEDs) for 20 interactions/Xin

Event Size versus Pile Up

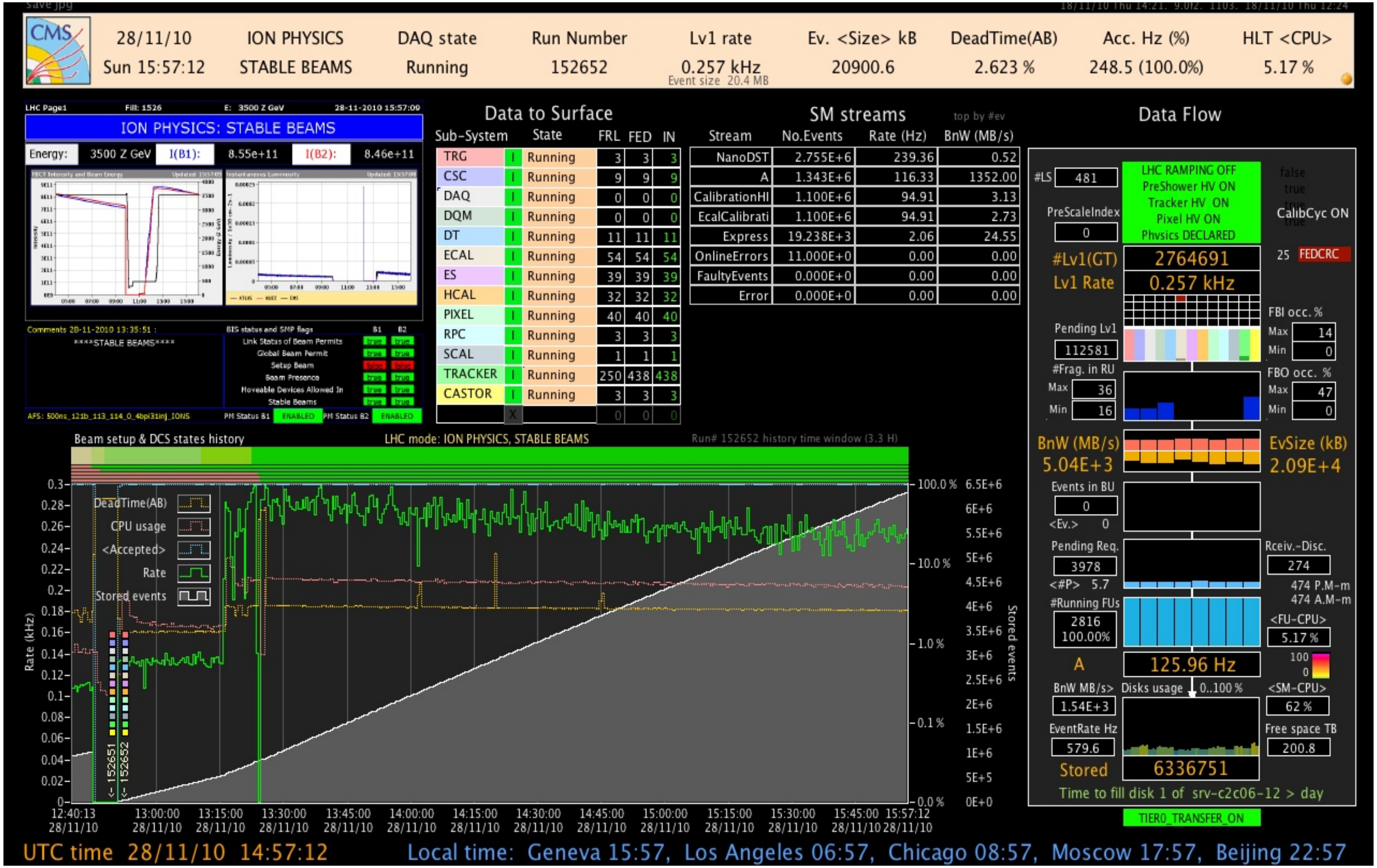


- Trigger and Muon: ~constant
- Tracker and Calorimeter: ~linear rise
- Note: recorded size ~factor 2 smaller due to compression in HLT

Pb-Pb collisions in CMS

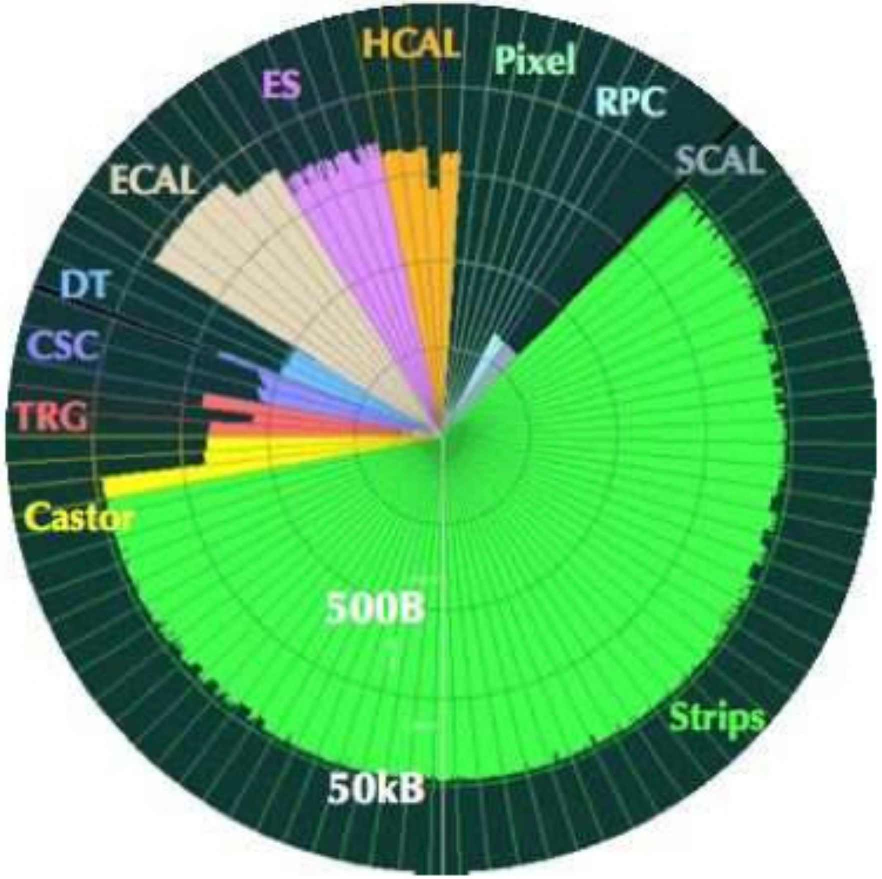


2010 DAQ for Pb-Pb



LHC fill 1526: 114 bunch colliding, lumi: L=~3e25/cm^2/s, L1 250 Hz, stream A 116Hz

Event Sizes (FED and EVB) in Pb-Pb



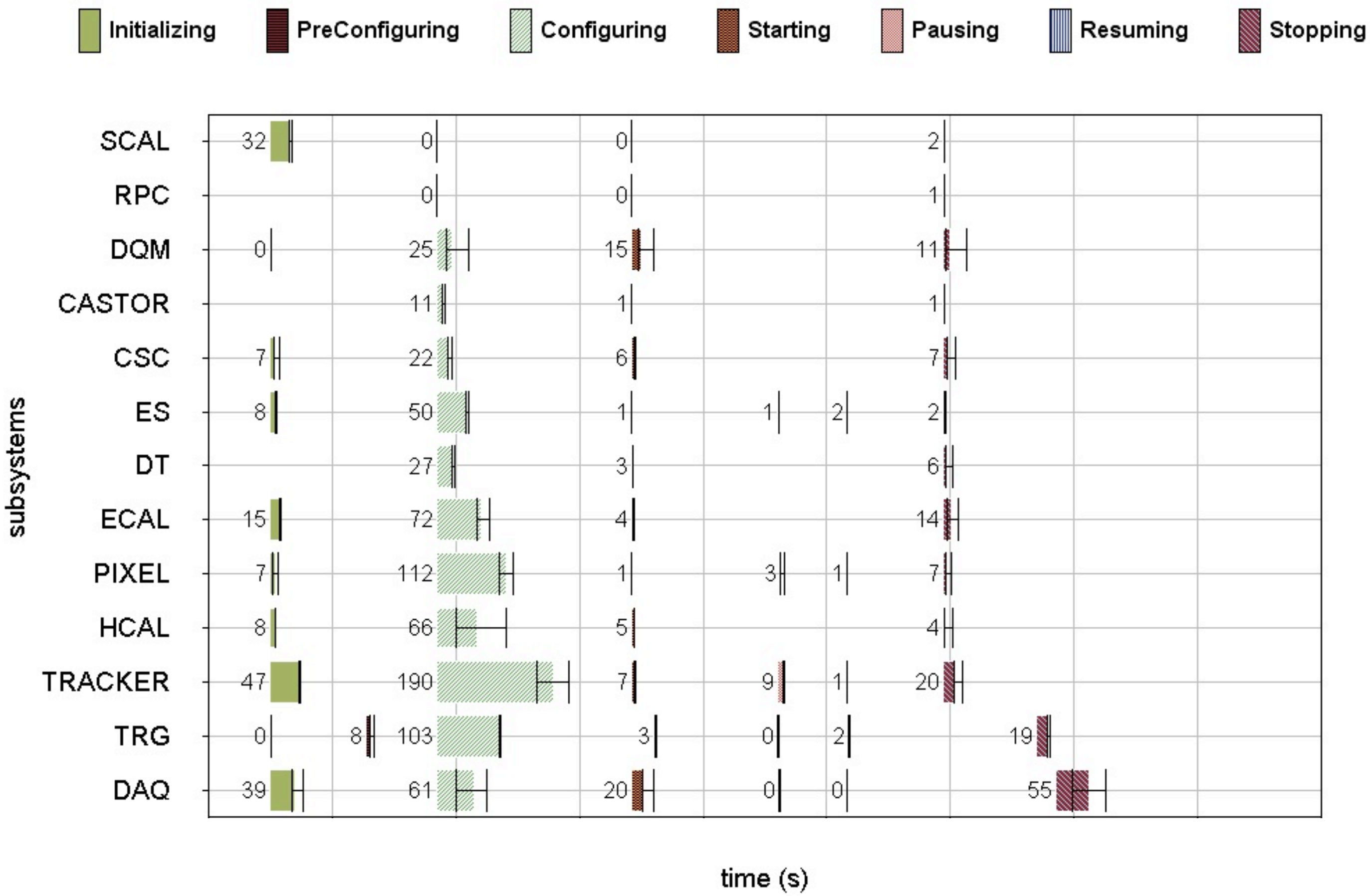
- FED and Event Builder
  - NO zero-suppression in FE
  - Total size ~20 MByte
  - TK FED 50 kByte, with merging 100 kByte per FRL: **50 times nominal**
- Storage Manager (16 SM nodes)
  - After (ROOT) compression ~11 MByte
  - Record + transfer at ~1.8 GByte/s

Data taking Efficiency

- CMS efficiency ~92%
  - **Dead time**: ~1% due to trigger rules
  - **Down time**:
    - “one-off” incidents with sub-det, trigger, Daq, HV, etc.
    - Sub-det electronics loses sync
      - Mostly recovered automatically with TTS system

- Central DAQ **availability** ~99%
  - Pathological events crashing HLT written to “error stream” and process restarted
  - Possibility to disable DAQ slices in case of problems
  - Guidance to operator
  - Diagnostic system analyzing monitoring data and proposing actions to operator. Expert system implemented in a script (perl)

Run configure/start/stop



- Cold start time is ~5 minutes
- Run start–stop time is ~2 min, Pause-resume is ~10 s.
- HLT loading of conditions from dBase (via frontier / squid) takes ~1 min

Large Scale

- Deal with large scale by ..
  - **Hierarchical** distribution / collection of data
    control, monitoring, dBase access, system installation
  - **Parallel** services (cluster services)

- Quantity of equipment:
  - ~3000 PC nodes:
    - failure ~daily/weekly
  - ~4000 Ethernet 1 Gbps copper links and switch ports:
    - failure ~monthly
  - ~6000 Myrinet fibre tranceivers:
    - failure ~monthly
- Observed subtle effects with large volume of equipment.
  - eg slower memory in pathological PCs
  - eg auto-negotiation problems

Remarks ..

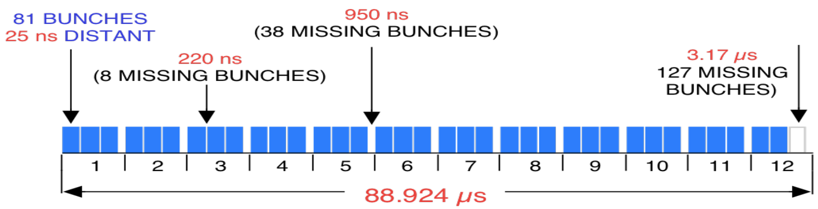
- Move from SLC4 to SLC5 / **64 bit** gained ~20% in HLT performance
- Migration to SLCx versions induced subtle networking performance effects
- **Emulator** mode for central DAQ proved extremely useful to be able to test large-scale system independent of sub-detectors
- The "**DAQ slice**" concept initially introduced for performance scaling proved also useful for fault mitigation

- The concept of "**Lumi-Section**" (LS)
  - was not foreseen in the Technical Design Report (TDR) 2003
  - period of ~23 s. of data taking (2^18 LHC orbits)
  - "atomic" unit for physics analysis
  - Can change the "soft" trigger configuration (by pre-scales) at these LS boundaries
  - Associated book-keeping and control

- "*The devil is in the details*"
- There is never enough monitoring / diagnostics

Prospect for 2011-2012 run (pp)

- **pp** at 2 x 3.5 TeV
  - Highest Lumi so far 1.3 x 10^33 with 1092 bunches (1042 coll.)
  - Lumi increase by factor 4 not excluded
    - 1380 bunches, smaller emittance, higher currents
    - ~30 events Pile-Up at 50 ns (exceeds TDR nominal conditions)
  - HLT now ~60 ms/evt (increases slightly faster than linear with PU)
  - Have a HLT CPU budget of ~90 ms/evt at 100 kHz
  - Tighten selection in L1 (to stay below 100 kHz accept) and HLT (below ~400 Hz accept)
- Can extend HLT farm further by adding PC boxes



Prospects 2011-2012 Heavy-Ion

- Nominal Heavy Ion **luminosity** is 8 kHz of Pb-Pb collisions,
- in 2011 it will be less than that, aim is to be **ready for up to 3 kHz**.
  - DAQ: tracker FRL with 2 FED merged 3 kHz x 100 kB = 300 MB/s (50% above 'nominal'), tested
- In 2010 **zero suppression** was done offline. For 2011 will be done in HLT farm (with NZS of ECAL and HCAL but with tracker ZS in HLT)
- HLT for Heavy Ion
  - will include **muon trigger** to select J/psi and Upsilon
  - Aim for HLT accept rate (recording): 100-150 Hz,
  - compressed event size 2 MByte
  - Maximum Storage Manager throughput ~2 GB/s. OK

FUTURE DAQ

LHC – CMS outlook

	2011	2012	2013	2014	2015	2016	2017	2018
	7 TeV		LS1			14 TeV		LS2
lumi 10^34 /cm2/s	0.2	0.5				~ 1 x		
events/xing 25/50 ns	4/8	10 / 20				~20 / 40		
Tracker						new Pixel, more channels		
Muons						complete forward muons		
CALO						new HCAL sensors and electronics		
Trigger						uTCA in parallel ('spectator')		

	2018	2019	2020	2021	2022
	LS2	14 TeV		LS3	14 TeV
lumi 10^34 /cm2/s		~2 x			~ 5 x
events/xing 25/50 ns		~ 40 / 80			~ 100 / 200
Tracker					new strips, 5x
Muons					
CALO					
Trigger		uTCA in production			tracking trigger

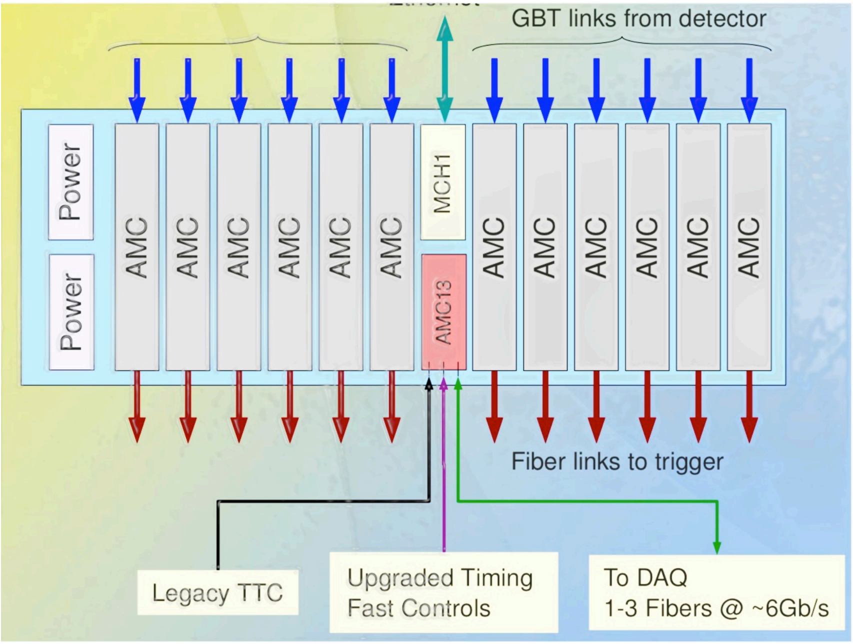
One of the possible scenarios being discussed

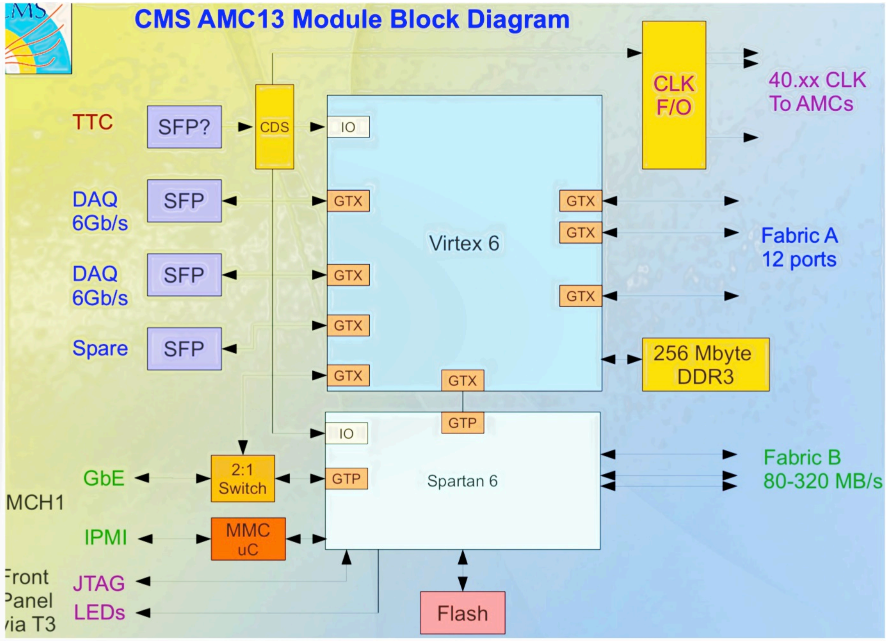
CMS DAQ 2014-2018

- Installation of new Pixel detector (more channels)
- Change of off-detector electronics for HCAL and Trigger
- Electronics will not change till >2021 for most sub-detectors (legacy FEDs)

- **Requirements for central DAQ**:
  - Readout of legacy FEDs (>90%)
  - Readout of new FEDs uTCA based and multi-gbits links (~10 Gbps)
  - Moderate increase in event sizes
    - Increase of number of channels
    - Possibly lumi ~2 10^34 and/or 50ns running
- DAQ leverages commercial networking and computing equipment technology
  - Take advantage of the rapid increase in price/performance
  - Typical lifetime ~5 years: In 2015 all equipment more than 5 years old

uTCA based off-detector electronics

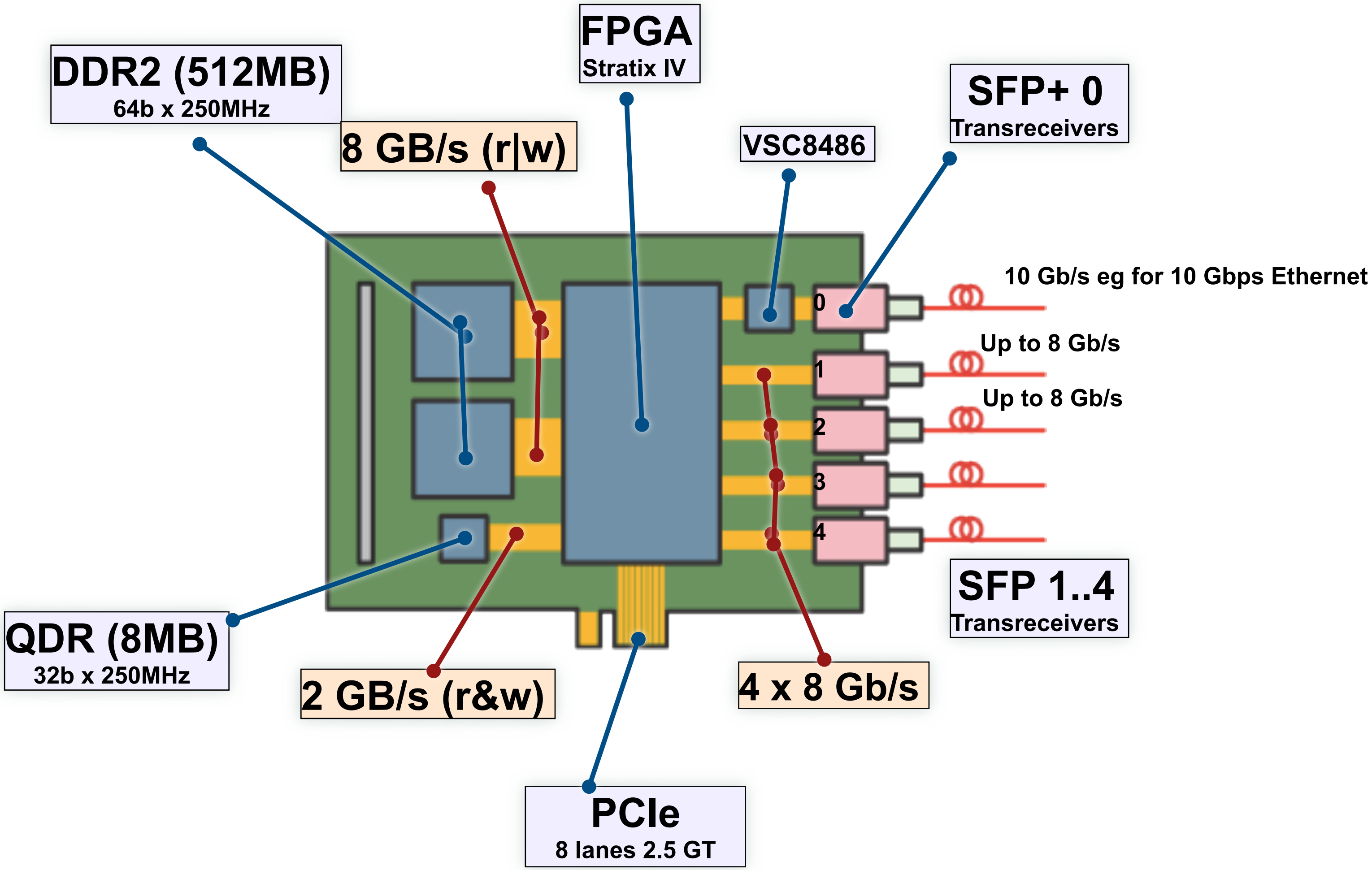




- Under development by BU (Boston University) for HCAL
- This structure is also considered for some of the Trigger sub-systems
- AMC13 might evolve in to CMS “common platform”

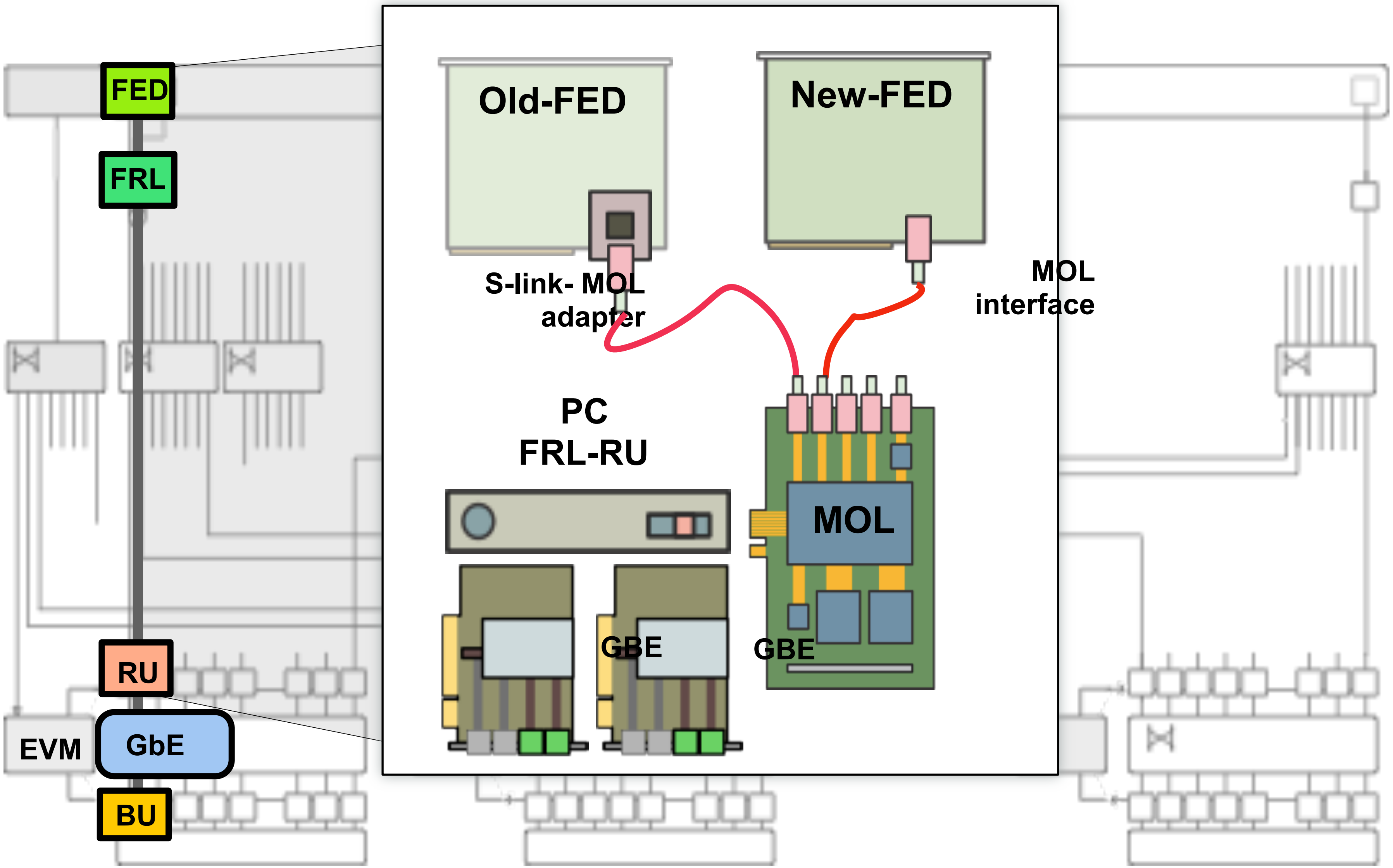
- AMC13 sends data to central DAQ over multi-gbps serial link (6 Gbps in prototype)
- Protocol for data link to central DAQ is under study

MOL DAQ interface board

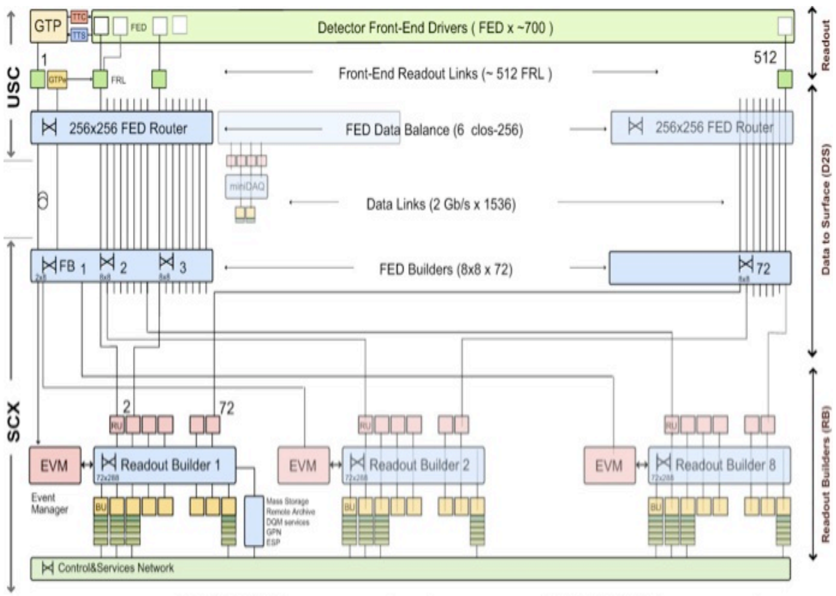


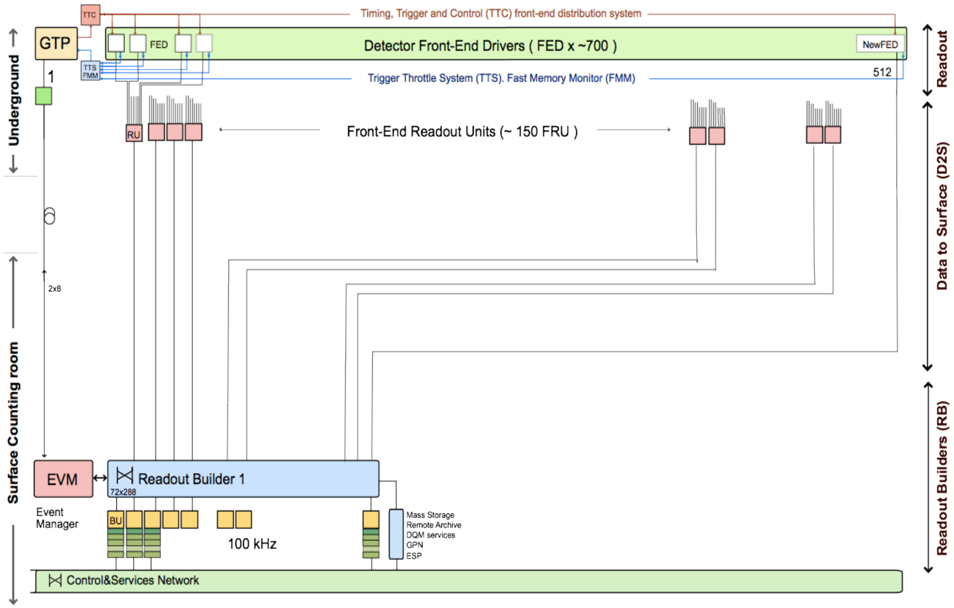
- Evaluation board currently under design in order to study various new options for the next generation DAQ

MOL readout and EVB input



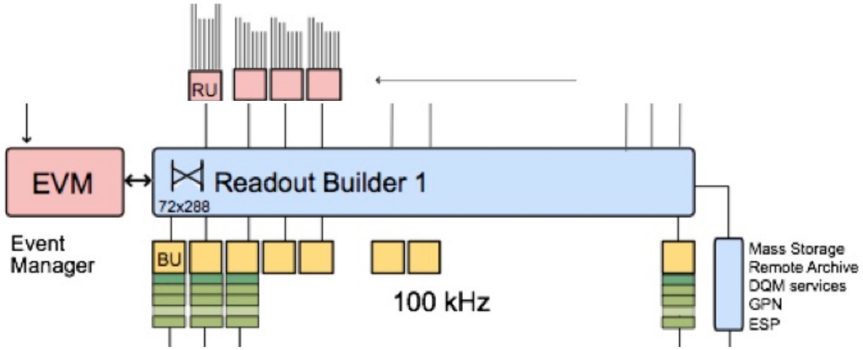
DAQ evolution





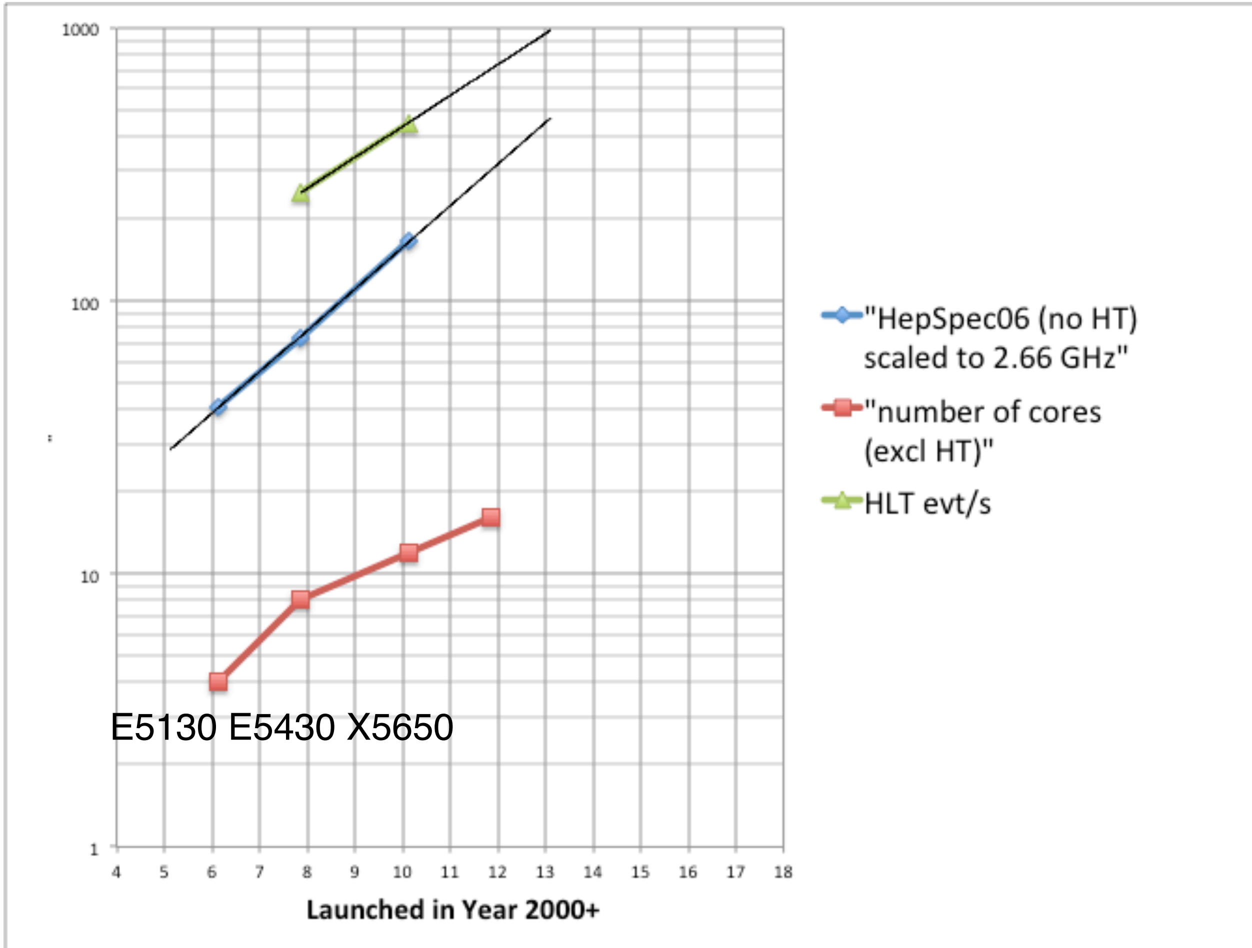
- One of the schemes under study is
  - Concentrate subdet-data into PCs
    - Merging by factor 3-4 of legacy FEDs to reach >= 10 Gbps I/O
    - Interface to new FEDs (minority)
  - EVB
    - EVB nodes with 10 Gbps I/O, eg 150x150 system
  - Distribute to HLT nodes, store accept events by Storage Manager

Next Generation EVB



- Use commercial equipment and standard protocols: PCs, network
- Concentrate sub-det data into PCs at the EVB sources
- With 10 Gbps links: 150 x 150 EVB provides 150 GByte/s EVB
  - TCP/IP: loss-less, fully standard
  - Switch: 300 (10 Gbps) port router, available today
  - Or, layer 2/3 cut-through switch might be interesting alternative (need to study traffic control against head-of-line blocking)
- Or, maybe Infiniband 40 Gbps links 40x40 EVB (needs study)
- Can be expanded if higher throughput required for larger event sizes
- Distribute events from BUs to HLT nodes (likely on 10 Gbps Ethernet)

Extrapolating PC performance



- Extrapolate performance dual-processor PCs
- In 2014 could have same HLT performance with 100 – 200 nodes
- Likely to have 10 GbE onboard

SUMMARY

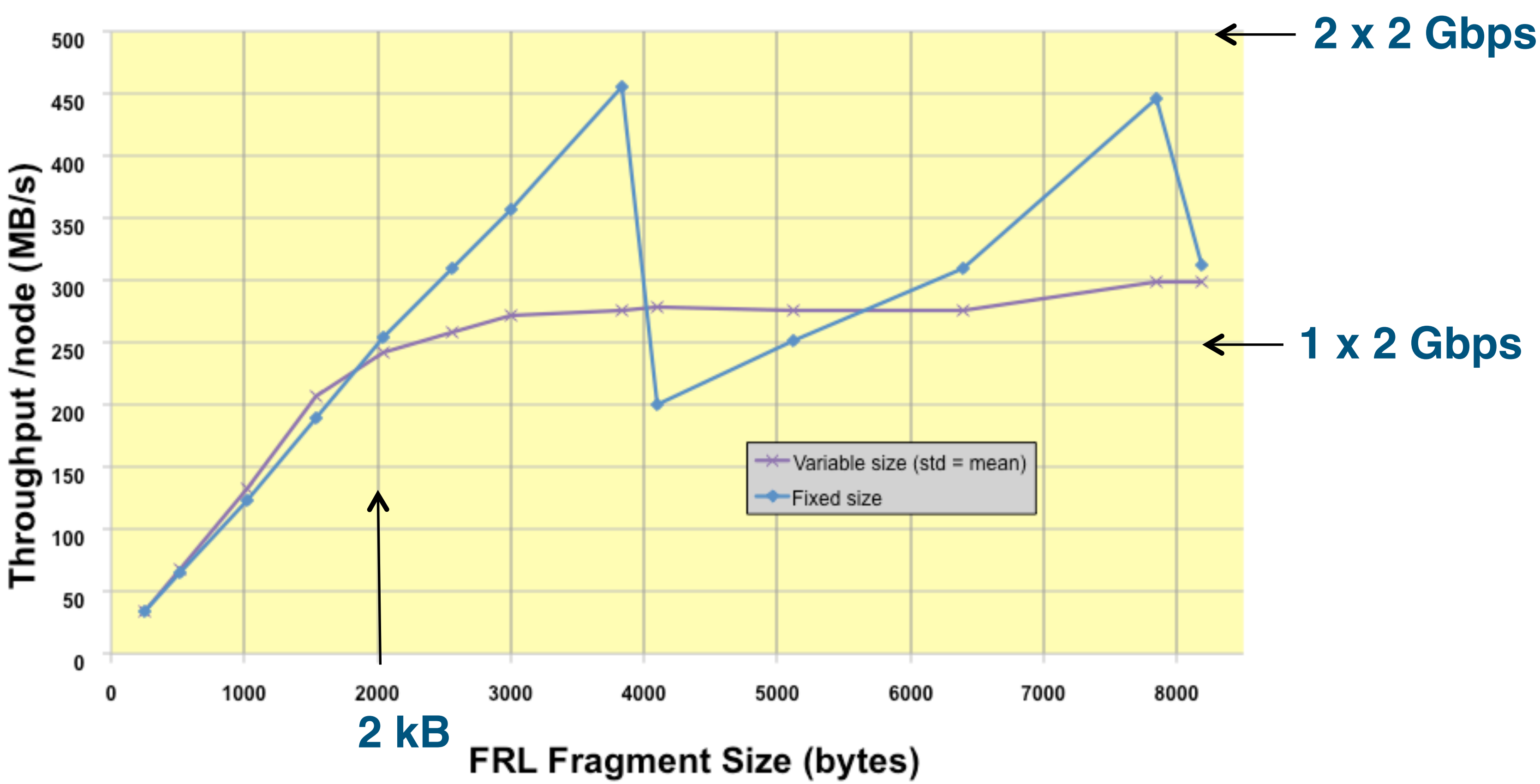
Summary

- CMS has a **flexible** DAQ system
  - Can be easily configured for high rates (pp) or large events (HI)
  - 150 GB/s Event Builder
  - HLT farm can be **expanded** as required
- Event Building at level-1 rate of 100 kHz
  - Sophisticated HLT algorithms can be employed on full events
- Reliable
  - Only minor down-time due to central DAQ failures
- Flexible **architecture**
  - Can be re-implemented with up-to-date networking and computing equipment for a simpler and more compact system

END

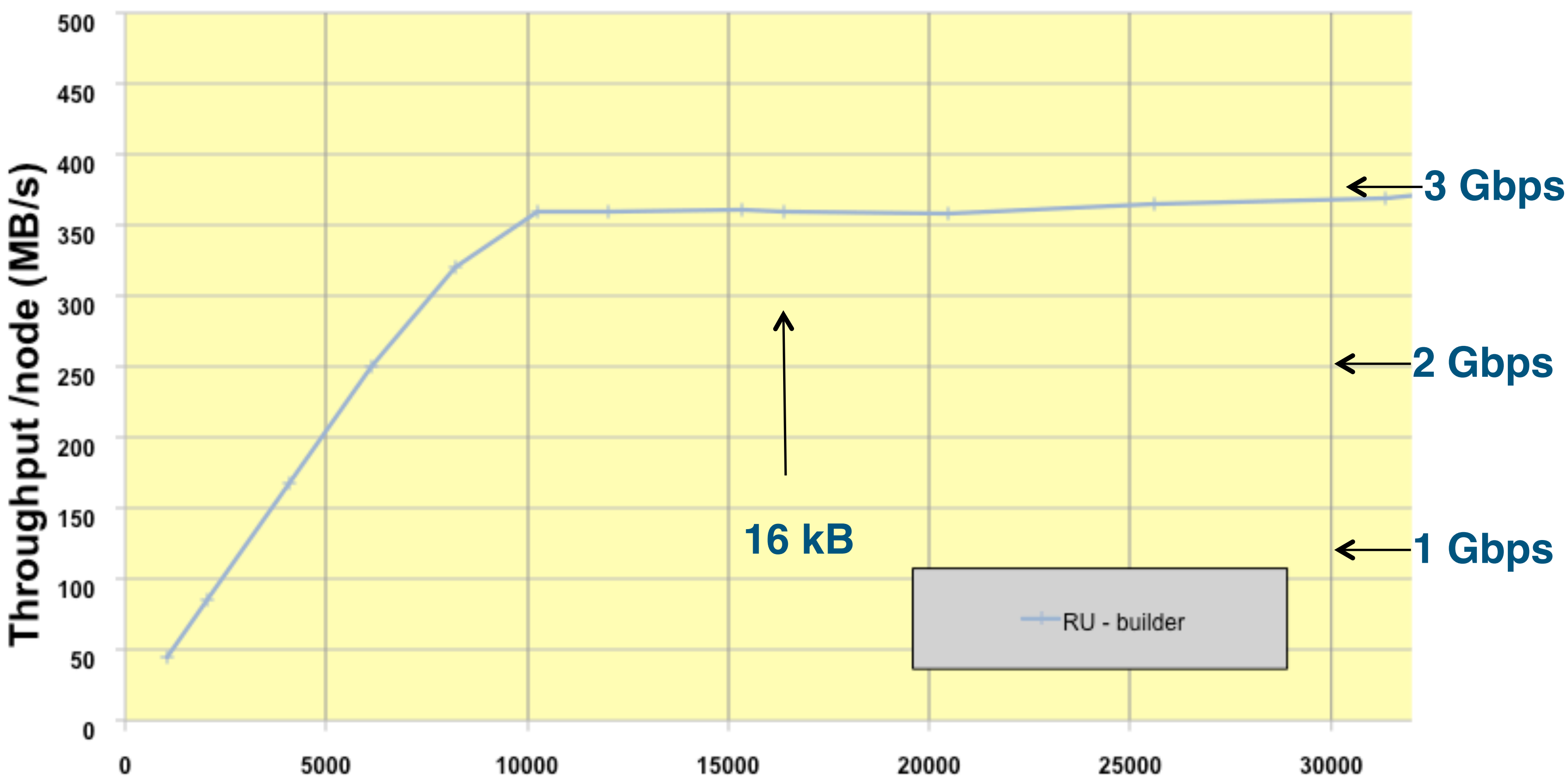
BACKUP MATERIAL

Performance Super-fragment builder



- **8x8 EVB**
- **Myrinet with 2 links (each 2 Gbps data rate)**
- **Can be improved, eg with traffic shaping**

TCP/IP EVB performance



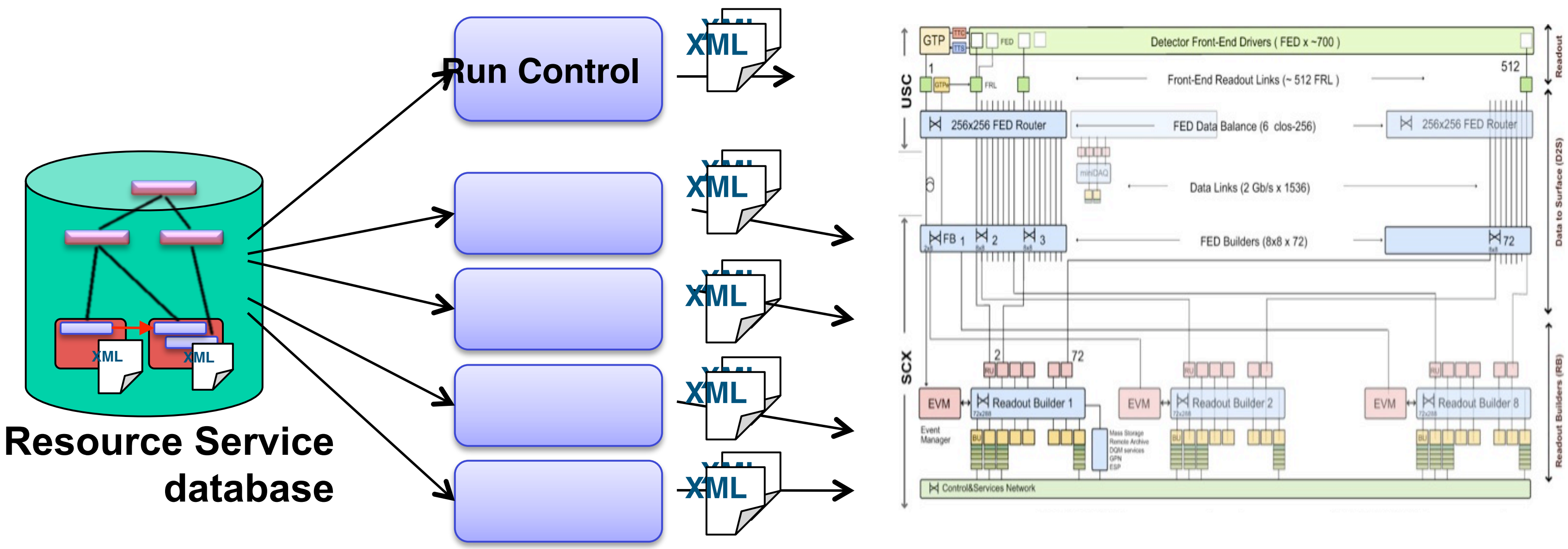
- **64 x 126 EVB**
- **TCP/IP, RU with 3 x 1 Gbps Ethernet links, MTU=1500**
- **Working point is 16 kB super-fragment = 8 x 2 kB fragment**

DAQ configuration

Calculate XML documents for all applications according to

- High level description (selected by user)
- Hardware info in Database
- Black list

Load and configure appllications



Loading and starting of O(10000) applications: ~30 sec