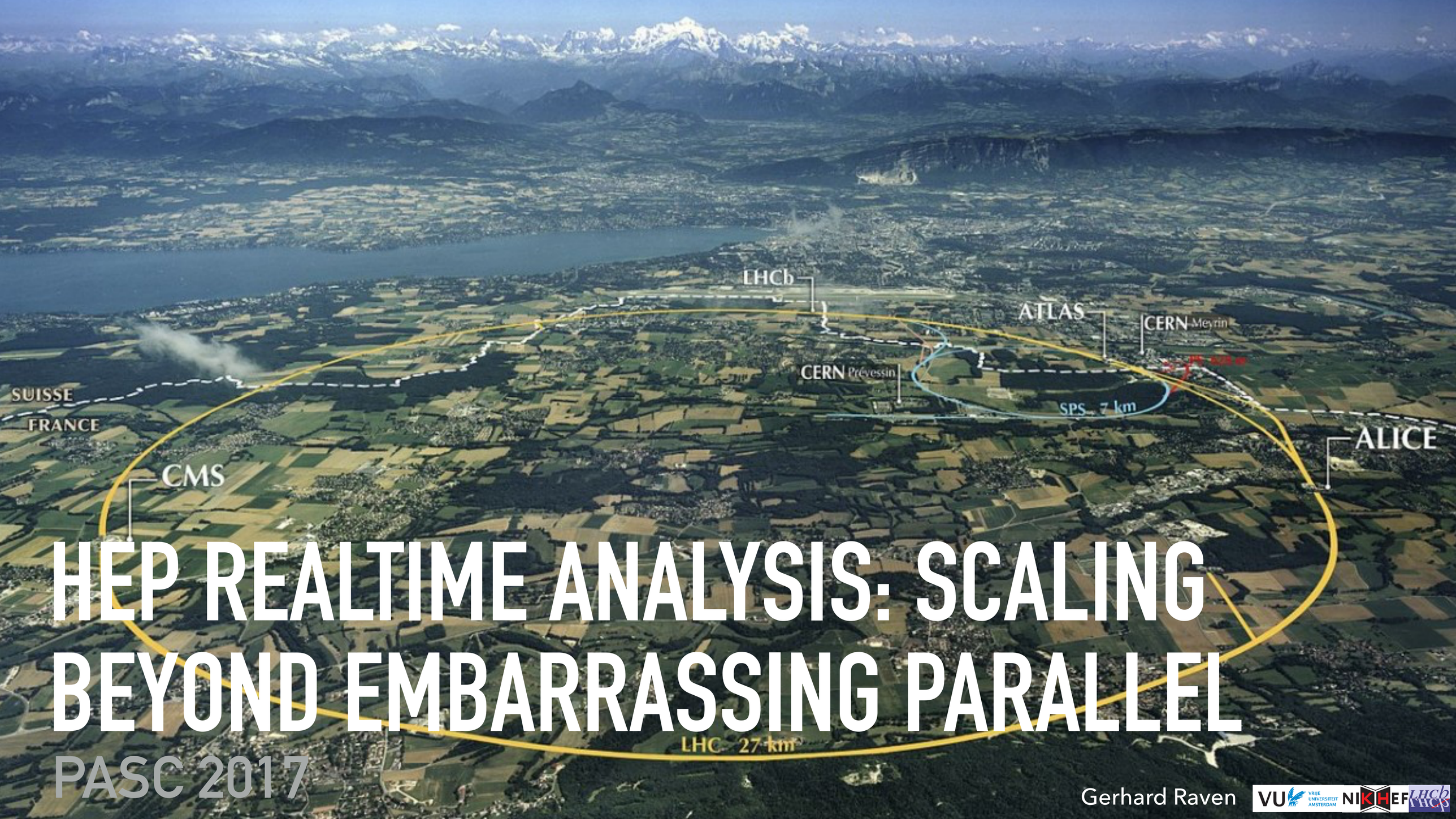


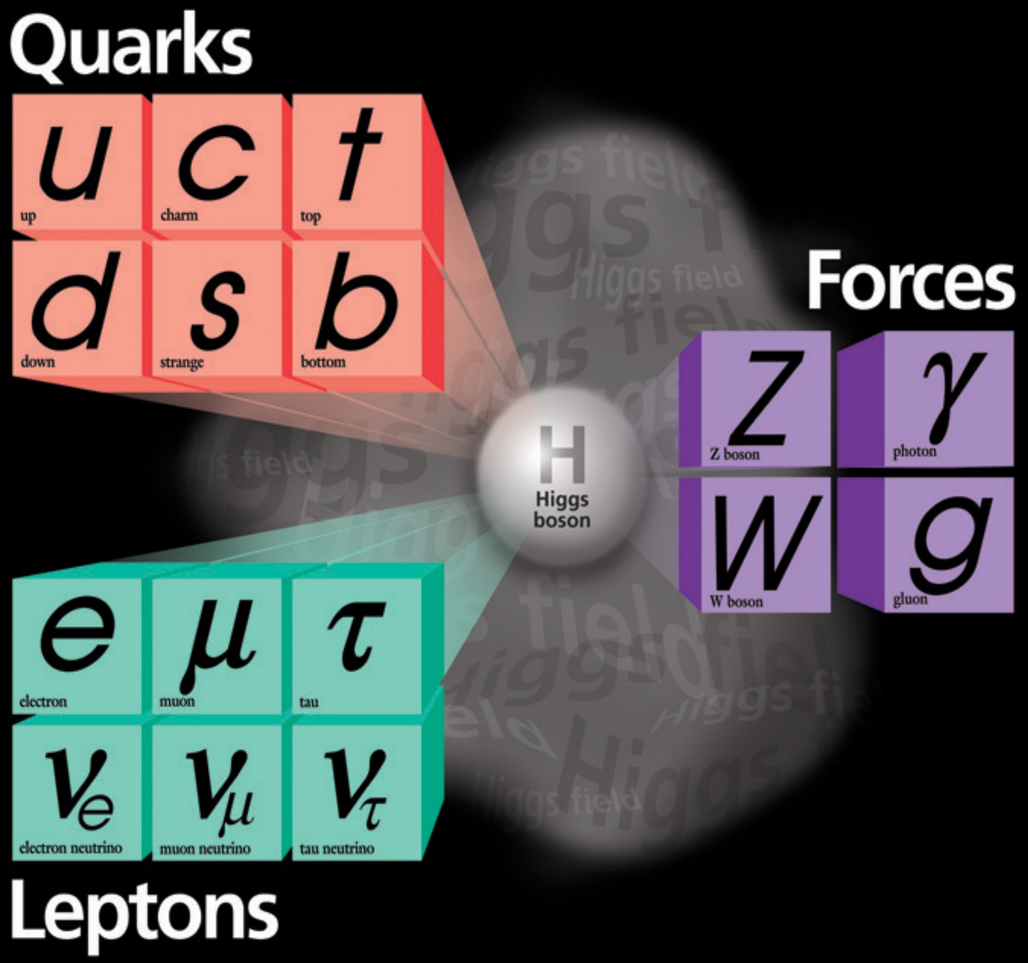
HEP REALTIME ANALYSIS: SCALING BEYOND EMBARRASSING PARALLEL

PASC 2017

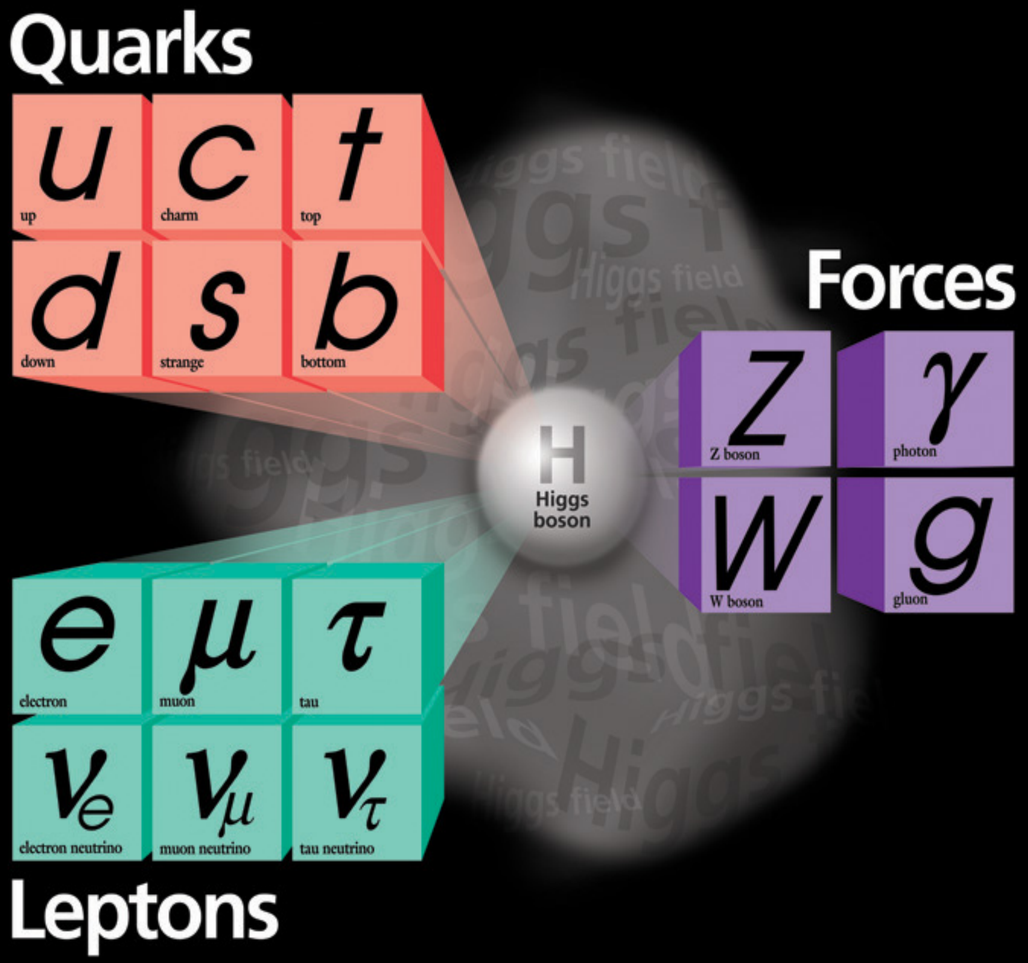
Gerhard Raven

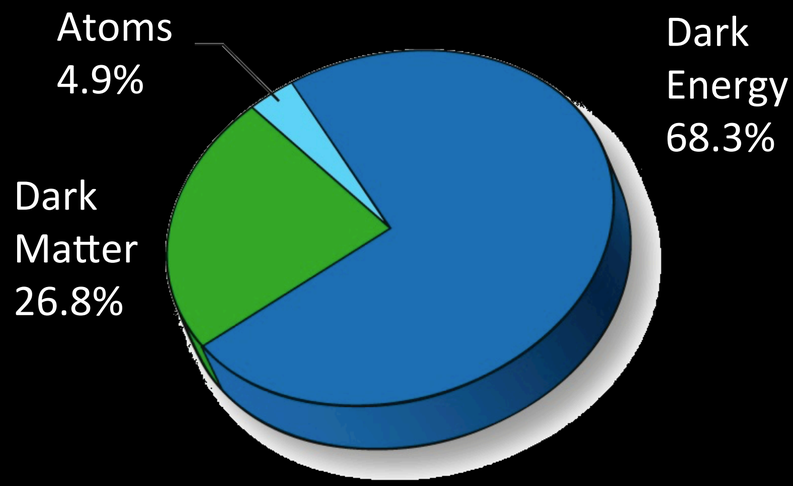
VU VRIJE UNIVERSITEIT AMSTERDAM NIKHEF LHCb

The Field of Particle Physics

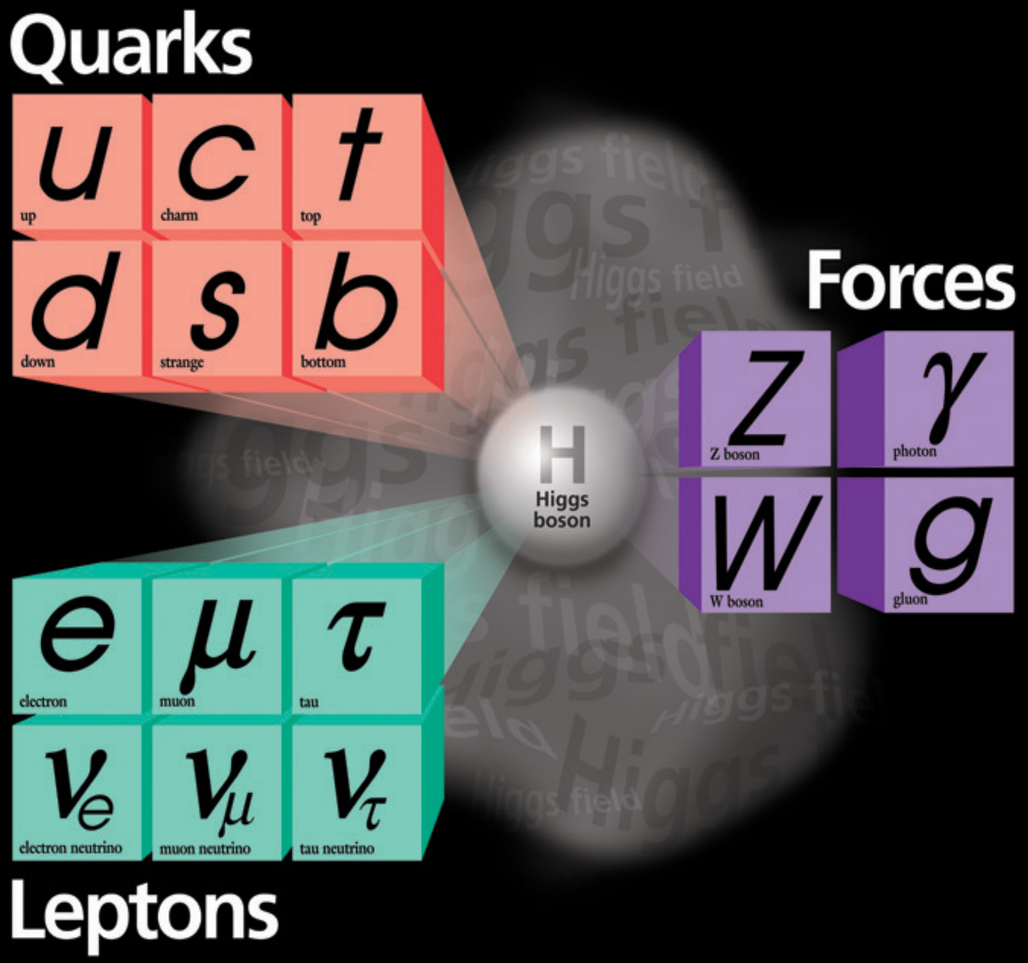


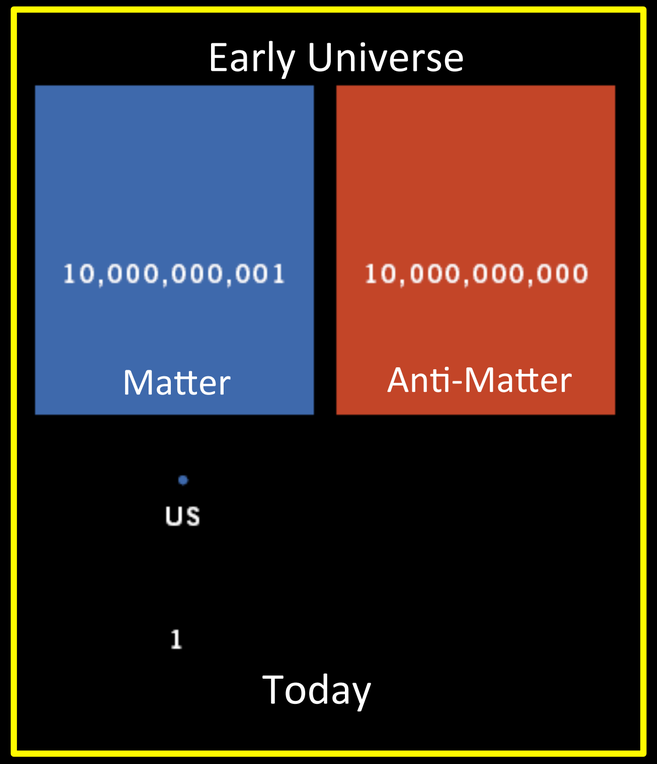
The Field of Particle Physics

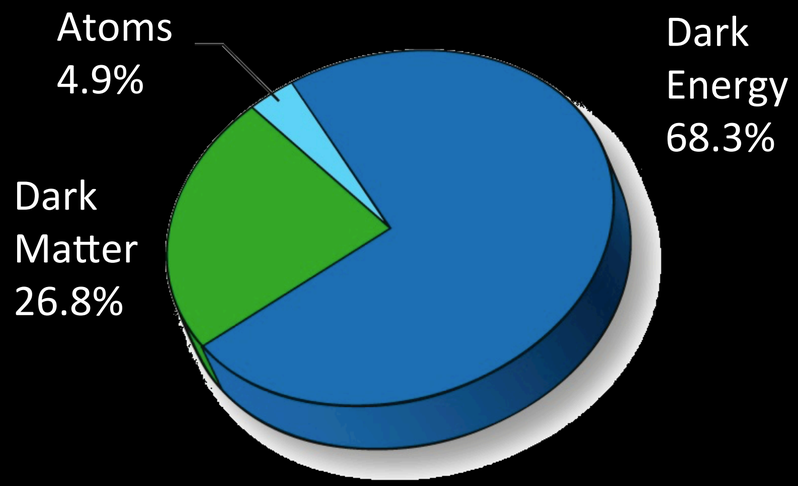




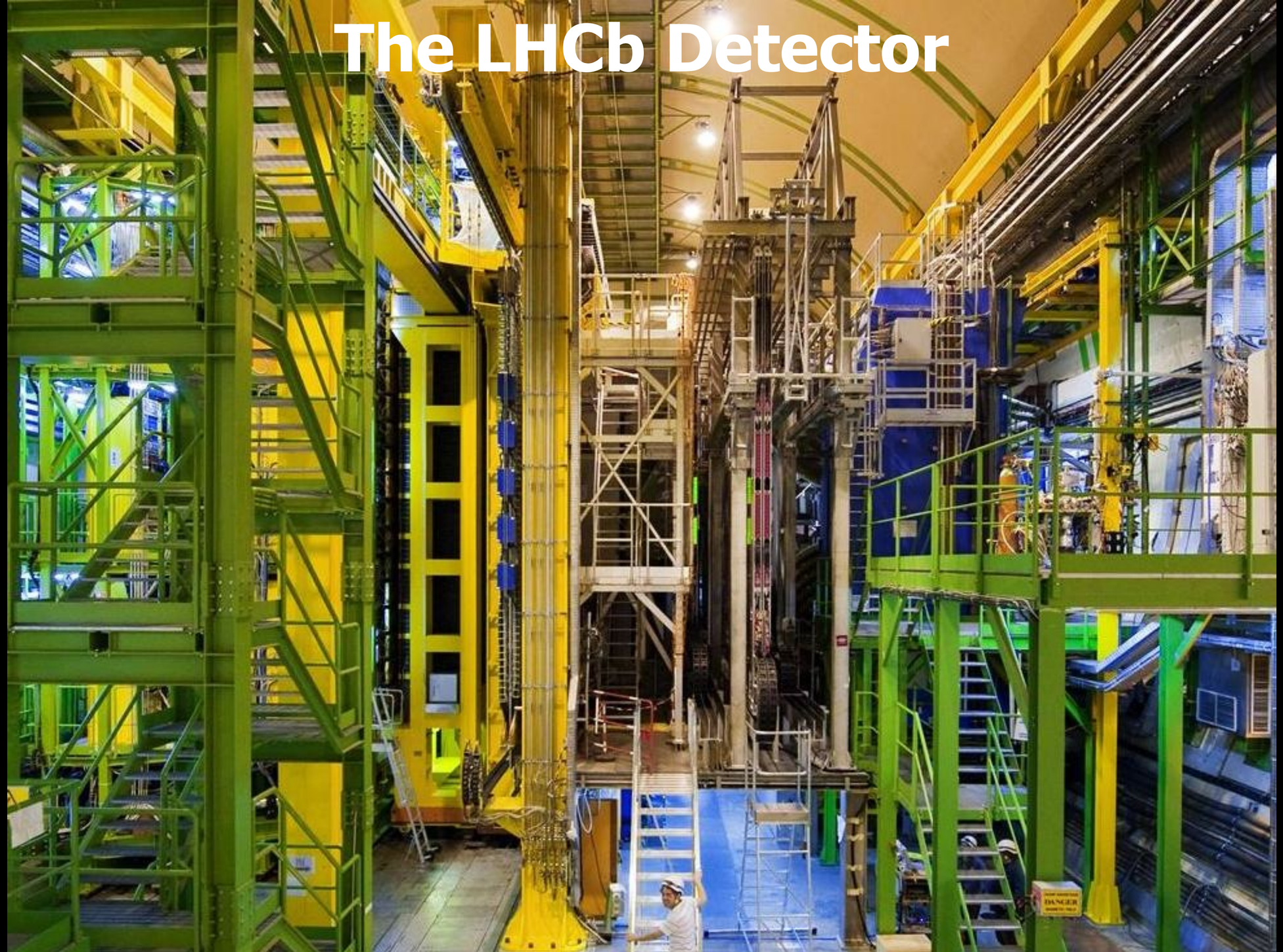
The Field of Particle Physics



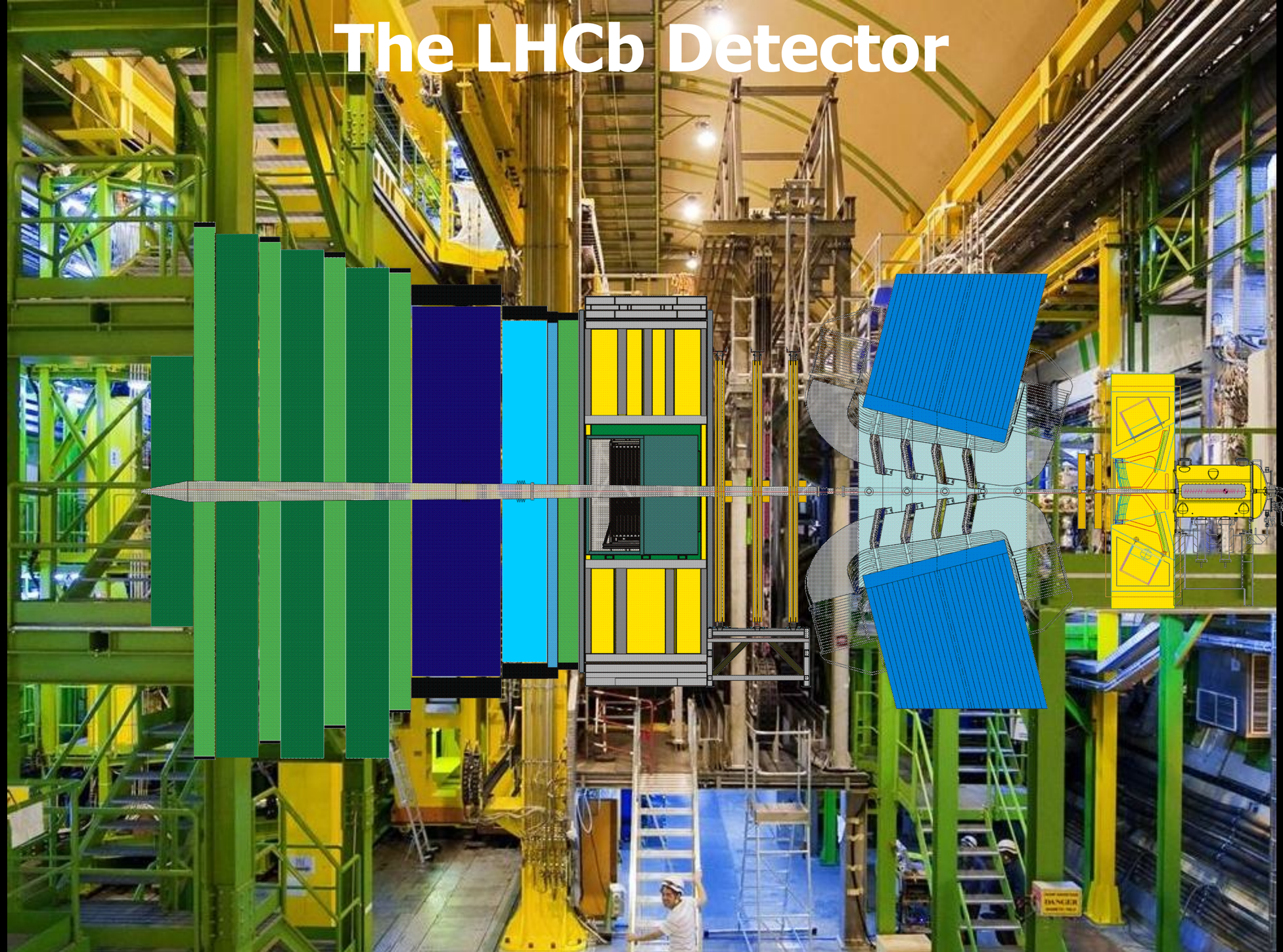




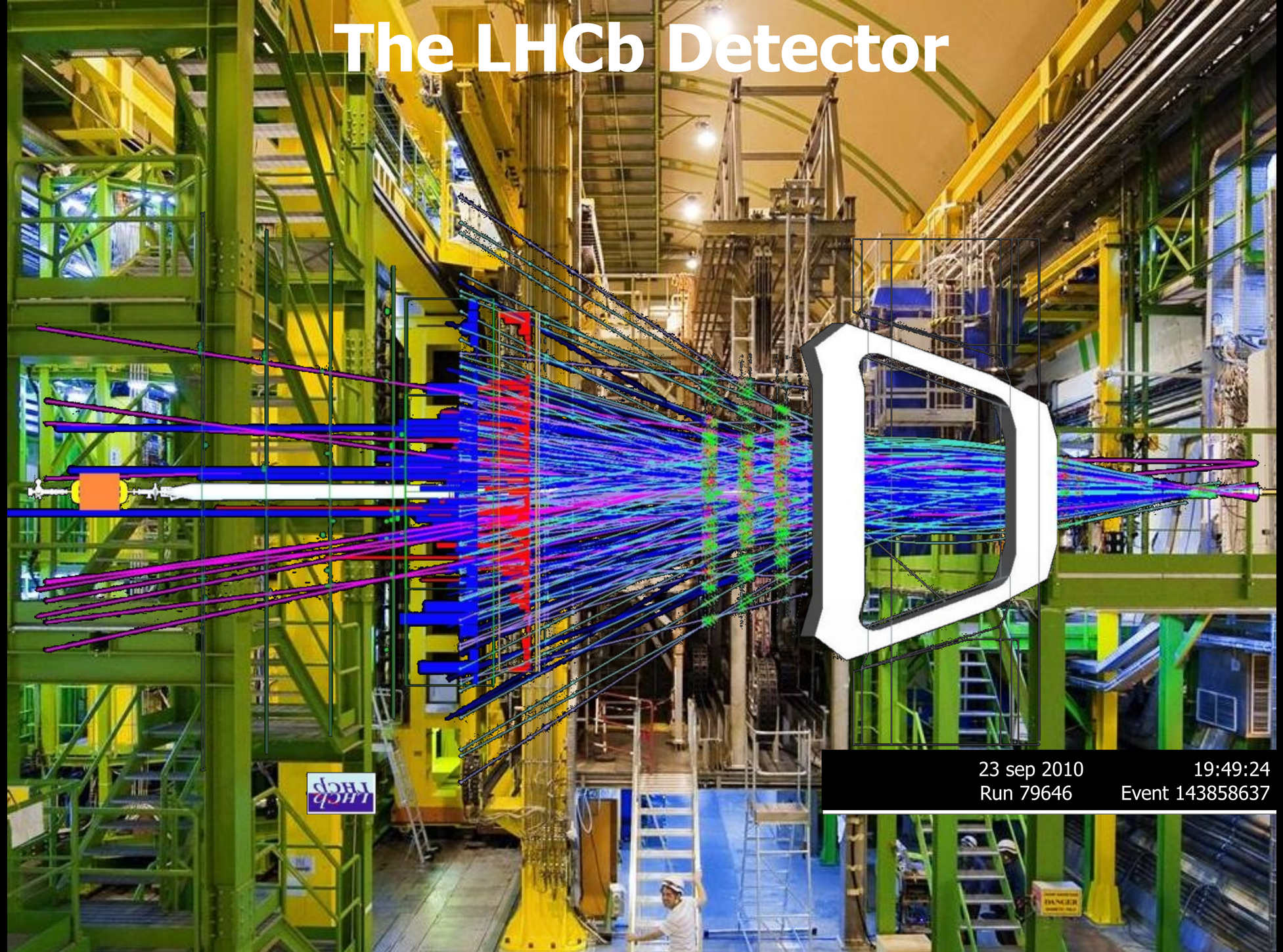
The LHCb Detector



The LHCb Detector



The LHCb Detector



The LHCb Detector

23 sep 2010 19:49:24
Run 79646 Event 143858637

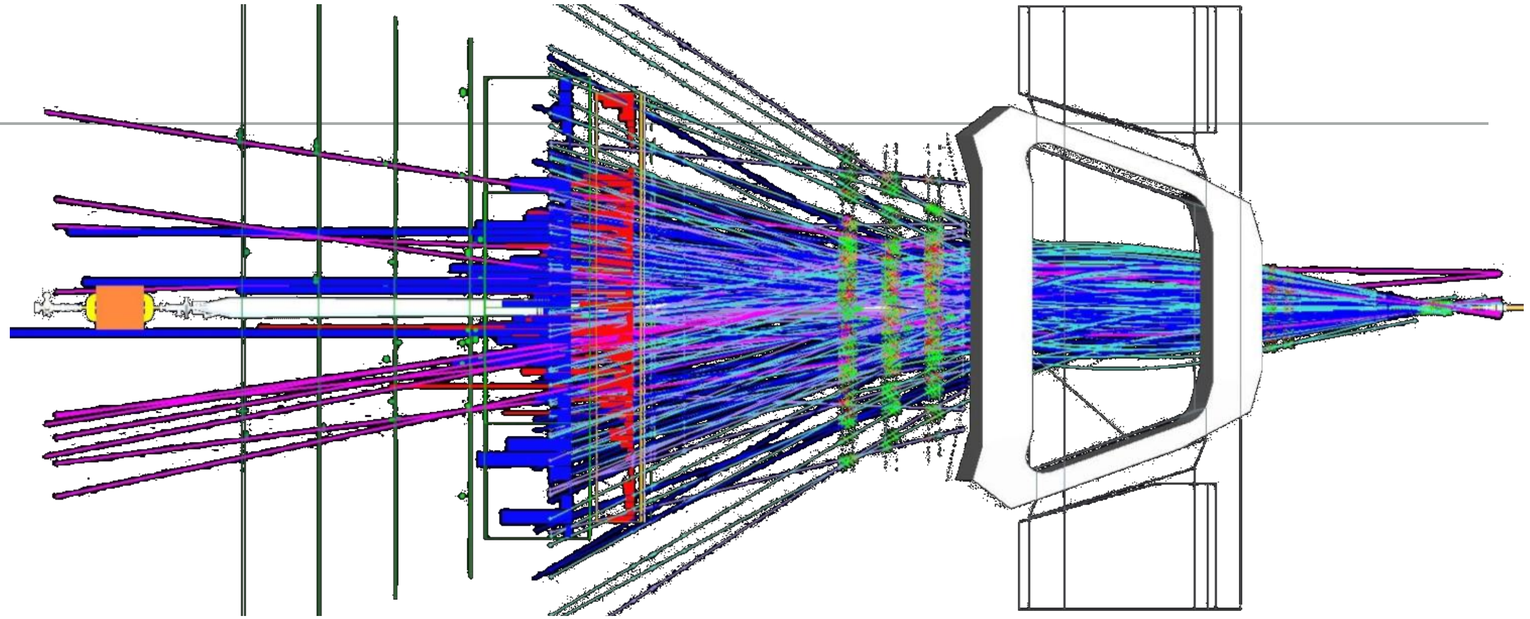
The LHCb Detector

23 sep 2010 19:49:24
Run 79646 Event 143858637

Compare trillions of matter and antimatter particle decays…
…and search for differences between matter and antimatter!

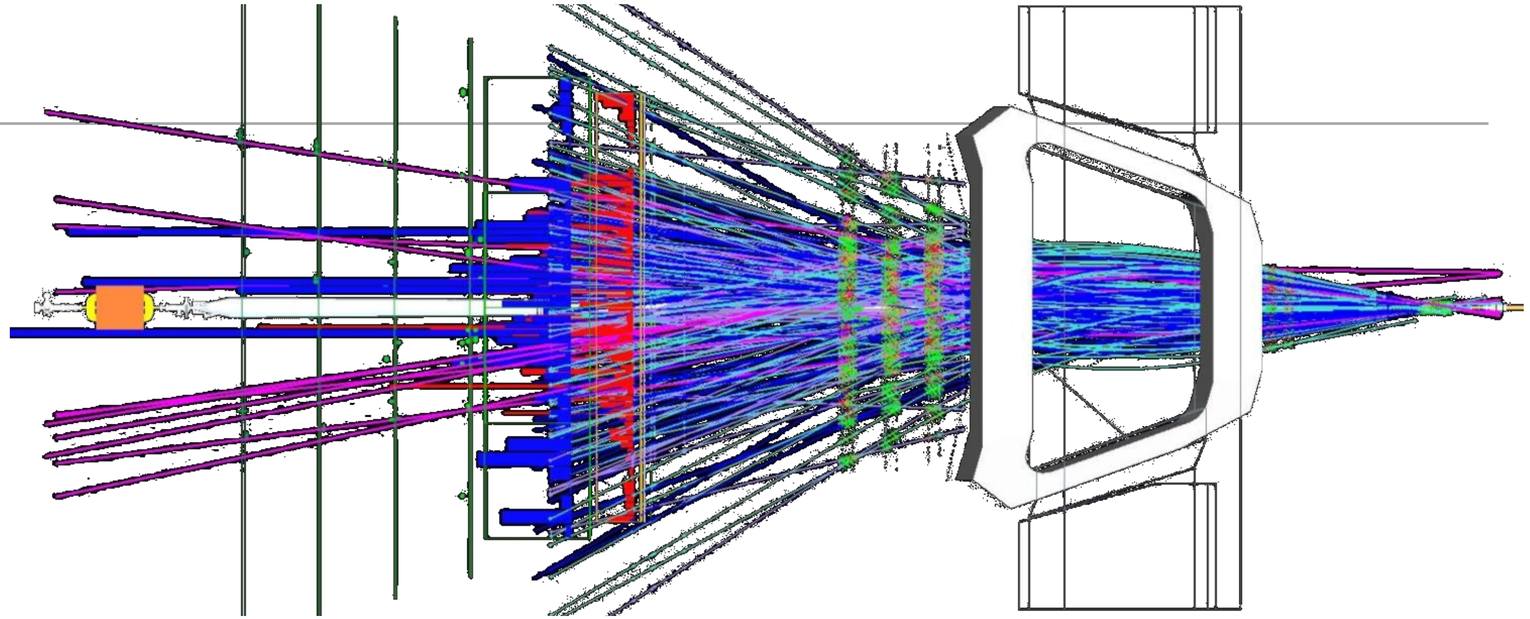
MATTER VS. ANTIMATTER

$$E = mc^2$$



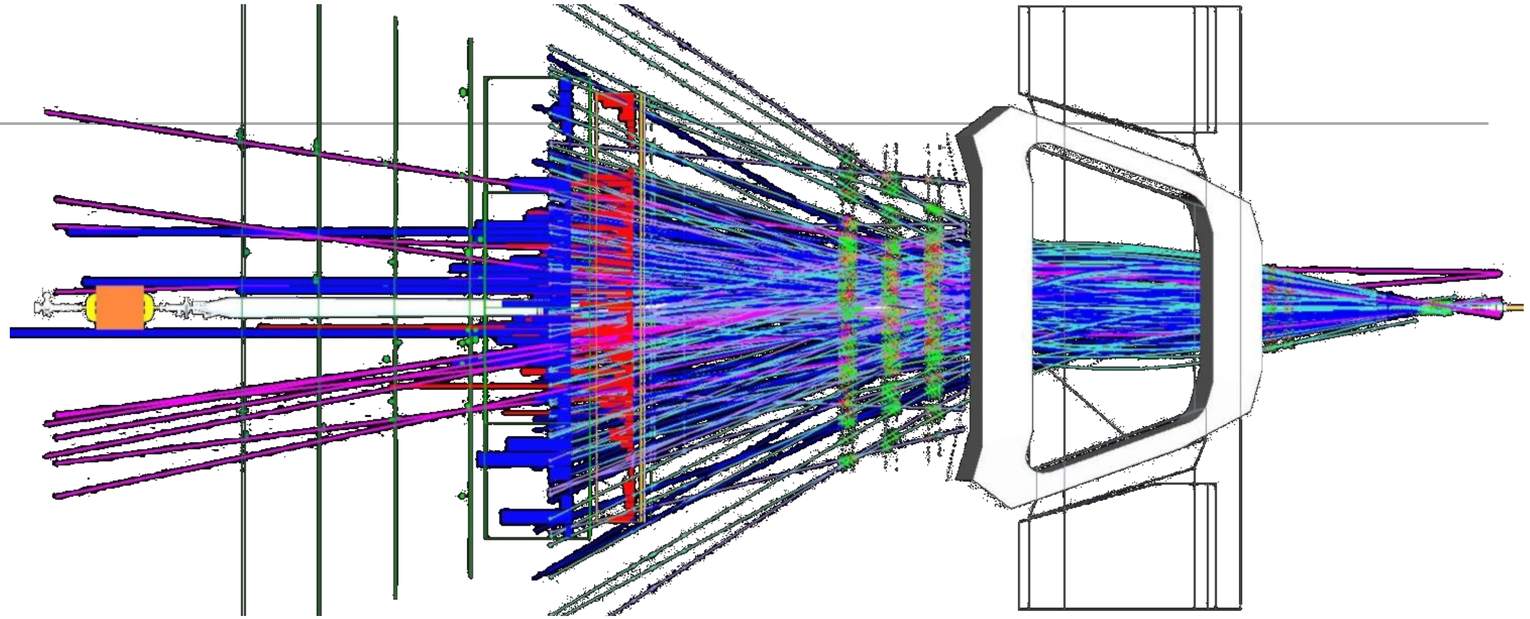
MATTER VS. ANTIMATTER

$$E = \sqrt{m^2c^4 + p^2c^2}$$



MATTER VS. ANTIMATTER

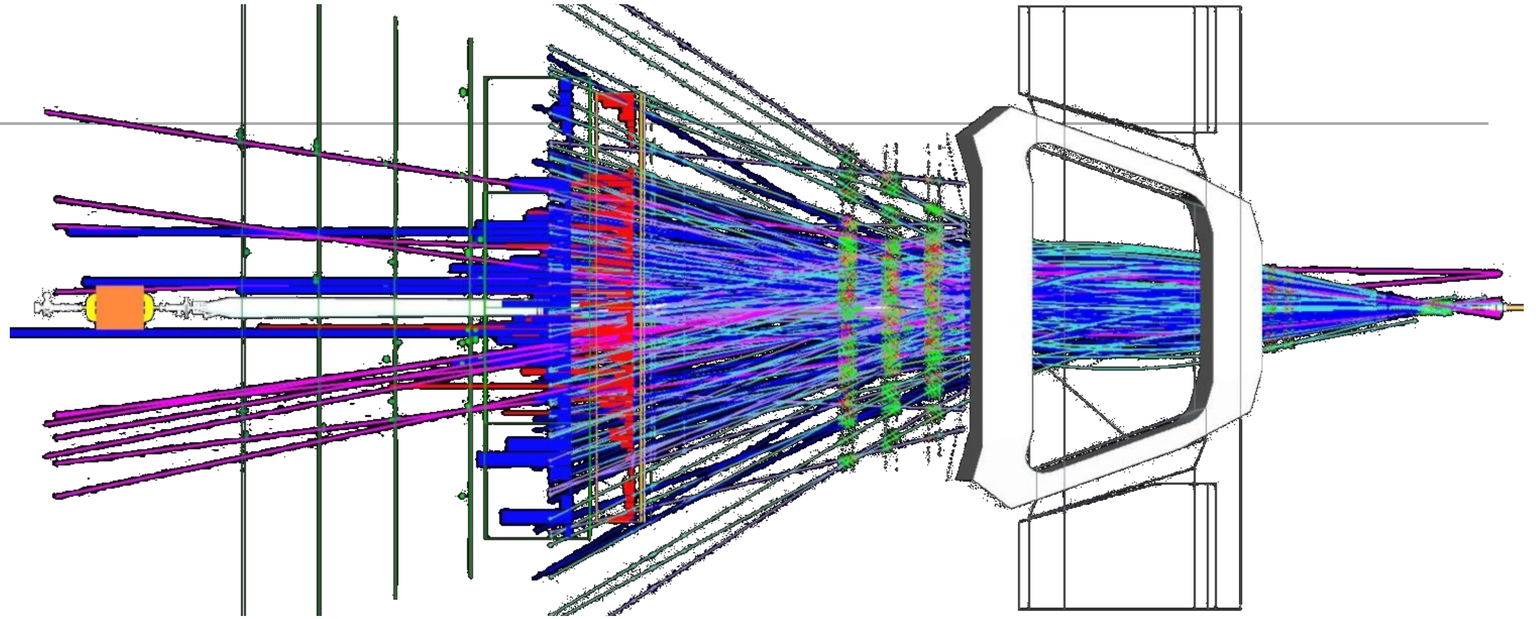
$$E = \sqrt{m^2c^4 + p^2c^2}$$



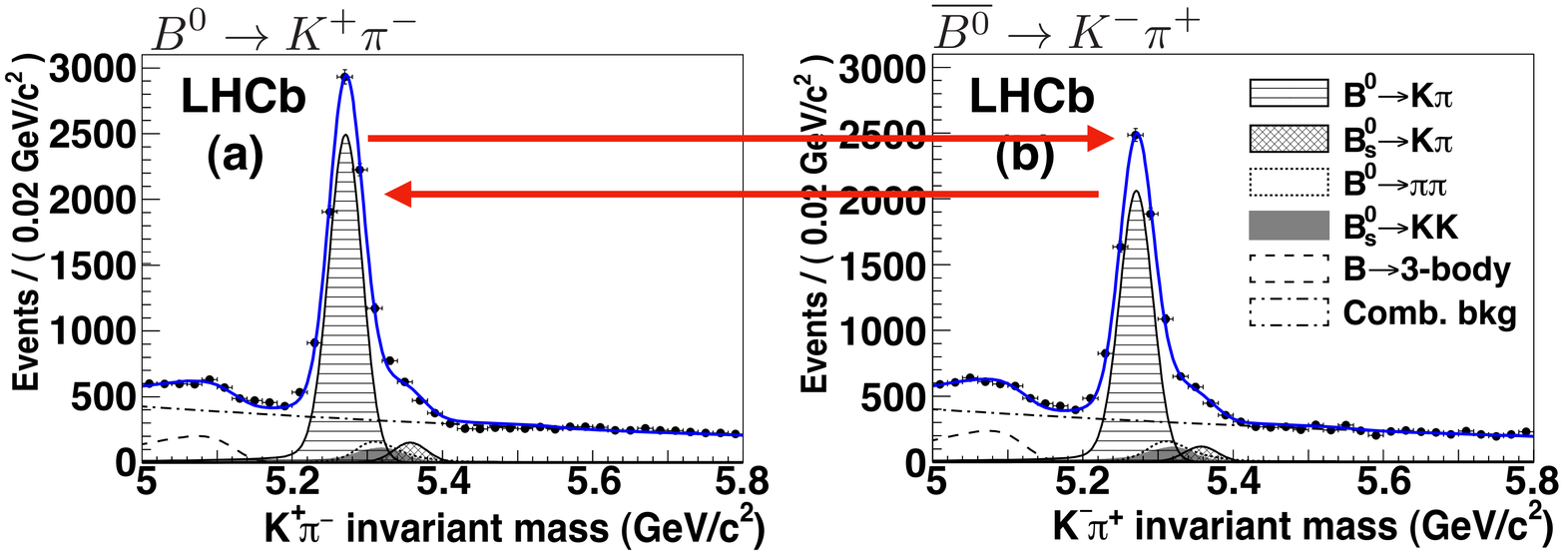
$$mc^2 = \sqrt{\sum_i E_i^2 - \sum_i p_i^2 c^2}$$

MATTER VS. ANTIMATTER

$$E = \sqrt{m^2c^4 + p^2c^2}$$



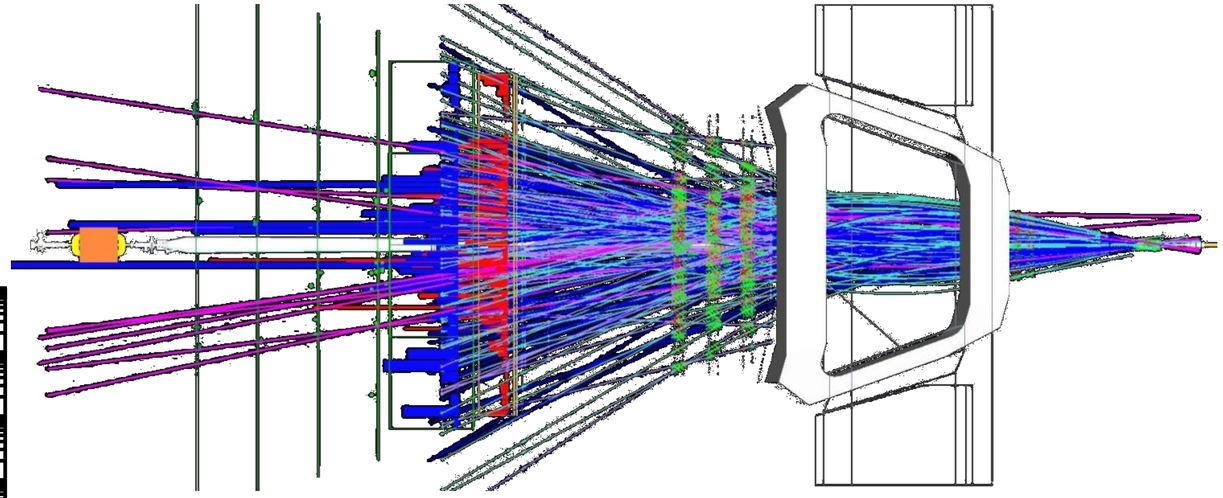
$$mc^2 = \sqrt{\sum_i E_i^2 - \sum_i p_i^2 c^2}$$

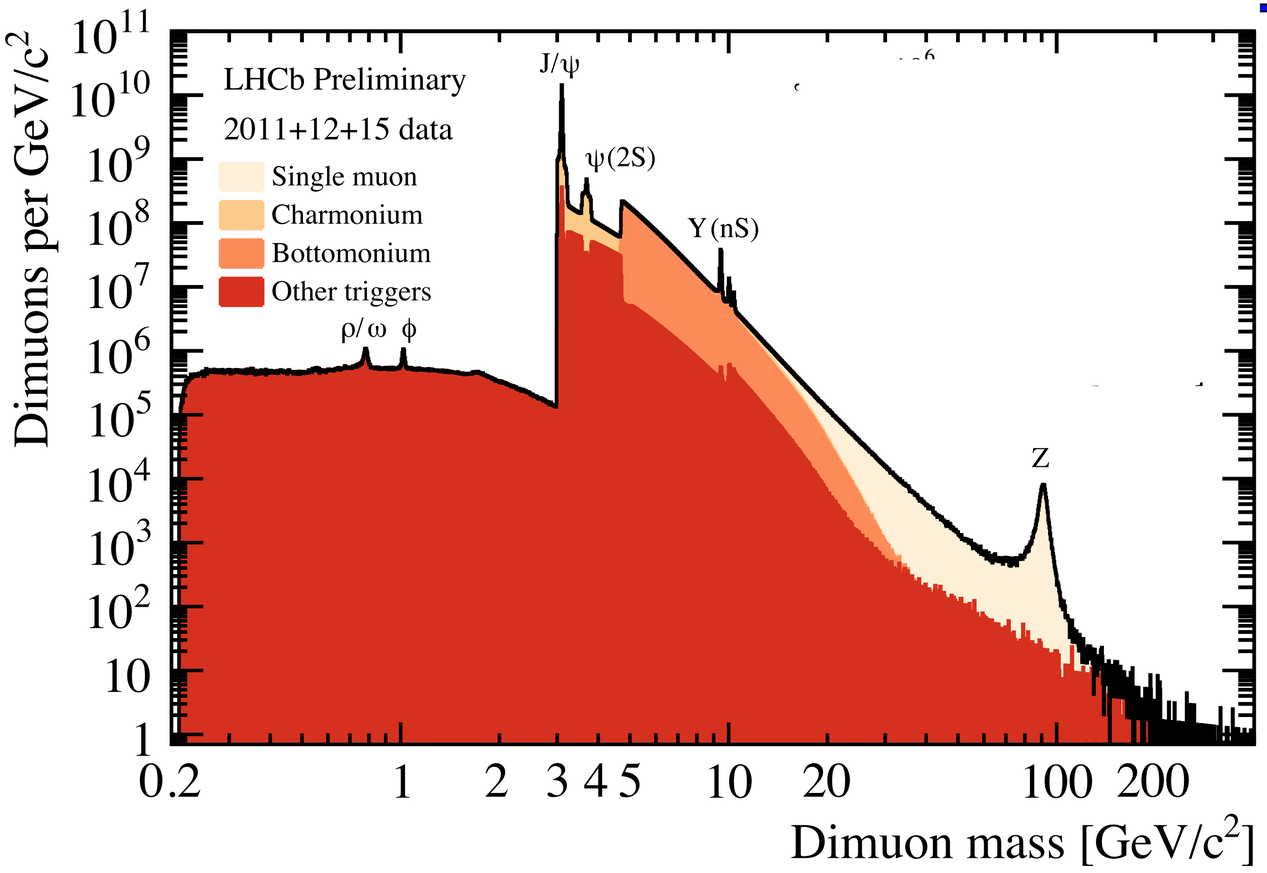


$$\frac{N_{B^0\to K^+\pi^-} - N_{\overline{B^0}\to K^-\pi^+}}{N_{B^0\to K^+\pi^-} + N_{\overline{B^0}\to K^-\pi^+}} = -0.088 \pm 0.011(stat) \pm 0.008(syst)$$

$$\Gamma(B \to K\pi)/\Gamma_{\rm tot} = (19.6 \pm 0.5) \cdot 10^{-6}$$

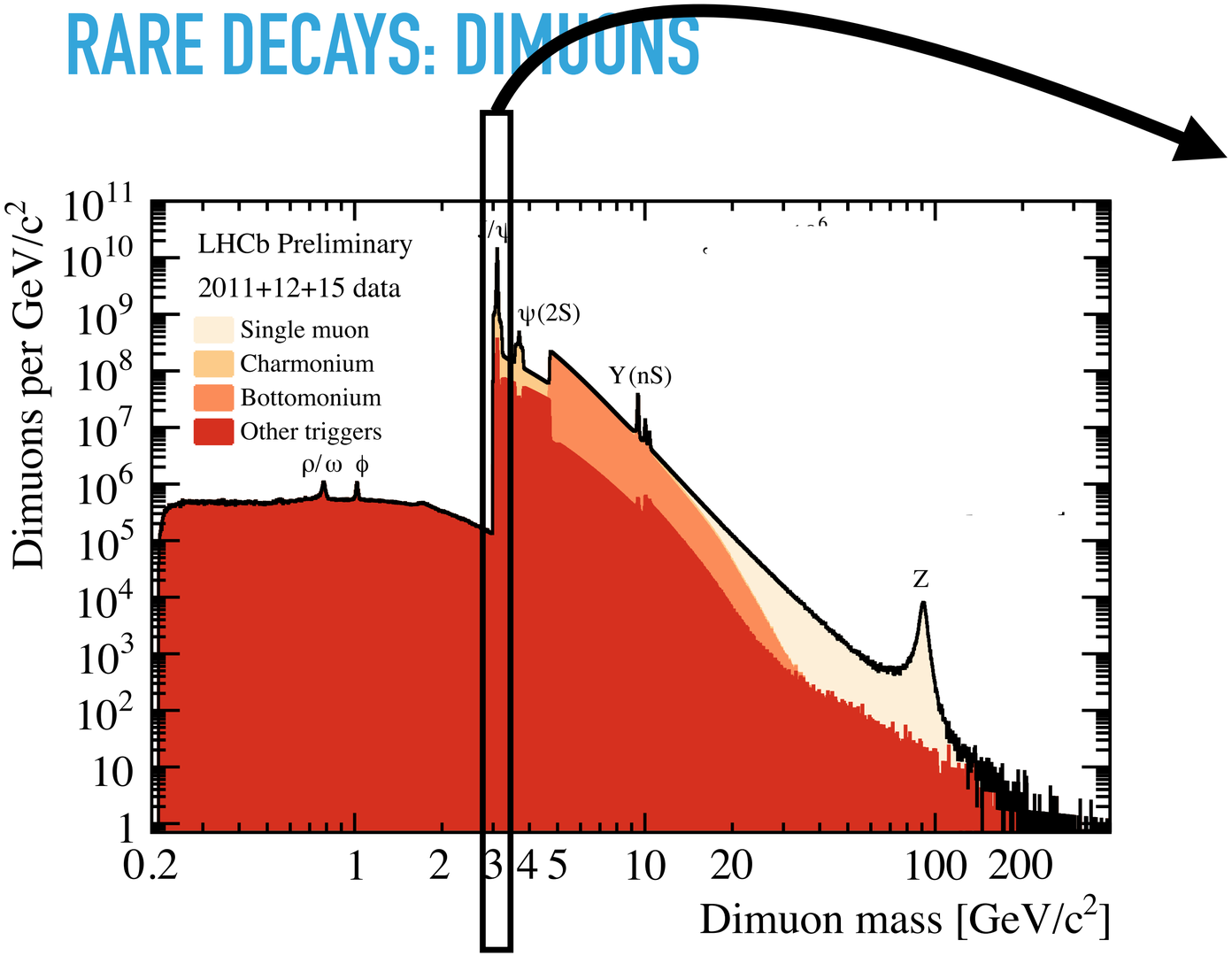
RARE DECAYS: DIMUONS

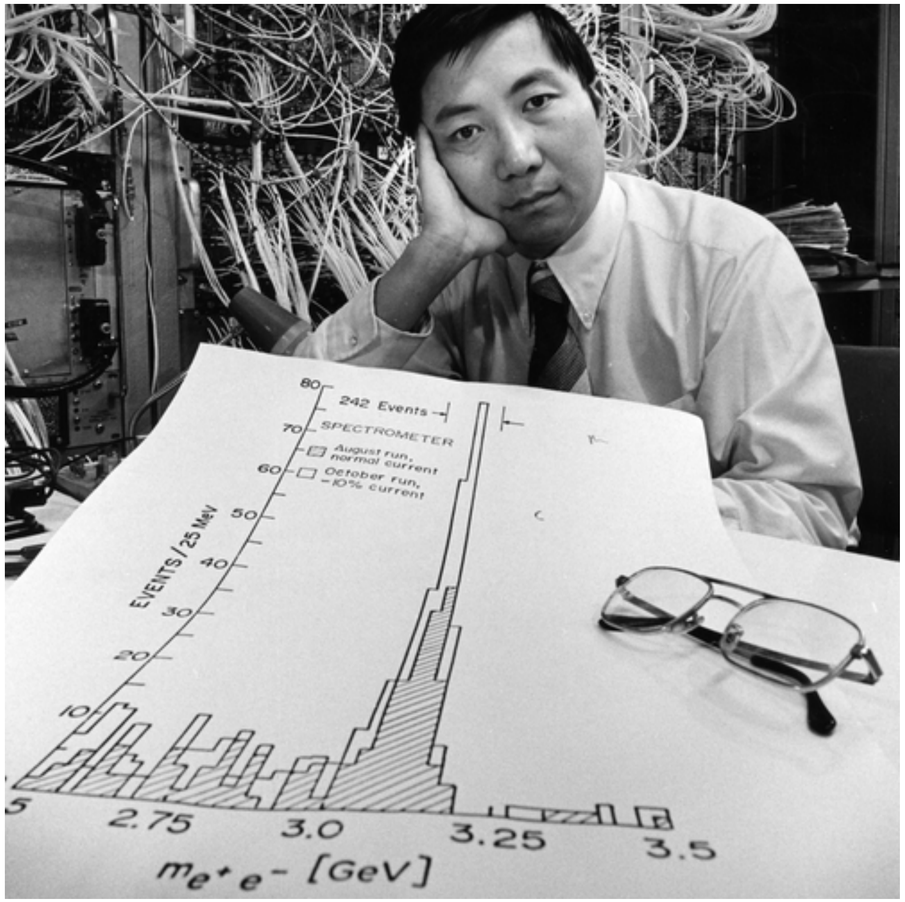




LHCb-CONF-2016-05

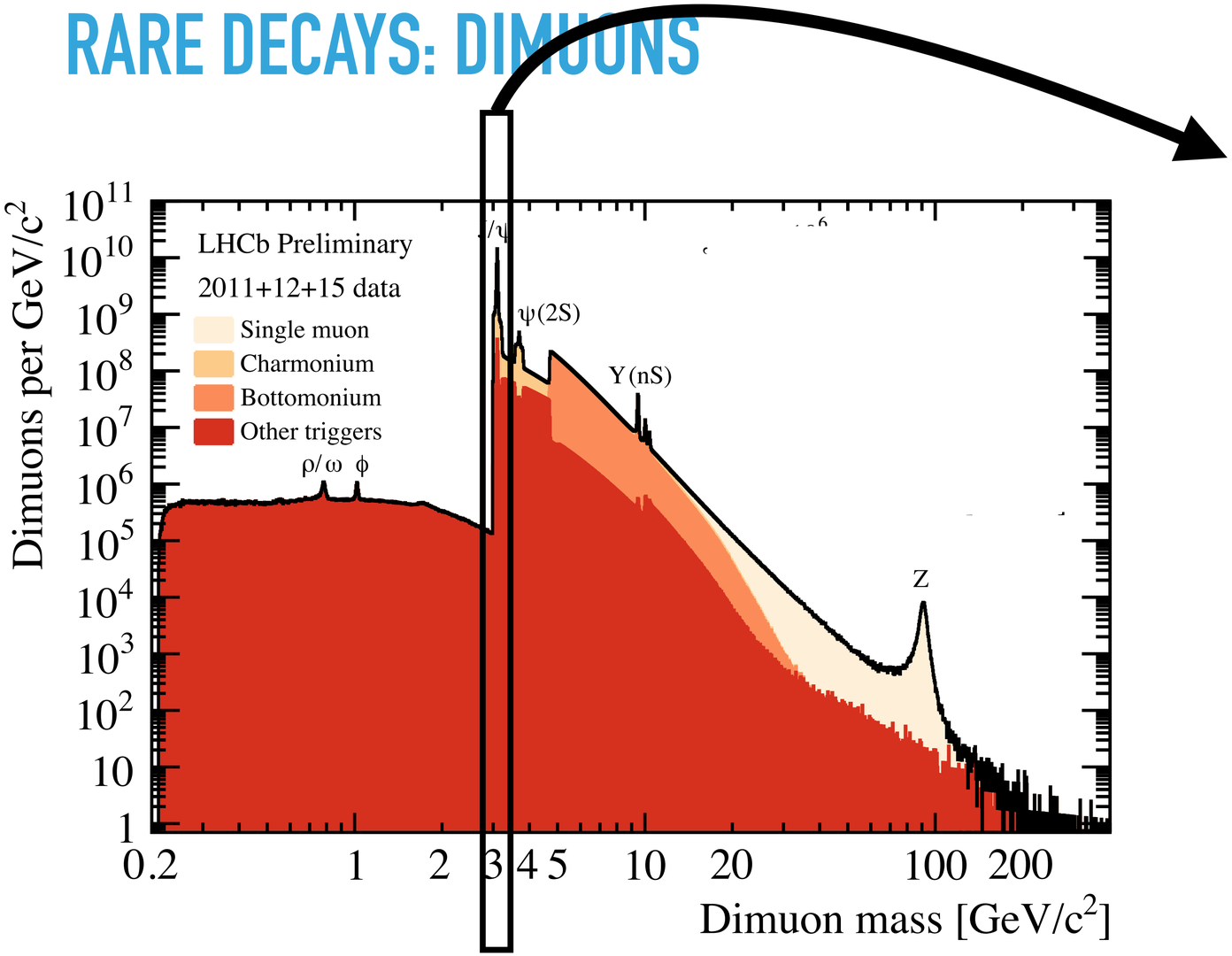
RARE DECAYS: DIMUONS



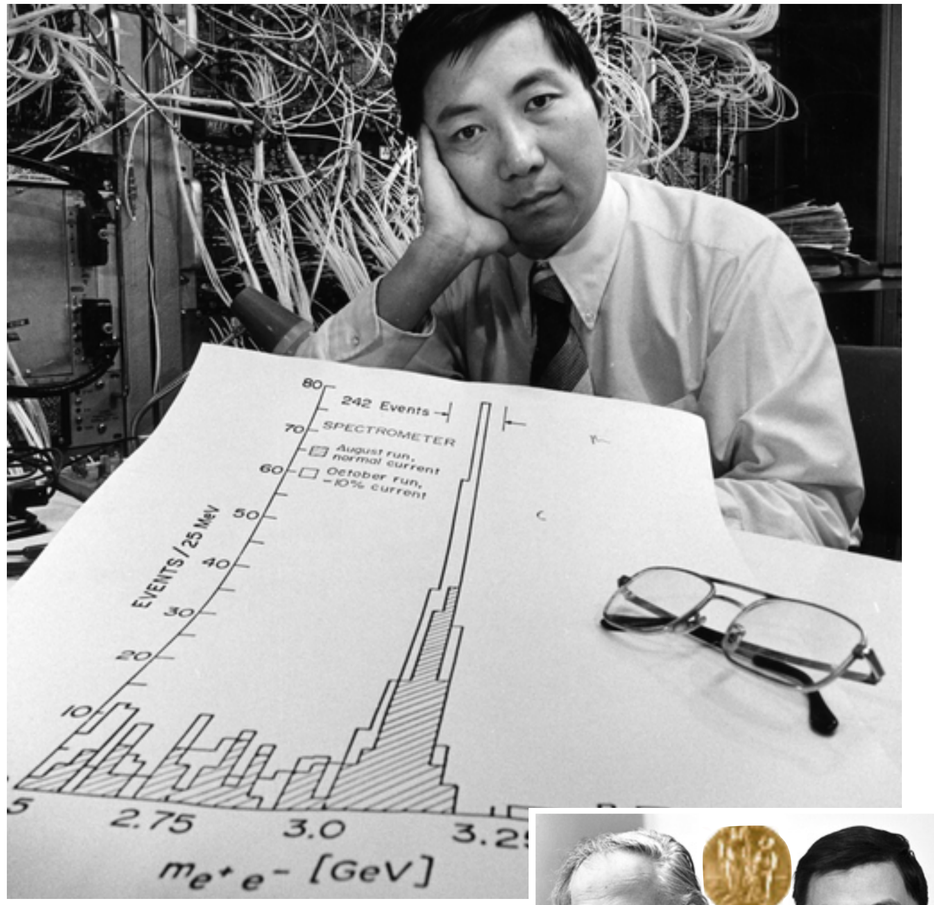


LHCb-CONF-2016-05

RARE DECAYS: DIMUONS



LHCb-CONF-2016-05

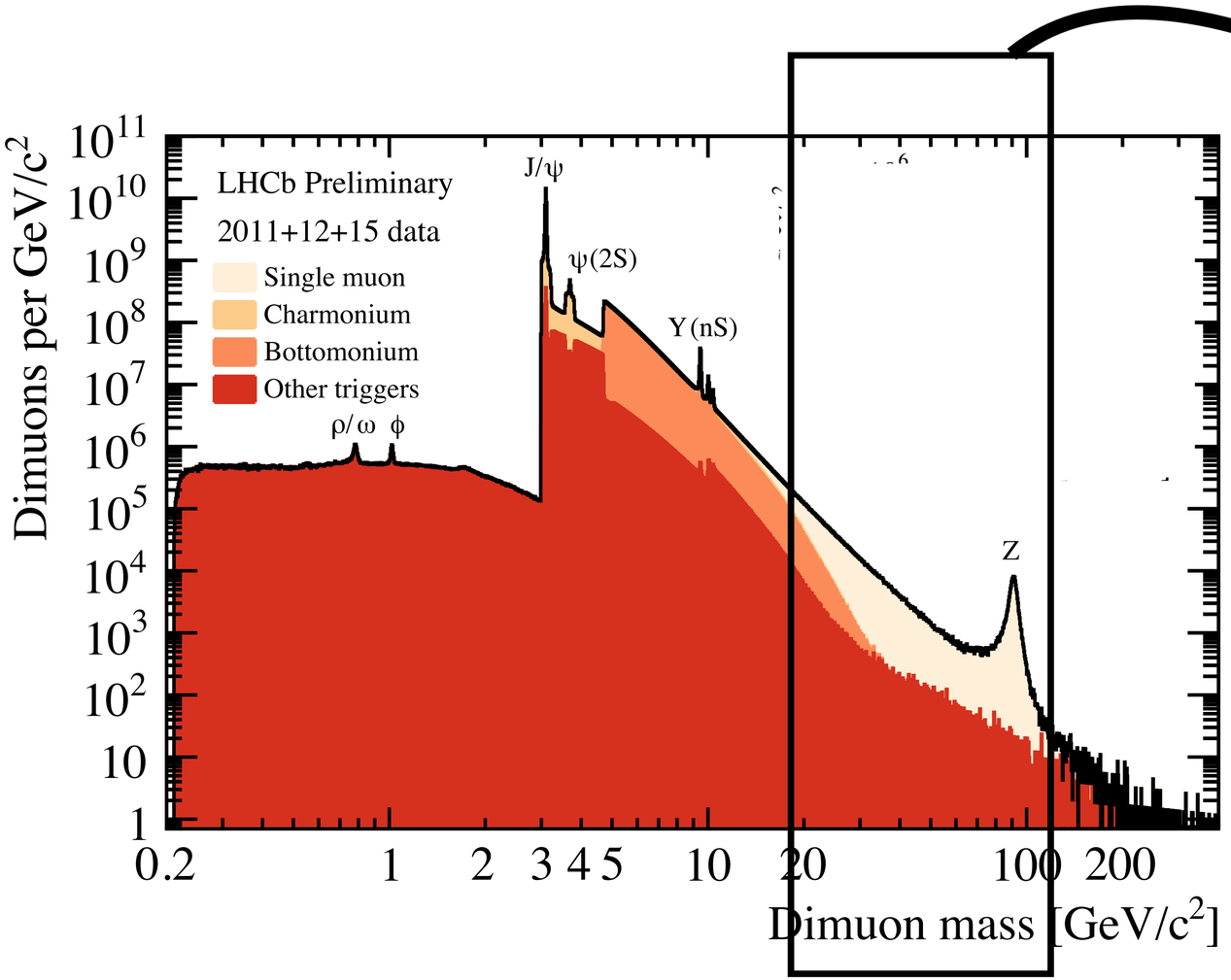


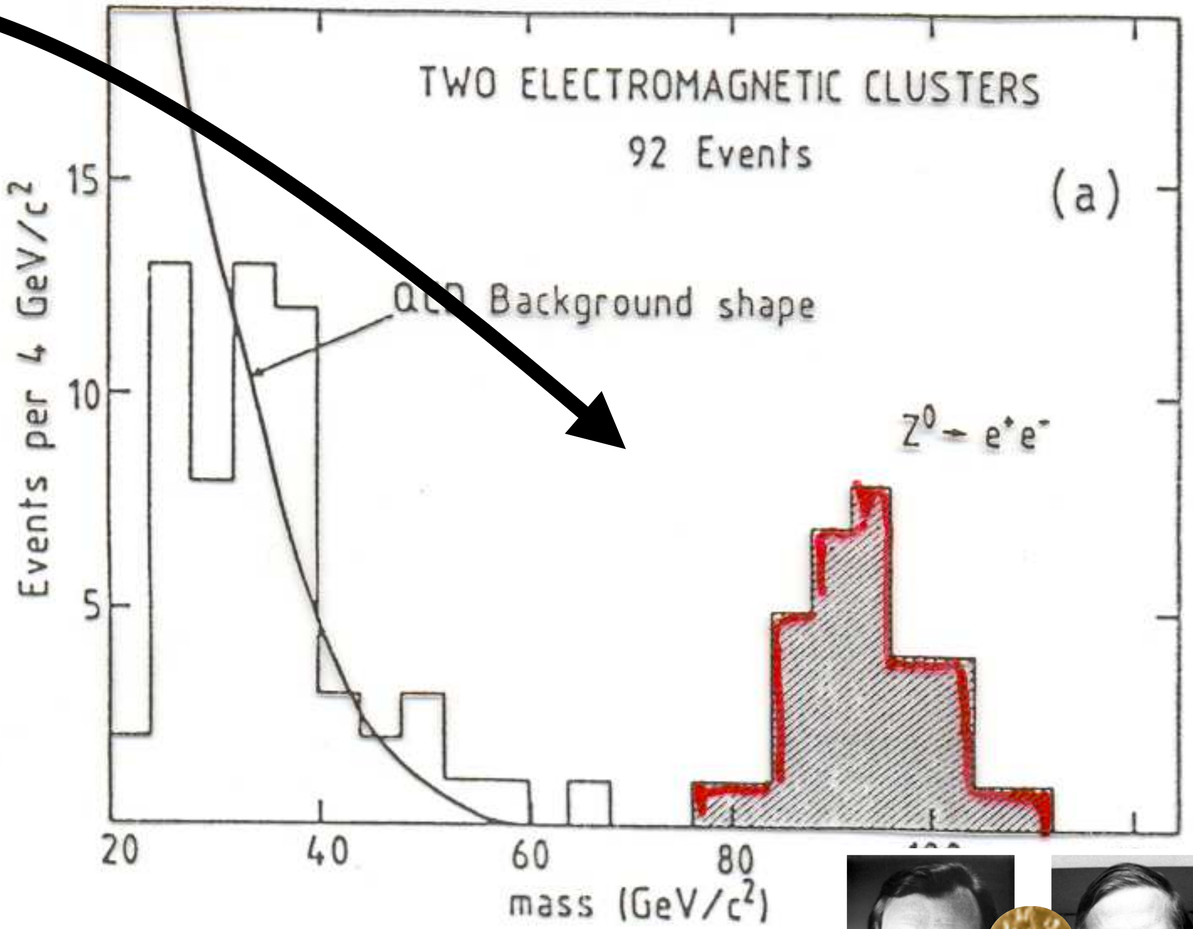
Nobel 1976

Burton Richter
Prize share: 1/2

Samuel Chao Chung Ting

DIMUONS



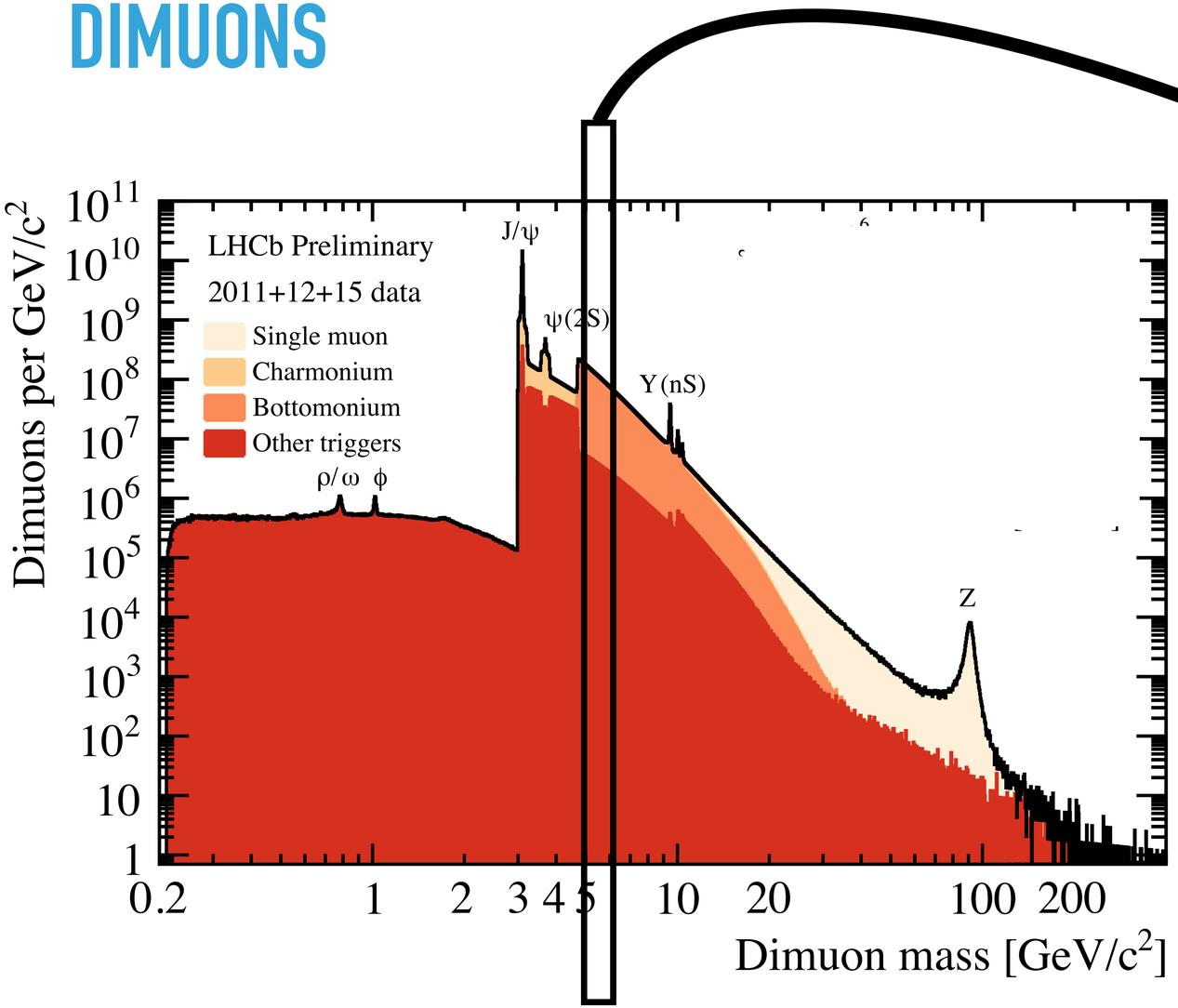


[LHCb-CONF-2016-05](#)

Nobel 1984

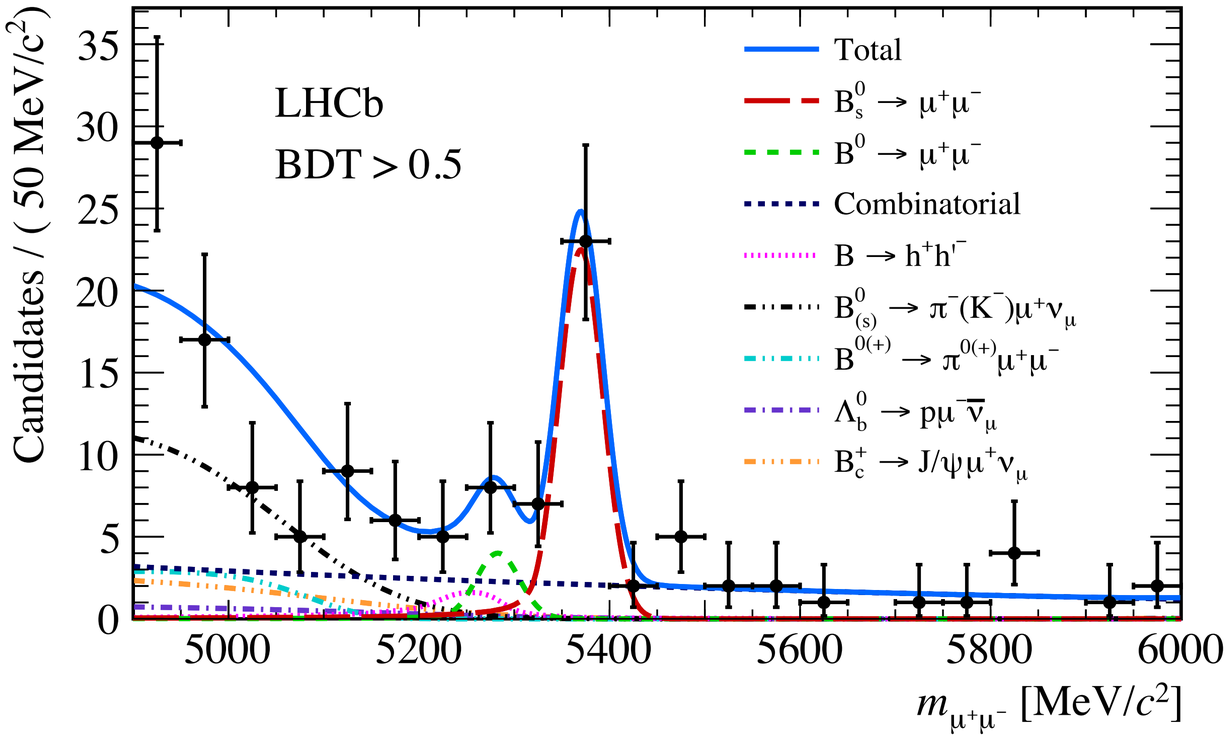
Carlo Rubbia Simon van der Meer

DIMUONS



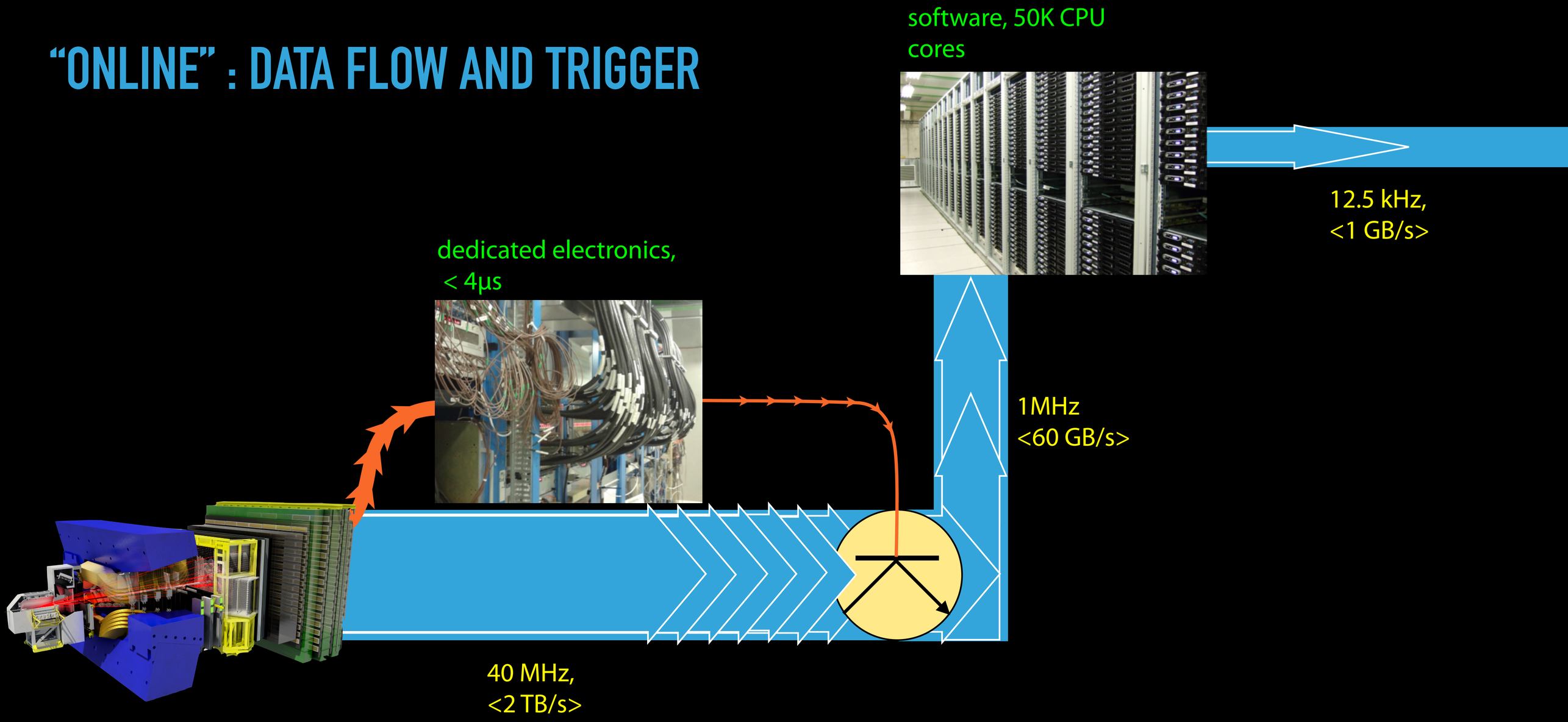
LHCb-CONF-2016-05

$Br(B_s \to \mu^+\mu^-) = (3.0 \pm 0.6\ ^{+0.3}_{-0.2}) \times 10^{-9}$



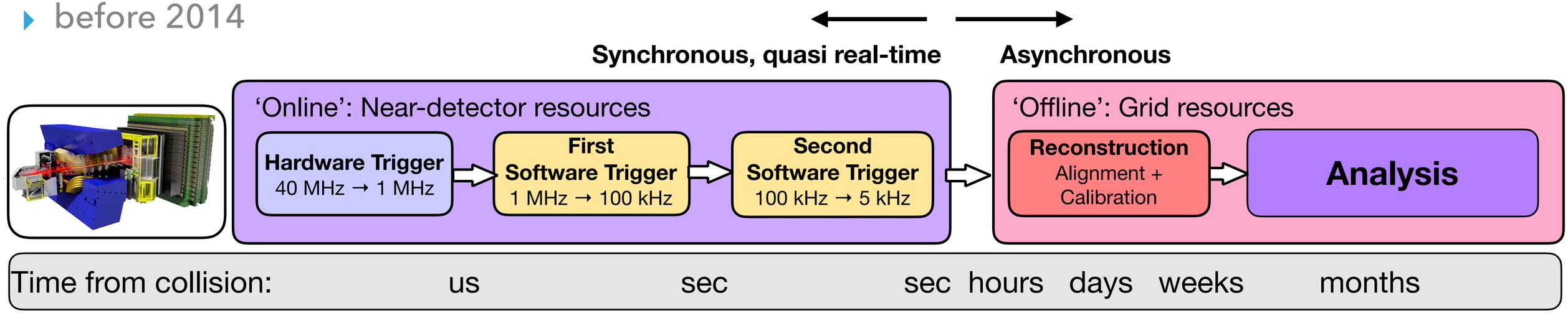
PhysRevLett.118.191801

"ONLINE" : DATA FLOW AND TRIGGER



DATA REDUCTION & DATA FLOW

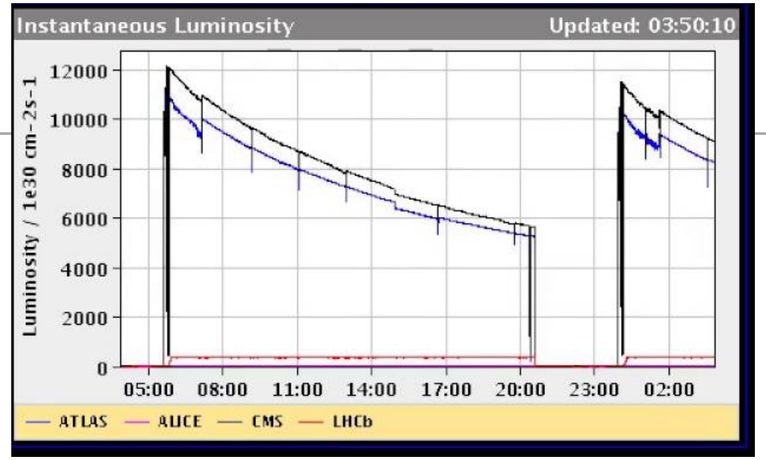
- before 2014

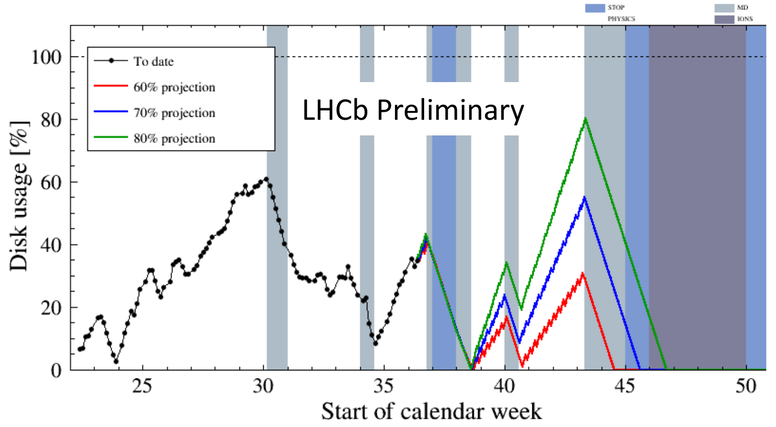


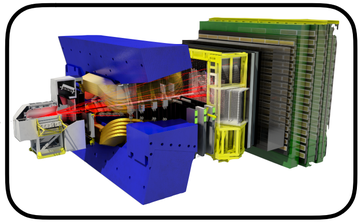
- Disadvantages:
  - time: alignment + calibration applied after data taking
  - money: uses a lot of computing resources to (re)process data
  - physics: imperfect reconstruction in trigger = loss of recorded signal

ALLOW FOR LATENCY!

- LHC does not always produce collisions
  (2012: 37%, 2016: 60%)
- 50% uptime → CPU capacity x 2
  (assuming sufficiently large buffer ;-)
- Install 10 PB buffer







Synchronous, quasi real-time ← → **Asynchronous**

‘Online’: Near-detector resources

Hardware Trigger
40 MHz → 1 MHz

First Software Trigger
1 MHz → 100 kHz

Second Software Trigger
100 kHz → 5 kHz

‘Offline’: Grid resources

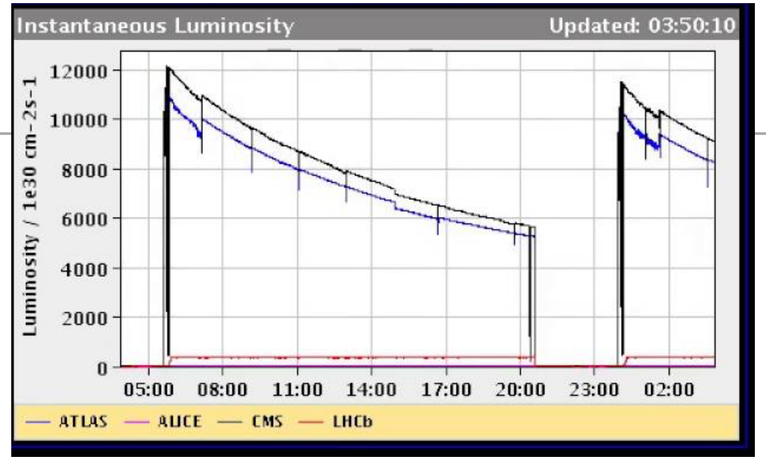
Reconstruction
Alignment + Calibration

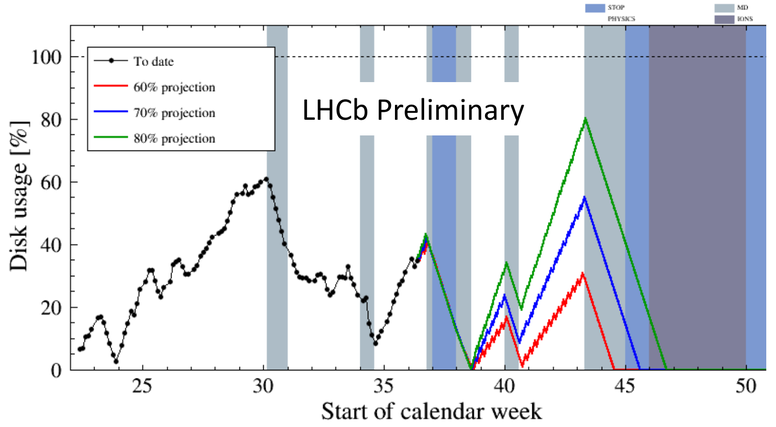
Analysis

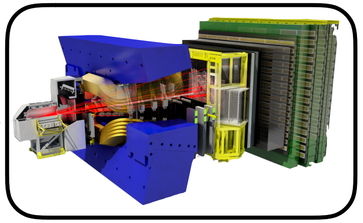
Time from collision: us seconds hours days weeks months

ALLOW FOR LATENCY!

- LHC does not always produce collisions
  (2012: 37%, 2016: 60%)
- 50% uptime → CPU capacity x 2
  (assuming sufficiently large buffer ;-)
- Install 10 PB buffer







Synchronous, quasi real-time ← → **Asynchronous**

‘Online’: Near-detector resources

Hardware Trigger
40 MHz → 1 MHz

First Software Trigger
1 MHz → 100 kHz

Second Software Trigger
100 kHz → 10 kHz

‘Offline’: Grid resources

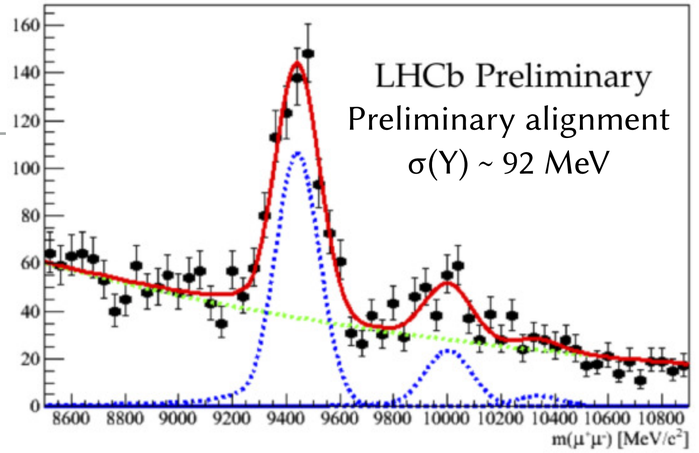
Reconstruction
Alignment + Calibration

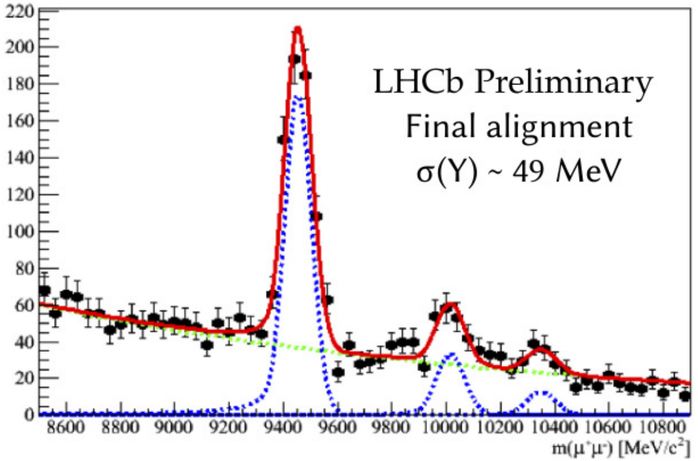
Analysis

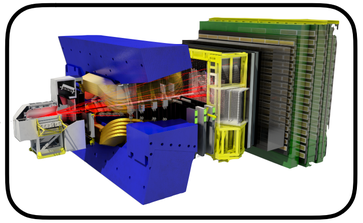
Time from collision: us seconds hours days weeks months

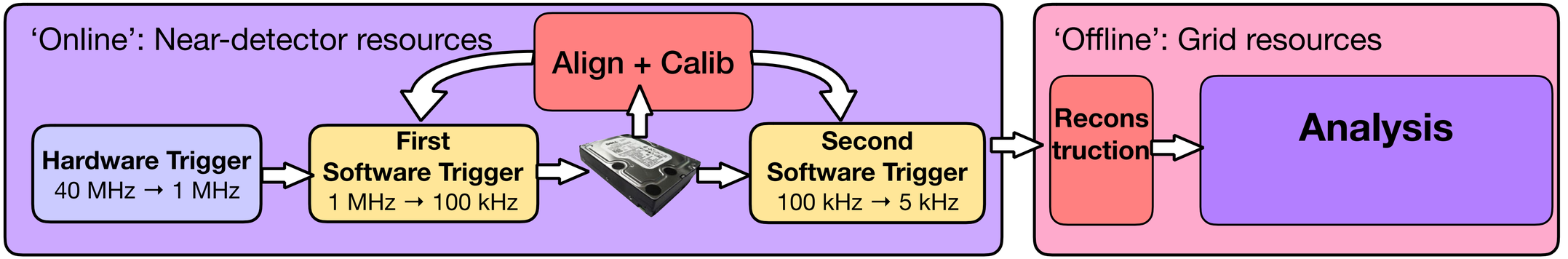
"REAL TIME" CALIBRATION

- Use the introduced delay to perform calibrations
- Software trigger has best possible calibrations available





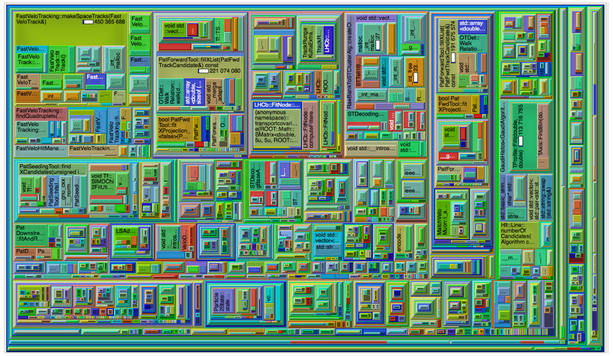




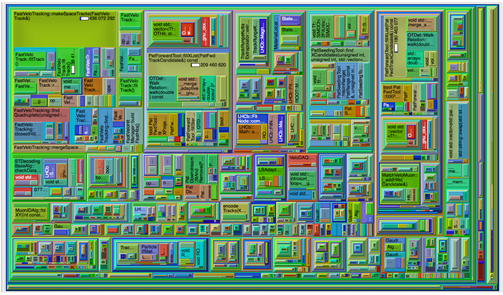
Time from collision: us seconds hours days weeks months

UNIFY ONLINE/OFFLINE RECONSTRUCTION

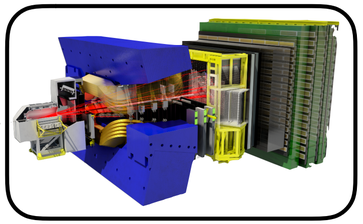
- Optimize offline code so it fits in ‘online’ budget
- *Very* inhomogeous workload: no single hotspot / magic bullet
- Improvements in hundreds of places

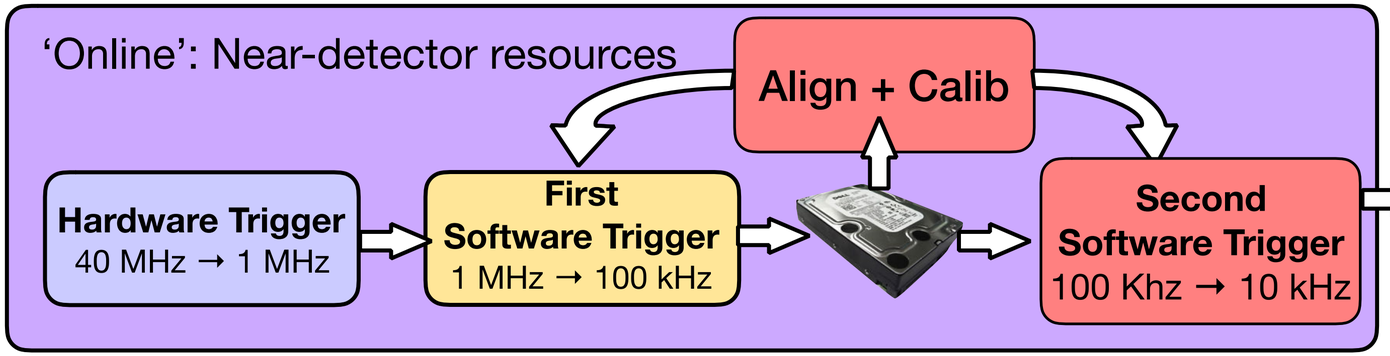


39.9M Cycles/event



27.9M Cycles/event

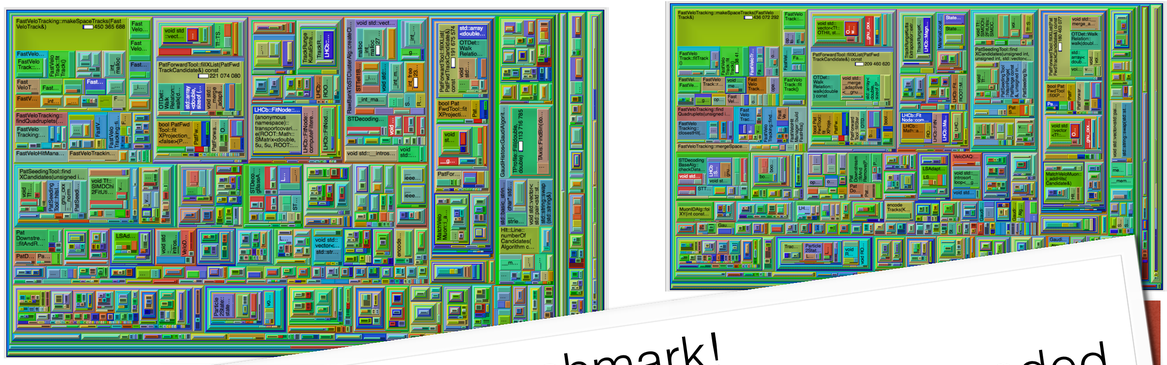




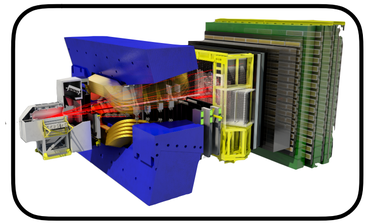
Time from collision: us seconds hours days weeks months

UNIFY ONLINE/OFFLINE RECONSTRUCTION

- Optimize offline code so it fits in ‘online’ budget
- *Very* inhomogeneous workload: no single hotspot / magic bullet
- Improvements in hundreds of places



1. Measure & Benchmark!
2. Don’t do more work than strictly needed
3. Improve memory usage
4. Vectorization — utilize SIMD



‘Online’: Near-detector resources

Hardware Trigger
40 MHz → 1 MHz

First Software Trigger
1 MHz → 100 kHz

Align + Calib

Second Software Trigger
100 Khz → 10 kHz

‘Offline’: Grid resources

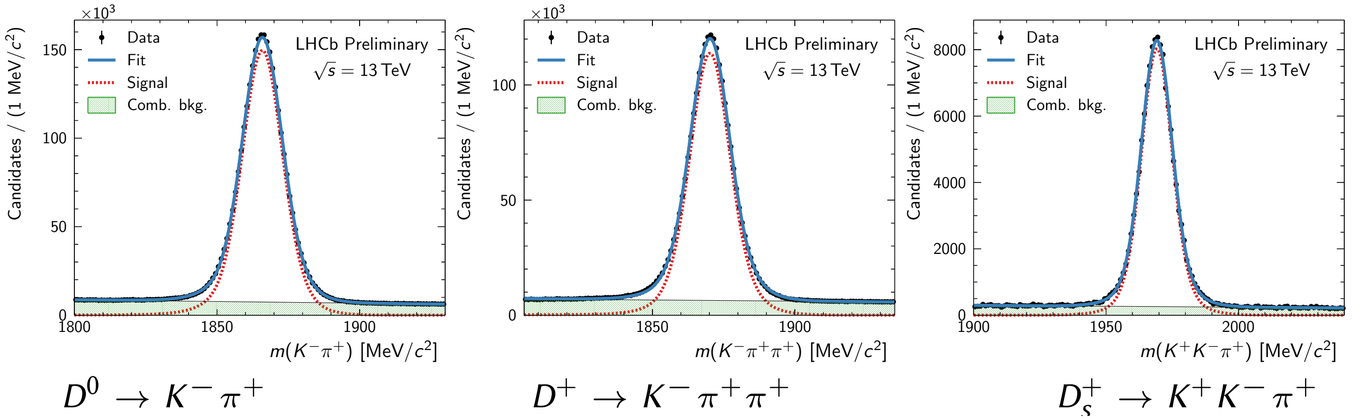
Recons truction

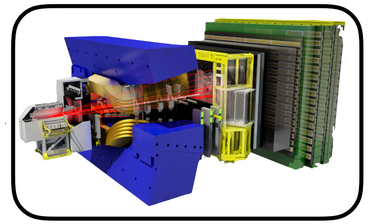
Analysis

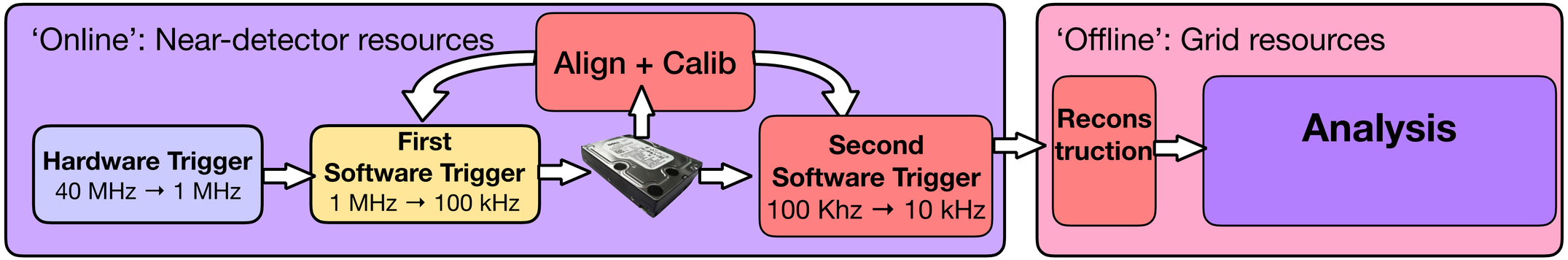
Time from collision: us seconds hours days weeks months

UNIFY ONLINE/OFFLINE RECONSTRUCTION

- For “high-rate” analysis, trigger can now reconstructs “offline quality” data in (quasi) real time!



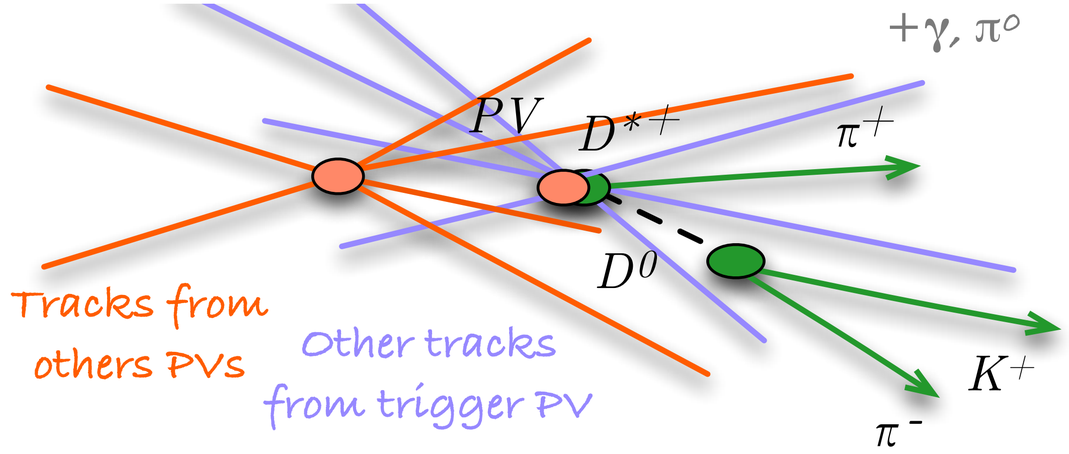


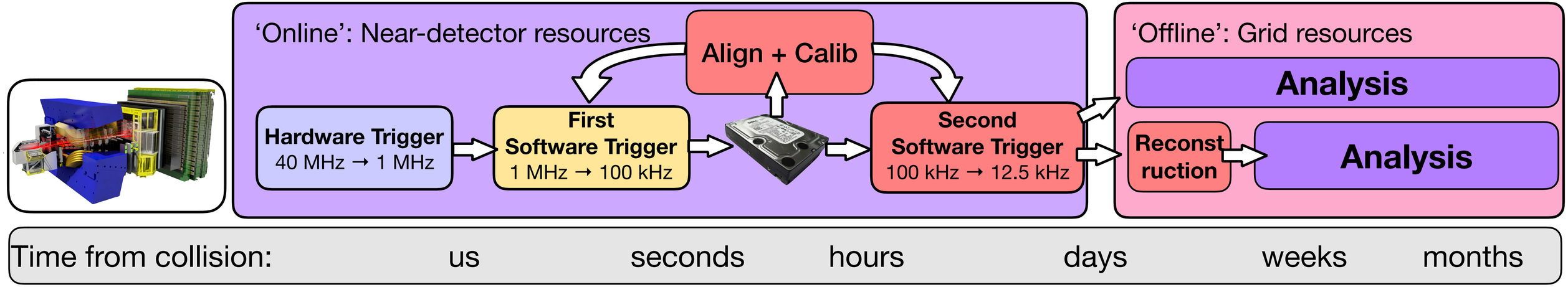


Time from collision: us seconds hours days weeks months

WRITE OUT LESS INFORMATION!

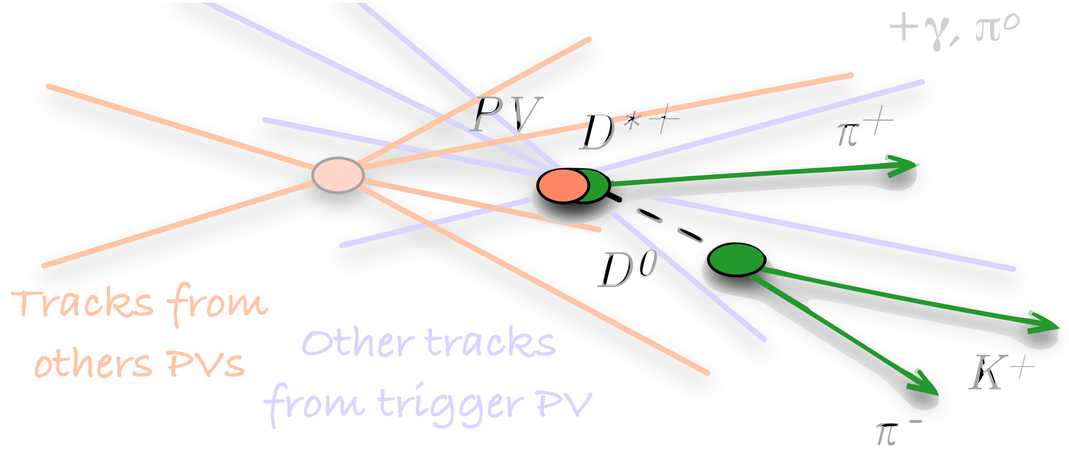
- For selected measurements, only write trigger-reconstructed signal data, *instead of* the sensor data
- For the same bandwidth, can allow more physics

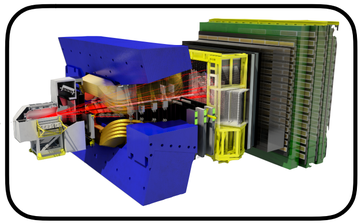


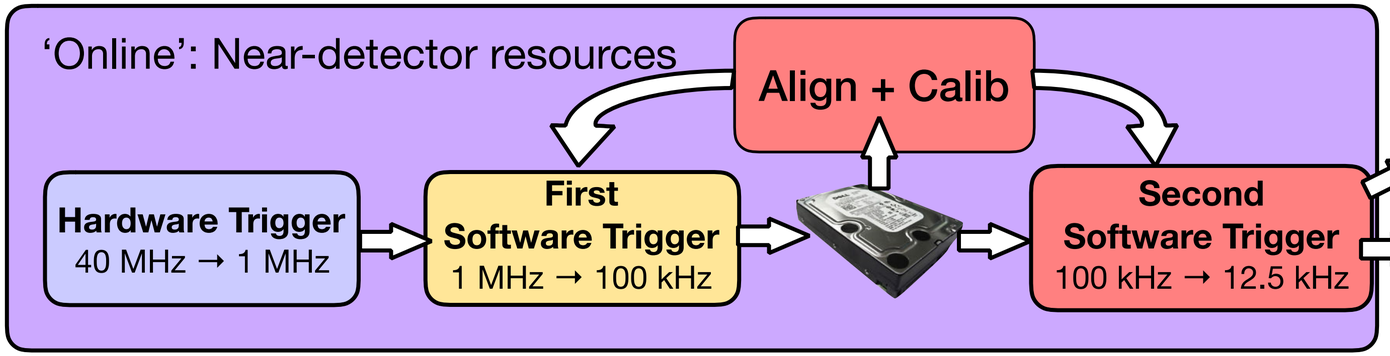


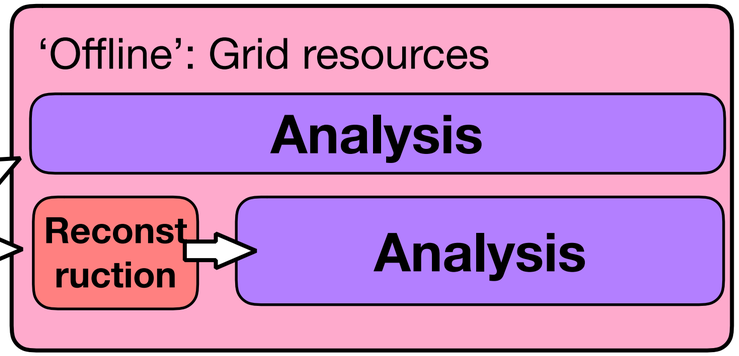
WRITE OUT LESS INFORMATION!

- For selected measurements, only write trigger-reconstructed signal data, *instead of* the sensor data
- For the same bandwidth, can allow more physics





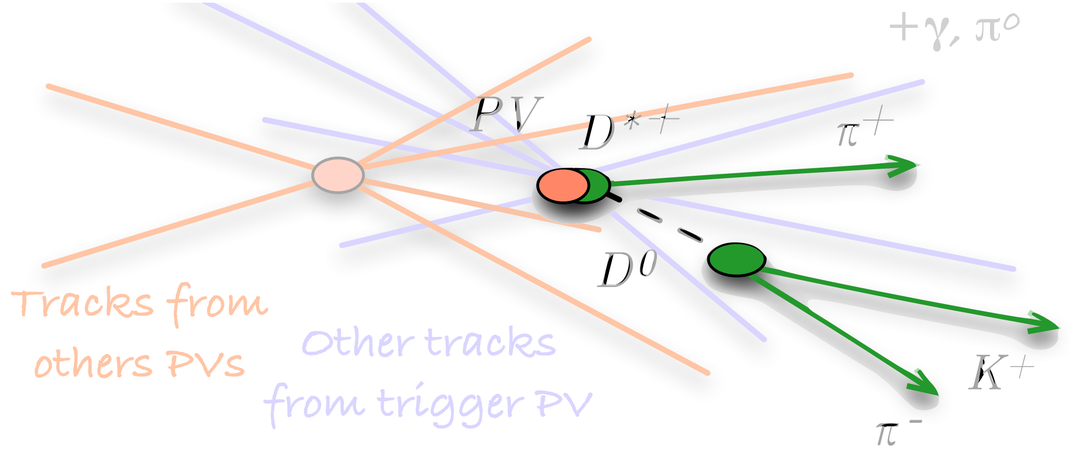




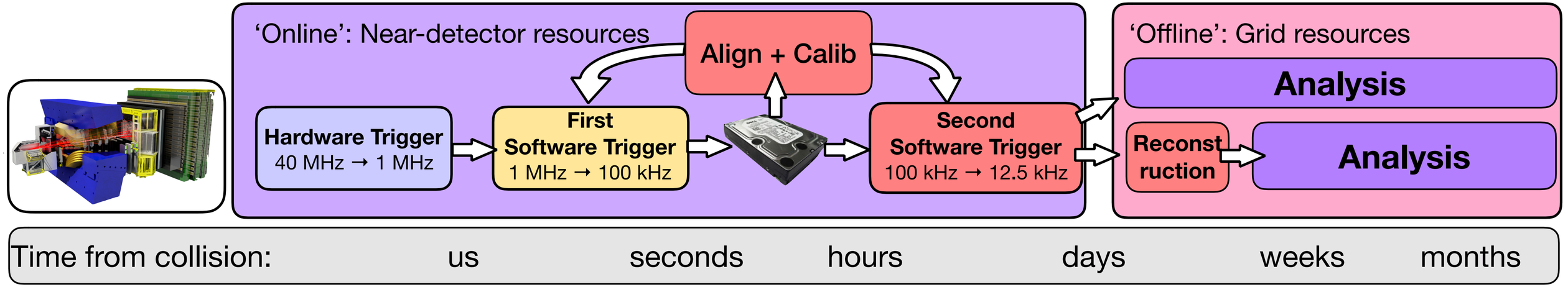
Time from collision: us seconds hours days weeks months

WRITE OUT LESS INFORMATION!

- For selected measurements, only write trigger-reconstructed signal data, *instead of* the sensor data
- For the same bandwidth, can allow more physics

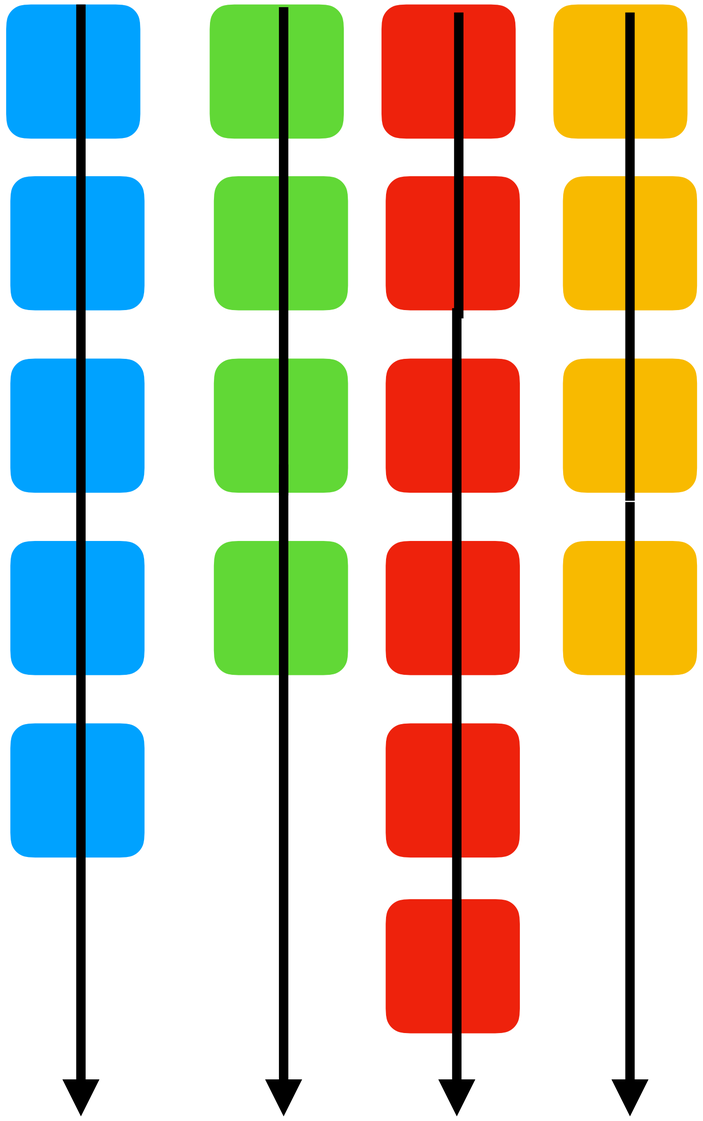


5 kB / event @ 2.5 kHz = 12.5 MB/s
70 kB / event @ 10 kHz = 700 MB/s



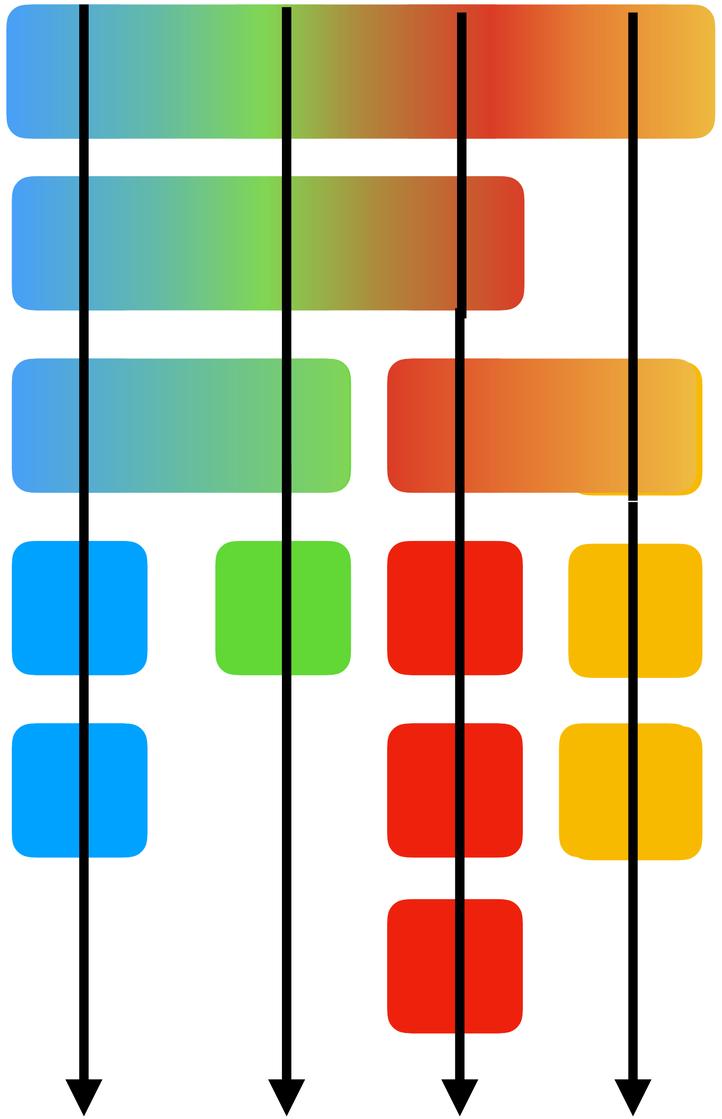
TRIGGER DATA PROCESSING

- Individual collisions are scheduled ‘round robin’ on single-threaded processes – approx. one per core – with *static* data/control flow
- Hlt1: O(50) decisions, Hlt2: O(500) decisions
- O(100) “algorithms” per decision
- Accept collision at each level if one (or more) positive decisions
- *All* decisions processed until ‘abort’ or ‘accept’ – no ‘early accept’!
- Each individual decision is based on different criteria, with (some) overlaps – but ‘logically independent’



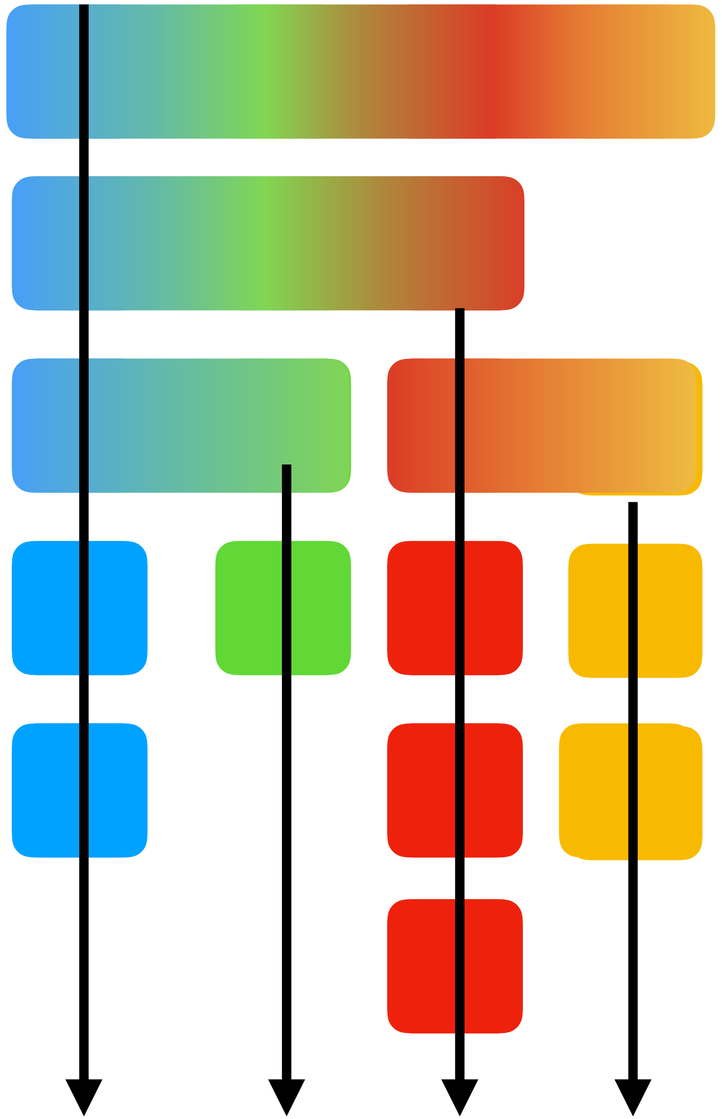
TRIGGER DATA PROCESSING

- Individual collisions are scheduled ‘round robin’ on single-threaded processes – approx. one per core – with *static* data/control flow
- Hlt1: O(50) decisions, Hlt2: O(500) decisions
- O(100) “algorithms” per decision
- Accept collision at each level if one (or more) positive decisions
- *All* decisions processed until ‘abort’ or ‘accept’ – no ‘early accept’!
- Each individual decision is based on different criteria, with (some) overlaps – but ‘logically independent’



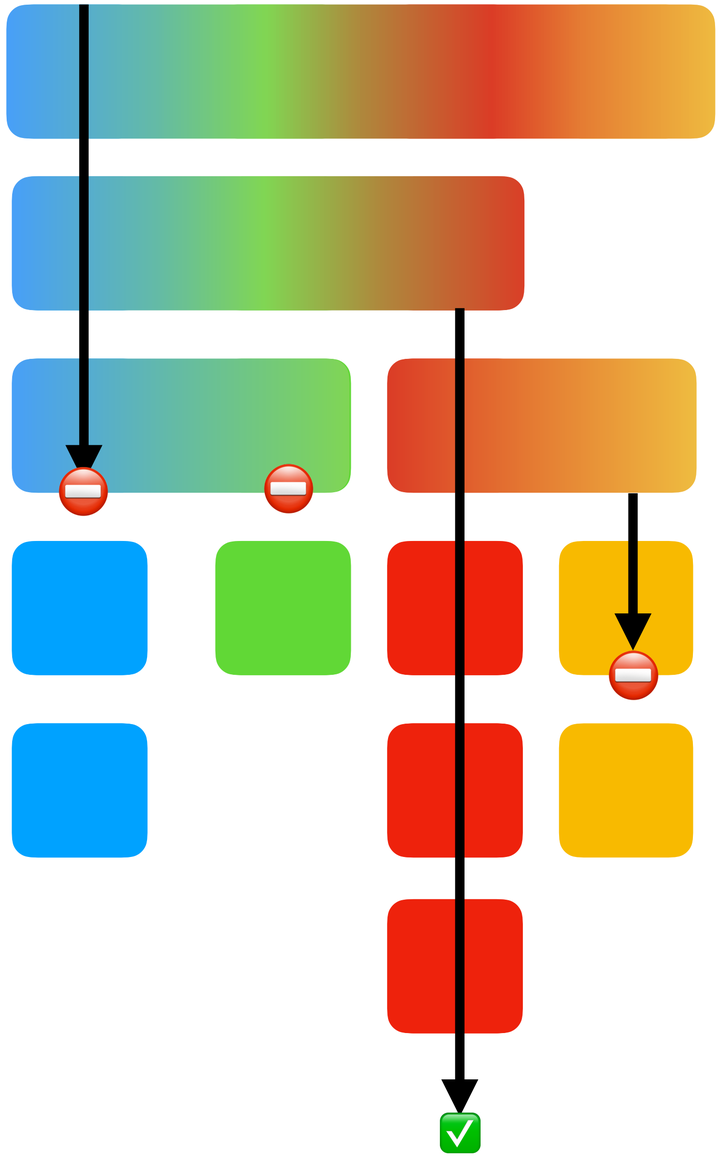
TRIGGER DATA PROCESSING

- Individual collisions are scheduled ‘round robin’ on single-threaded processes – approx. one per core – with *static* data/control flow
- Hlt1: O(50) decisions, Hlt2: O(500) decisions
- O(100) “algorithms” per decision
- Accept collision at each level if one (or more) positive decisions
- *All* decisions processed until ‘abort’ or ‘accept’ – no ‘early accept’!
- Each individual decision is based on different criteria, with (some) overlaps – but ‘logically independent’



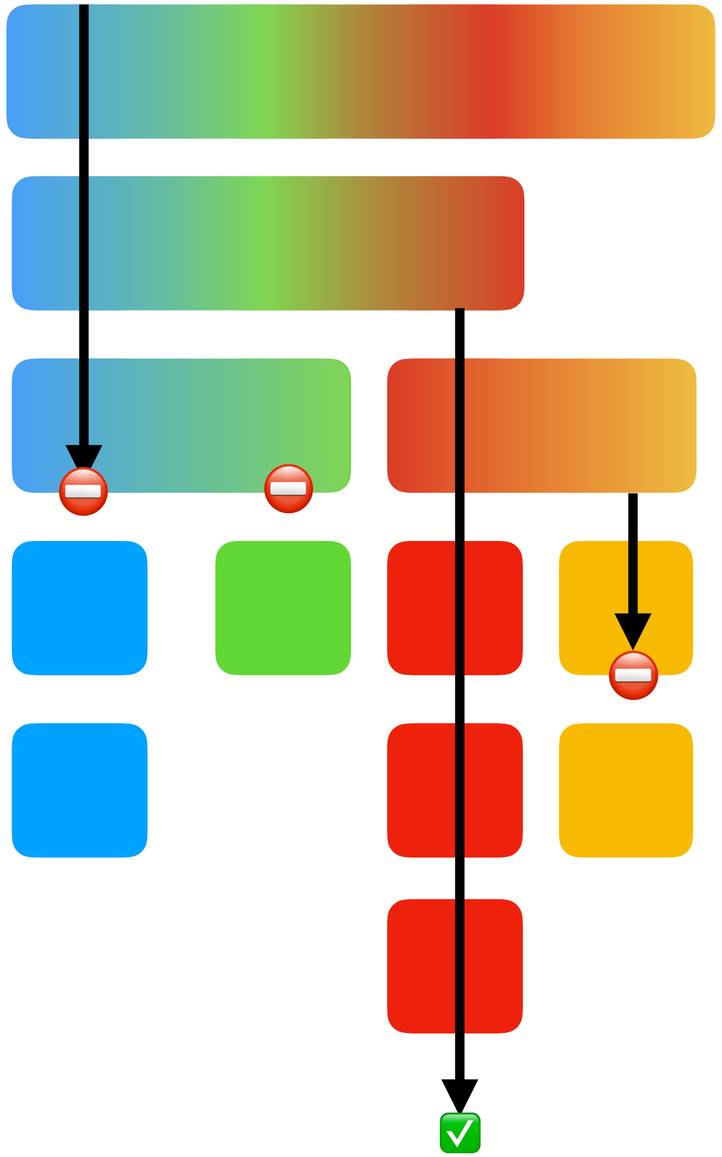
TRIGGER DATA PROCESSING

- Individual collisions are scheduled ‘round robin’ on single-threaded processes – approx. one per core – with *static* data/control flow
- Hlt1: O(50) decisions, Hlt2: O(500) decisions
- O(100) “nodes” per decision
- Accept collision at each level if one (or more) positive decisions
- *All* decisions processed until ‘abort’ or ‘accept’ – no ‘early accept’!
- Each individual decision is based on different criteria, with (some) overlaps – but ‘logically independent’



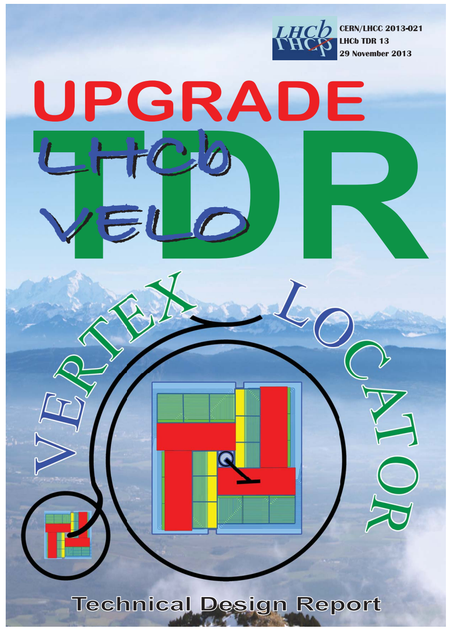
WORK IN PROGRESS

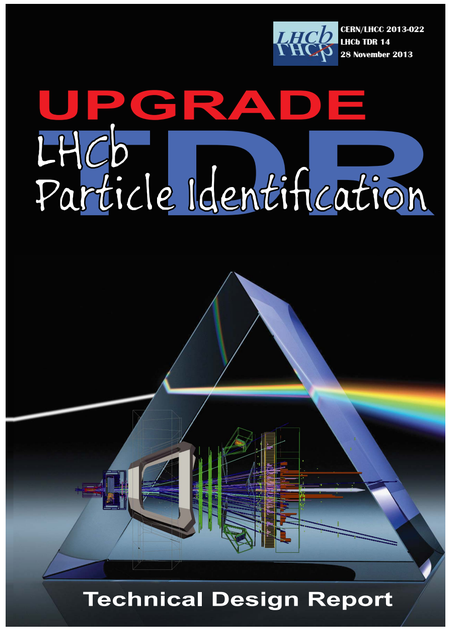
- Map to tasks & build a graph
  - Require explicit data dependency declarations & control flow definitions
- Redesign/refactor code for thread-safety
- Dynamically scheduling (for now: TBB tasks)
  - allow for latency (hiding) – necessary (but not sufficient!) for using accelerators
  - Allow multiple collisions ‘in flight’ to further increase parallelism and workload per task

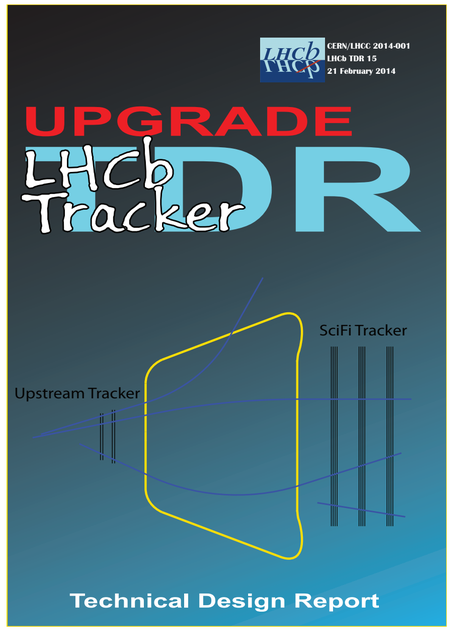


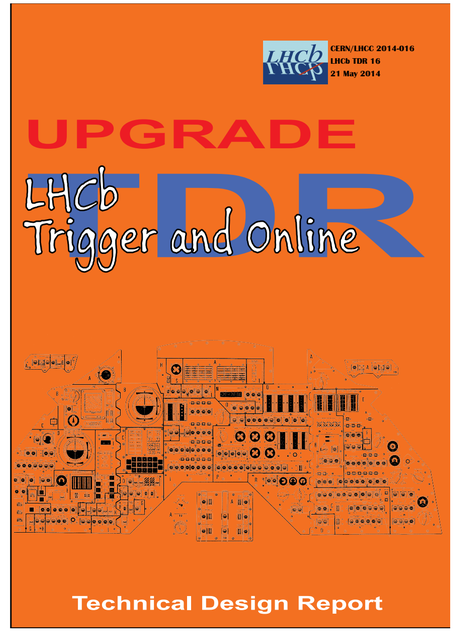
WHY? THE LHCB UPGRADE!

- To (continue to) make progress in the future:
  - Increase signal rate by (at least!) an order of magnitude
  - Increase luminosity x5, (trigger) efficiency x2 (depending on mode)

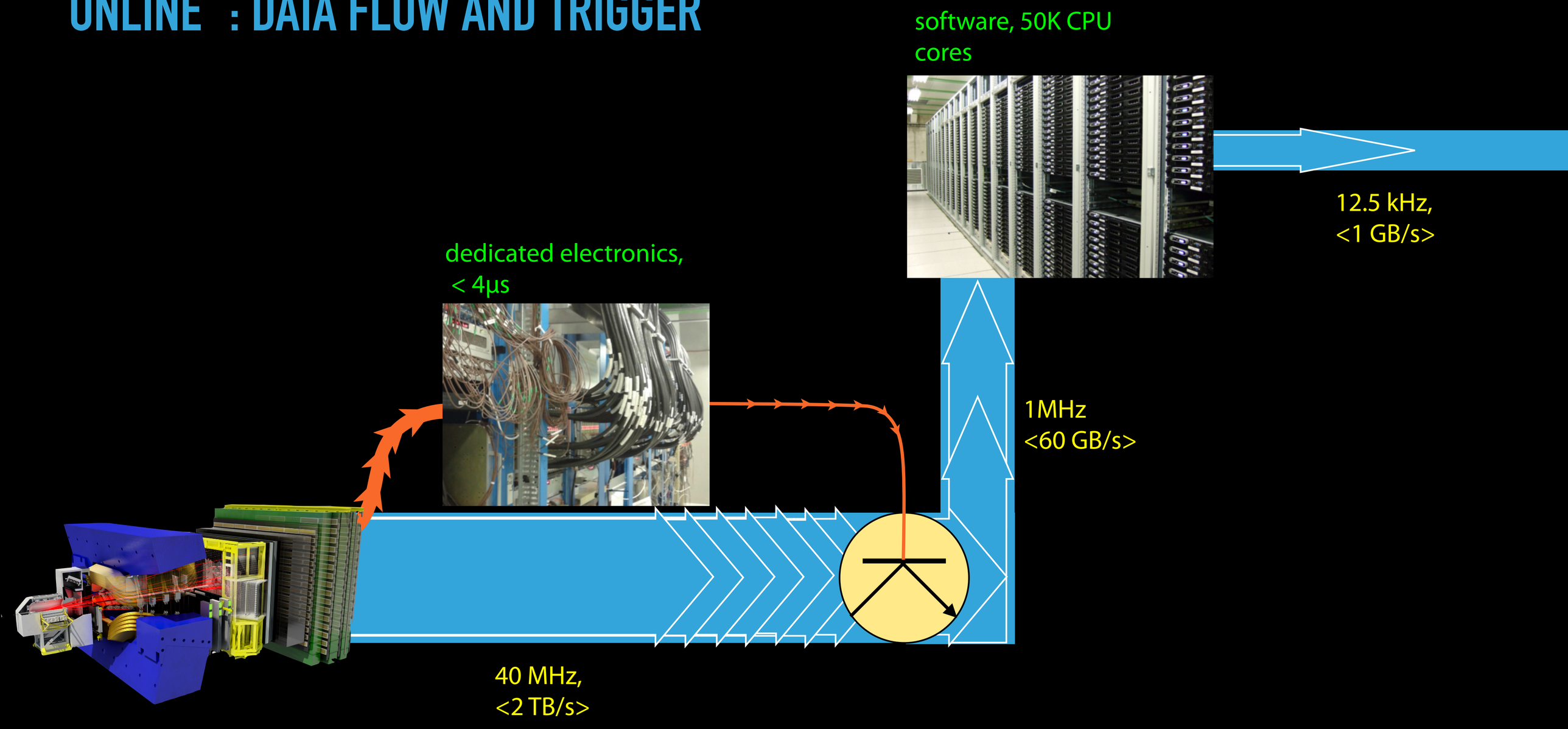




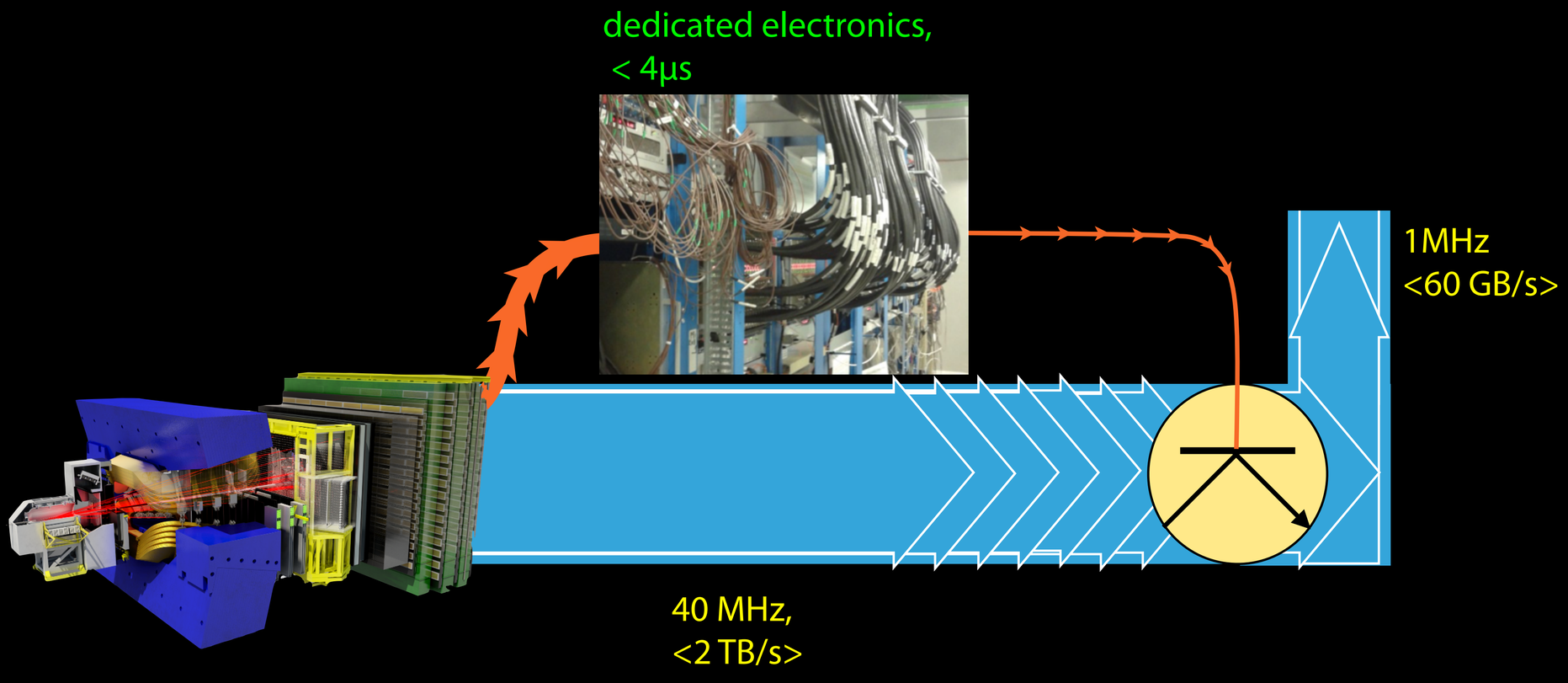




"ONLINE" : DATA FLOW AND TRIGGER

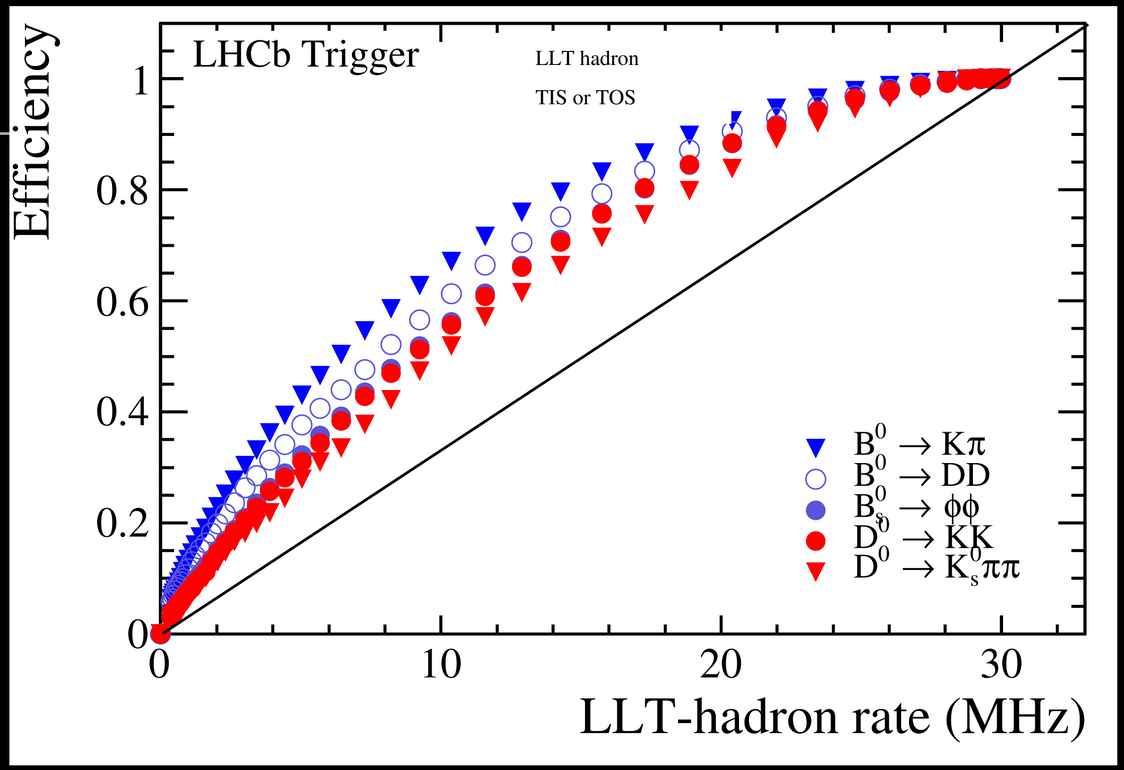


“ONLINE” : DATA FLOW AND TRIGGER

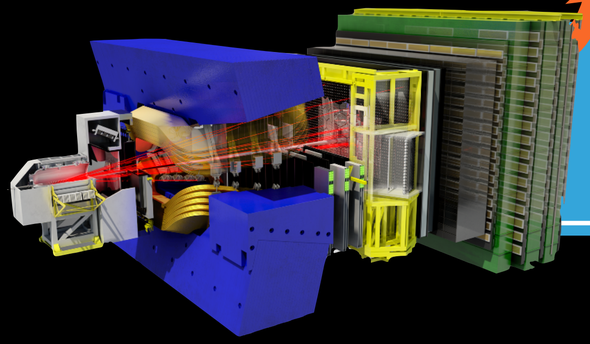


"ONLINE" : DATA FLOW AND TRIGGER

- At 5x luminosity, the 1 MHz readout rate becomes a bottleneck
  - Signal no longer identifiable by 'simple', fixed latency hardware processing



dedicated electronics,
< 4μs

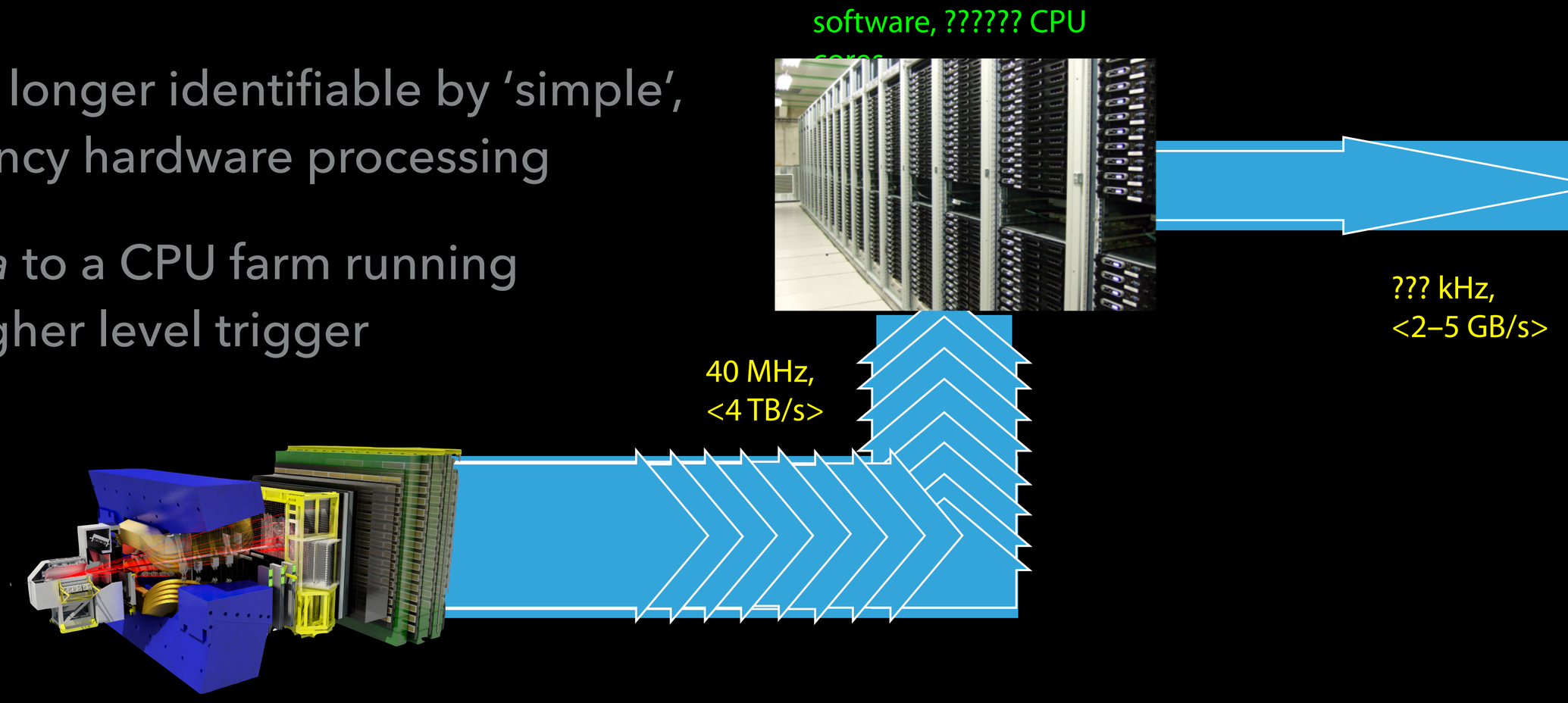


1MHz
<60 GB/s>

40 MHz,
<2 TB/s>

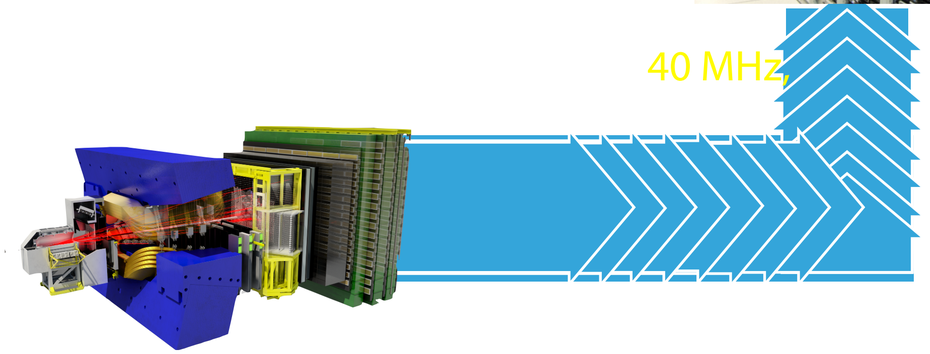
"ONLINE" : UPGRADE DATA FLOW AND TRIGGER

- At 5x luminosity, the 1 MHz readout rate becomes a bottleneck
  - Signal no longer identifiable by 'simple', fixed latency hardware processing
- Ship *all data* to a CPU farm running software higher level trigger



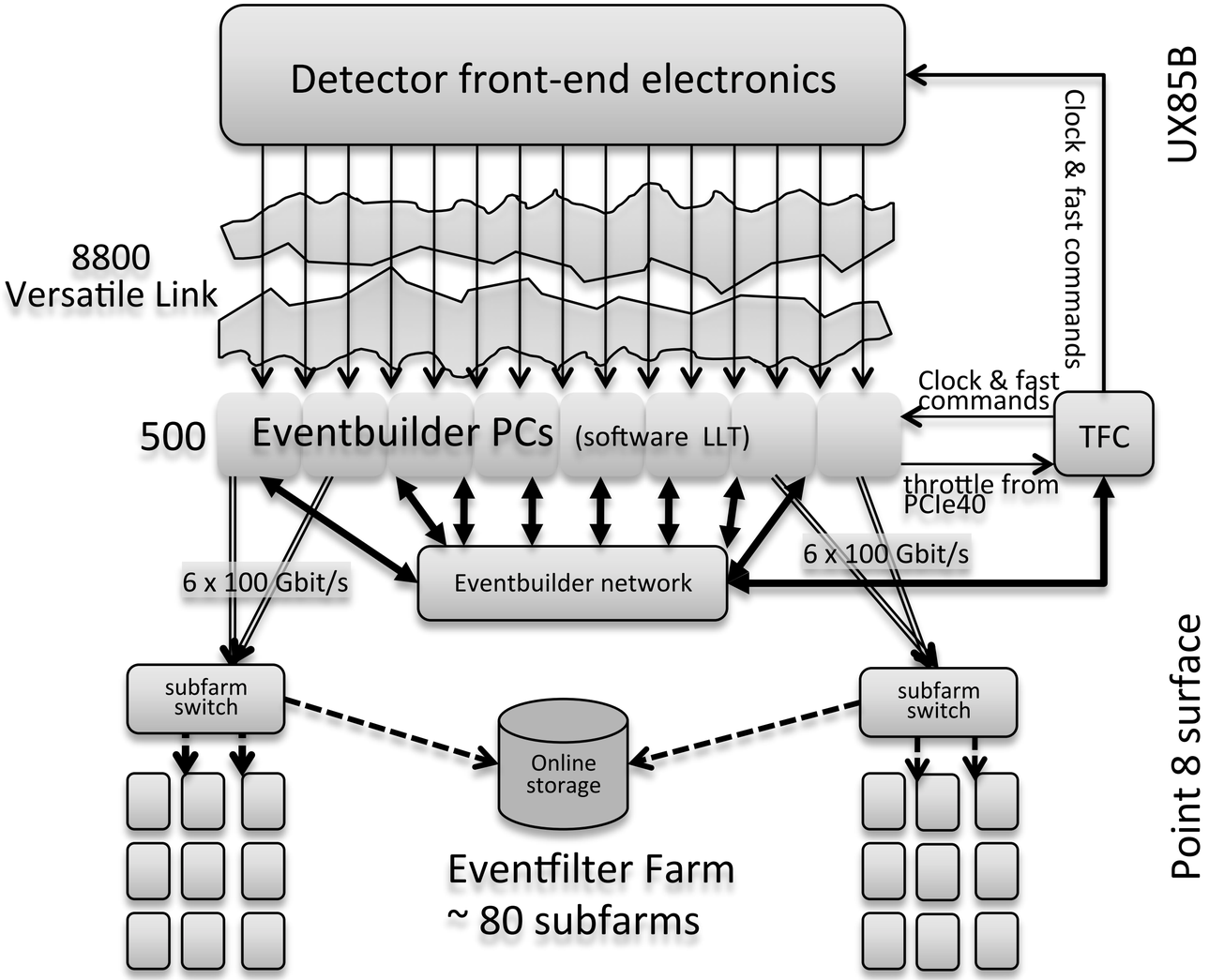
EVENT BUILDING @ 40 MHZ

- 32 Tbit/s
- “All data to the surface”



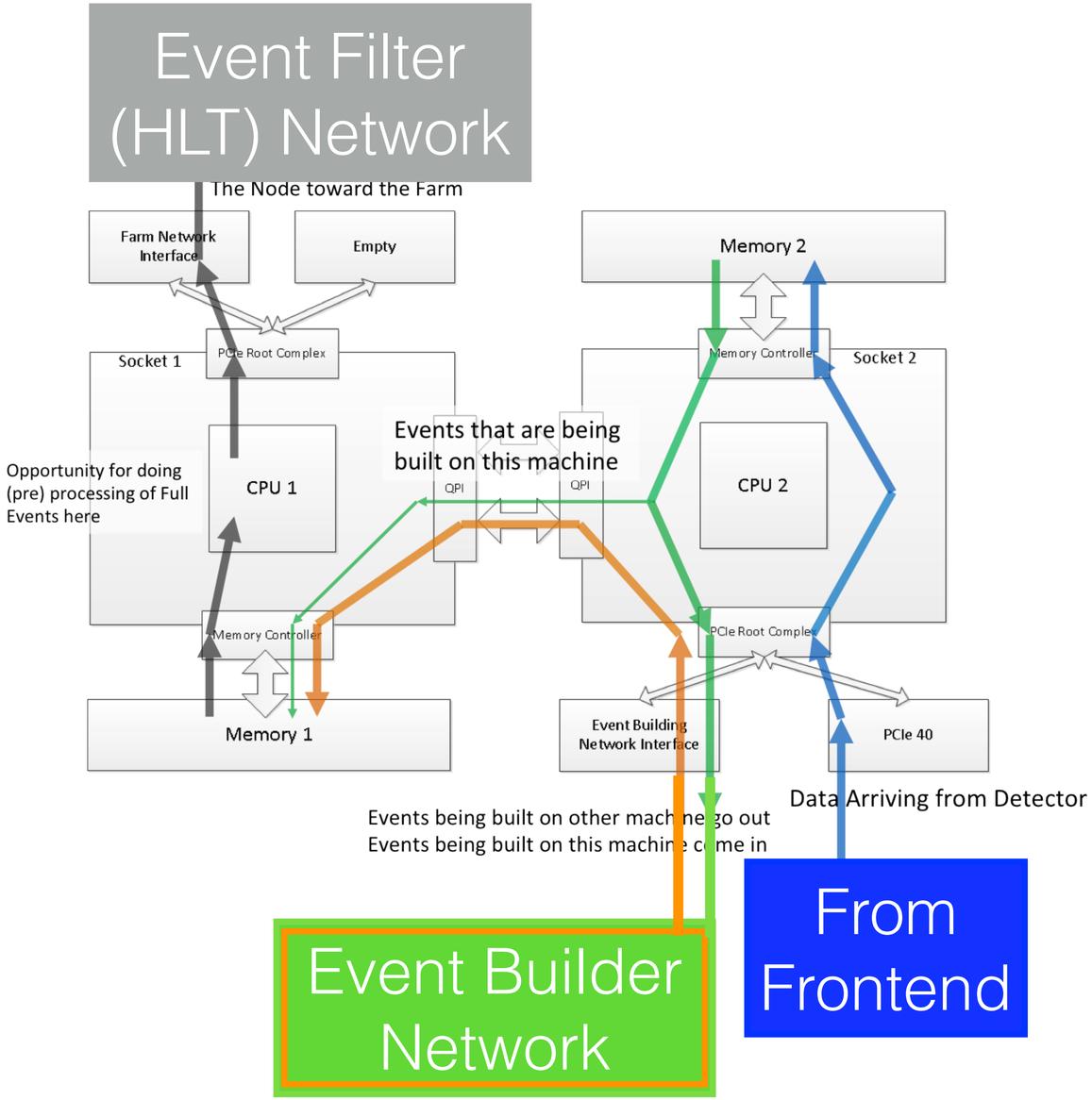
EVENT BUILDING @ 40 MHZ

- 32 Tbit/s
- “All data to the surface”
- Decouple front-end electronics from event builder network
  - Frontend→GBT link →PCIe
  - GBT link: rad-hard, integrated into front-end, so no commodity solution possible…
- Buffering in PC memory



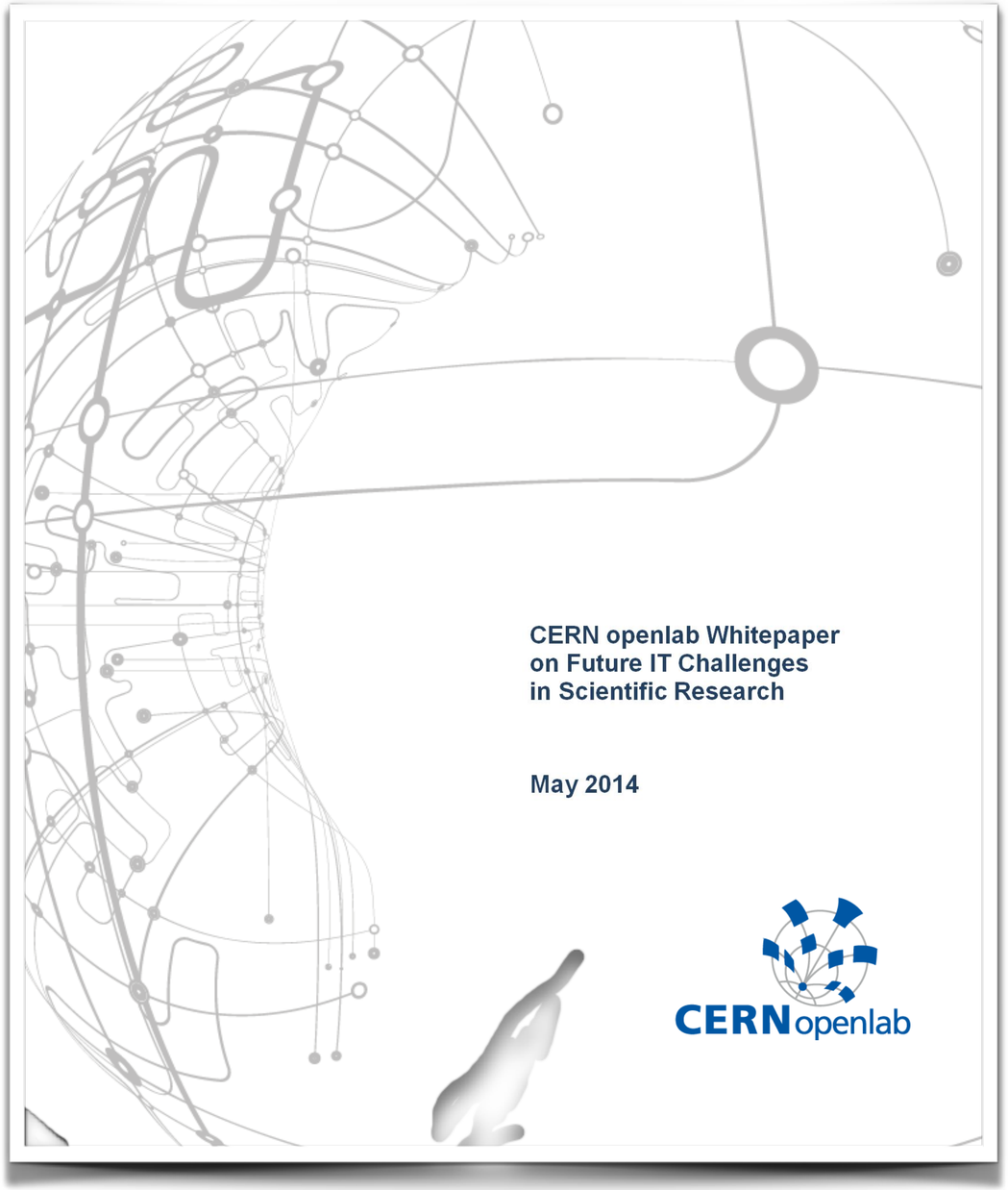
EVENT BUILDING @ 40 MHZ

- "COTS" as soon as possible
- O(500) servers for event building
- "Data Center" ("thin" switch, Infiniband/Ethernet/...) instead of "Telecom" (ATCA, "fat" switch)
- Event Filter: O(1000) servers



CERN OPENLAB WHITEPAPER

- “Data acquisition is where instruments meet IT systems. “
- “Costs and complexity must be reduced by replacing custom electronics with high-performance commodity processors and efficient software.”



SUMMARY

LHCb physics covers a large “dynamic range”

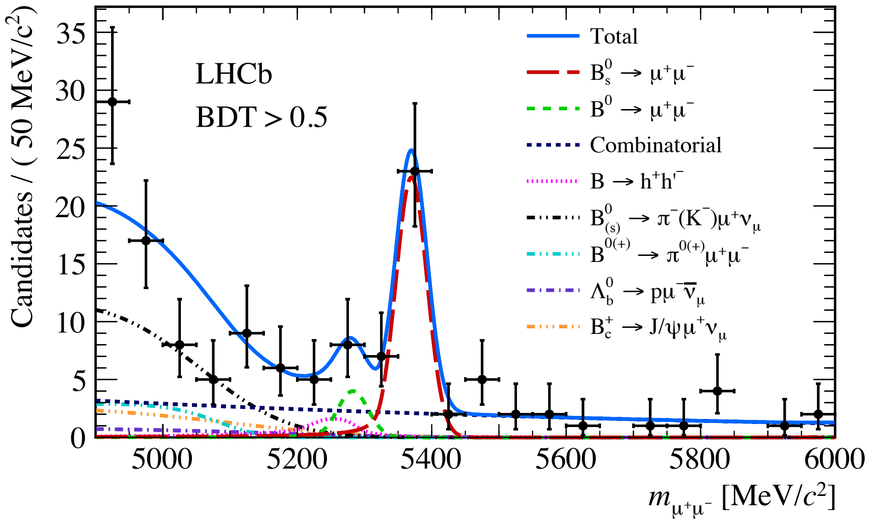
- high efficiency for the rarest B decay
- high purity for the largest charm samples

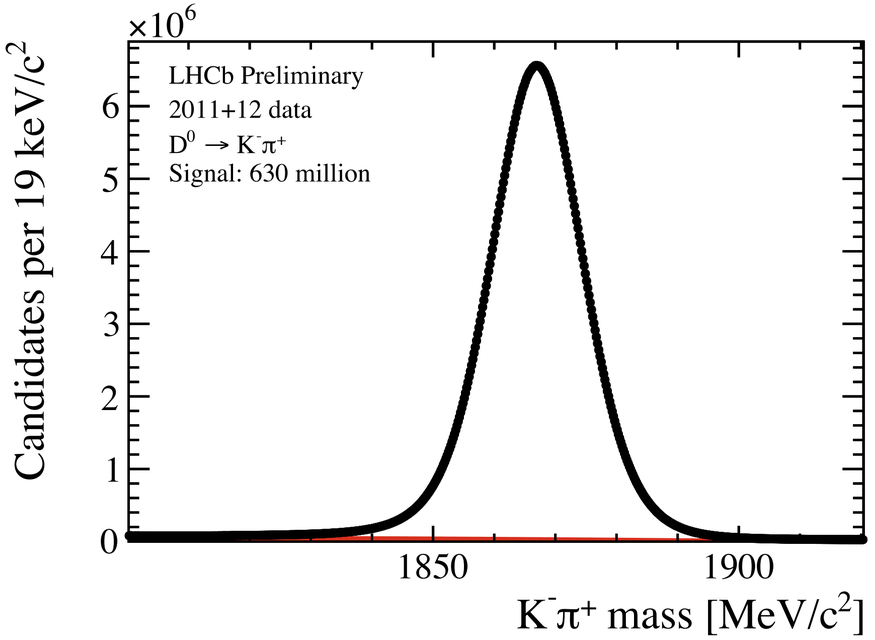
“Real-time” processing crucial for the physics reach

- In the future: software processing 30 MHz of collisions

Writing the required software will be a challenge!

- Robust – crashes will lead to data loss
- Correct – mistakes will render the data ‘useless’
- Efficient, both in reconstruction & selection, and use of computing resources





(a) $D^0 \to K^-\pi^+$