

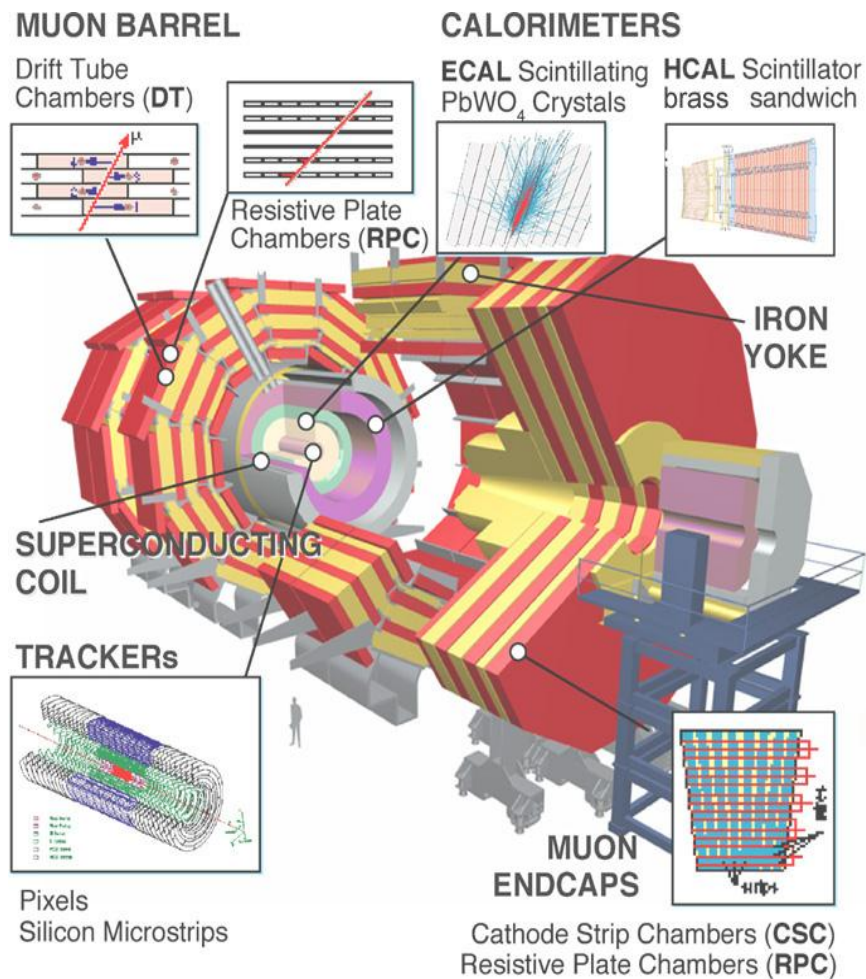
Operational Experience with the CMS Data Acquisition System

Hannes Sakulin, CERN/PH
on behalf of the CMS DAQ group



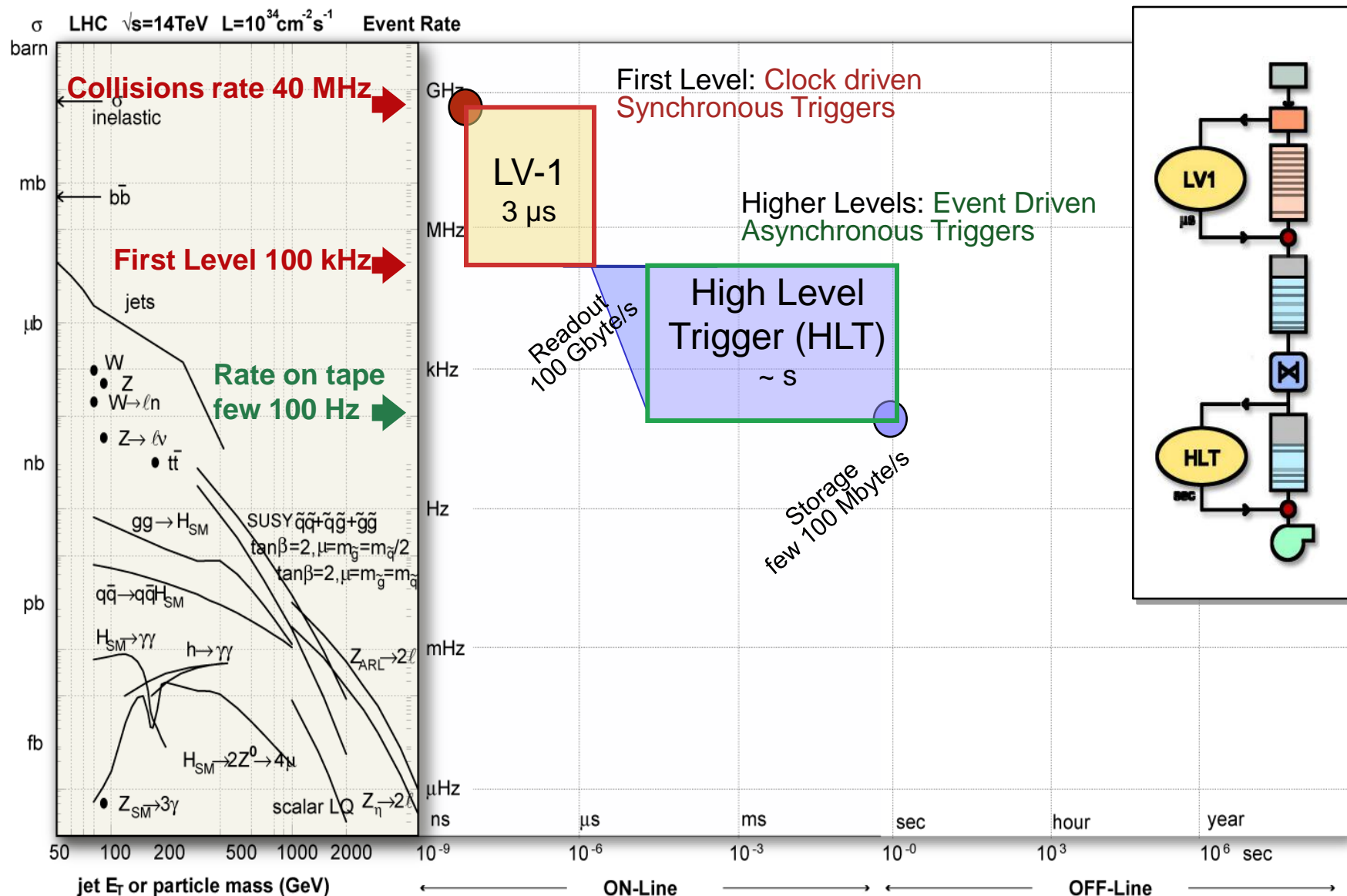


Compact Muon Solenoid



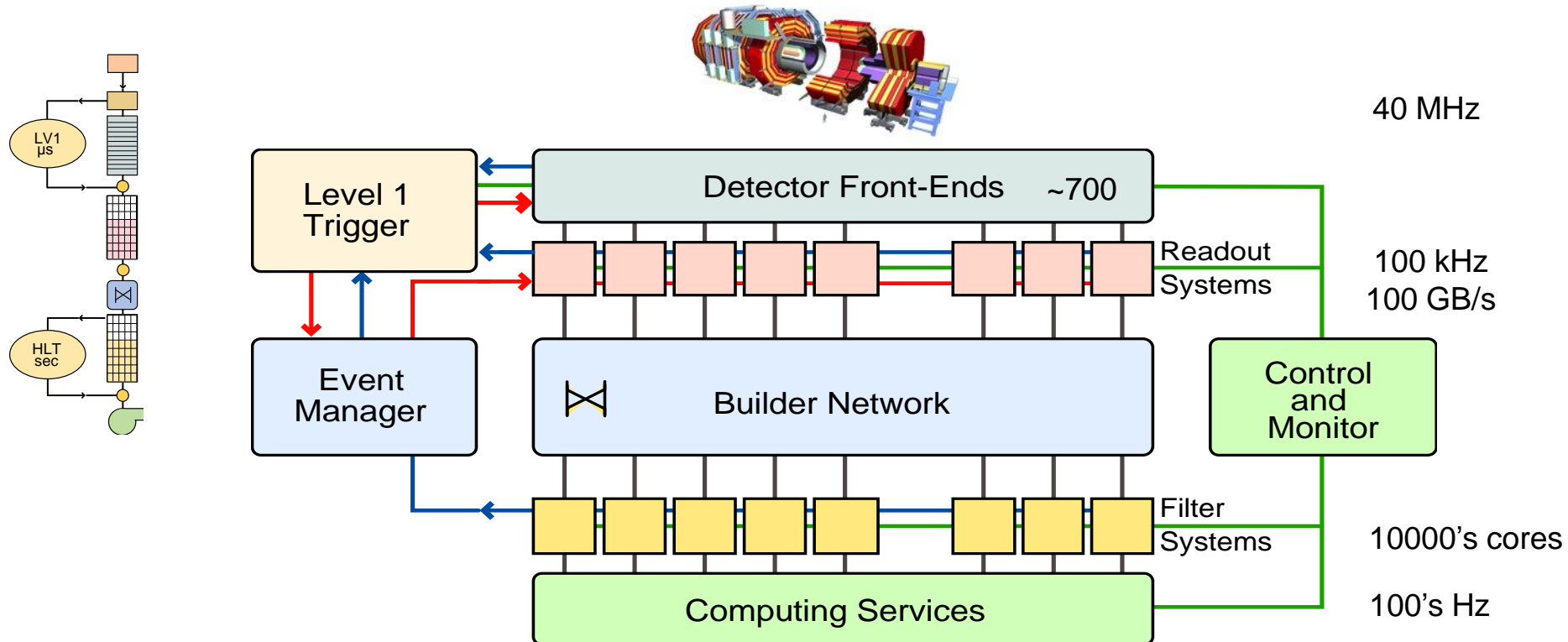
- General purpose detector at the LHC
- 55 million readout channels
 - Event size of 1MB
- Proton physics
 - At 7 TeV in 2010/11
 - At 8 TeV in 2012
- Heavy Ion physics
 - In 2010 & 2011

Two-level trigger concept



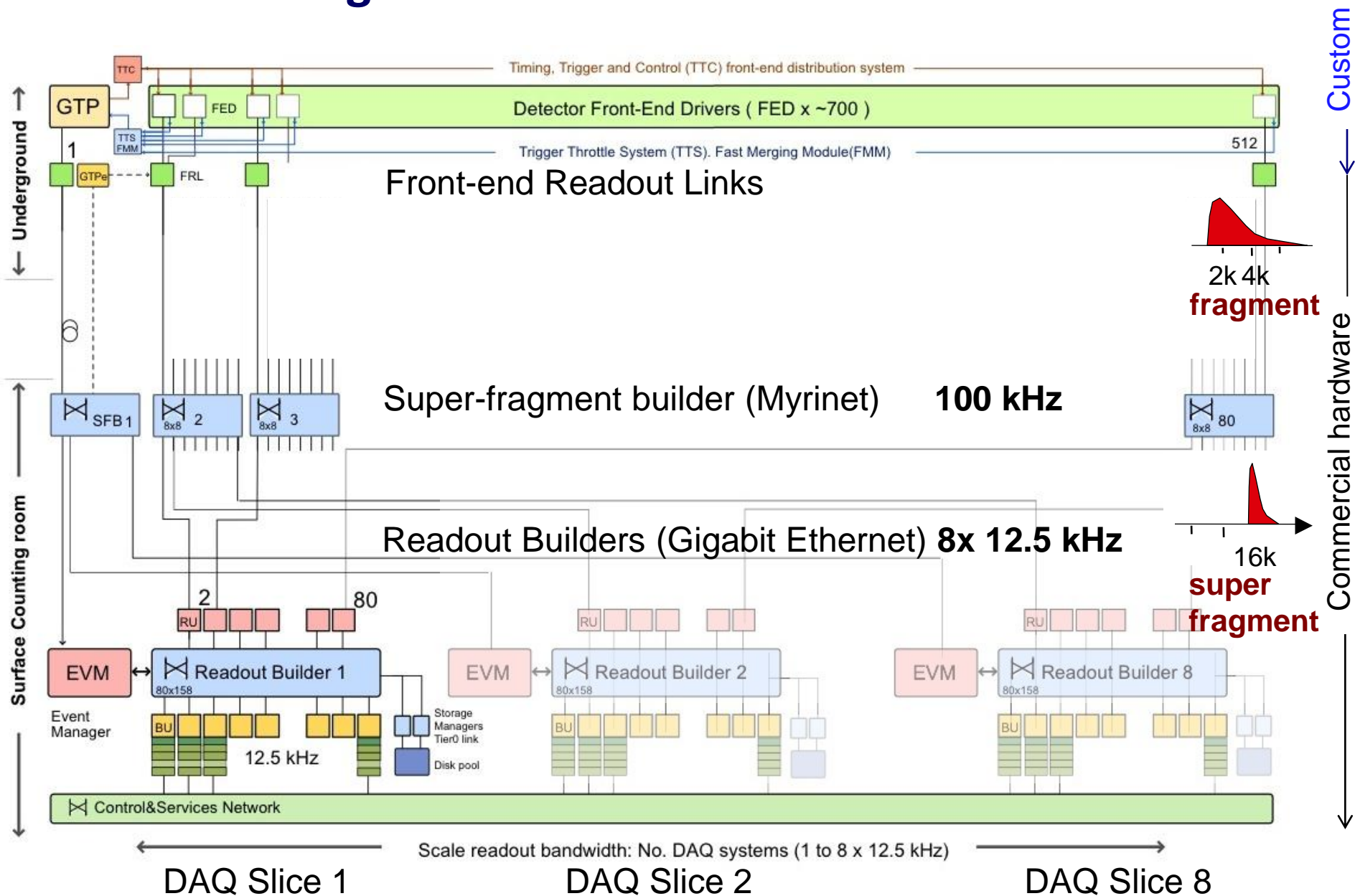


CMS DAQ requirements

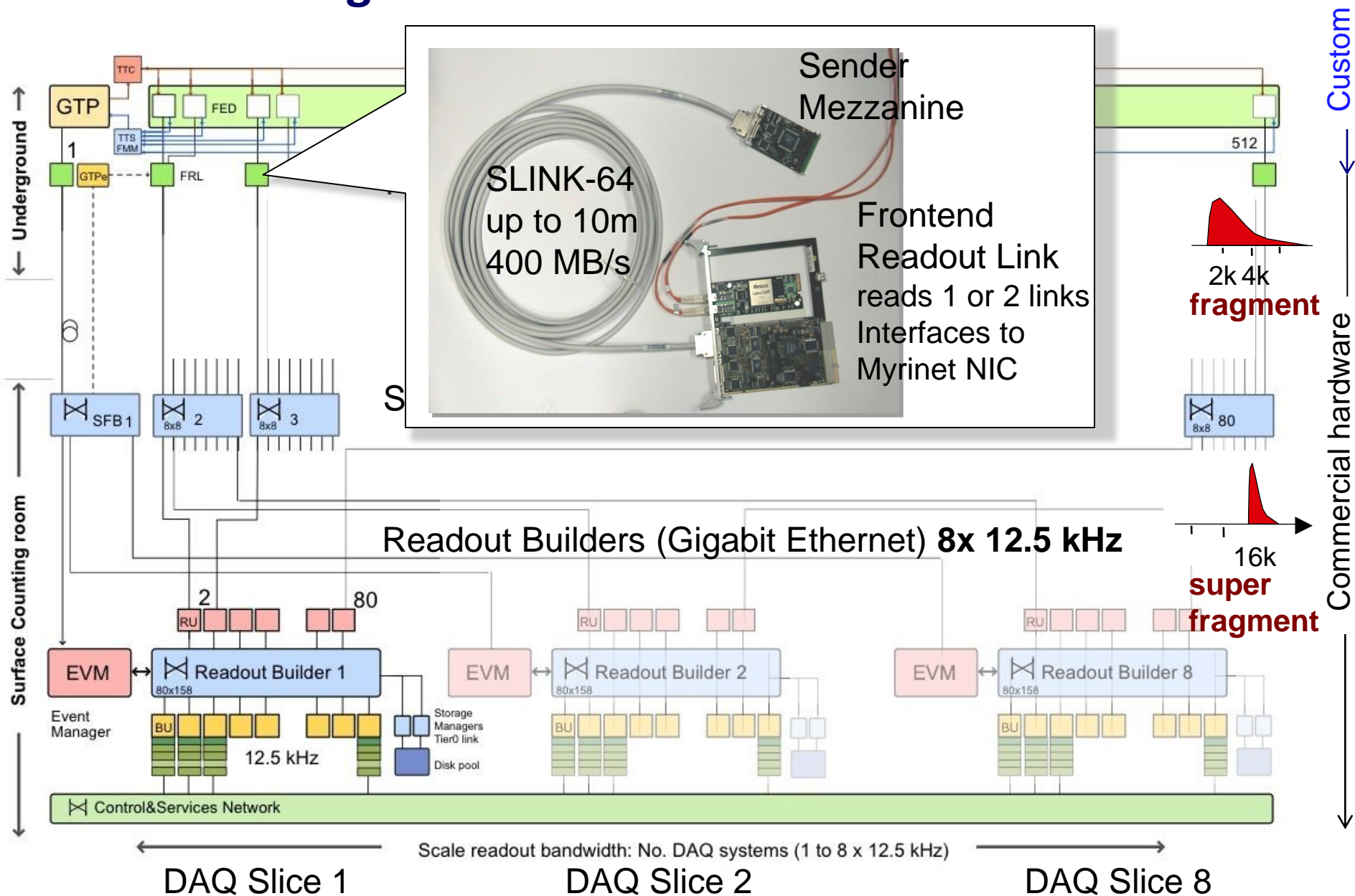


- Read out 700 detector front-ends (max. average fragment size 2 kB)
- Build complete events at 100 kHz (L1 trigger rate)
- Make them available to a filter farm of $O(10000)$ cores
- Store 100's of Hz to disk (10's of TB/day)
- Scalable system employing commercial components wherever possible
 - **Proprietary** / Commercial: **Front-Ends**, VME, PCI, PC servers, networks, Protocols, OS

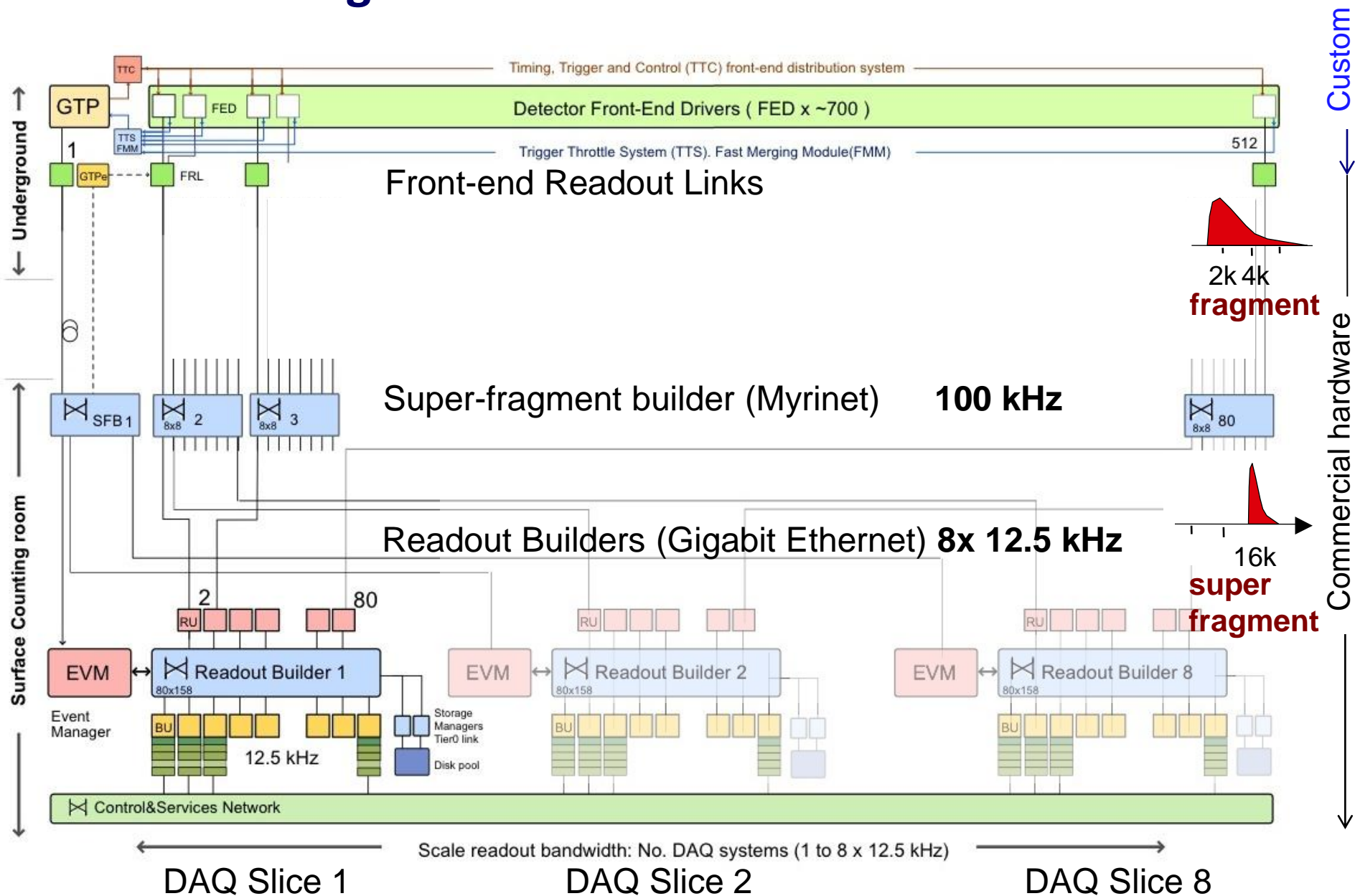
CMS two-stage event builder



CMS two-stage event builder

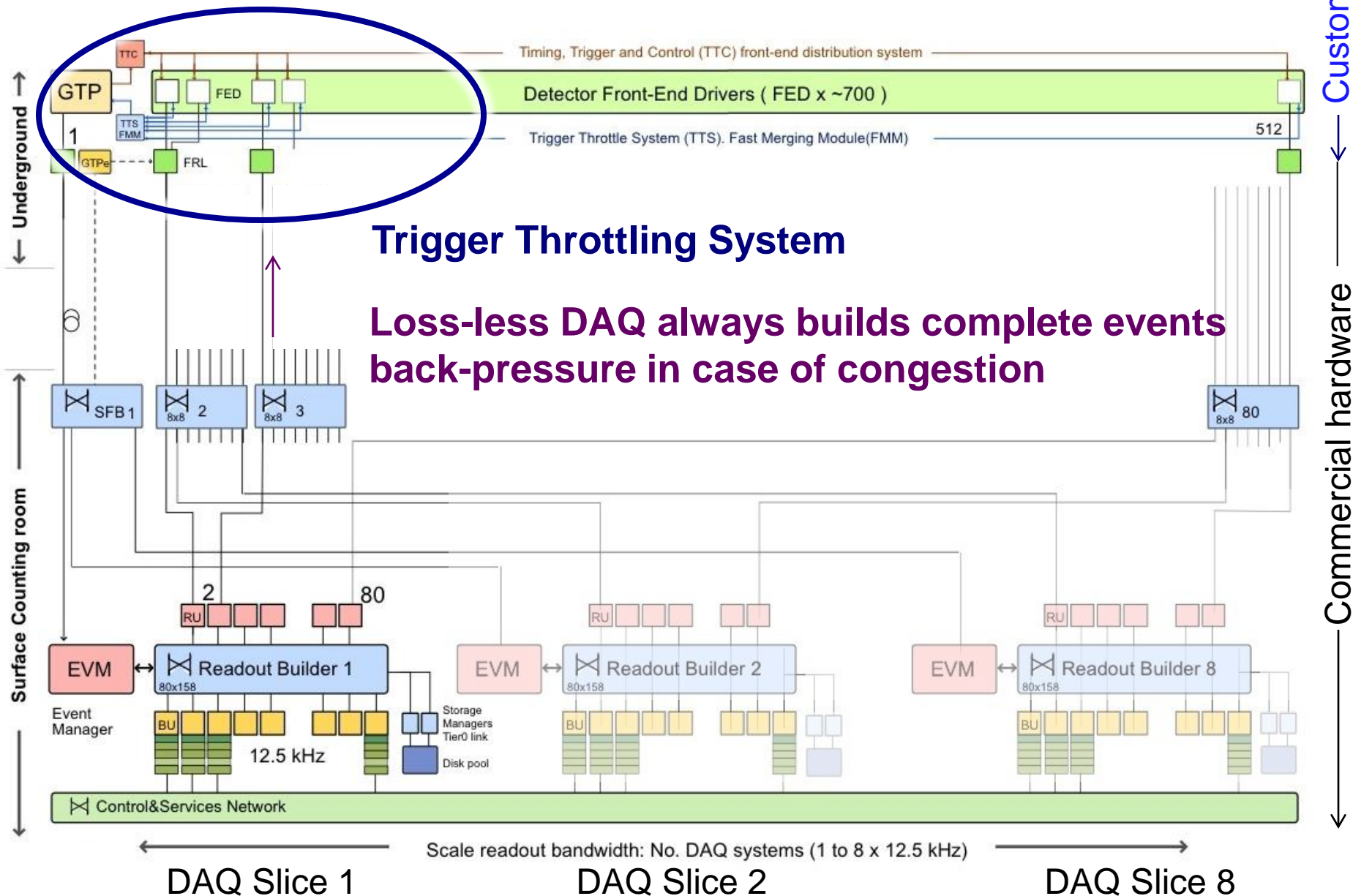


CMS two-stage event builder

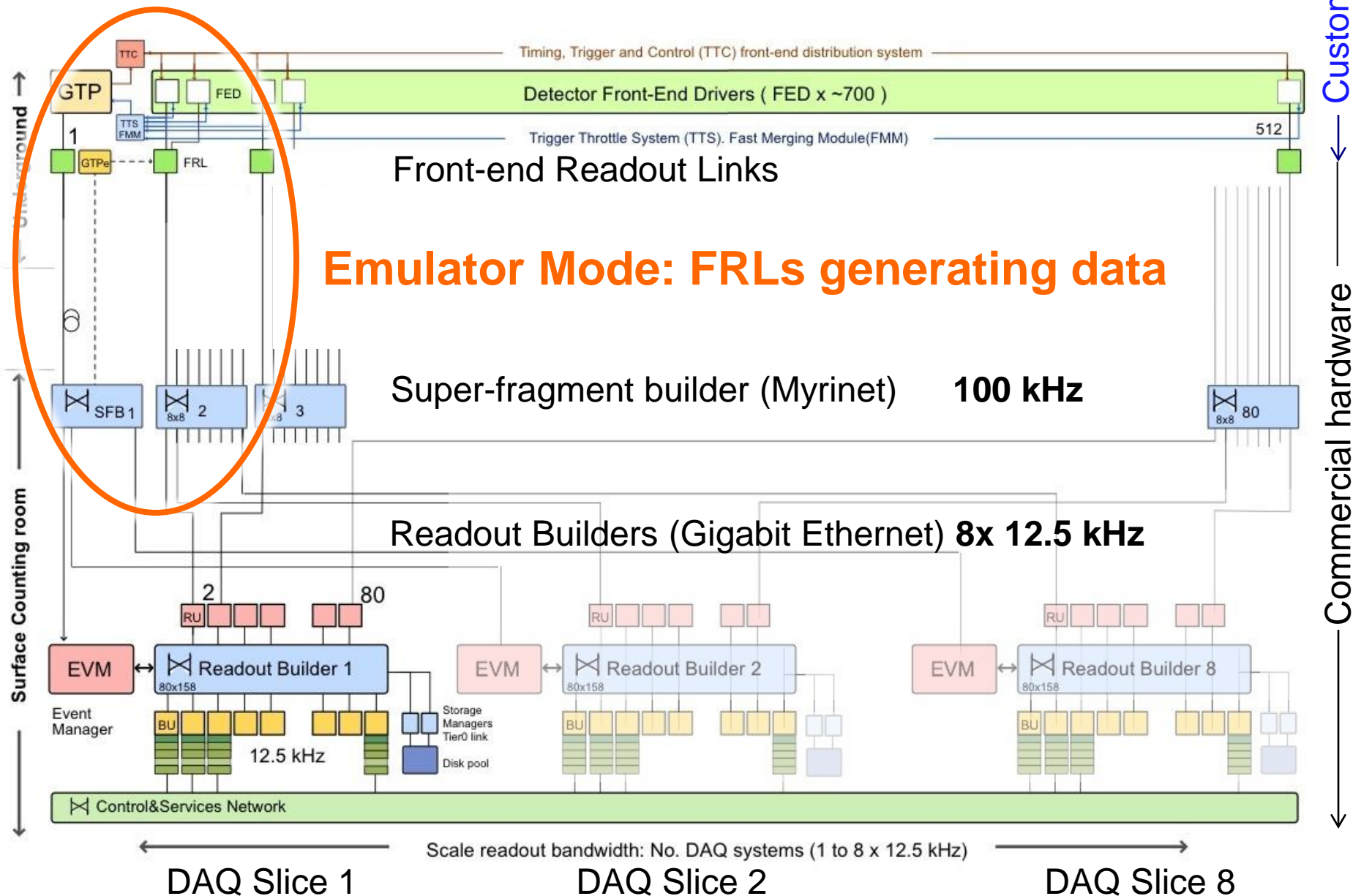




CMS two-stage event builder

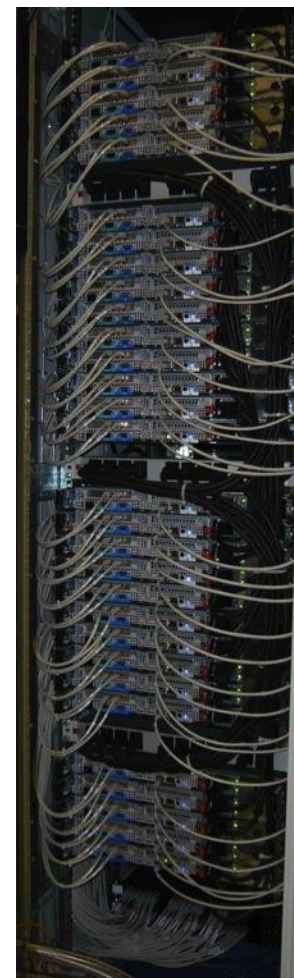


CMS two-stage event builder



Installed hardware

- Custom compact PCI Modules
 - 512 Frontend Readout Links
 - 60 Fast merging modules (trigger throttling)
- Myrinet Switches
 - 12 clos-256 enclosures
 - 1536 2.5 Gb/s links underground to surface
- “Readout Unit” PC nodes
 - 640 times dual 4-core E5130 (2007)
 - Each node has 3 links to GbE switch
- Gbe Switches
 - 8 times F10 E1200 routers
 - In total ~4000 ports (1 Gb/s)
- Event builder–output + HLT nodes (“BU-FU”)
 - Currently ~13000 cores, 26 TB RAM
 - Extensible – see later
- Storage Manager
 - 16 PCs
 - Storage Area Network (NexSan SataBeasts), 300 TB
 - 2.1 GB/s write speed (2.6 GB/s w/o Tier0-Transfers)





CMS DAQ Software

Run Control System – Java, Web Technologies

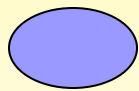
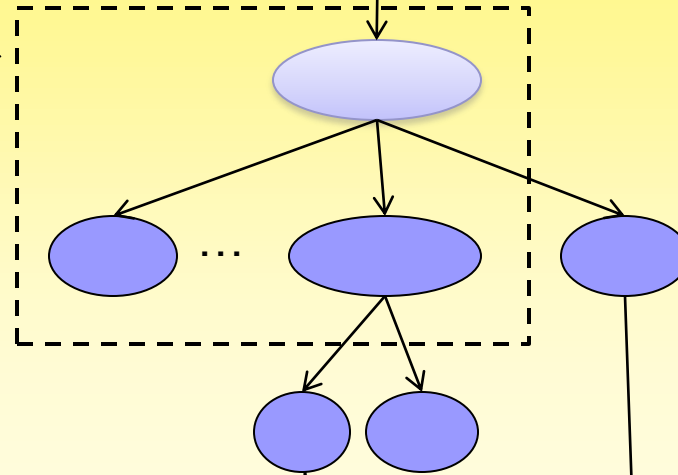
Defines the control structure



GUI in a web browser

HTML, CSS, JavaScript, AJAX

Run Control Web Application
Apache Tomcat Servlet Container
Java Server Pages, Tag Libraries,
Web Services (WSDL, Axis, SOAP)



Function Manager

Node in the Run Control Tree
defines a State Machine & parameters
User function managers dynamically
loaded into the web application

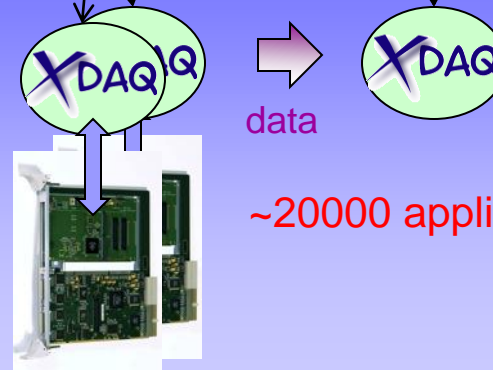
XDAQ Framework – C++, XML, SOAP

XDAQ applications control hardware and data flow

XDAQ is the framework of CMS online software
It provides Hardware Access, Transport Protocols,
Services etc.



XDAQ Application





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)
- Control of individual sub-systems for fast recovery
- Cross-checks and warnings to help the DAQ shifter

Running 00:53.7

Configuration: /approPublicGlobal/levelZeroFM

Run Number: **147003**

SID: 149234

Seq Name: GLOBAL-RUN

Global Key: /GLOBAL_CONFIGURATION_MAP/CMS/CENTRAL/GLOBAL_RUN

HLT Key from trigger mode: /daq/special/CastorTestHLT_BasicV8

L1 Trigger Key from trigger mode: L1_20101001_100925_2613

Clock source: LHC => MI KEY: beam1-manual

HWCFG Key: /cms/eq_100923/RUN_2010/eq_sl_rev100923/eq_slx_b43_84BU1MFU_16SM-0

Level-0 Action: Tasks completed

Level-0 Error:

CMS is configured with LHC clock but LHC clock stability is no longer / not yet guaranteed.
NEED TO RE-CONFIGURE ALL OF CMS WITH LOCAL CLOCK.

Subsystem	ECAL	ES	HCAL	[LUMI]	TRACKER	TRG	DT	CSC	RPC	DAQ
State	Running	Running	Running	Running	Running	Running	Running	Running	Running	Running
Time:	00:06.3	00:02.0	00:12.0		00:46.9	00:06.2	00:10.2	00:06.2	00:00.4	00:05.7

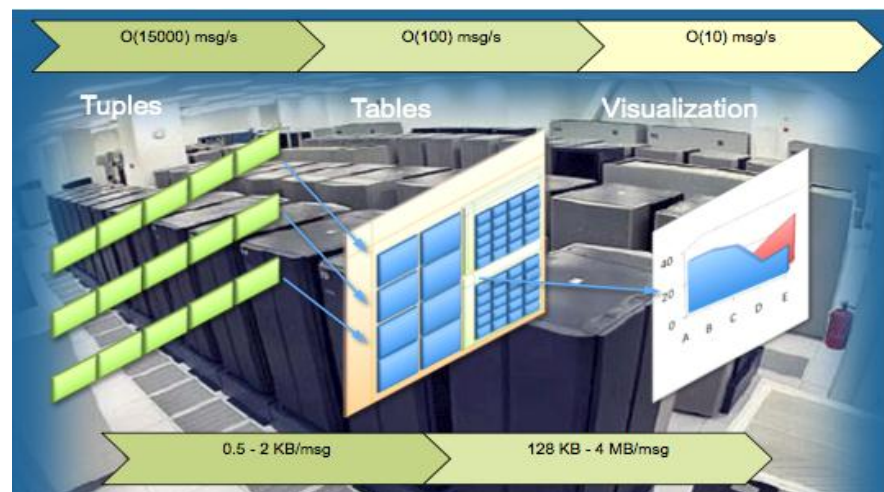
Enabled Slices: [Buttons]

Run Key: [Buttons]

Commander: [Buttons]

Monitoring

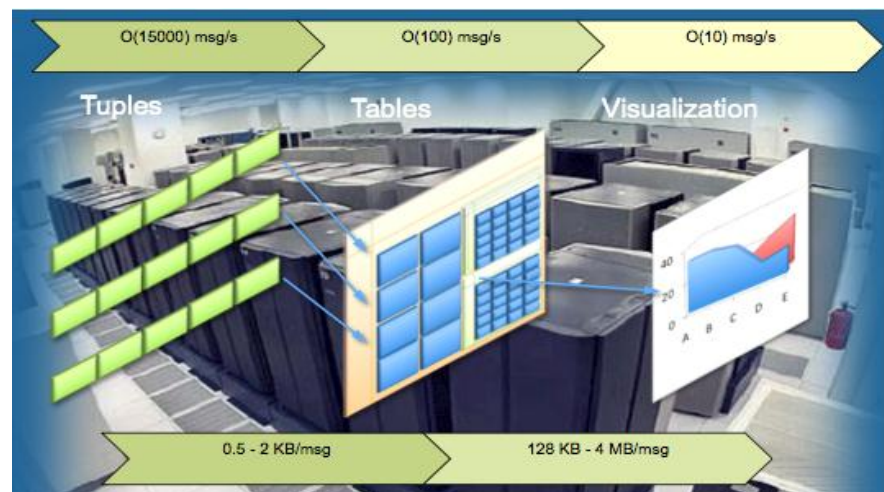
- Monitoring tuples and error messages
 - O(2000) PCs
 - O(20000) applications
- Collect and aggregate
 - Hierarchy of collectors
 - Load balancing
 - Latency ~seconds
- Access service for
 - Error reporting GUI
 - Visualization applications
 - DAQ Doctor (“expert system”)



Poster #139 / session 2: Distributed error and alarm processing in the CMS data acquisition system

The DAQ doctor

- Constantly analyzes monitoring information
- Detects abnormal situations
 - ☐ Warns the shift crew with Text & Audio alerts
 - ☐ Gives recovery instructions
 - ☐ Now also creates new DAQ configurations
 - ☐ Dumps diagnostic info for post-mortem analysis
- All diagnostic information is archived & categorized by sub-system



An extract of the DaqDoctors Notes for the year 2012

Note that not all subsystem errors are contained in this list. For a full error list browse the [archive of the DaqDoctor!](#)

Date	LHC state	System	Error	Remarks
2012-05-18 18:18:26	STABLE BEAMS	PIXEL	FSMShutdown	oldstate: Faulty
2012-05-17 17:48:28	SETUP	PIXEL	FSMShutdown	oldstate: Initializing
2012-05-17 03:39:08	STABLE BEAMS	PIXEL	FSMShutdown	oldstate: FixingSoftware
2012-05-17 03:39:23	STABLE BEAMS	PIXEL	FSMShutdown	oldstate: FixingSoftware
2012-05-14 13:50:20	SETUP	PIXEL	FSMShutdown	oldstate: Configuring
2012-05-14 13:50:20	SETUP	PIXEL	FSMShutdown	oldstate: Configuring
2012-05-14 13:48:32	RAMP DOWN	PIXEL	FSMShutdown	oldstate: Initializing
2012-05-14 13:47:37	RAMP DOWN	PIXEL	FSMShutdown	oldstate: Halted
2012-05-08 14:31:41	SETUP	PIXEL	SyncLostDraining	37, 37, 38, 33, 39, 32, 35, 36, 26, 29, 25, 23, 20, 30, 17, 18, 24, 31, 28, 16, 19, 27, 5, 9, 14, 15, 3, 0, 10, 13, 7, 4, 1, 2, 8, 12, 6, 11
2012-05-07 01:39:50	SETUP	PIXEL	FSMShutdown	oldstate: Initializing
2012-05-06 23:17:12	PREPARE RAMP	PIXEL	FSMShutdown	oldstate: Pausing
2012-05-02 18:04:37	INJECTION PHYSICS BEAM	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-28 00:32:32	NO BEAM	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-27 15:23:59	NO BEAM	PIXEL	FSMShutdown	oldstate: Starting
2012-04-27 15:19:38	NO BEAM	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-20 15:22:08	SETUP	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-19 16:09:52	SETUP	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-19 12:56:32	SETUP	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-19 00:49:53	STABLE BEAMS	RPX	TTSystem	1,20
2012-04-19 00:48:30	STABLE BEAMS	RPX	TTSystem	1,20
2012-04-19 00:48:06	STABLE BEAMS	RPX	TTSystem	1,20
2012-04-19 00:47:30	STABLE BEAMS	RPX	TTSystem	1,20
2012-04-19 00:46:36	STABLE BEAMS	RPX	TTSystem	1,20
2012-04-18 20:30:42	STABLE BEAMS	PIXEL	SyncLostDraining	37, 37, 38, 33, 39, 32, 35
2012-04-15 12:54:52	SETUP	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-15 09:44:28	SETUP	PIXEL	FSMShutdown	oldstate: Initializing
2012-04-12 23:36:13	SETUP	PIXEL	FSMShutdown	oldstate: Initializing

System status display

01/05/11 Sun 02:43 | LHC: STABLE BEAMS | DAQ: Running, Run#163761 | EvSize 373.8 kB, Rate 27.321 kHz(9869.1 MB/s), DeadTime(AB) 0.572 % | #HLT 140730630, <Acc>27416.9Hz (100.0%), <CPU> 16.01%



Read Builders

Slice	RunBU	kHz	CR/s	#Acc	#Runnin	CPU %
0	69x123	3.352	1.224	175909	993	16.118
1	69x125	3.328	1.223	175913	1015	16.083
2	69x124	3.340	1.222	175911	1004	15.843
3	69x126	3.353	1.225	175911	1026	15.138
4	69x124	3.347	1.221	175909	1004	15.896
5	69x121	3.379	1.221	175911	979	16.600
6	69x120	3.318	1.222	175912	960	16.736
7	69x125	3.358	1.224	175915	1015	15.858
Totals	552x98	26.775	9.762	140729	7996	16.009

SM-Stream	No.Events	Rate (Hz)	BW (MB/s)
FaultyEvents_Nor	0.000E+0	0.00	0.00
A	1.424E+6	281.28	55.30

Monitors

Enabled	App. states
Enabled	MyriNet/FBIO
Enabled	FU states
Disabled	Flash plots
Enabled	Data Flow
Enabled	FMM/FED
Enabled	Faults

Application states

failed	failed
halted	halted
enabled	enabled
ready	ready
mismatch	mismatch
time-out/tolerating	time-out/tolerating
configuring	configuring
lost/synch	lost/synch
stopping	stopping

----> Subsystem ERROR/FAULTY: TRG

Data to Surface	
CRG	Sub-System
TRG	3 3 3
CSC	9 9 9
DAQ	0 0 0
DQM	0 0 0
DT	6 6 6
ECAL	54 54 54
ES	39 39 39
HCAL	26 26 26
HFLUMI	6 6 6
PIXEL	40 40 40
RPC	3 3 3
SCAL	1 1 1
TRACKER	438 250 438
CASTOR	3 3 3

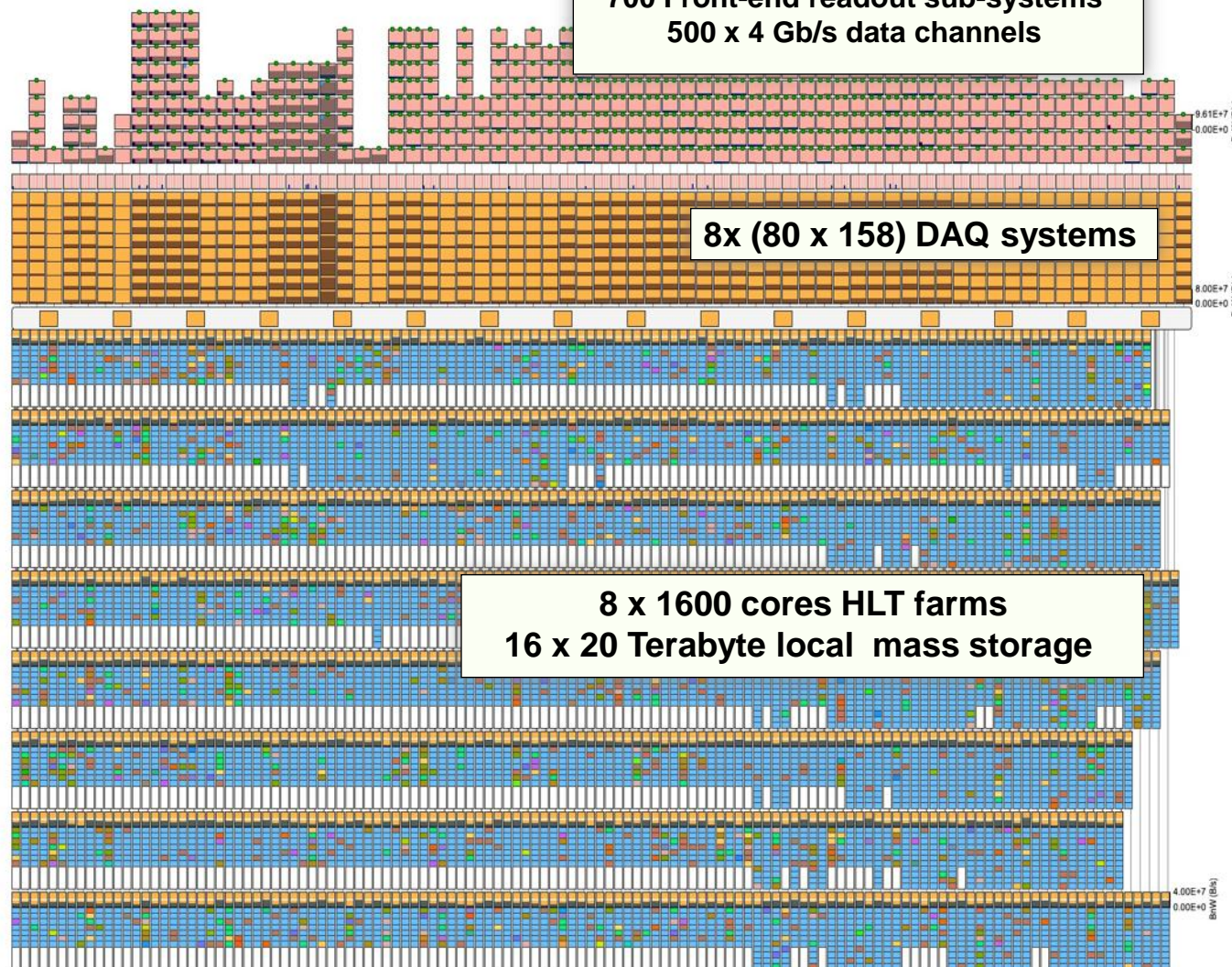
Flash list summary	
Totals	FMM FED FRL EVM RU BU mFU SM
# inFlash	1279 628 440 8 544 988 0868 16
# Enabled	779 628 440 8 544 988 0868 16
# Missing/OldStamp	0 0 0 0 0 0 0 0
Flash Dts(s)	8 3 4 3 4 4 3 1
Latency (s)	4 3 2 4 4 4 6 5

Sub-Sys	MB/s	Tot (kB)	<#Bytes>	Max#B	FED-ID	Min#B
TRG	105.4	3.9	1338	2040	745	408
CSC	290.4	10.8	1230	1889	760	1013
DT	61.4	2.3	390	475	770	49
ECAL	1504.5	56.0	1061	1289	623	874
ES	342.0	12.7	334	708	554	226
HCAL	997.2	37.0	1459	1740	714	977
HFLUMI	483.2	18.0	3071	3437	722	2916
PIXEL	518.0	19.2	493	631	23	364
RPC	3.0	0.1	38	39	793	38
SCAL	33.9	1.3	1296	1296	735	1296
TRACKER	5522.5	205.3	480	621	53	280
CASTOR	145.6	5.4	1851	1960	691	1736
Totals	10007	372.1	1087			

700 Front-end readout sub-systems
500 x 4 Gb/s data channels

8x (80 x 158) DAQ systems

8 x 1600 cores HLT farms
16 x 20 Terabyte local mass storage





Data acquisition in operation



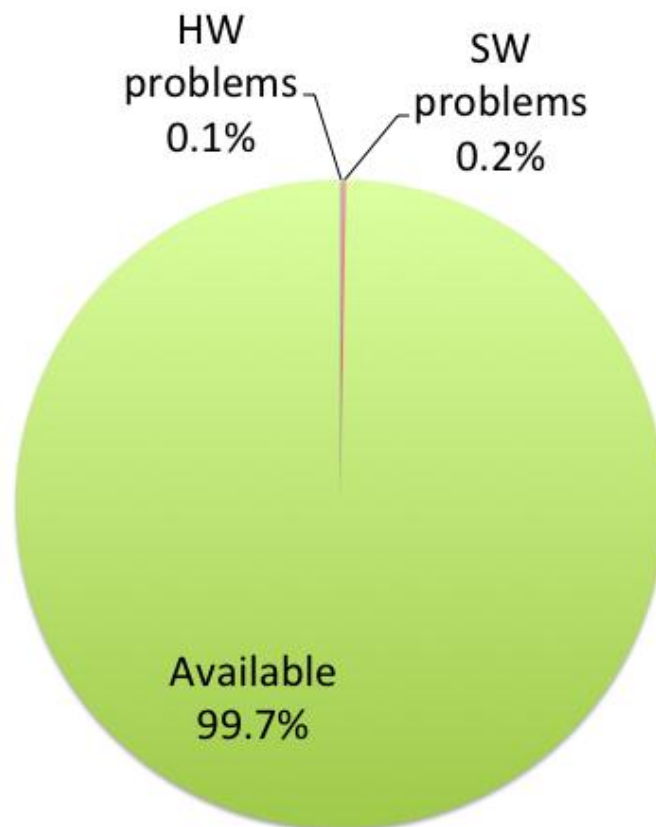


Operational Efficiency

CMS control room, Cessy, France ...



CMS Central DAQ efficiency, 2011 - pp



CMS central DAQ availability during stable beams: 99.7 %

CDAQ down time : < 4 hours

Luminosity lost: ~ 0.5% of delivered



Central DAQ Down times

- Software (24 down times, 3 hours)
 - Due to surfacing and newly introduced bugs
 - Often related to features that were added to the original design
 - Usually fixed as soon as identified
- Hardware (8 down times, 1 hour)
 - 1 Broken Myrinet link
 - 1 Broken Gigabit-ethernet switch line card
 - 1 Broken control network switch
 - 203 PC failures

**Only 1 hour of down time due to HW ?
=> Resilience**



Resilience features of CMS DAQ

- **Automatic restart** of crashed Event Filter processes during an ongoing run
- **Tolerance** against crashed processes & machines
 - Data flow applications / machines
Builder & Filter Units, Storage Manager
run continues with reduced throughput
 - Applications controlling custom hardware
run continues with degraded monitoring
- **Slice Masking**: fast workaround for single points of failure in a DAQ Slice (Readout Units, GbE switches, etc.)
 - mask the slice and continue with 7/8th of capacity
 - requires stop/start of the run



Resilience features of CMS DAQ (2)

■ Fast Configuration Change

- ☐ Mask a broken machine
(except those controlling custom hardware)
- ☐ Mask a rail in one leg of the Myrinet Super-Fragment Builder
- ☐ Use only 1 out of 2 racks of Storage Managers

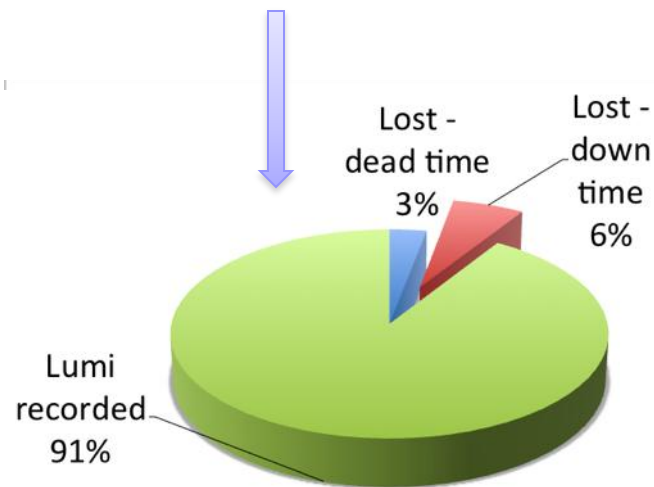
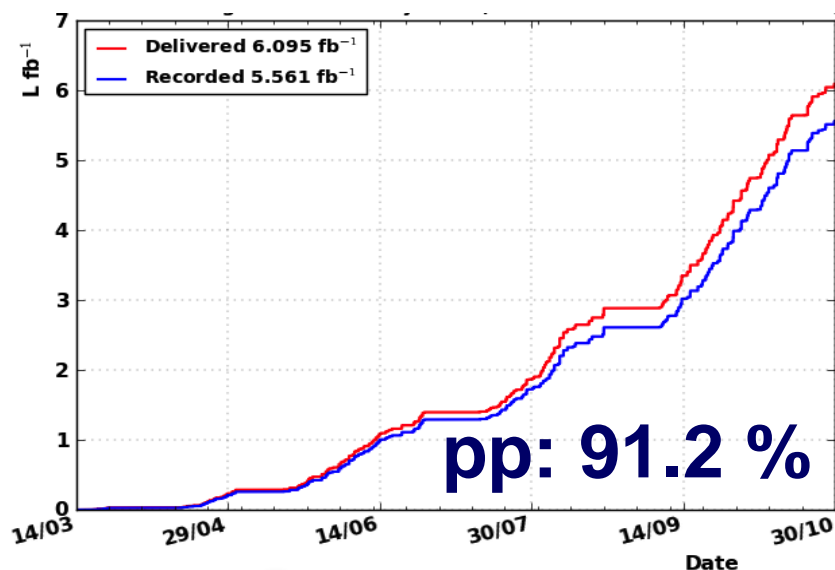
■ Tool: CMS DAQ Configurator

- ☐ Until mid 2010: Several tools needed, manual bookkeeping
new configuration in ~10 minutes
- ☐ mid 2010 - 2011: One-Step tool with blacklist database
new configuration in ~2 minutes
- ☐ Since 2012: One-Step tool automatically launched by DAQ Doctor
new configuration in ~ 40 seconds

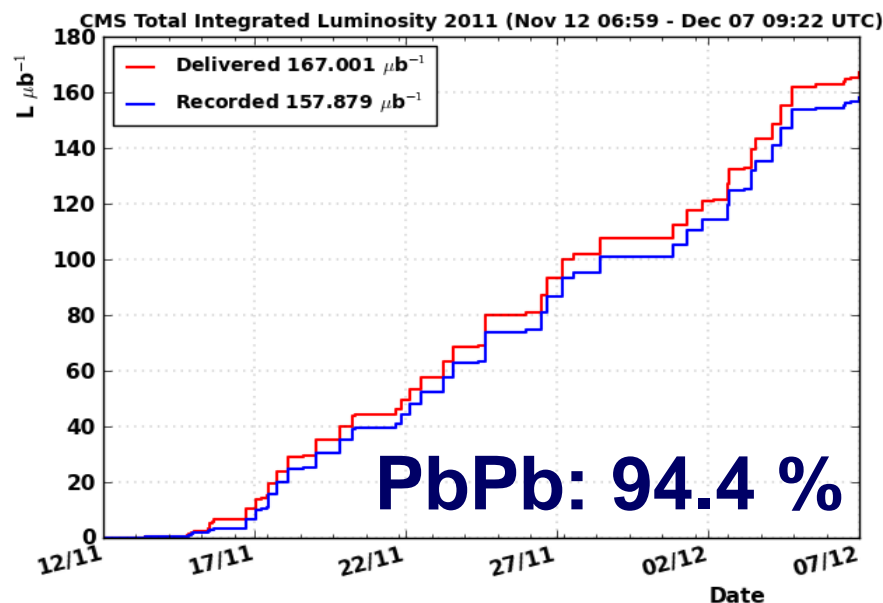
■ Configuration change requires a run stop/start



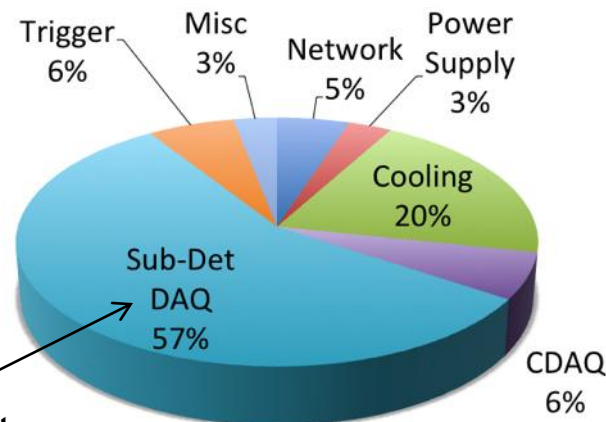
Over-all CMS data taking efficiency 2011



pp



In part due to
Single-Event upsets



Lumi lost in down times (pp)

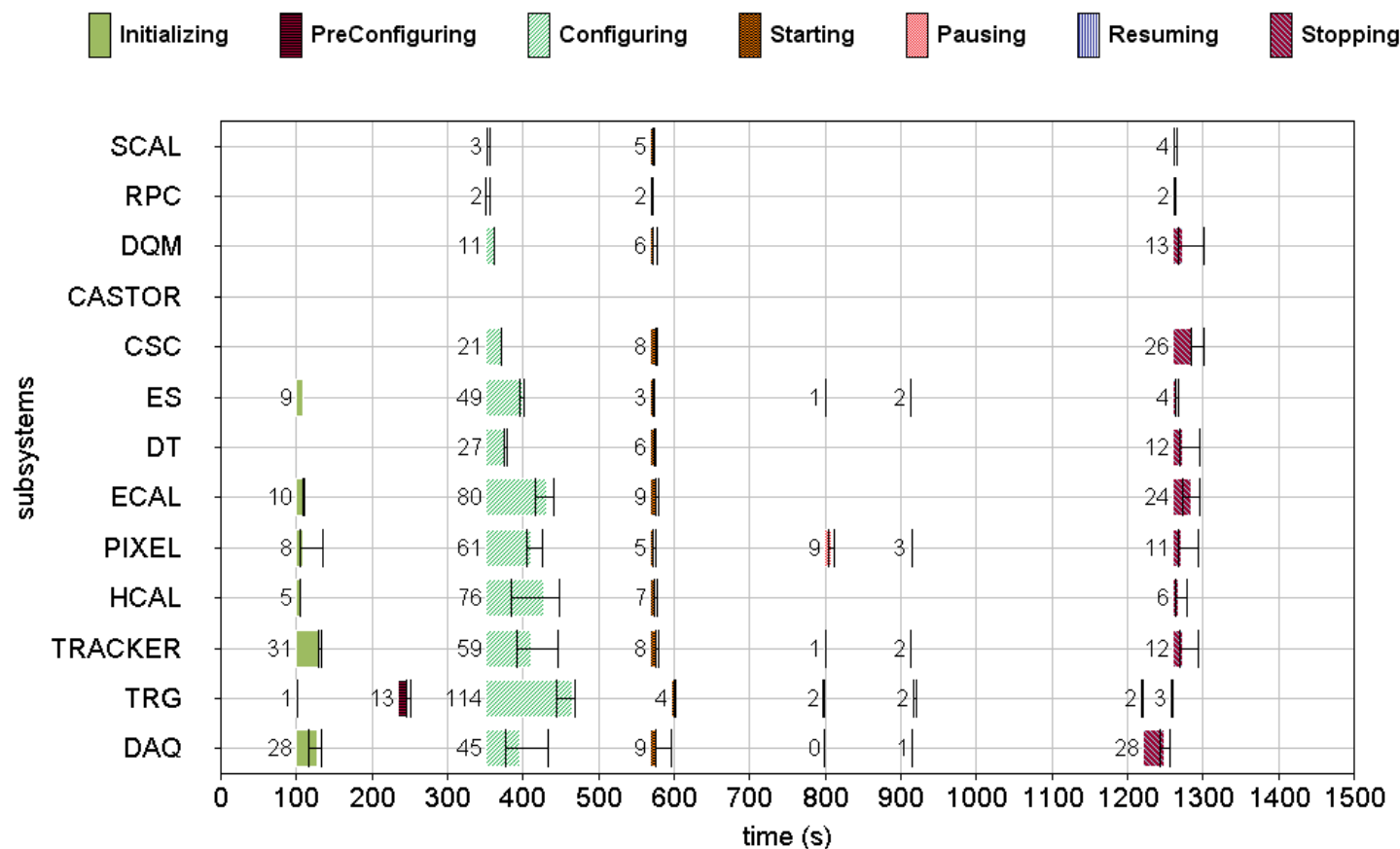


New: Automatic Recovery from Single Event Upsets

- Frequent sub-detector DAQ failures due to Single-Event upsets observed towards the end of 2011 with increasing instantaneous luminosity
- Recovery typically needed re-configuration of the system
- New in 2012: Automatic Single-event-upset Recovery Mechanism
 - Coordinated by top-level run control
 - Sub-detector detects SEU problem and notifies top-level run control
 - Top-level Run Control
 - Invokes a recovery transition
 - On the requesting sub-system
 - Other sub-systems may do preventive actions in the shadow



Impacting over-all efficiency: startup time



During stable beams, Apr 13 – May 2, 2012

- Start of data taking session (starts all software): < 3 minutes
- Run stop & start: 1 min 15 seconds



Evolution of operating conditions



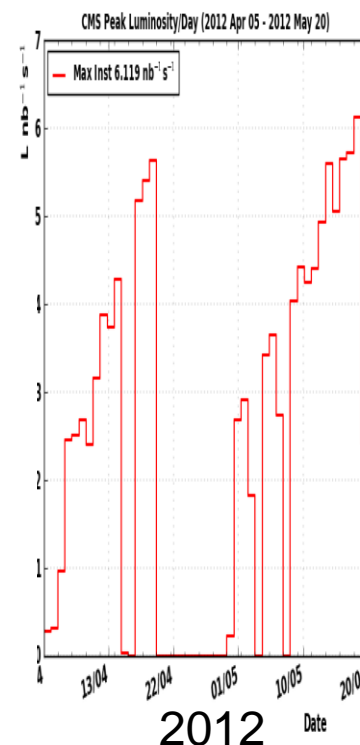
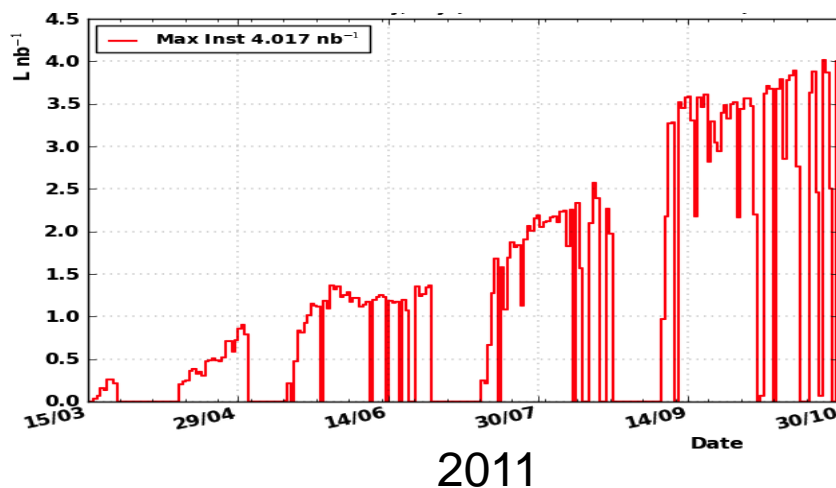
Evolution of operating conditions

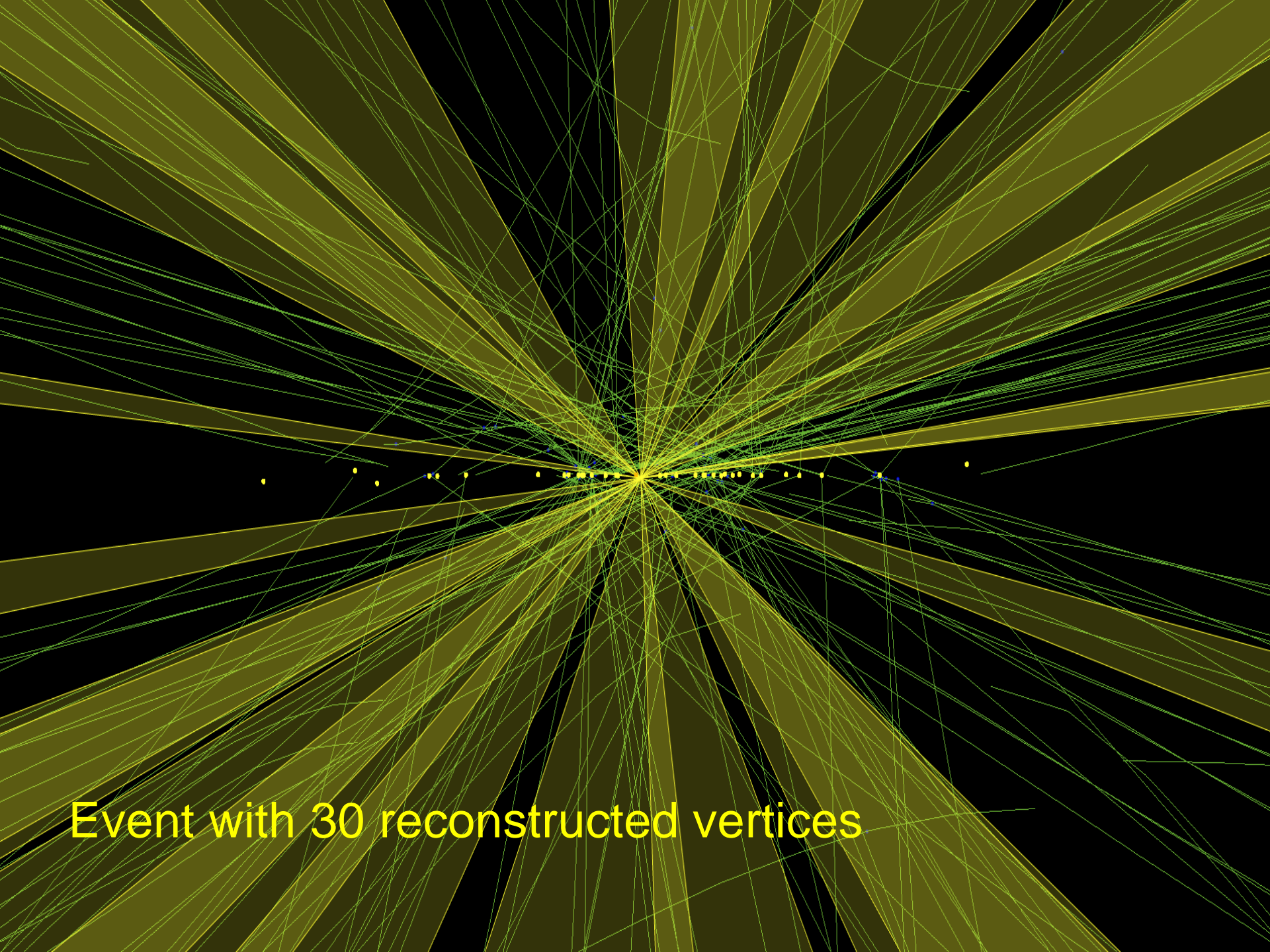
■ Design

- $L = 10^{34} / \text{cm}^2\text{s}$, 25 ns bunch spacing, 14 TeV
- Pile-up of 20
- DAQ at 100 kHz

■ 2012

- $L = 7 \times 10^{33} / \text{cm}^2\text{s}$ (expected),
50 ns bunch spacing, 8 TeV
- Pile-up of 35 (~2x design)
- DAQ at 100 kHz





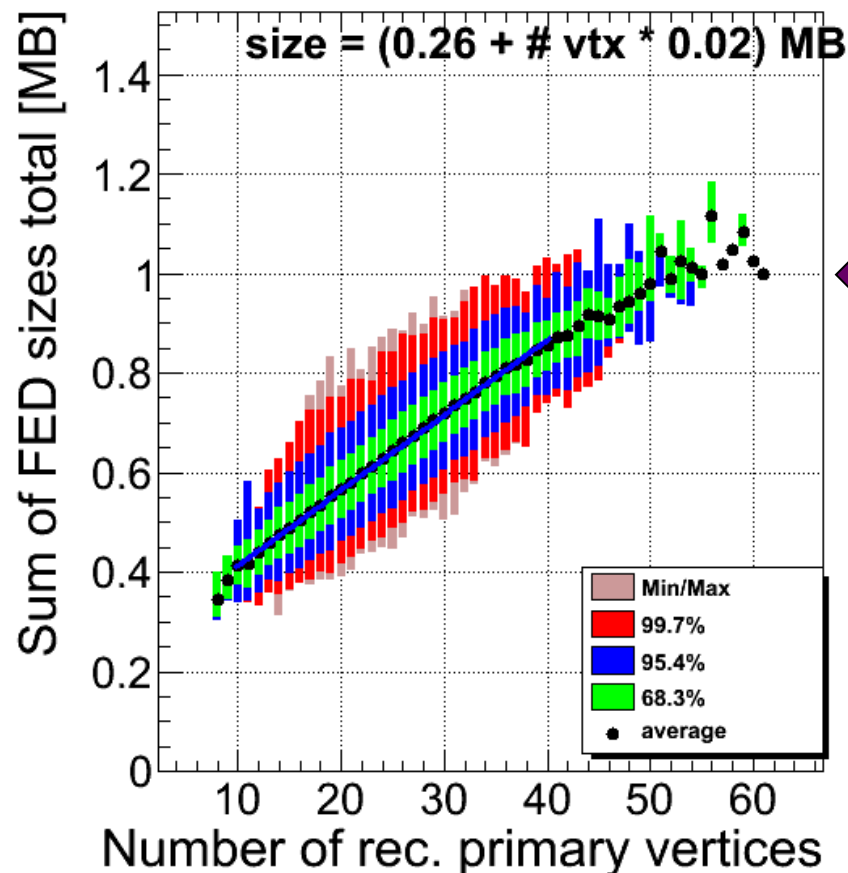
Event with 30 reconstructed vertices



Can we handle the event size?



Can we handle the event size ?



← CMS design event size: 1 MB

- Globally yes, but have to look at individual Inputs & Super-fragment builders

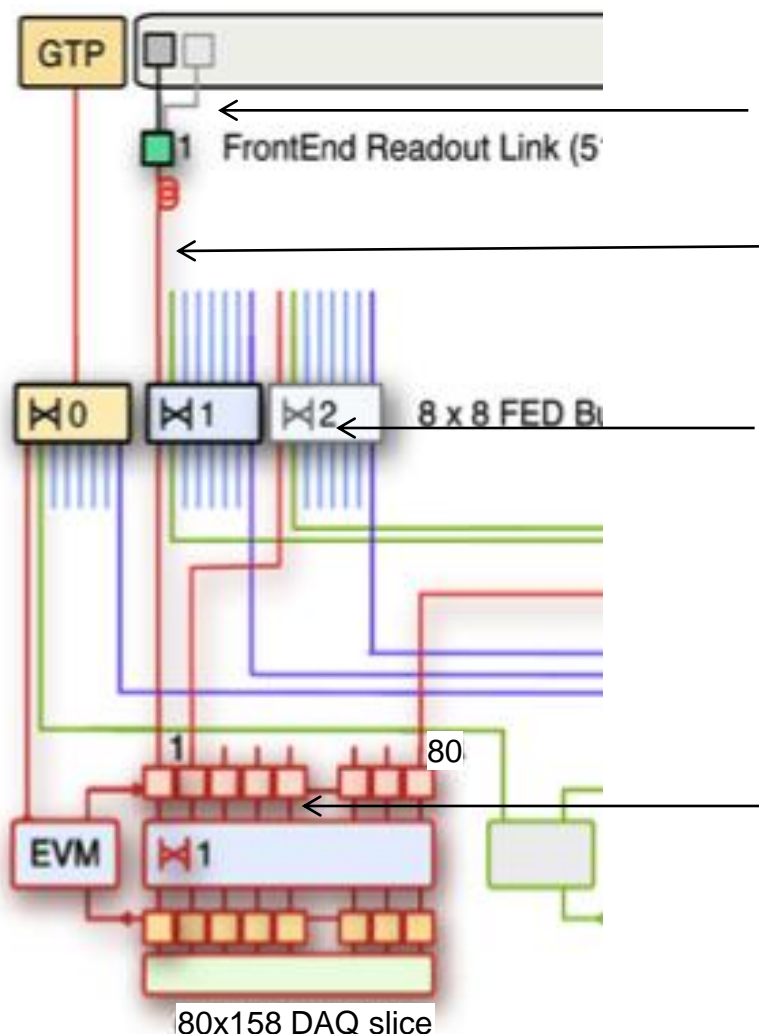
High pile-up fill
Run 179827

Expected 2012

30 primary vertices corresponding to pile-up of 35



Bandwidth at various stages



SLINK: **400 MB/s** (64b @ 50 MHz)

✓ No problem

Myrinet link: **500 MB/s** (2 rails of 2.5 Gbit/s)

✓ No problem

Myrinet Cross-bar switch: **~260 MB/s**

Wormhole-routed

No buffering in switch

Head-of line blocking reduces throughput by up to 50% when no traffic-shaping applied

Some Super-Fragment Builders critical

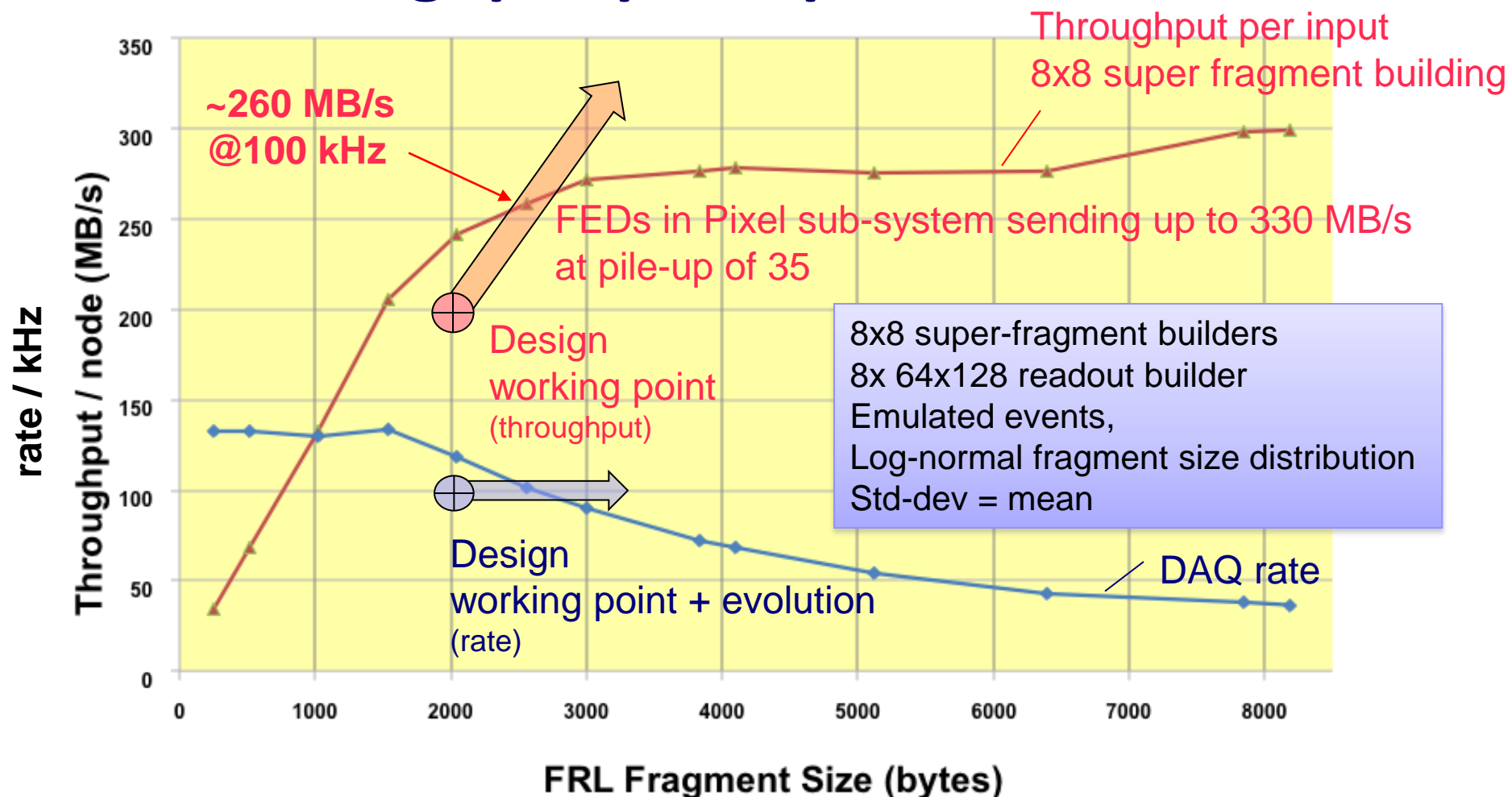
Gigabit Ethernet: 3 rails: **375 MB/s**

Ethernet switches have internal buffer shared memory – no HOL blocking

✓ No problem



DAQ throughput per input



- **32 inputs (Pixel sub-system) may exceed available throughput at pile-up of 35**

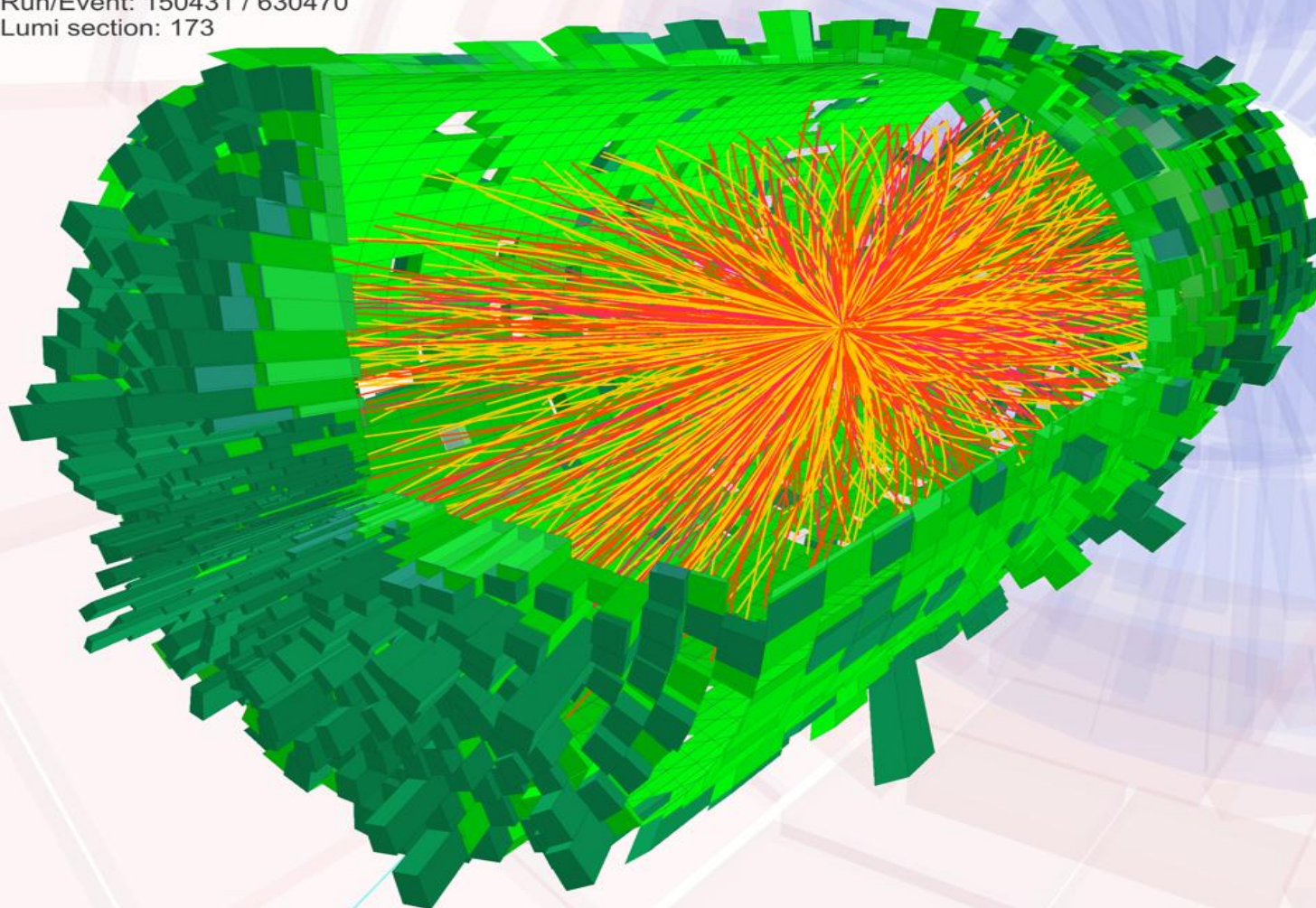
✓ Solution: super-fragment builders with fewer than 8 inputs for pixel combine some smaller super-fragment builders,



Throughput in Heavy-Ion Operation



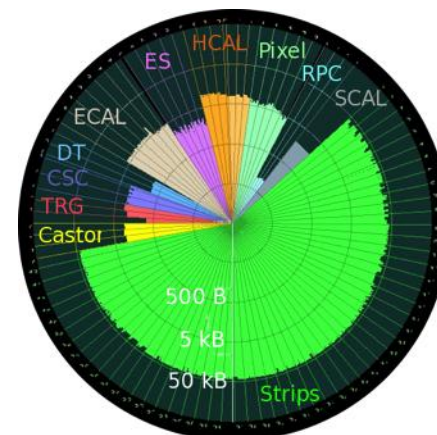
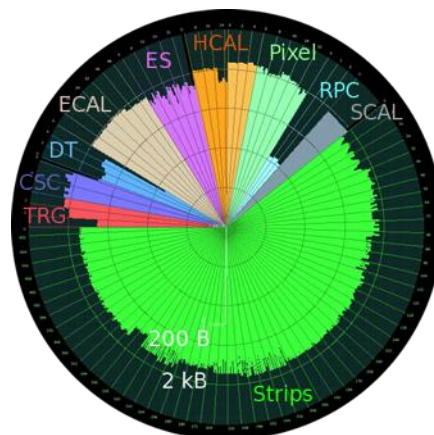
CMS Experiment at LHC, CERN
Data recorded: Mon Nov 8 11:30:53 2010 CEST
Run/Event: 150431 / 630470
Lumi section: 173





Proton physics – Ion physics

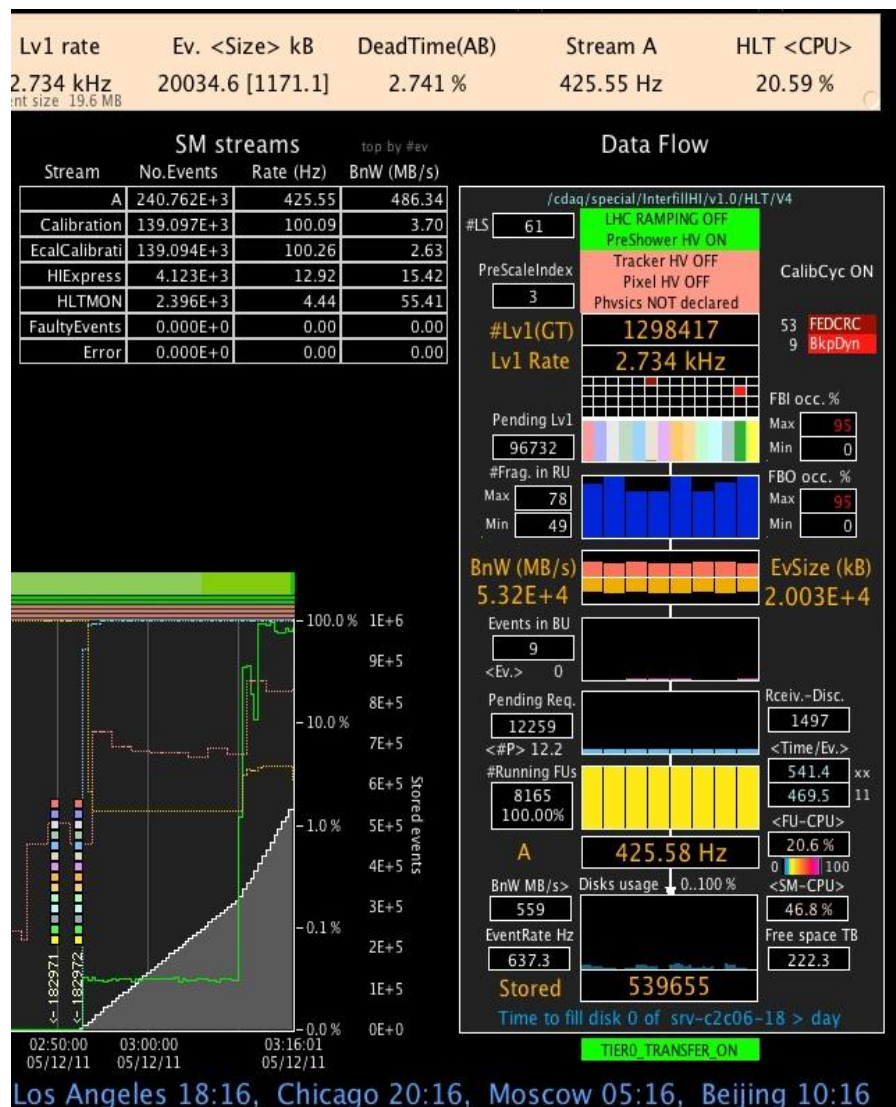
	Proton physics	Ion Physics
Zero suppression for Si-strip tracker	In FED (hardware)	In HLT farm (software)
Fragment size	2 kB	50 kB (100 kB after merging)
Event size	1 MB	20 MB
Max trigger rate	100 kHz	3.5 kHz
Max. DAQ throughput per input (8x8 super-fragment building)*	260 MB/s	350 MB/s (DAQ settings tuned for large fragments)



*log-normal distributed event size
std-dev = average



DAQ performance at start of 2011 HI fill



2.7 kHz L1 rate

20 MB / event

Zero-suppression in HLT farm -> 1MB

560 MB/s to disk

2010 HI run: ZS offline /
ROOT compression in HLT
11 MB / event, 1.8 GB/s to disk



High-Level Trigger

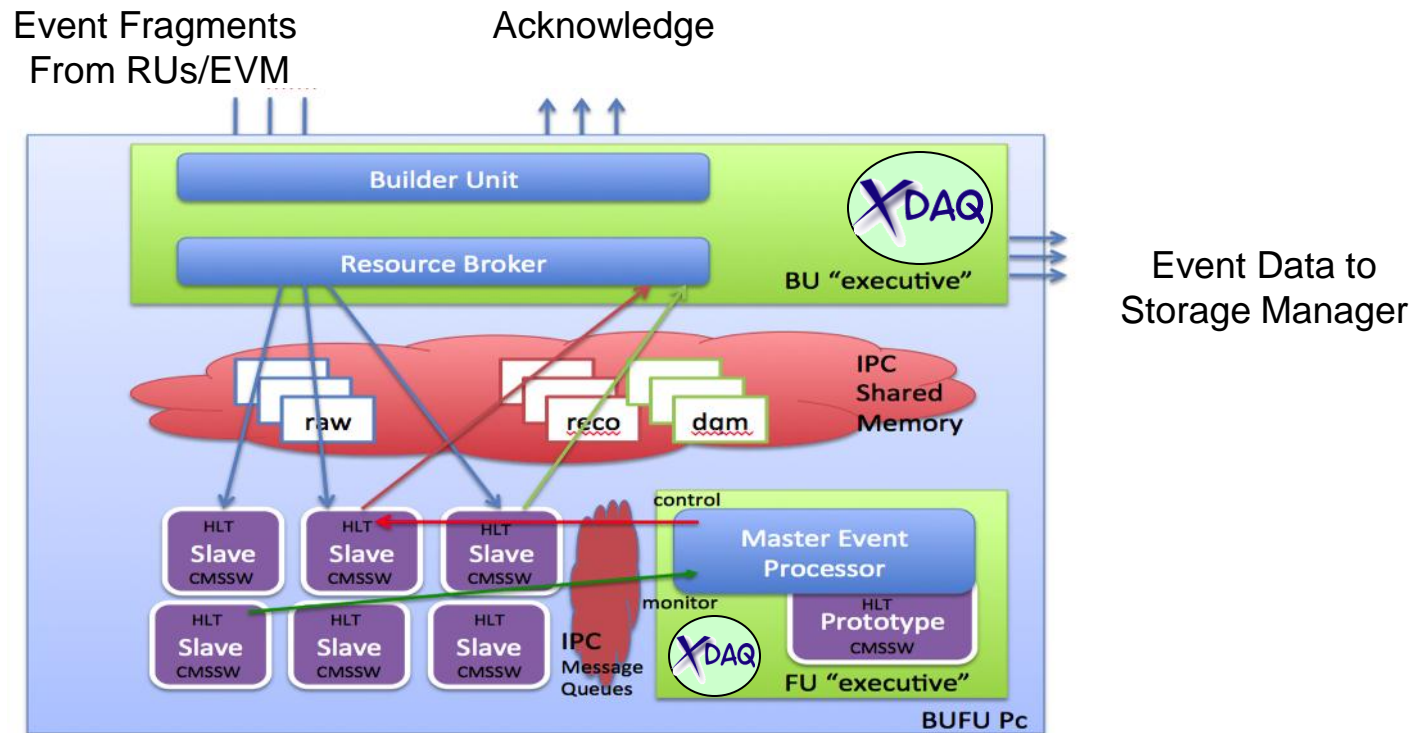


Filter Farm deployment strategy



- High-level trigger based entirely on commodity hardware
- Buy the processing power just in time
 - Better value for money
- Computing requirements evolve with LHC luminosity
 - Higher luminosity requires higher selectivity
→ more complex algorithms
 - Higher luminosity → more pile-up → more time consuming tracking
- Challenge: increasing number of cores per machine

High-Level Trigger Software



- Trigger algorithms are processed with CMS offline software framework CMSSW
 - 1 Process per core / per hyperthread but limited memory available
 - Copy On Write:
 - 1) Prototype process loads configuration and conditions
 - 2) Child processes are forked
 - Coupling between XDAQ and CMSSW very tight
 - same compiler, same process
- Poster #219 / session 2
CMS High Level Trigger
System: Experience and

Poster #219 / session 2: The CMS High Level Trigger System: Experience and Future Development

HLT farm evolution

2009:

720x



May 2011

add:

72x



May 2012

add:

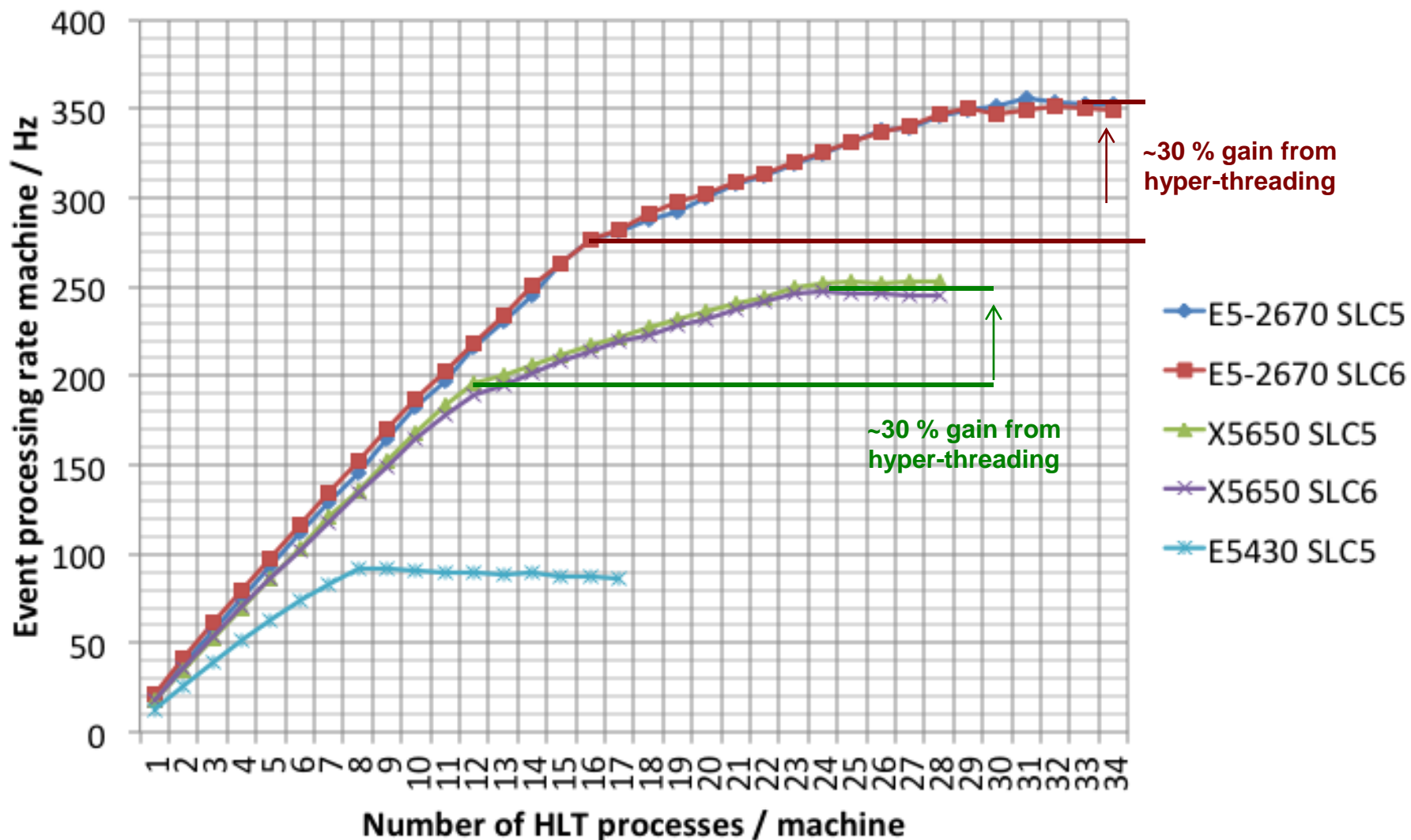
64x



	Original HLT System Dell Power Edge 1950	2011 extension Dell Power Edge c6100	2012 extension Dell Power Edge c6220
Form factor	1 motherboard in 1U box	4 motherboards in 2U box	4 motherboards in 2U box
CPU's per motherboard	2x 4-core Intel Xeon E5430 Harpertown , 2.66 GHz, 16GB RAM	2x 6-core Intel Xeon X5650 Westmere , 2.66 GHz, hyper-threading, 24 GB RAM	2x 8-core Intel Xeon E5-2670 Sandy Bridge , 2.6 GHz, hyper-threading, 32 GB RAM
#boxes	720	72 (=288 motherboards)	64 (=256 motherboards)
#cores	5760	3456 (+ hyper-threading)	4096 (+ hyper-threading)
cumulative #cores	5.6k	9.1k	13.2k
cumulative #CMSSW	5k	11k	20k



HLT machine performance with HLT playback



HLT menu for $5 \times 10^{33}/(\text{cm}^2\text{s})$, recent data sample & software

HLT farm evolution

2009:

720x



May 2011

add:

72x



May 2012

add:

64x



	Original HLT System Dell Power Edge 1950	2011 extension Dell Power Edge c6100	2012 extension Dell Power Edge c6220
Form factor	1 motherboard in 1U box	4 motherboards in 2U box	4 motherboards in 2U box
CPU's per motherboard	2x 4-core Intel Xeon E5430 Harpertown , 2.66 GHz, 16GB RAM	2x 6-core Intel Xeon X5650 Westmere , 2.66 GHz, hyper-threading, 24 GB RAM	2x 8-core Intel Xeon E5-2670 Sandy Bridge , 2.6 GHz, hyper-threading, 32 GB RAM
#boxes	720	72 (=288 motherboards)	64 (=256 motherboards)
#cores	5760	3456 (+ hyper-threading)	4096 (+ hyper-threading)
cumulative #cores	5.6k	9.1k	13.2k
cumulative #CMSSW	5k	11k	20k

Per-event
CPU budget
@ 100 kHz:

2009:

~50 ms / evt

2011:

~100 ms / evt

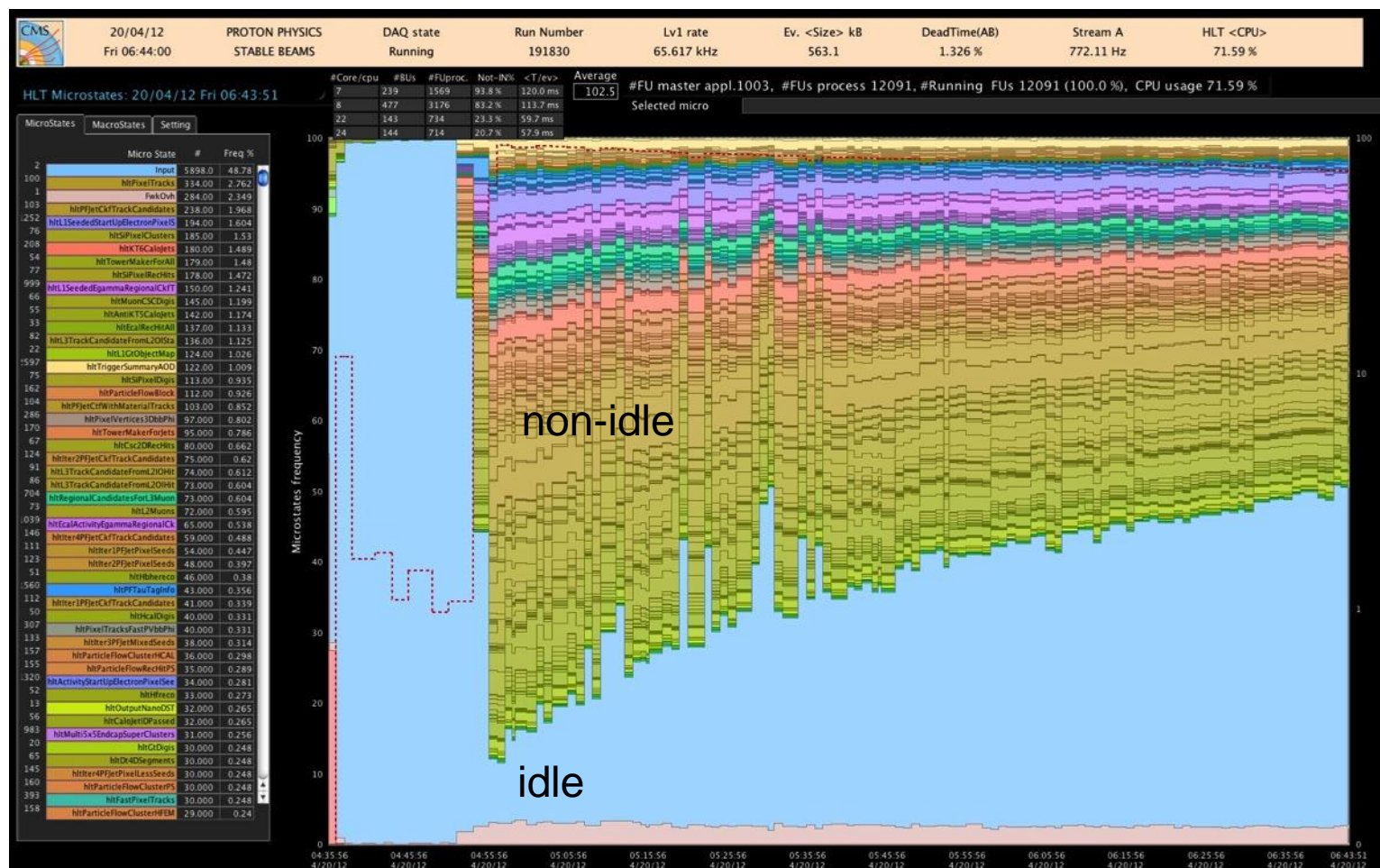
2012:

~150 ms / evt

(CPU budgets are on 1 core of an Intel Harpertown)



States of HLT nodes at start of a pp fill before extension 2



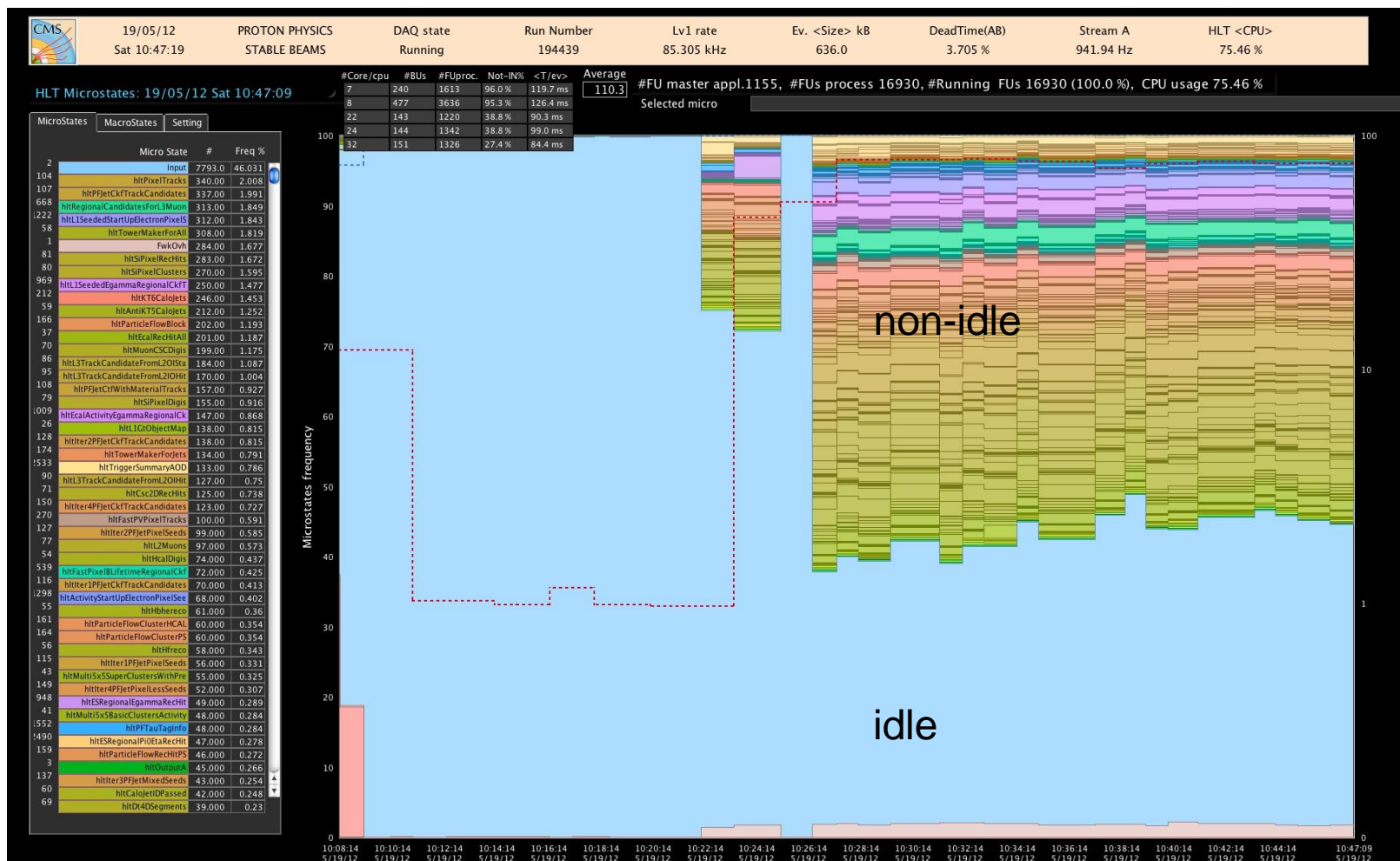
Time into fill

Fill 2536, 20 Apr 2012
 $L_{\text{peak}} = 6.1 \times 10^{33}/(\text{cm}^2\text{s})$

HLT farm almost fully utilized at start of fill (since September 2011)
 Algorithms are tuned for available computing power



HLT states with HLT extension 2



HLT extension-2 in 5 out of 8 DAQ slices

Time into fill

- Ready for higher instantaneous luminosity and more complex algorithms

Fill 2645, 19 May 2012
 $L_{\text{peak}} = 6.1 \times 10^{33} / (\text{cm}^2 \text{s})$



Summary

- CMS DAQ system building events at 100 kHz in 2 stages
 - 1MB event size, 100 GB/s throughput
- Central DAQ availability 2011: 99.7 %
- Continuous effort to improve CMS over-all efficiency
- Increased data volume due to higher pile-up with 50 ns LHC bunch spacing can be handled
- HLT farm being extended as required
 - reached 13000 cores this month. Ready for higher luminosity.

Beyond 2012: see next talk ...



Thank You



Bonus track



Comparison of HLT machines

	Harpertown	Westmere	Sandy Bridge
	Xeon E5430, 2.66 GHz	Xeon X5650, 2.66 GHz	Xeon E5-2670 2.6 GHz
#cores	8 (2x4)	12 (2x6) + HT	16 (2x8) + HT
SPEC int (max)	25	37 (= 25 * 1.5)	52 (= 25 * 2.1)
HEP Spec	73	208	386
CPU burner test*	1.0	3.6	5.4
Eg Action 11 test (CPU + memory)	1.0	2.2	3.3
HLT 2011	1.0	2.4	-
HLT playback*	1.0	2.8	3.9

Performance
per
motherboard

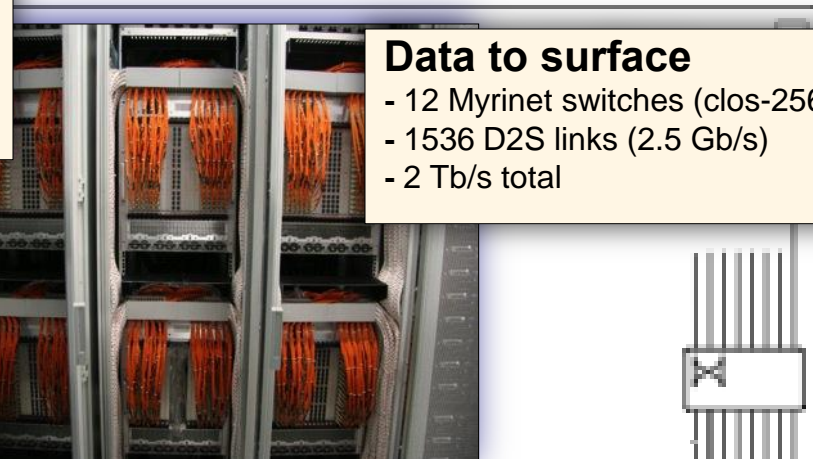
* Does not include event building

CMS DAQ installation



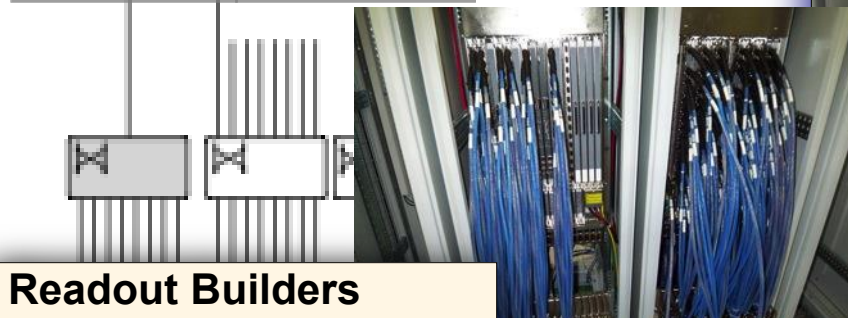
Detector readout.

- 650 Slink/FMM cables
- 500 FRL + 60 FMM modules
- 60 FRL/FMM crates
- 200 DAQ/DCS PCs



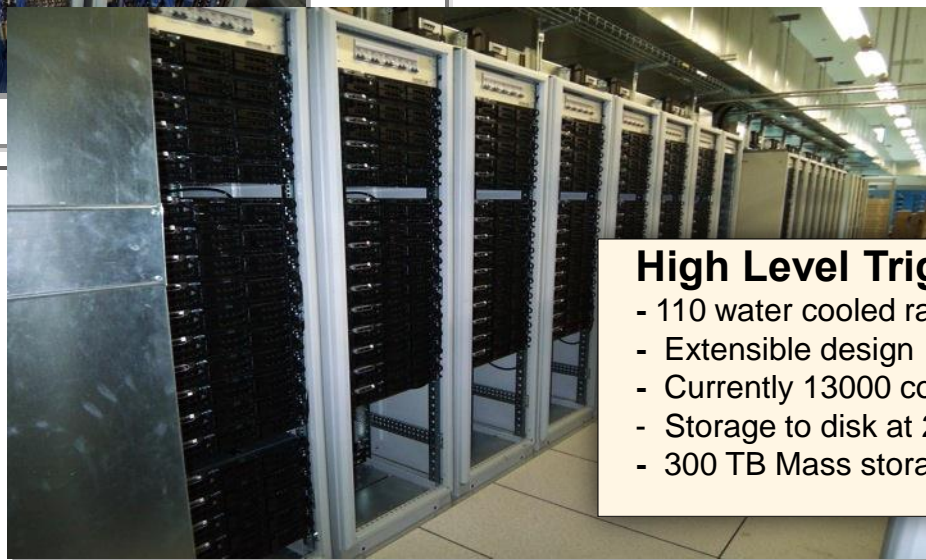
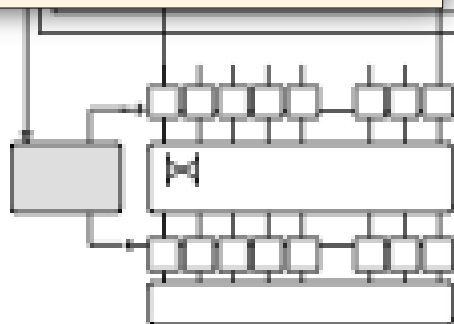
Data to surface

- 12 Myrinet switches (clos-256)
- 1536 D2S links (2.5 Gb/s)
- 2 Tb/s total



Readout Builders

- 640 Readout Unit PCs (4-core)
- 8 Force-10 GBE switches
4000 ports in total



High Level Trigger Farm

- 110 water cooled racks
- Extensible design
- Currently 13000 cores, 26 TB RAM
- Storage to disk at 2.1 GB/s
- 300 TB Mass storage

Experiment control and monitor system and WWW services

Cessy: Master&Command control room



Fermilab: Remote Operations Centre



Meyrin: CMS DQM Centre

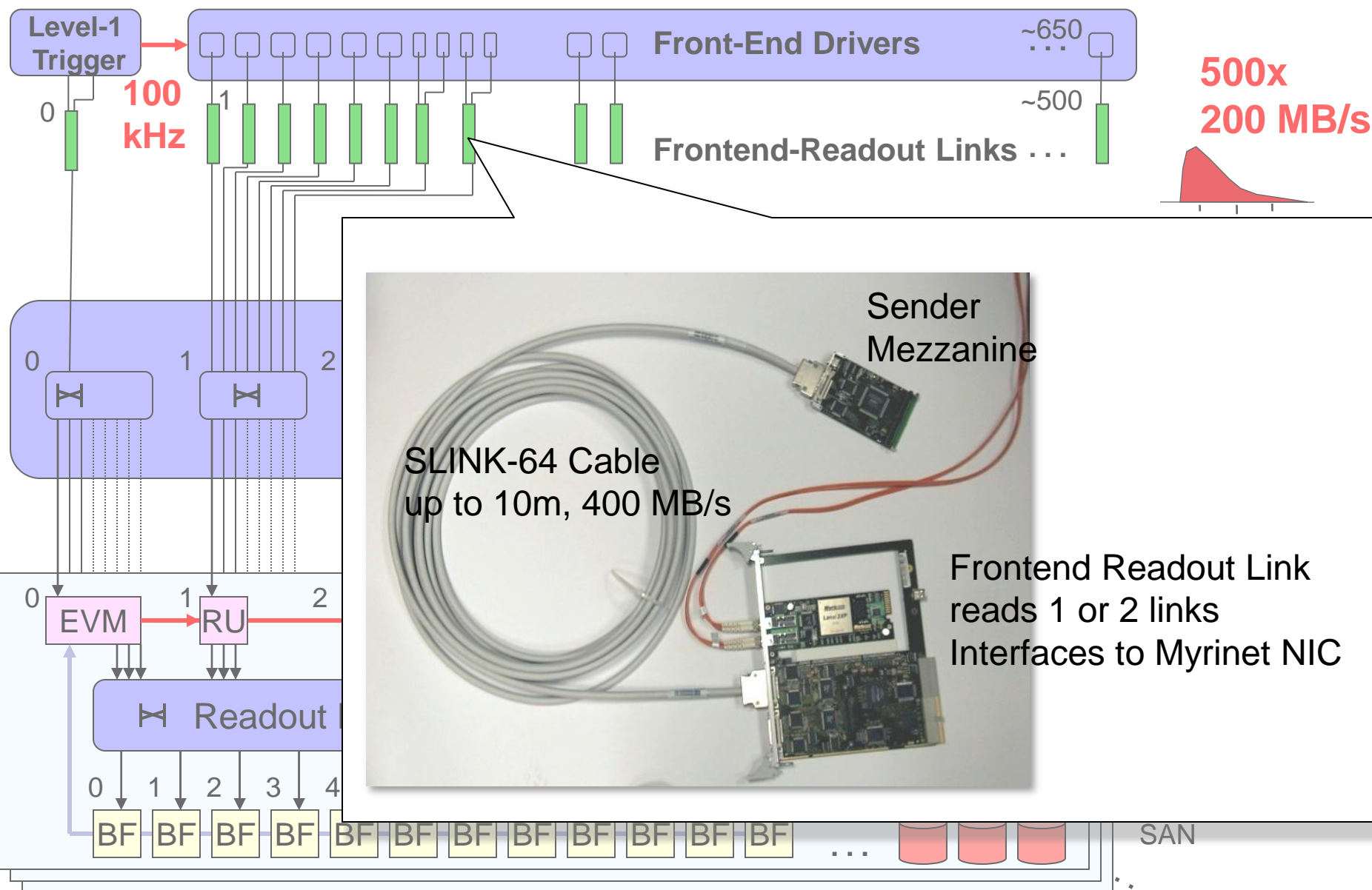


CR: Any Internet access.....

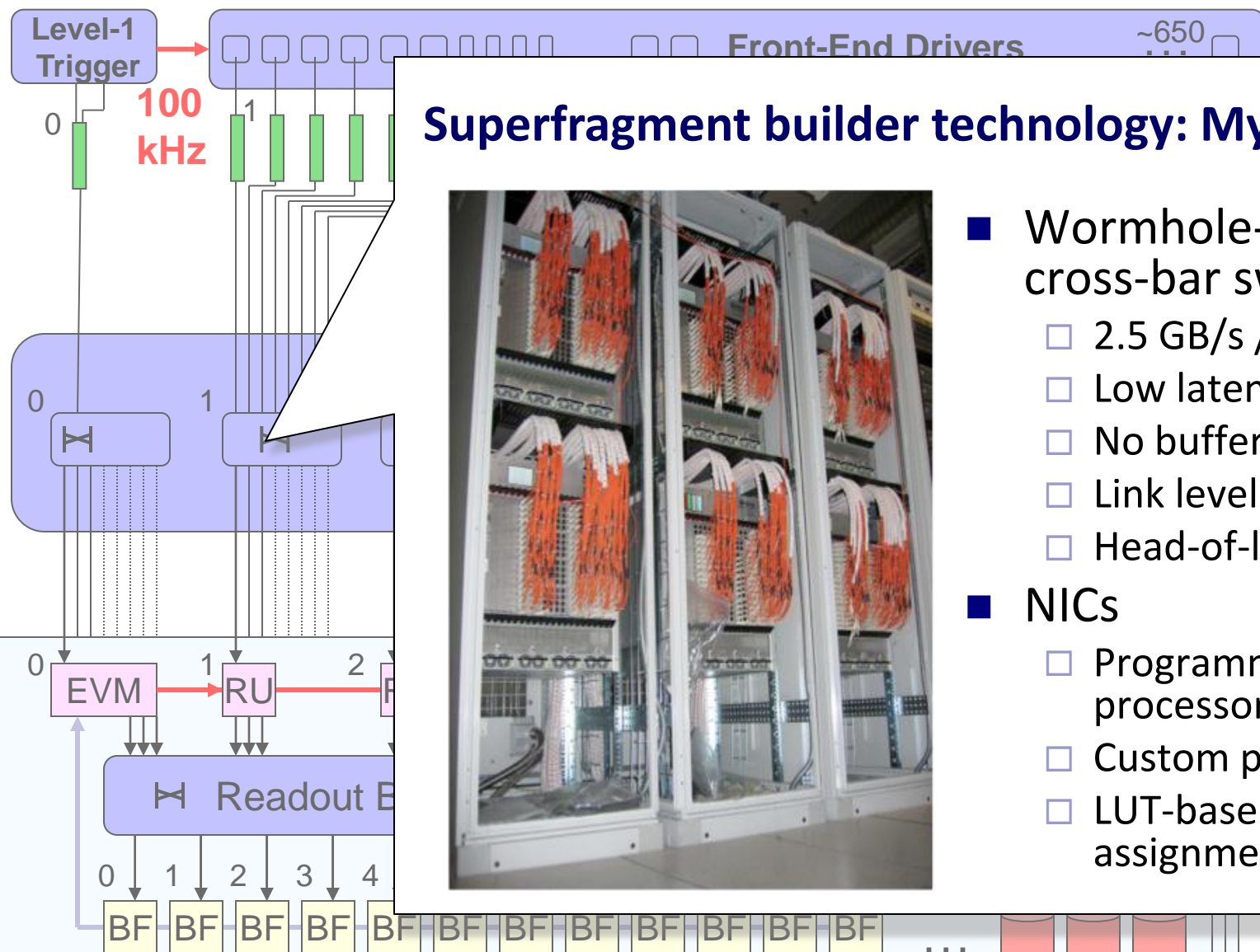


A general and expandable architecture has been deployed for the **experiments'**
Run control and monitoring largely based on the emerging Internet technology
developed in the field of **WWW services**

Two-stage event building architecture



Two-stage event building architecture



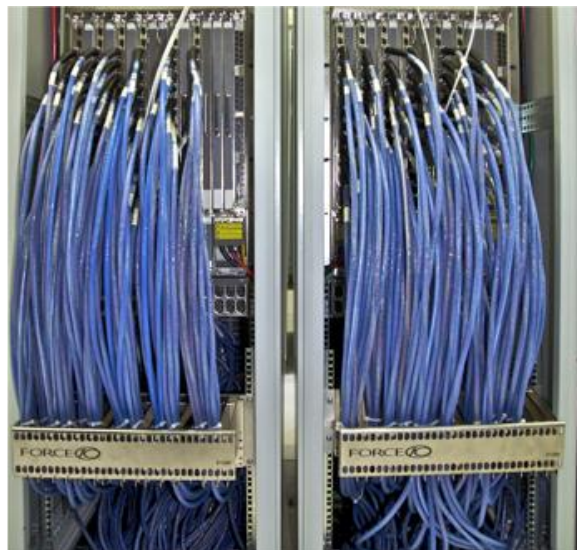
Superfragment builder technology: Myrinet



- Wormhole-routed cross-bar switch
 - 2.5 GB/s / link
 - Low latency
 - No buffering
 - Link level flow control
 - Head-of-line blocking
- NICs
 - Programmable RISC processor
 - Custom protocol
 - LUT-based destination assignment

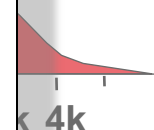
Two-stage event building architecture

Event builder technology: Gigabit ethernet



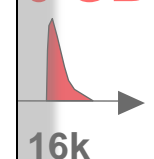
- Standard 1 Gb/s Ethernet
- 8 switches (by Force-10)
 - 1 per slice
 - 4000 ports in total
- 3 rails per Readout Unit PC
- 1 or 2 rails per Builder/Filter PC according to performance

500x
200 MB/s

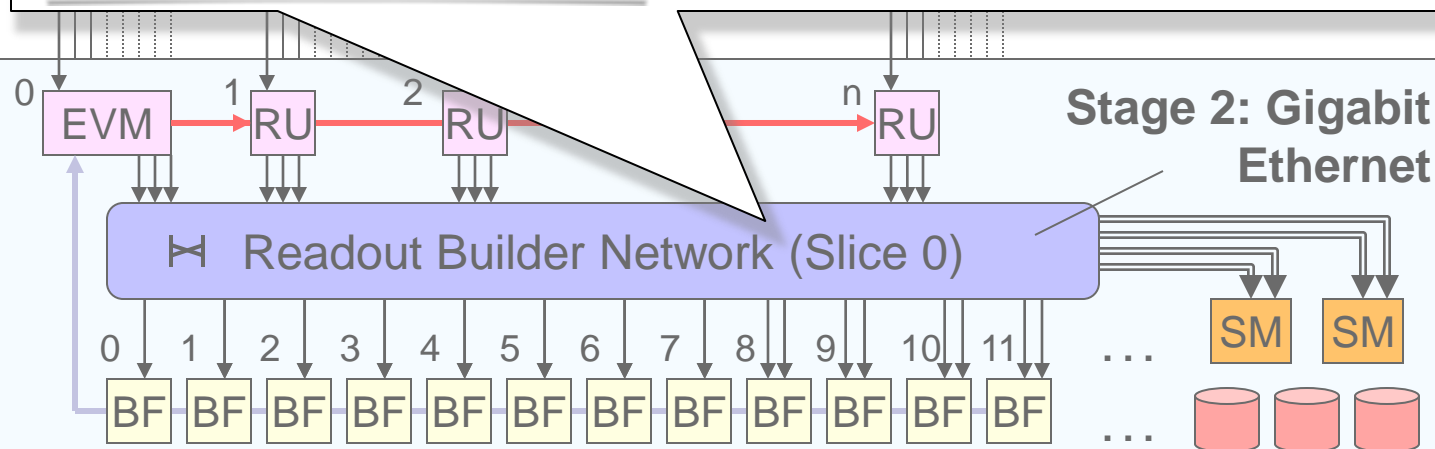


4k

8x8 Super-
ment Builder
0 GB / s



16k



8 independent
DAQ Slices:
~72 x (90 .. 160)

8x 12.5 GB/s

Storage Manager

SAN

Two-stage event building architecture

Storage Managers



- 2 Storage Manager PCs per slice
- NexSan SataBeasts (RAID-6 disk array) connected through redundant Fibre Channel switches
- Max write speed 2.1 GB/s with simultaneous transfer to Tier-0
2.6 GB/s w/o transfer
- Local storage 300 TB (several days)

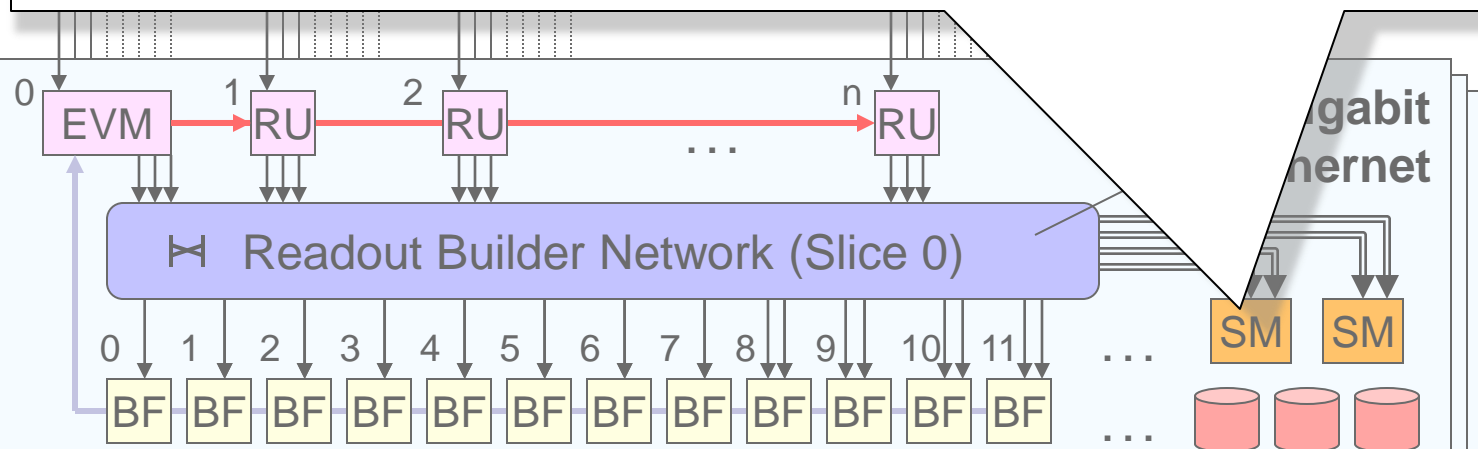
500x
200 MB/s

4k

8x8 Super-
segment Builder

10 GB / s

16k



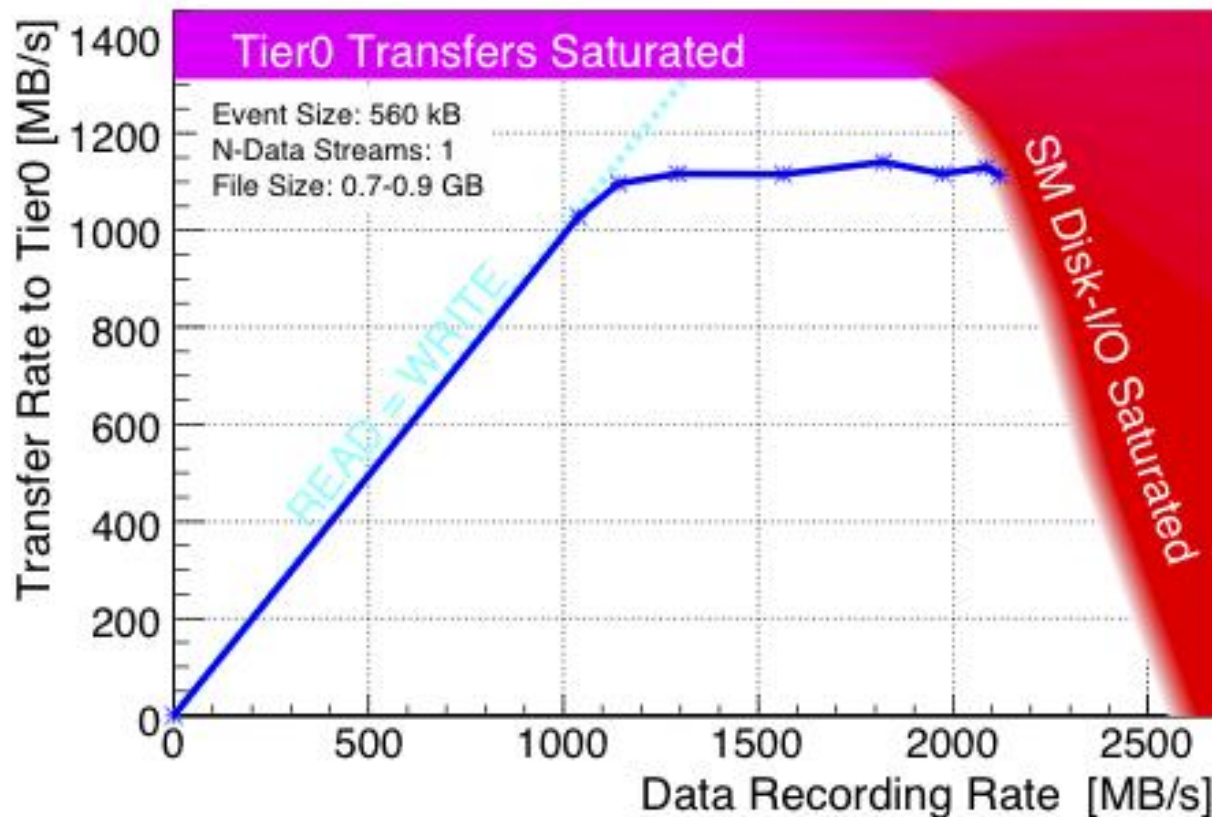
8 independent
DAQ Slices:
~72 x (90 .. 160)

8x 12.5 GB/s

Storage Manager

SAN

Storage Manager Performance

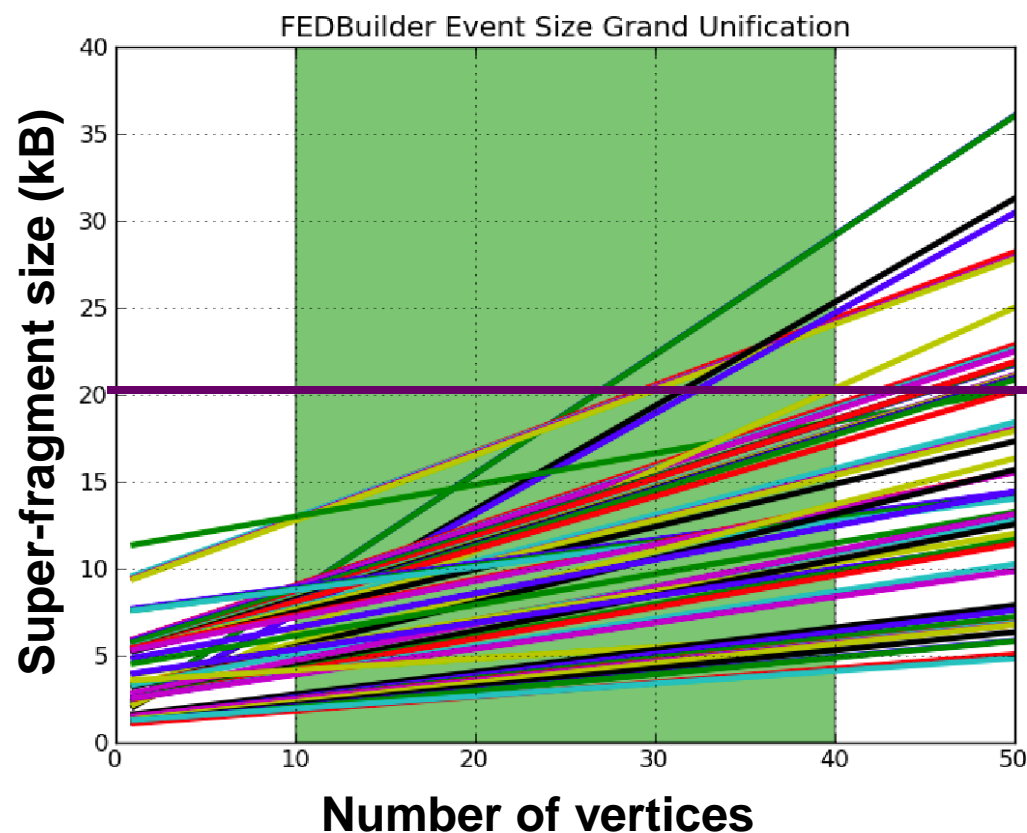


- Total capacity: 300 TB (several days of data taking)
- HLT compresses event data (root); reduction by factor ~2
- Event data to disk
 - pp; ~200 MB/s, design 600 MB/s
 - Heavy Ions: ~1.4 GB/s (up to 2.8 GB/s w/o transfer)



Super-fragment size in pp runs (n_{vertex})

Super-fragment size
at 30 vertices / kB



BPIX s1d06-30 (0.687 kB/vertex)	22.3
BPIX s1d06-29 (0.686 kB/vertex)	22.3
EB- s2d10-03 (0.382 kB/vertex)	20.5
EB+ s2d10-10 (0.373 kB/vertex)	20.4
EB+ s2d10-02 (0.377 kB/vertex)	20.4
EB- s2d10-11 (0.376 kB/vertex)	20.3
BPIX s1d06-29 (0.596 kB/vertex)	19.3
BPIX s1d06-30 (0.576 kB/vertex)	18.9
HF s2d10-07 (0.182 kB/vertex)	16.6

At 100 kHz can take 2.5 kB per FED or
20 kB per super-fragment

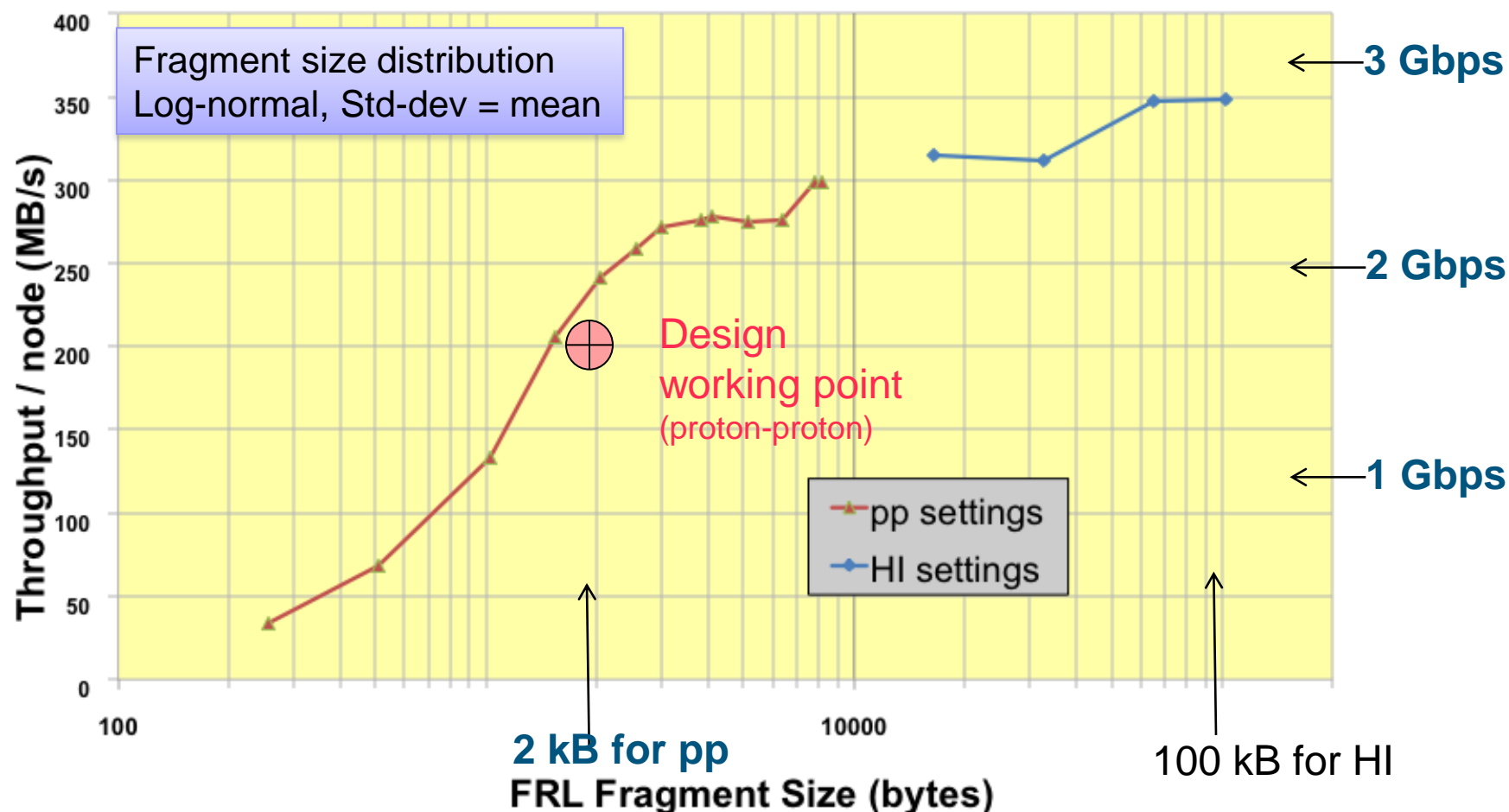
- Some super-fragment builders at the limit with 2011 configuration
- ✓ Fixed by re-arrangement of super-fragment composition

↑ expected 2012



DAQ throughput per input / pp and HI

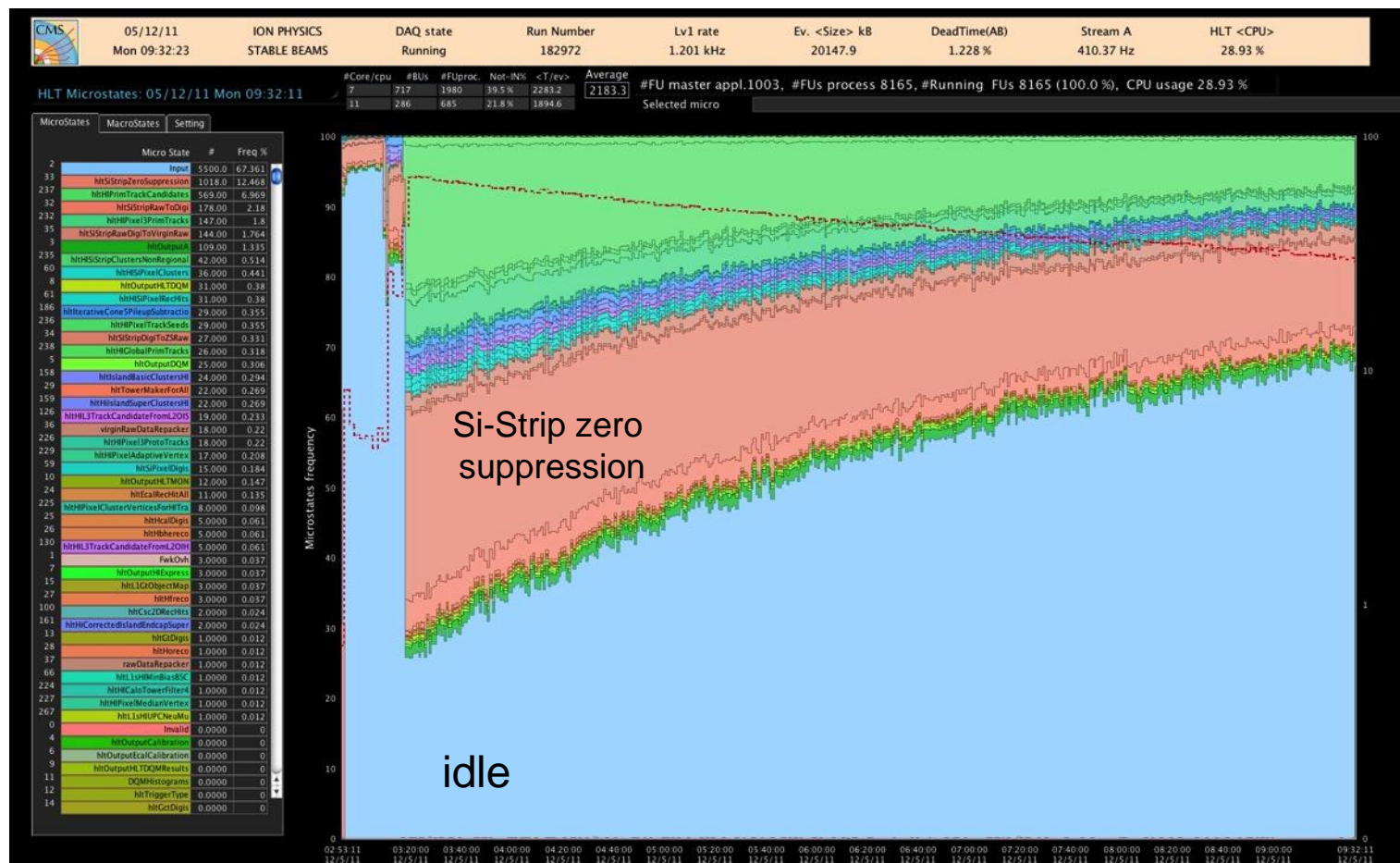
(generated events)



- DAQ optimized for large fragment sizes: reach 350 MB/s (limited by GBe)
- Max rate at 100 kB/FRL: 3.5 kHz
- Max aggregate EVB throughput: ~150 Gbyte/s (436 x 350 MB/s)



HLT states during 2011 Heavy Ion run



- In 2011, Tracker zero-suppression done in HLT farm