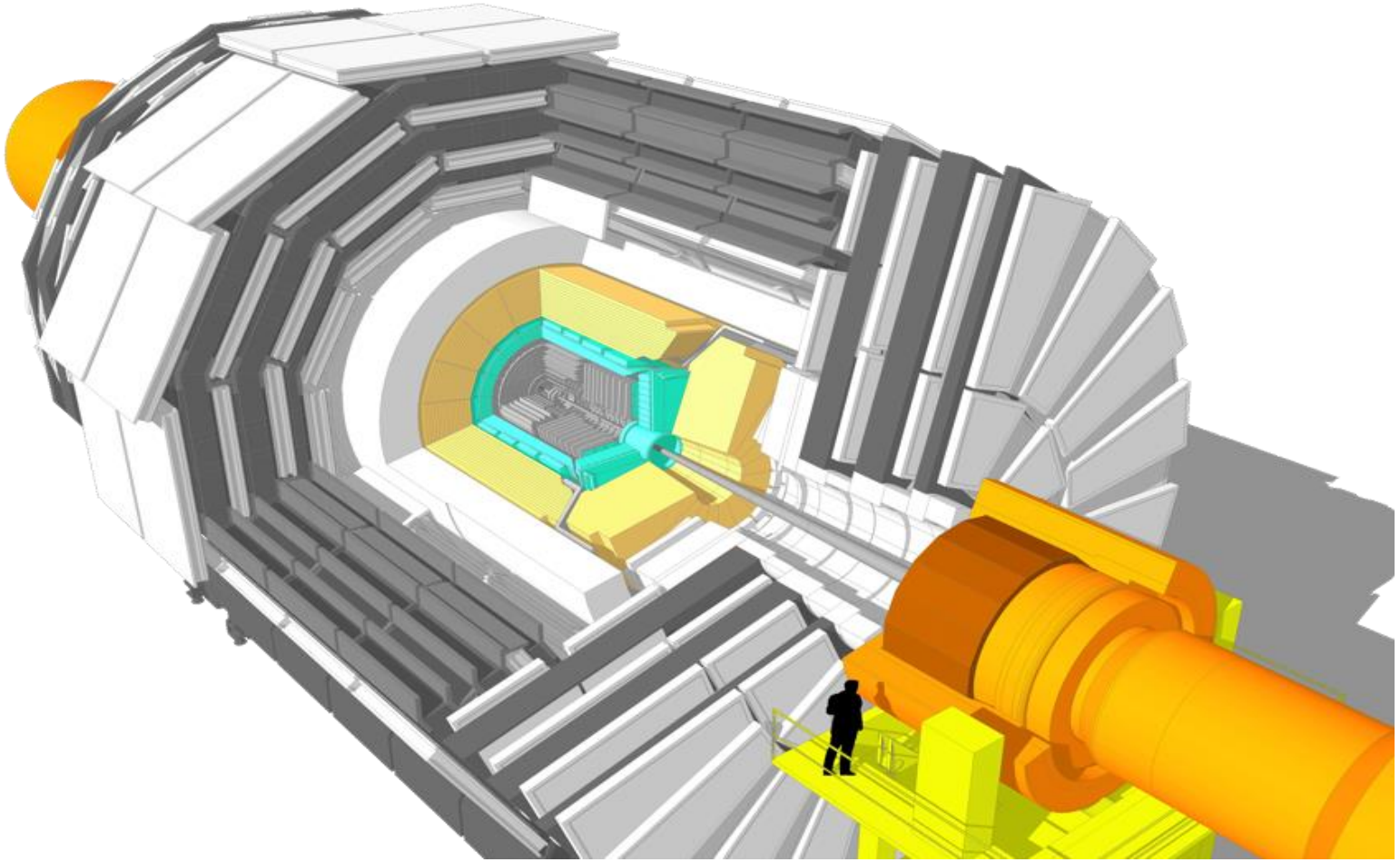
Modern DAQ Architectures for HEP Experiments

Intelligent Detectors and Smart Data Reduction in the Era of Big Data

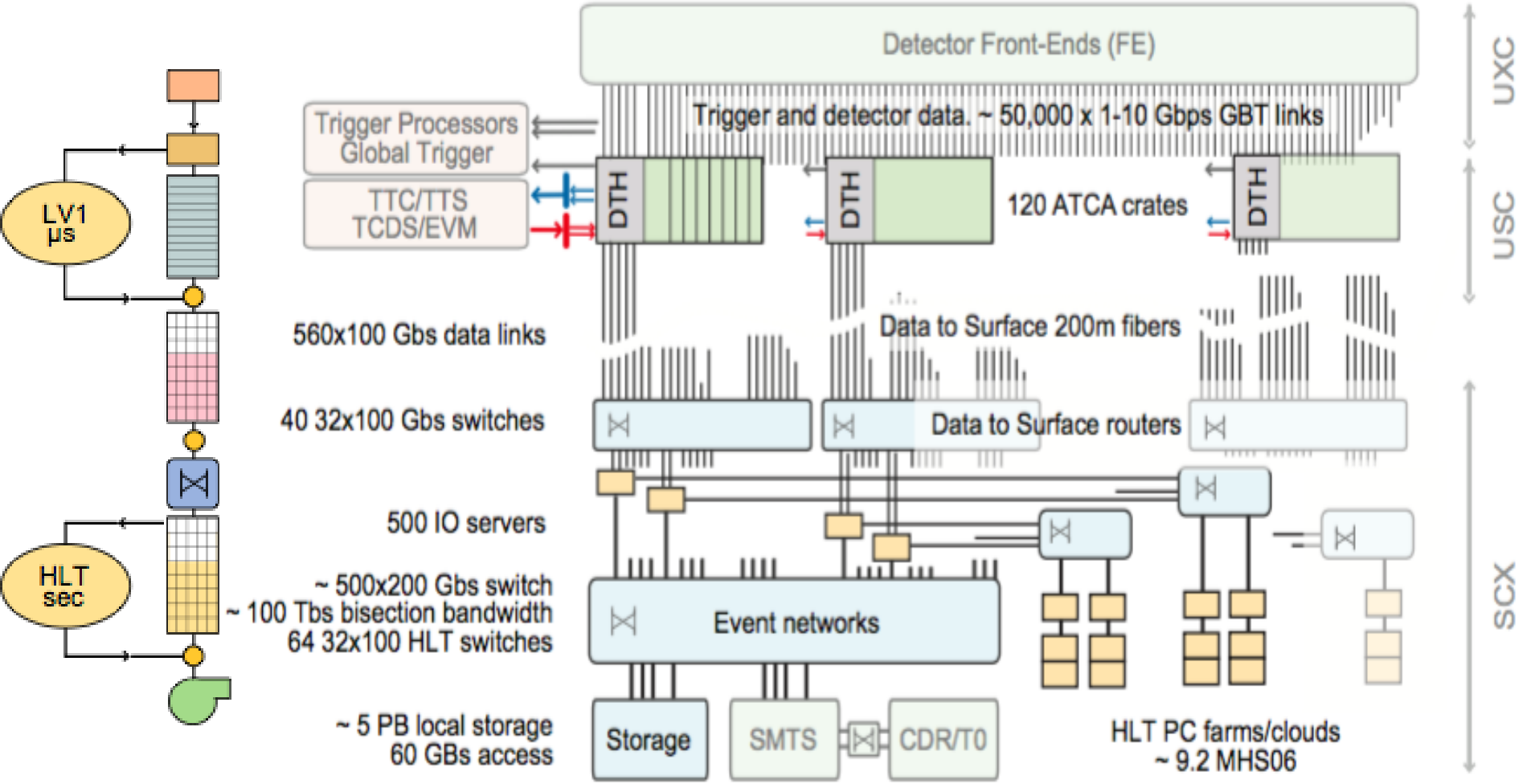
Emilio Meschi
CERN/EP

IFDEPS Workshop
Annecy 11-14 March 2018

Detector



DAQ


LV1
µs
HLT
sec
Detector Front-Ends (FE)
Trigger and detector data. ~ 50,000 x 1-10 Gbps GBT links
Trigger Processors
Global Trigger
TTC/TTS
TCDS/EVM
DTH
DTH
DTH
120 ATCA crates
560x100 Gbs data links
Data to Surface 200m fibers
40 32x100 Gbs switches
Data to Surface routers
500 IO servers
~ 500x200 Gbs switch
~ 100 Tbs bisection bandwidth
64 32x100 HLT switches
Event networks
~ 5 PB local storage
60 GBs access
Storage
SMTS
CDR/T0
HLT PC farms/clouds
~ 9.2 MHS06
UXC
USC
SCX

Menu

- Trends in HEP
- Readout and Data Acquisition
- Data Reduction: Intelligent Detectors
- Putting it all together

Trends

- Rare phenomena require larger and larger integrated luminosity

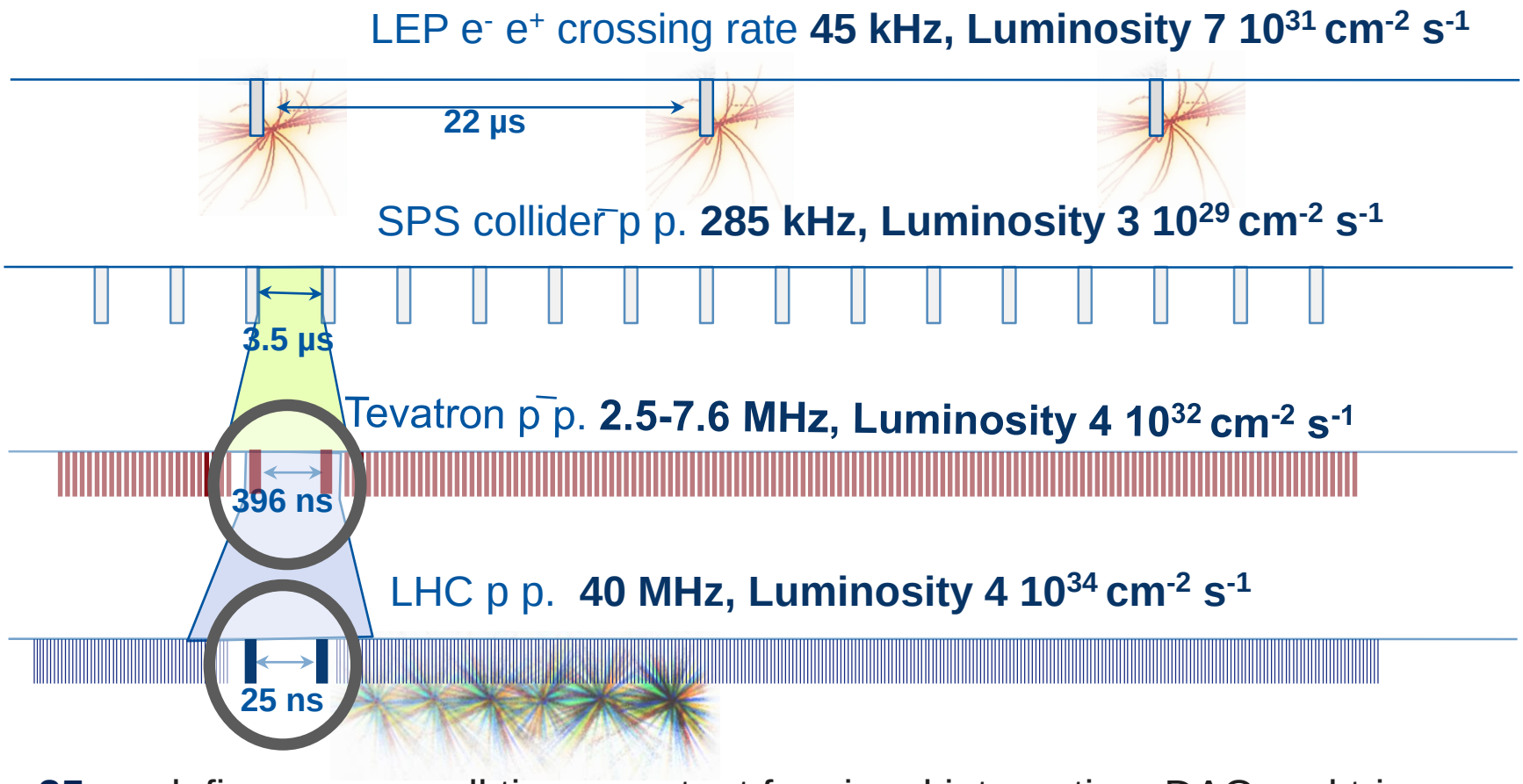
$$\frac{dR}{dt} = \mathcal{L} \cdot \sigma_p \qquad \mathcal{L} = \frac{N_1 N_2 f N_b}{4\pi\sigma_x\sigma_y}$$

  - Shorter and shorter time between collisions
  - Large numbers of interactions per crossing

$$\langle PU \rangle = dR/dt \; / \; N_b f$$

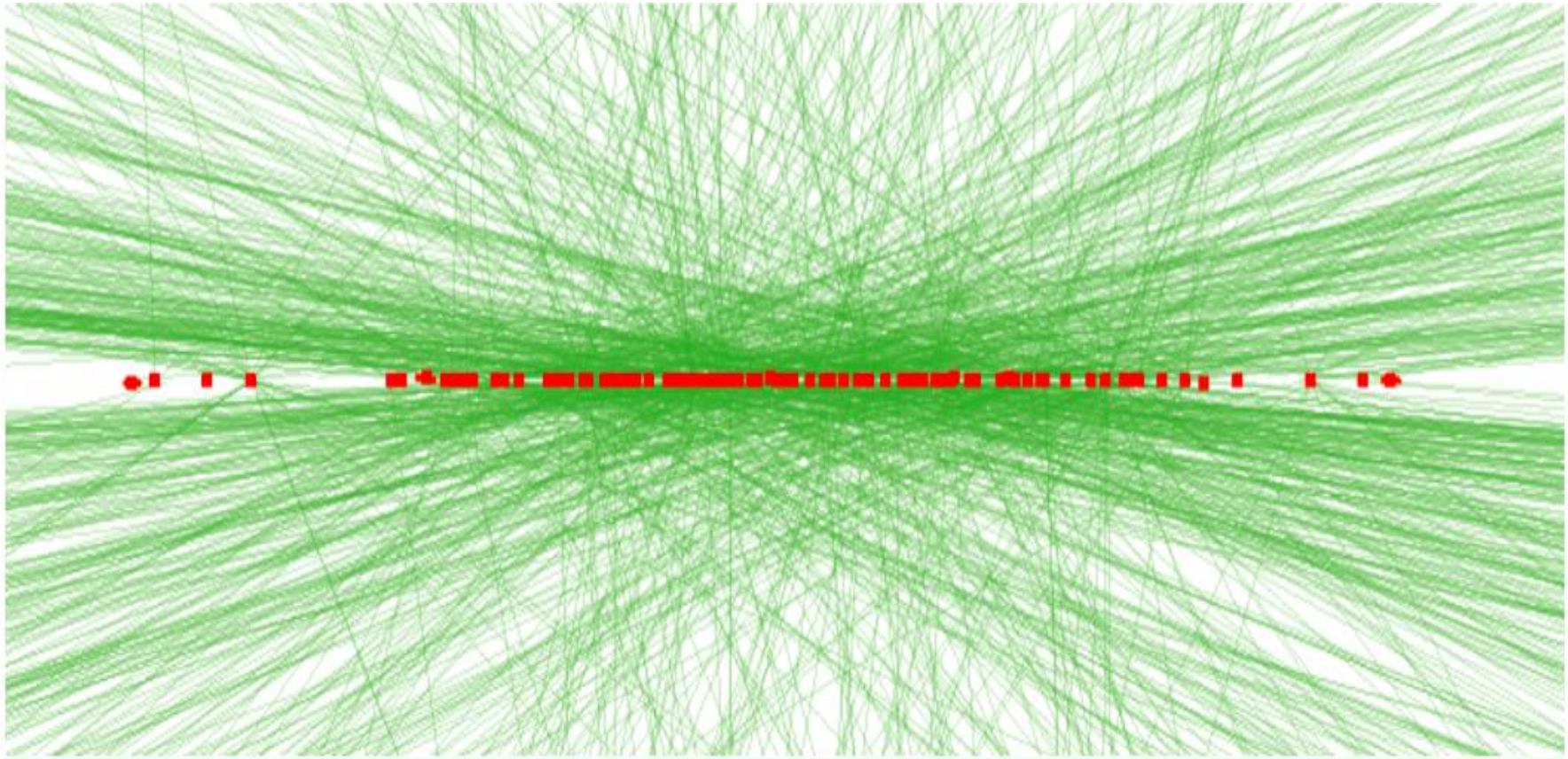
  - Interesting physics in regions with high occupancy

Colliders bunch crossing frequencies



- **25 ns** defines an overall time constant for signal integration, DAQ and trigger.
- Nota Bene: The LHC rate of collisions (**40 MHz**) was not affordable by any data taking system at the time of the first design of the LHC experiments.
- The off-line computing budget and storage capacity limit the **output rate**

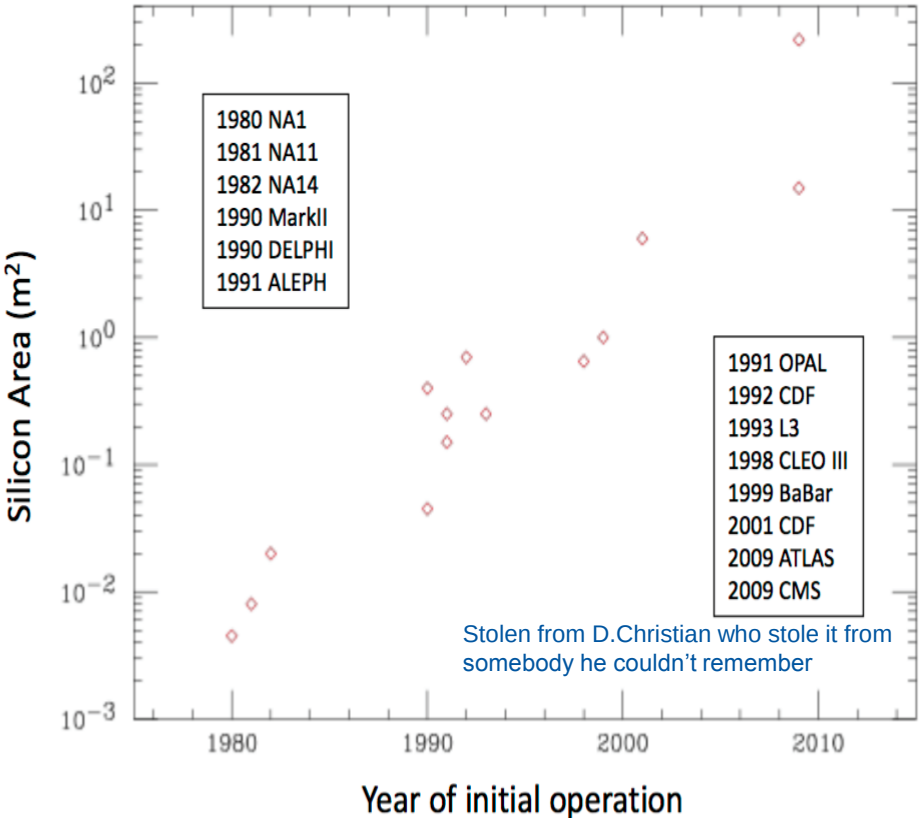
HL-LHC



Top pair event + 140 additional low energy interactions
"Classical" spatial view of the vertices

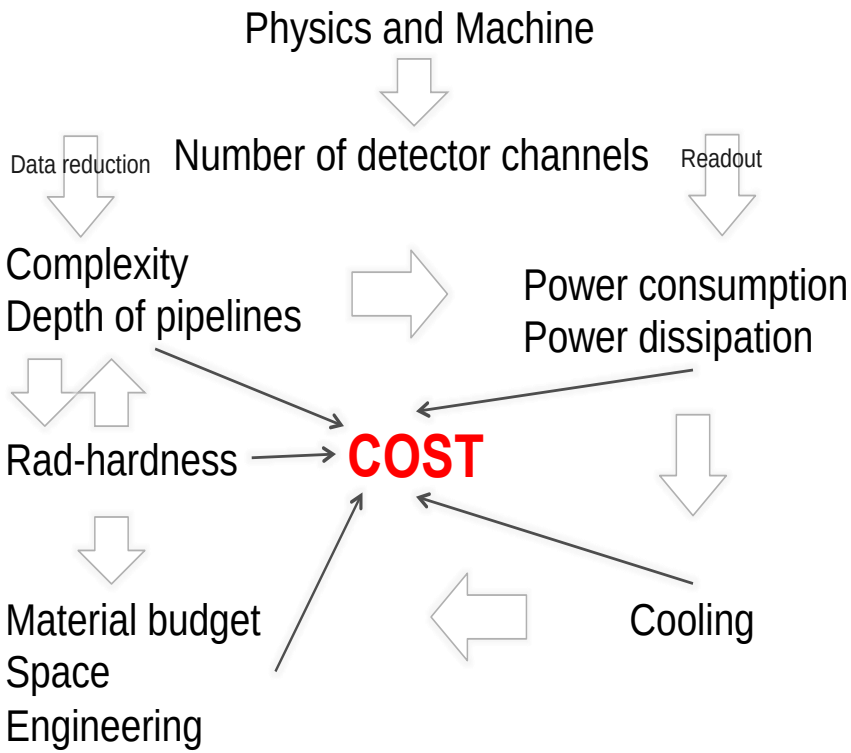
Detectors

Silicon Trackers



Stolen from D.Christian who stole it from somebody he couldn't remember

A similar exponential trend for the number of channels, from ~10000 to ~10 M over 30 years



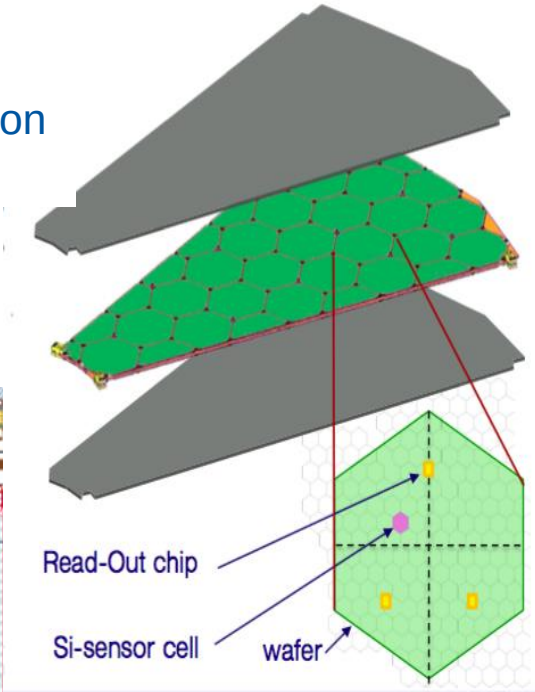
Trends

- Must cope with inherent physical limitations of calorimetry
  - Exponential increase of granularity
  - Increasingly important role of tracker-driven corrections and identification (particle flow)
  - Solid state calorimeters the next jump
- Muon identification vs. momentum accuracy
  - Modern gaseous detectors
    - To cope with larger and larger occupancy

High-Granularity Calorimeter (CMS)

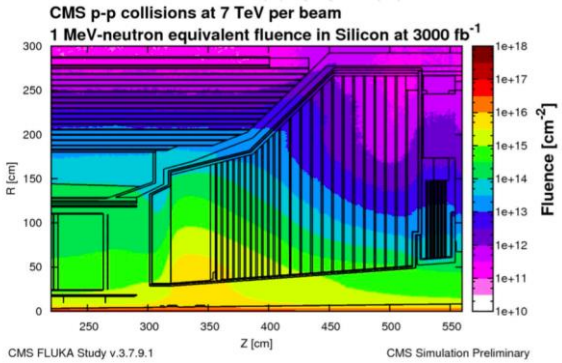
Active Elements:

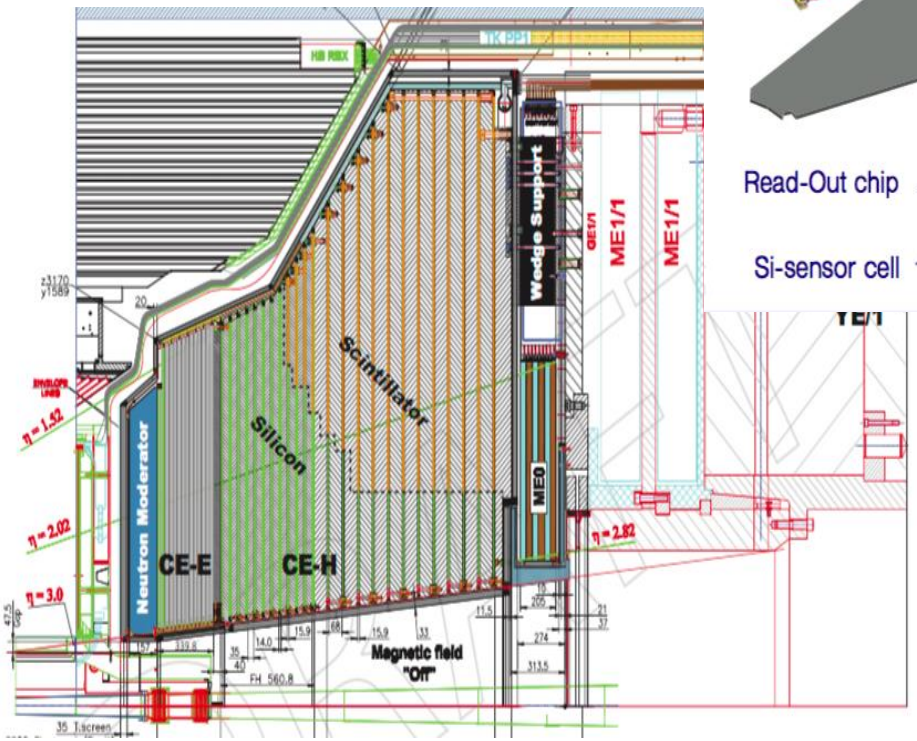
- Hexagonal modules based on Si sensors in CE-E and high radiation regions of CE-H



Key Parameters:

- EC covers $1.5 < |\eta| < 3.0$
- Full system maintained at -30 C
- ~600 m^2 of silicon sensors
- ~500 m^2 of scintillators
- 6M Si channels, 0.5 or 1 cm^2 cell size
- ~27000 Si modules
- Power at end of HL-LHC: ~100 kW per endcap





Electromagnetic calorimeter (CE-E: Si, Cu & CuW & Pb absorbers, 28 layers, 25 X0 & ~1.3 l
Hadronic calorimeter (CE-H): Si & scintillator, steel absorbers, 24 layers, ~8.5 l

Trends

Pileup ➔Number of charged tracks

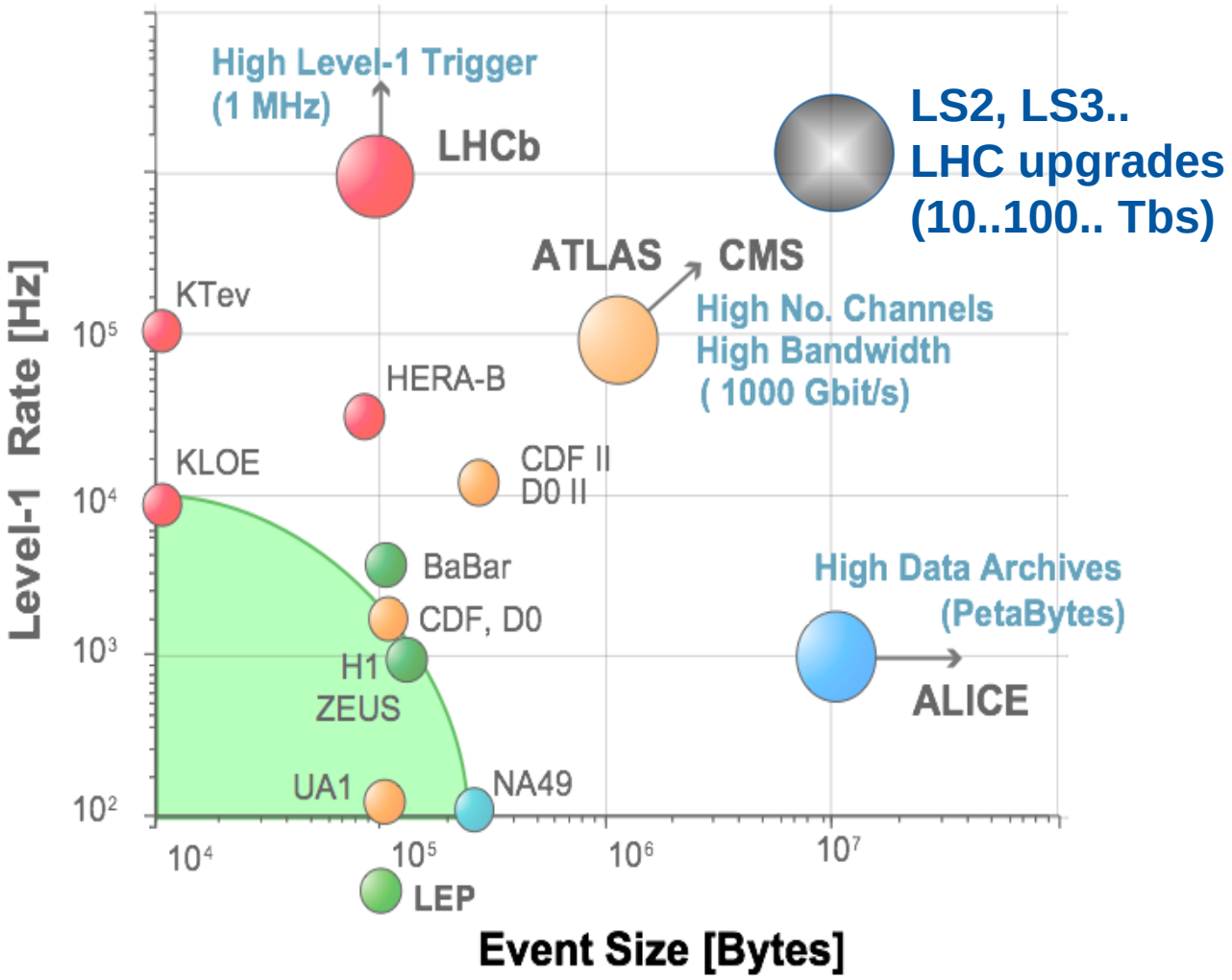
Granularity

Crossing rate / Interbunch

50Tb/s @HL-LHC after L1 trigger

Another 50 Tb/s to L1 trigger (e.g. CMS) – raw hit rate is another order of magnitude

Rate / data volume



Trends: HEP vs. Datacom

- HEP experiments become exponentially more complex
  - Predominant role of precision tracking
    - Driven by the need to provide accurate charged track momenta and vertex position (and displaced decay vertices)
    - Made possible by solid state detectors and very large scale integration
  - Huge number of channels (solid state detectors >10^7)/ Huge masses of data
- Exponential increase of bandwidth requirements
  - Serial Optical Links to the rescue
- Datacom explosion (driven by commercial traffic)
  - Higher per-link data-rates (25, 50, 100 Gb/s)
  - Multi lane links (x4, x8, x12, x16)
  - Tighter integration with electronics
  - Lower power

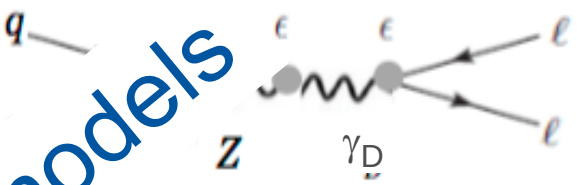
New Physics is hard to chase

- Quasi-massless:
  - H->$\gamma\gamma_D$: monoenergetic photon + MET
- Light:
  - collimated lepton pair (Lepton-Jet)
- Heavy(er):
  - non-collimated, e.g. ZZ_D to leptons

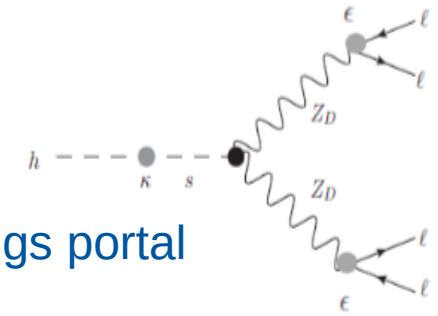
- Zero or short lifetime
- Medium lifetime
  - Prompt or displaced (LJ or lepton pairs)
- Stable

(Almost) no guidance from models

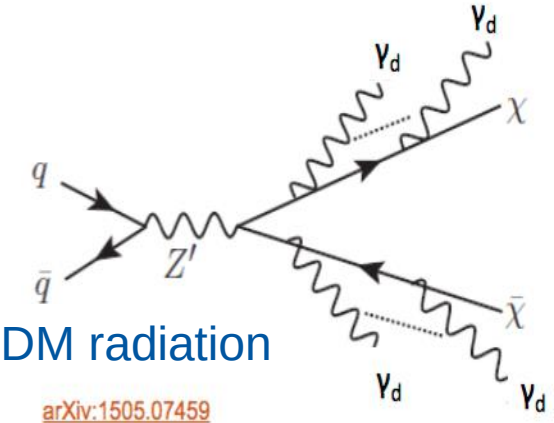
Vector portal



Higgs portal



DM radiation



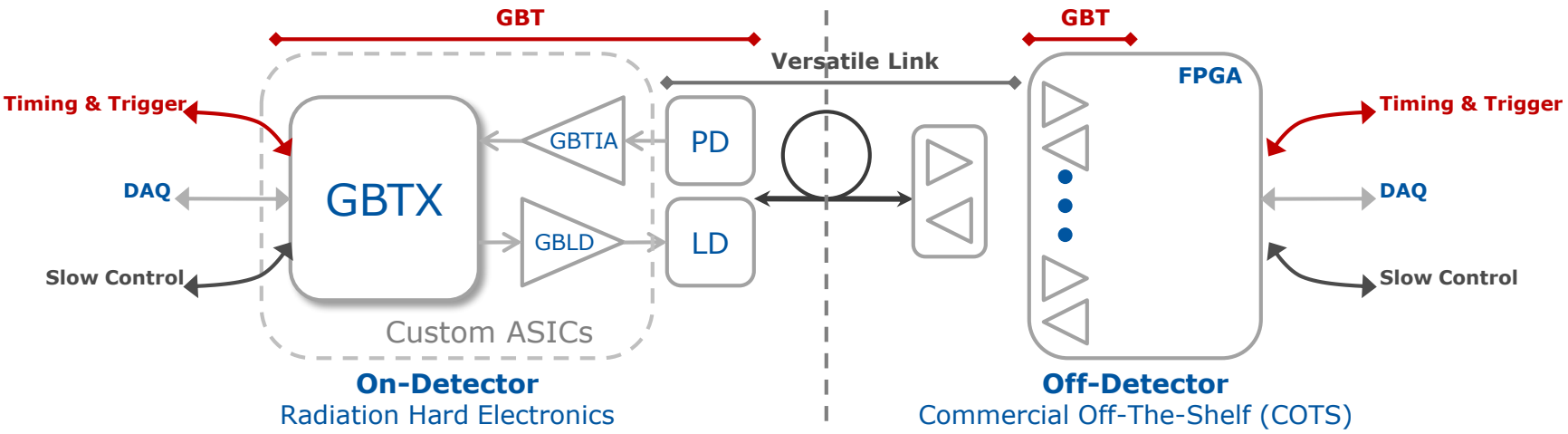
arXiv:1505.07459

Readout and Data Acquisition

Readout

- At the LHC, today's general purpose experiments each have 50-100k links
- LHC Phase I upgrades based on **Datacom technology**
- Optical Data Transmission technology is key in readout of modern HEP detectors
  - High Bandwidth (multi-Gb/s per link) **low mass, low power**
  - Immune to **electromagnetic interference** (+ isolation between power and readout)
  - Sufficiently **radiation tolerant**
- Typical rad-hard optical links of today @5Gb/s

GBT / VL



- Upstream 5 / 3.2 Gbps (low power) or 10 / 6.4 Gbps signalling / effective
- Downstream 5 Gbps

Readout

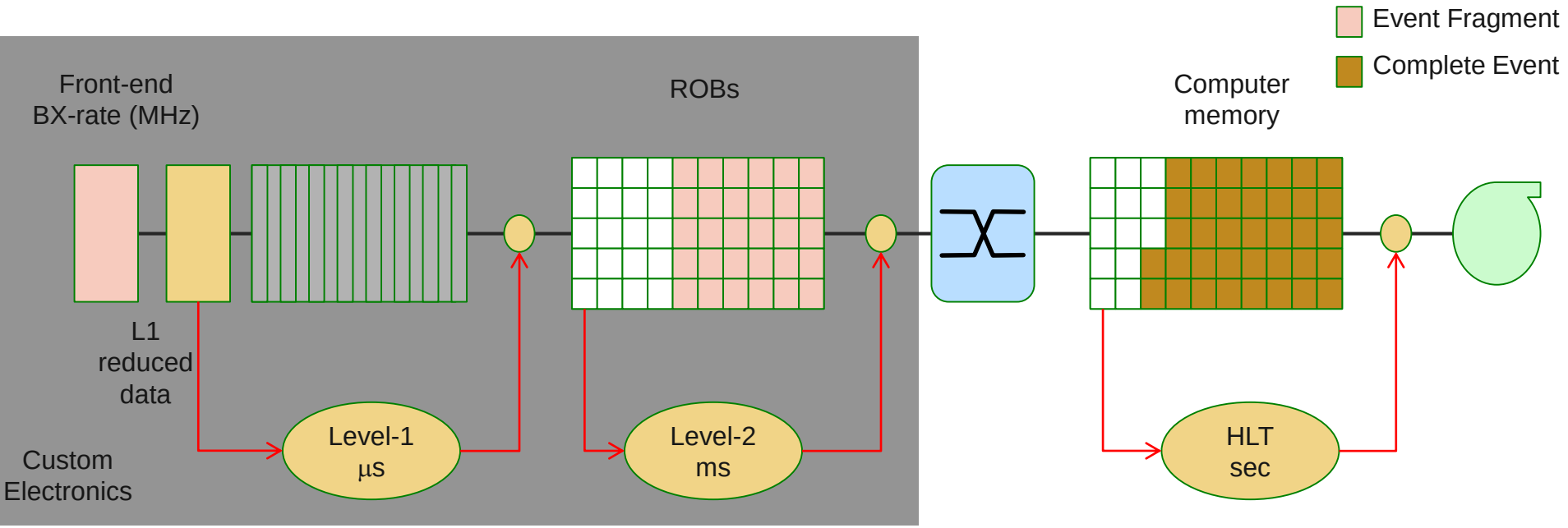
Optical layer is only part of the story

- **Acquisition, aggregation, and serialisation** before transmission over the link
- **Interconnectivity between front- and back-ends**

- **ASICs** need to be specified and designed to meet system requirements
  - Increasingly complex over generations

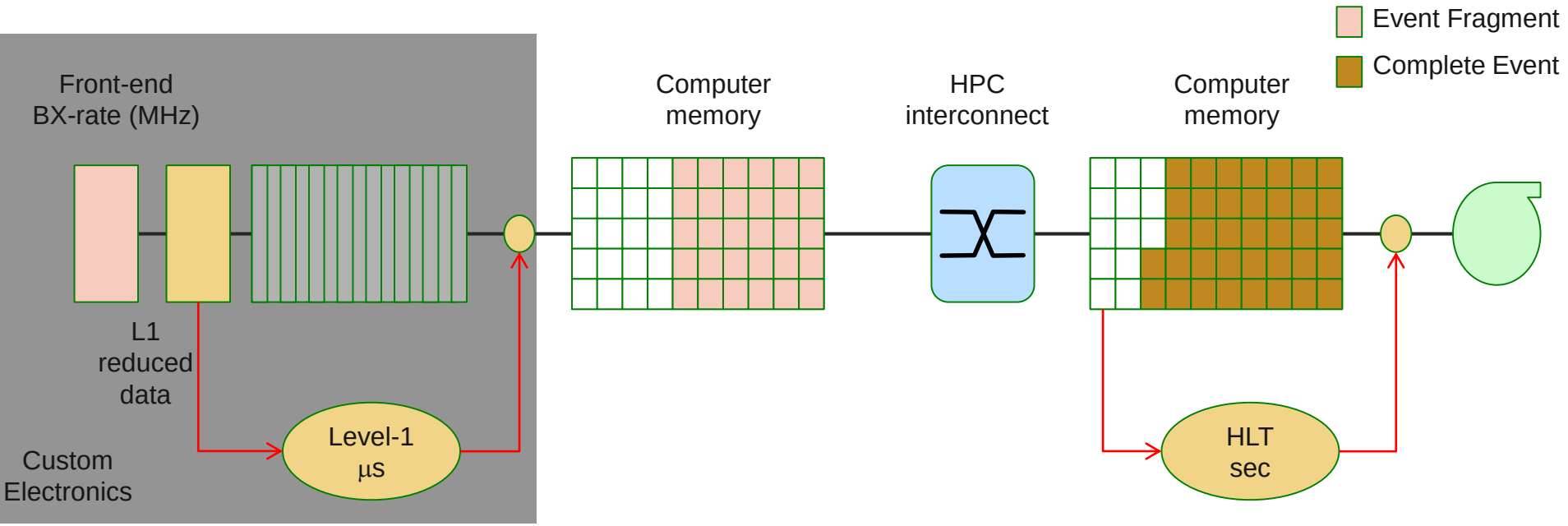
DAQ in 2000

- Pipelines, Custom multi-level trigger processors
- Simple first level with limited latency (because of pipeline depth) and input throughput
- Readout Buffers limitations require a second level before event building



DAQ in 2009

- Only one Level of Custom electronics trigger
- Algorithms limited by input throughput and number of gates in FPGA
- Memory bandwidth allows concurrent readout and event building on commercial computers

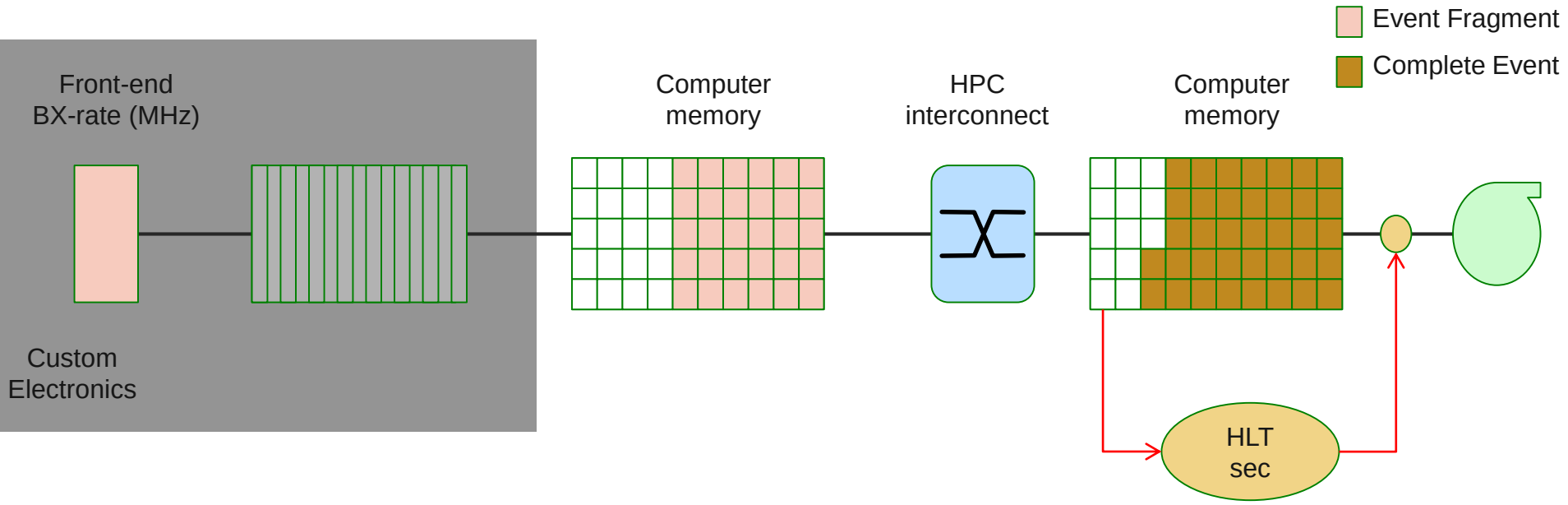


What next ?

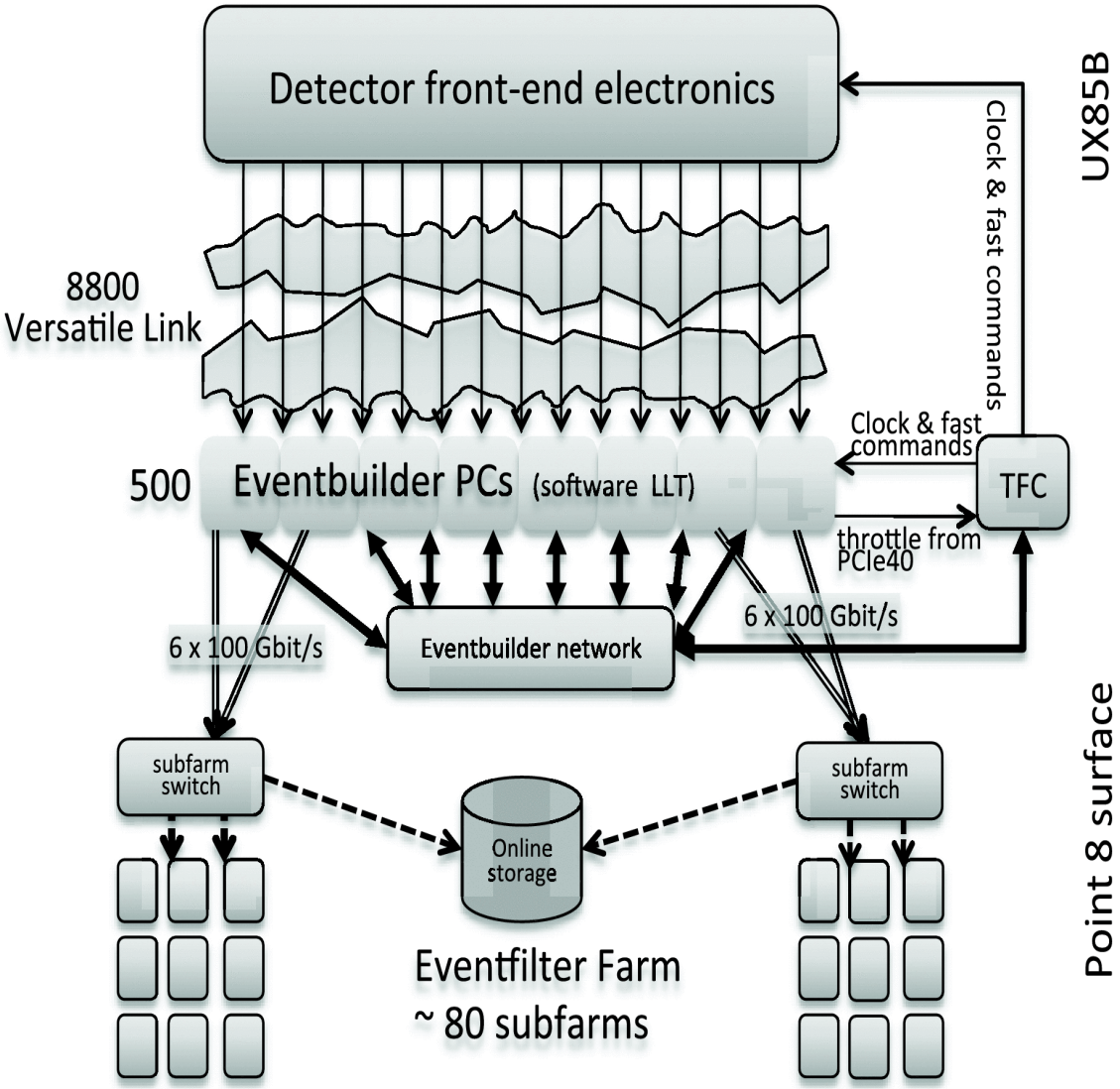
Can we get rid of custom trigger electronics ?

- Not quite… but

- We have high-speed (multi-25 Gb/s) serial links into powerful FPGAs

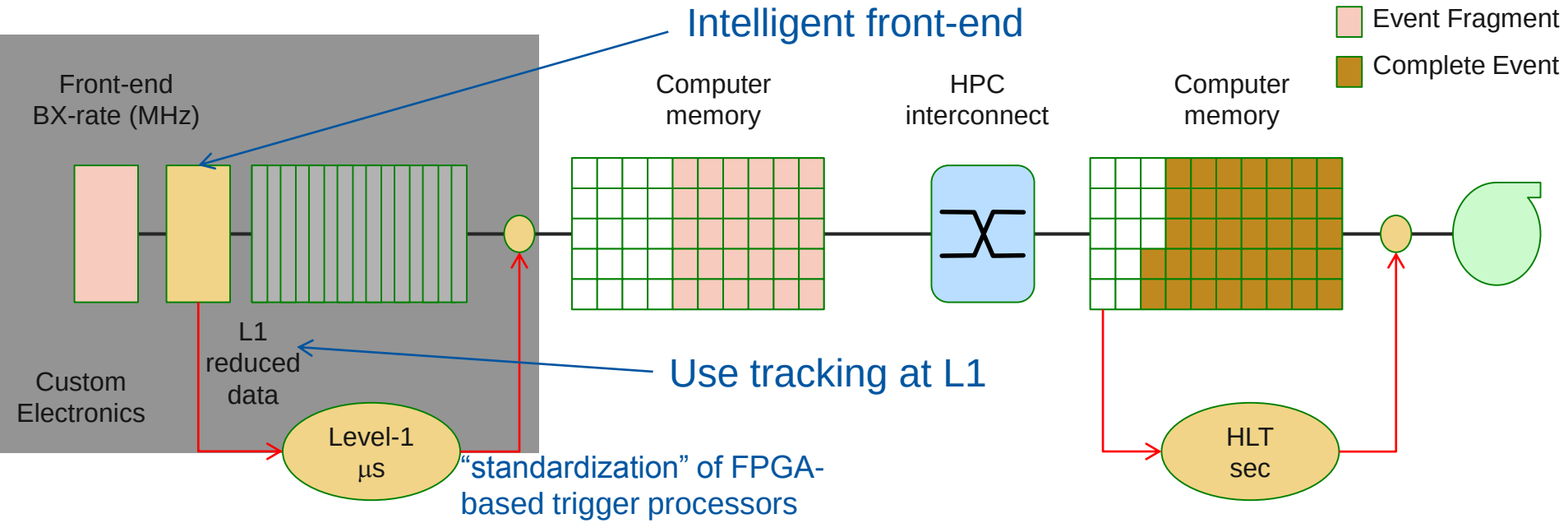


LHCb Run 3 DAQ (2021)



What Next ?

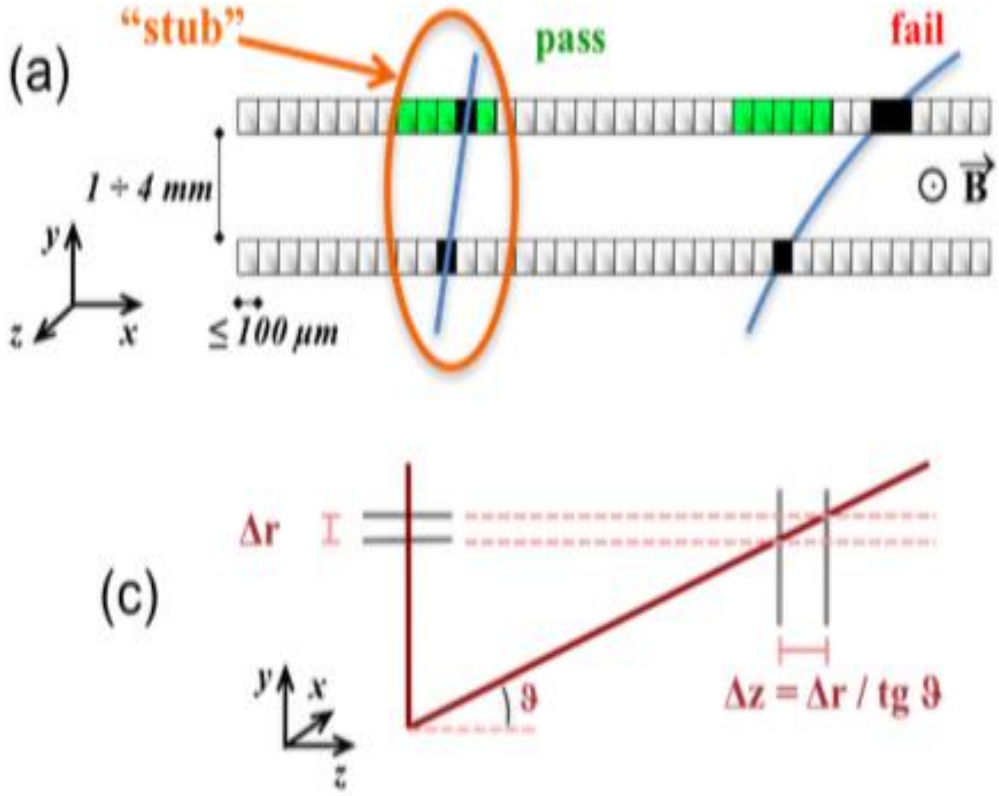
- Brute-force readout at crossing rate still difficult but…
  - Make the most of multi-Gb/s optical links
- Use “intelligent” front-end (at least to reduce L1 input)
- Large FPGA & high-level synthesis tools
  - FPGA-based “standardized” processor boards can run sophisticated trigger algorithms using tracking



Data Reduction: Intelligent detectors

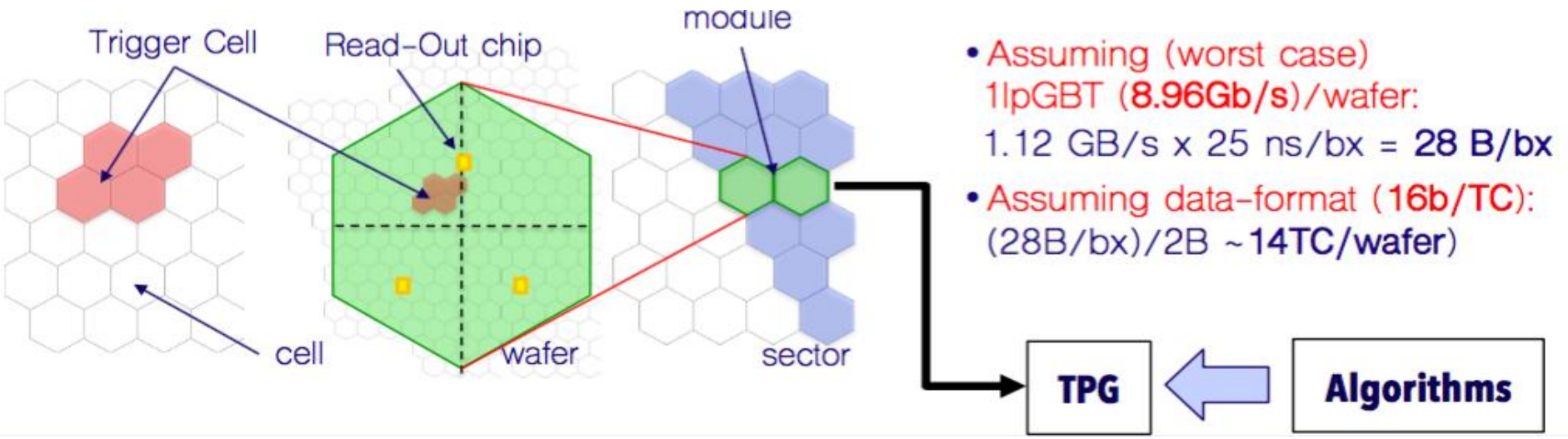
Put some intelligence in the detector…

CMS: CERN-LHCC-2017-009



Pt > 2 GeV ➔ Data reduction by one order of magnitude

CMS HGCal L1-Trigger

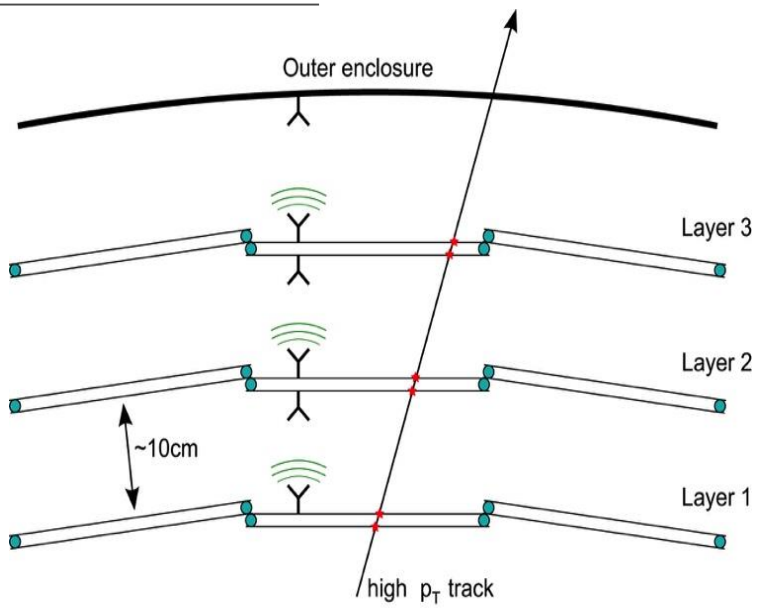


- Trigger Cell (TC) = 4 or 9 hexagonal cells
- Online data reduction at front-end
  - select average 10% TCs per module
  - The selected TCs are used as input to back-end algorithms to produce the trigger primitives (clusters)

- More intelligent detectors
  - Capable of “reconstructing themselves”
  - Require high speed connectivity from layer to layer
- Complex algorithms at front-end
  - Difficult in high-radiation environment
  - Exploit alternative technologies

More intelligent detectors

- Wireless data transmission
- 60 GHz large bandwidth potential
- Small form factor antenna
- Driven by industry



NIM A 830, 417-426 (2016)

e.g. WADAPT (S.Dittmaier et al.)
Ideal for layer distances of 1-10 cm
Must deal with reflections, cross talk, signal induced on silicon

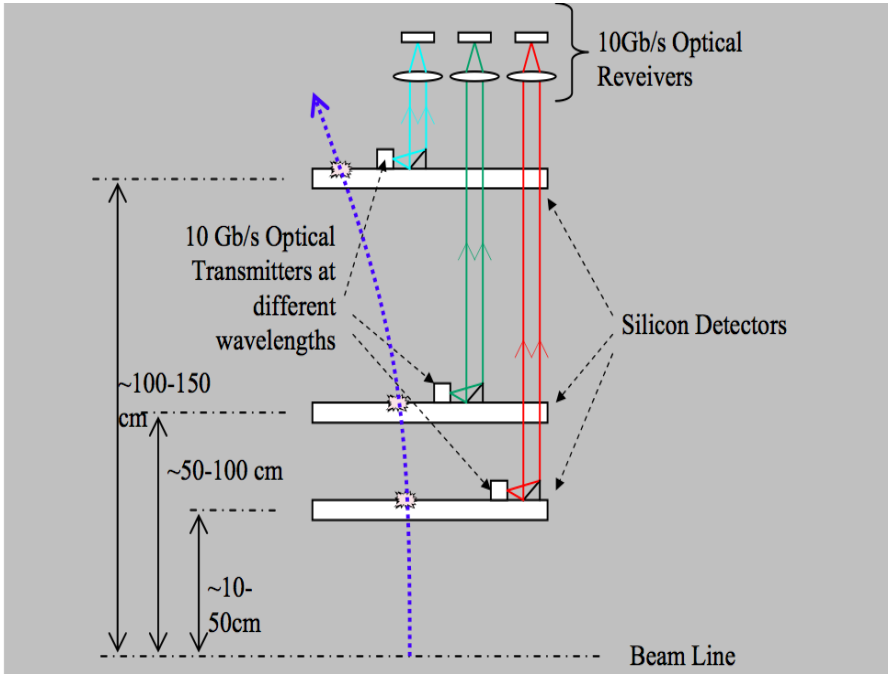
More intelligent detectors

- Electro-optical modulation
- MEMS mirrors and lenses can apply transforms

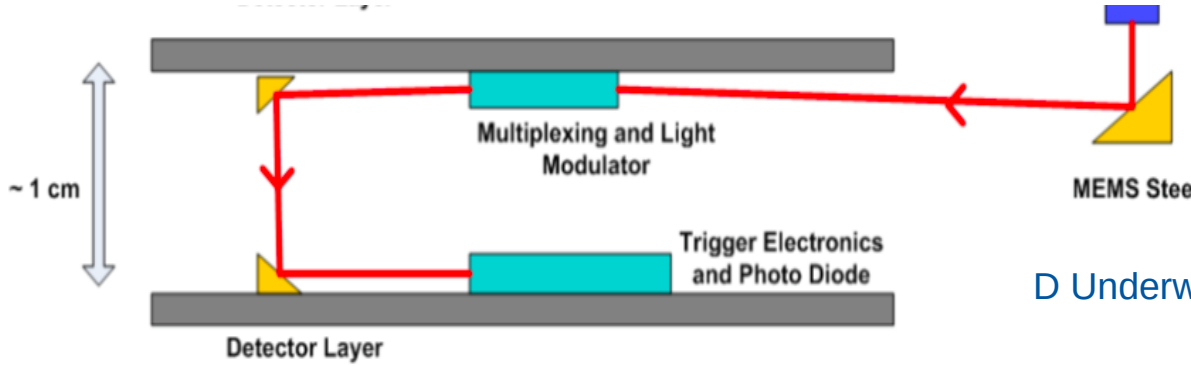
Phys. Proc. 37, 1805 (2012)

- Must deal with
  - Signals induced in silicon by laser beam
  - Source and alignment
  - Reflections

JINST 10 C08003 (2015)



http://www.sciencedirect.com/science/article/pii/S1875389212018998



Lightweight layer interconnects for trigger processing

D Underwood et al 2010 JINST 5 C07011

Optical Processing: modulators & lenses
Transforms: e.g. filter high-Pt components (in x-y)

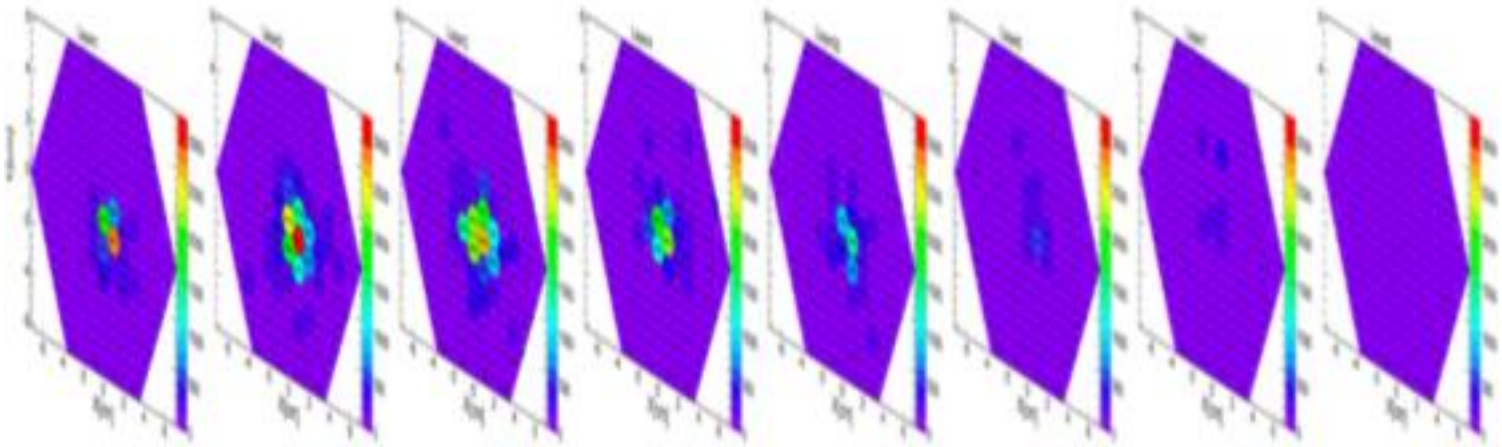


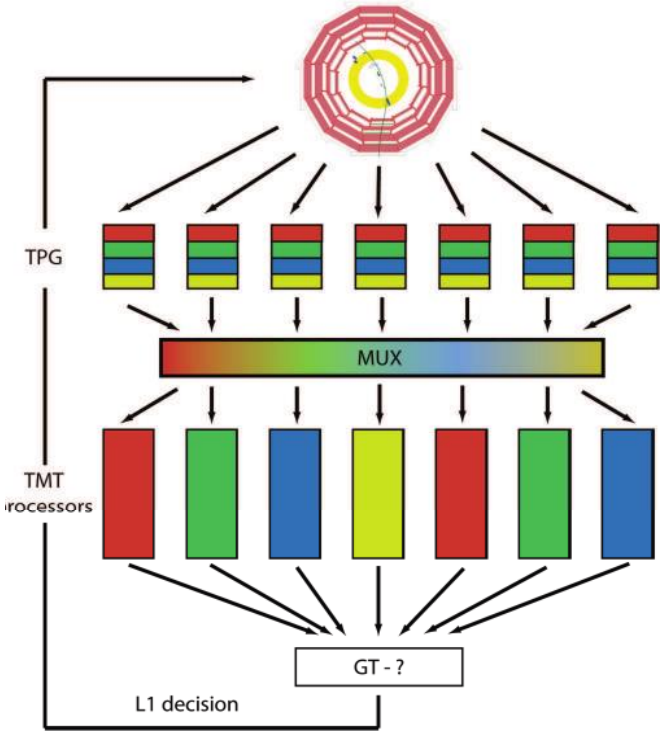
Image processing techniques at front-end
Compressive sampling

Error-correction in rad-hard environment

Data Reduction:
Preemptive Reconstruction

Time multiplexing: a new old idea

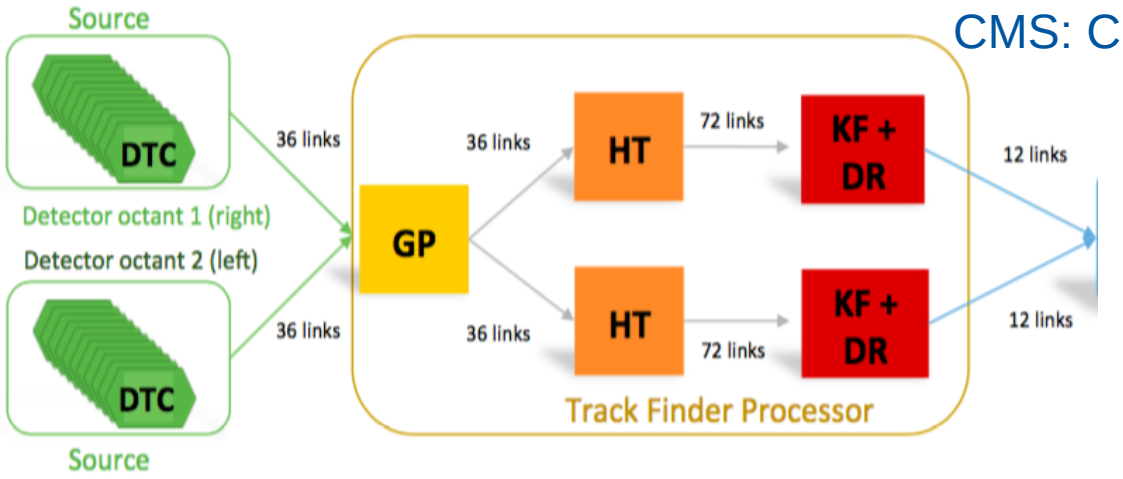
- Best when
  - input data scattered over many sources
  - But computing resources concentrated and in “small” number
  - algorithms are non-local
  - computing resources are “homogeneous”
- N.B. DAQ Event Building is (equivalent to) Time Multiplexing

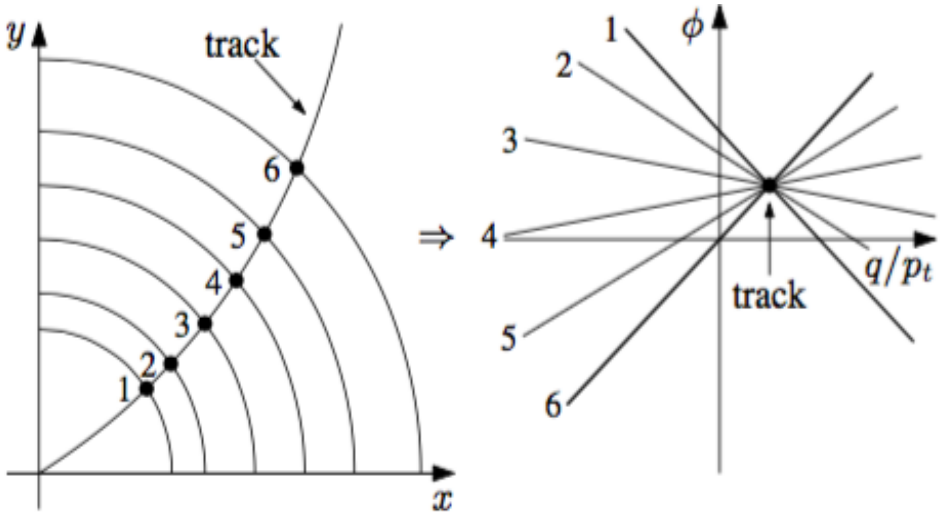


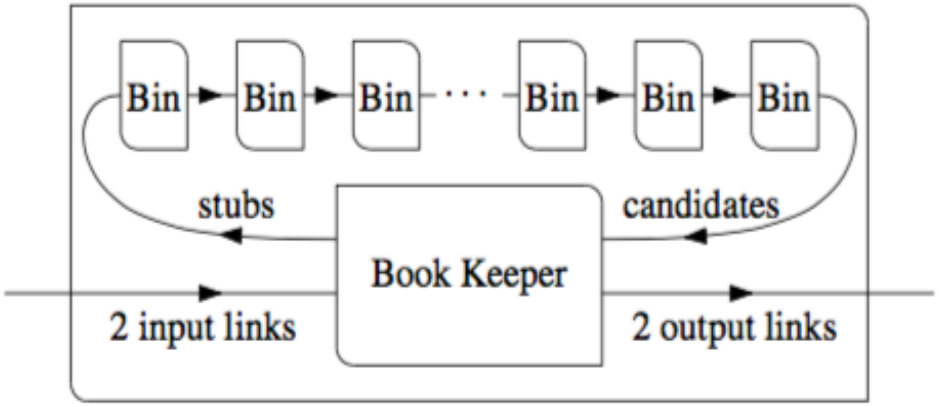
G. Hall et al 2014 JINST 9 C10034

At L1 (HL-LHC)

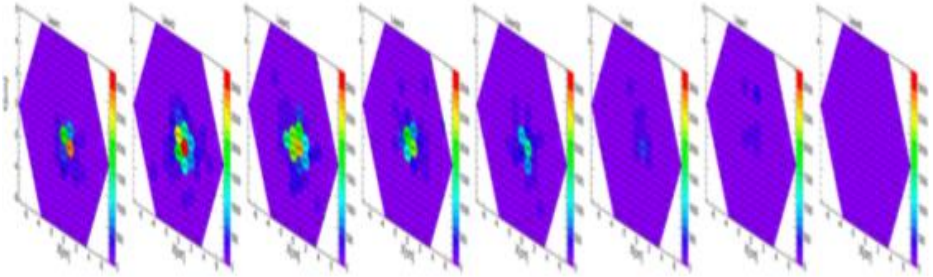
CMS: CERN-LHCC-2017-009





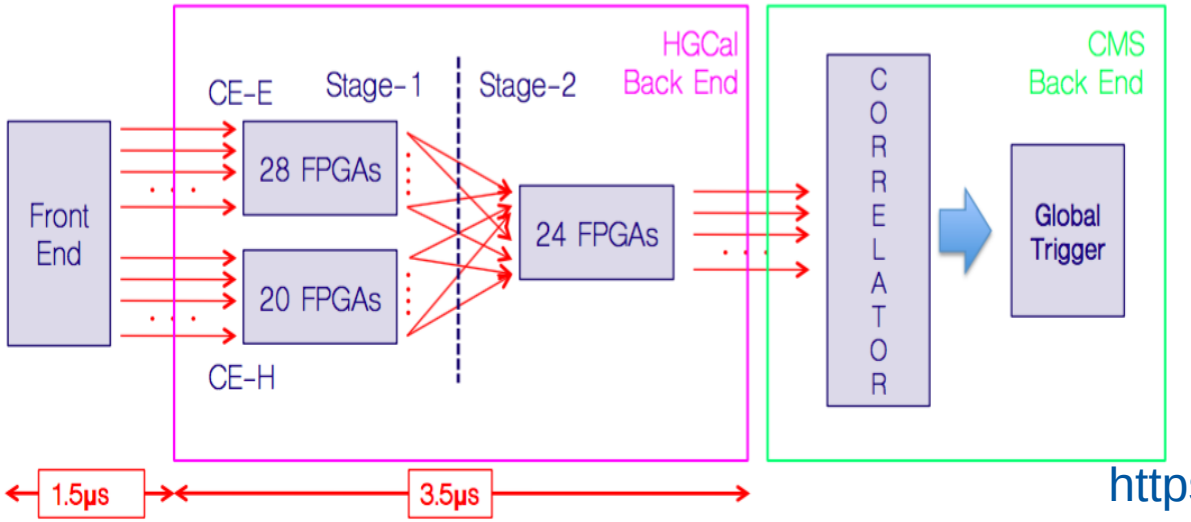


Example: CMS HGCAL Level-1 clustering



An Image processing problem

- **Stage-1:** Dynamical 2D-clustering of TCs in each layer using DBSCAN
- **Stage-2:** 3D-clustering relying on the longitudinal correlation of 2D clusters, by projecting the position of each 2D-cluster
- Stage-1 to Stage-2 x24 time-multiplexed (all data from one endcap processed by one FPGA)

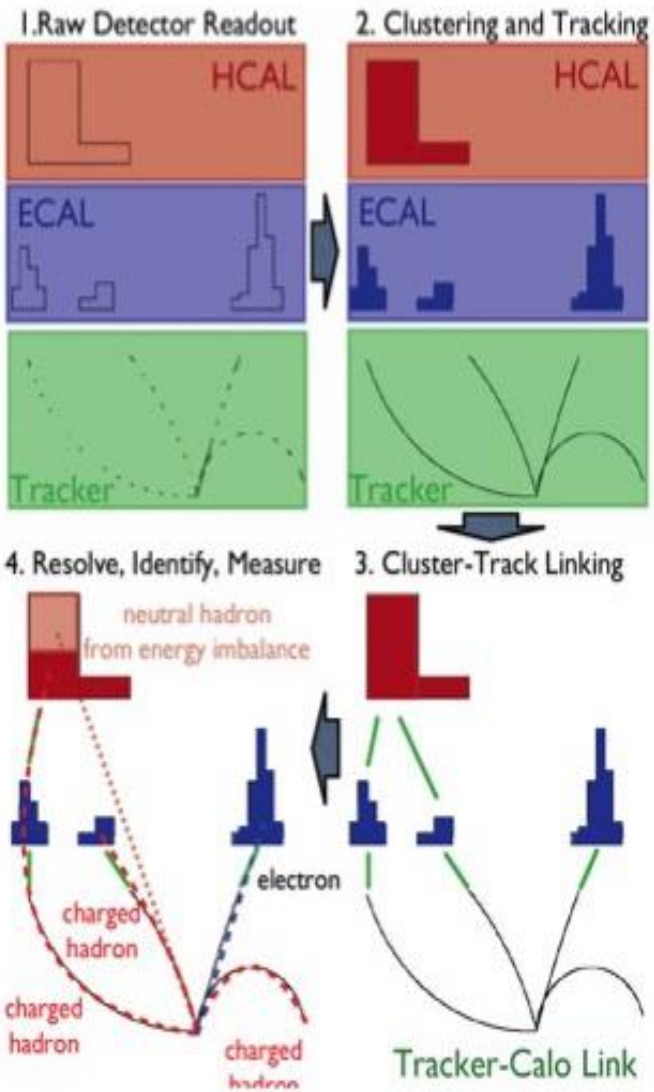


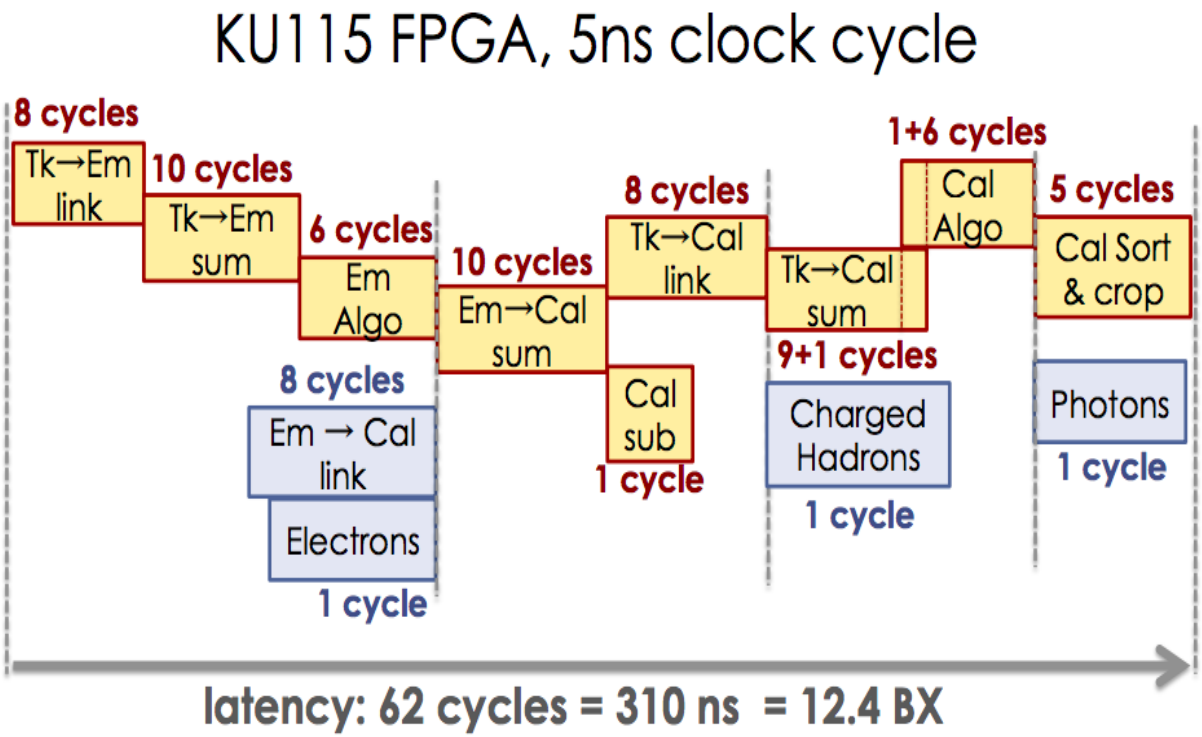
ML Alternatives (e.g. CNNs)
Hexagonal structure !

https://arxiv.org/pdf/1708.00647.pdf

Pflow @L1

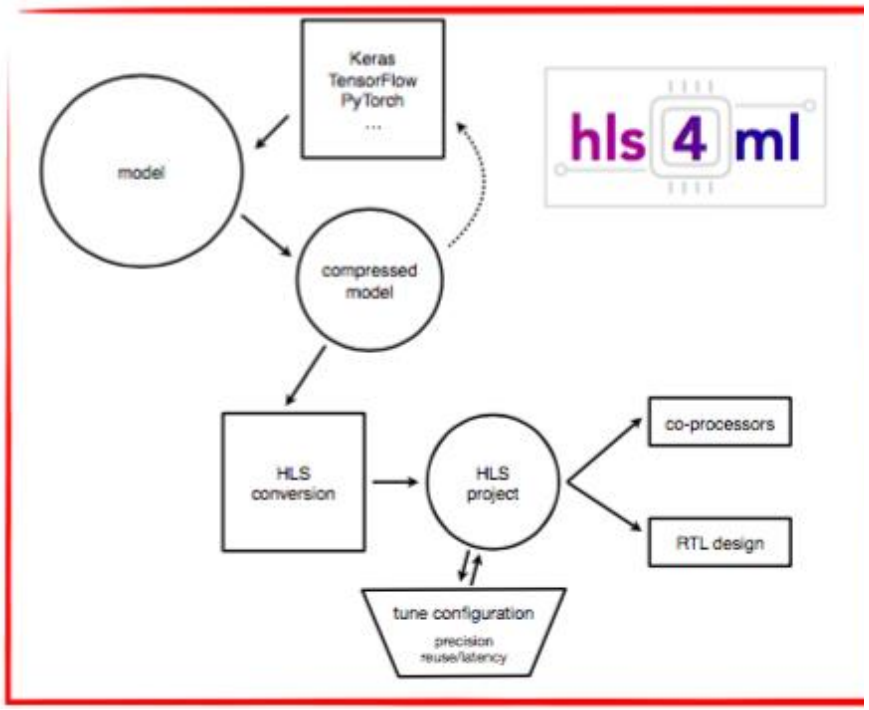
G.Petrucciani et al. (CMS)

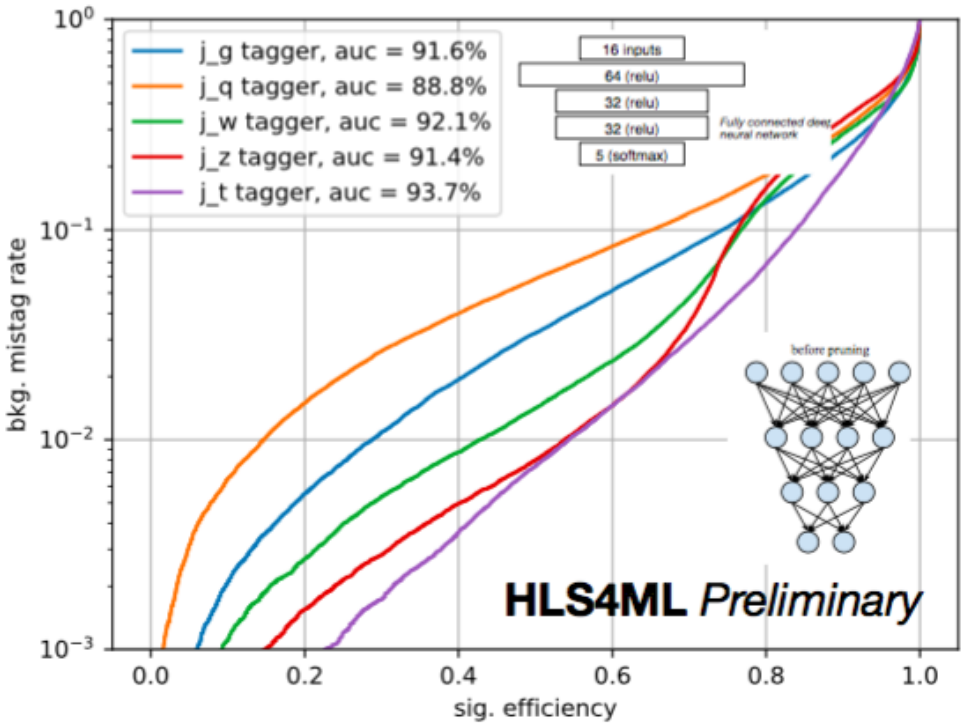




Using High Level Synthesis
Open fw programming of sophisticated algorithms to a wider community
Can be tested bit-by-bit against pure sw version

Deep learning for L1 (and HLT)





HLS4ML: CERN/FNAL/MIT joint effort
To debut at Connecting The Dots 2018 in Seattle (March 2018)

20

The Level-1 Trigger system of a HEP experiment is already a large, preemptive, parallel inference and classification engine

In summary we (will) have…

Intelligent readout, carrying out data reduction (aka low-level reconstruction)

- Help reduce the throughput requirements

Sophisticated reconstruction and selection algorithms at the first (synchronous) level

- Can profit of progress in hardware from “big data” apps
- Neural networks, deep learning
- Dedicated hardware for classification

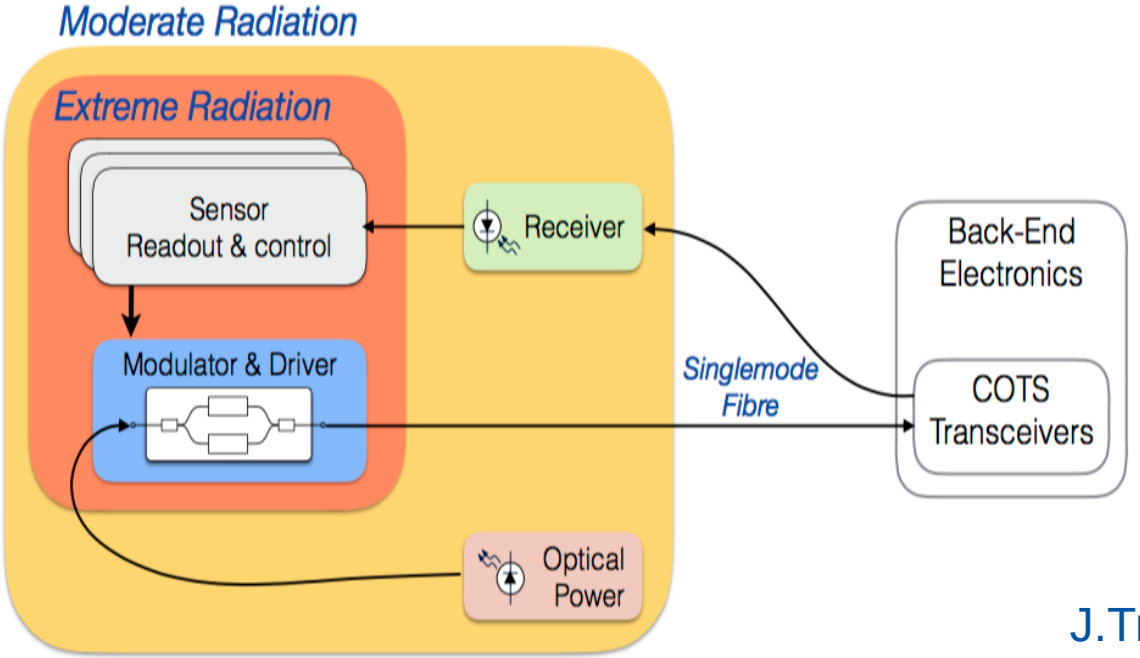
Is low-power, high bandwidth readout still worth pursuing?

Is it feasible or even desirable ?

Yes…

Silicon Photonics

- Use of silicon substrate to manipulate optical field
  - Promise of lower power and cost
- Industry: telecom applications, cluster interconnects, chip-to-chip…
- HEP



J.Troska, WIT 2017

- But…

High Level Trigger (aka Online Selection)

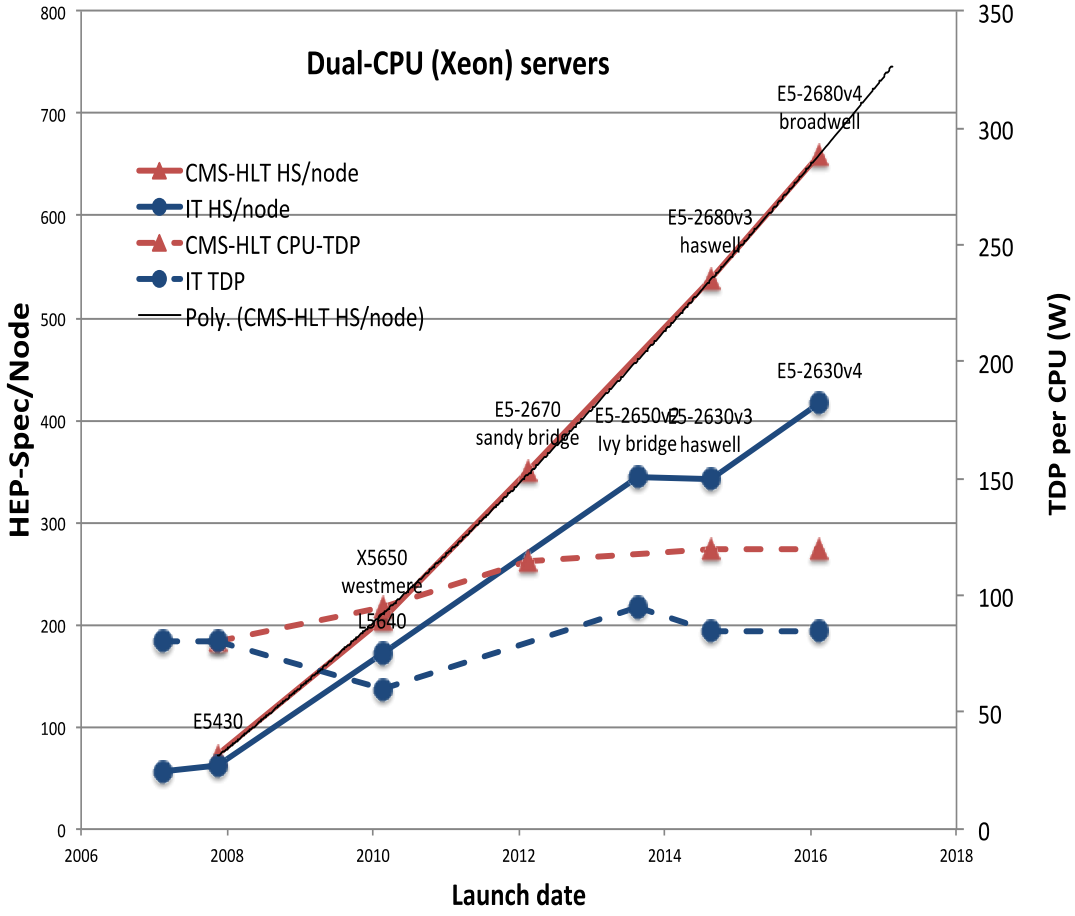
CMS: CERN-LHCC-2017-014

Traditionally working on **general-purpose CPU**

- Performance improvement per year roughly linear (or slightly quadratic)
- 2017 to 2026: Factor 2.7
- Estimated number of servers for HL-LHC order of 10000 (after L1)
- A 40 MHz system would require O(100000) servers (assuming – **optimistically** – that Level1 rate reduction is five times “easier”)

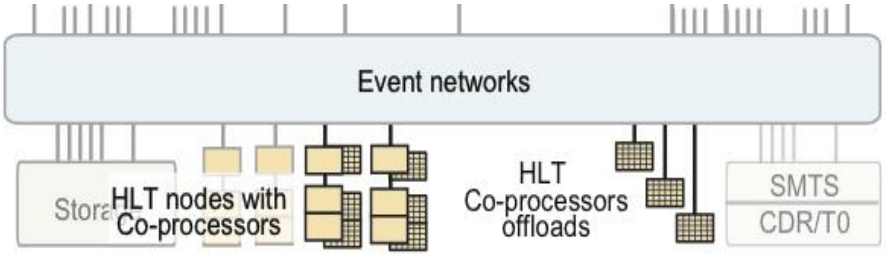
Performance per $: **optimistically** 10-15% progression per year

TDP would also be a problem

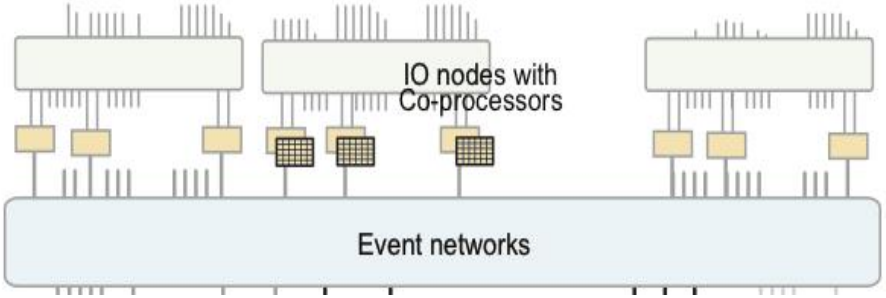


The answer: Heterogeneous Computing

Offload to optimized co-processors



- GPGPU for memory-local algorithms
  - Pattern recognition
  - Space partitioning containers
  - Image processing
- FPGA for transforms, feature extraction and inference/classification
- …

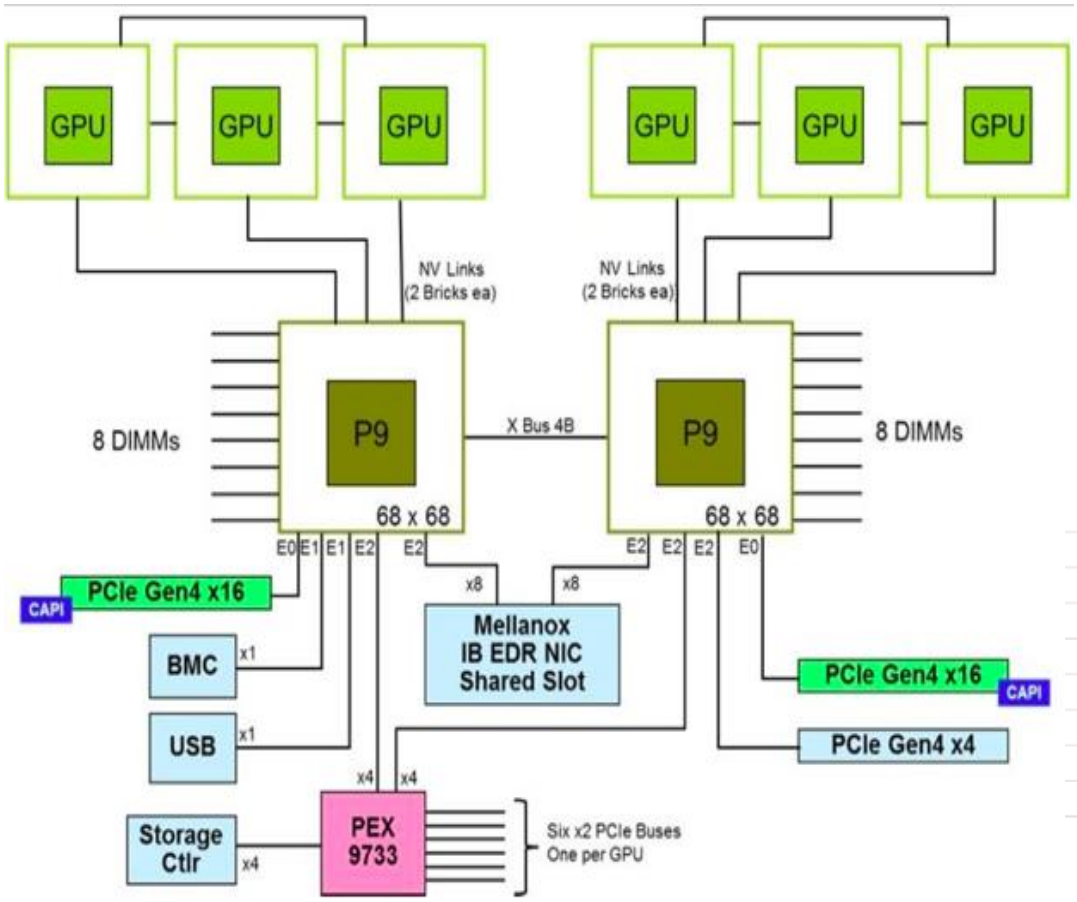


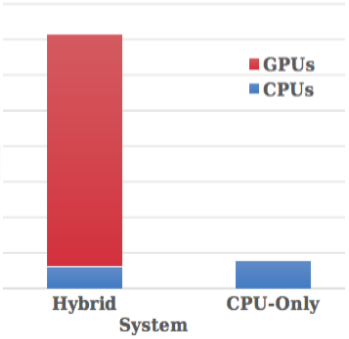
- Traditional approach to instrument processor with offload engines
  - Limits choice of hardware
  - Forces early choice and backward compatibility
  - Creates “live-locks”
- **Preemptive reconstruction** in dedicated co-processor “farms”
  - Choose the best hardware for the task
  - Add the right type of resources when needed
  - Easier upgrade

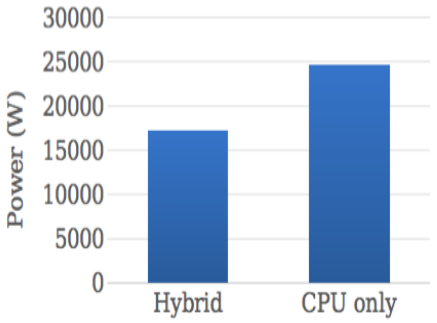
Processor

- High throughput market driven by AI applications
  - PCIe gen4: 15.7 Gbps per lane
  - Lots of GPU, dedicated links, coherence, remote DMA
  - Integration of HPC interconnects (InfiniBand, Omnipath…)
  - Large clusters with >100 Gb/s interconnects
- Classification/inference engines run best on FPGA
  - Need cache coherence
- Large memory throughput and volumes
  - New memory architectures
  - NVRAM with reasonable latency, with sizes in the TB range

AI market

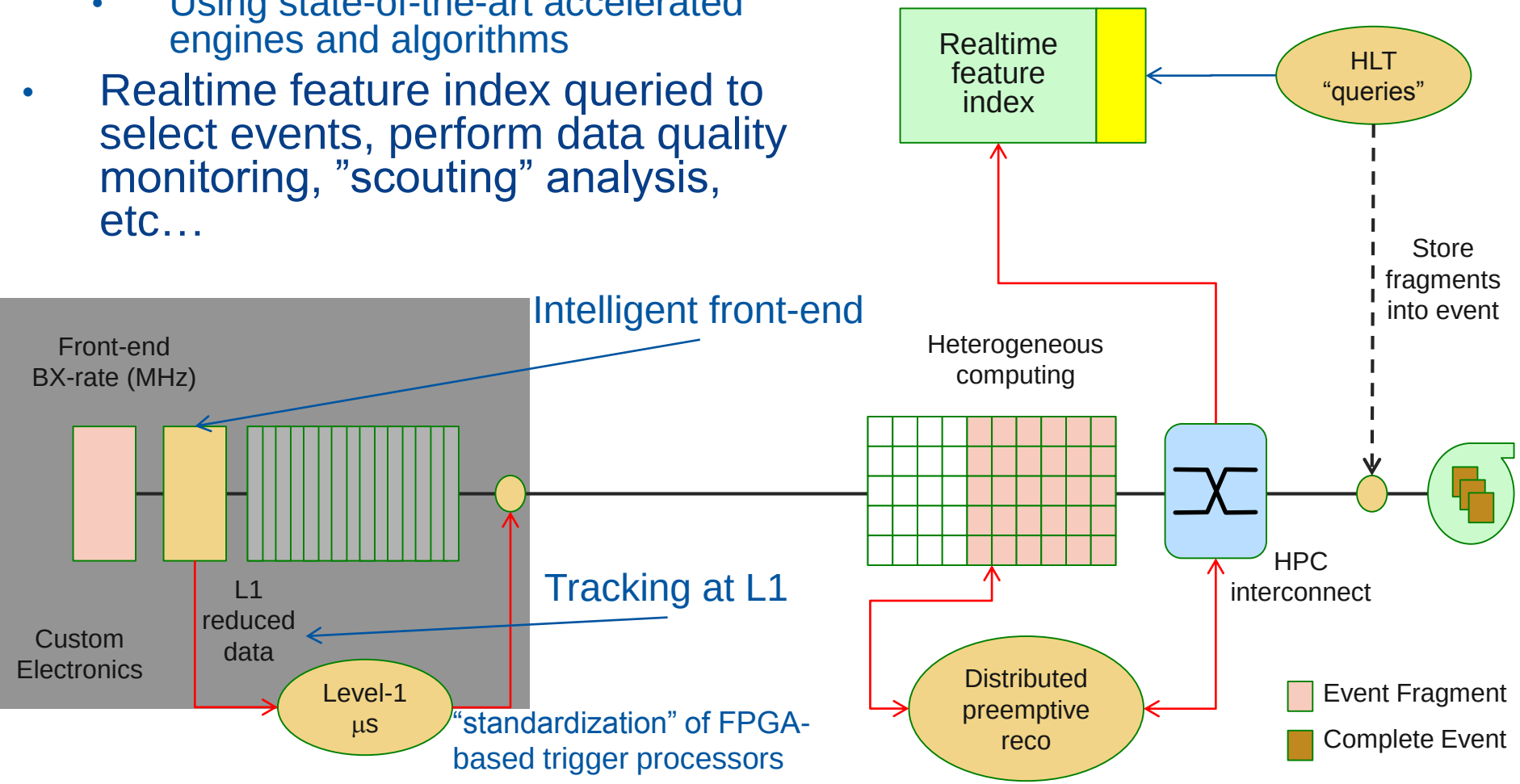

GPU
GPU
GPU
GPU
GPU
GPU
NV Links
(2 Bricks ea)
NV Links
(2 Bricks ea)
8 DIMMs
P9
X Bus 4B
P9
8 DIMMs
68 x 68
68 x 68
E0 E1 E1 E2
E2
E2 E2 E2 E0
PCIe Gen4 x16
CAPI
x8
x8
Mellanox
IB EDR NIC
Shared Slot
BMC
x1
USB
x1
PCIe Gen4 x16
CAPI
PCIe Gen4 x4
x4
x4
Storage
Ctlr
x4
PEX
9733
Six x2 PCIe Buses
One per GPU


GPUs
CPUs
Hybrid
CPU-Only
System

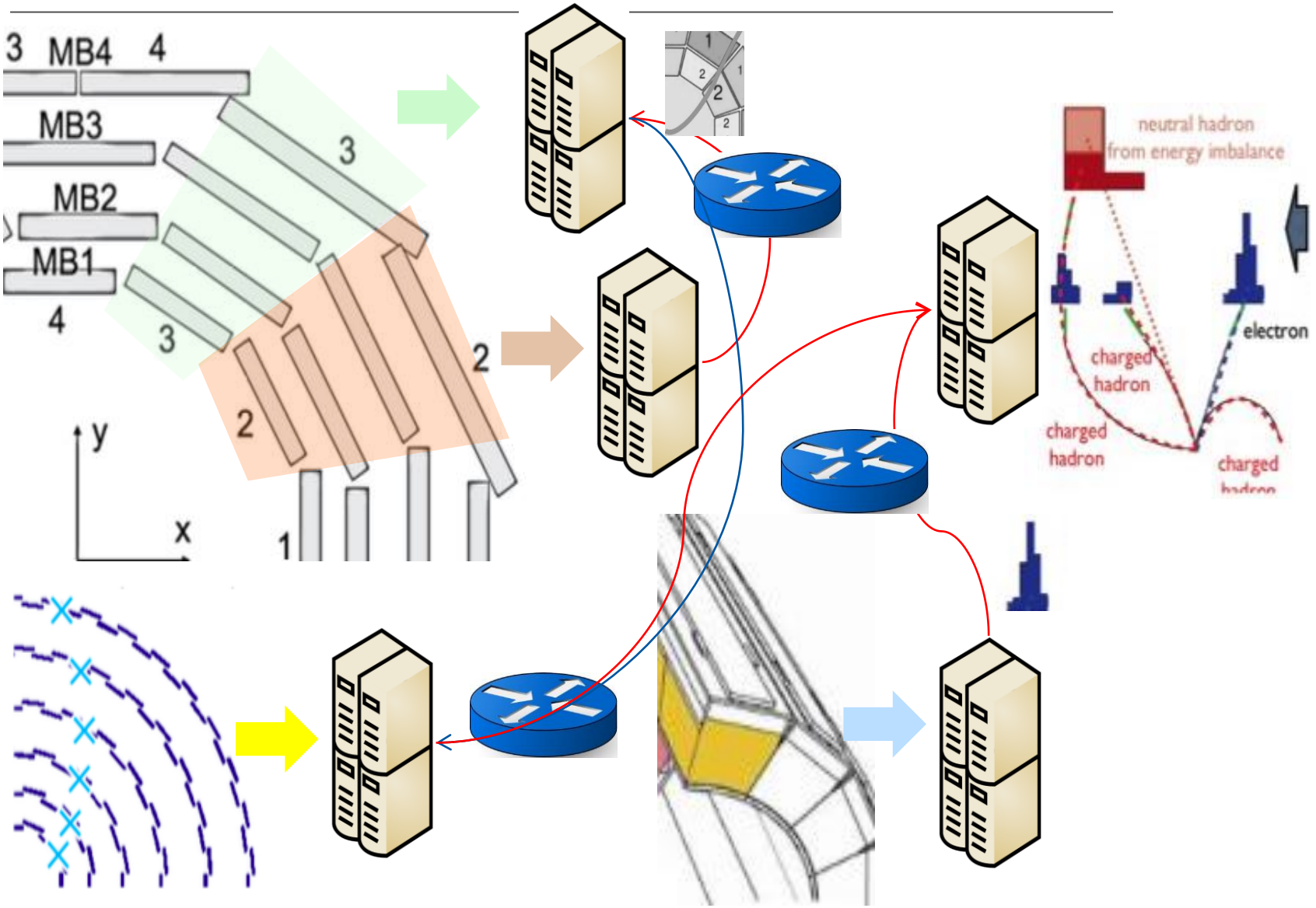

30000
25000
20000
15000
10000
5000
0
Power (W)
Hybrid
CPU only

A possible way forward

- Distributed preemptive reconstruction via HPC interconnect
  - Using state-of-the-art accelerated engines and algorithms
- Realtime feature index queried to select events, perform data quality monitoring, ”scouting” analysis, etc…

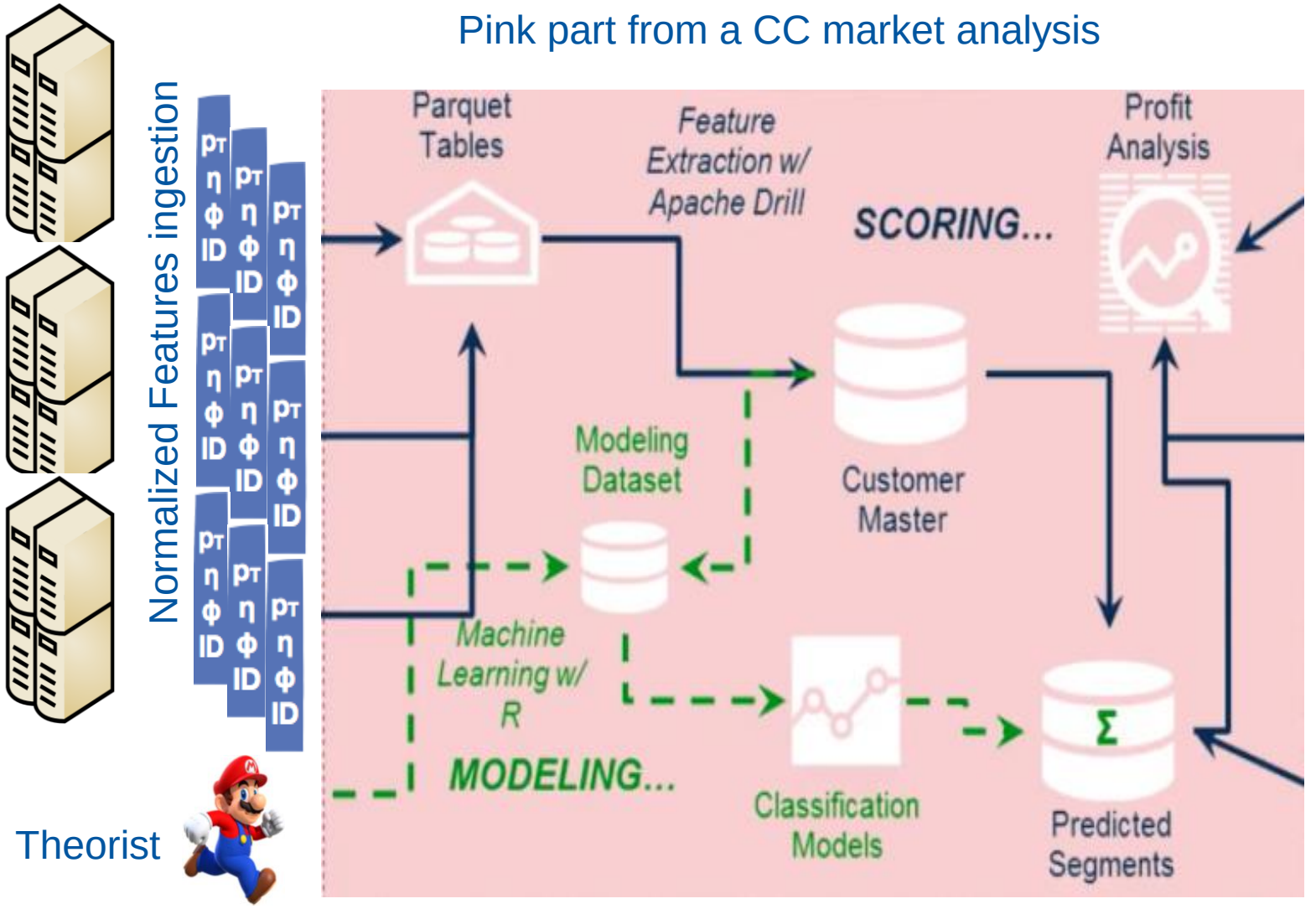


The idea: bring the algorithm to the data



Realtime indexing, query, classification

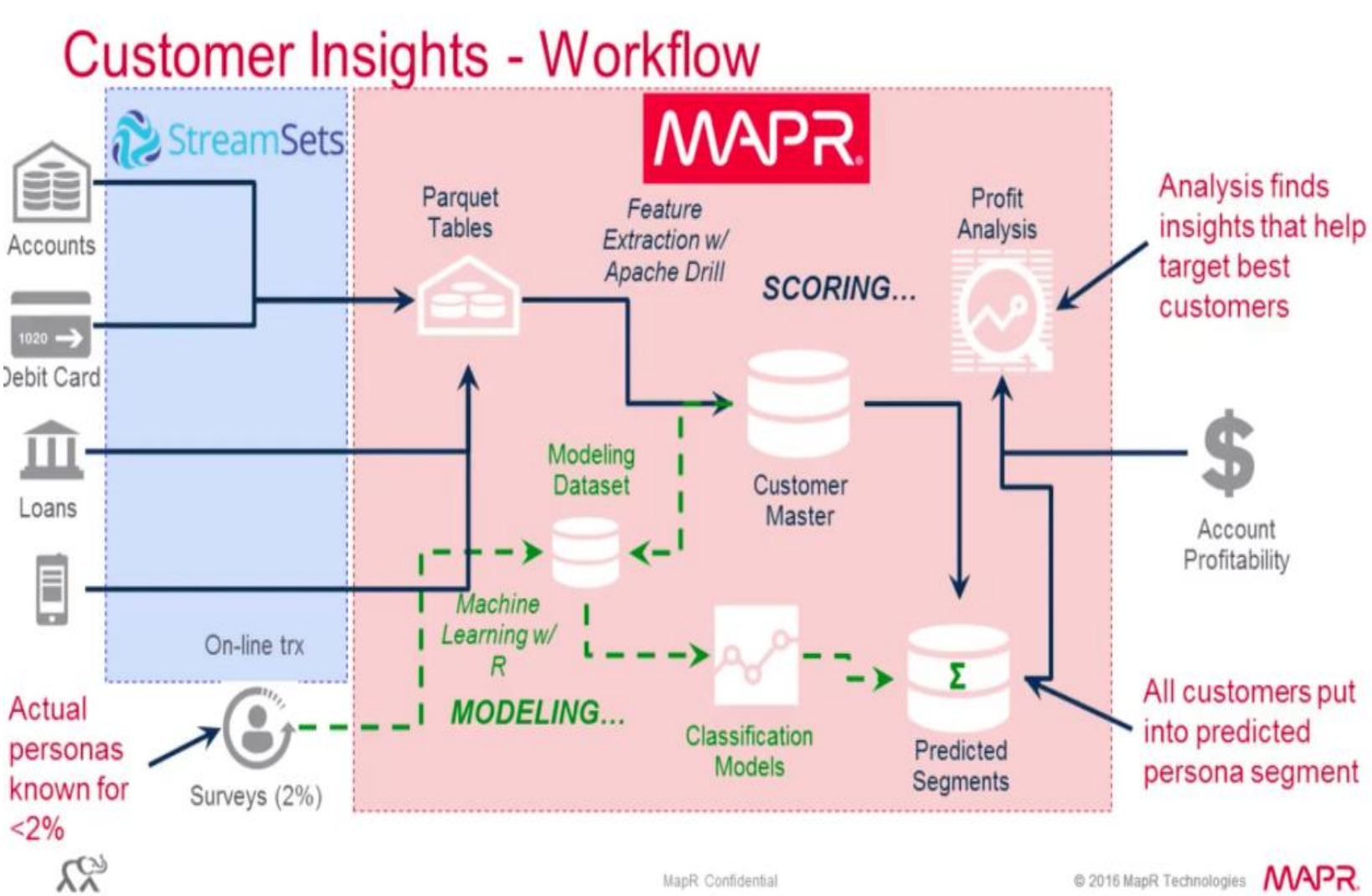
Pink part from a CC market analysis



Analysis finds Insights in data

Anomalies may hint at new physics (and tell us where to look)

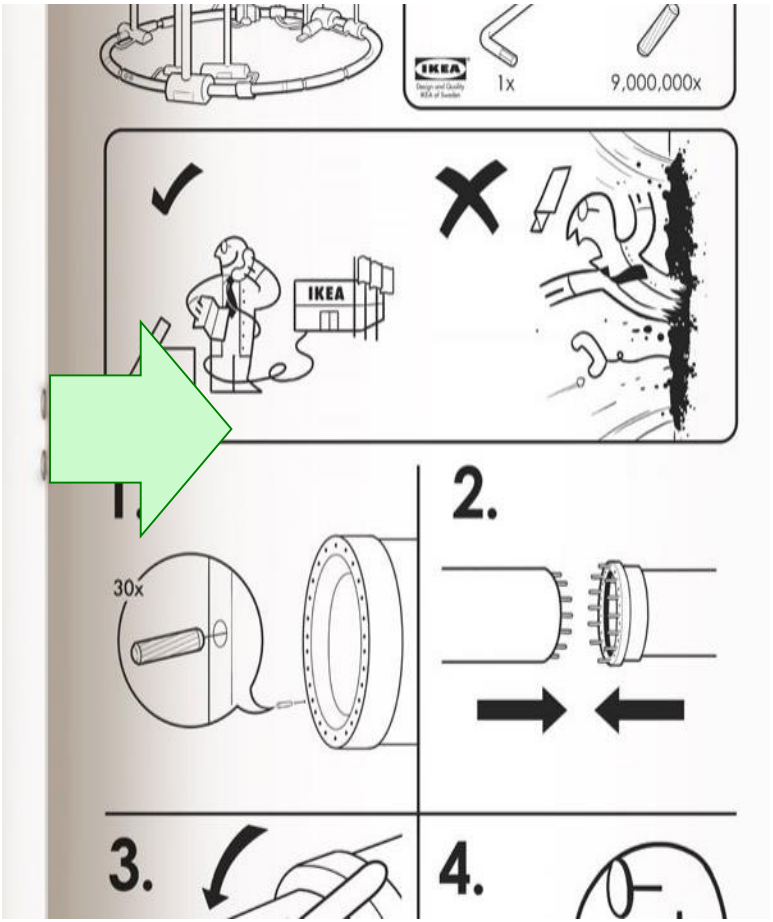
All events Classified Into model(s)

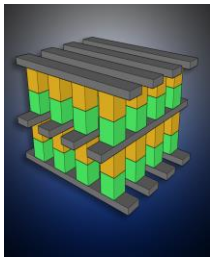

Customer Insights - Workflow
StreamSets
MAPR
Accounts
Debit Card
Loans
On-line trx
Parquet Tables
Feature Extraction w/ Apache Drill
SCORING...
Profit Analysis
Analysis finds insights that help target best customers
Modeling Dataset
Customer Master
Account Profitability
Machine Learning w/ R
MODELING...
Classification Models
Predicted Segments
All customers put into predicted persona segment
Actual personas known for <2%
Surveys (2%)
MapR Confidential
© 2016 MapR Technologies
MAPR
27

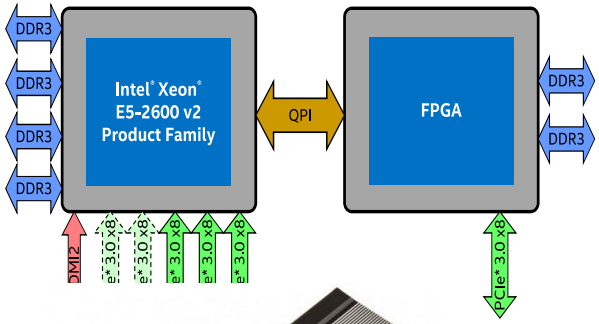
Putting it all together

Assemble forefront technologies

(developed by others)
(for commercial applications)







Preemptive online processing
Fast feature search
Real-time indexing
Query-based selection

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

Outlook

- HEP progress towards a trigger-less system is slow
  - It may not be necessary or even desirable to go all the way
  - The appearance of FPGAs with large I/O throughput, and the use of time-multiplexing, allow a “standardization” of the approach
  - And make possible the application of inference engines and other prevalent techniques
- At the subsequent step, i.e. HLT, a paradigm shift is necessary to accommodate and make the most of those technology advances in HTC
  - It is just as much about how we gather, manage, and store data as it is about the algorithms and hardware we use to process them
  - While L1 hardware and approach can benefit HLT as well, by boosting parallelism in a distributed heterogeneous computing environment…
- Need to deal with resource orchestration, and storage/access of the resulting feature collections
  - Real-time indexing and query-based systems a promising possibility