DAQ, Online, and Software Triggers summary

V.V. Gligorov, CERN

On behalf of the Trigger/Online/Offline/Computing preparatory group

ECFA HL-LHC workshop, Aix-les-Bains, 23/10/2014

Talk overview

A summary of the DAQ and software trigger plans for the experiments in HL-LHC (n.b. LHCb/ALICE upgrades coming in Run3)

1) Overview of DAQ architectures

2) Common assumptions and technologies

3) Software reconstruction in the HL-LHC era

4) Software triggers and real-time data analysis

As Wesley already said, a big thank you to all the working group members whose slides/results I have stolen!

What is a "software trigger"?

=> A trigger implemented in "COTS" commodity processors, generally CPUs but possibly with GPU/FPGA or other "coprocessors" to help

=> Generally taken to mean a trigger which can perform something close to a "full event reconstruction" even if it doesn't in practice.

Another way to say this : anything which is not fixed-latency custom electronics. Important to realize though that in the multi-core era the actual underlying hardware may well be far from homogenous.

Architecture overview

MORE
IS
MORE

The basic approach of all four collaborations can be summarized as follows : put as much as DAQ will allow into software triggers

Nevertheless “physics” and hardware constraints are leading to implementation differences

DAQ overview

	ALICE	LHCb	CMS	ATLAS
Hardware trigger	No	No	Yes	Yes
Software trigger input rate	50 kHz Pb-Pb 200 kHz p-Pb	30 MHz	500/750 kHz for PU 140/200	0.4 MHz
Baseline processing architecture	CPU/GPU/FPGA/ Cloud&Grid	CPU farm (+coprocessors)	CPU farm (+coprocessors)	CPU farm (+coprocessors)
Software trigger output rate	50 kHz Pb-Pb 200 kHz p-Pb	20-100 kHz	5-7.5 kHz	5-10 kHz

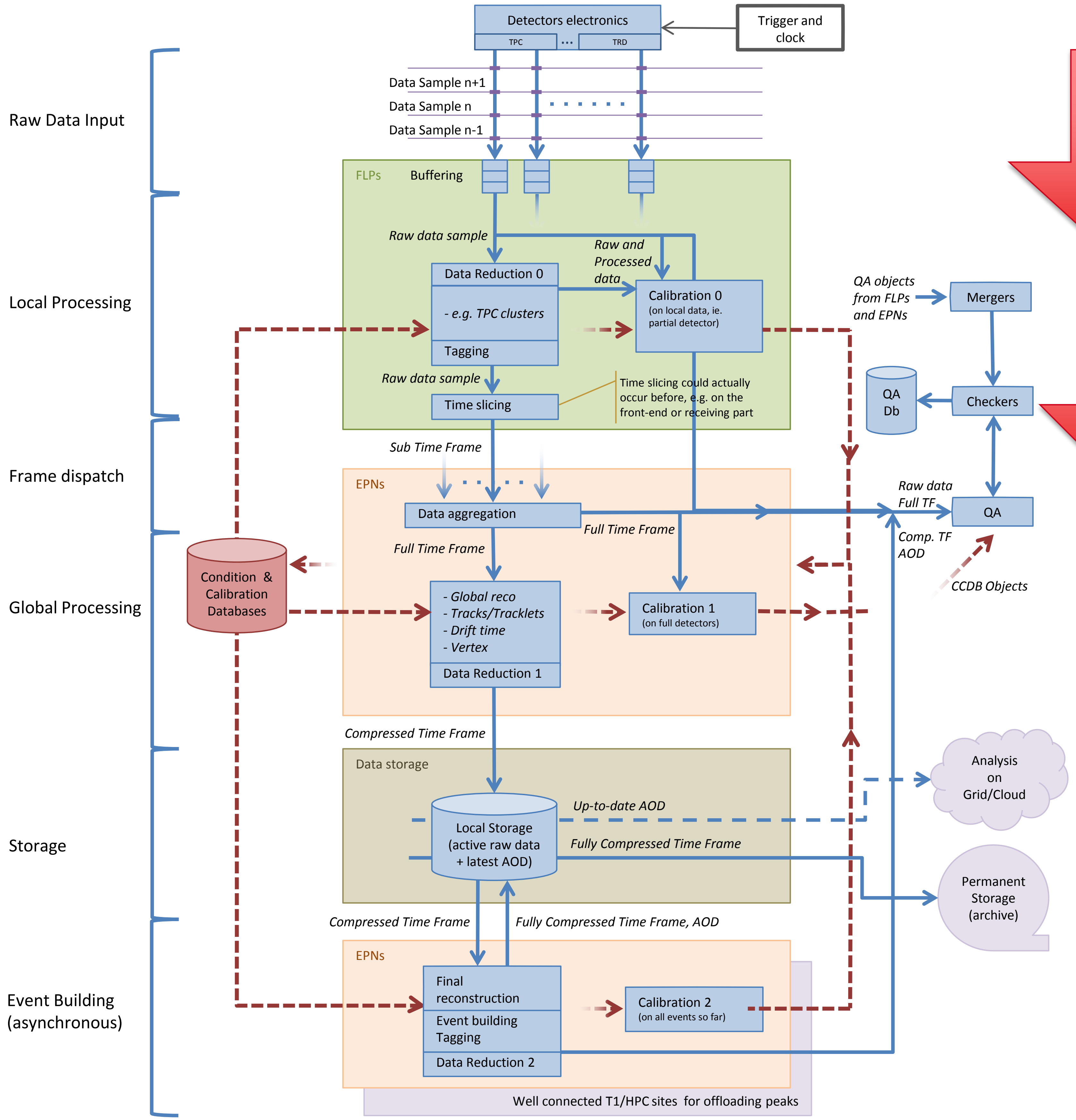
DAQ overview

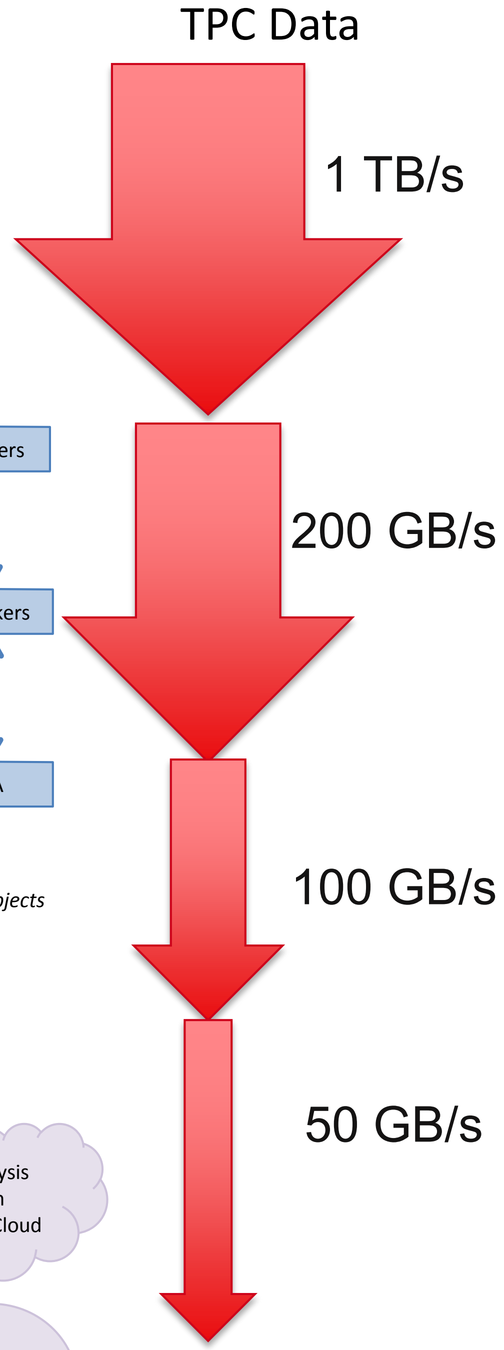
	ALICE	LHCb	CMS	ATLAS
Hardware trigger	No	No	Yes	Yes
Software trigger input rate	50 kHz Pb-Pb 200 kHz p-Pb	30 MHz	500/750 kHz for PU 140/200	0.4 MHz
Baseline processing architecture	CPU/GPU/FPGA/ Cloud&Grid	CPU farm (+coprocessors)	CPU farm (+coprocessors)	CPU farm (+coprocessors)
Software trigger output rate	50 kHz Pb-Pb 200 kHz p-Pb	20-100 kHz	5-7.5 kHz	5-10 kHz

ALICE DAQ

ALICE's online and offline data processing integrated into a single workflow

Aim is to compress events, not throw them away : driven by the fact that traditional "physics" probes have low S/B, hence event filtering not an efficient approach.





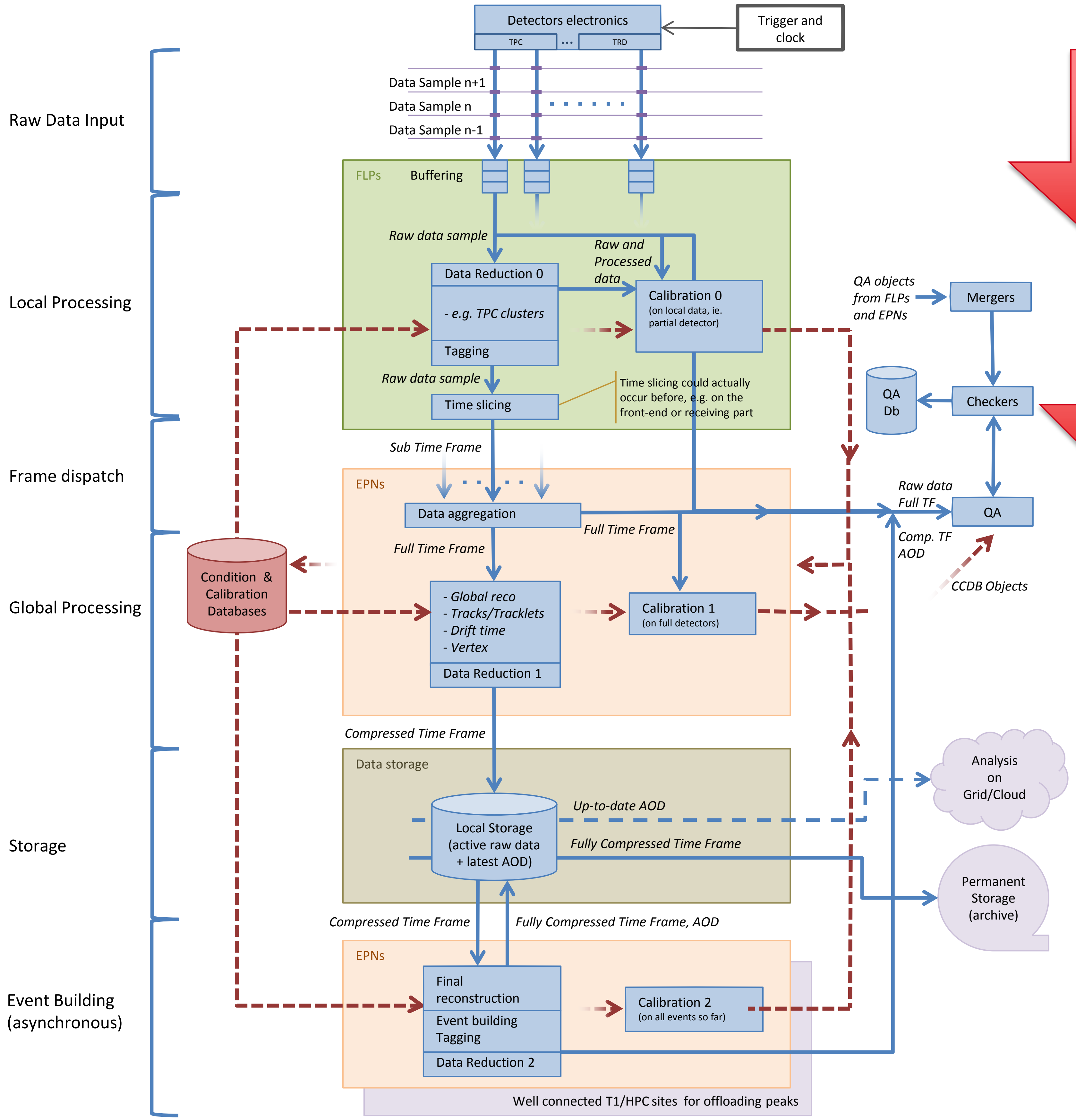
ALICE DAQ

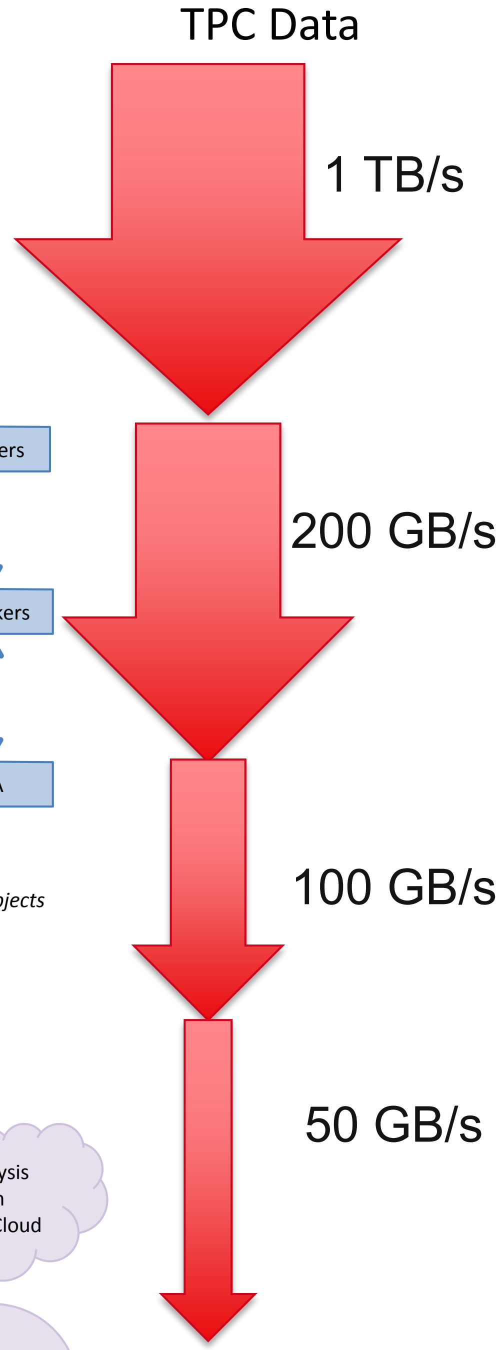
Detector	Input to Online System (GByte/s)	Peak Output to Local Data Storage (GByte/s)	Avg. Output to Computing Center (GByte/s)
TPC	1000	50.0	8.0
TRD	81.5	10.0	1.6
ITS	40	10.0	1.6
Others	25	12.5	2.0
Total	**1146.5**	**82.5**	**13.2**

Input rate 1TByte/s

Goal is to achieve around 100x compression

Later compression stages perform detector calibrations which are fed back into earlier stages. The compression explicitly preserves the ability to recalibrate offline.

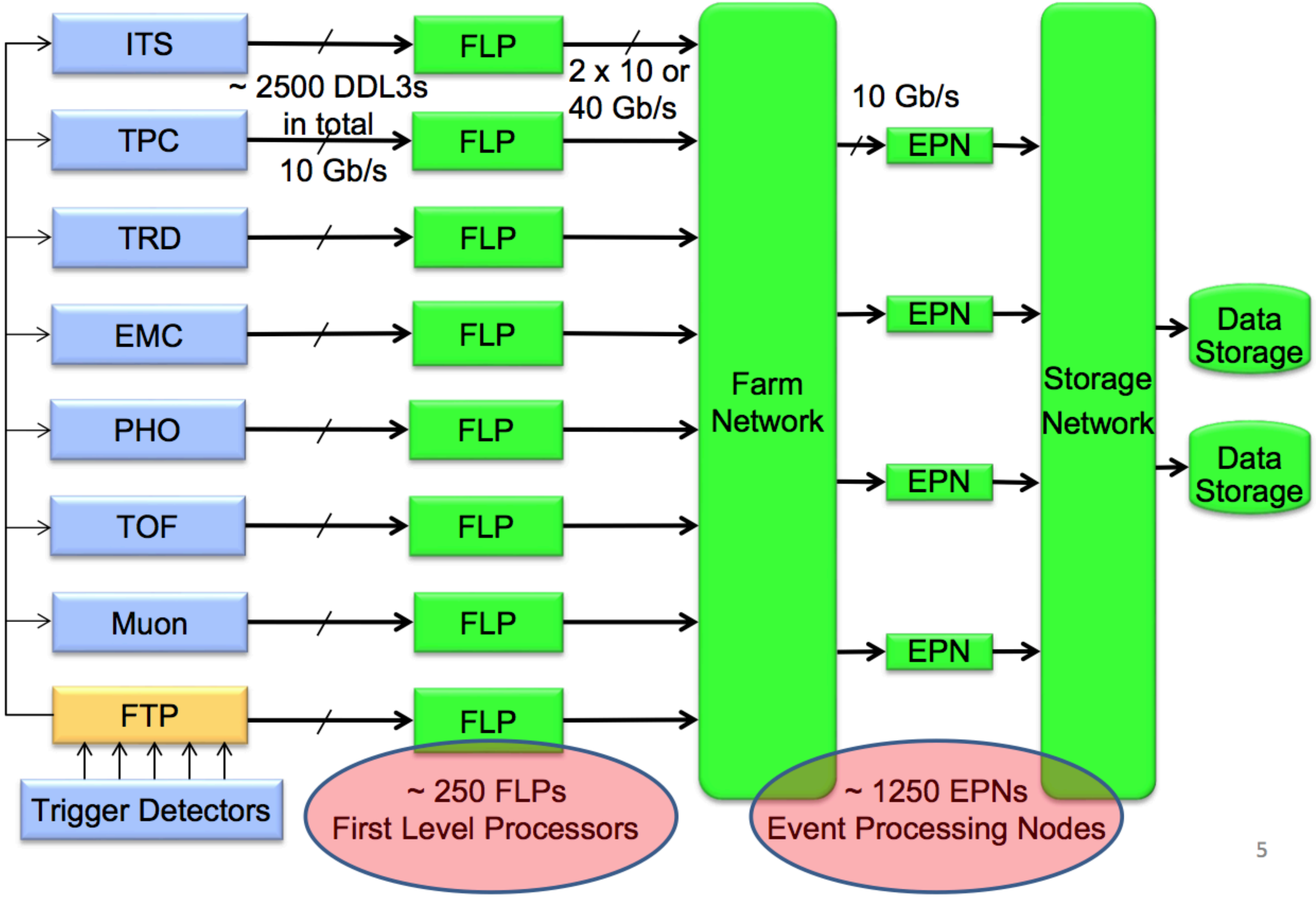




ALICE DAQ

Detector	Input to Online System (GByte/s)	Peak Output to Local Data Storage (GByte/s)	Avg. Output to Computing Center (GByte/s)
TPC	1000	50.0	8.0
TRD	81.5	10.0	1.6
ITS	40	10.0	1.6
Others	25	12.5	2.0
Total	**1146.5**	**82.5**	**13.2**

The data compression begins separately within each subdetector (the First Level Processors) and then continues once the whole event is built within the Event Processing Node farm.



5

ALICE performs event compression, not selection, in their software "trigger"

DAQ overview

	ALICE	LHCb	CMS	ATLAS
Hardware trigger	No	No	Yes	Yes
Software trigger input rate	50 kHz Pb-Pb 200 kHz p-Pb	30 MHz	500/750 kHz for PU 140/200	0.4 MHz
Baseline processing architecture	CPU/GPU/FPGA/ Cloud&Grid	CPU farm (+coprocessors)	CPU farm (+coprocessors)	CPU farm (+coprocessors)
Software trigger output rate	50 kHz Pb-Pb 200 kHz p-Pb	20-100 kHz	5-7.5 kHz	5-10 kHz

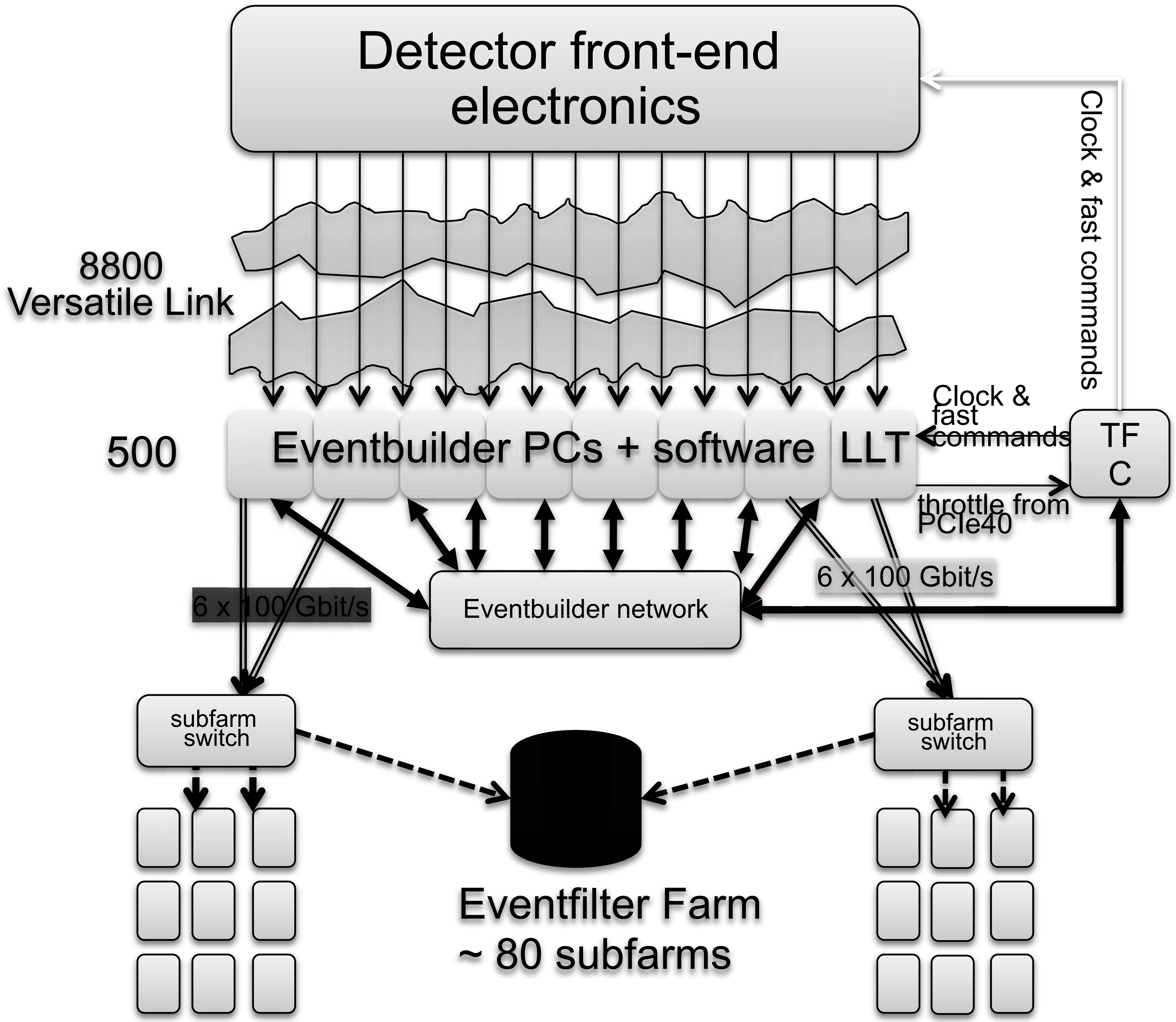
LHCb DAQ

LHCb's DAQ network built around a bidirectional eventbuilding farm.

Note that about 80% of the CPU in the event-building PCs remains free for implementing the "low-level trigger" (selecting on muon and CALO primitives) and/or the first stages of the event reconstruction.

Low-level trigger to be implemented in software, will NOT act on the front-end. Must read all events out regardless.

Need to transport/build 40 Tbit/s

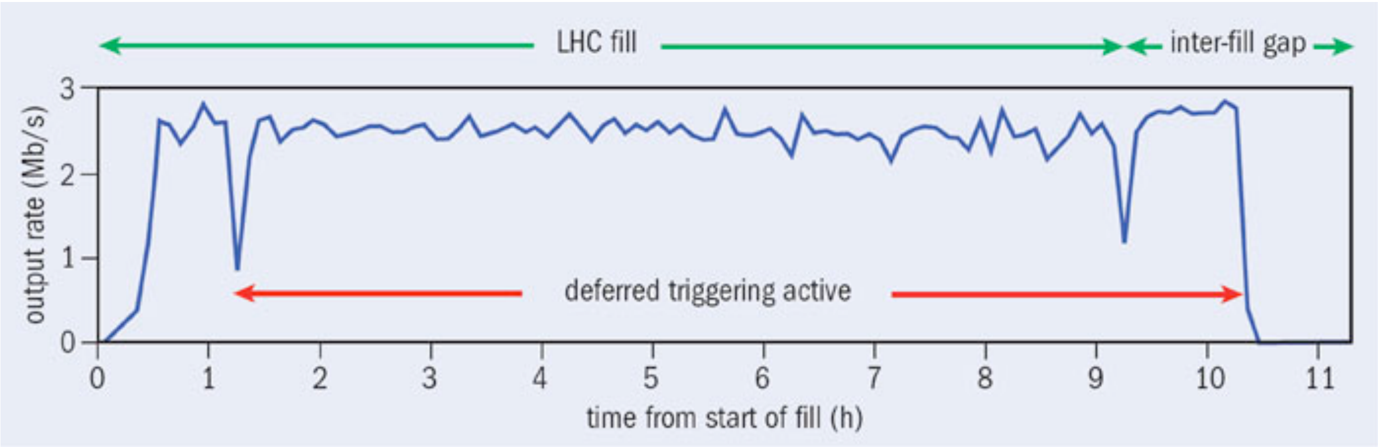


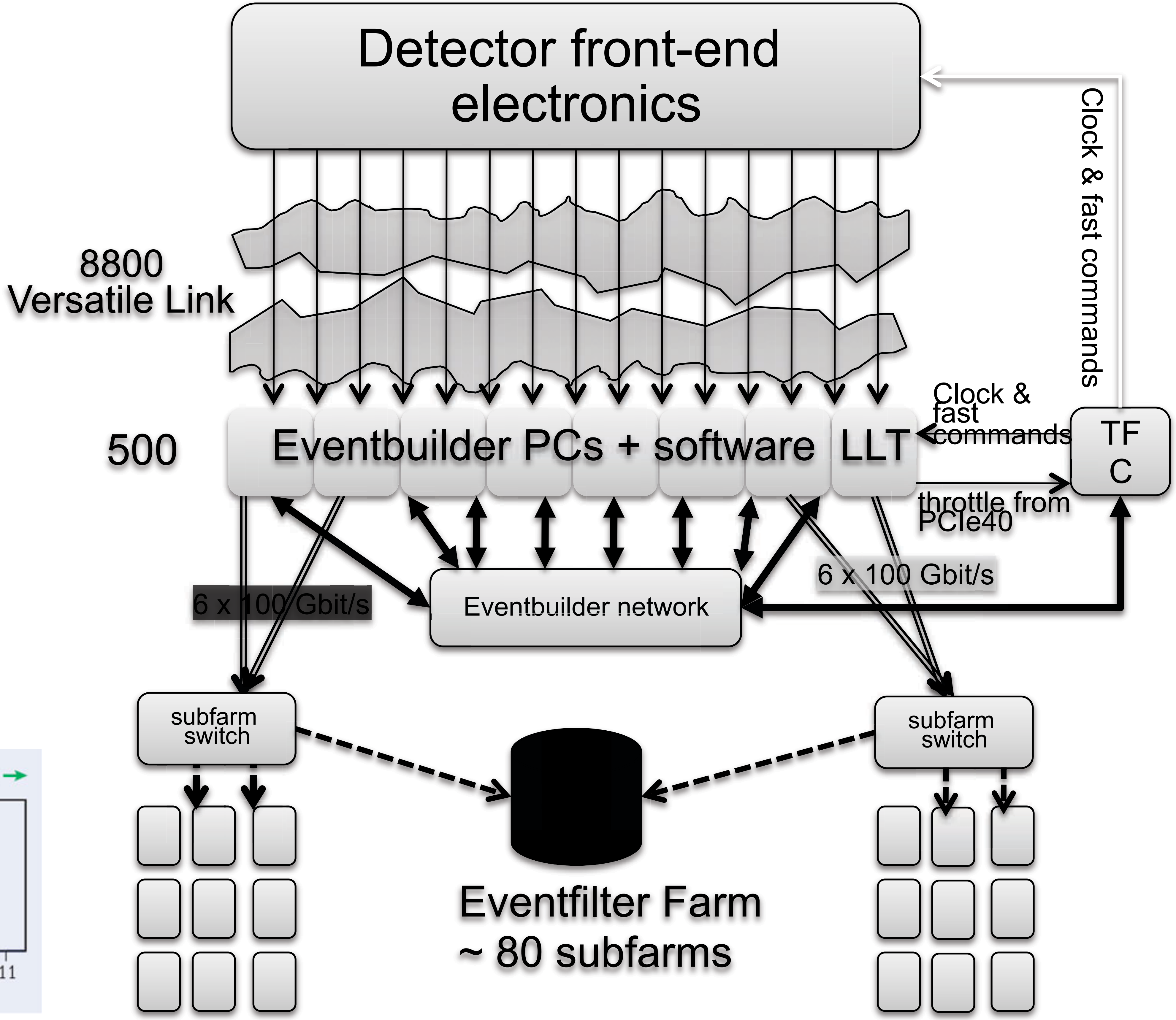
LHCb DAQ

A critical part of the DAQ is the ability to buffer events onto hard disks located in the EFF nodes (“deferred triggering”).

Serves two purposes : multiply the available processing time, and allow real-time detector calibration/alignment.

Deployed in Run1 gaining 20% in HLT processing time, will be used more aggressively in Run2.





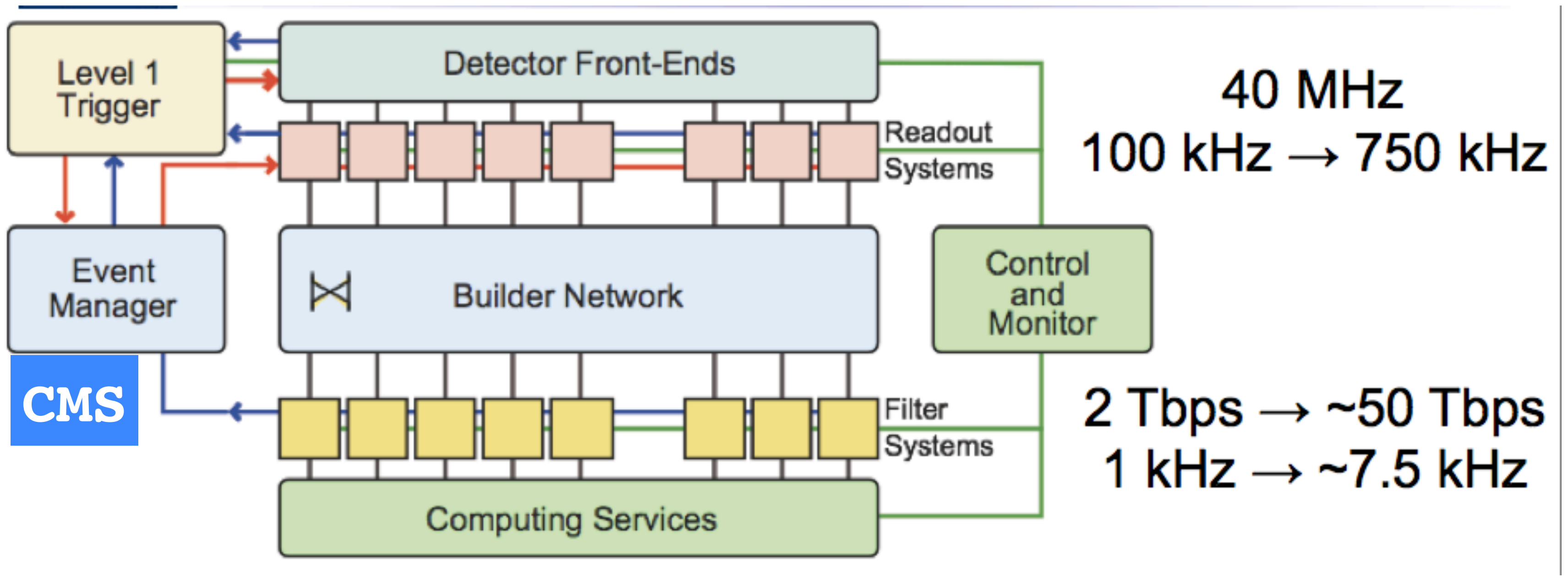
DAQ overview

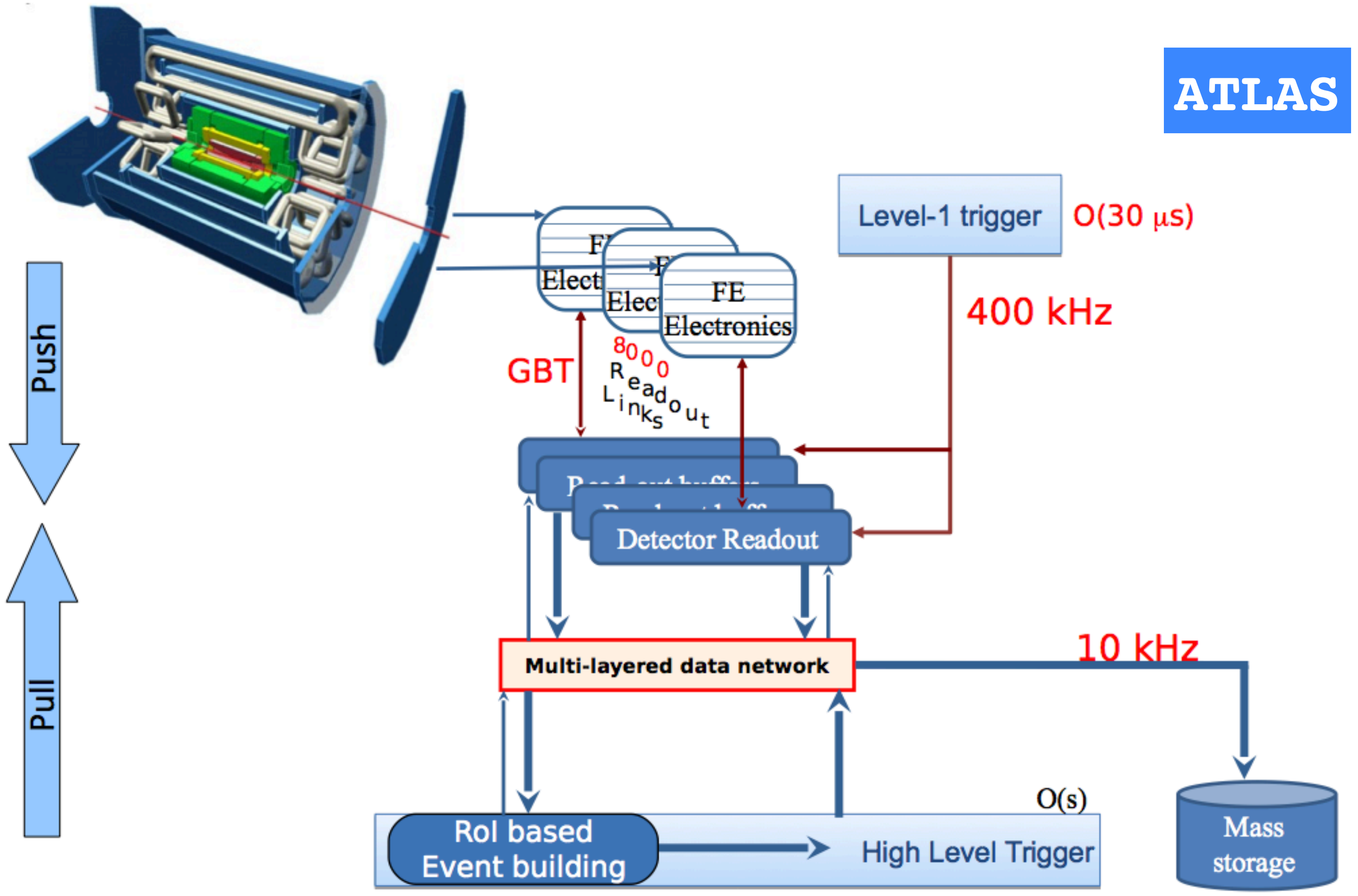
	ALICE	LHCb	CMS	ATLAS
Hardware trigger	No	No	Yes	Yes
Software trigger input rate	50 kHz Pb-Pb 200 kHz p-Pb	30 MHz	500/750 kHz for PU 140/200	0.4 MHz
Baseline processing architecture	CPU/GPU/FPGA/ Cloud&Grid	CPU farm (+coprocessors)	CPU farm (+coprocessors)	CPU farm (+coprocessors)
Software trigger output rate	50 kHz Pb-Pb 200 kHz p-Pb	20-100 kHz	5-7.5 kHz	5-10 kHz

CMS/ATLAS DAQ

Hardware trigger aside, the CMS architecture is not far from what LHCb is planning. Important to note that the L1 tracking trigger will provide seeds for the HLT reconstruction however, which should significantly reduce the computing burden.

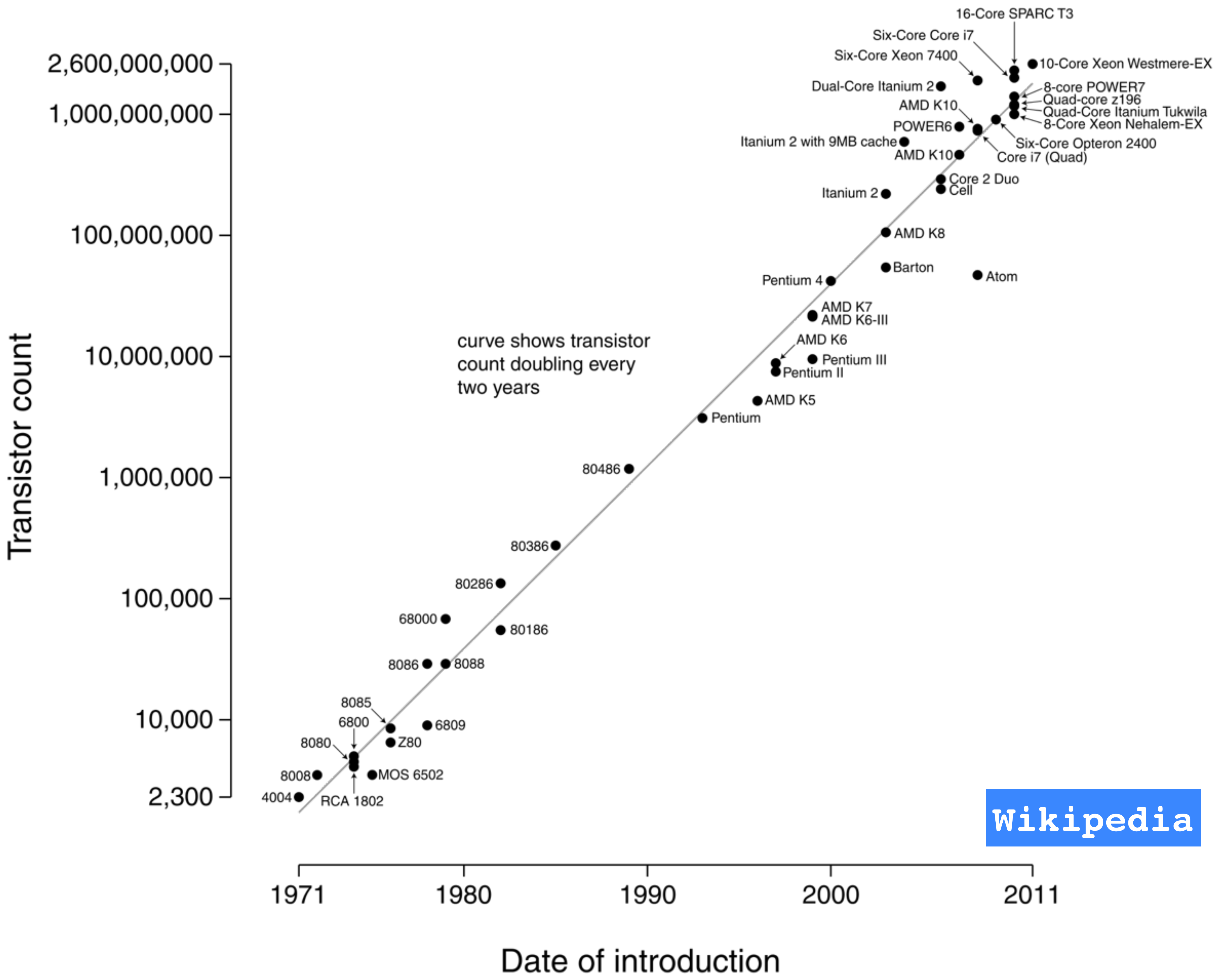
ATLAS plans for a slightly smaller HLT input rate due to two-stage hardware trigger design.





Common assumptions and technologies

Microprocessor Transistor Counts 1971-2011 & Moore's Law



Actually a bit more complicated

Architectural change		Fabrication process	Microarchitecture	Codenames	Release date	Processors	
						8P/4P Server	4P/2P Server/WS
Tick	Die shrink	65 nm	P6, NetBurst	Presler, Cedar Mill, Yonah	January 5, 2006		
Tock	New microarchitecture		Core	Merom	July 27, 2006	Tigerton	Woodcrest Clovertown
Tick	Die shrink	45 nm		Penryn	November 11, 2007	Dunnington	Harpertown
Tock	New microarchitecture		Nehalem	Nehalem	November 17, 2008	Beckton	Gainestown
Tick	Die shrink	32 nm		Westmere	January 4, 2010	Westmere-EX	Westmere-EP
Tock	New microarchitecture		Sandy Bridge	Sandy Bridge	January 9, 2011	(Skipped)	Sandy Bridge-EP
Tick	Die shrink	22 nm		Ivy Bridge	April 29, 2012	Ivy Bridge-EX	Ivy Bridge-EP
Tock	New microarchitecture		Haswell	Haswell	June 2, 2013		We are here!
Tick	Die shrink	14 nm		Broadwell	2014		

Stolen from Beat Jost

Future microprocessor evolution?

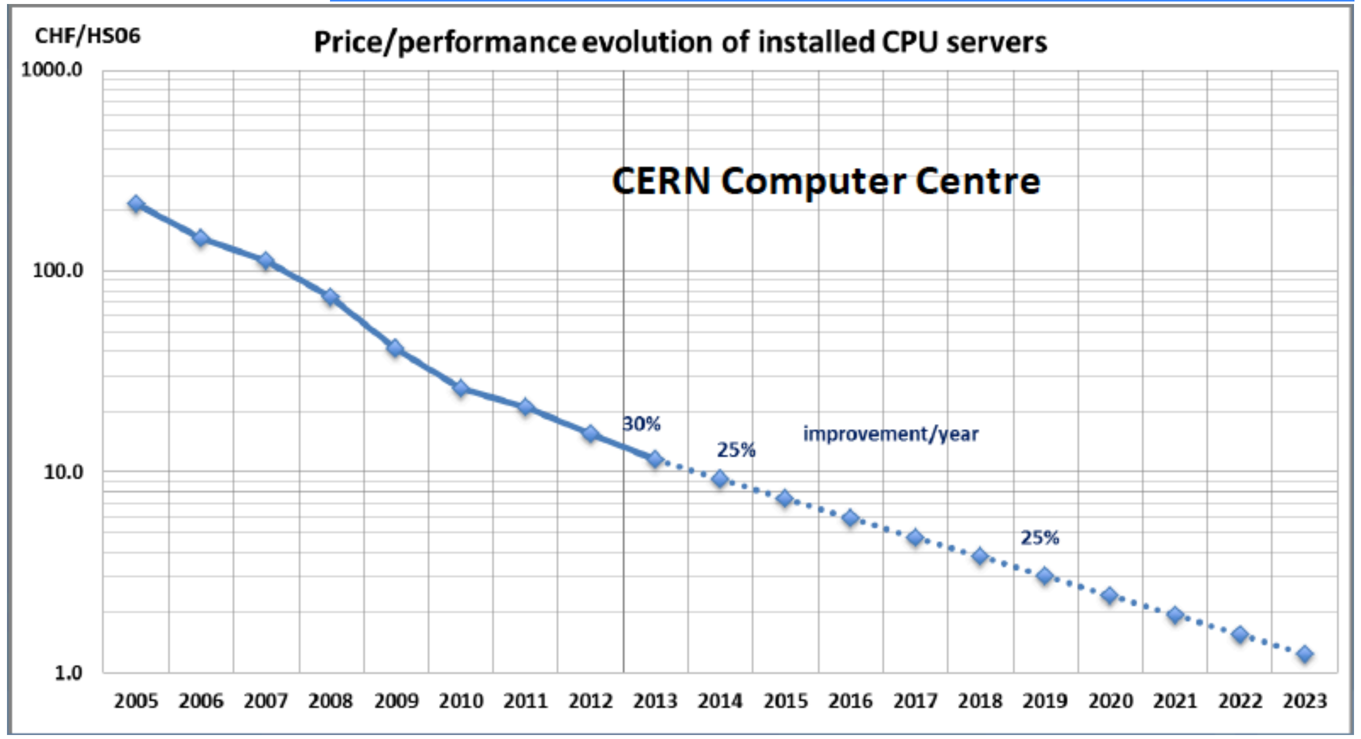
Architectural change		Fabrication process	Microarchitecture	Codenames	Release date	Processors	
						8P/4P Server	4P/2P Server/WS
Tick	Die shrink	14 nm	Haswell	Broadwell	2014		
Tock	New microarchitecture		Skylake	Skylake	2015		
Tick	Die shrink	10 nm		Cannonlake	2016		
Tock	New microarchitecture						
					2017		
Tick	Die shrink	7 nm			2018		
Tock	New microarchitecture				2019		
Tick	Die shrink	5 nm			2020		
Tock	New microarchitecture				2021		

Take home message: expect tick-tock and die shrinking to continue for the next years

Extrapolating to the future

B.Panzer, shown by N. Neufeld, ECFA 2013

Clearly 25% performance improvement per year is not the same as doubling the performance every 2 years (more like 3).

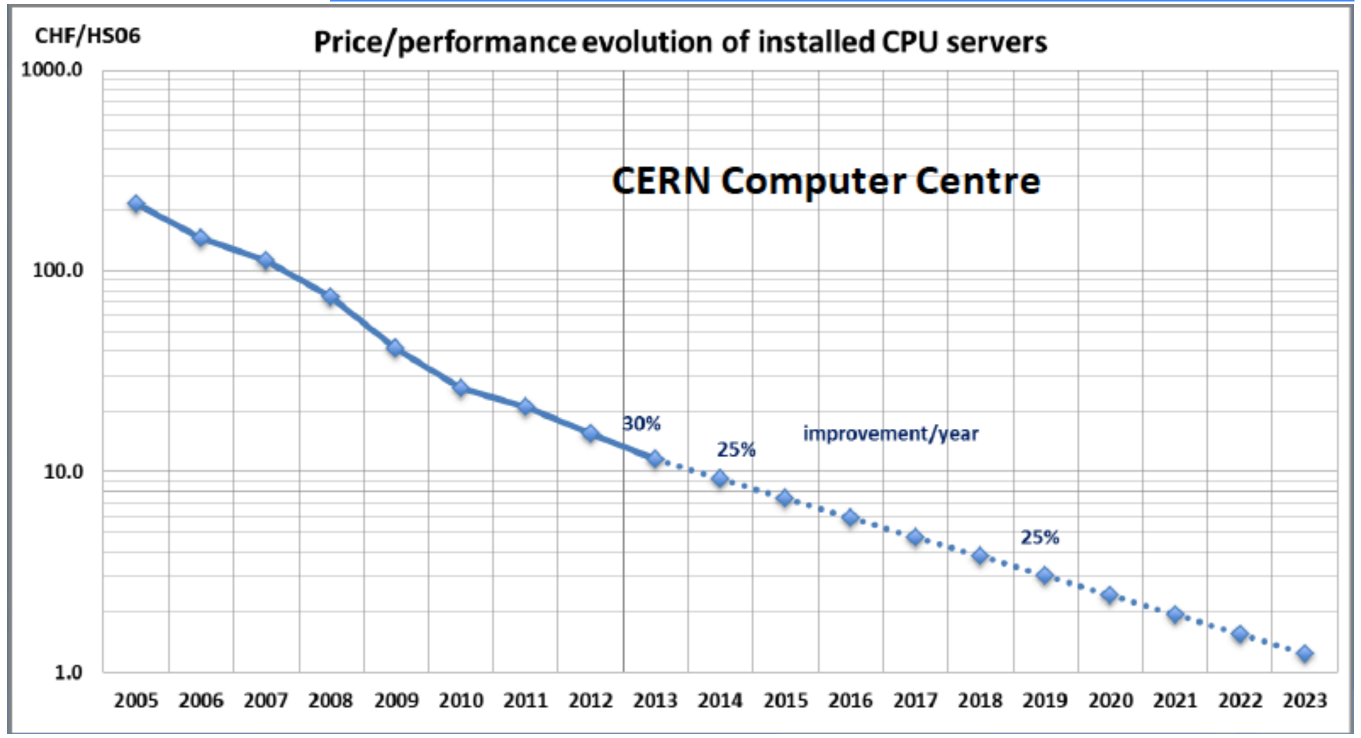


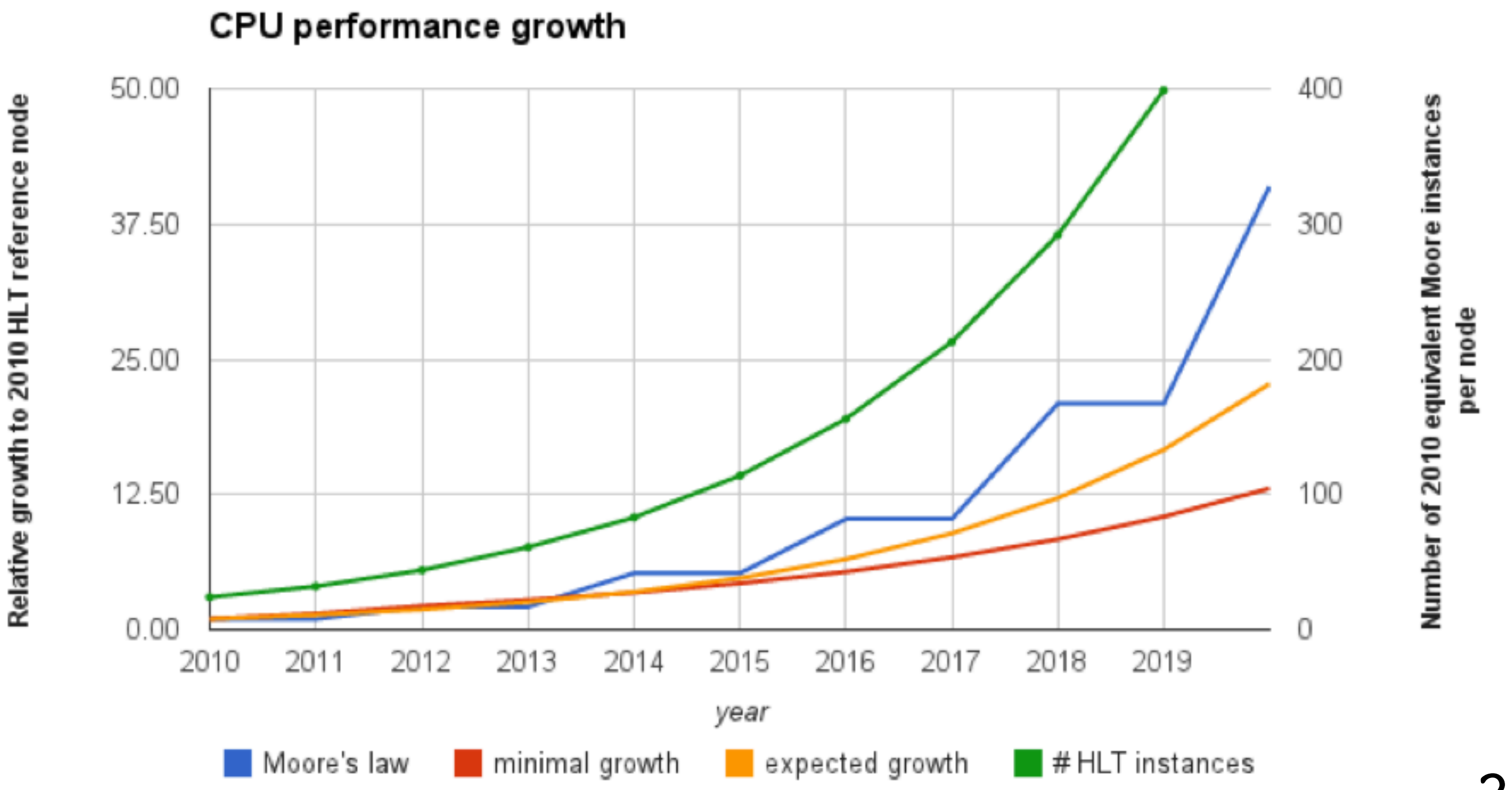
Extrapolating to the future

Clearly 25% performance improvement per year is not the same as doubling the performance every 2 years (more like 3).

However also important to notice that this is a power law, so small changes in the assumed %/year lead to big differences on a 10-20 year timescale.

B.Panzer, shown by N. Neufeld, ECFA 2013





Extrapolating to the future

Clearly 25% performance improvement per year is not the same as doubling the performance every 2 years (more like 3).

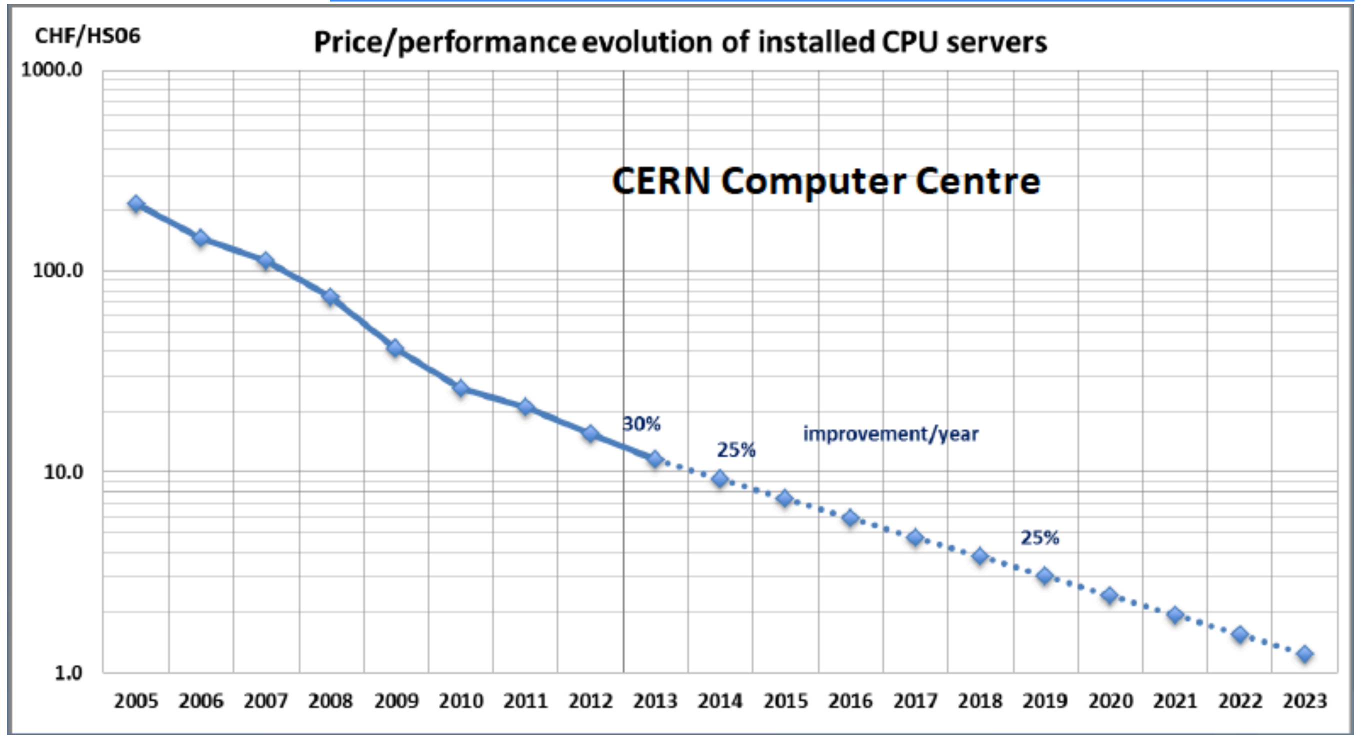
However also important to notice that this is a power law, so small changes in the assumed %/year lead to big differences on a 10-20 year timescale.

CMS and LHCb somewhat more optimistic than CERN computing, backed up by observed performance improvements. But nobody betting the farm on ±5%.

Critical point : must fully exploit the new many core architectures!

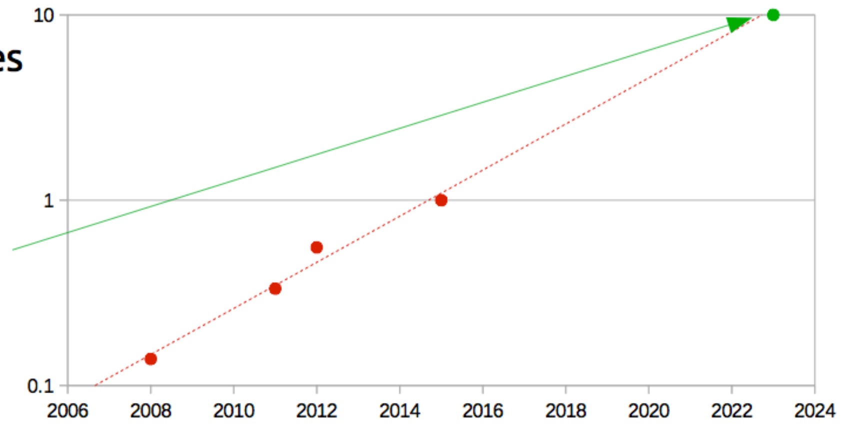
	ALICE	LHCb	ATLAS	CMS
Assumed online performance gains	25%/year	35%/year	25%/year	35%/year

B.Panzer, shown by N. Neufeld, ECFA 2013



CMS observed performance improvements

- look at the power of the HLT nodes
  - bought in 2008, 2011, 2012
  - and foreseen for 2015
- extrapolating to 2023 we could estimate increase by a factor ×10
- this still leaves a factor ×2 (x4)



Software event reconstruction

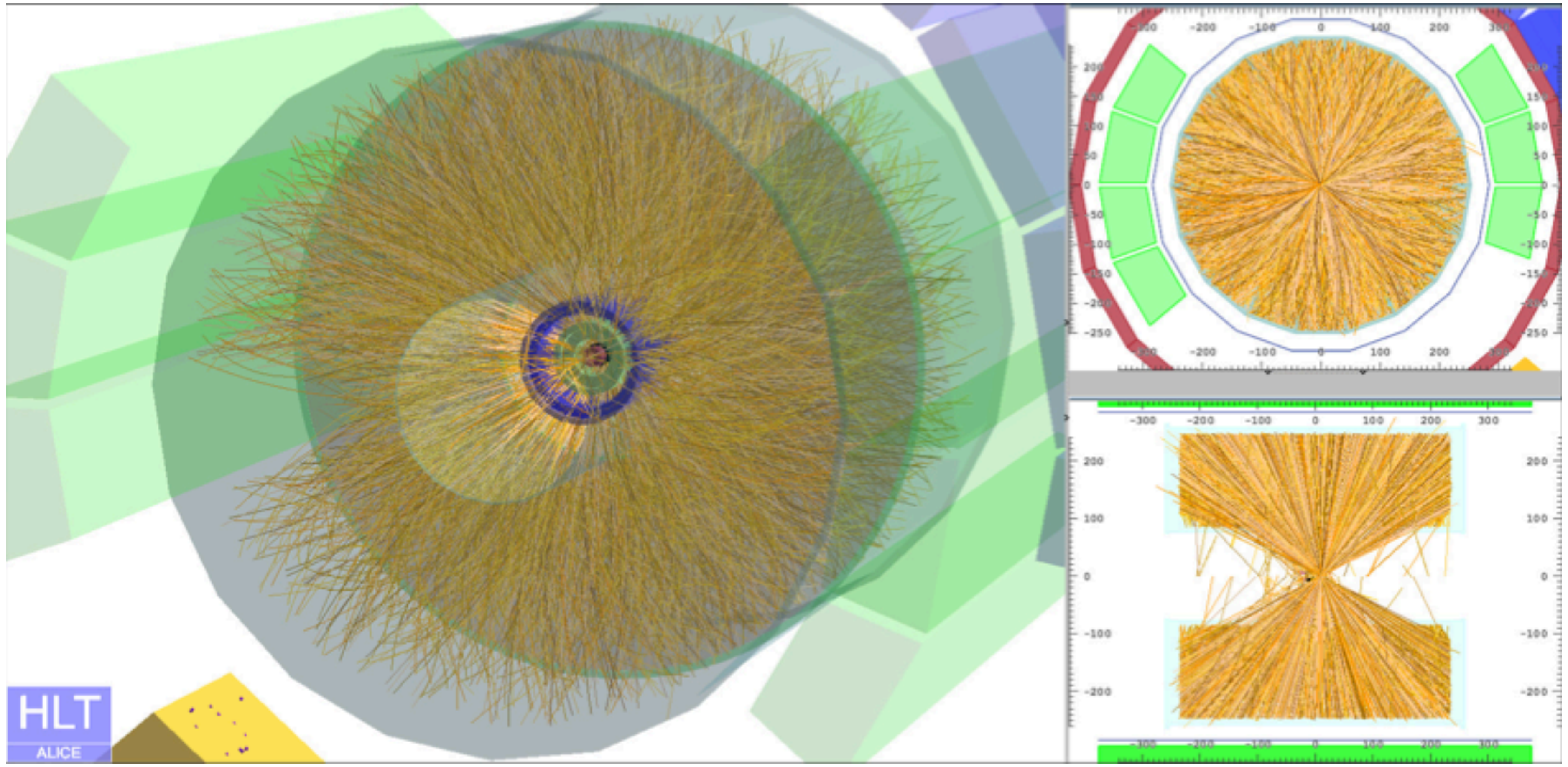
What remains after Moore’s law

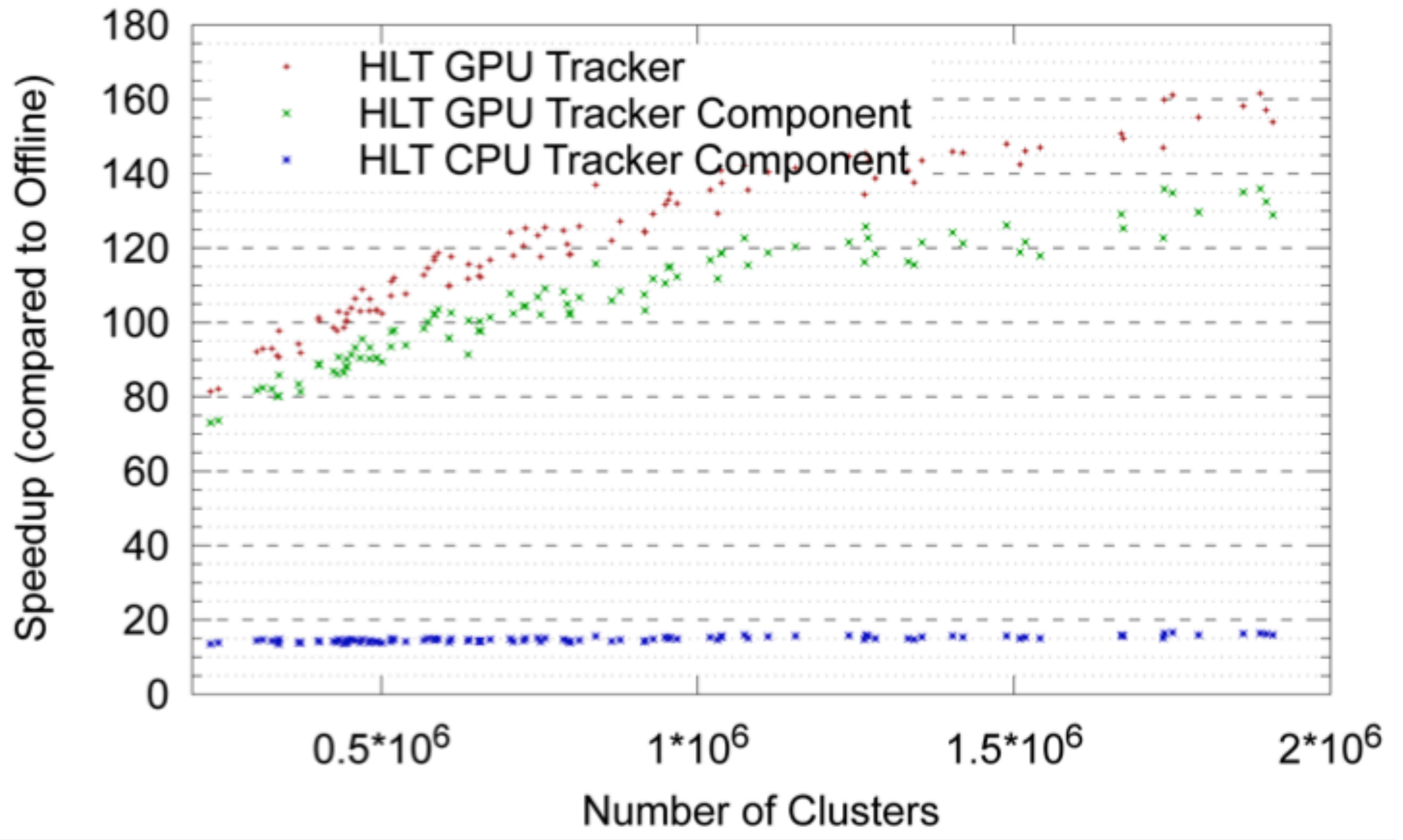
Will need to make significant gains in computing performance on top of Moore's law projections, typically another factor 2-5.

This comes down to exploiting the many-core architectures more intelligently.

A personal comment : we often discuss absolute performance in terms of algorithm speed, but for software triggers latency is basically irrelevant. We should focus on physics/CHF.

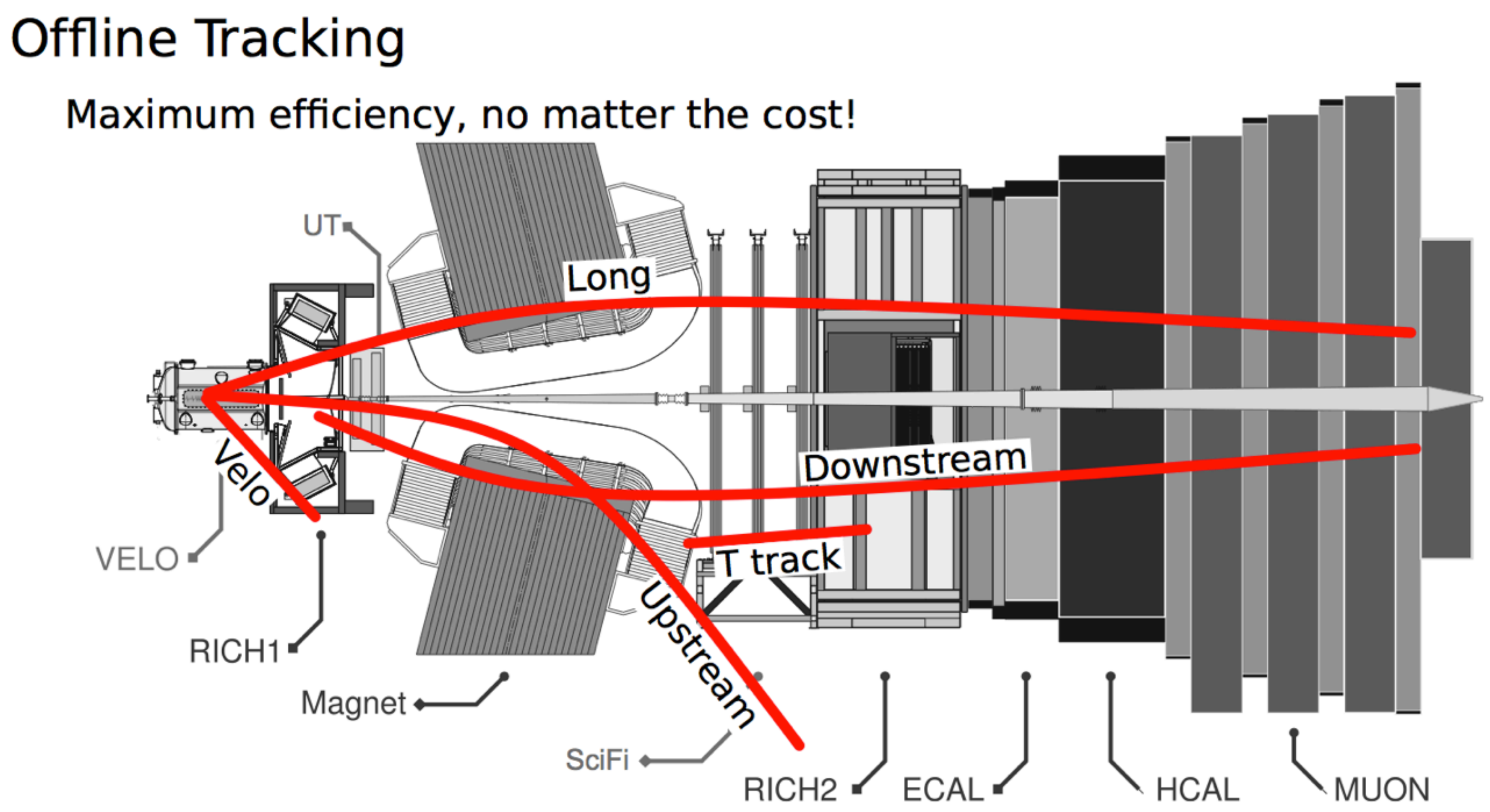
ALICE's GPU tracking



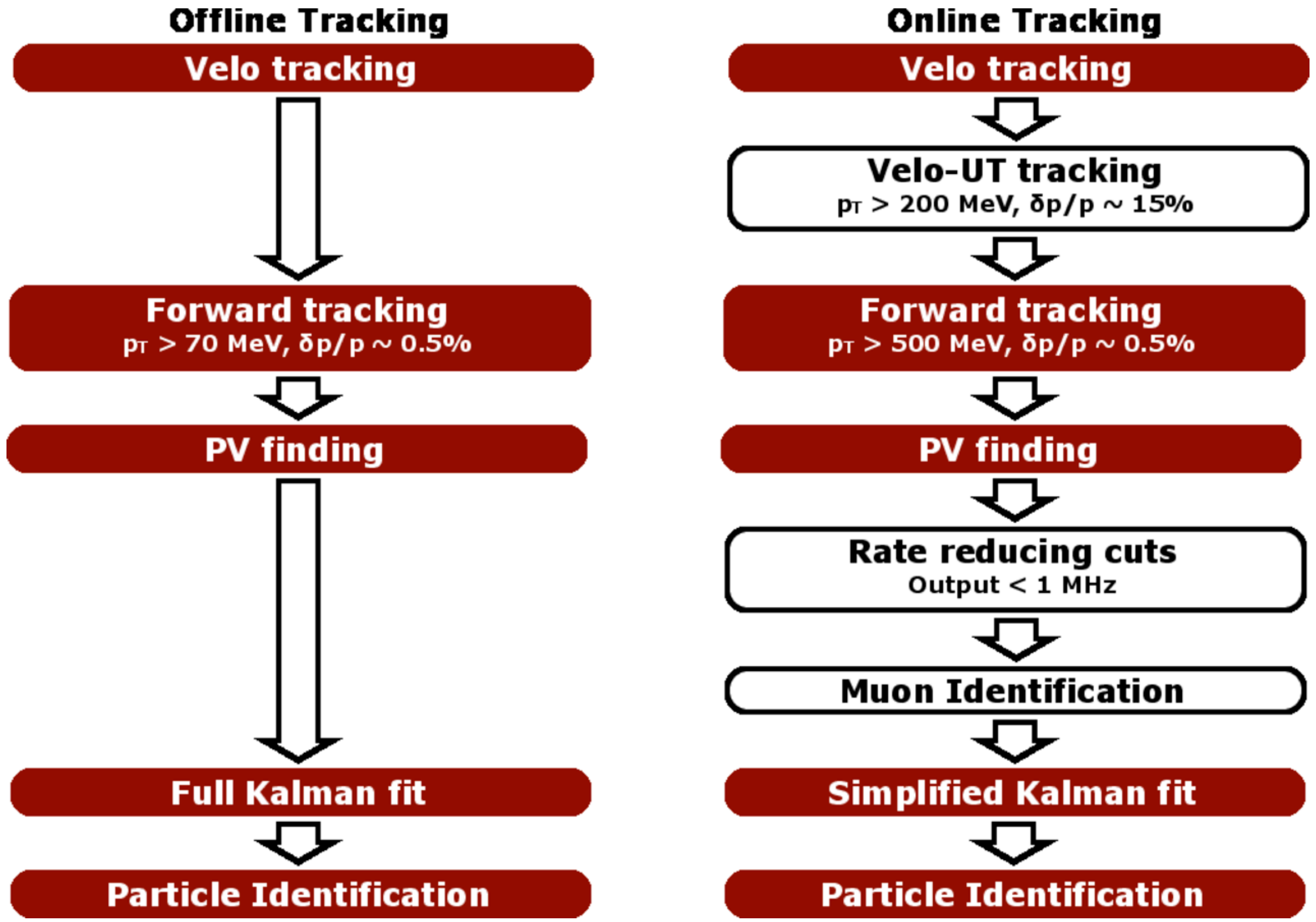


ALICE are fully committed to a GPU reconstruction for the TPC in particular. Already commissioned in Run I! Achieves a threefold increase in performance compared to CPU.

LHCb's 30 MHz reconstruction

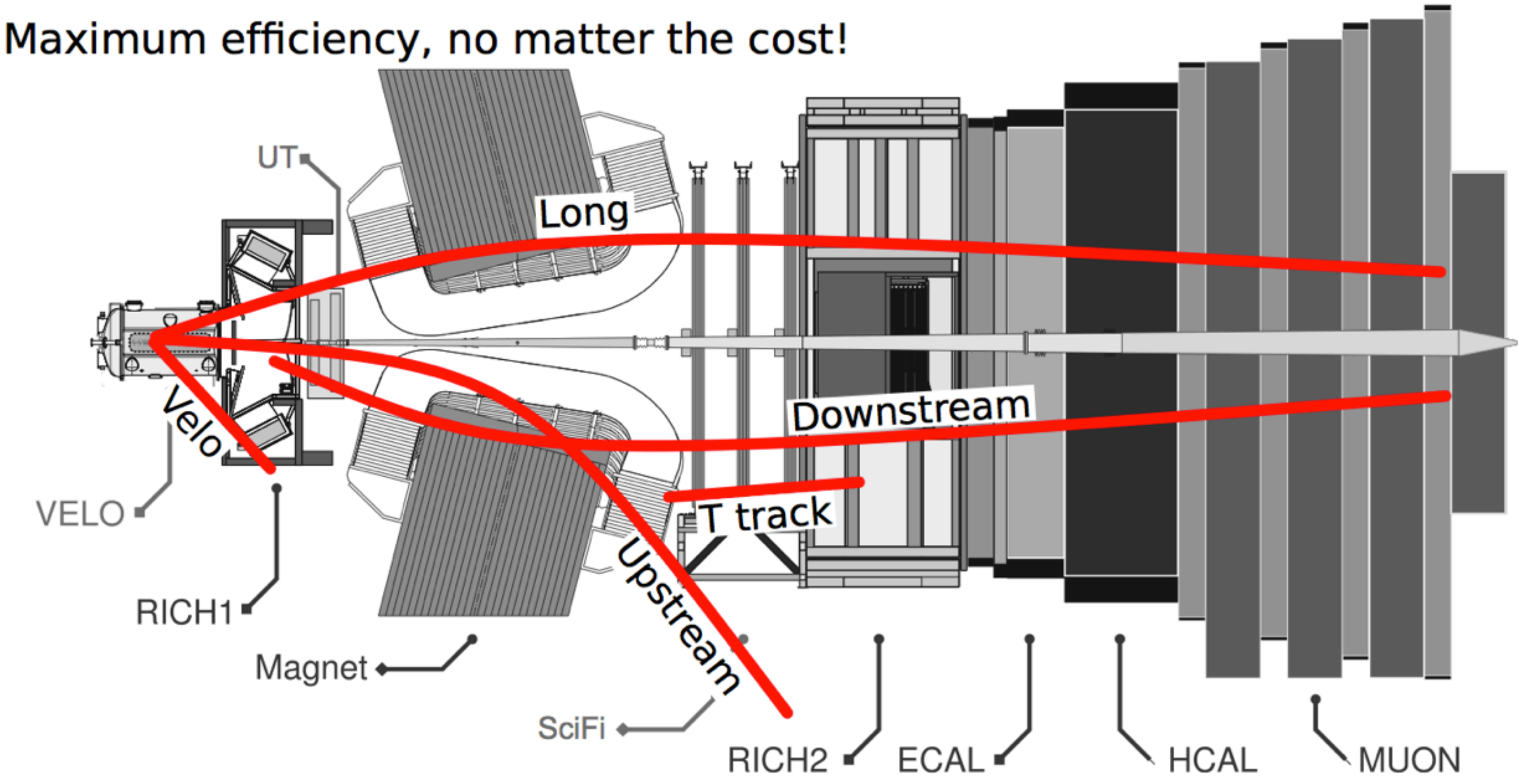


LHCb's 30 MHz reconstruction



Offline Tracking

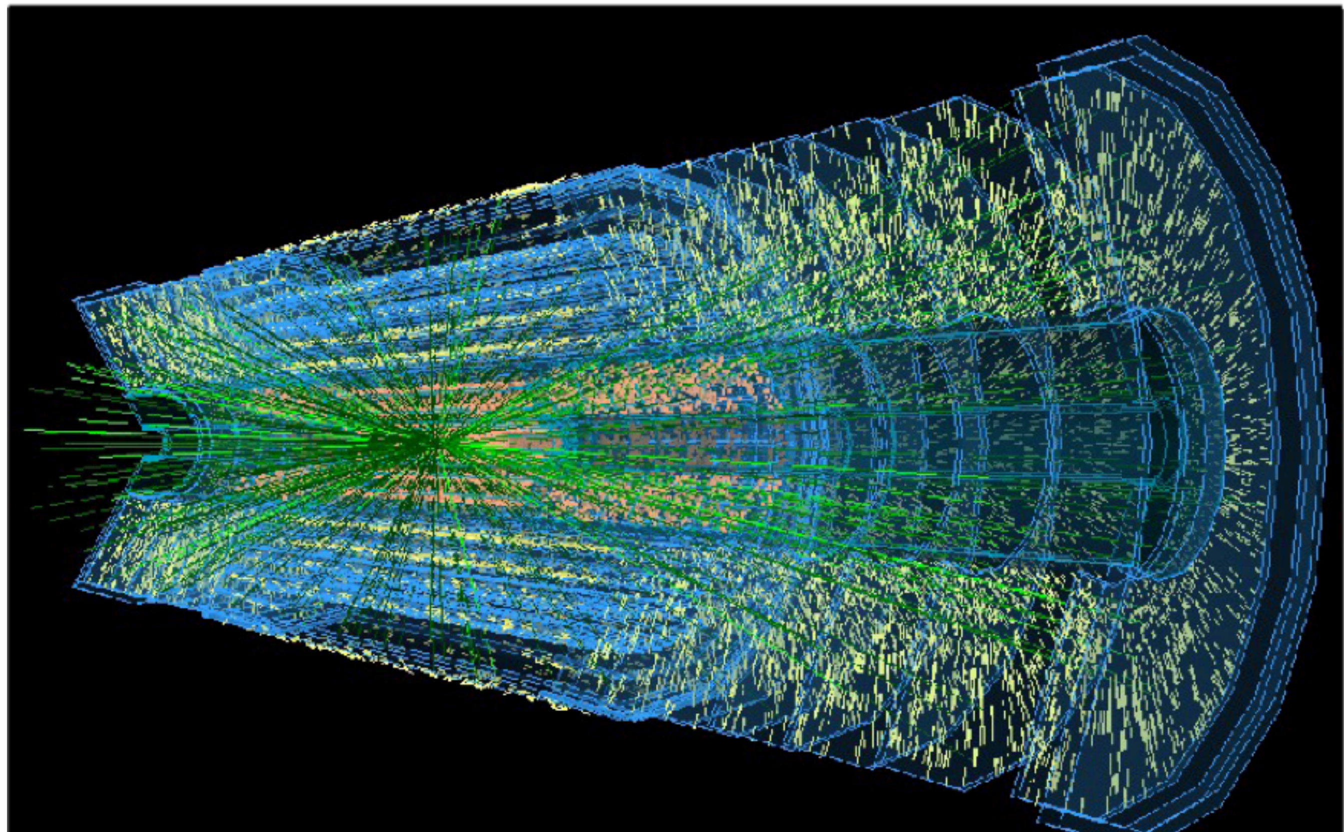
Maximum efficiency, no matter the cost!



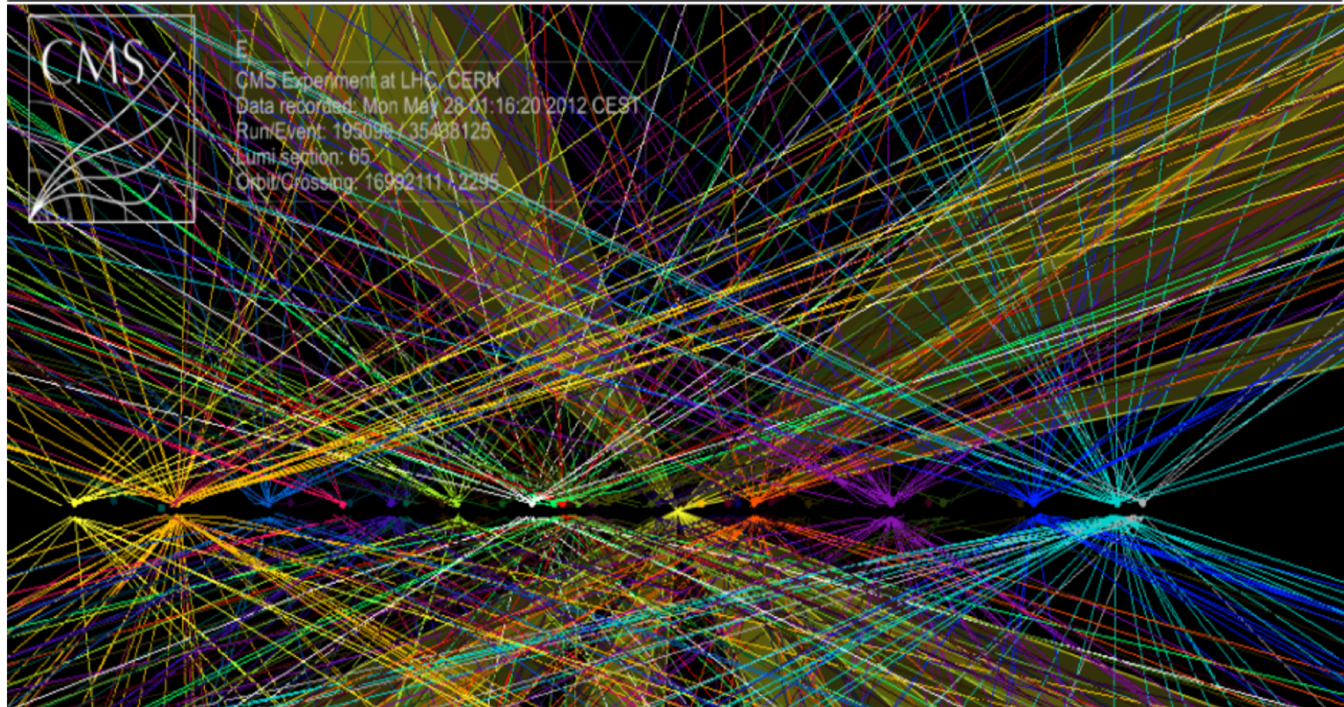
LHCb's vertex detector outside the dipole magnet makes it a slightly special case. Reconstruction timing is basically linear with instantaneous lumi/pileup. Because we want to catch low momentum tracks crossing the full detector volume it is not trivial to parallelize the track finding, although a lot work is ongoing into GPU coprocessors.

ATLAS/CMS reconstructions

Enormously challenging environment, and both experiments are significantly upgrading the tracking hardware to cope (not topic of this talk)



ATLAS HL-LHC event in new tracker (pileup of 140)



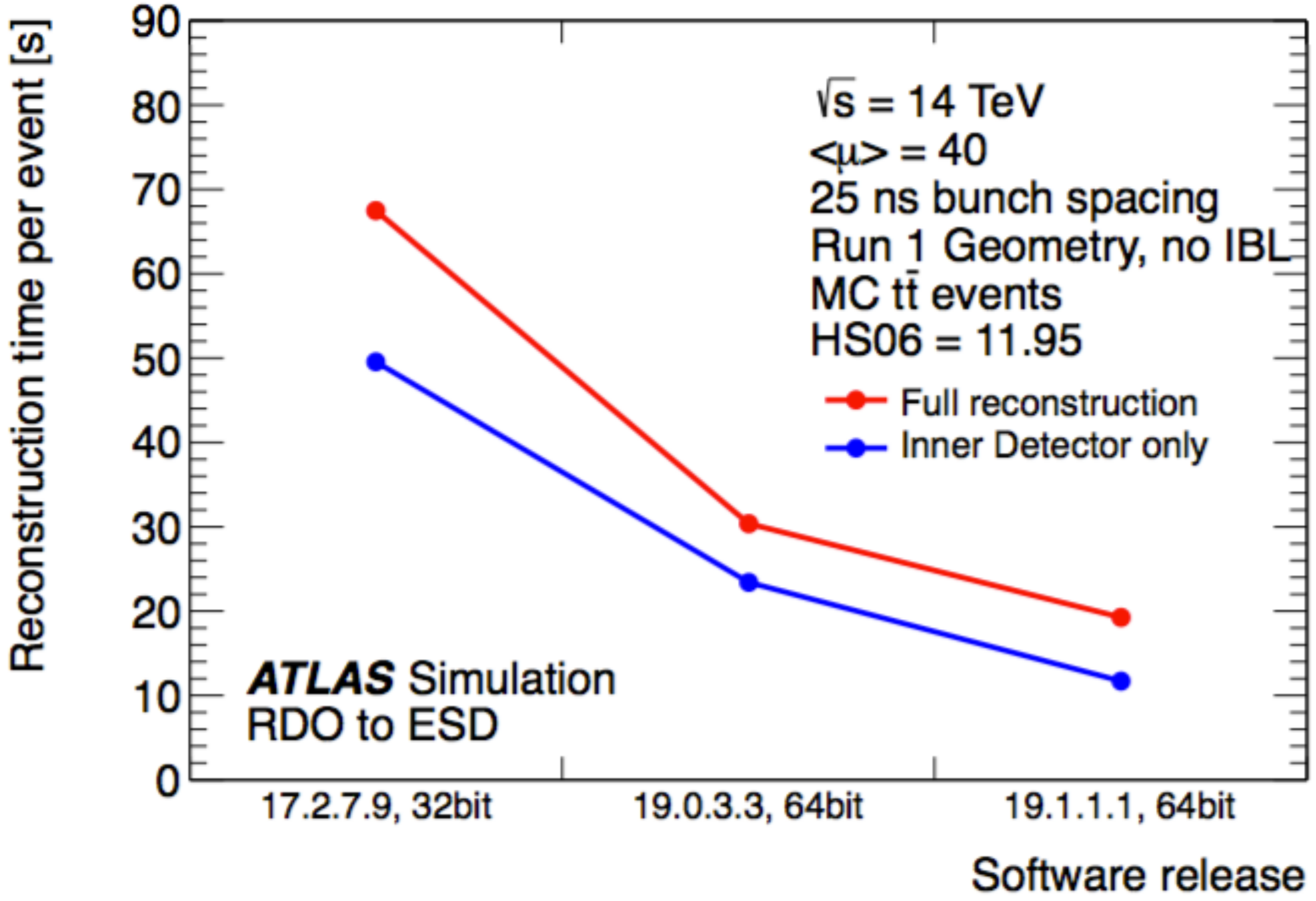
CMS event with 50 pileup

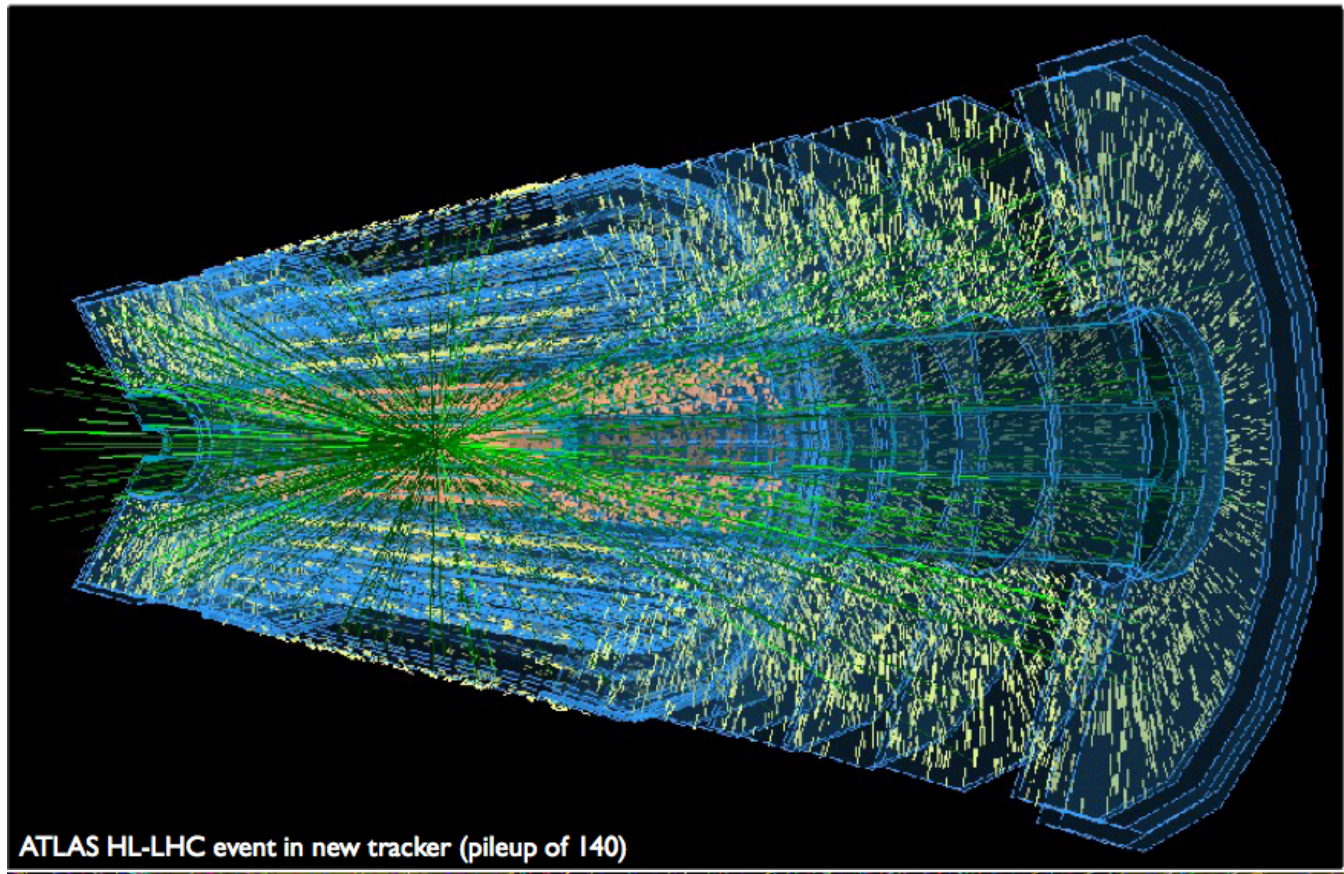
ATLAS/CMS software trigger tracking will be seeded by the L1 track trigger candidates

ATLAS/CMS reconstructions

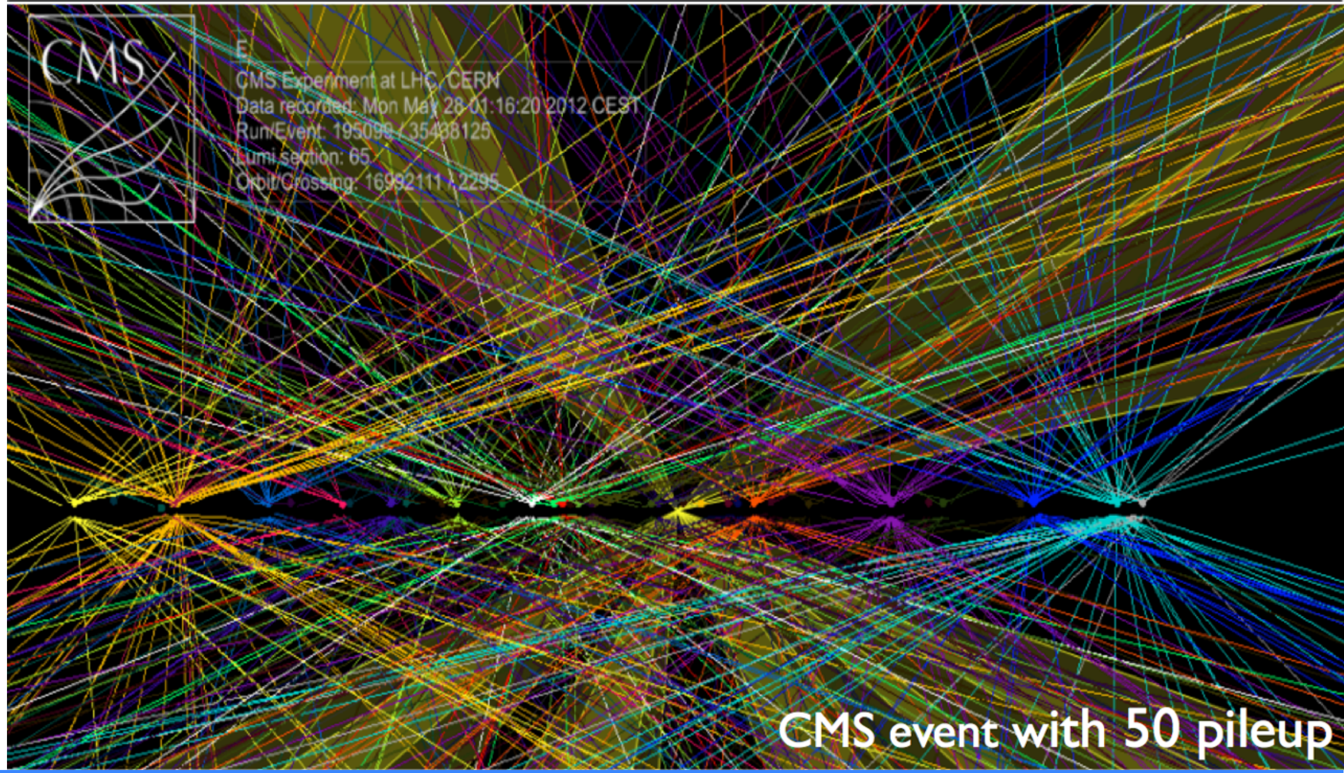
Already a lot of work for Run2, vectorizing code is a hot topic (also on LHCb/ALICE). Also lots of work on optimal tracking algos for pileup.

ATLAS reports x3 gain for CPU, CMS x2. Will need more gains like that going towards HL-LHC!





ATLAS HL-LHC event in new tracker (pileup of 140)



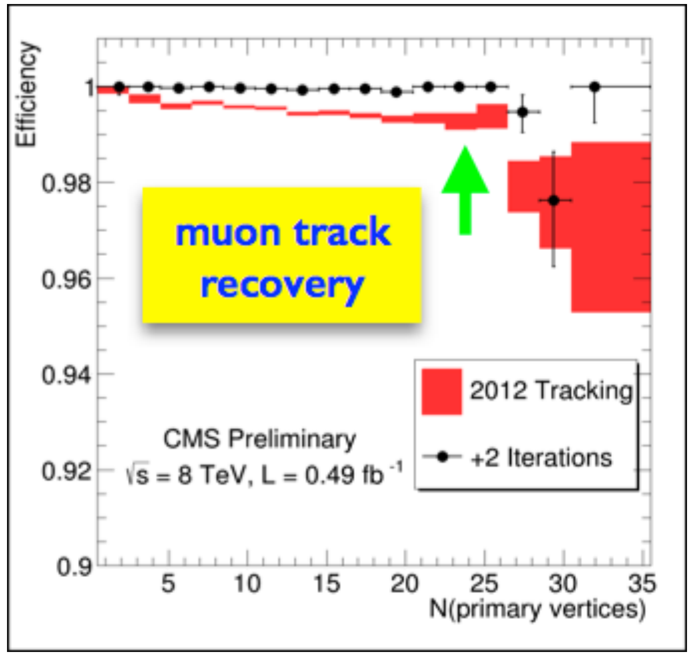
CMS event with 50 pileup

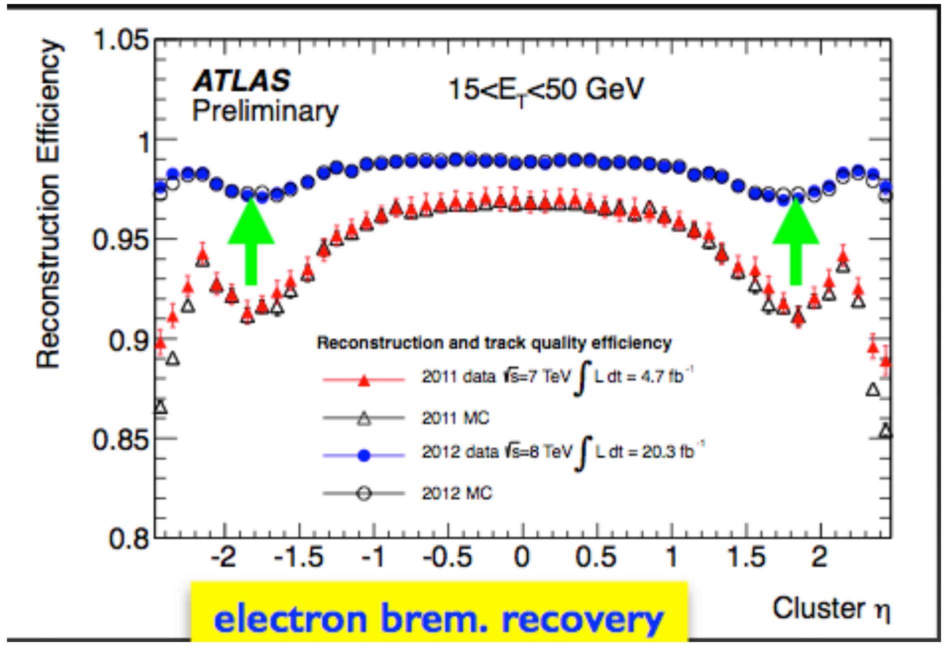
ATLAS/CMS software trigger tracking will be seeded by the L1 track trigger candidates

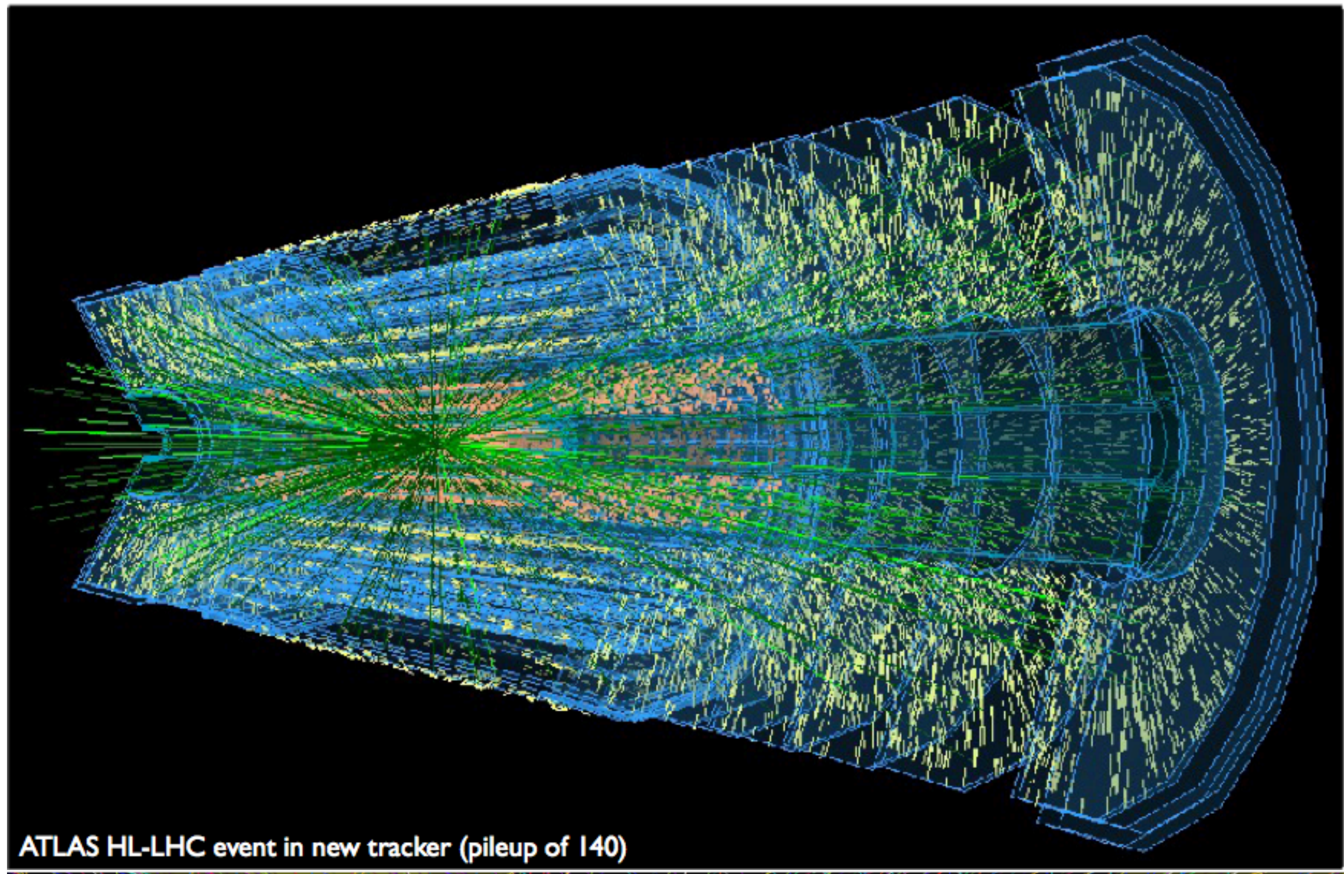
ATLAS/CMS reconstructions

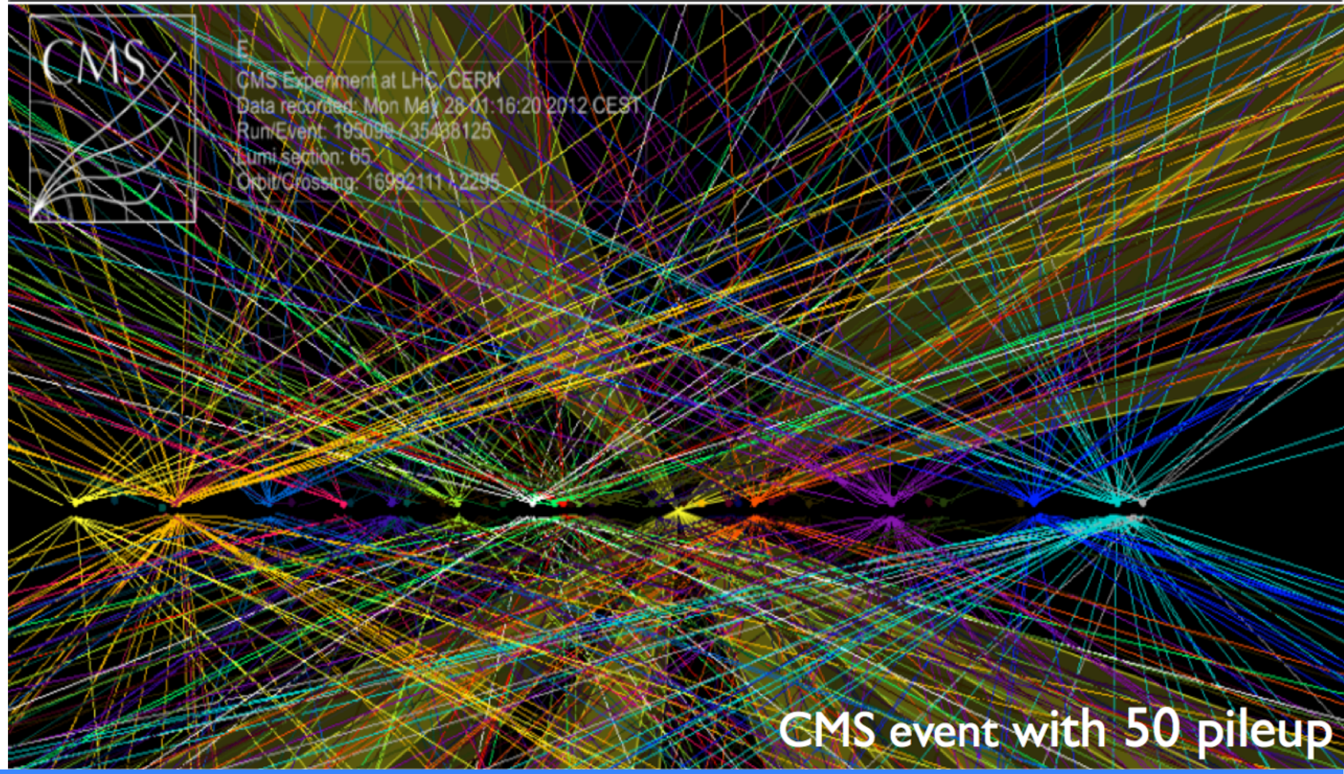
Also more aggressive ideas being studied, e.g. different tracking inside/outside the signal ROI.

Already used in RunI for brems/muon efficiency recovery. Expect to expand on these strategies.






ATLAS HL-LHC event in new tracker (pileup of 140)



CMS event with 50 pileup

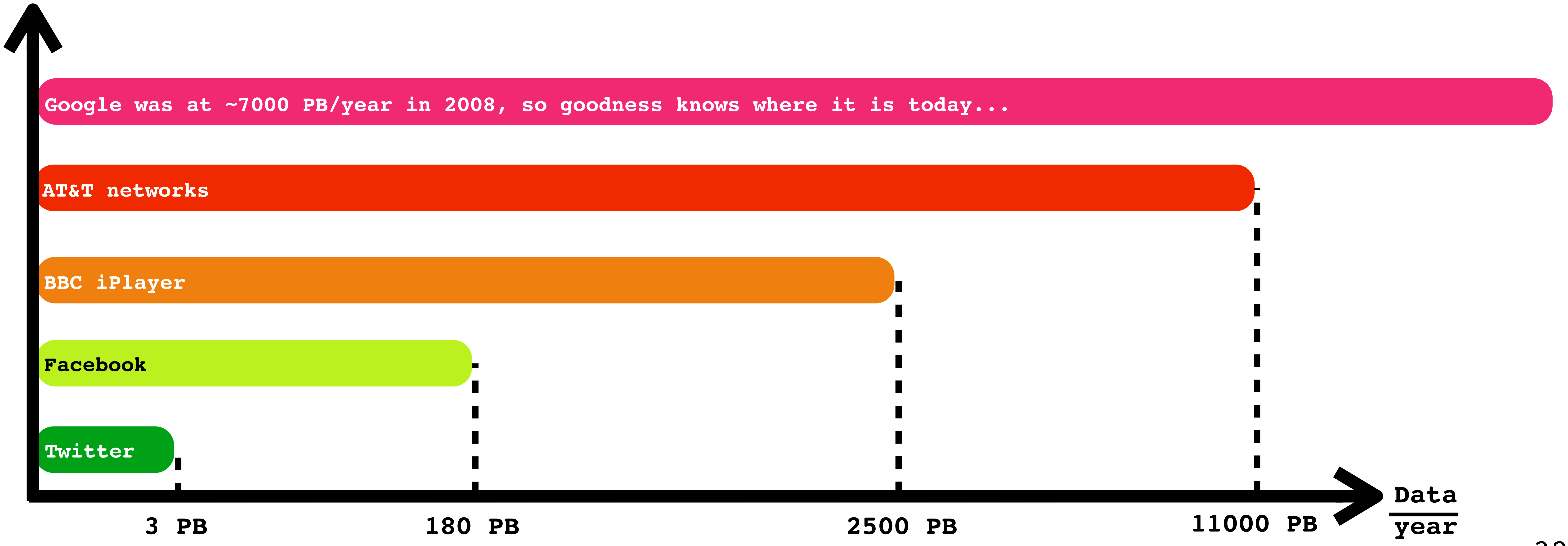
ATLAS/CMS software trigger tracking will be seeded by the L1 track trigger candidates


Software trigger menus
and real-time analysis

Big data, big opportunities

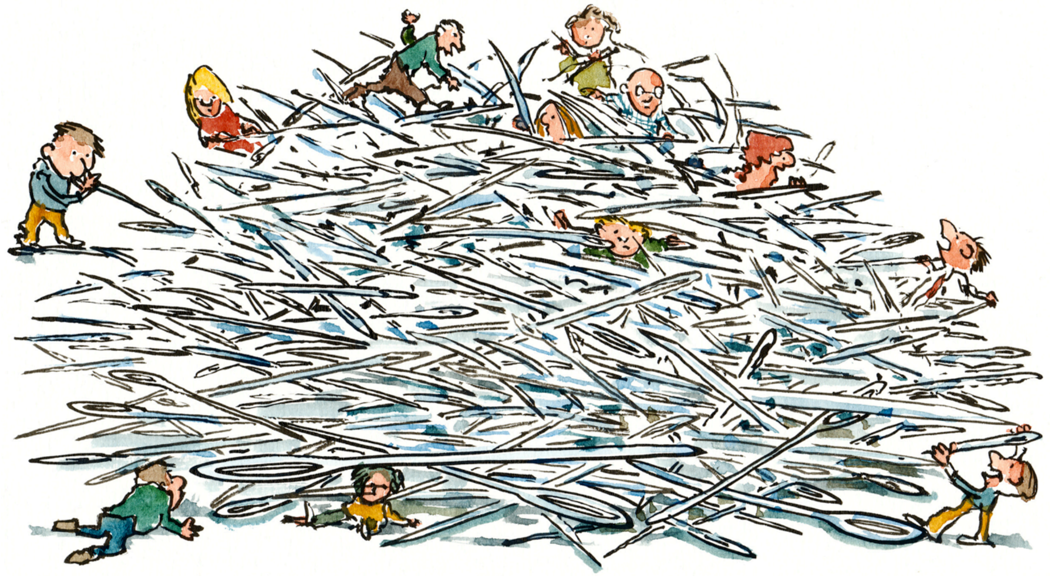
Input data rate of the LHCb upgrade post LS2 = 5 TB/second

This means ~20000 PB of data every year



A pinch of salt is needed but…

Triggers today



Triggers in the future

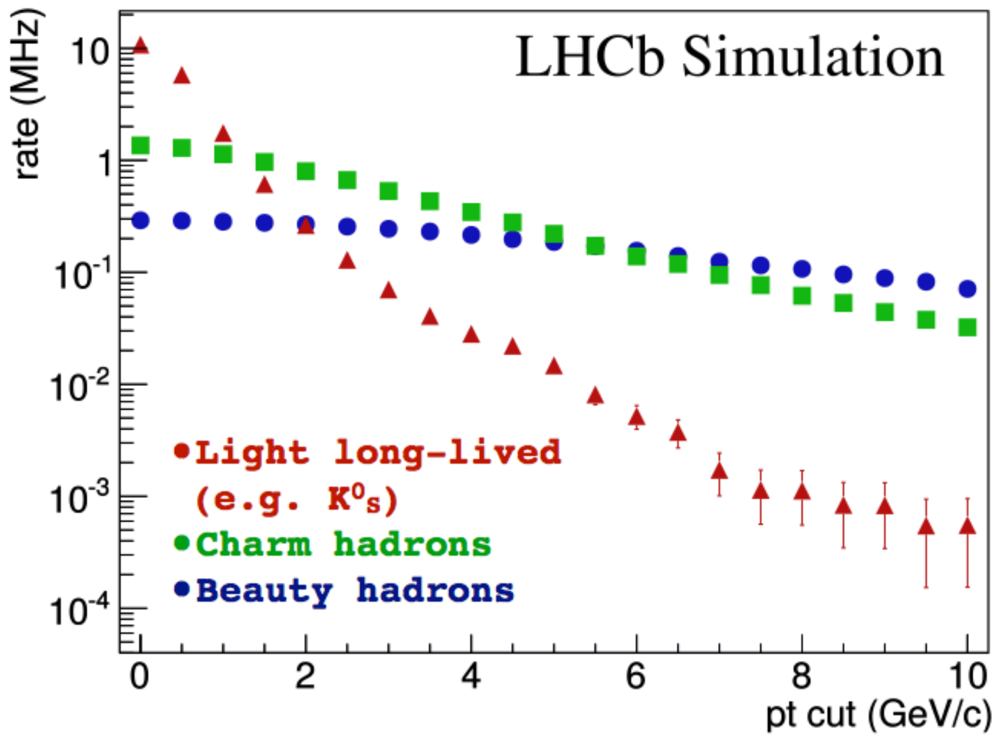
While I am going to mention menus, there are enormous "parasitic" opportunities for physics beyond the core programmes at the HL-LHC, and we should expect these to evolve and compete for output bandwidth with the "core" physics for both ATLAS/CMS and LHCb as we approach the HL-LHC era.

Remember : ALICE keeps all interactions, hence no HLT "menu" as such.

LHCb HLT menus

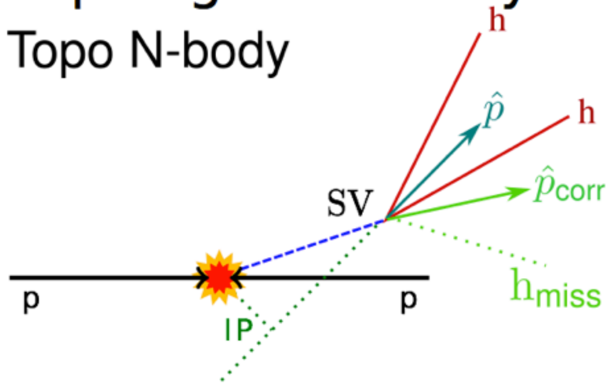
Because of the offline-like reconstruction, can in principle select any Beauty/Charm decay to charged tracks (and some with neutrals) at HLT level.

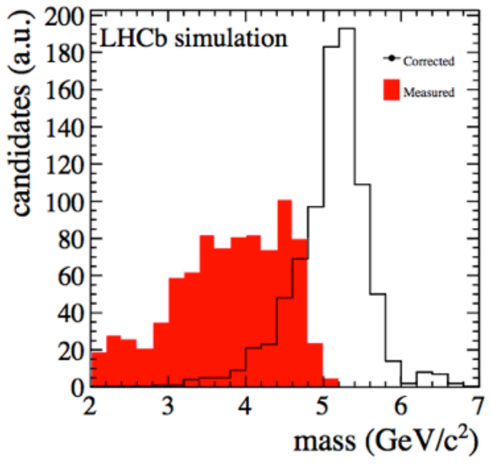
Several output rate scenarios being considered, main driver is what we want to do with charm physics. 2-10 Gb/s output rate foreseen.



Topological *N*-Body Trigger

Topo N-body

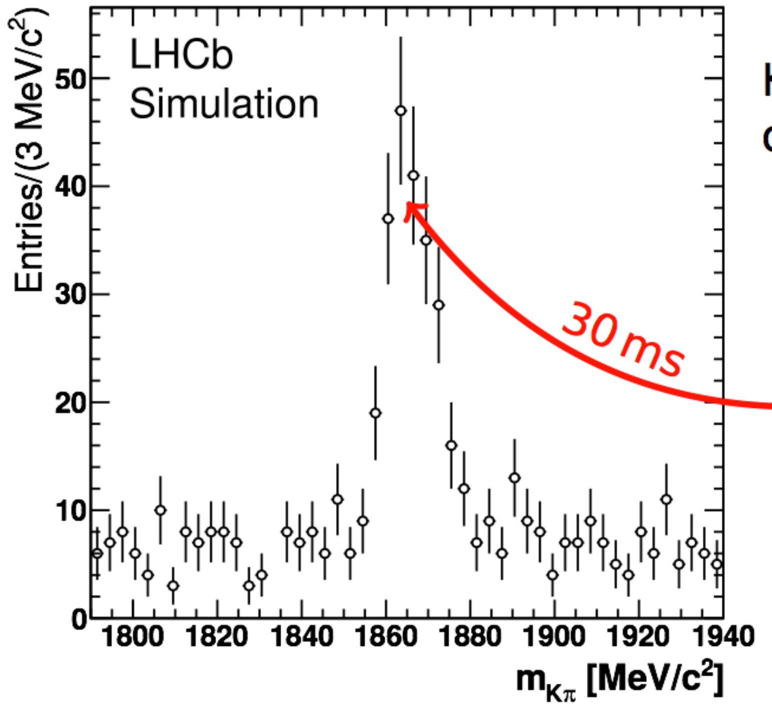




- Main trigger for B decays is based on a Boosted Decision Tree
- Inclusive trigger for 2, 3, 4-body detached vertices
- Preselect tracks based on distance to PV, scalar and vector sum of p_T
- BDT inputs: p_T, IP_{χ^2}, flight distance χ^2, mass and corrected mass:

$$m_{corr} = \sqrt{m^2 + |p_{T_{miss}}|^2} + |p_{T_{miss}}|$$

Exclusive selections



Key challenges: combinatorics and output rate

- $B^0, D^0 \to h^+h^-$
  - Timing: 0.13 ms

$B^0 \to h^+h^-$	~ 1 kHz
$D^0 \to K^-\pi^+$	~ 20 kHz
$D^0 \to K^+\pi^-, \pi\pi$	~ 40 kHz
$D^0 \to KK$	~ 2 kHz

- $B_s \to \phi(\to KK)\phi(\to KK)$
  - Timing: 0.1 ms, Rate: ~ 12 Hz

ATLAS/CMS menus

CMS Category	L1 Triggers	L1 rate (w/ overlaps)	Required reduction	HLT rate
Muons	μ, μμ	21 kHz	~ 21	1 kHz
E/Gamma	e, ee, iso-e, γ, γγ	102 kHz	~ 102	1 kHz
Taus	τ, ττ, e+τ, μ+τ	75 kHz	~ 75	1 kHz
Hadronic	jets, e+MHT, μ+MHT, HTT	138 kHz	~ 138	1 kHz
Others	MET, others	160 kHz	~ 160	1 kHz
Total rate (w/o overlaps)		500 kHz	100	5 kHz

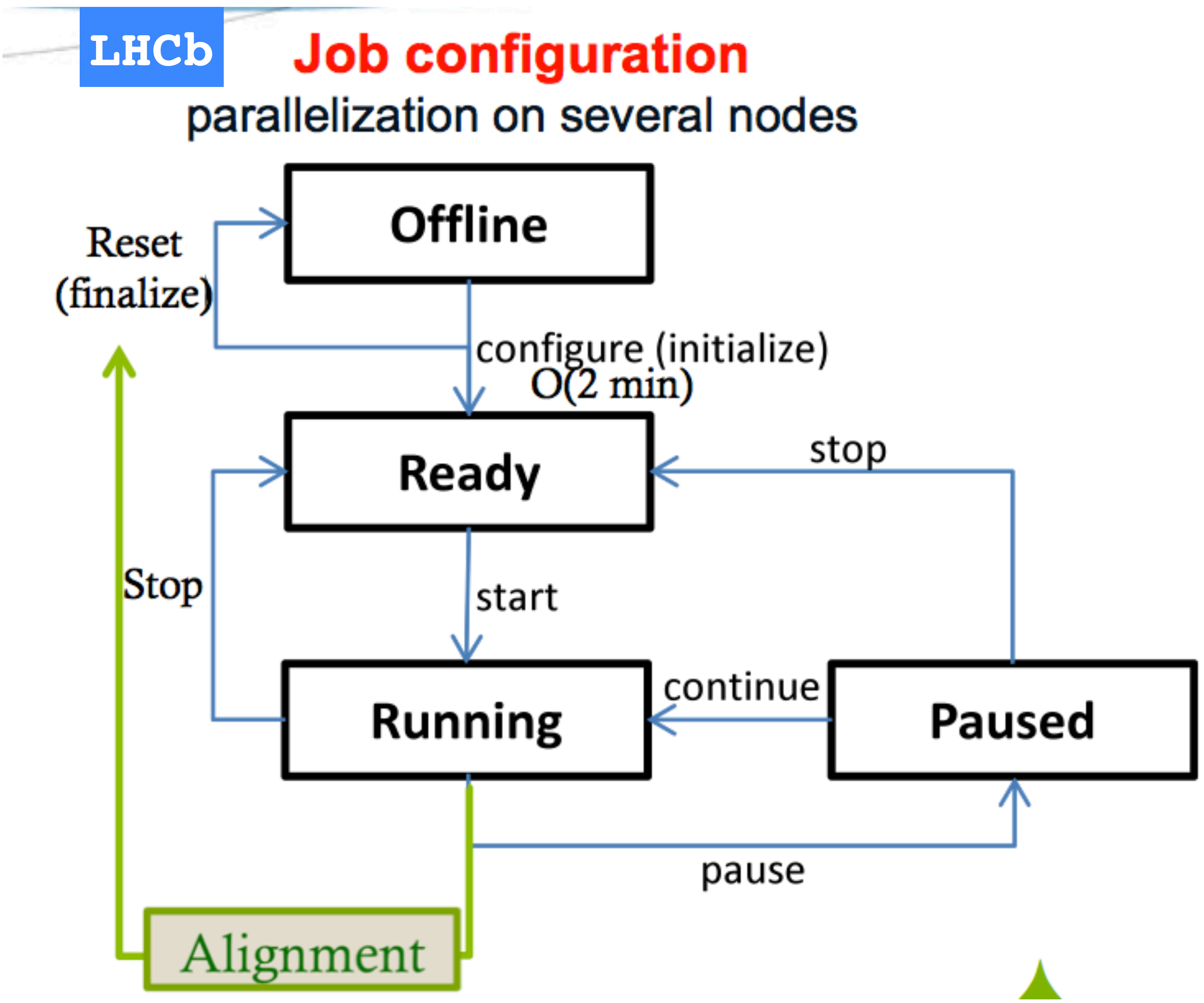
VERY PRELIMINARY!

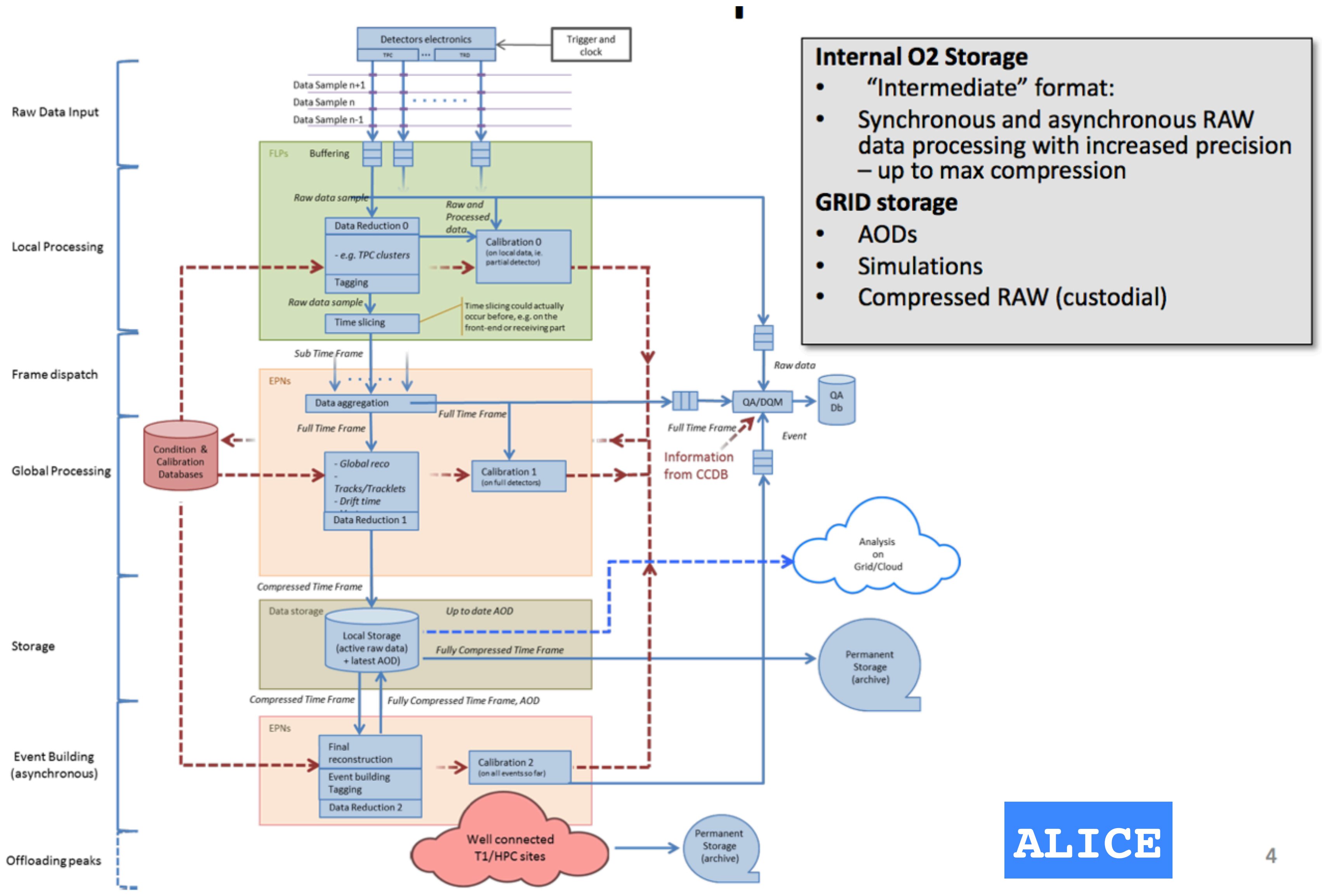
Somewhat different foreseen HLT rejection rates

100:1 for CMS and 40:1 for ATLAS.

Menus very sketchy at present, which is understandable because really the reconstruction questions are more pressing.

Real time detector calibration

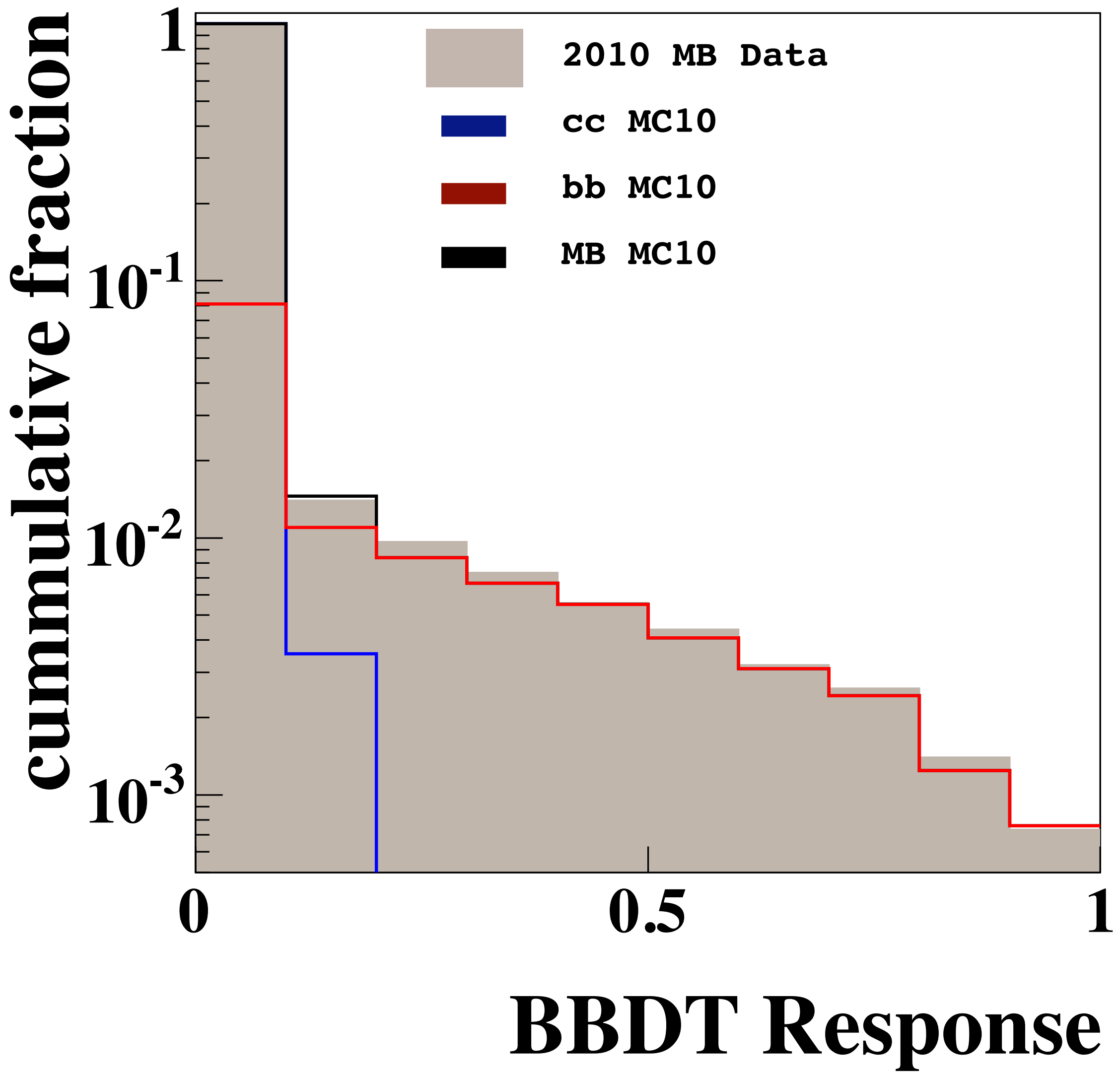




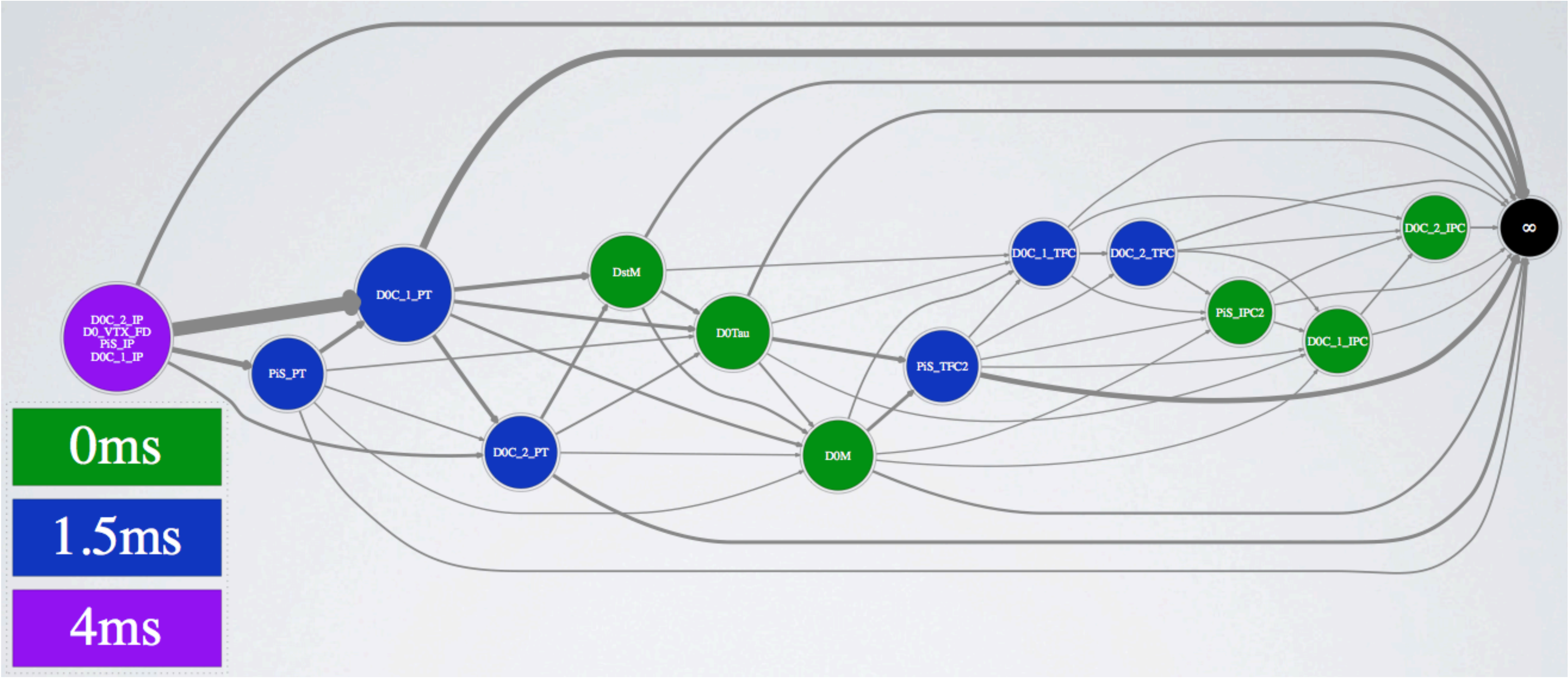
Both LHCb and ALICE plan a real-time detector alignment and calibration. In the LHCb case this is absolutely critical because it enables hadronic particle identification to be used in the trigger. Not clear whether CMS/ATLAS need or want to go down this road.

Real time multivariate analyses

LHCb topological trigger



MDDAG, Benbouzid, Kegl et al.



Well known that multivariate analyses perform better than so-called “cut-based” approaches. Now making their way into HLT algorithms, e.g. LHCb’s inclusive b-physics trigger in Run I. Real-time data analysis is an area where the private sector invests a lot, expect significant improvements as a result of collaborations over coming years.

Ceterum censeo…

MORE
IS
MORE

The basic approach of all four collaborations can be summarized as follows : put as much as DAQ will allow into software triggers.

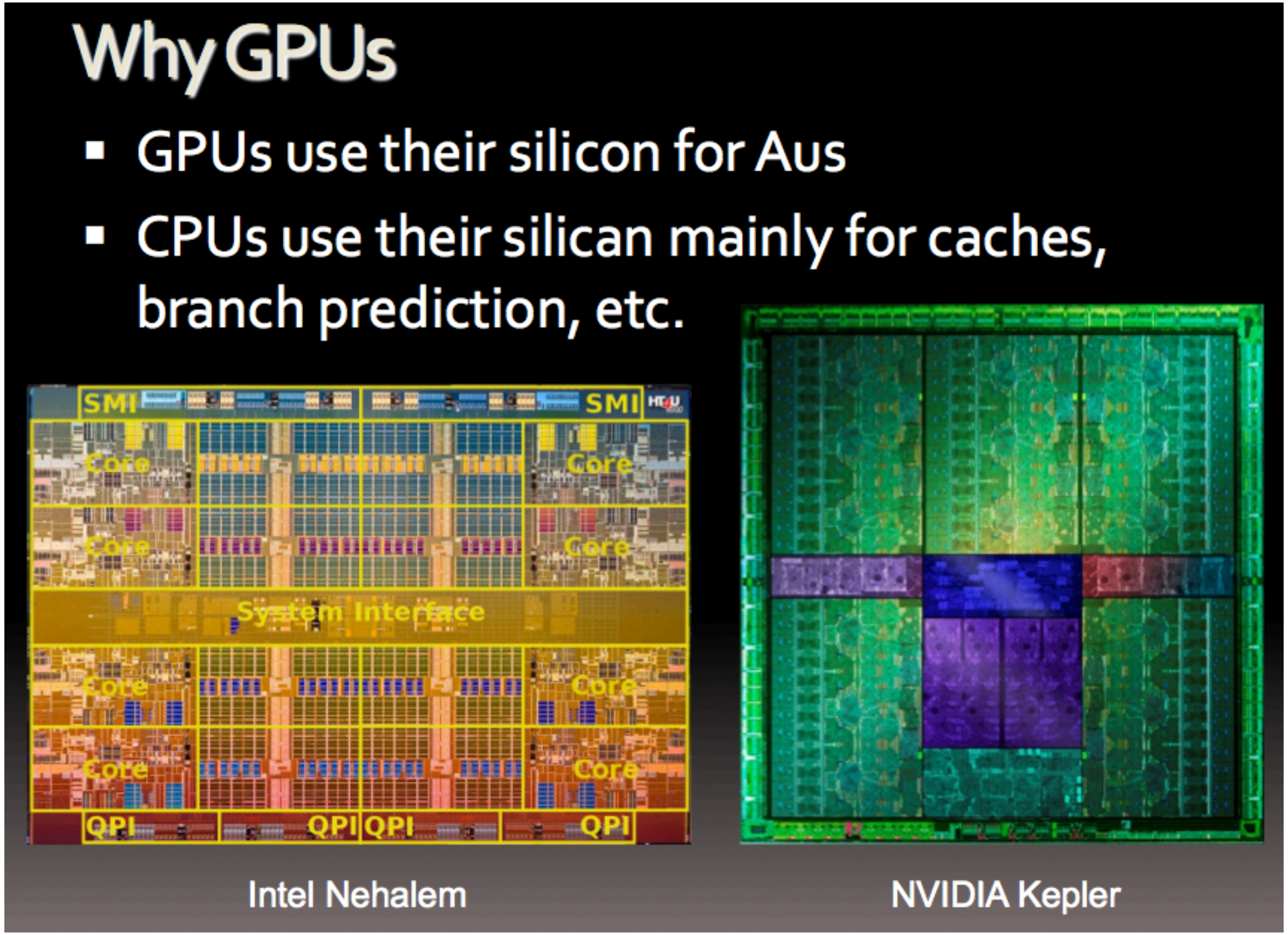
Nevertheless “physics” and hardware constraints are leading to implementation differences.

Will be critical to fully exploit multi-core architectures and opportunities for parallelism in algorithms if software triggers are to reach their full potential!

Another big thank you to all the working group members whose slides/results I have stolen!

Backups

ALICE’s GPU tracking



Naively, GPUs gain as long as the cores don’t have to talk to each other.

LHCb DAQ

LHCb's DAQ network built around a bidirectional eventbuilding farm.

Note that about 80% of the CPU in the event-building PCs remains free for implementing the "low-level trigger" (selecting on muon and CALO primitives) and/or the first stages of the event reconstruction.

Need to transport/build 40 Tbit/s

	LHCb Run1 & 2	LHCb Run 3
Max. inst. luminosity	4 x 10^32	2 x 10^33
Event-size (mean – zero-suppressed) [kB]	~ 60 (L0 accepted)	~ 100
Event-building rate [MHz]	1	40
# read-out boards	~ 330	400 - 500
link speed from detector [Gbit/s]	1.6	4.5
output data-rate / read-out board [Gbit/s]	4	100
# detector-links / readout-board	up to 24	up to 48
# farm-nodes	~ 1000 (+ 500 in 2015)	1000 - 4000
# links 100 Gbit/s (from event-builder PCs)	n/a	400 - 500
final output rate to tape [kHz]	5	20 - 100