Presenter: Hannes Mohr - KIT, IPE

Contributors: L. Ardilla Perez, M. Balzer,
M. Caselle, S. Chilingaryan, T. Dritschler,
A. Kopmann, L. Rota, T. Schuh, M. Weber

Karlsruhe Institute of Technology

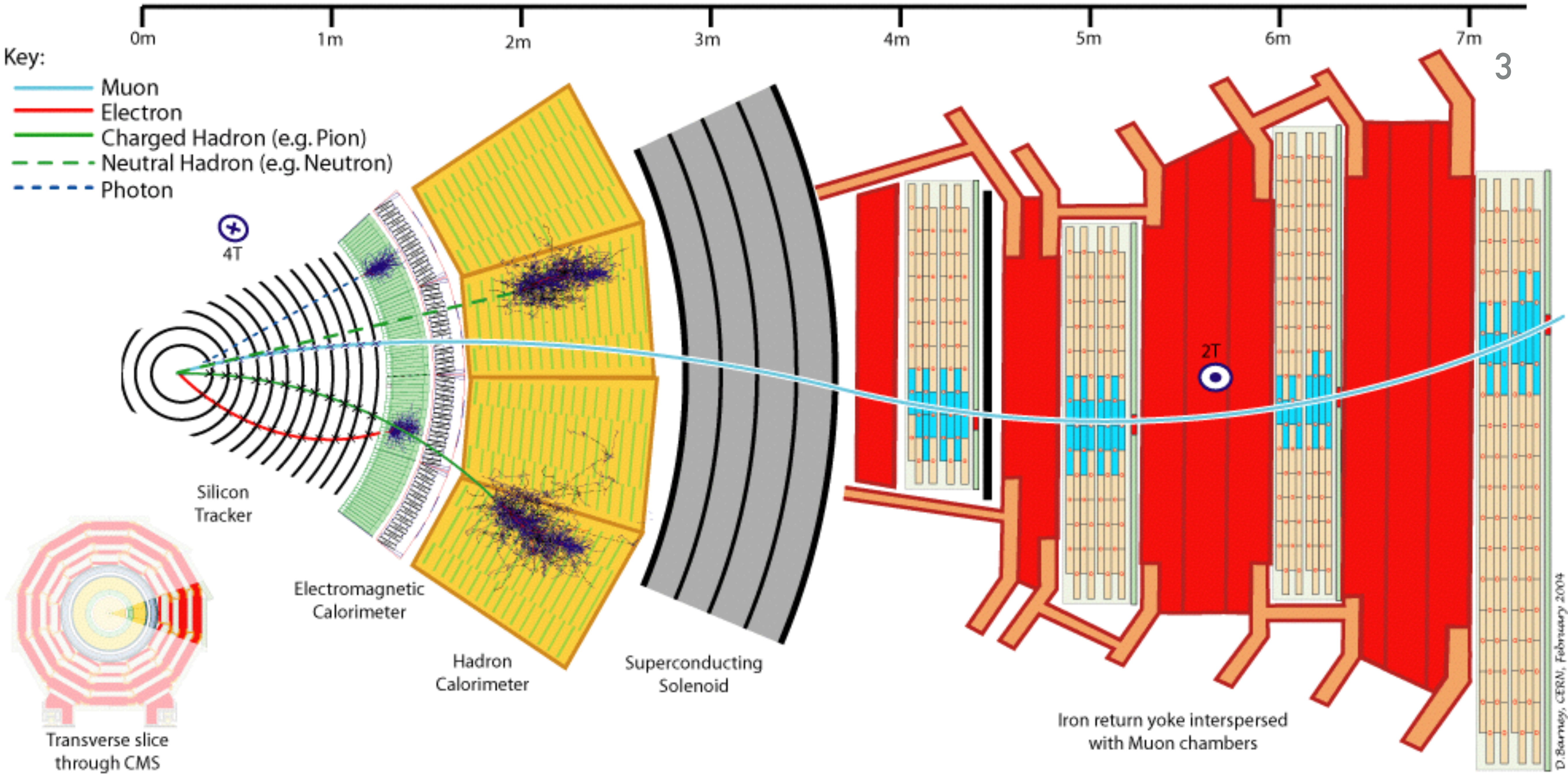
29TH SEPTEMBER, TWEPP 2016

EVALUATION OF GPUS FOR HIGH-LEVEL TRIGGERS IN HIGH ENERGY PHYSICS

OUTLINE

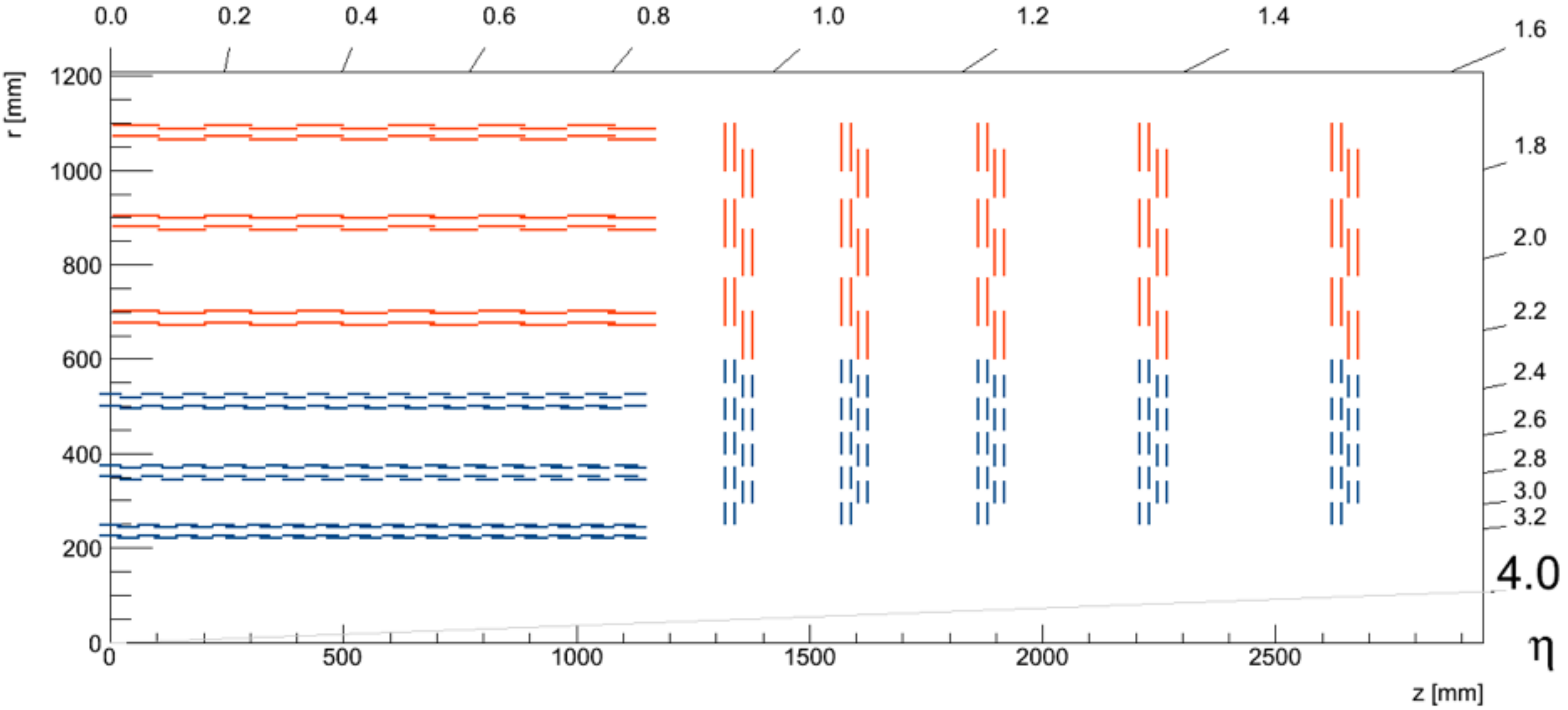
- Implement track trigger using GPUs
- Use established methods for seeding
- Present our own version of the Hough transformation
- Compare different GPUs/vendors
- Investigate data transfer/latencies
- Estimate impact of technological advances

OUR GOAL IS TO ACHIEVE COMPETITIVE RESULTS, WHILE GAINING FLEXIBILITY


0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004

CMS DETECTOR

*Image CMS Collaboration



BASELINE GEOMETRY - 6 LAYERS 5 DISKS

SILICON TRACKER

*Image CMS Collaboration

UPGRADE TIMELINE

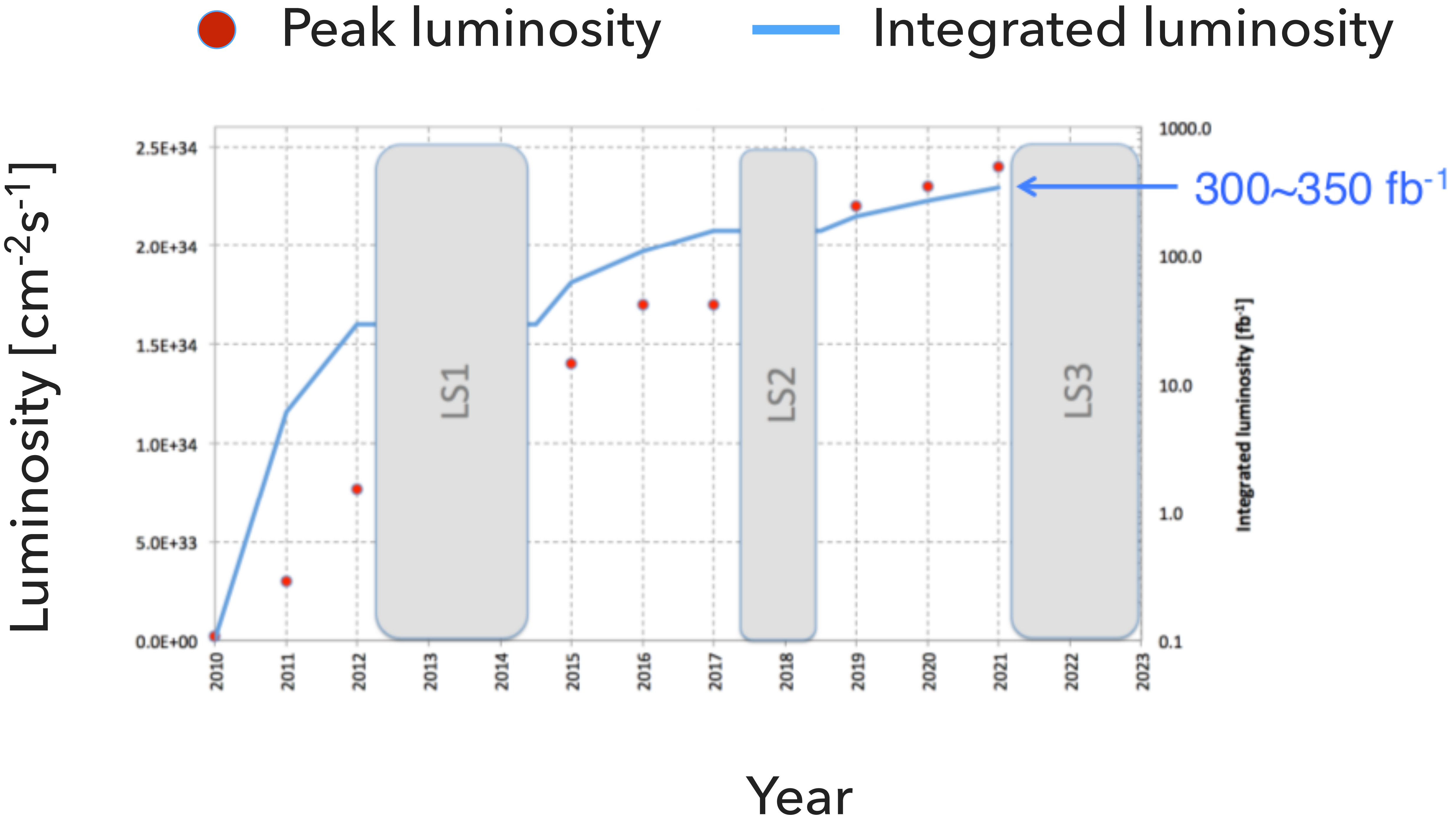


*Image CMS Collaboration

THE LS3 UPGRADE

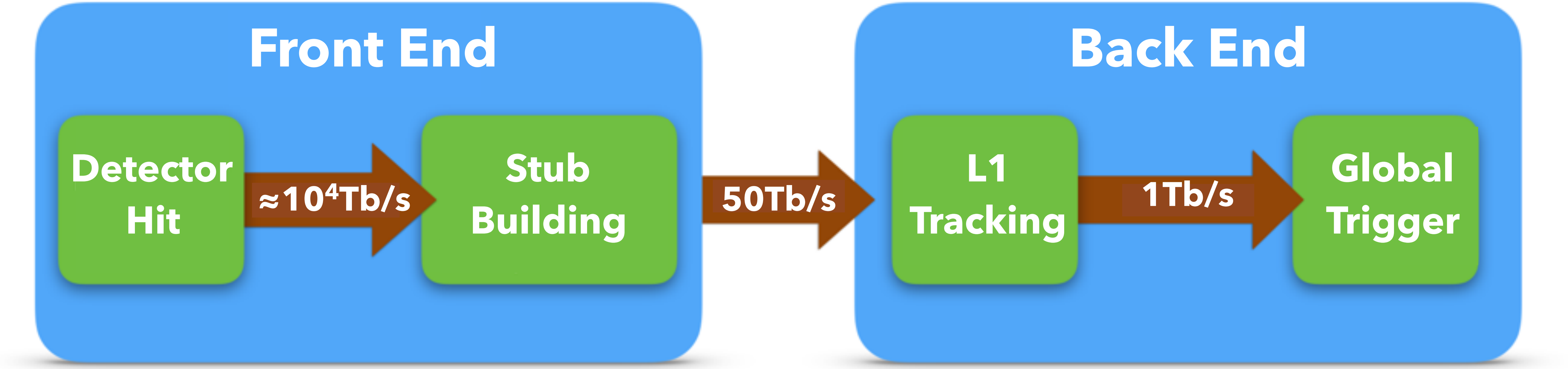
- **Current CMS trigger won’t be able to handle:**
  - Increased data rates
  - Increased pile-up

?

- **Currently proposed solution:**
  - Data reduction on detector
  - Raise latency of trigger from 3.4 to 12.5 µs
  - L1 track trigger

!

EXPECTED DATA RATES



Front End

Detector Hit

≈10^4Tb/s

Stub Building

50Tb/s

Back End

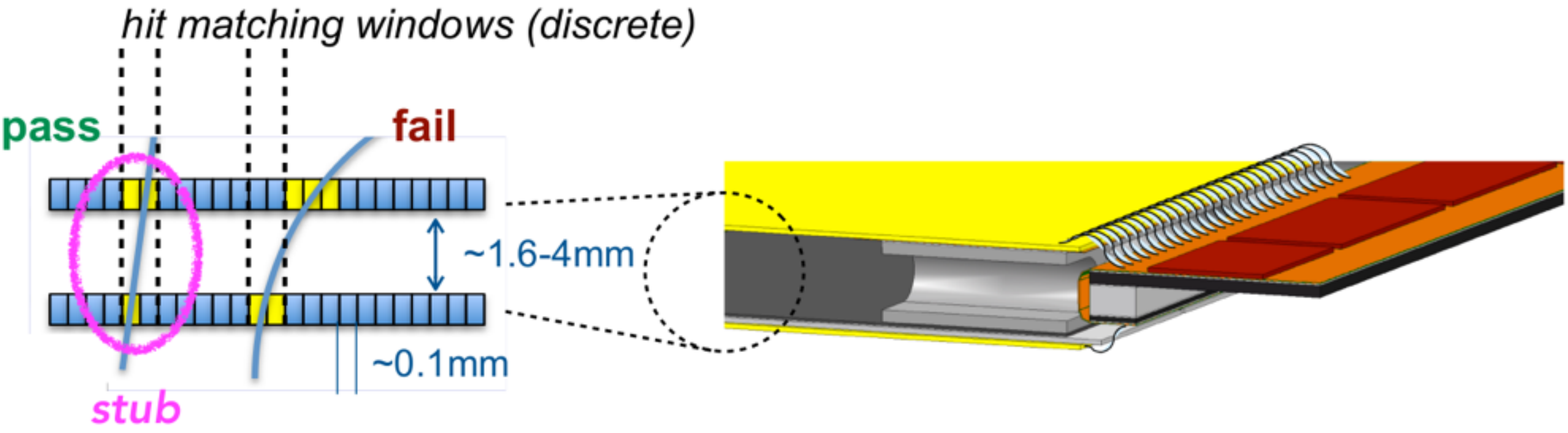
L1 Tracking

1Tb/s

Global Trigger

- Readout at 40 MHz, BX every 25 ns
- 6 μs each for L1 Trigger and Global Trigger
- L1 Tracking to combine Track seeding and Fitting

STUB BUILDING



- Applies momentum cut to hits
- Delivers estimate on track bend
- Drastically decreases number of hits by a factor of 100

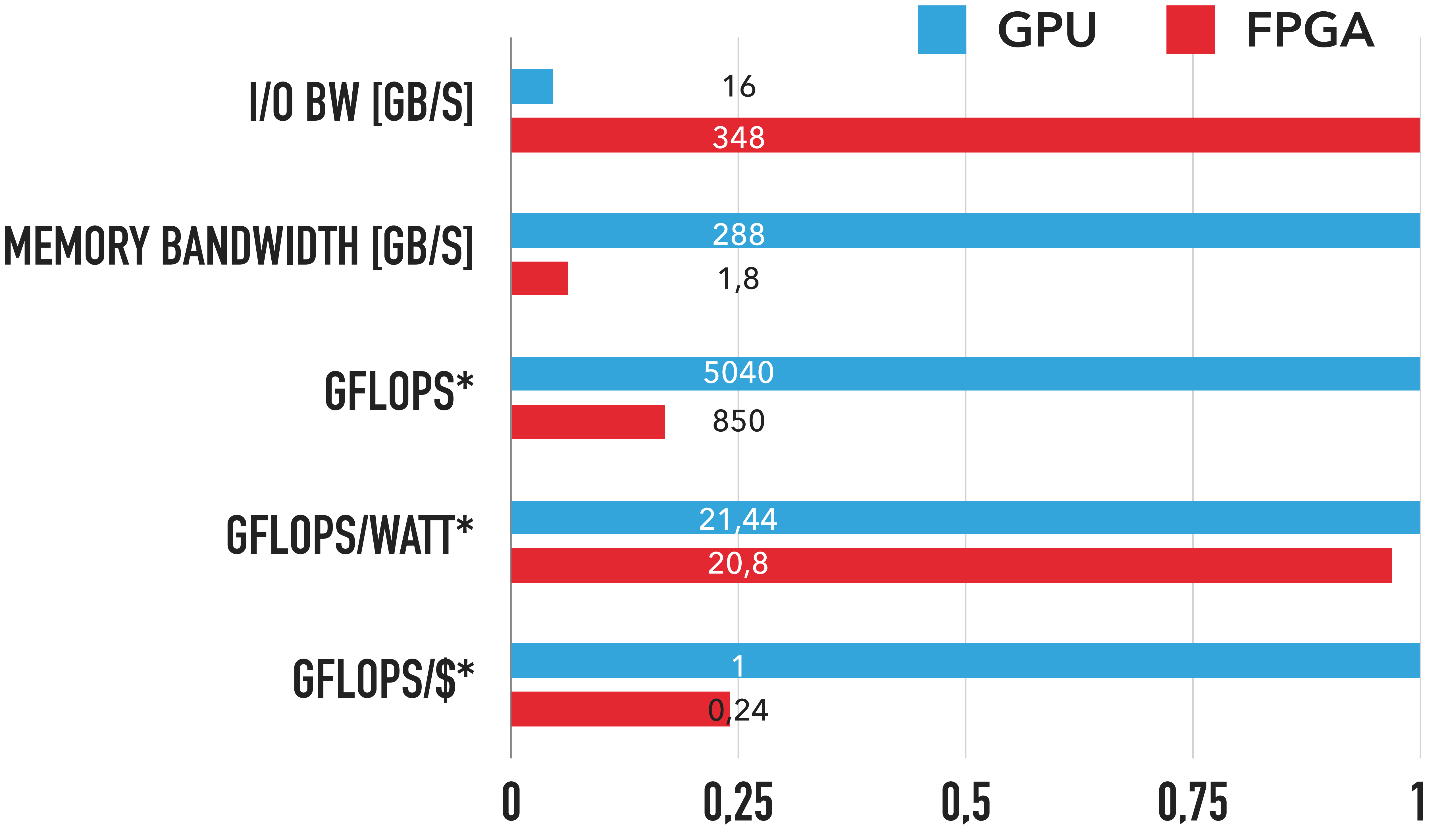
CURRENT APPROACHES

- Associative Memory approach (ASICs)
- Time-multiplexed FPGA Hough transformation
- …

CURRENT APPROACHES USE SPECIALIZED HARDWARE

COMPARISON GPU VS. FPGA

Nvidia Tesla K40c vs. XILINX VIRTEX-7 XC7VX1140T (both 28nm)



*SINGLE PRECISION,FLOATING POINT

QUALITATIVE COMPARISON OF STRENGTHS

GPU	FPGA
Rapid development cycles and high flexibility	Huge I/O Bandwidth
Large bandwidth to external memory	Deterministic timings/ runtimes
High Floating-point performance	High bit-level performance

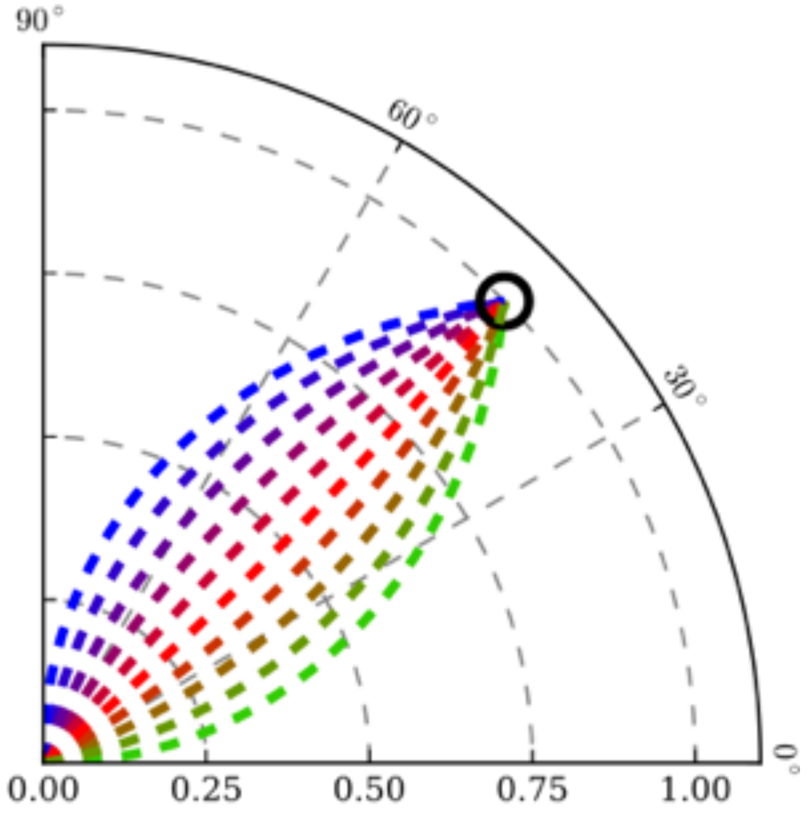
HOUGH TRANSFORMATION

Equivalent to FPGA approach, used by collaboration of KIT and UK Track Trigger Group*

- Uncompress data
- Perform Hough transformation
  - Uses module bend information
- Apply layer condition
- Reject or return track candidates

*An FPGA-Based Track Finder for the L1 Trigger of the CMS Experiment at the High Luminosity LHC (DOI: 10.1109/RTC.2016.7543102)

HOUGH TRANSFORMATION

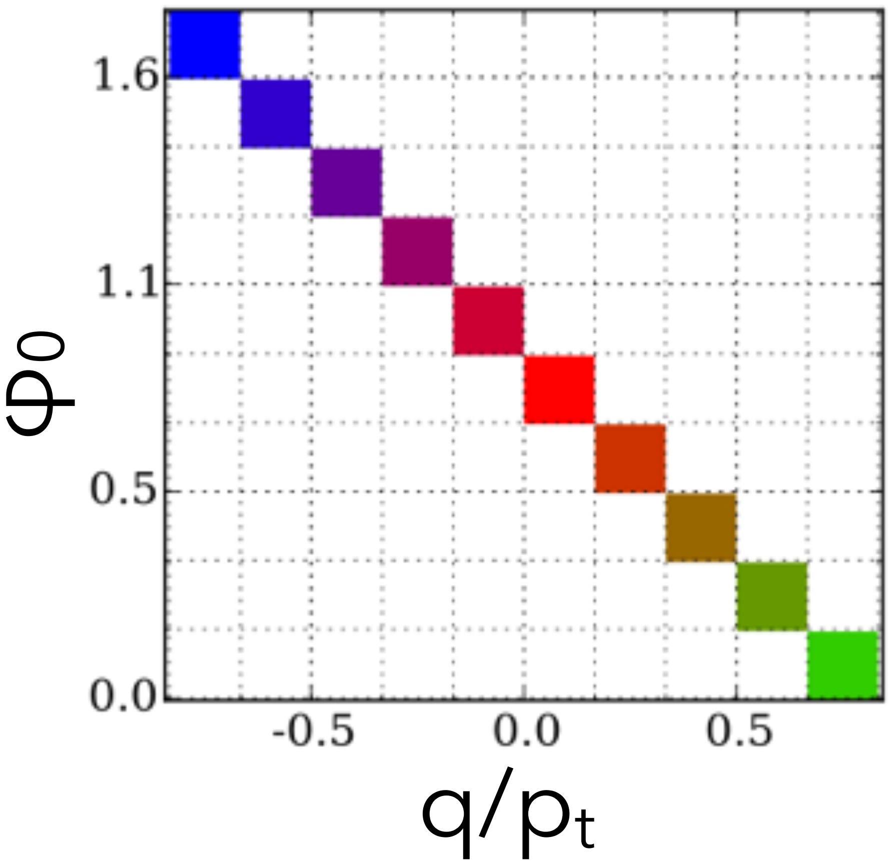


Calculate possible

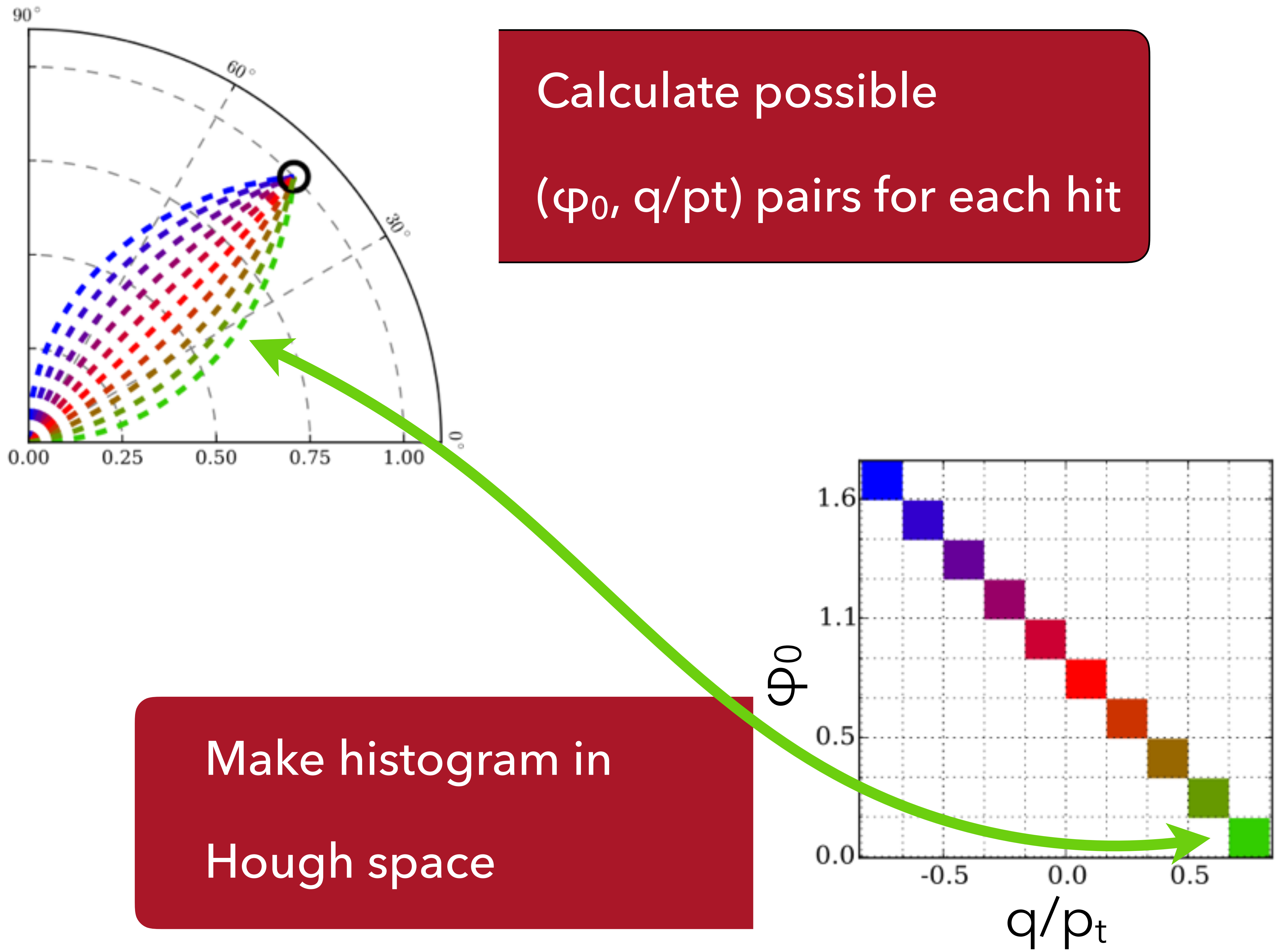
(φ$_0$, q/pt) pairs for each hit

Make histogram in

Hough space



HOUGH TRANSFORMATION



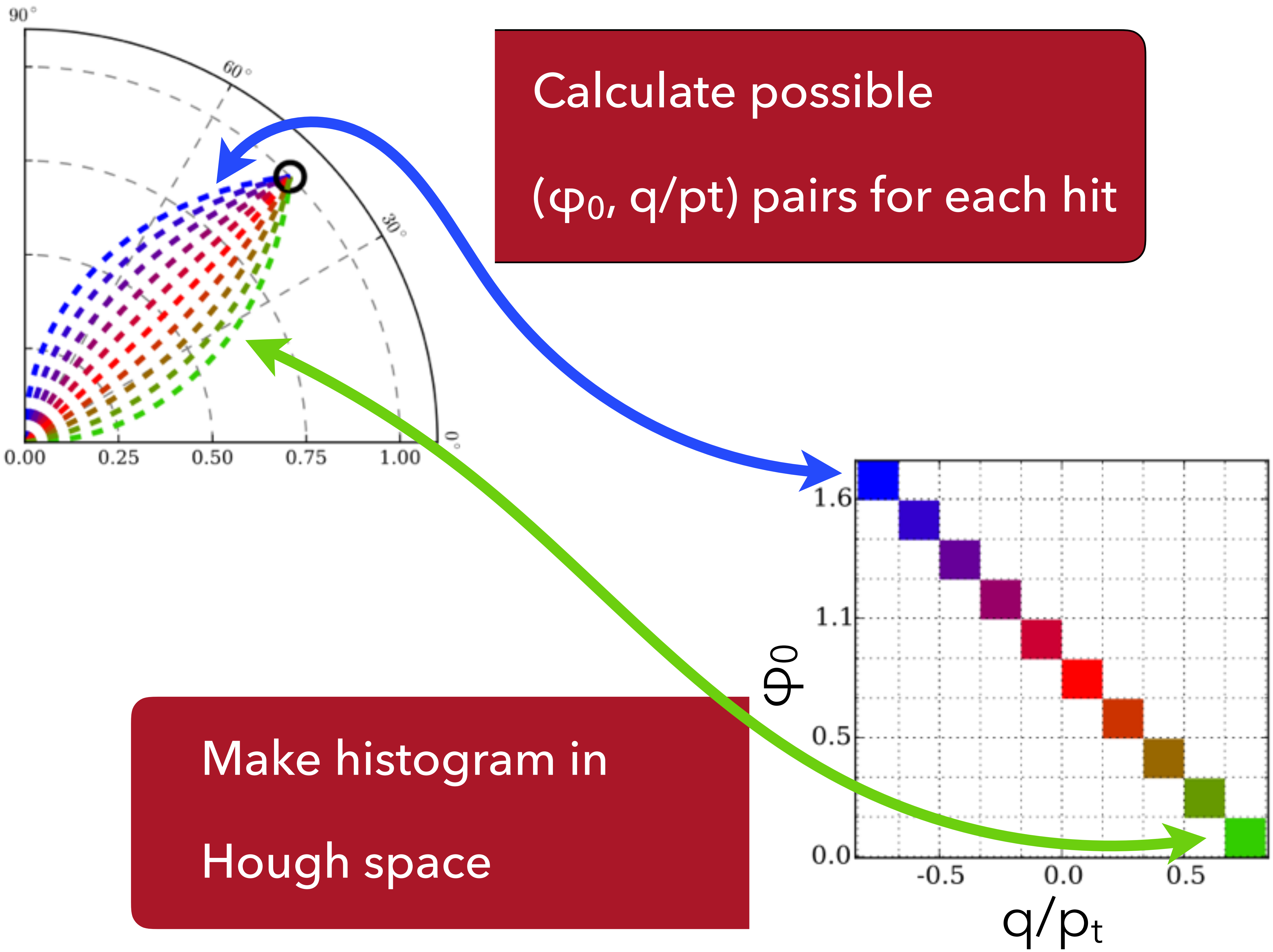
Calculate possible

(φ_0, q/pt) pairs for each hit

Make histogram in

Hough space

HOUGH TRANSFORMATION



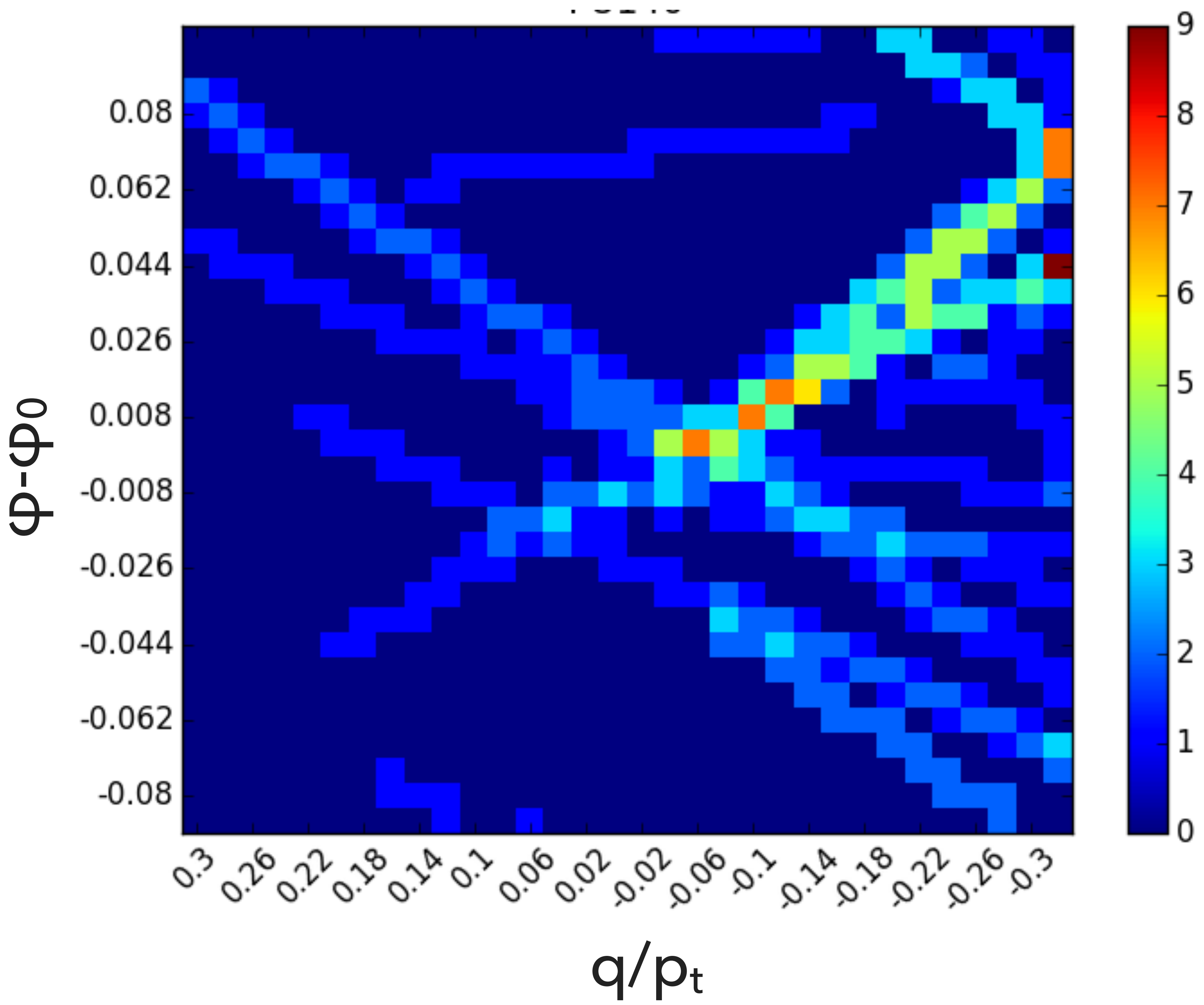
Calculate possible

(φ_0, q/pt) pairs for each hit

Make histogram in

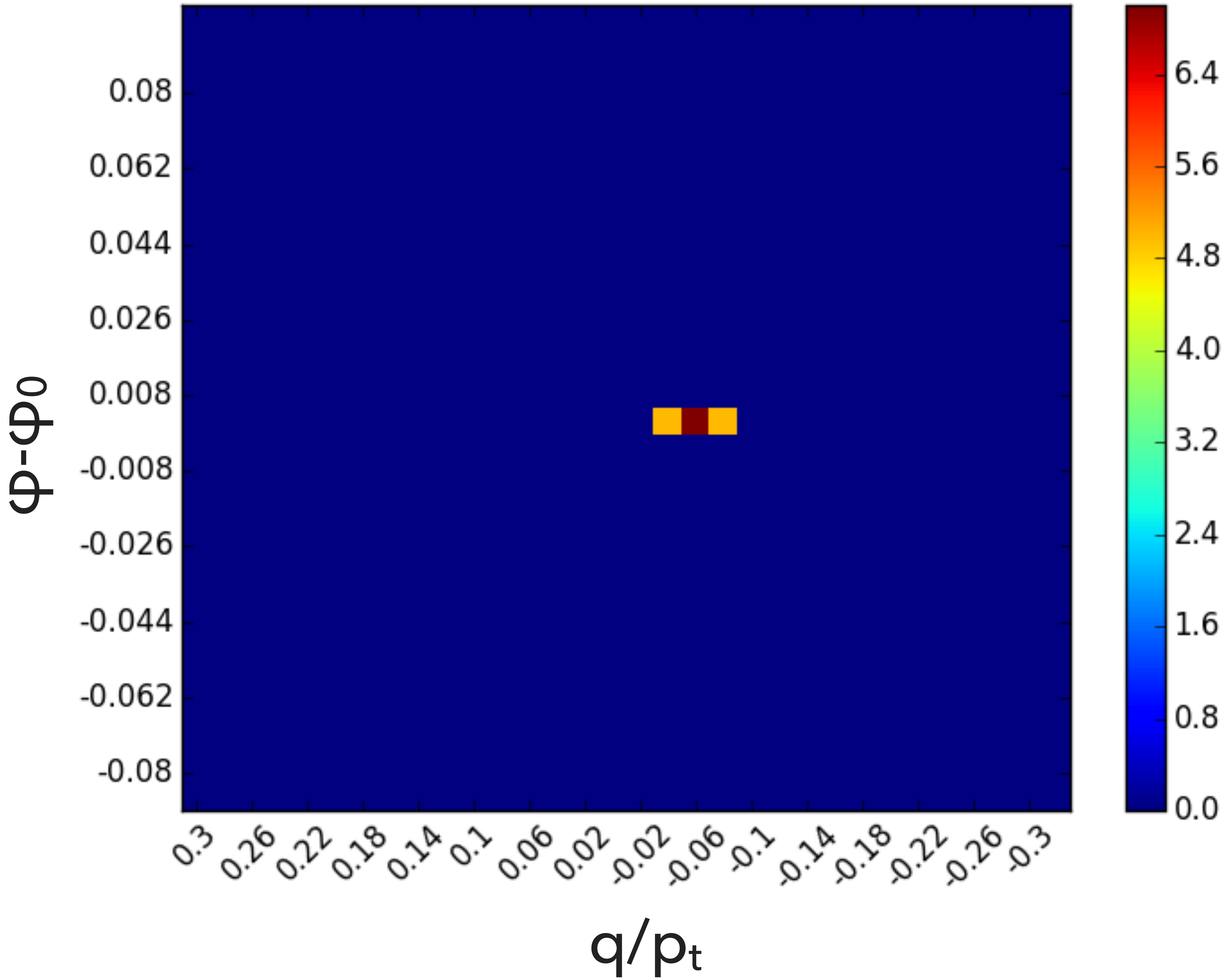
Hough space

CLUSTERING POINTS ARE TRACK CANDIDATES



TTBar event - PU 140

FILTERING BY LAYER CONDITION



TTBar event - PU 140

OUR IMPLEMENTATION

GPU implementation specifics:

- Optimized for minimum latency
- Computes q/p_t-bins in parallel
- Almost no dependence on number of stubs

OVERHEADS…

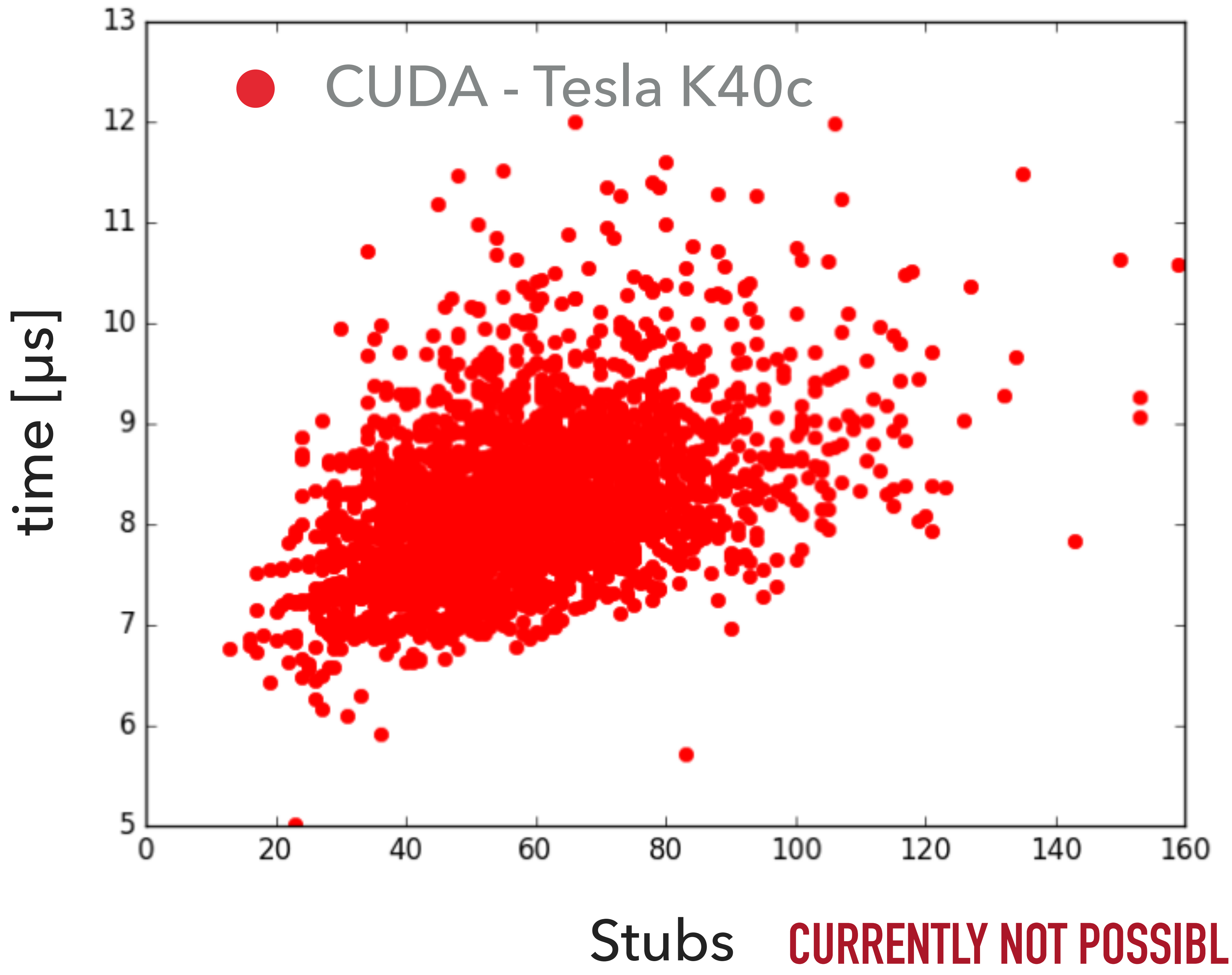
- Kernel scheduling
- Kernel launch time
- Allocation of shared memory
- …?

Invocation and setup of kernels is too costly,

we need to keep it running continuously

→ **NEEDS SPINNING KERNEL**

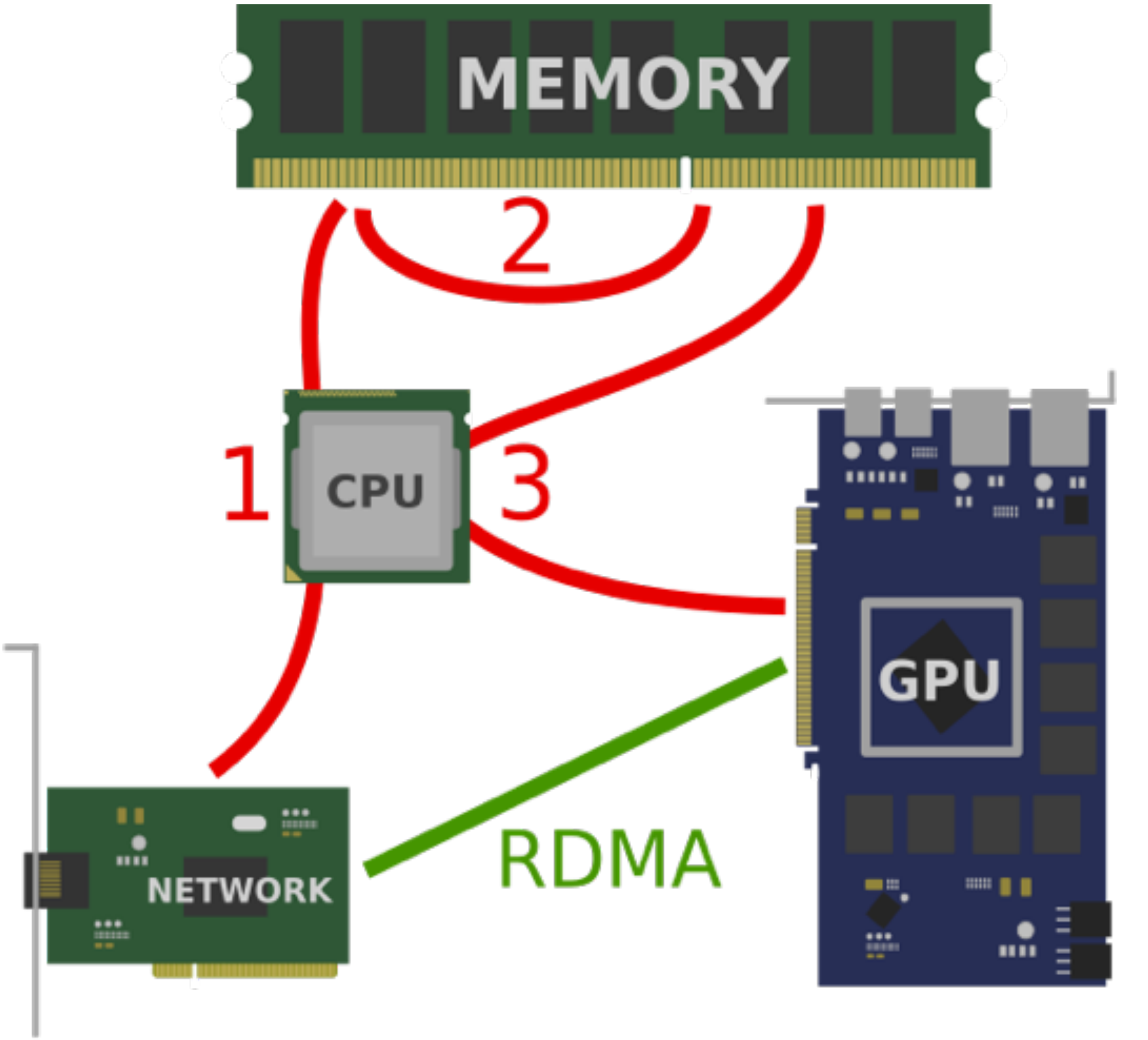
BENCHMARK KERNEL RUNTIME – SPINNING



CURRENTLY NOT POSSIBLE IN OPENCL: CACHE CAN'T BE FLUSHED FROM KERNEL

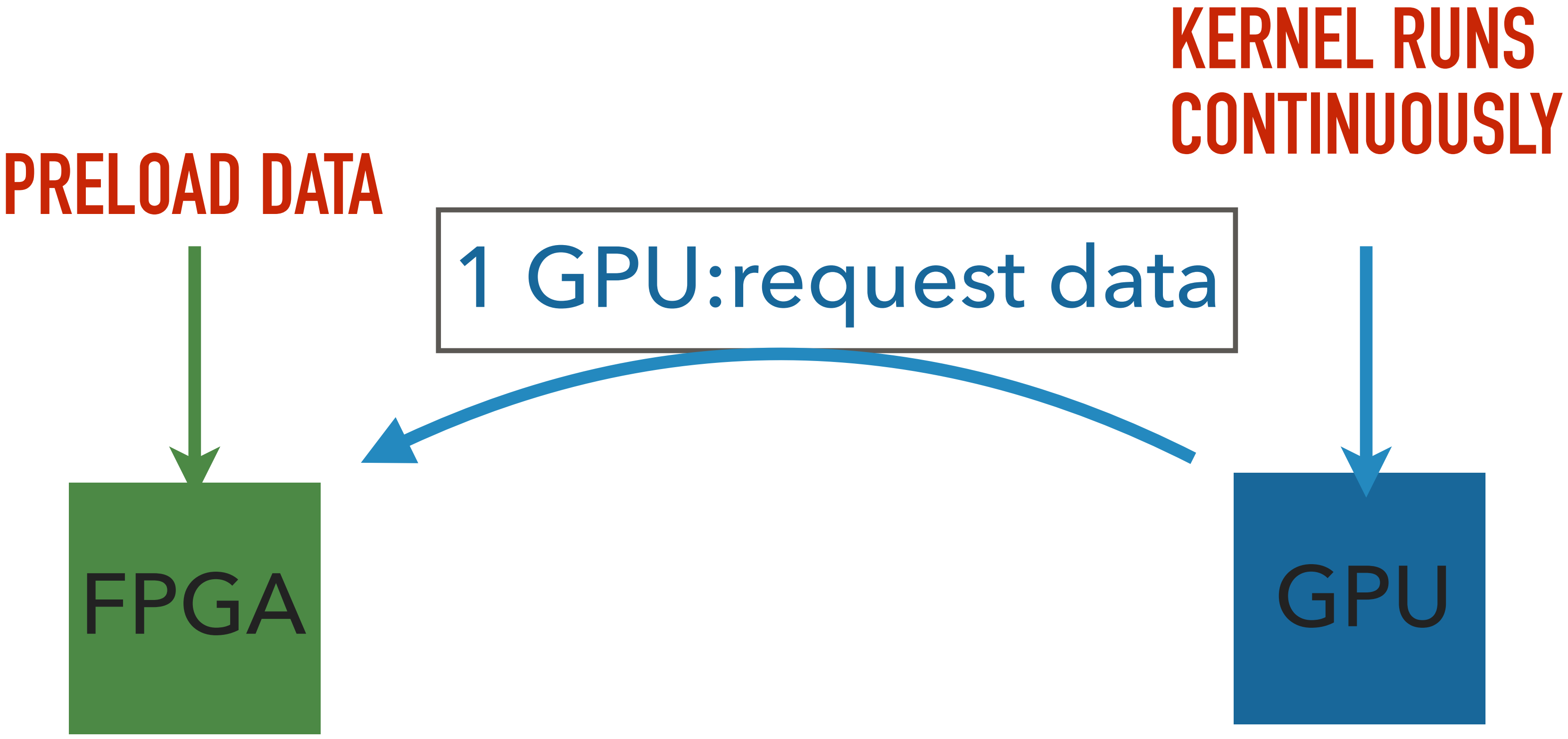
WHAT ABOUT DATA TRANSFER?

DMA SETUP – CPU ONLY STARTS THE KERNEL



(Red) Conventional transfer (Green) RDMA transfer

DMA MEASUREMENT SETUP



KERNEL RUNS CONTINUOUSLY

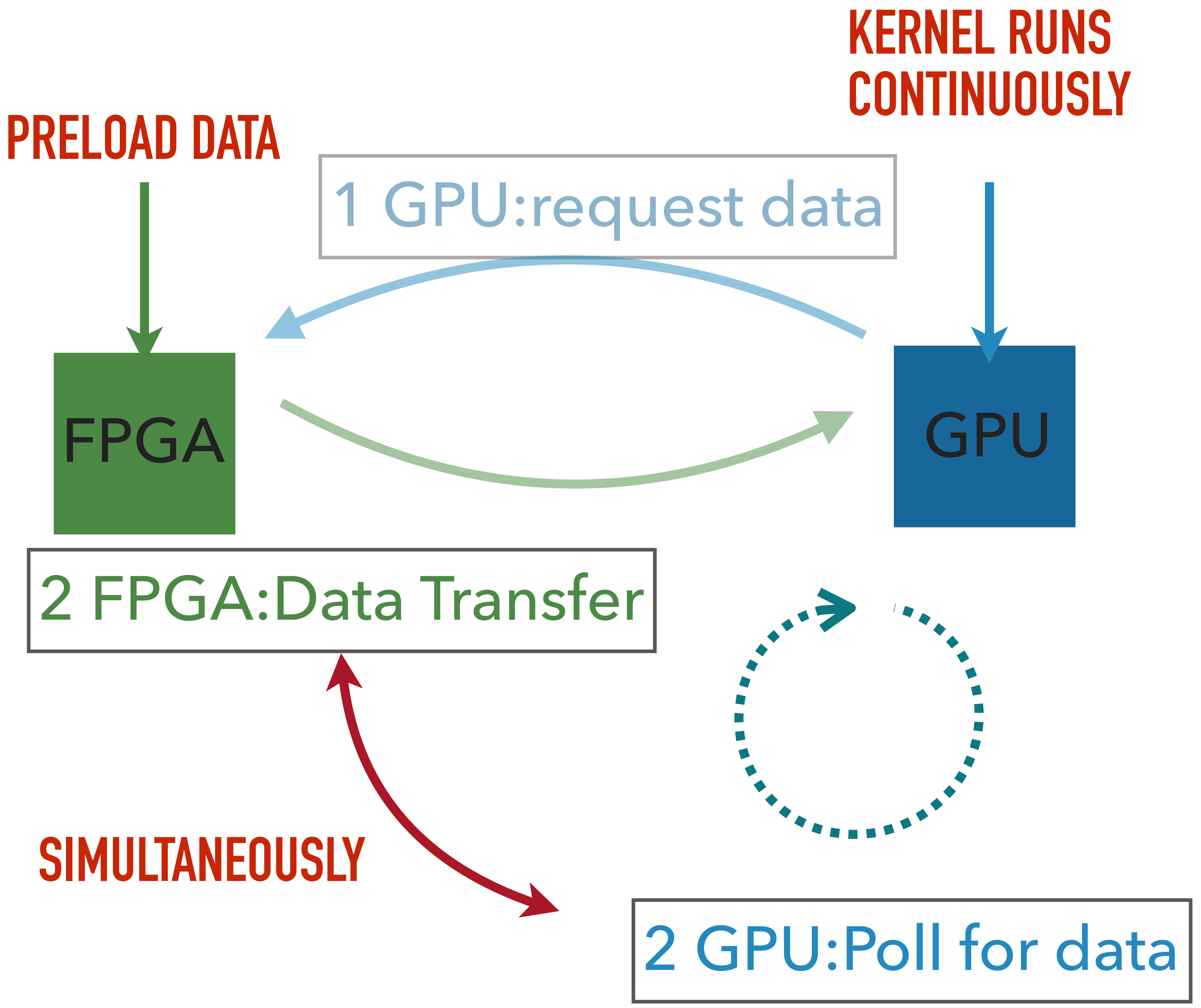
PRELOAD DATA

1 GPU:request data

FPGA

GPU

DMA MEASUREMENT SETUP



KERNEL RUNS CONTINUOUSLY

PRELOAD DATA

1 GPU:request data

FPGA

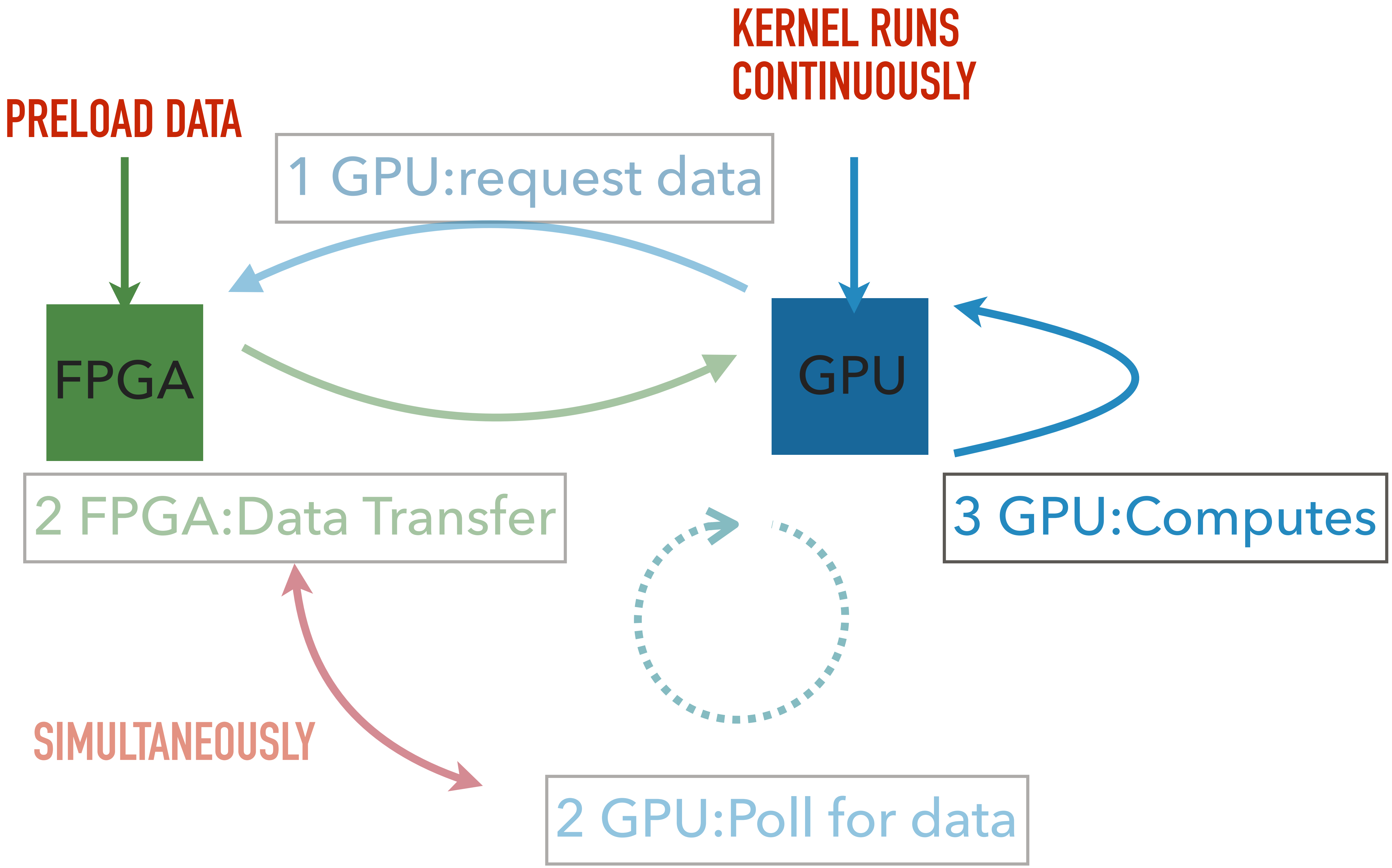
GPU

2 FPGA:Data Transfer

SIMULTANEOUSLY

2 GPU:Poll for data

DMA MEASUREMENT SETUP



KERNEL RUNS CONTINUOUSLY

PRELOAD DATA

1 GPU:request data

FPGA

GPU

2 FPGA:Data Transfer

3 GPU:Computes

SIMULTANEOUSLY

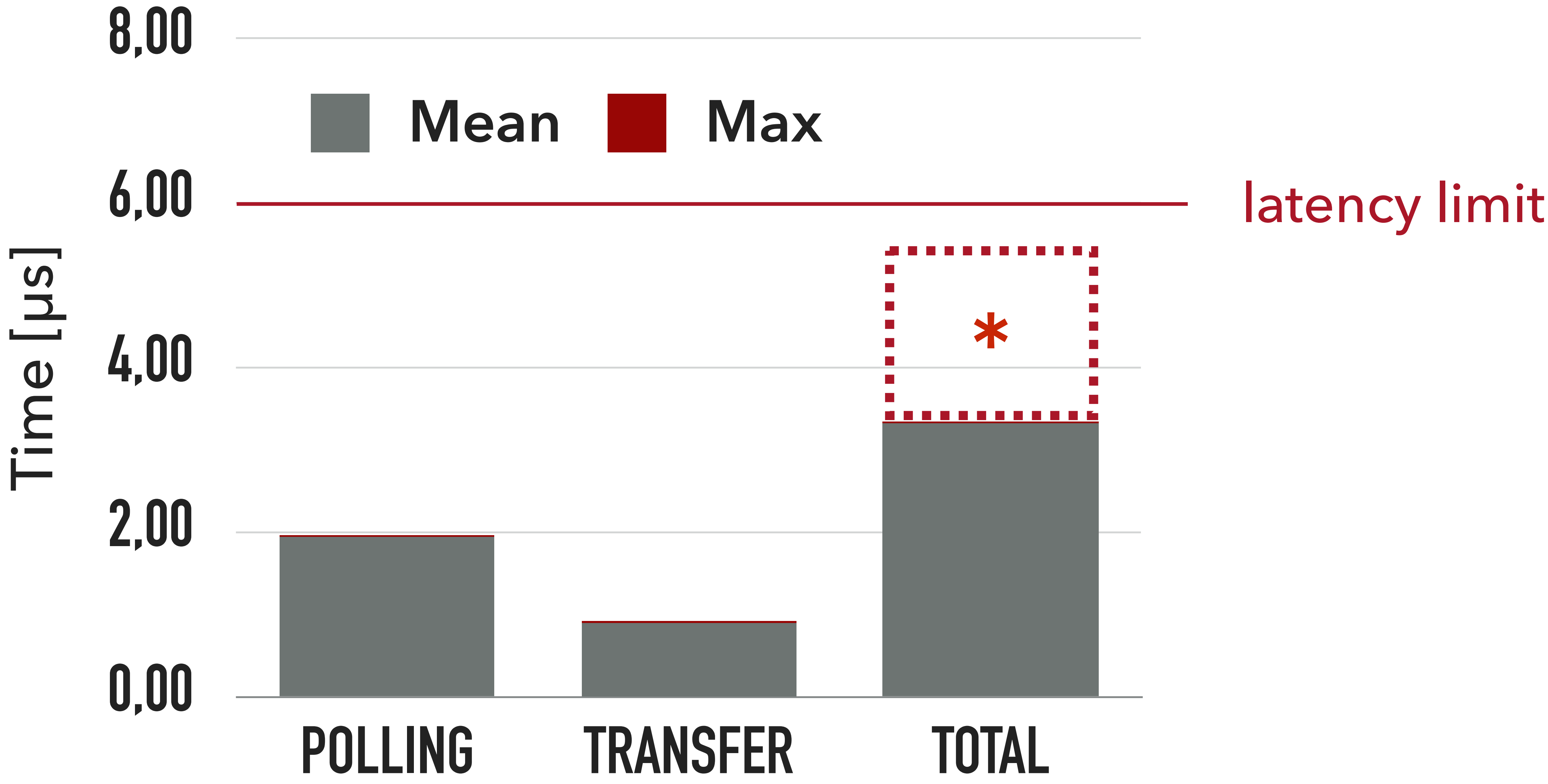
2 GPU:Poll for data

At the moment we don’t write back into the FPGA,

DMA BENCHMARK: POLLING – SPINNING KERNEL

read and write 160 stubs (64 bits each)

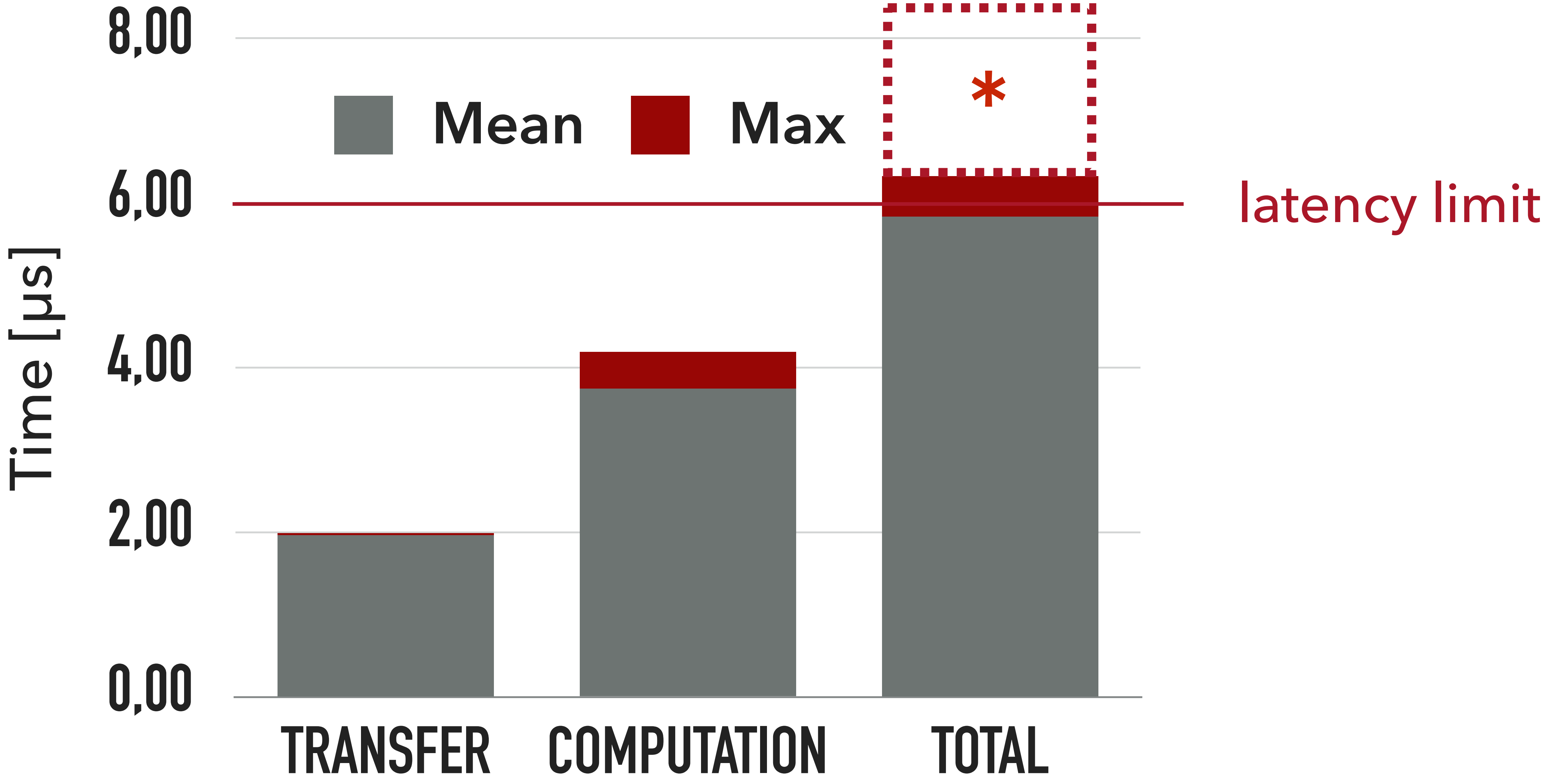
➤ Start transfer ➤ Poll for data ➤ Write back result



* ···· estimated response time

DMA BENCHMARK – HOUGH TRANSFORMATION

➤ Read/Uncompress data ➤ Compute ➤ Poll



160 stubs, 1 sector

* .... estimated response time

INCREASING THROUGHPUT

- Computation time is higher than data transfer
- We can hide the transfer behind the computation

INTERLEAVED APPROACH

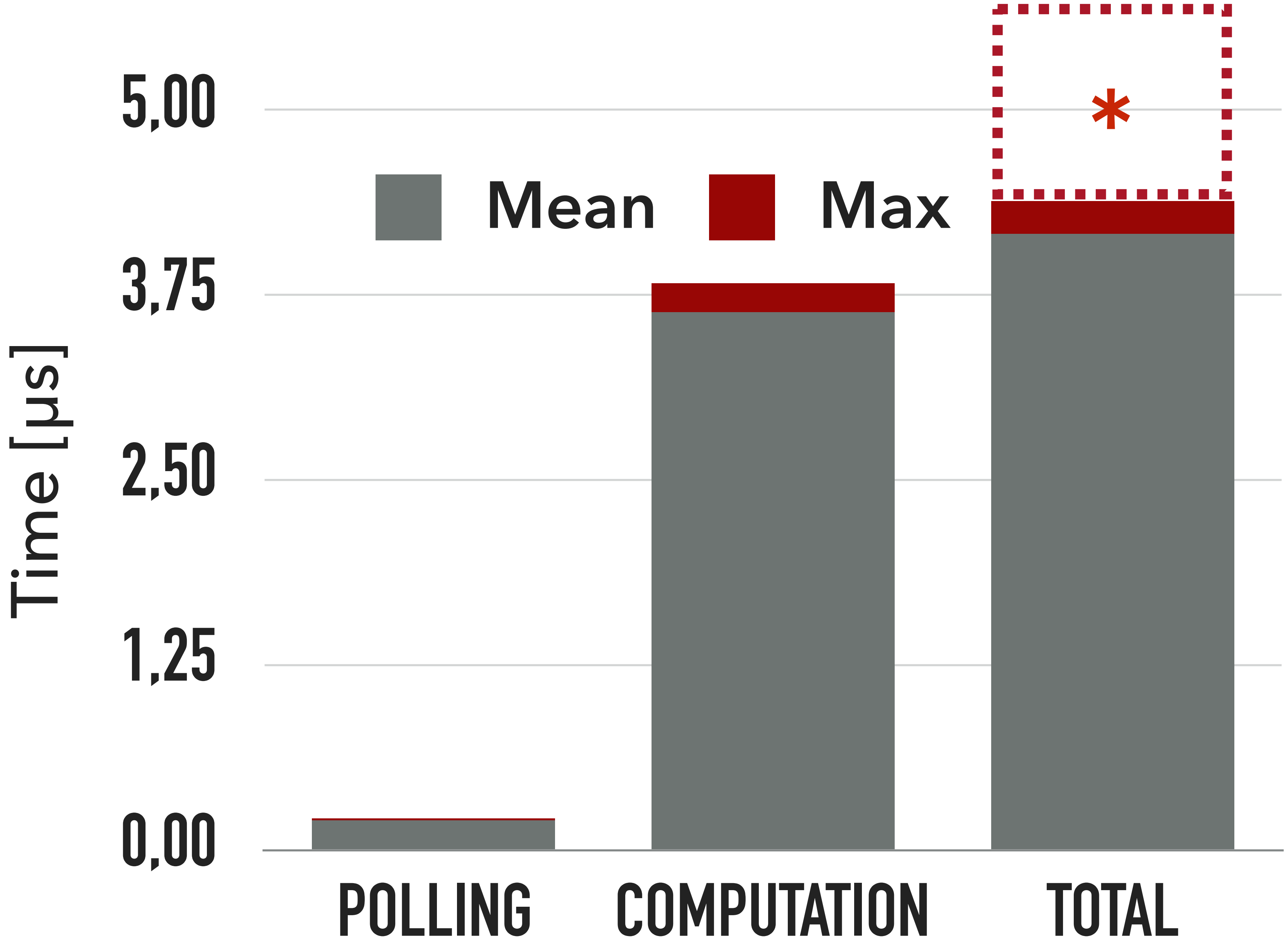
- Start data transfer for current data set
- Do calculations on previous dataset (lies in register memory)
- Poll new data (should take less time)

INCREASES THROUGHPUT AT COST OF LATENCY

DMA BENCHMARK: INTERLEAVED HT

Data older, throughput higher

➤ (poll) Read/Uncompress data ➤ Ask for data ➤ Compute

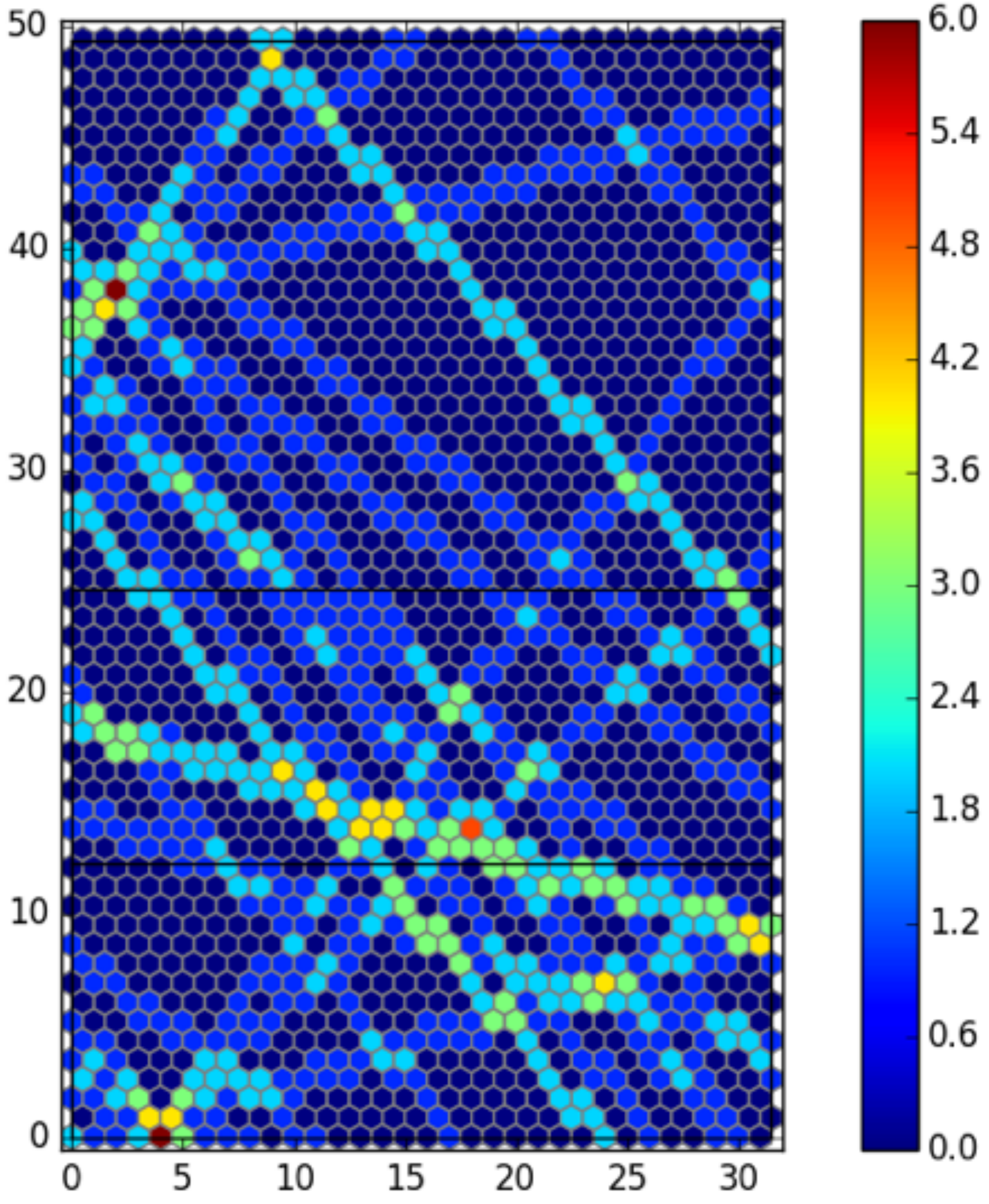


* ···· estimated response time

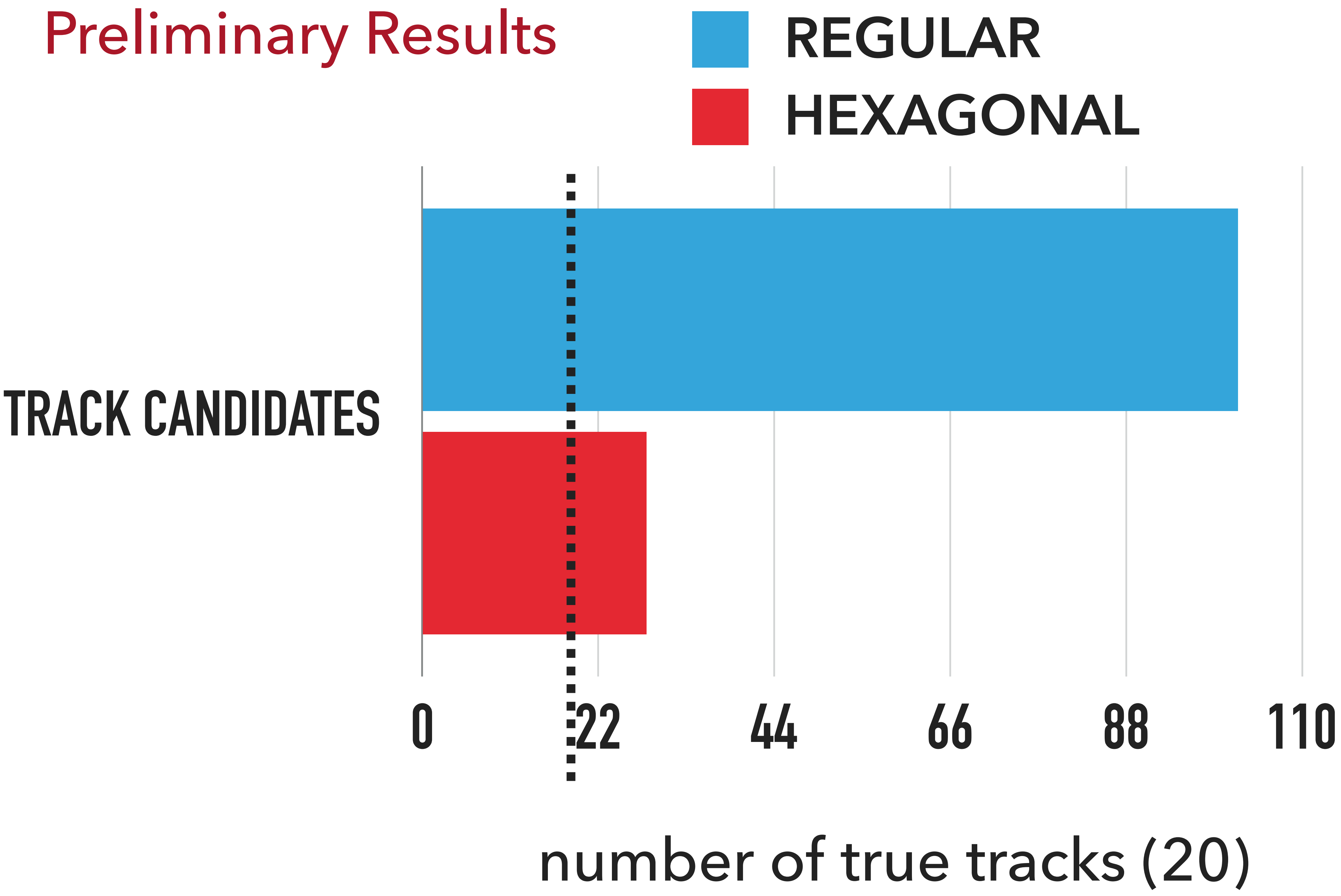
NEW APPROACH - HEXAGONAL HOUGH-SPACE

- Hexagonal bins in hough space
- Suppresses fake candidates
- Runtime comparable
- only 1 possible bin per row
- less algorithmic branching

CAVEAT: NEEDS MORE BINS
(FACTOR OF 2)



COMPARISON OF FAKE RATES – REGULAR VS. HEXAGONAL

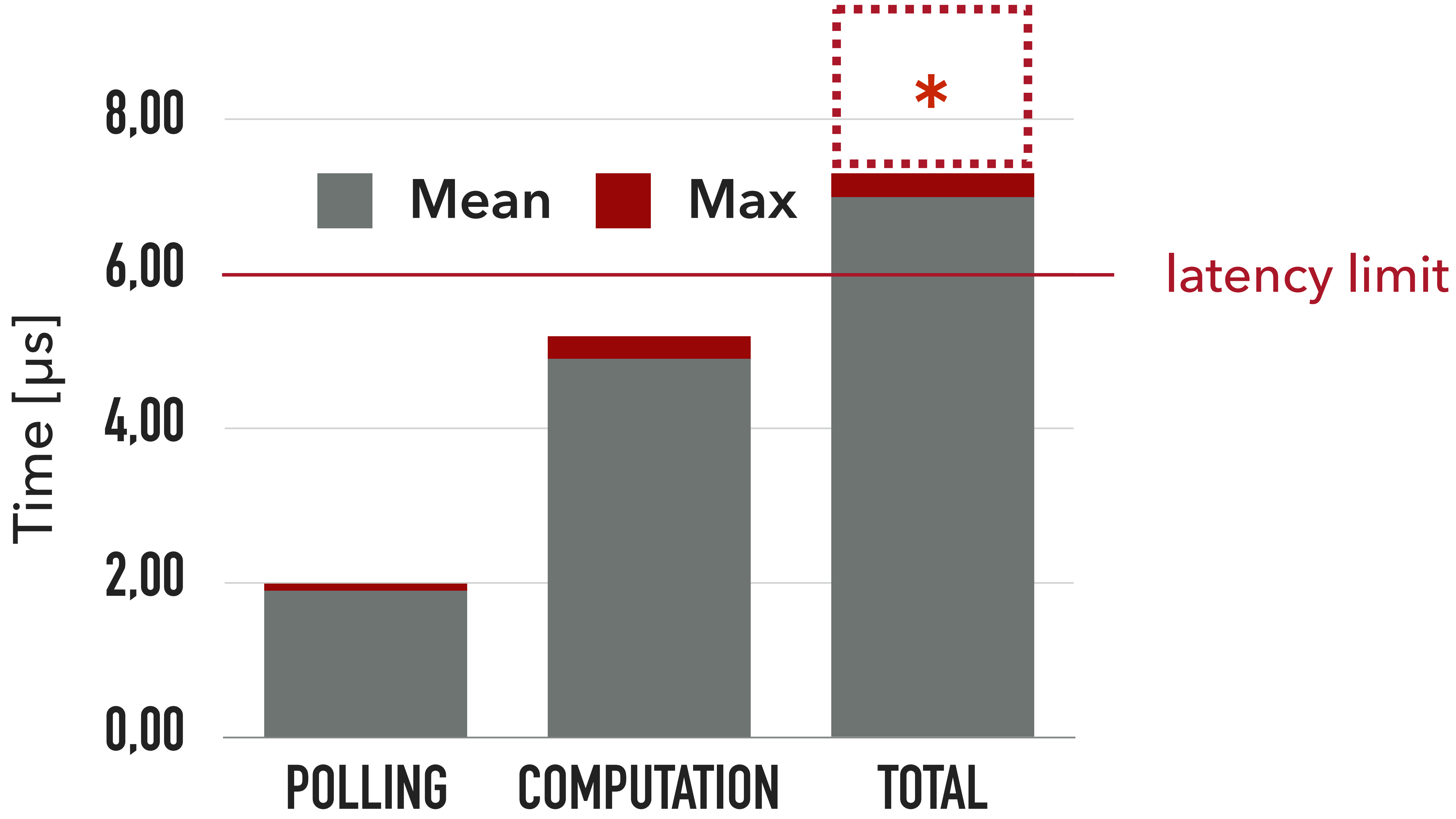


Preliminary Results

Results for TTBar Dataset PU140, whole detector, 1 event

DMA BENCHMARK: HEXAGONAL HT

- (poll) Read/Uncompress data ➤ Ask for data ➤ Compute



* ···· estimated response time

CONCLUSIONS

Performance:

- Computational time of around 4 μs
- Transfer time of around 2 μs

Surpassed our expectations

Development is faster

More complex algorithms are possible:

- Example: hexagonal approach

Data transfer using standard interfaces is challenging

OUTLOOK

- Need to process multiple sectors per card in future
- Look at performance of newer cards
  - High Bandwidth Memory,

    already in consumer model cards,

    promises 2-4x better throughput
- Investigate new transfer technologies
  - PCIe 4.0 (2x faster)
  - nv-link (5-10x faster)

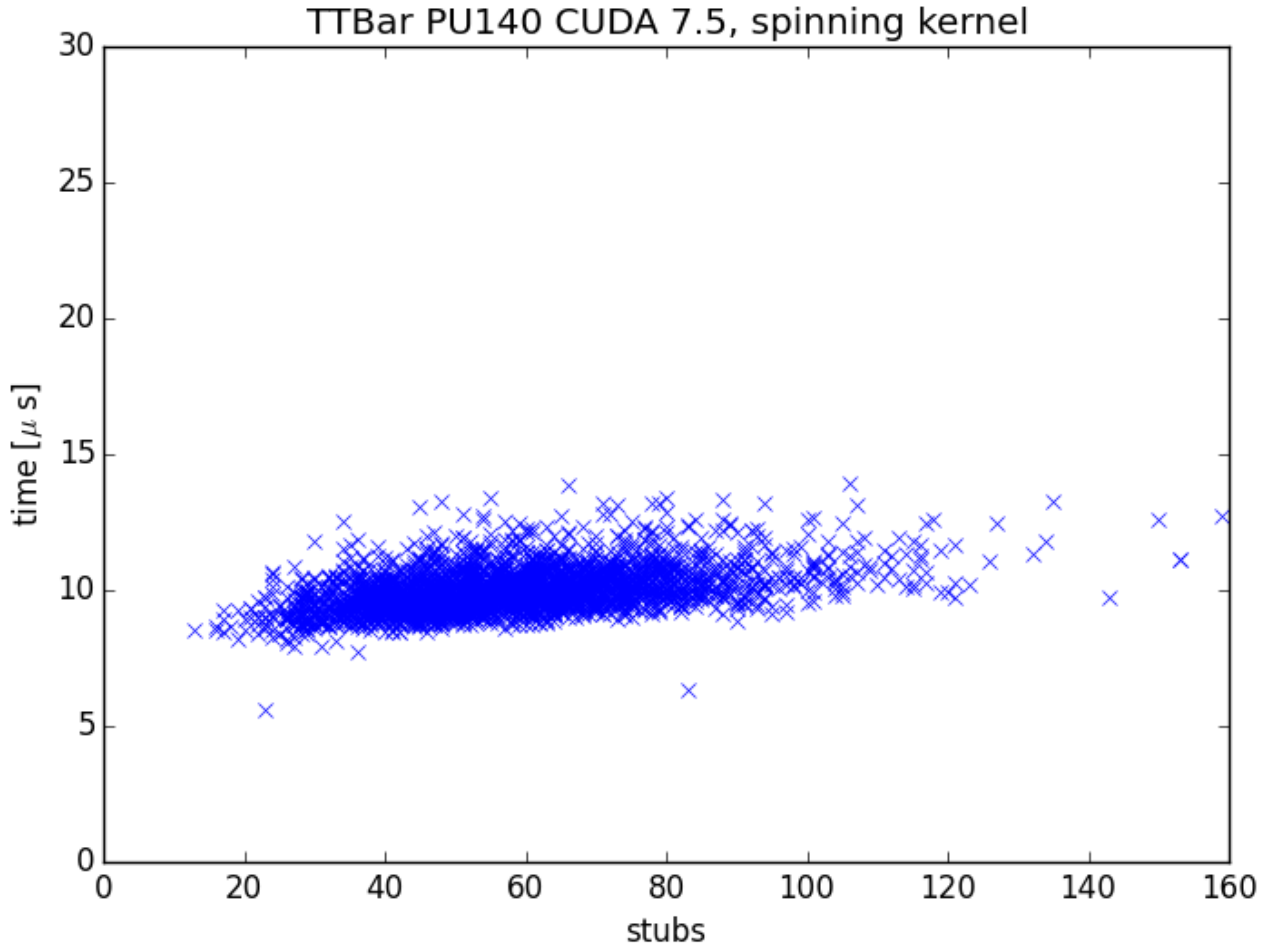
MOORS LAW IS OUR FRIEND!

THANK YOU!

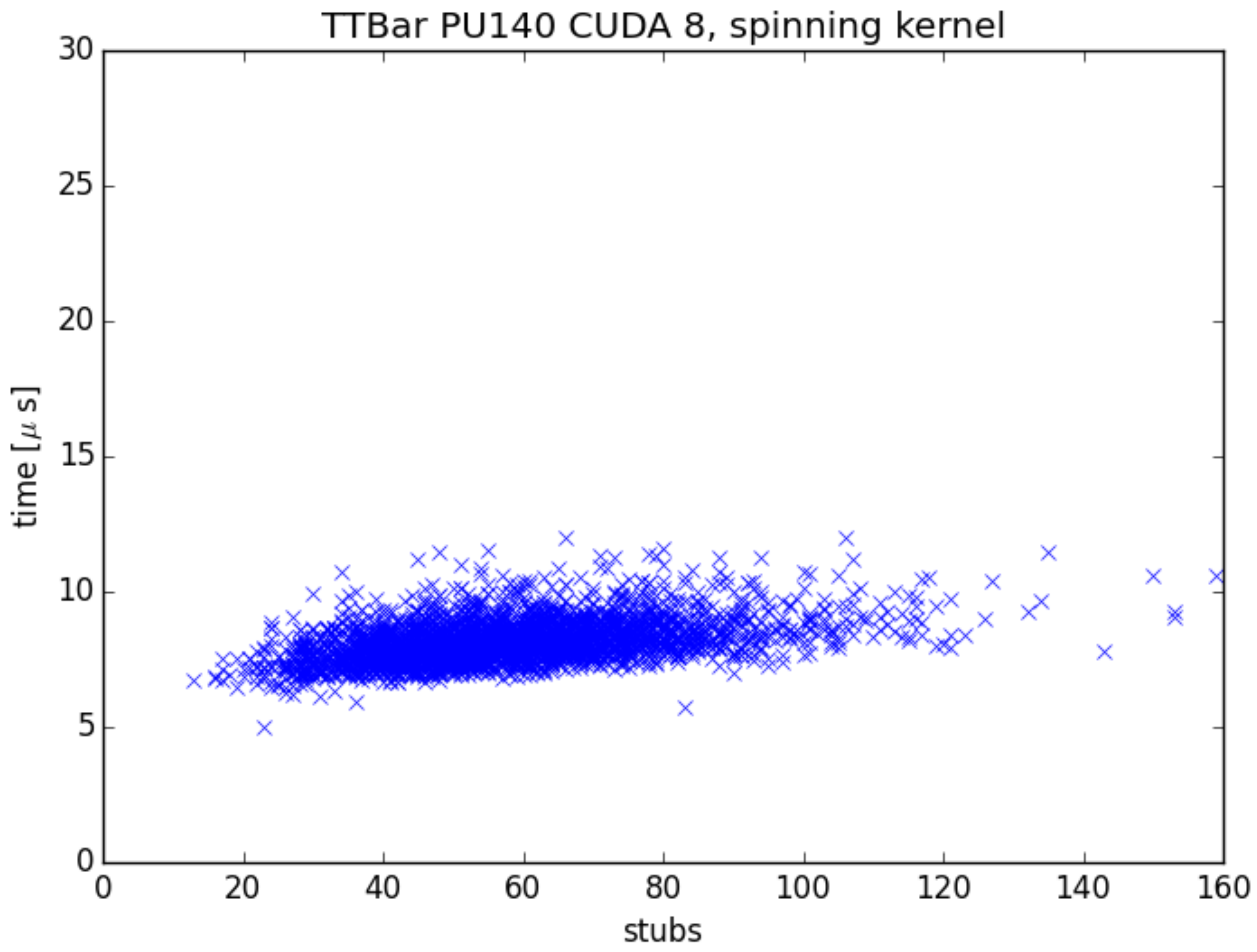
QUESTIONS?

Karlsruher Institut für Technologie

CUDA 7.5 – TESLA K40C



CUDA 8 - TESLA K40C


TTBar PU140 CUDA 8, spinning kernel
time [μ s]
stubs