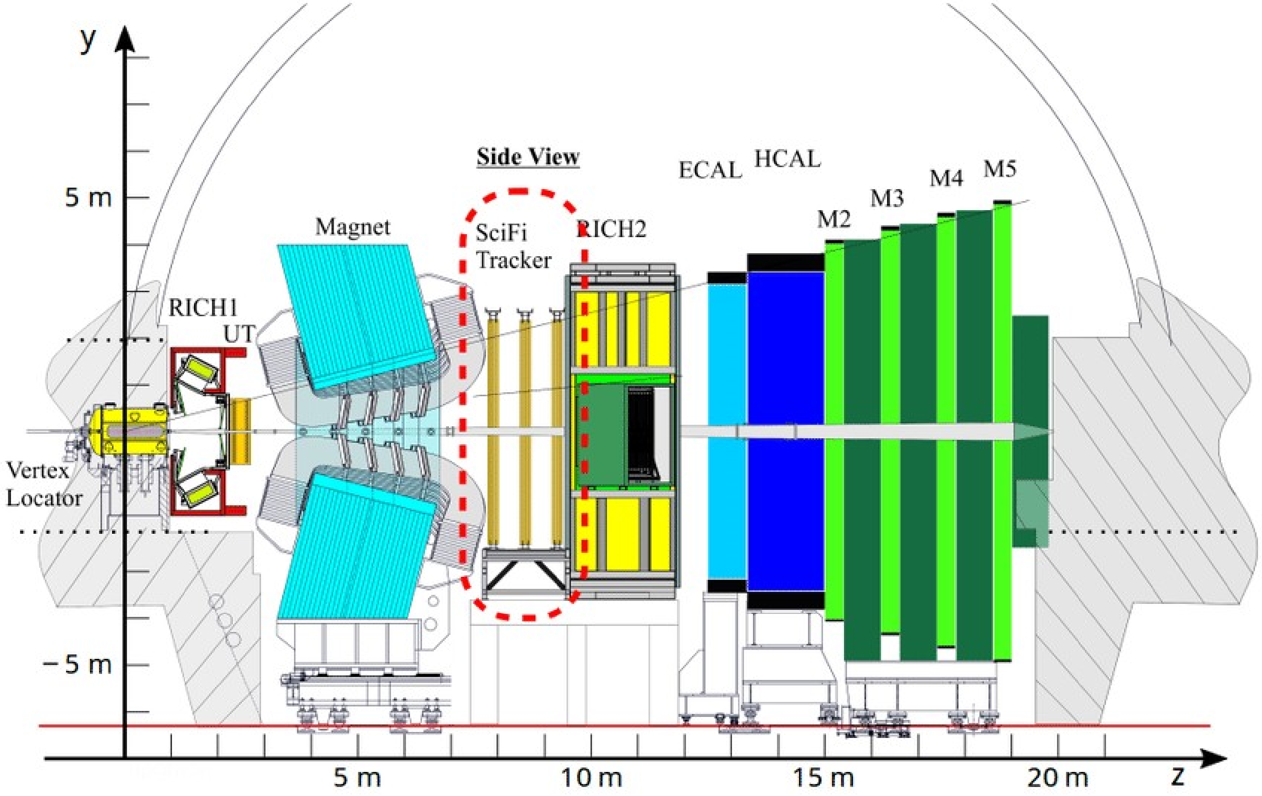
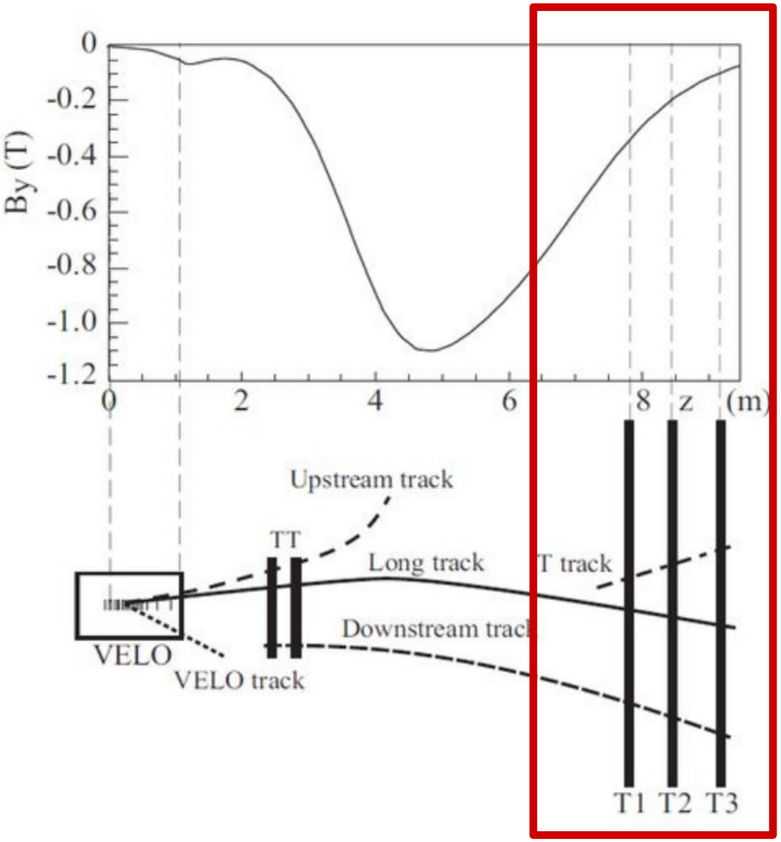
Hybrid seeding at LHCb: flexibility begets flexibility

Louis Henry
Swiss Physics Society Joint Annual Meeting, Basel, 06/09/2023

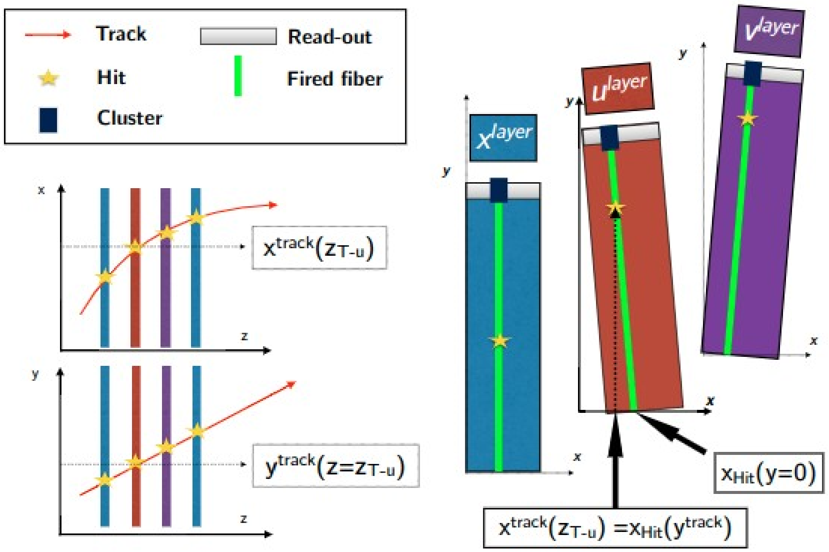
The hybrid seeding at LHCb

- LHCb is a detector along the LHC, specialised in the study of beauty and charm hadrons [JINST 3 (2008) S08005]
- The hybrid seeding is the **stand-alone reconstruction algorithm of the SciFi tracker.**
  - Needs to runs inside the online trigger (total throughput needed ~ kHz/node).





Hybrid seeding: overall strategy



- SciFi: three stations arranged in a x-u-v-x geometry, u and v being layers titled by a +/- 5° stereo angle.
  - Easier to get x coordinate than y coordinate.
  - But ~only residual B_y field → simpler y trajectory (line).

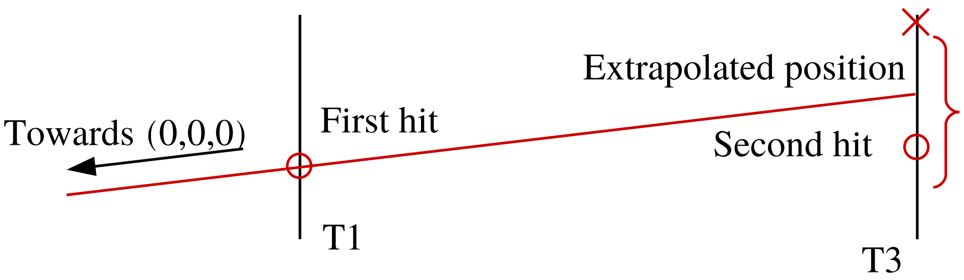
- Combinatorics too large to tackle all at once → **iterative strategy**: go for high-momentum first (~straight lines), cleanup the environment progressively.

- Each iteration starts with different pair of layers in T1 & T3.
  - Covers for hit inefficiency → modest theoretical cap on efficiencies.

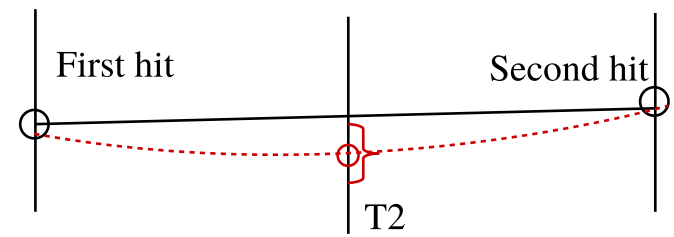
Hybrid seeding: the gist

Principle of the search:

- Starts with doublet search in T1 & T3, windows depending on minimum p, taking charge asymmetry in consideration.



- For each doublet, already a charge-momentum estimation → narrower windows to look for 3rd hit in T2 station, taking bending into account.



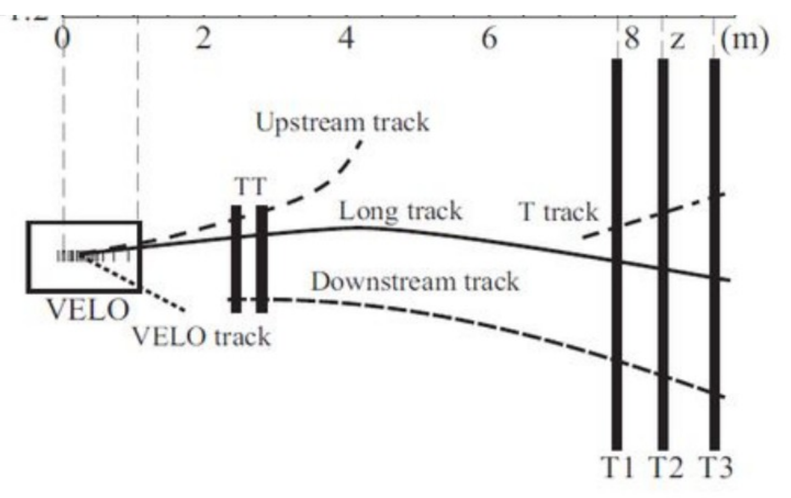
- Triplet → track model. We look for at least 2 remaining hits → **XZ segment**.
- Real tracks have **~ constant t_y = y/z** if no scattering and come from close to the origin.
  - Solution: **discretised Hough cluster** search in bins of t_y = y/z.

	t_y bin 0		t_y bin n-1	t_y bin n
T1U	1	...	1	0
...	...	...	...	
T3V	0	...	1	1
Total	3	...	5	2

Porting the seeding to GPU: a whole new world

- The LHCb strategy in the upgrade relies on a **full-software** trigger.
  - HLT1 = partial signatures indicating heavy-flavour decay. Runs at **30 MHz** on GPUs.
  - HLT2 = complete reconstruction of events. Runs at **1 MHz** on CPUs.
  - HLT2 aims at maximum efficiency, HLT1 can focus on easier tracks.

- The hybrid seeding has historically been developed as an HLT2-only algorithm.
  - Useful to reach maximum efficiency on Long tracks.
  - Critical for the reconstruction of downstream and T tracks.



- But what if we could run it in HLT1?
  - Increased statistics on modes with displaced vertices; maximum efficiency reached earlier.

Porting the seeding on GPUs: how-to for non-computing experts

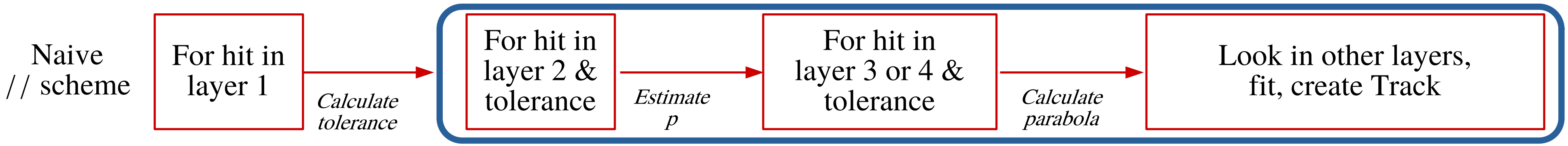
- GPUs have 3 levels of memory, with decreasing (increasing) space (access) speed: global, shared, and registers.

- Change algorithm to **reduce memory usage**
  - Replace the Hough cluster with an algorithm similar to XZ search
  - Precalculations are actually harmful if we store more than the allowed numbers of registers!

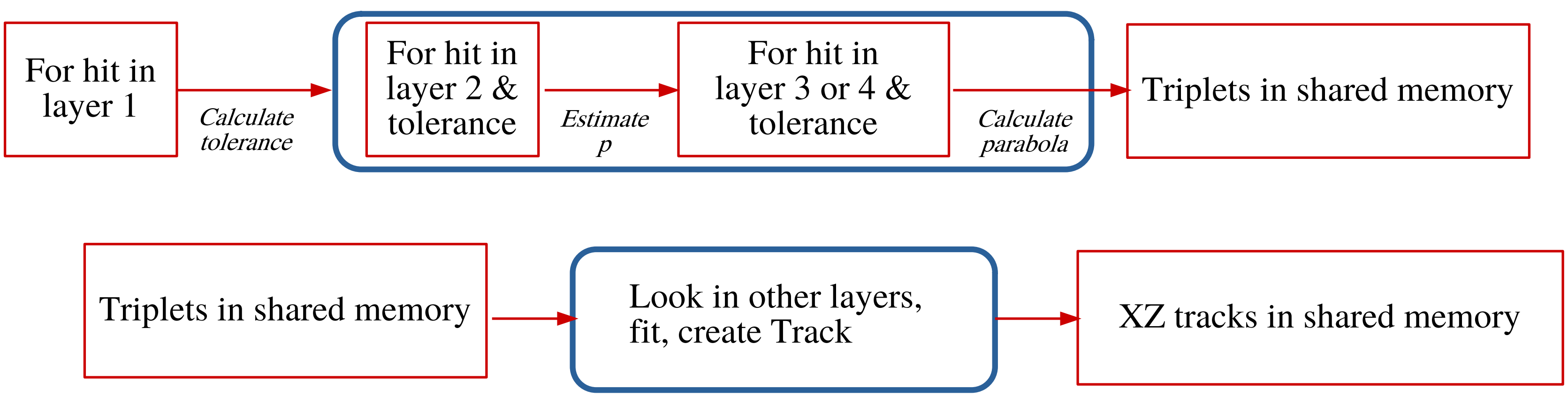
- **Reduce conditionals** and early breaks.
  - Parallelisation works best when threads are doing **something** and doing the **same** thing.
  - “Hit flagging” of the original seeding relies on conditionals too much → run twice the algorithm with different sets of initial layers.

Porting the seeding to GPU: the result

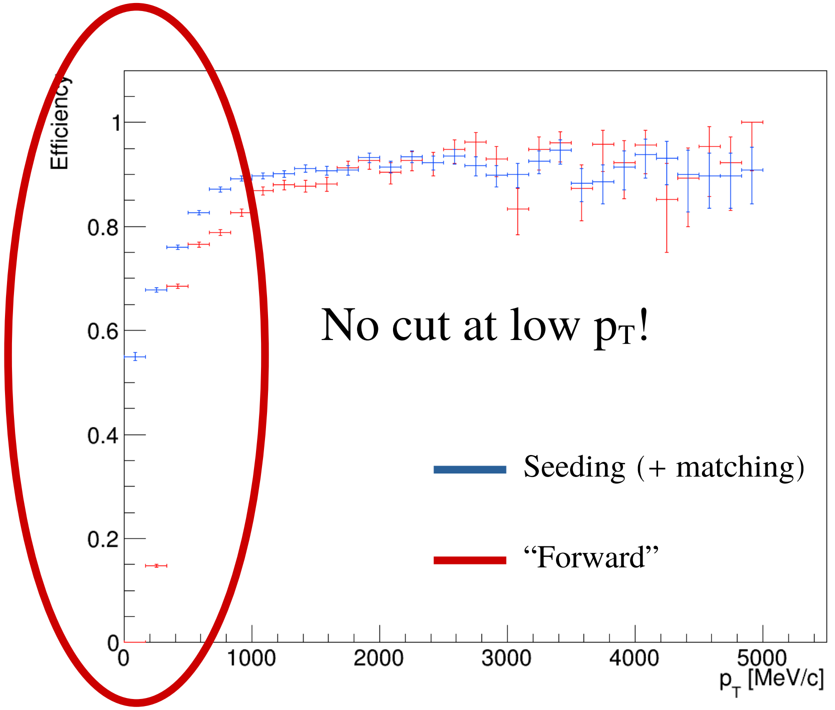
- XZ search: naive sequence would use one level of parallelisation (over first hits)

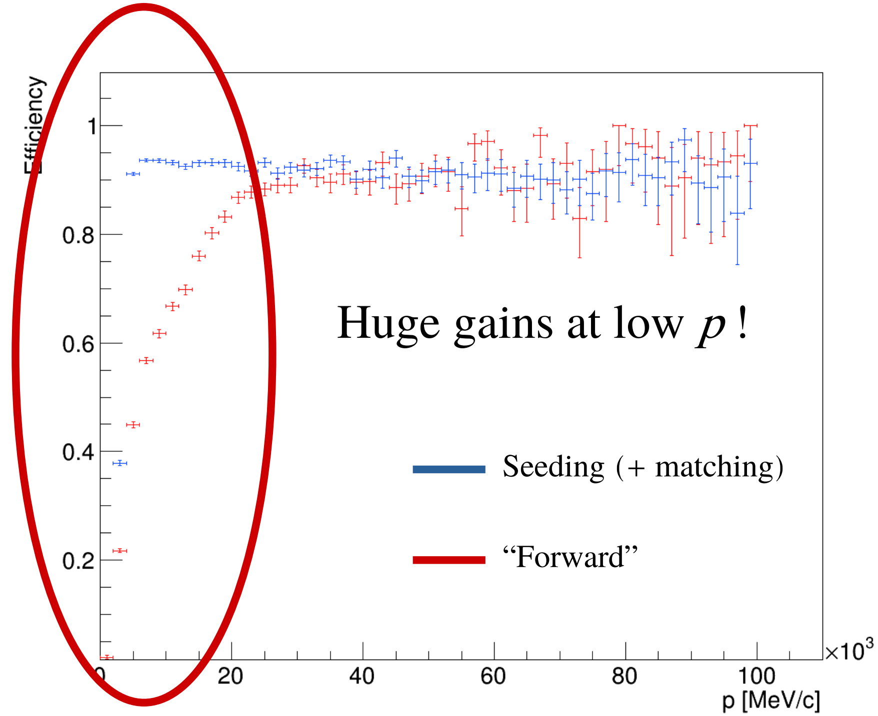


- Studies on MC show that 80% of triplets get promoted to full track → costly for many threads without a triplet to wait for the other ones to finish.
  - New scheme is in two parallel sequences
  - First sequence is fast, and hit-or-miss (many doublets do not have a triplet); second sequence is slow but high occupancy of threads.



Porting the seeding on GPUs: more than a technical feat





- Early trigger on displaced vertices benefits novel physics programmes, like searches for long-lived particles.
- Monolithic reconstruction holds promises for possible FPGA-based trackings (Upgrade 2?)

Through a flexible approach and a porting to GPUs, the Hybrid Seeding is in position to facilitate a whole sector of searches at LHCb

Questions?

CPU vs GPU: different machines, different inputs, different uses