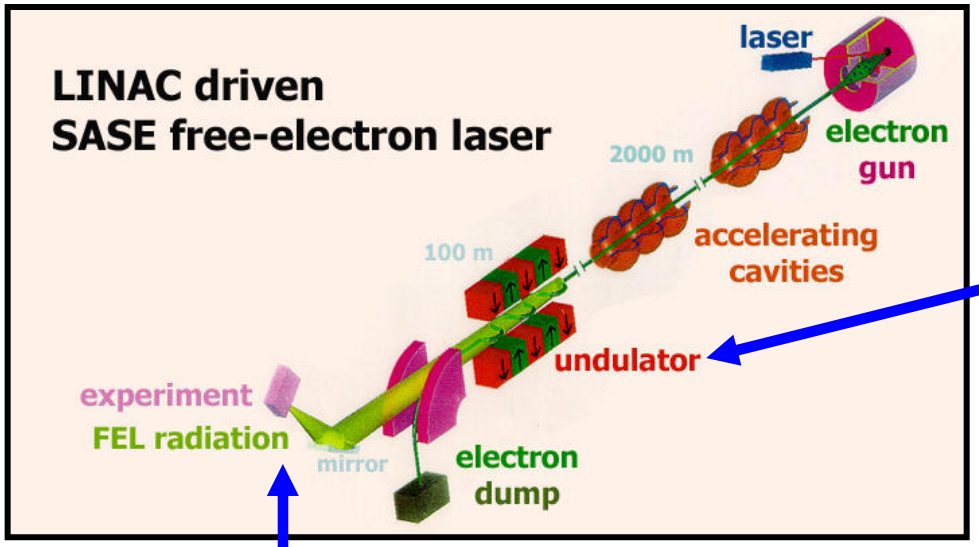
DAQ and control at Free Electron Lasers

- Contents:

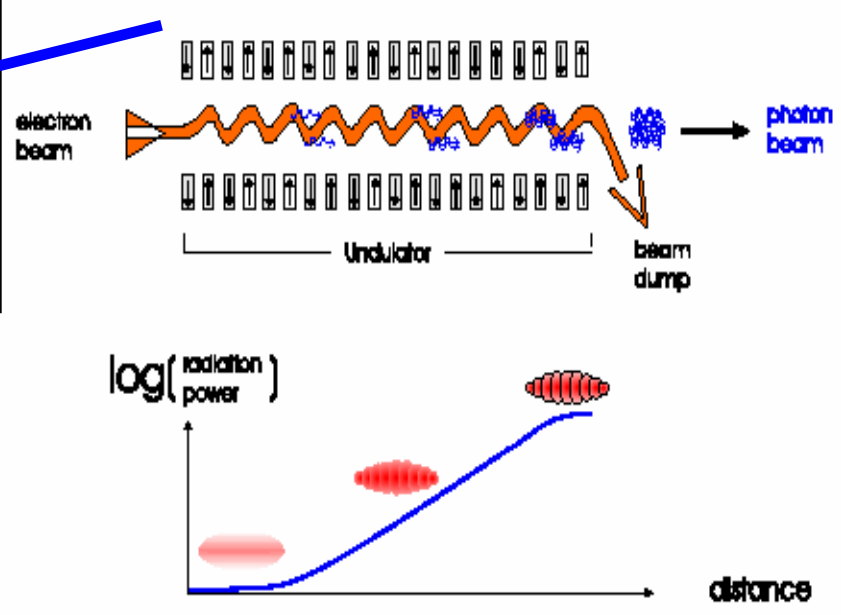
  - Background information
  - DAQ rate and bandwidth requirements
  - Data readout architectures
  - The trigger or data reduction problem
  - Data management, archiving and analysis
  - Summary of challenges
  - Acknowledgements

What is a Free Electron Laser

SASE = Self Amplified Spontaneous Emission

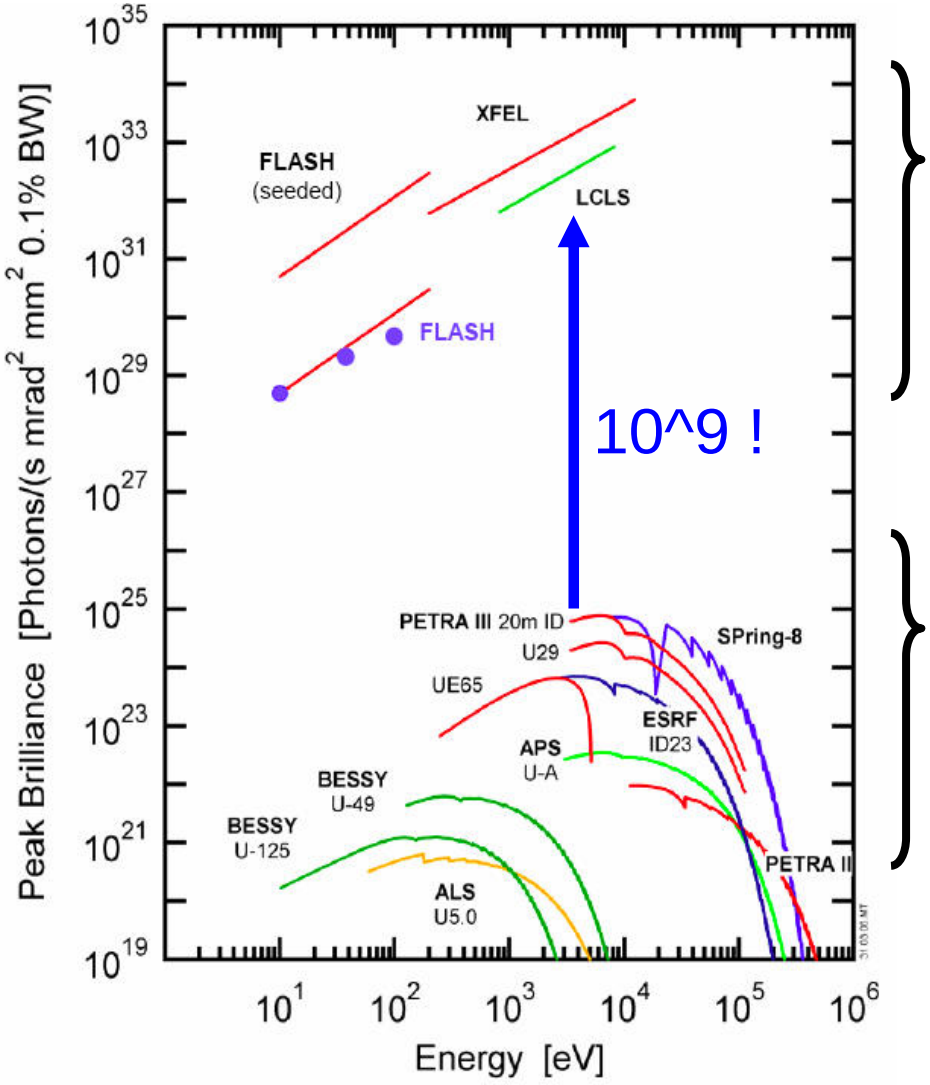


Interaction of linac electron beam with undulator radiation field produces micro-bunching at the resonant wavelength



Experiments benefit from, high power, increased brilliance, highly coherent short pulses

X-ray FEL unprecedented leap in brilliance



X-ray FEL facilities

= 4th generation light sources
(e-linac + undulators)

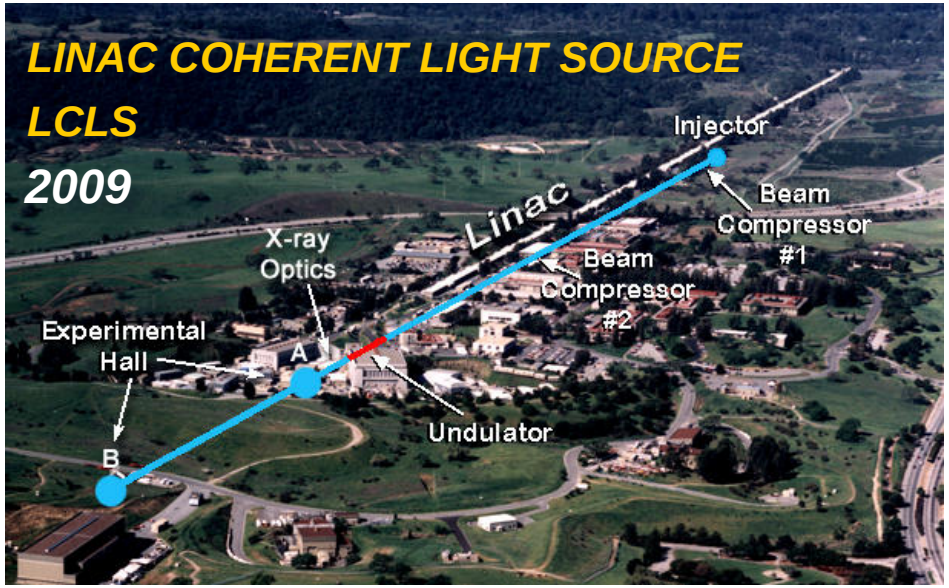
Synch. Radn (SR) facilities

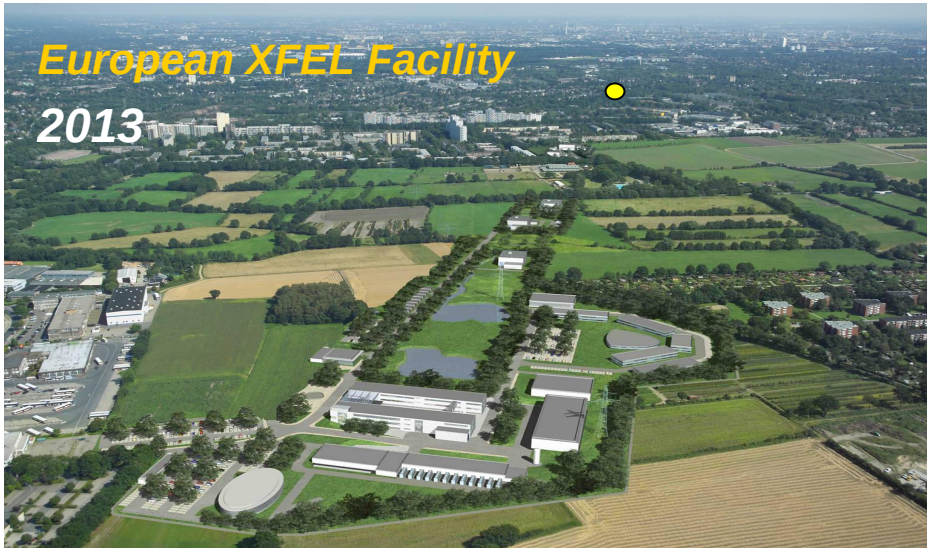
= 3rd generation light sources
(e-ring + undulators + wigglers)

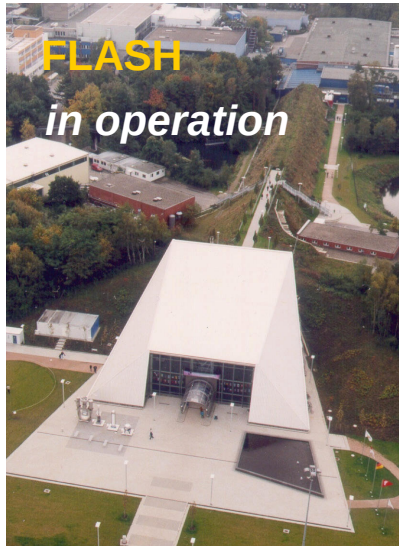
2nd generation = purpose built e-ring

1st generation = parasitic use of HEP e-ring

Hard X-ray SASE Free Electron Lasers







Comparison of the X-ray FEL Light Sources

	FLASH	*LCLS*	*SCSS*	*XFEL-SASE1*
Min. Wavelength (nm)	**6.5**	**0.15**	**0.1**	**0.1**
Peak Brilliance	**10^30**	**8.5 10^32**	**5 10^33**	**5 10^33**
Average Brilliance		**2.4 10^22**	**1.5 10^23**	**1.6 10^25**
Transverse coherence		**0.83**	**0.24**	**0.1**
Pulse duration (fs)	**10-50**	**230**	**500**	**100**
First Beam for Expts.	**2005**	**2009-2010**	**2010**	**2013-2014**

Now:
Older implementation
Use as prototype

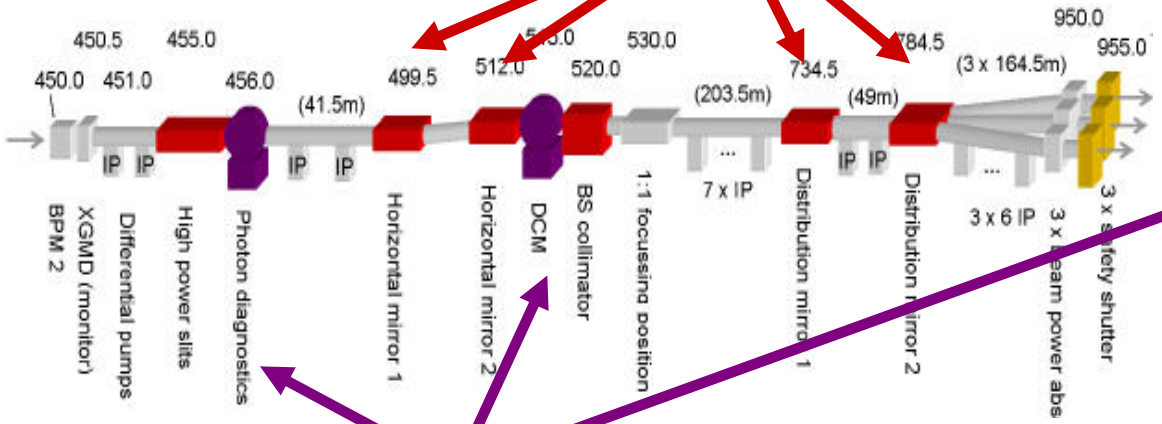
Soon:
Newer implementation

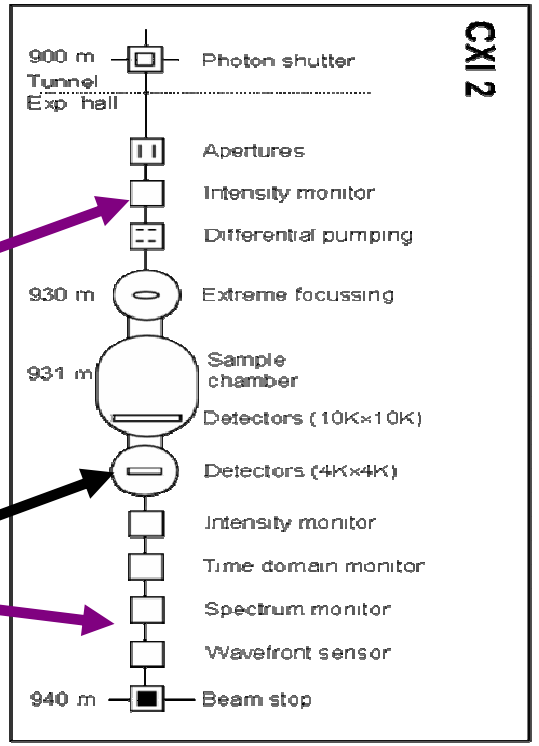
Later

Today's DAQ component data sizes

Beam = mirrors, monochramators, shutters, …

~few channels per component, e.g. motor control





Diagnostic detectors = pulse position, intensity,

<1kB data per pulse, e.g. intensity monitor

Expt. Detectors = cameras, …

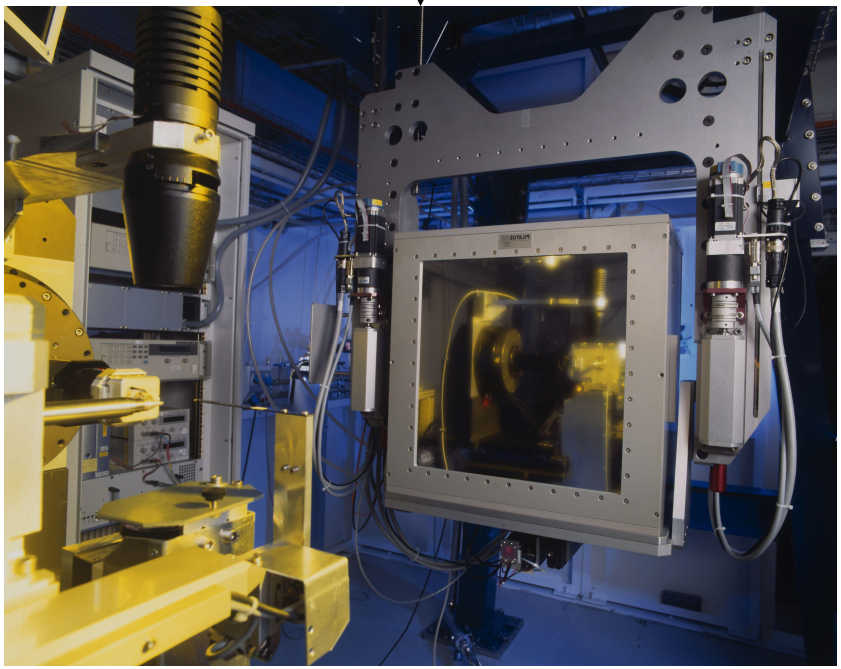
~2MB data per pulse, e.g. 2D pixel detector

Data size per pulse = 2D Detector : Diagnostic : Beam = MB : kB : ~0 Bytes

⇨ DAQ: satisfying 2D requirements will automatically satisfy others

What do 2D detectors (cameras) look like

- Synchrotron rings = CCD cameras
  - ~few Mpixels
  - ≤ 100 Hz readout
  - Commercially available



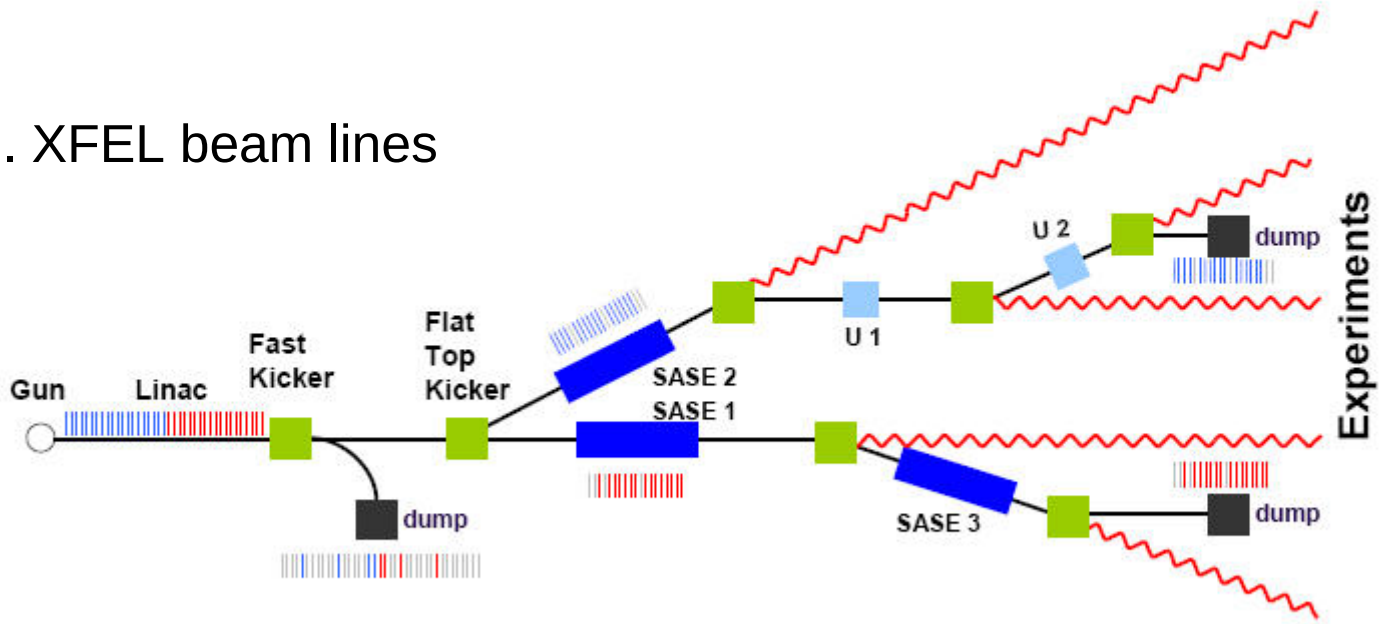
6Mpixel Pilatus (PSI) 10Hz

- LCLS = Brookhaven + Cornell cameras
  - ~1 Mpixel
  - Pixel Sensor bonded to ASIC
  - 120 Hz readout
  - Custom development, now being built

- XFEL = HPAD, LPD and DEPFET
  - ~1 MPixel
  - Pixel detectors, sensor bonded to ASIC)
  - 5 MHz readout (⇨ more demanding 50 – 150 x LCLS 2D rate)
  - Custom development, now being designed

X-ray FEL beam lines

e.g. XFEL beam lines

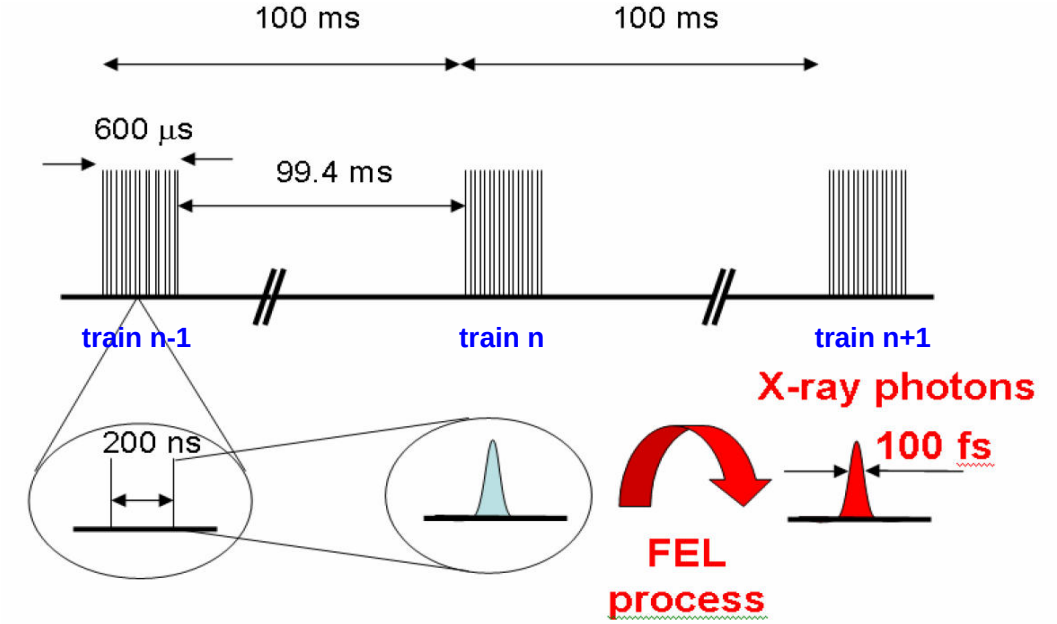


- Multiple beam lines
  - e-bunches reused to generate new photon bunch trains
  - Number of experiments ~ Number of beamlines
  - ~1 detector taking data per beamline
- Easier to add beam lines than interaction points ?

⇨ Facility data volume = N x data volume per bunch train

Nominal X-ray FEL pulse structures

e.g. XFEL =



	FLASH	*LCLS*	*SCSS*	*XFEL*
Nominal Linac rates:				
Train repetition rate (Hz)	**5**	**120**	**60**	**10**
Pulses per train	**800**	**1**	**1**	**3000**
Pulse rate in train	**1 MHz**	**-**	**-**	**5 MHz**

FLASH = No dedicated large 2D detectors capable of handling multiple bunches per train deployed.

X-ray FEL DAQ rates and bandwidths

Using pulse structures, detector size, etc. can now guess rough bandwidths

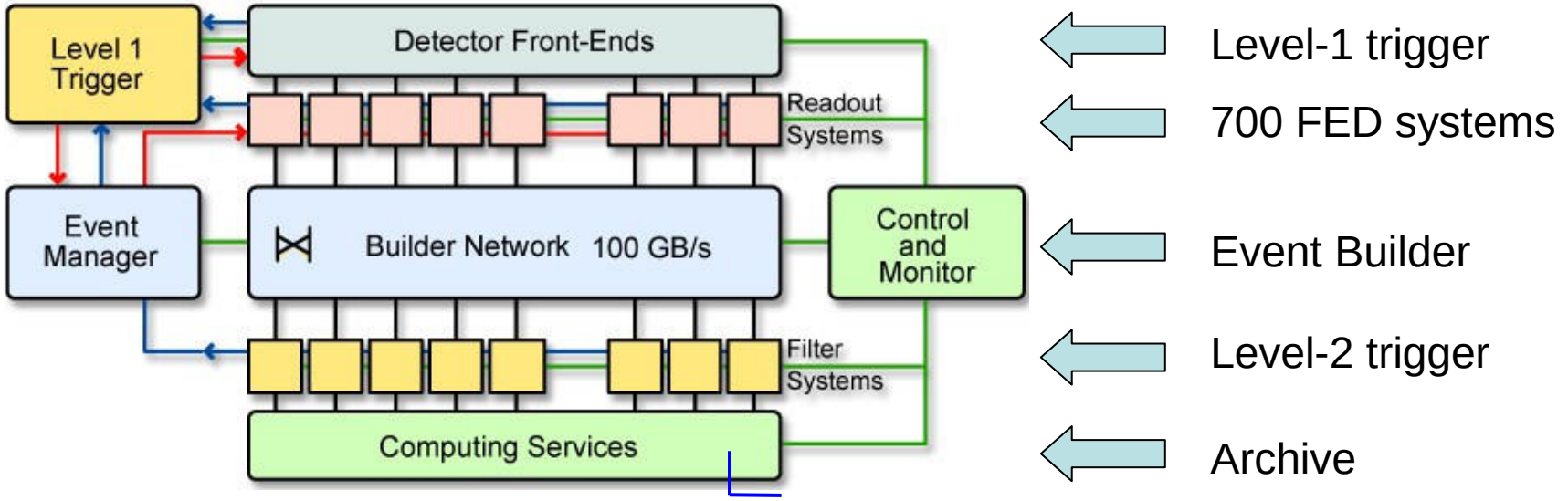
	LCLS	*SCSS*	*XFEL*
Max. DAQ rates per train:			
1Mpixel pulses digitized	**1**	**1**	**512**
Diagnostic pulses digitized	**1**	**1**	**~1500**
Ballpark detector bandwidths:			
1Mpixel(=2MB) data digitized (MB/s)	**240**	**120**	***10240***
Diagnostic data digitized	**0.001**	**0.001**	**1.5**
Ballpark facility bandwidths:			
Beam lines (N)	**3**	**1**	**5**
machine x detector (0.7 x 0.7) eff.	**0.5**	**0.5**	**0.5**
N x 1Mpixel bandwidth (MB/s)	**360**	**120**	***25600***
TB of data per day	**31**	**6**	***2211***

The 512 pulses digitized is a limit set by current Mpixel FEE design.
The 1500, not 3000, pulses is due to the beam line setup.

LCLS and SCSS bandwidths are within current technological solutions: 10 GE, switch, and storage technologies. XFEL bandwidths are more difficult, hence reduction in pulses digitized per train

Requirement: handle as much readout bandwidth as possible and provided as much data reduction/rejection as possible before archiving.

HEP Trigger & DAQ – here CMS



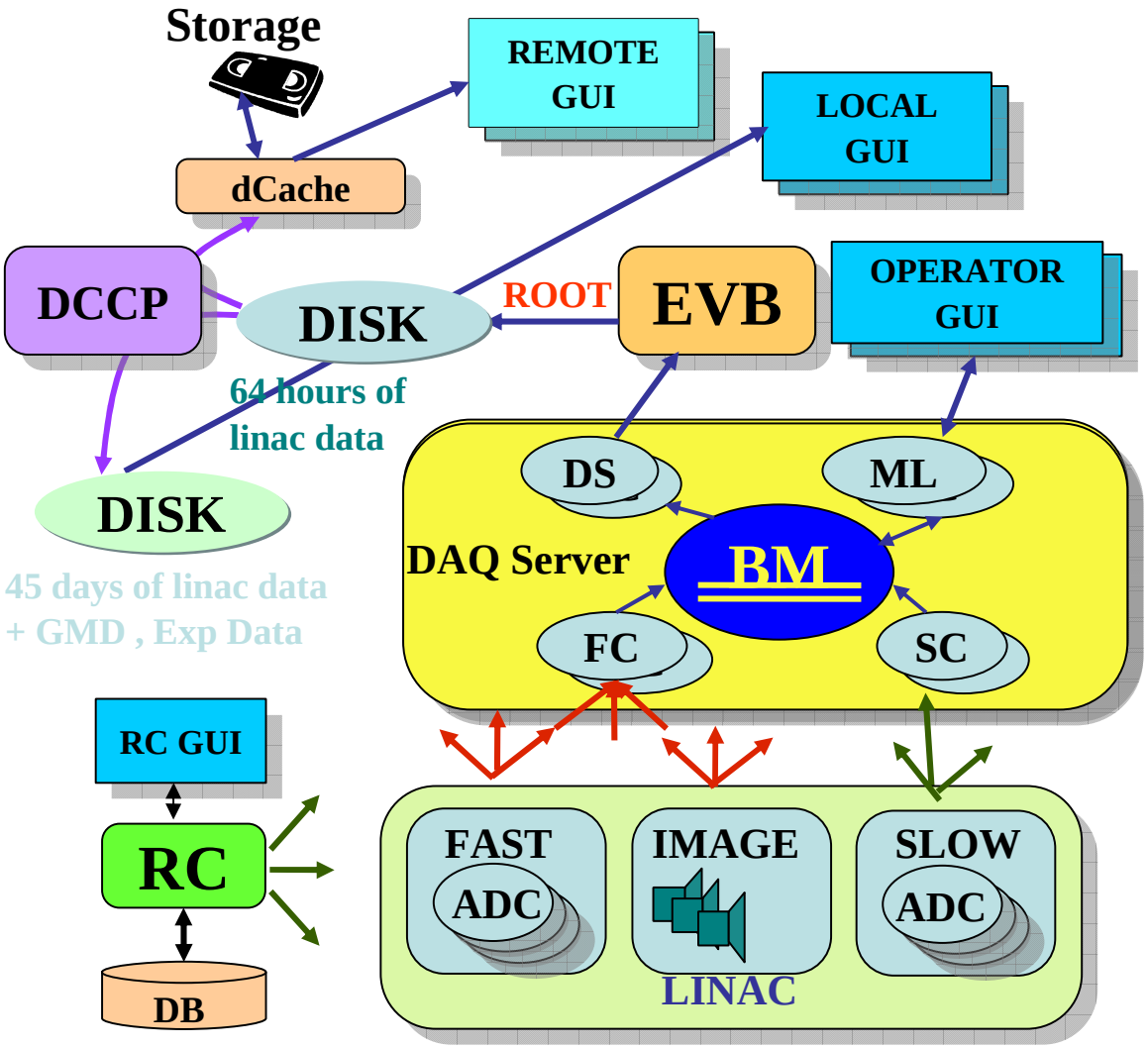
- Overall Trigger & DAQ Architecture has 2 Levels:
- Level-1 Trigger: 25 ns input, 3.2 μs latency = 128 deep analogue pipelines
- Level-1 Output: initially 50 kHz increasing to 100 kHz final
- On L1 accept ALL detector data sent (FEDs) to L2 PCs O(2000) nodes
- Level-2 Trigger: ~10ms, find more complex signatures in full event data
- On L2 accept send data to archive
- Archive: event size 1MBytes, rate = 100Hz, 9 TBytes/day

HEP versus PS trigger & DAQ

- In HEP sizing the DAQ is relatively simple
  - Accelerator and detectors are designed for “expected” physics events
  - Simulations of “expected” plus “backgrounds” are accurate and give data sizes at sub-detectors and allow trigger simulations
  - Trigger cutting on small number of “simply” identifiable primitives possible (high P leptons, missing energy,…)
  - Rates: input rate known (lumiosity,expected+noise cross-sections), output rate is fixed to a sensible value (technology and event sample)
  - Divide the DAQ into trigger/readout layers which will be technically feasible (pipeline lengths, transfer rates between layers,…).
  - Detector sizes do not change during experiment, nor do the trigger rates.
- In photon science sizing the DAQ is not so simple
  - No single “expected” experiments, uncertainty in
  - Only rudimentary simulations hence poor understanding of what is coming, e.g. noise effects on data reduction
  - No “simple” trigger primitives are present, hence uncertain data rejection.
  - Detector sizes are not fixed and will increase

⇨ Makes DAQ design for higher pulse rate machines difficult

FLASH readout architecture



multicast

Fast data (every micropulse)

Beam relevant info:

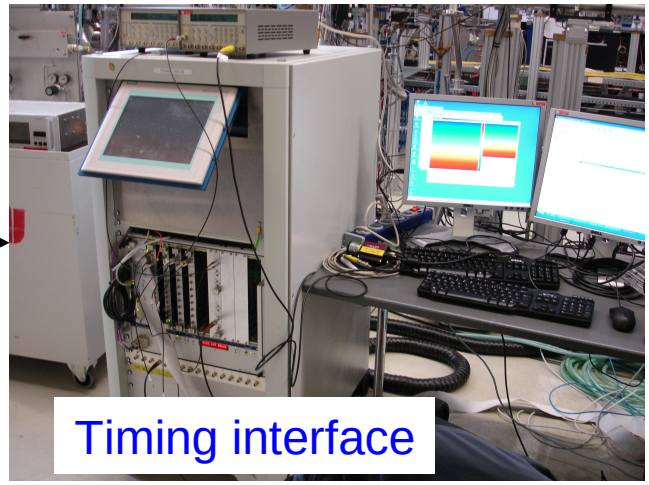
- ADCs (BPM, BLM, TOR, etc)
- CAMERAs

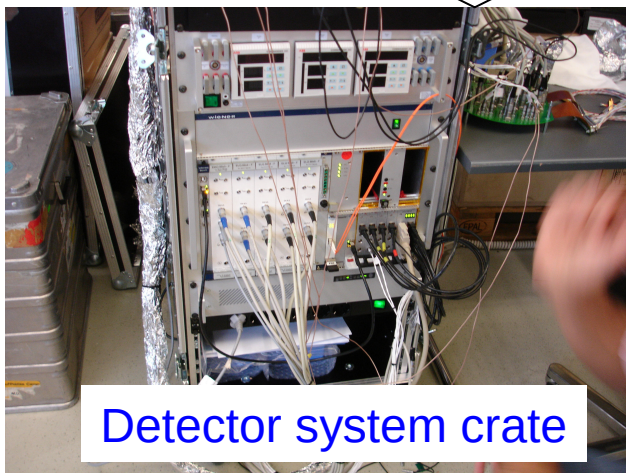
DOOCS (TINE)

Slow data (max 1Hz)

- Data from slow ADCs (MAG, V, etc.)
- DOOCS channels (Masks, params, etc.)

FLASH detector systems

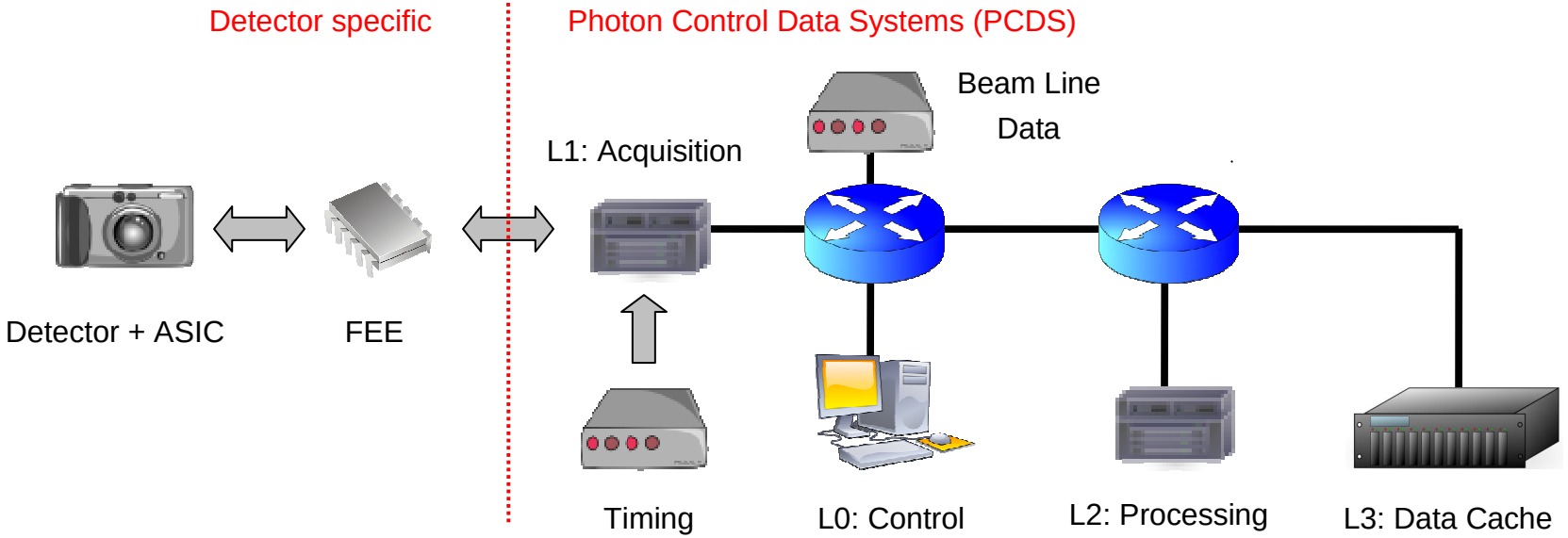




- VME crate system (e.g. photon intensity monitor GMD, etc.)
  - Single board computer running Solaris
  - Sequencer, ADCs, timing system interface, …
  - Data forwarded to buffer manager system, RC console on PC, …
- Custom solution (normally for single bunch mode operation)
  - Bring own DAQ system (e.g. in crate) and use start signal from timing interface.
  - If needed connect to BM of FLASH readout system to archive data, or write to local PC disk.

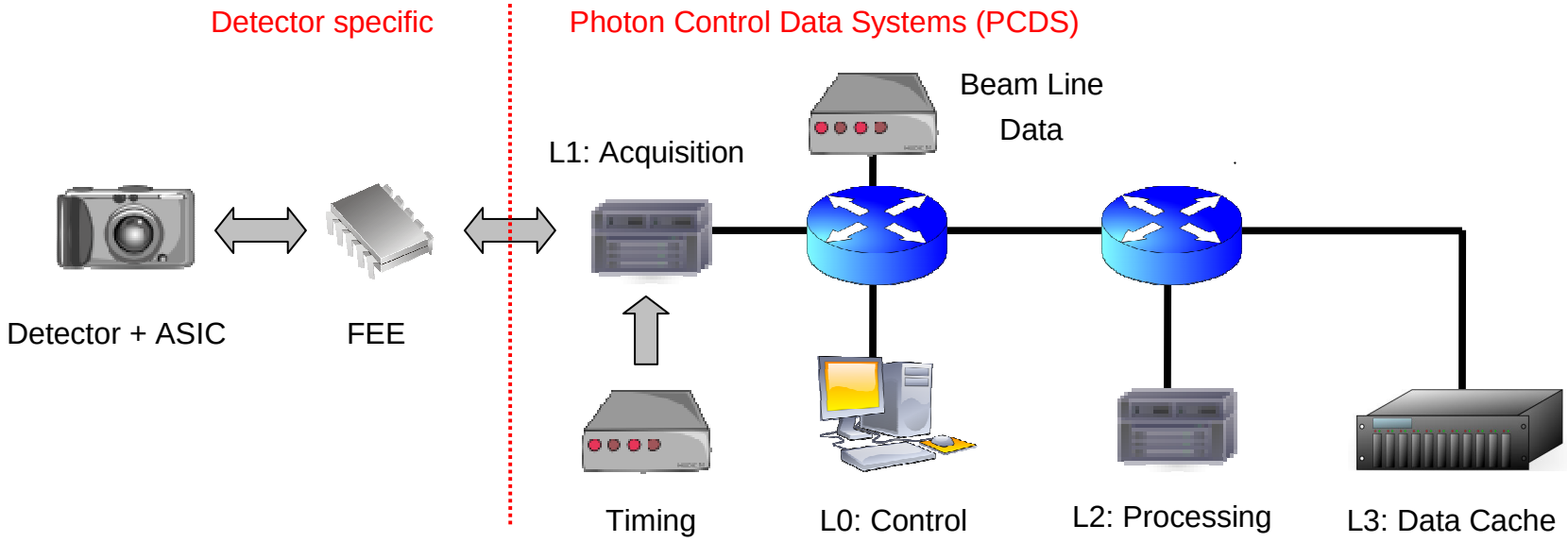
⇨ FLASH DAQ: initially for prototyping, since extended

LCLS data system architecture



- Detector + Asic = experiment specific
- Front End Electronics (FEE)
  - Provides configuration registers and state machines to run detector
  - Provides ADC if ASIC does not
  - FPGAs transit data and control on one fibre between FEE and L1 node
    - PGP small footprint reliable protocol developed and used (not UDP, aurora..)

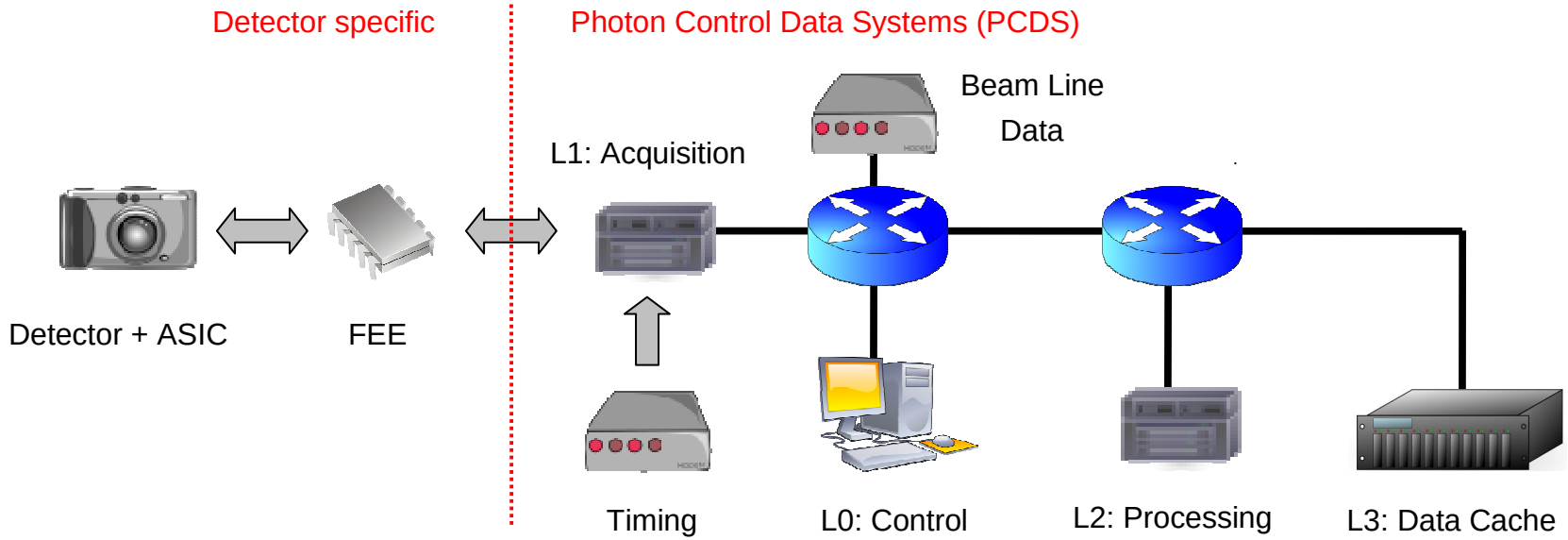
LCLS data system architecture



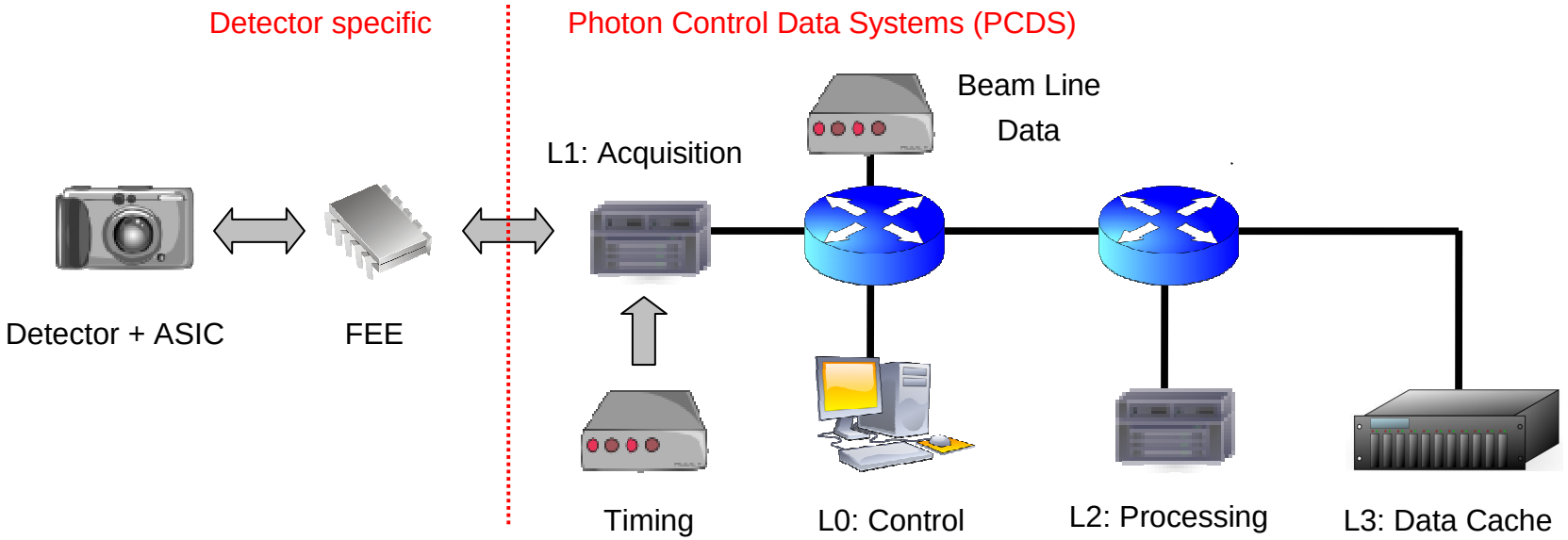
- Level 1 (L1) acquisition node = DAQ custom ATCA board
  - Receives 120Hz timing signal, sends trigger to FEE, receives data
  - Sends configuration and control data to FEE
  - Event builds: detector data and beam line data
  - Drives calibration and image processing
    - Pedestal subtraction and cross talk using calibration constants
    - Data reduction (compression) and rejection using beam line data (vetos?)
    - Processing planned on FPGAs and CPUs (Vitrex 4)
  - Send data to L2 node using 10GE links

DAQ and control at Free Electron Lasers

- Level 2 (L2) nodes = commercial ATCA nodes
  - Higher level processing
    - Pattern recognition filter, alignment monitoring, reconstruction…
  - Send processed data to L3 using 10GE link
- Level 3 (L2) = commercial data cache
  - Provides data storage in experimental hall
    - Covers ~4 day down time in tape storage system
  - Sends data for archiving to SLAC computer center



LCLS data system architecture



- Level 0 node (L0) = Experiment control
  - Manages all L1, L2, L3 nodes in a given partition
  - Data taking run control system
  - Detector configuration control (modes, biases, thresholds, etc.)
  - Run and environment monitoring

LCLS DAQ xTCA hardware

- LCLS is designed around ATCA crate standard

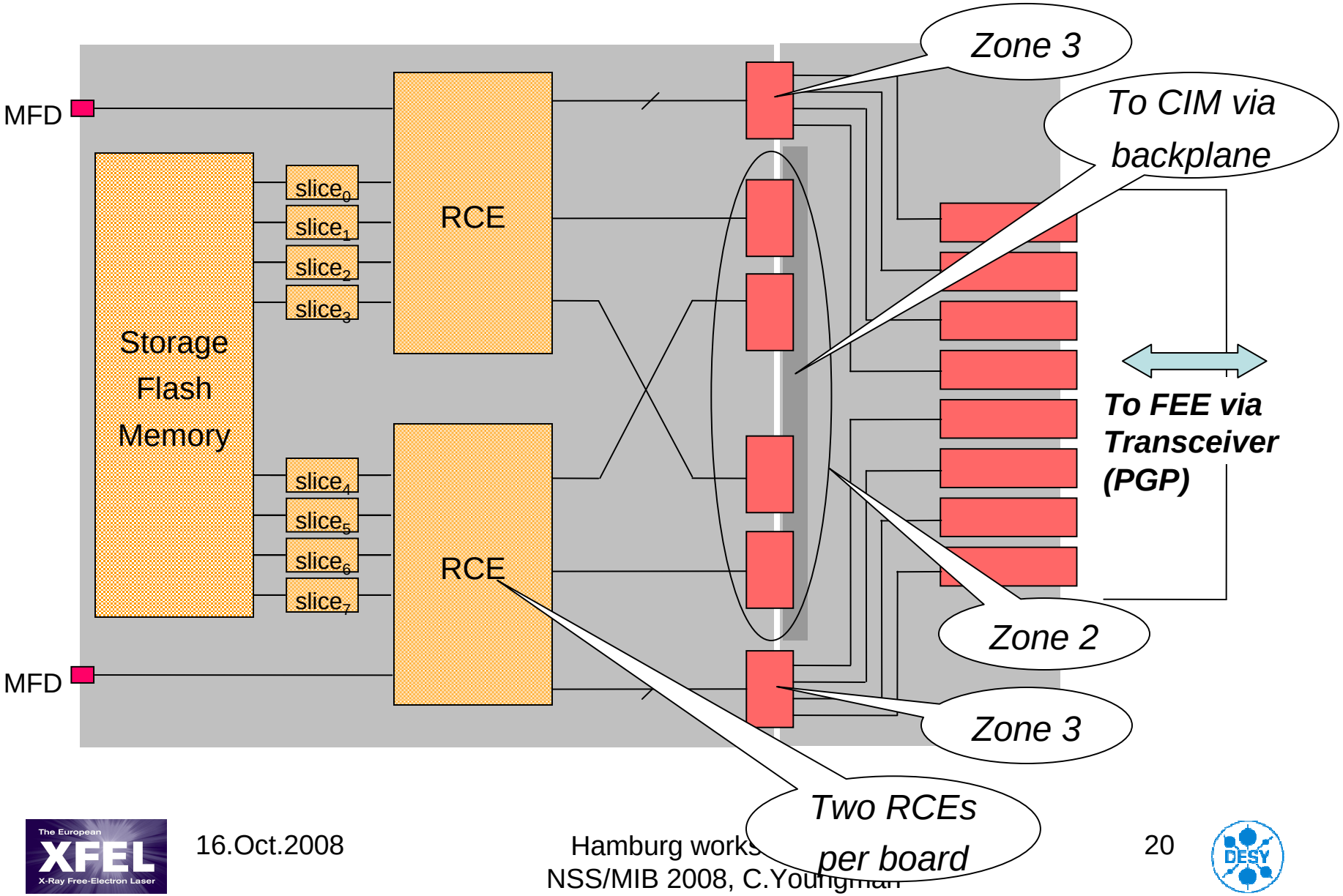
- RCE or Reconfigurable Cluster Element (ATCA form factor)
  - L1 node = Front end to the FEE
  - Idea: All L1 nodes are RCE
  - RCE - FEE connected via single fibre connection (allows physical separation)
  - Data and control (register r/w, configuration) via PGP protocol
  - All FEEs have to obey protocol
  - Can cluster RCEs into one crate

- CIM or Cluster Interconnect Module (ATCA form factor)
  - 10 GE switch
  - Connects RCE within crates
  - Connects crates (clusters of RCEs) to other crates

- Both RCE and CIM initially developed at SLAC for Bhabha DAQ Petadata system

⇨ Looks like a HEP solution – standard (API and crate)
Detectors/experiments are the components of the LCLS experiment

RCE Board Block Diagram

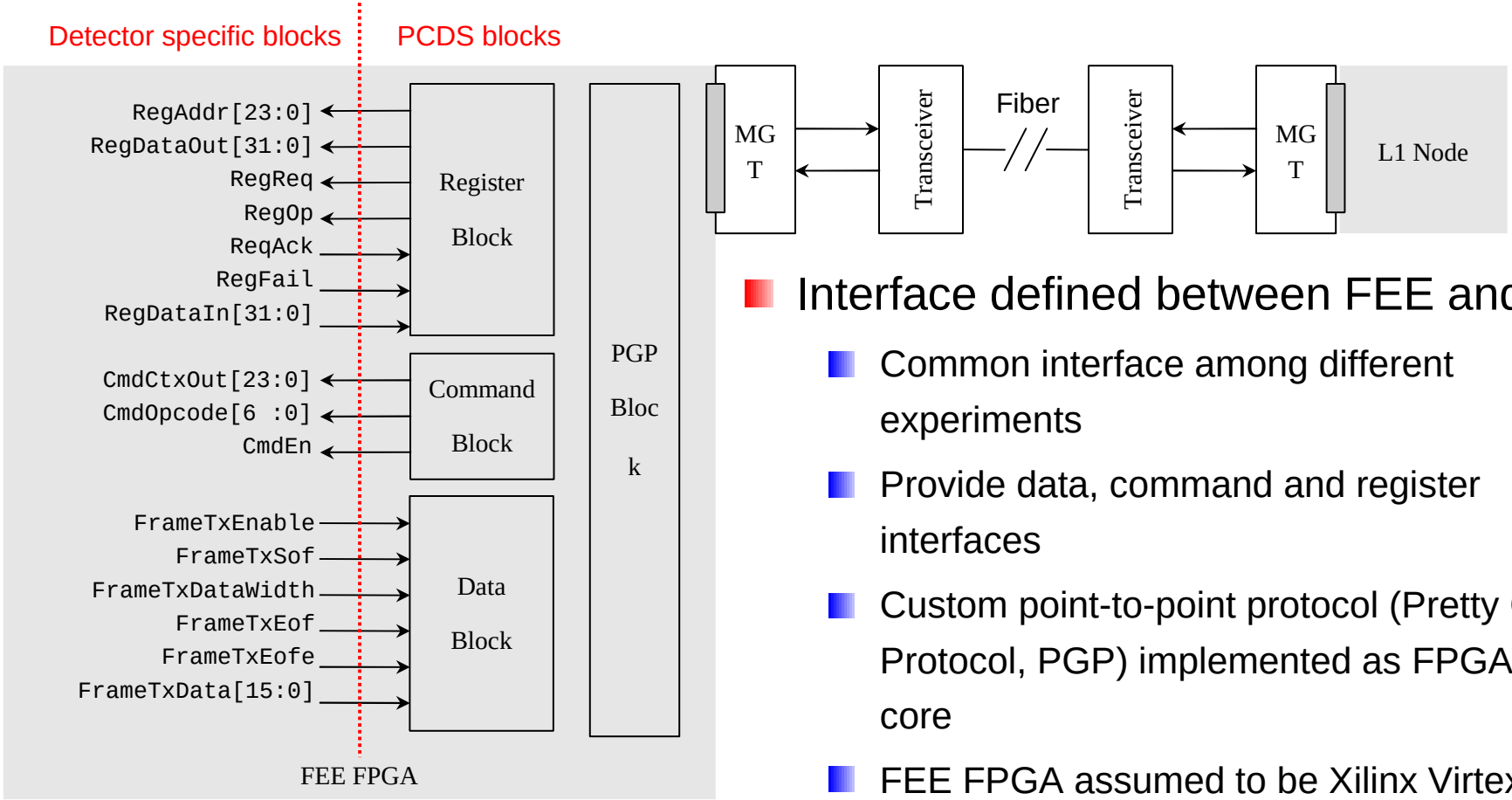


Reconfigurable Cluster Element

- Based on System On Chip (SOC) Technology
  - Currently implemented with Xilinx Virtex 4 devices, FX family
    - Targeting XC4VFX60
  - Xilinx devices provide
    - Reconfigurable FPGA fabric
    - DSPs (200 for XC4VFX60)
    - Generic CPU (2 PowerPCs 405 running at 450 MHz for XC4VFX60)
    - TEMAC: Xilinx TriMode Ethernet Hard Cores
    - MGT: Xilinx Multi-Gigabit Transceivers 622Mb/s to 6.5Gb/s (16 for XC4VFX60)
- FPGA fabric
  - Interfaces to: memory subsystems, JTAG debug port, ..
  - Generic DMA Interface (PIC) designed as set of VHDL IP cores
    - Up to 16 PIC channels
  - PIC in conjunction with Multi-Gigabit Transceivers and protocol cores, provide many channels of generic, high speed, serial I/O
    - 10Gb Ethernet and PGP
  - PIC in conjunction with TriMode Ethernet Hard Cores also provide commodity network interfaces
    - 1Gb Ethernet

RCE Board with RTM

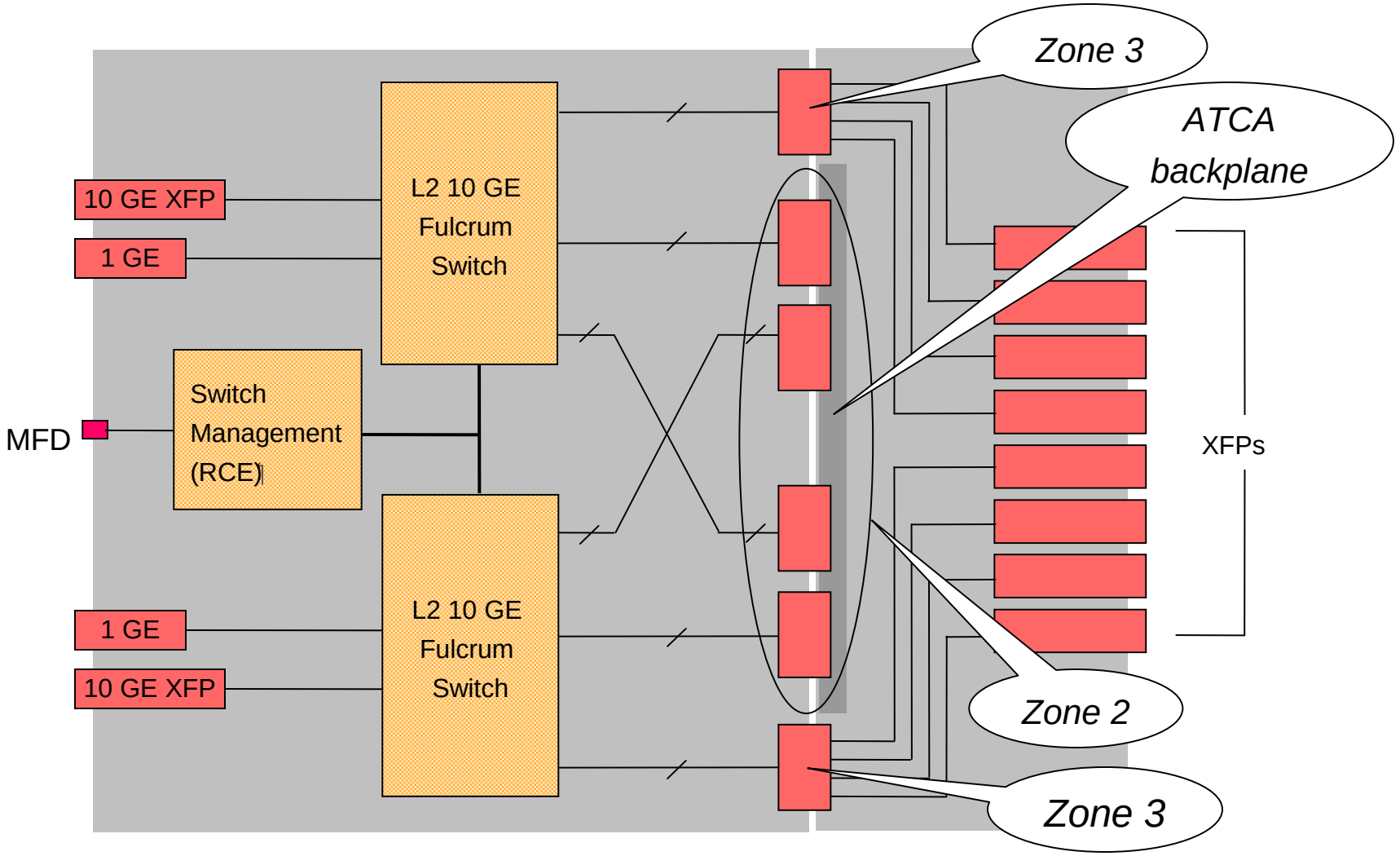
Register Command Data Interface



- Interface defined between FEE and L1
  - Common interface among different experiments
  - Provide data, command and register interfaces
  - Custom point-to-point protocol (Pretty Good Protocol, PGP) implemented as FPGA IP core
  - FEE FPGA assumed to be Xilinx Virtex-4 FX family with Multi Gigabit Transceivers (MGT)

⇨ All detectors look like RCE interfaces.
Implementation of FEE behind the definition is fn. of detector

CIM Board Block Diagram

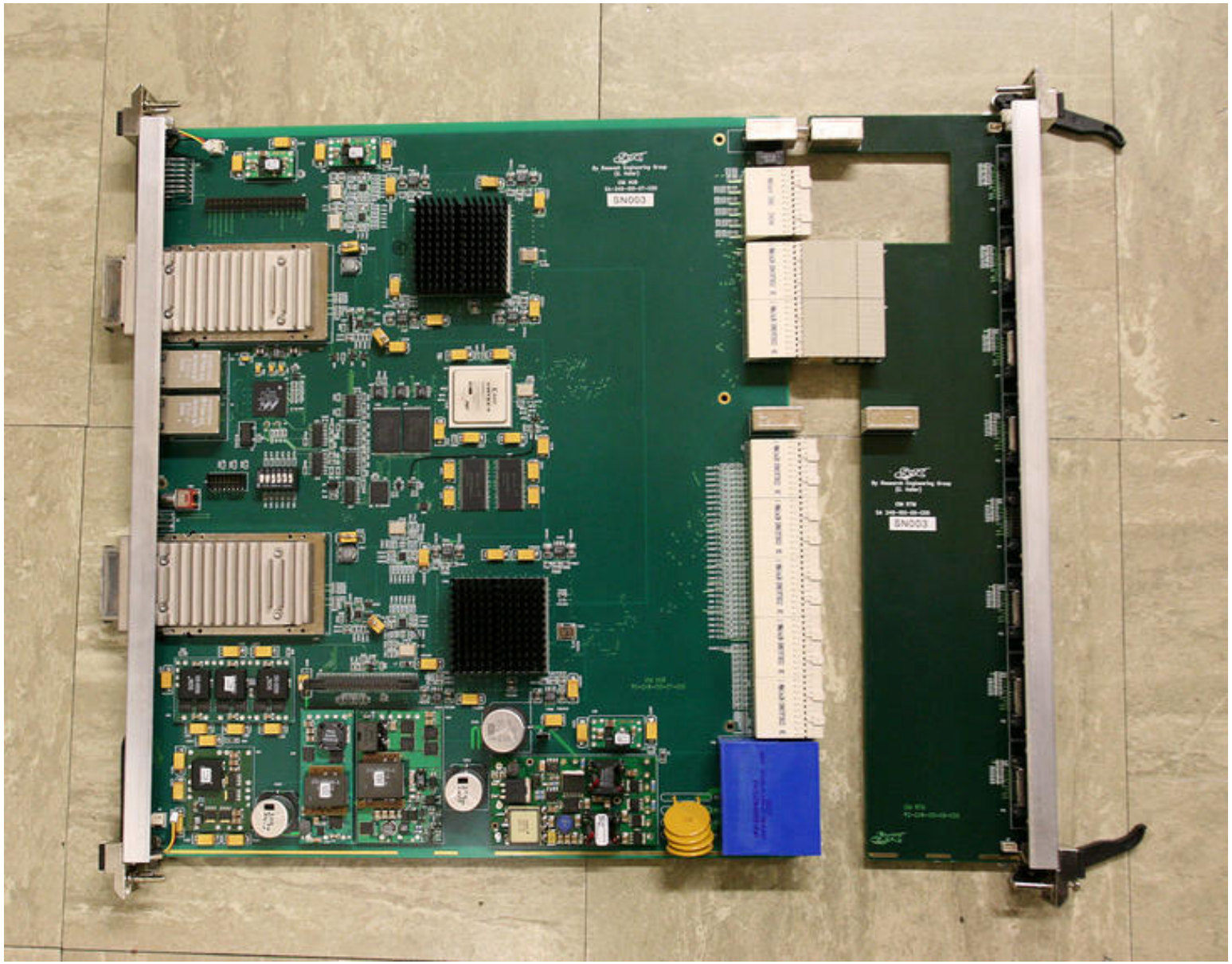


Cluster Interconnect Module

- ATCA network switch
  - Based on two 24-port 10Gb Ethernet switch ASICs from Fulcrum
    - Up to 480 Gb/s total bandwidth
  - Managed via Virtex-4 devices
    - Currently XC4VFX20
  - Interconnect up to 14 in-crate RCE boards (i.e. 28 RCEs)
  - Interconnect multiple crates for additional scalability
    - This is how L1 and L2 node crates are connected (and L3 to L3?)

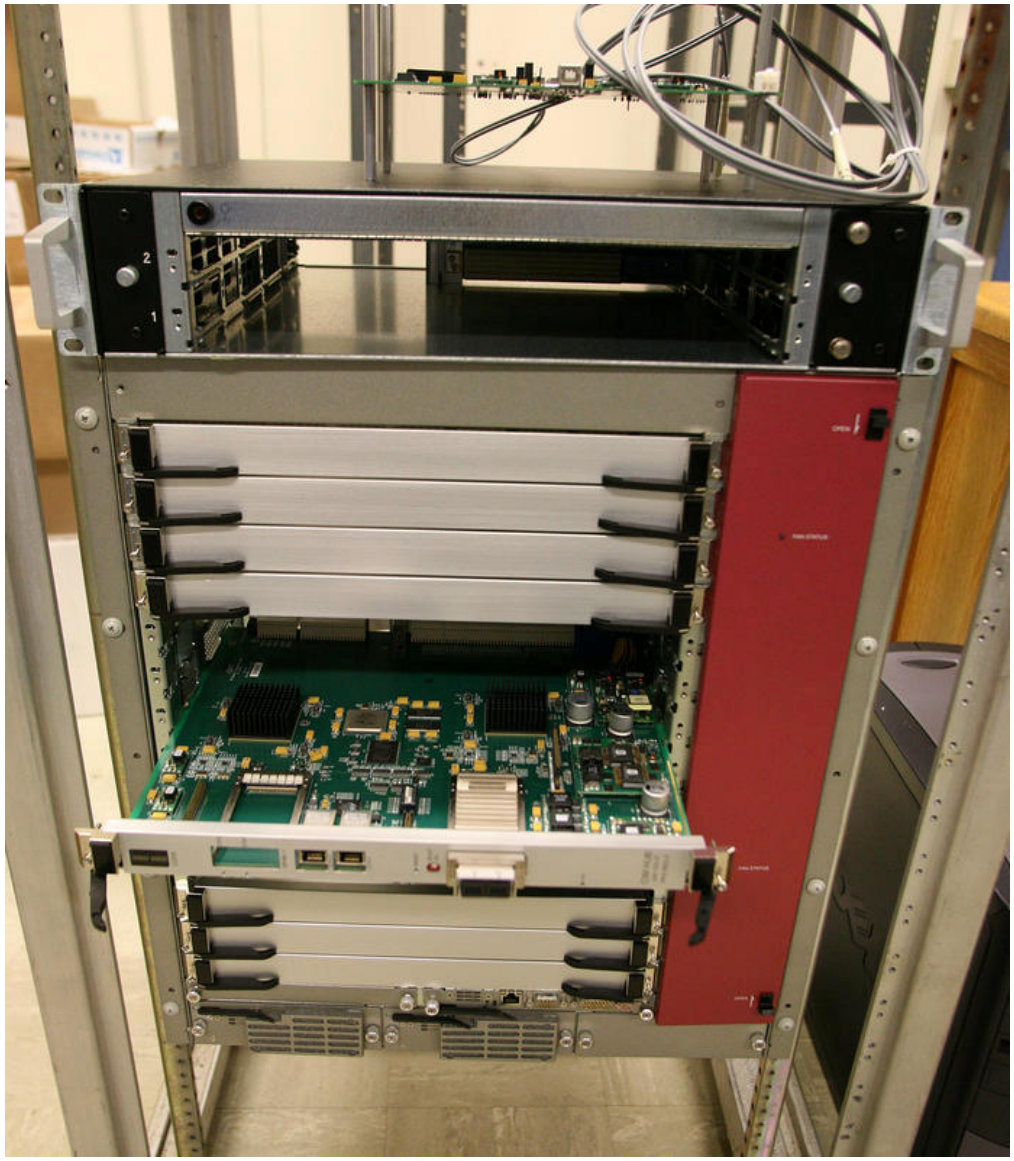
- Fully configurable
  - Designed to optimize crates populated with RCE boards
    - Ability to use ATCA redundant lanes for additional bandwidth if desired
    - Ability to use 2.5Gb/s connections in place of standard 1Gb/s Ethernet
  - At the same time may be configured to connect standard ATCA blades

CIM Board with RTM



Front-end Development Board

ATCA 14-slot Chassis



Spring 8 data system architecture

On-site Blade server
<100 nodes

Off-site Super
computer center
http://www.nsc.riken.jp/index-eng.html

storage

Pre-process: gain, blemish pixel
Calibration. Data Compression

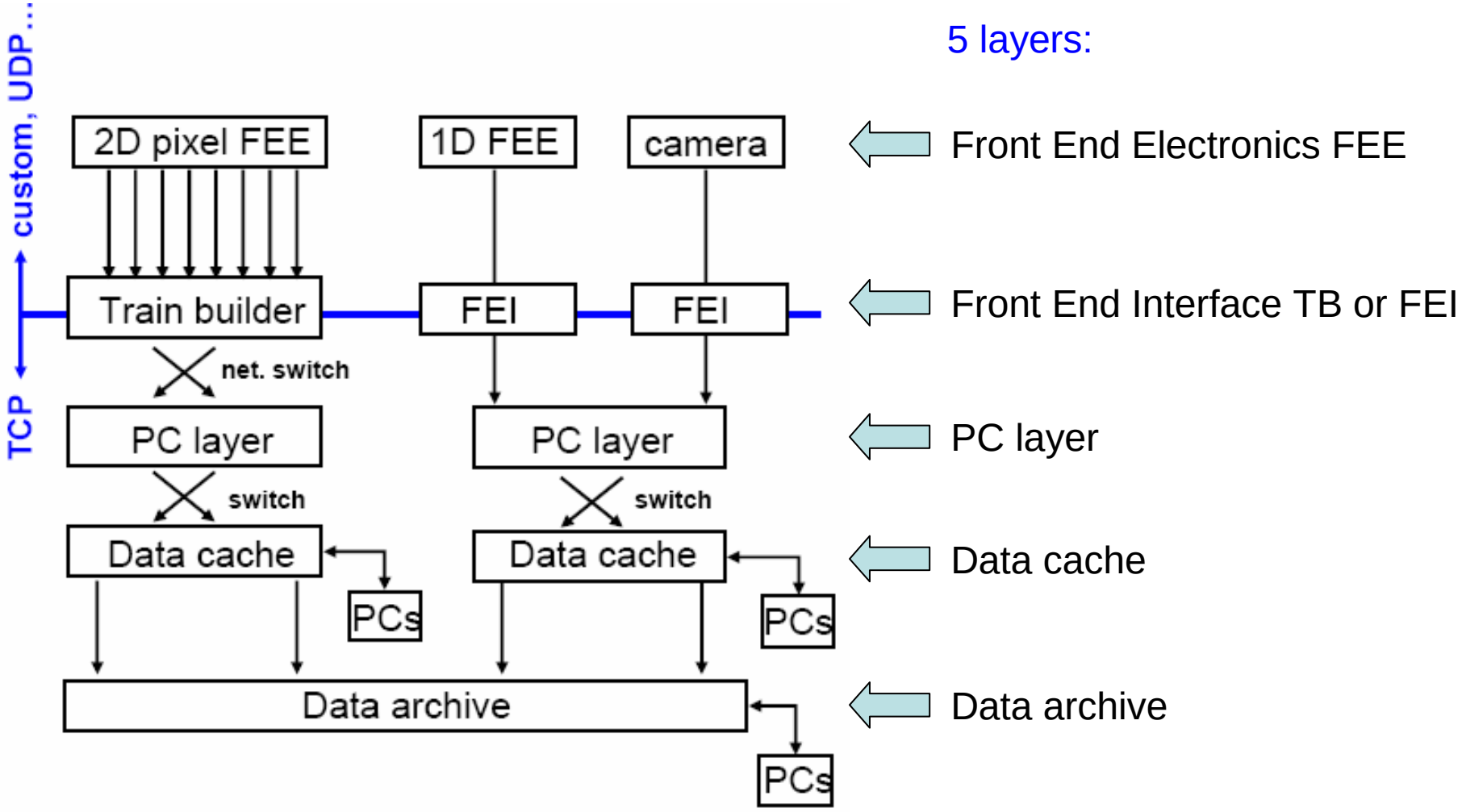
P2P parallel storage network (TBD)
<100 MB/s/line
Secure data transfer

SCSS data system architecture

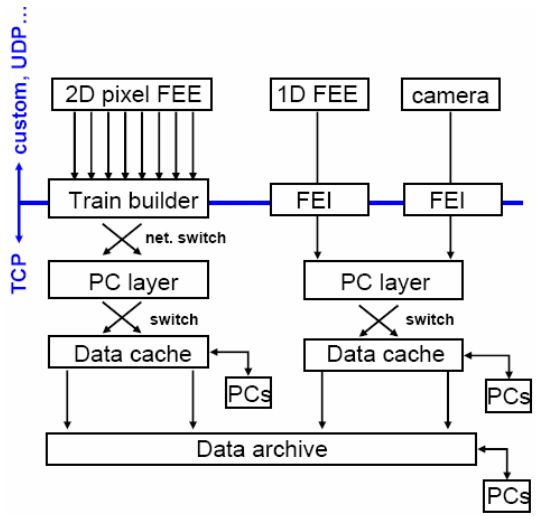
- Detector expected to be commercial camera systems
  - industrial standards readout (cameralink, firewire, etc.)
  - 4 Mpixel MP-CCD cammera being developed
  - Investigating fully depleted monolithic active pixel sensor in SOI technology
- Readout system
  - Typically VME based frame grabbers
  - Frame data is tagged with time stamp
  - Compressed
  - Sent to local storage cluster
- Processing
  - Limited data processing on local 100 blade farm
- Archiving
  - Data stored for 3 months, thereafter user responsibility
- Custom calculation boards for coherent x-ray imaging planned

⇨ Assume that significant implementation work is still in the pipeline.

XFEL readout architecture



XFEL data system architecture



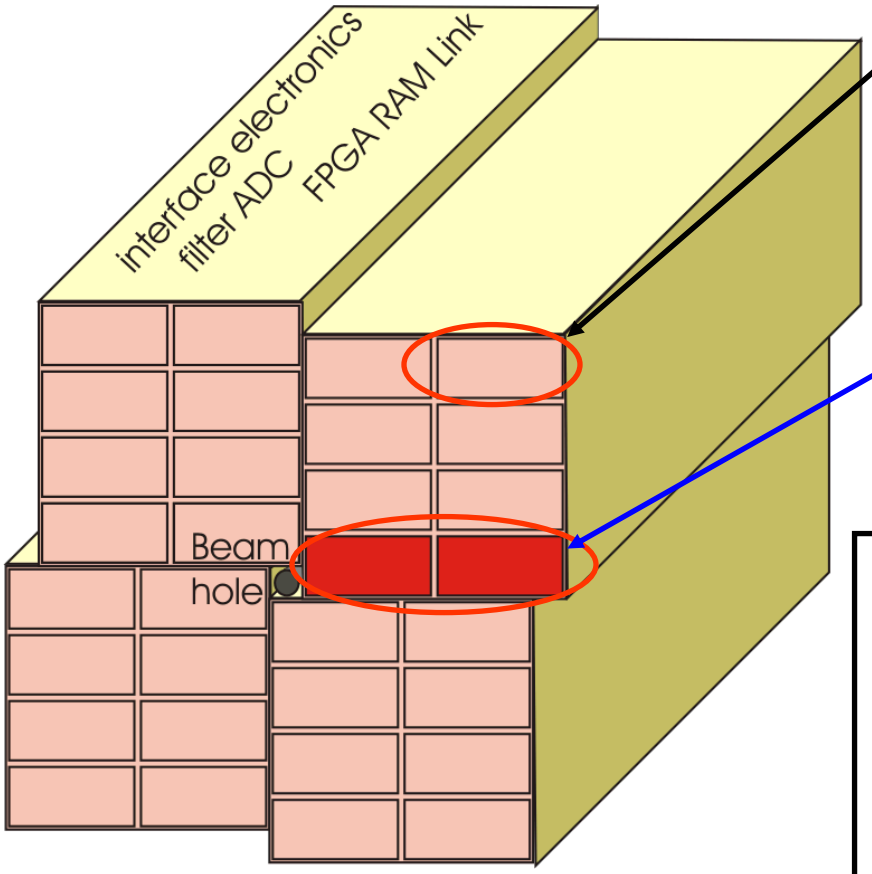
- 5 layers – current concept
  - Detector + Asic = experiment specific
  - TB + FEI = interface systems (data readout + control)
    - Train builder (TB) builds trains – need due to 2D pixel detector segmentation
    - FEI non 2D – simplified train building – for 1D and commercial cameras
  - PC layer = processing and interface to data cache
  - Data cache = processing and temporary storage
  - Data archive = tape storage and analysis
- Development driven by large data sizes from 2D pixel detectors
- Full bandwidth to data cache, but allow processing at all layers
  - Allows data reduction and rejection when analysis available
- 10 GE standard link connections
- Basic data unit corresponds to a train (i.e. all related frames)
- Archive to tape storage

HPAD FEE quick review – surrogate for XFEL 2D

32 Sensor module:

128x256pixel^2

2.5x5cm^2



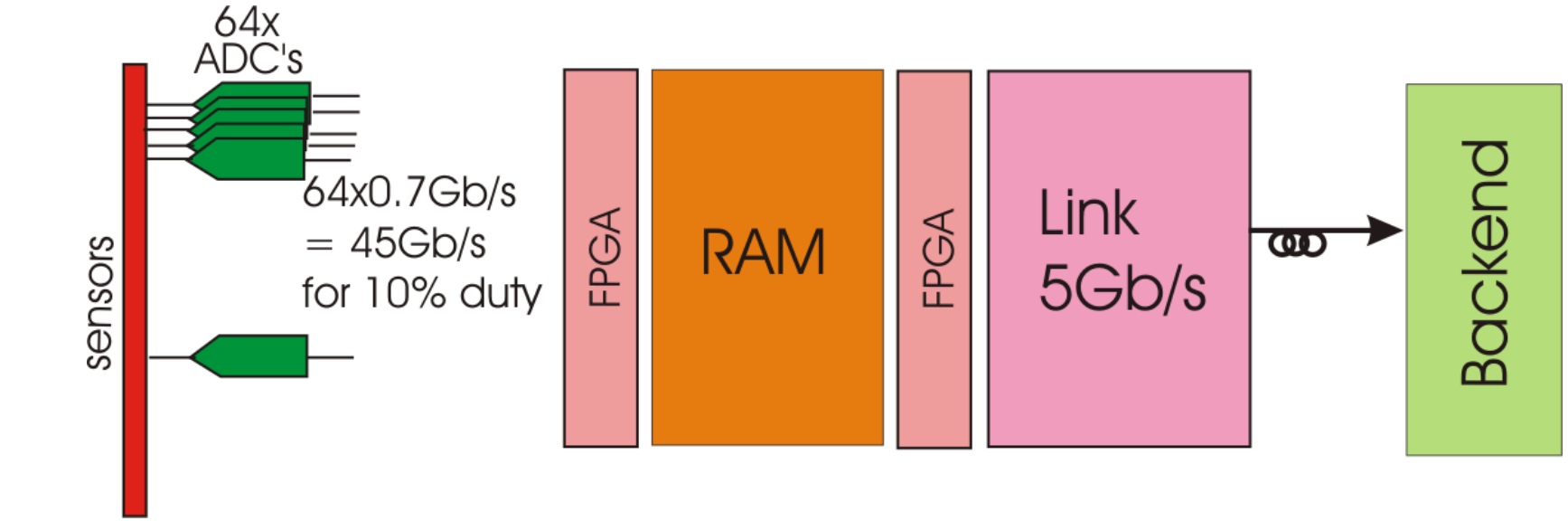
16 x Links to backend

TB (10Gb)

Pixel:

- 200µm^2
- dynamic gain: 4 states or 2bit
- 400 storage-cap: bunches/train
- 14 bit ADC's, 10-12 required

Detector head data processing



XFEL trains

Transfer 10Gb links

Formatting

ADC

Sensor and bump bonded ASIC

Data process in train gaps

Link to backend TB numbers

Data volume/train	**items…factors**	**data rate**
	10 trains/sec	
400picture	400 picture/train	4Kpicture/sec
6.4Gbit/Mpixel	16(M)bit/picture/(M)pixel	64Gbit/sec/Mpixel
0.4Gbit/link	16links/Mpixel	4Gbit/sec/link
0.5Gbit/link	+ some overhead	5Gbit/sec/link

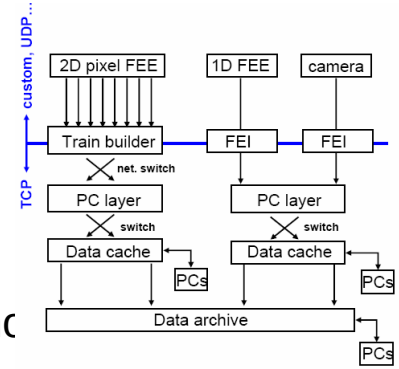
Mostly: k=1024, $M=1024^2$, $G=1024^3$

.... Links occupancy ~50% with 10Gbit/sec

.... Data volume reasonable

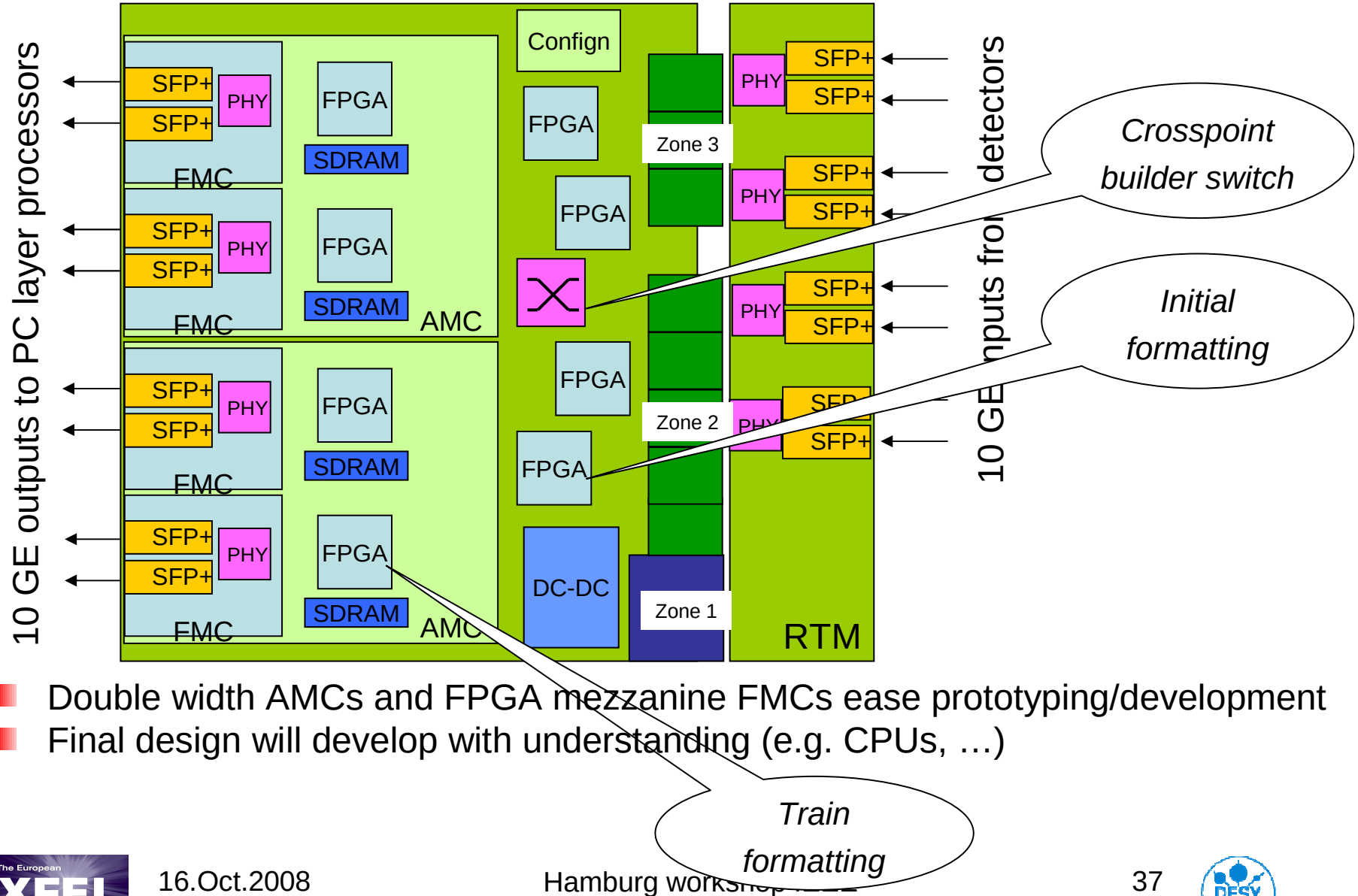
.... Data access rate defines the design (for interface electronics)

XFEL DAQ xTCA hardware



- Crate standard at XFEL is xTCA
- TB or Train Builder
  - Want to rebuild trains from frame fragments of 2D pixel detector mc
  - Development project started and has fixed some design issues
    - Input and output links are 10GE SFP+, use same transceiver/PHY layer at TB + FEE
    - 16 inputs from 1Mpixel FEE, fixes modularity questions at FEE
    - Initial design targetting ½ Mpixel detector board (5GB/s TB) started
    - Component selection started
    - Target form factor ATCA
    - Break development into smaller parts – transceiver+PHY, FPGAs+crosspoint switch – which can be used elsewhere in DAQ
  - Scalability issues remain 1Mpixel by ~2011 may be 4Mpixel 2013 !
  - Keep open for new technology developments
- 1D and commercial camera developments
  - Need system similar to RCE development at LCLS
    - FEI is functionally identical to RCE
    - Reuse developments from TB
    - Data sizes per train small – aim at sending ordered trains to PC layer.
- Different timing system interface control needs to be handled
  - Unlike HEP the detectors will have to move between LABs for testing

A ½ Mpixel TB ATCA implementation

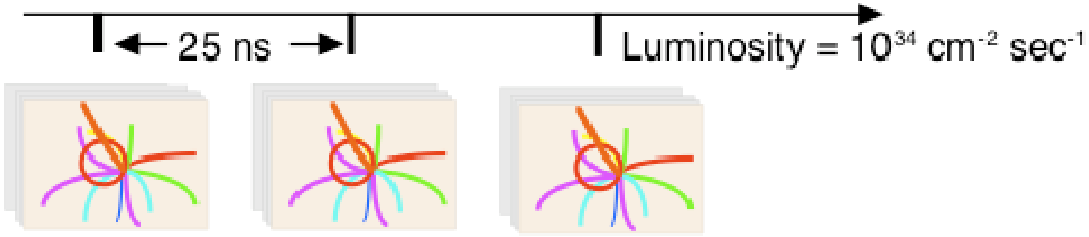


- Double width AMCs and FPGA mezzanine FMCs ease prototyping/development
- Final design will develop with understanding (e.g. CPUs, …)

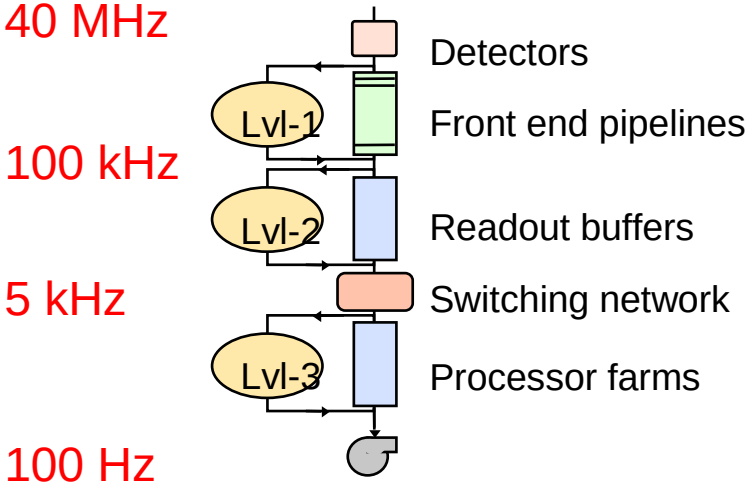
The trigger or data reduction problem

≈ 30 Collisions/25ns
(10^9 event/sec)

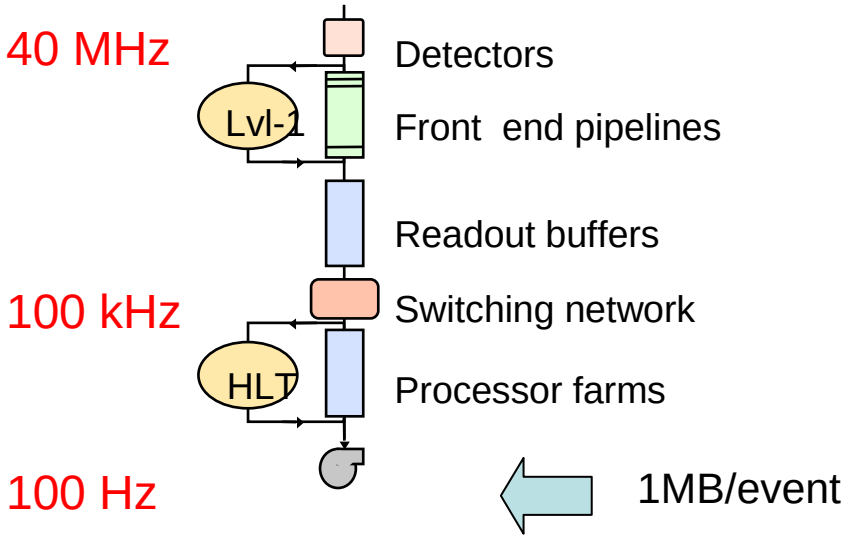
10^7 channels
(10^{16} bit/sec)



ATLAS: 3 trigger levels

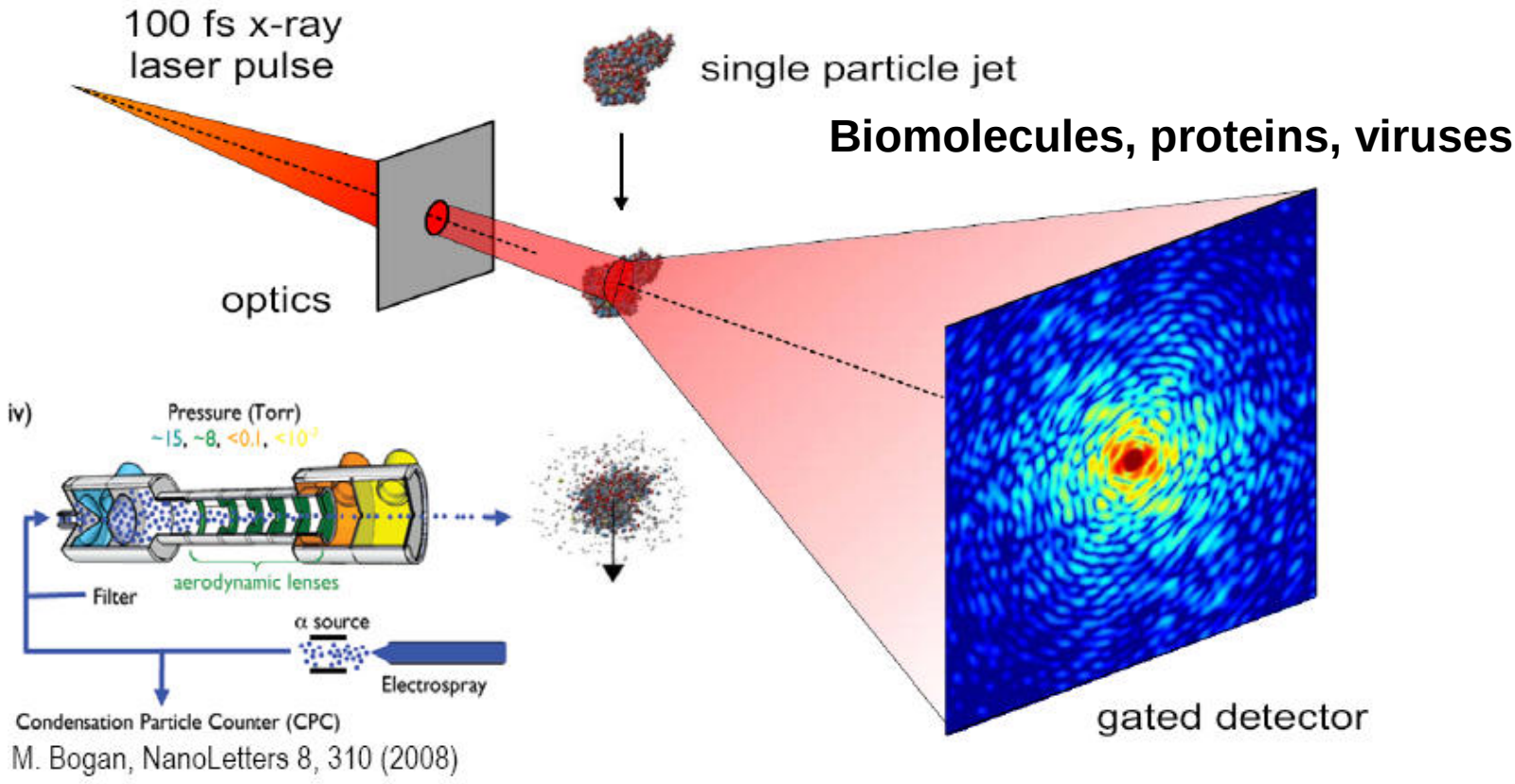


CMS: 2 trigger levels



HEP experiments employ successive trigger levels to reduce the DAQ rate. The works because “relatively” simply available trigger primitives (high transverse momentum leptons, missing energy balance, …)can be cut on.

The trigger or data reduction problem



At light sources there are usually no simple primitives = no HEP reject trigger.

Currently forced to readout all the data – this works at low rate machines (LCLS and SCCS), but is going to be the big problem at machines with large pulse rates.

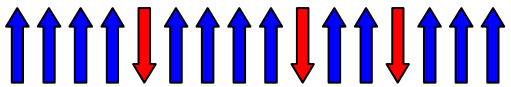
Data management, archiving and analysis

- Data caching
  - Today's disk cache systems
    - 1 GByte/s concurrent in and out rates
    - 10-400 TB storage - file system handle in hardware (FPGA)
    - These will satisfy the needs of LCLS and Spring8
  - By 2013 costs will reduce by ~5 and performance should improve (faster, more store)
    - Expect to be able to build a data caches for 1 Mpixel detectors foreseen at XFEL (≤512 frames/train, 10 GB/s input rate.

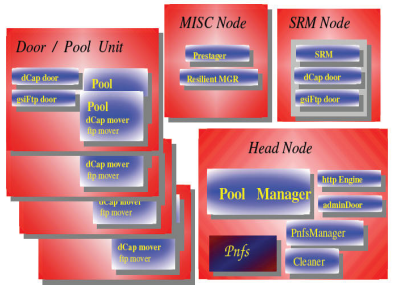
Data archiving and analysis

HEP large data bandwidth archiving and analysis solution

Archiving + Analyzers



Staging disk cache – improves
Hit rate on frequently used files



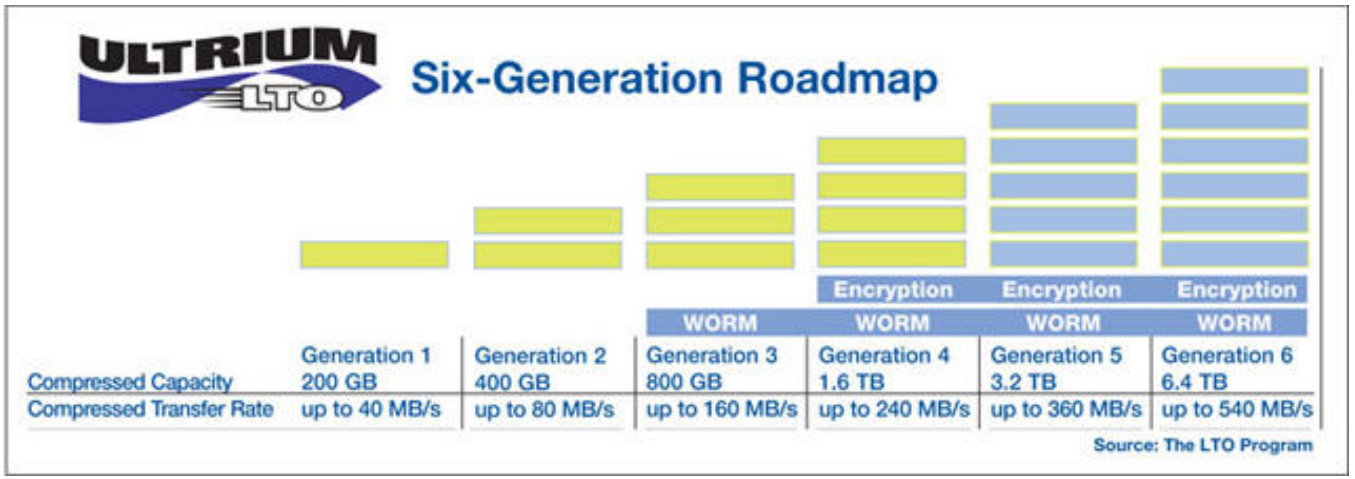
LHC now:
~? TB/day
~30-50 TB/day

Tape robot data archive

LHC now:
~? TB/day
~30-50 TB/day

Use this pattern at x-ray FELs with large data rates

performance outlook and challenge



- 2012 (for 1 silo ?)
  - 26 PB capacity (same library size as today)
  - >= 5 GB/sec (same # of drives as today)
- 2016 - factor 4 - at least by technology = new robot
  - >= 200 PB capacity
  - >= 40 GB/sec (including media change, other inefficiencies)
- Summary: storing is NOT the real problem !
  - more dependent on available money (media + personnel)

⇨ HEP experience shows: *Real Challenge* is the data access (Read) part (i.e. Analysis)

The GRID infrastructure

Huge amount of computing resources has been built and shared in the context of the computing Grid

Grid projects:

- EGEE
- NorduGrid
- OpenScience Grid

Data management will become a bigger issues – maybe less than HEP.

File catalogues, access to large numbers of files, etc. Can be solved by tools like the GRID – this is only now becoming a known name at Light Sources.

Summary of challenges

- The data volume problem
  - Volumes will increase: bigger 2D detectors, more bunches acquired, more beam lines, how much is taken and stored – scientist and detector developer input
  - Requires suitable scalable DAQ systems – but there is a limit !
  - Requires trigger processing (selection by quality) – scientist input
  - Requires scalable archiving and data access – needs money
  - Requires data and users access management – e.g. GRID type tools
  - Who is going to analyze all the data – scientist input

- DAQ hardware
  - Have arbitrary limits on bunches acquired, what is the limit, maybe lower – scientist input needed. Limited by FEE layer.
  - Advent of new technology could be useful, e.g. 100GE links just move the problems – needs data reduction and rejection

- DAQ problem is data rejection and reduction.

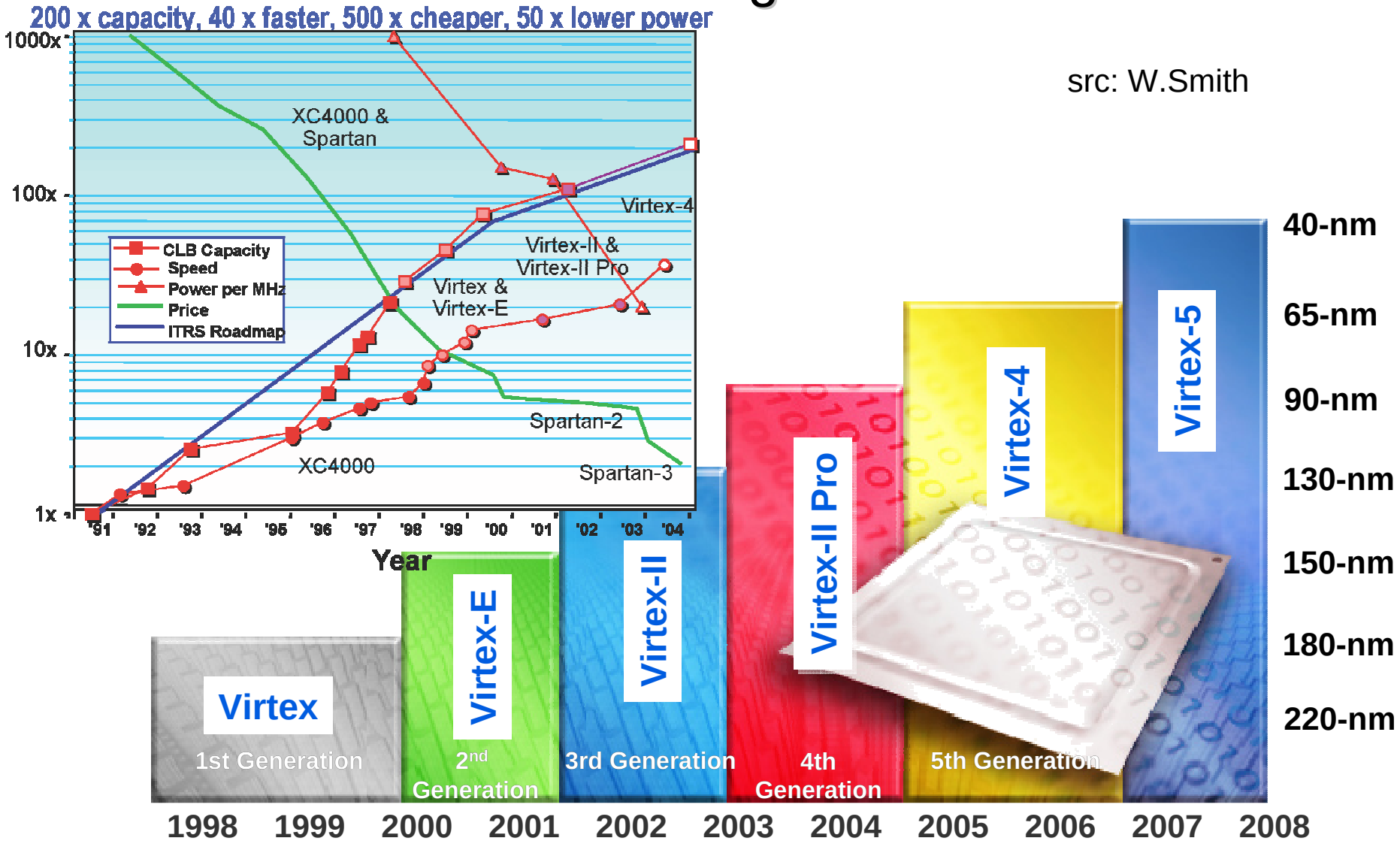
Acknowledgements

- Thanks for input and slides to:
  - Amedeo Pezzaro (LCLS)
  - Volker Guelzow,
  - Heinz Graafsma,
  - Kryzystof Wrona,
  - Patrick Fuhrmann
  - Kay Rehlich,
  - Sergey Esenov,
  - Vladimir Rybnikov (FLASH)
  - John Coughlan (TB)
  - Alessandro Polini
  - Andreas B. Meyer
  - Wesley Smith (LHC trigger)
  - Takaki Hatsui (Spring 8)

- And thanks for listening

Spares

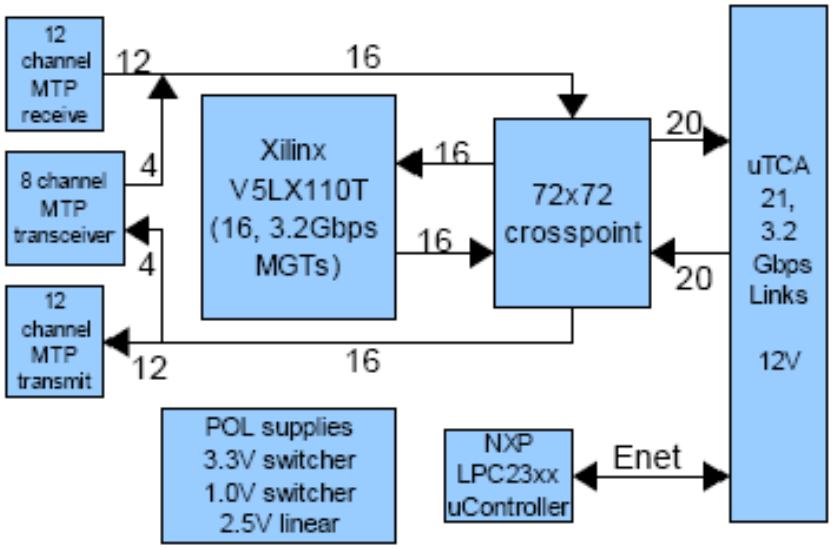
FPGA Progress



src: W.Smith

Proto. Generic Trigger System

Concept for Main Processing Card



uTCA Crate and Backplane

- The Main Processing Card (MPC):
  - Receives and transmits data via front panel optical links.
  - On board 72x72 Cross-Point Switch allows for dynamical routing of the data either to a V5 FPGA or directly to the uTCA backplane.
  - The MPC can exchange data with other MPCs either via the backplane or via the front panel optical links.
- The Custom uTCA backplane:
  - Instrumented with 2 more Cross-Point Switches for extra algorithm flexibility.
  - Allows dynamical or static routing of the data to different MPCs.