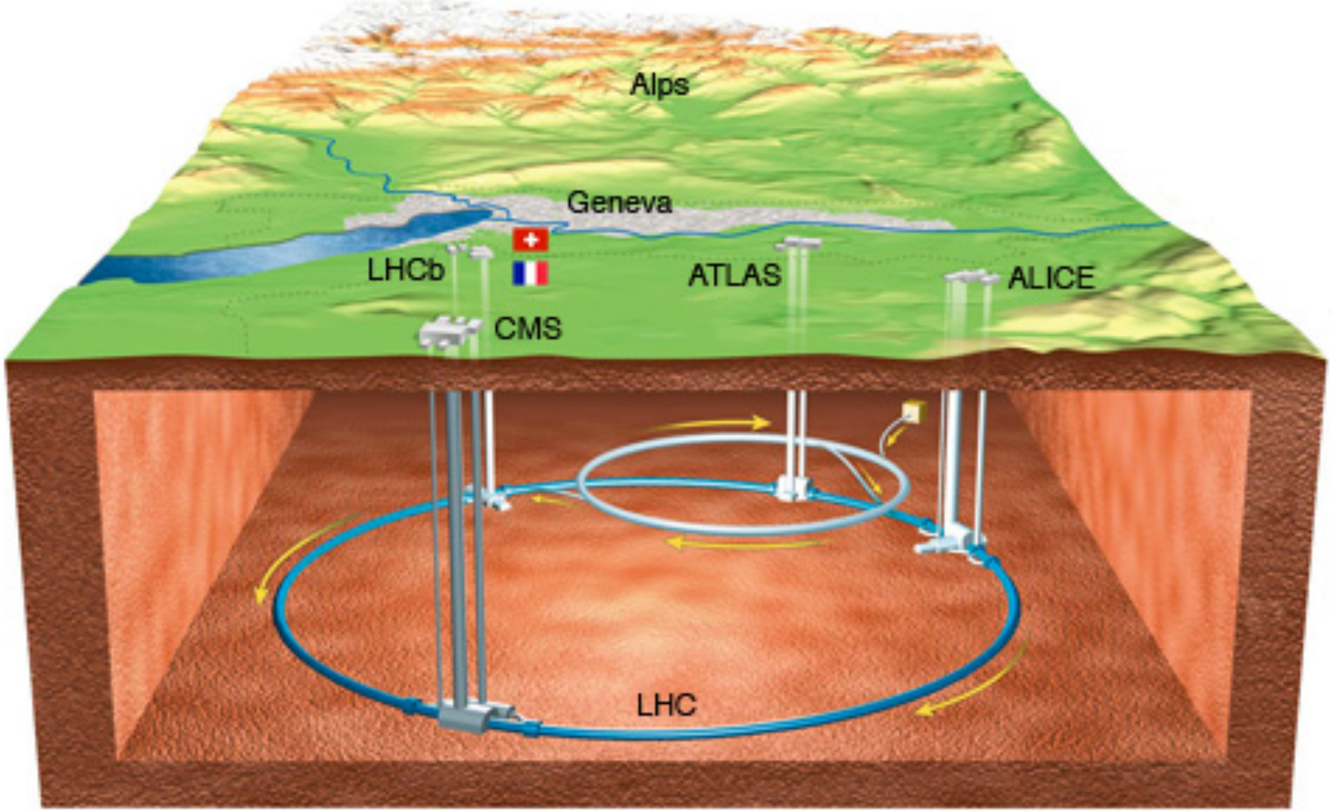
High-Throughput Computing Collaboration:
An overview and future directions

28.2.2017

Omar Awile (omar.awile@cern.ch),

Background image: Shutterstock

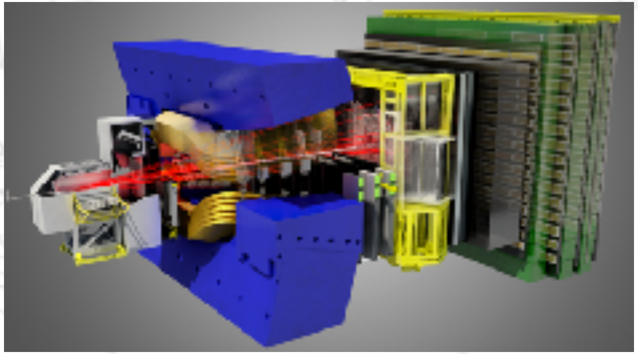
The Large Hadron Collider

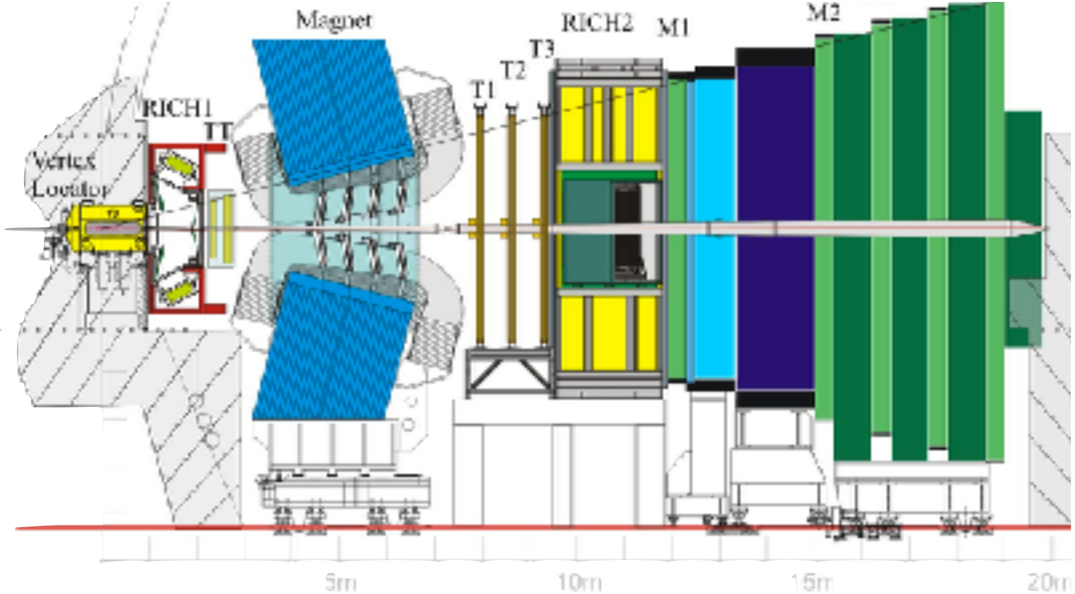


Background image: Shutterstock

The LHCb experiment

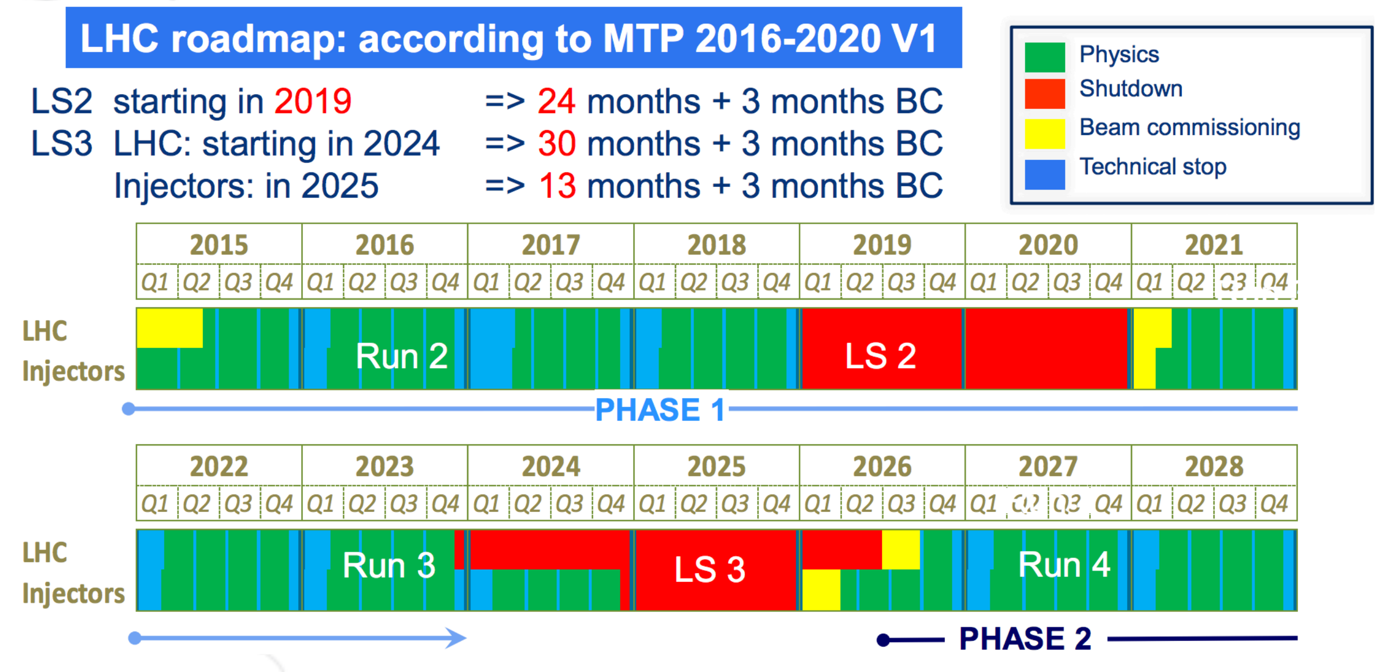
- Specialized “b-physics” experiment
  - Helps us understand what happened after big bang that allowed matter to survive (leading to… us here)





Background image: Shutterstock

LHC long term planning



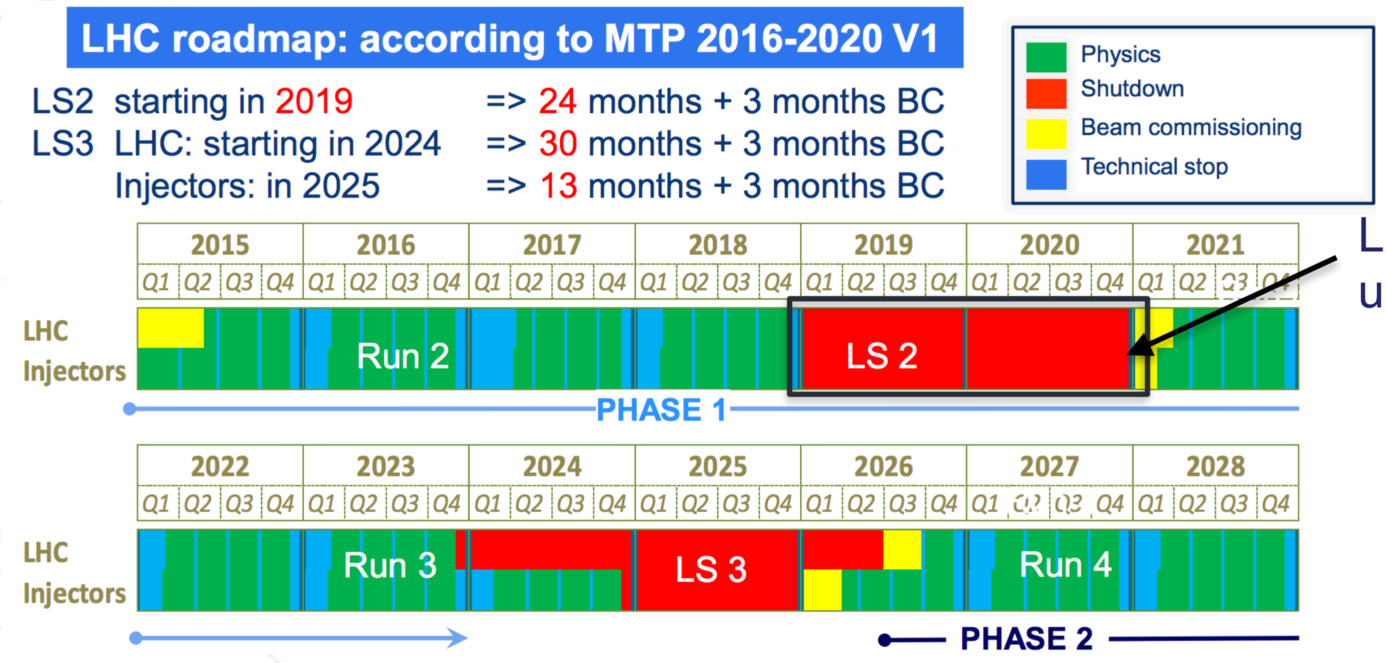
LHC roadmap: according to MTP 2016-2020 V1

LS2 starting in 2019 => 24 months + 3 months BC
LS3 LHC: starting in 2024 => 30 months + 3 months BC
Injectors: in 2025 => 13 months + 3 months BC

🟩 Physics
🟥 Shutdown
🟨 Beam commissioning
🟦 Technical stop

Background image: Shutterstock

LHC long term planning



LHC roadmap: according to MTP 2016-2020 V1

LS2 starting in 2019 => 24 months + 3 months BC
LS3 LHC: starting in 2024 => 30 months + 3 months BC
Injectors: in 2025 => 13 months + 3 months BC

🟩 Physics
🟥 Shutdown
🟨 Beam commissioning
🟦 Technical stop

LHCb upgrade

Background image: Shutterstock

HTCC in a nutshell

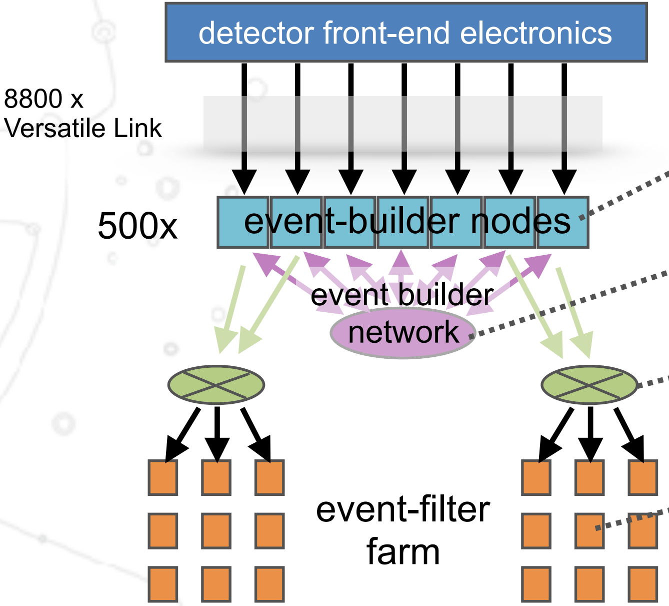
- High-Throughput Computing Collaboration:

LHCb
intel
CERN
IT Department

- Apply upcoming Intel technologies in an “Online” computing context at the Large Hadron Collider
  - Data Acquisition (DAQ) and event-building
  - Accelerator-assisted decision-taking on collected data
- Use LHCb upgrade as an example, but applicable and useful for other experiments too!

Background image: Shutterstock

LHCb TDAQ Architecture Using Intel



Intel® Xeon®
Intel® Xeon® + FPGA w/
Intel ® Omni-Path
Intel® Intel® Xeon Phi™ &
Intel® Omni-Path

Intel® Omni-Path

Intel® Omni-Path and/or
100 GbE

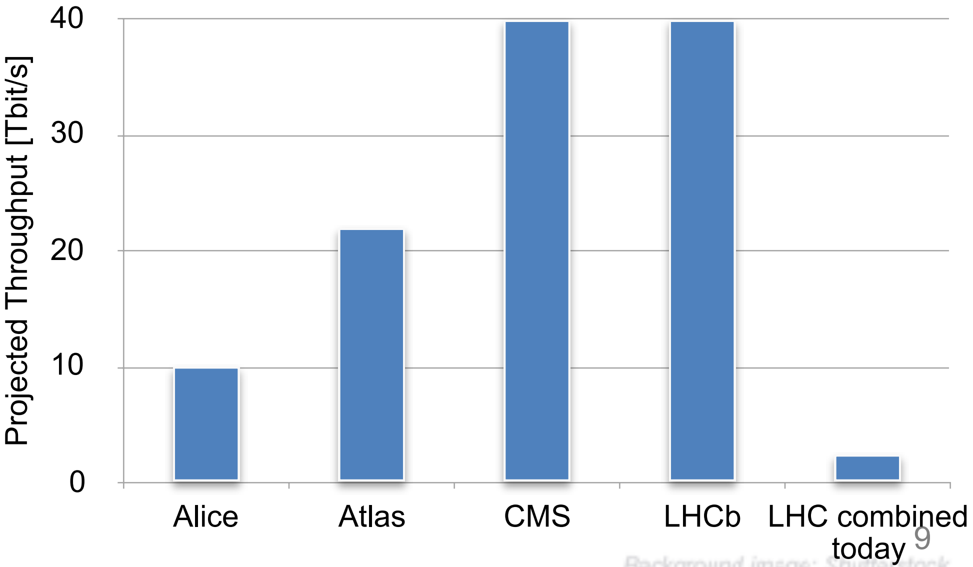
Intel® Xeon®
Intel Xeon® + FPGA
Intel® Xeon Phi™
3D XPoint™

Background image: Shutterstock

The data-acquisition network

DAQ Challenges

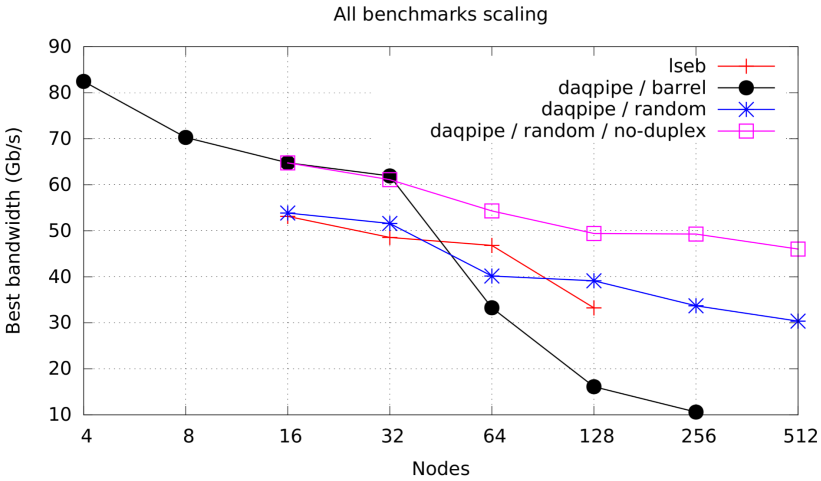
- Transport multiple Terabit/s reliably and cost-effectively
- 500 port full duplex, full bi-sectional bandwidth network, aiming at 80% sustained link-load @ >= 100 Gbit/s / link
- Integrate the network closely and efficiently with compute resources
- Multiple network technologies should seamlessly co-exist in the same integrated fabric



Background image: Shutterstock

How to evaluate a 40 Tbit/s interconnect (without buying one)

- Many high-bandwidth, low-latency interconnects exist already! specifically in HPC systems (see Top500)
- **DAQPIPE** is a highly portable software package for emulating data-acquisition systems on an HPC site
  - Supports multiple protocols and network technologies
  - Allows one to scan for many relevant parameters (message size/rate, buffers, push/pull, scheduling etc…)

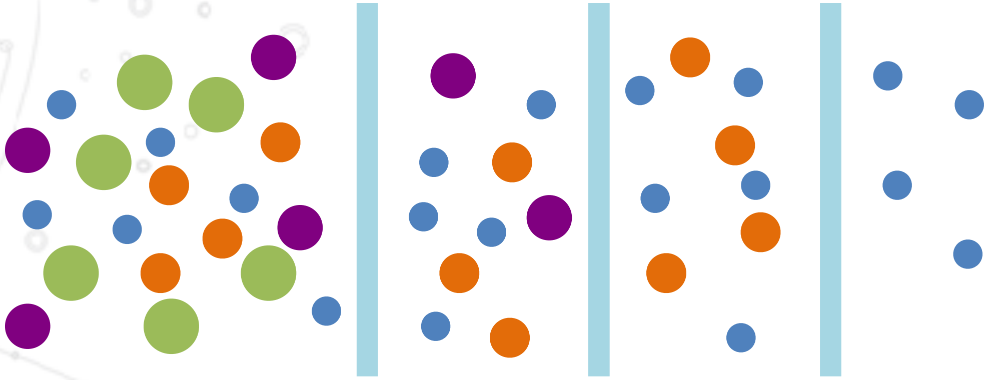


Data processing and filtering

Background image: Shutterstock

How to deal with 40 Tbit/s of incoming data

- Not all particle collision data coming from the detector is relevant.
    - In fact we throw away most of the data!
- How do we choose which data to keep and which to discard?



Background image: Shutterstock

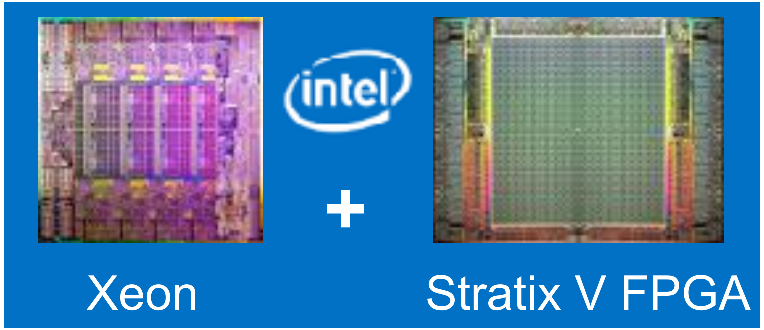
Accelerating the software filtering

- 5 million lines of C++ code
- Three Intel technologies:
  - Intel® Xeon®: Baseline; code offers room for a lot of optimizations to take advantage of modern Intel hardware platforms!
  - Intel Xeon® + FPGA: Code contains many computationally expensive algorithms well suited for FPGA!
  - Intel® Xeon Phi™: Much of incoming data is independent. Massive parallelization is possible!

Background image: Shutterstock

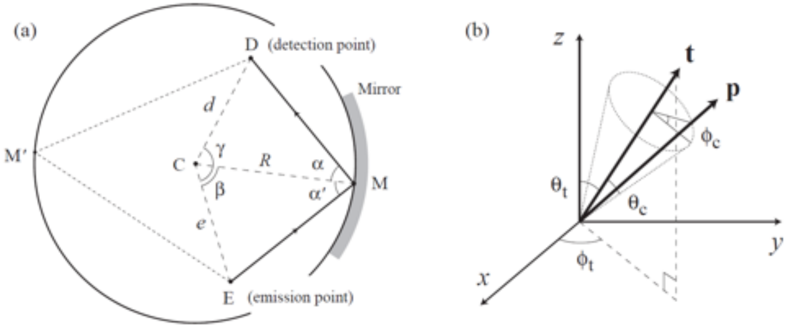
New (and old) challenges on FPGA

- Sophisticated algorithms need more time, bigger FPGAs, more data
- Long-term maintenance issues with custom hardware and low-level firmware
  - Upgrades usually mean replacing all the hardware
- Exact reproducibility of results without the custom hardware challenging and/or computationally intensive
- HTCC question: Can we build something similar with the integrated FPGA on Xeon® Platform?



Xeon-FPGA for filtering

- We developed one important prototype application (RICH PID) to understand the advantages and drawbacks of Xeon-FPGA
    - Achieved a significant speedup (up to x28)
    - Fast development time of code thanks to convenient tools and languages (OpenCL)
    - Low power consumption compared with e.g. GPGPUs
- Conclusion: Using FPGA technology *could* allow us to use more precise algorithms more often for higher-quality data filtering!



Background image: Shutterstock

HTCC results

- Scaling results for Omni-Path look very promising:
  - 15 TB/s aggregate bandwidth on 512 nodes!
  - Open questions remain on how Omni-Path compares with Infiniband EDR
- Great results on Xeon-FPGA StratixV for RICH particle ID prototype code
  - x35 (x26 using OpenCL implementation) In near future: Xeon + on package Arria10 FPGA.
  - How should the filter farm be designed to take advantage of Xeon-FPGA nodes while keeping purchase costs manageable?
- KNL so far looks like a strong alternative to Xeon.
  - When using AVX512 & effective multi-threading speedups of more than x6 have been shown.
  - How far can HLT software scale on KNL and how can the KNL memory model be effectively used?

Background image: Shutterstock

Openlab VI

Future directions

Background image: Shutterstock

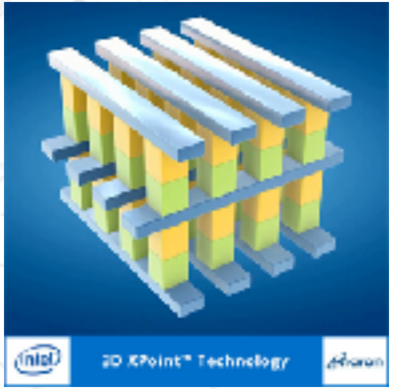
Datacenter needs for ALICE and LHCb

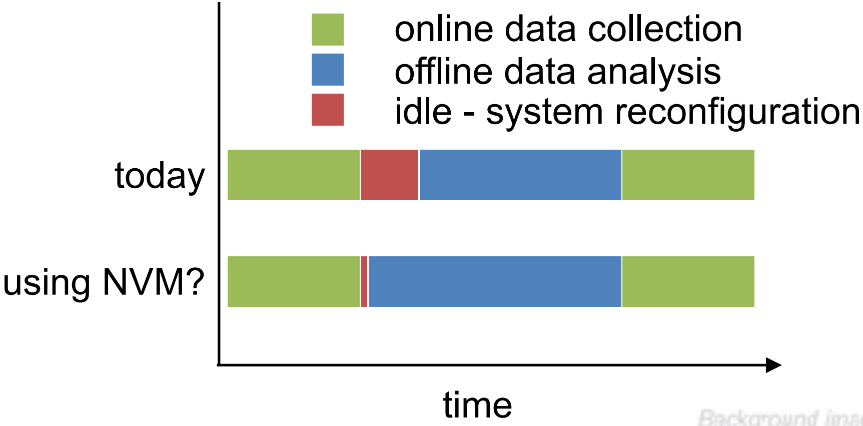
- New datacenter to be constructed (2 MW)
- Require about 2500 4000 U
- Needs to house the computing infrastructure for the software filtering:
  - Network
  - Storage
  - Compute: Xeon + Accelerators (FPGA, GPGPU, KNL?)
- CERN investigates two options: local datacenters close to the experiments or large central DC, requiring long-distance (< 10 km) data-transport

Background image: Shutterstock

Persistency & Non-volatile memory

- Quick turn-around of large, massive applications is crucial for efficient usage of Online farms
- A combination of Operating System Level check-pointing and non-volatile memory could be interesting needs investigations
- Related: some applications could profit from massive amount of memory (time-slice processing in ALICE), to be investigated





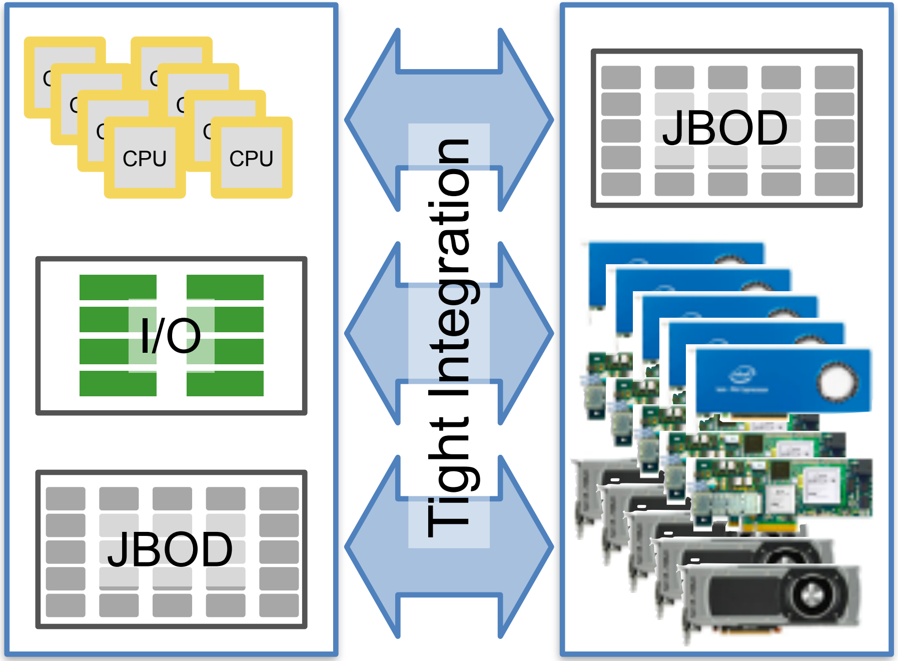
High Efficiency

- LHC up-time integrated over the year is "only" about 30%
- Tight integration of off- and online facilities needed.
- Idle cycles used for analysis, simulation, etc...

Background image: Shutterstock

Flexible infrastructure through RackScale architecture

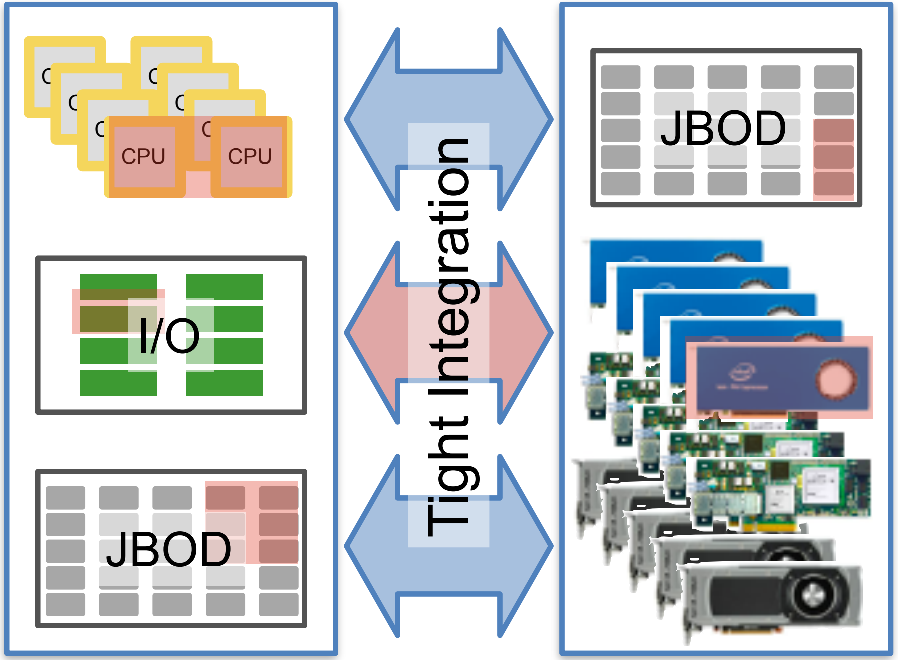
- Ideally could seamlessly run batch-type (simulation, analysis) and (near) real-time workloads.
- easy access to disk-storage
- Housing of custom I/O cards (PCIe)
- Flexible amount of accelerators
- High speed network between (some) servers
- rack-level and datacenter oriented design?



Background image: Shutterstock

Flexible infrastructure through RackScale architecture

- Ideally could seamlessly run batch-type (simulation, analysis) and (near) real-time workloads.
- easy access to disk-storage
- Housing of custom I/O cards (PCIe)
- Flexible amount of accelerators
- High speed network between (some) servers
- rack-level and datacenter oriented design?



Background image: Shutterstock

Thank you!

Who are we:

CERN openlab High Throughput Computing Collaboration

Olof Bärring, Niko Neufeld

Luca Atzori, Omar Awile, Paolo Durante, Christian Färber, Placido Fernandez, Jon Machen (Intel), Rainer Schwemmer, Sébastien Valat, Balázs Vőneki

Background image: Shutterstock

Background image: Shutterstock

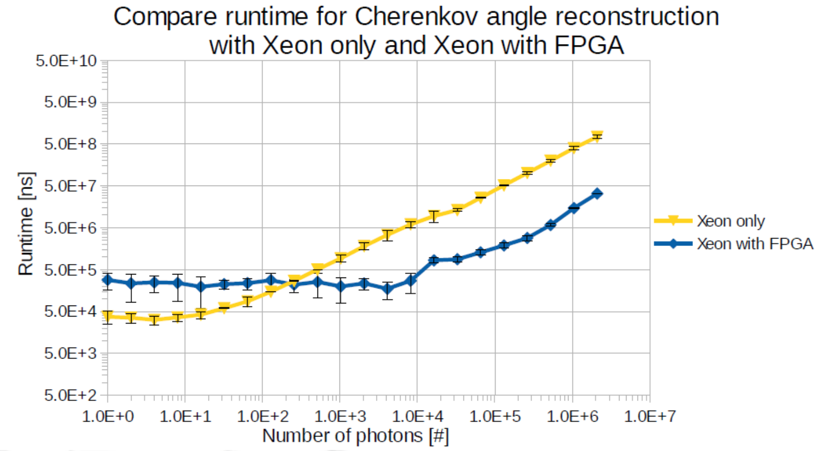
communication and outreach

- Other experiments have expressed interest in Xeon+FPGA and Omni-Path.
  - Discussions are ongoing on DAQ evolution and offline computing using Omni-Path, Xeon+FPGA, and KNL.
  - Potential for using Xeon+FPGA to build a new readout unit with simplified custom hardware.
- Presentations:
  - 09/2016 – CHEP 2016: Acceleration of Cherenkov angle reconstruction with the new Intel Xeon/FPGA compute platform for the particle identification in the LHCb Upgrade
  - 09/2016 – CHEP 2016: LHCb Kalman Filter cross architectures studies
  - 06/2016 - PASC 2016 : Experiments with multi-threaded velopixel track reconstruction
  - 06/2016 - Real Time Conference 2016 : Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade.
  - 06/2016 - Real Time Conference 2016 : Particle identification on an FPGA accelerated compute platform for the LHCb Upgrade.
  - 04/2016 - Open Fabric Alliance workshop 2016 : Building a 4 TB/s event building
  - 04/2015 - ICHEP 2015 : A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications.

Background image: Shutterstock

Cherenkov angle reconstruction on FPGA

- Implementation in Verilog and OpenCL.
  - OpenCL allowed for faster development time (2 weeks vs. 2.5 months) at comparable performance

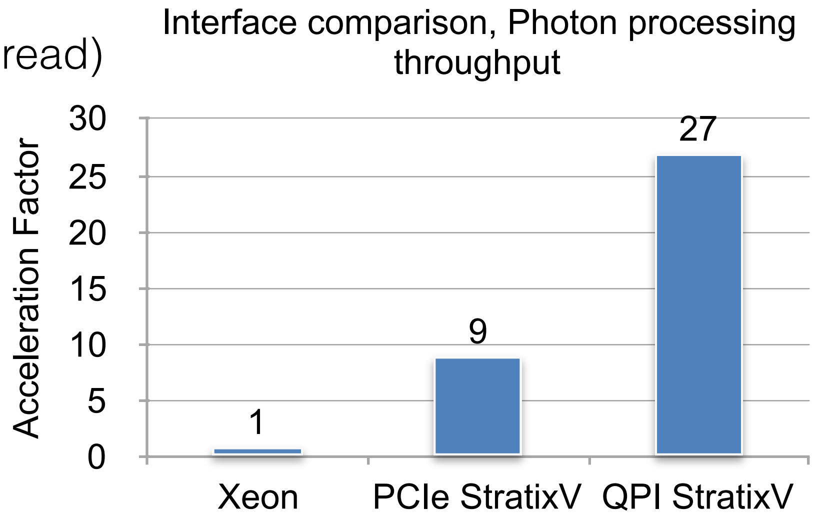


- Acceleration of factor up to 35 (26 using OpenCL) with Intel® Xeon/FPGA
- Theoretical limit of photon pipeline: a factor 64 with respect to single Intel® Xeon® thread
- Bottleneck: Data transfer bandwidth to FPGA

The case for QPI

- StratixV programmed in OpenCL
- Compared to vectorized E2630v2 (single-thread)
- Still room for improvement (pipeline could do at least 2x more)

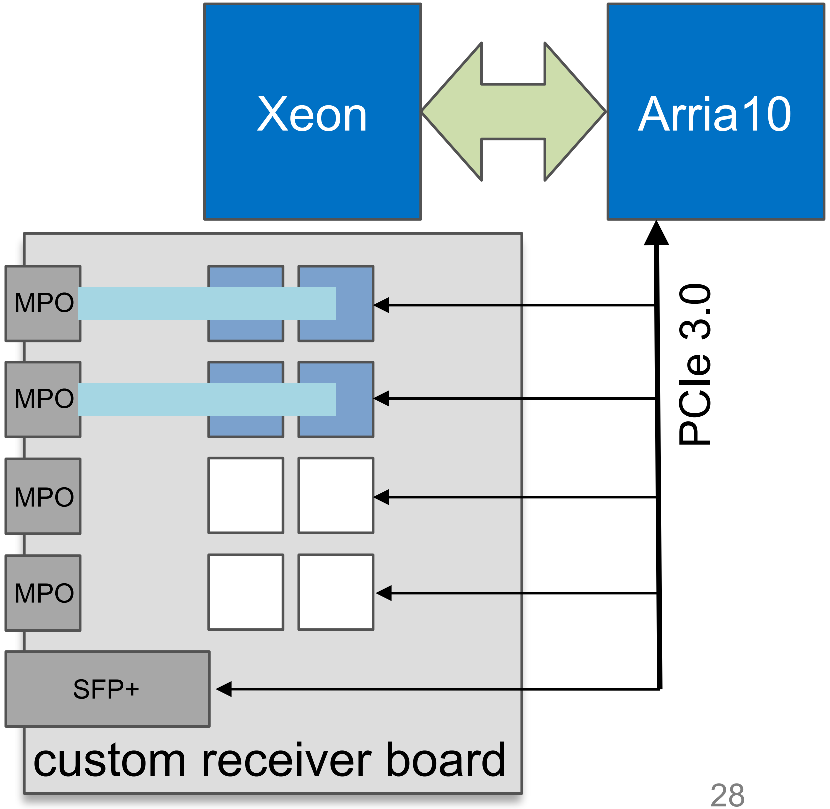
- Currently testing Broadwell+Arria10
  - New interconnect has significantly increased bandwidth
  - Doubled ALMs and registers
  - increased DSPs (implementing hardened FP blocks) by a factor 6
- We expect a further increase of FP performance and ability to implement more (and more complex) algorithms.



Background image: Shutterstock

A simplified PCI-40 card

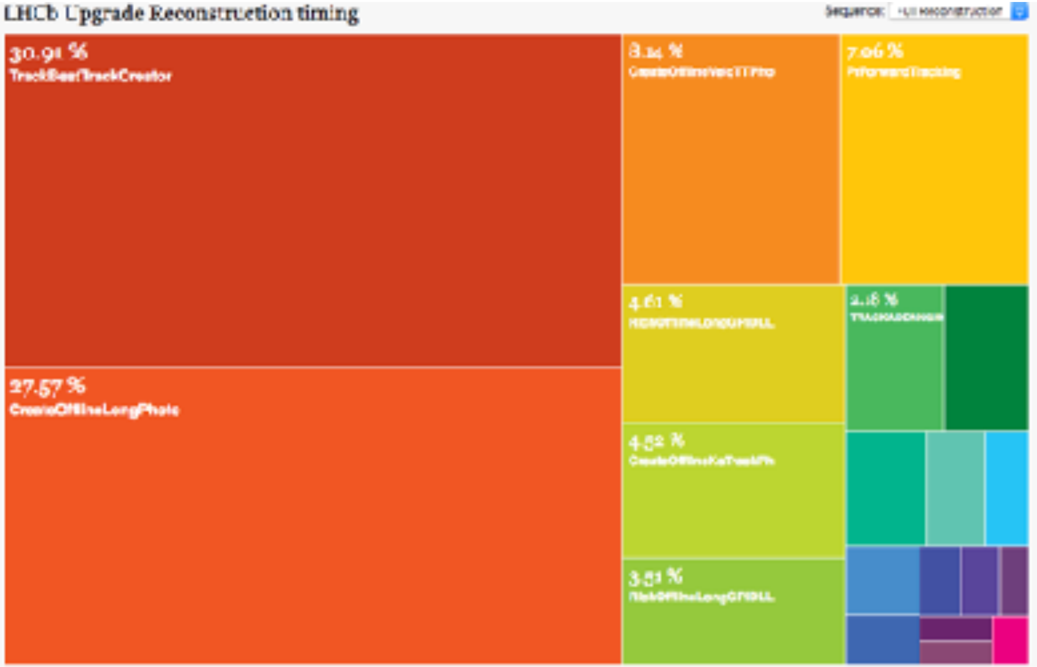
- Simplified receiver card without an FPGA.
- Use Xeon+FPGA for data processing
- Currently being discussed with CMS and Alice.



Background image: Shutterstock

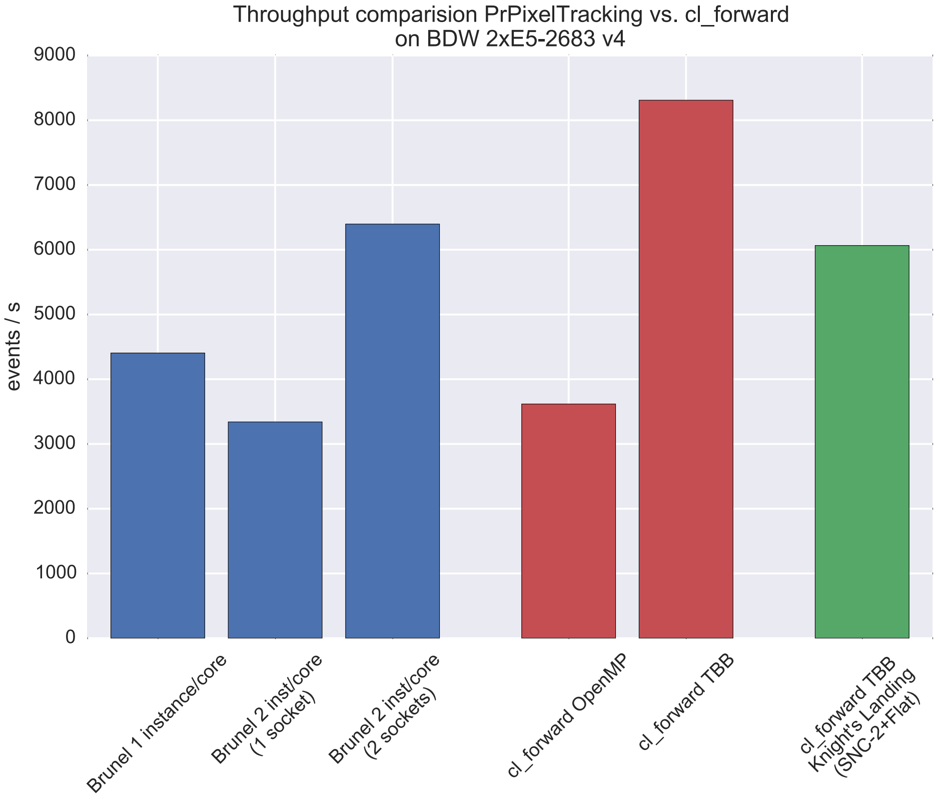
Accelerators for the HLT

- More than 5 MLOCs of C++. Currently under redesign for SIMD and shared-mem parallelism.
- Baseline remains Xeon CPUs
- New framework uses TBB to dispatch algorithms to process events in a multi-threaded fashion.
- Two ways to accelerate algorithms:
    - Offload critical functions to FPGA
    - Rewrite most time-consuming algorithms in a parallel fashion and use Xeon-Phi (Knight’s Landing)



Background image: Shutterstock

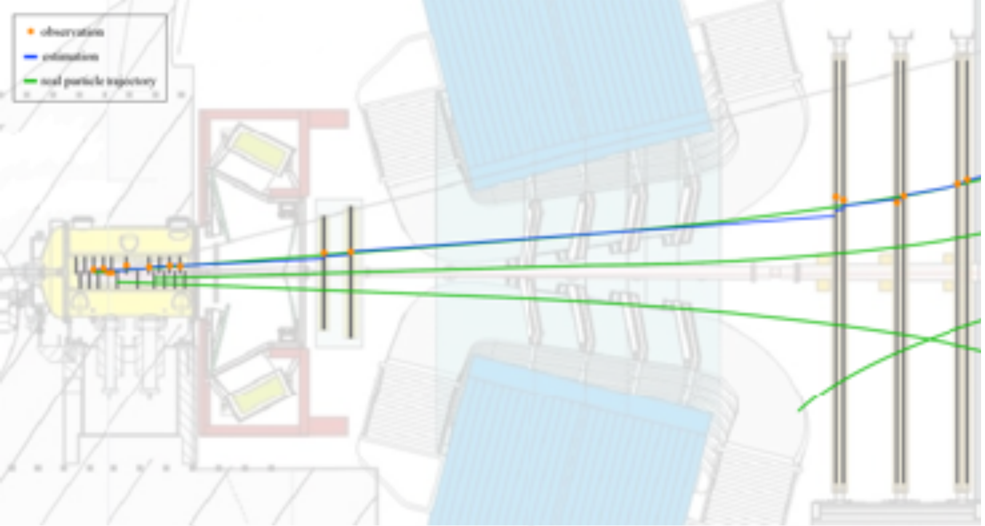
TBB for accelerating track-reconstruction



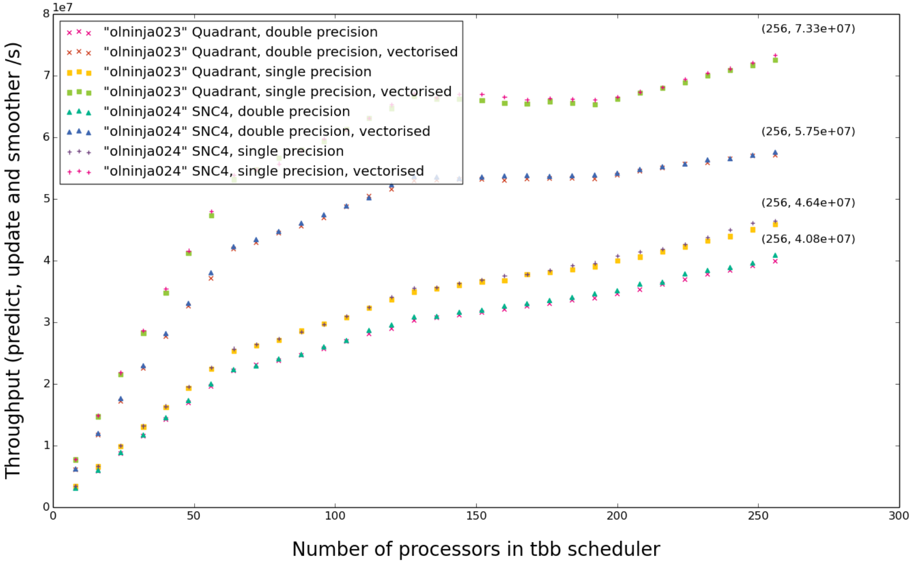
- Straight-line track reconstruction in the velopixel subdetector.
- Comparing TBB with official reconstruction (run in multiple instances without HT)
- tbbPixel speedup on BDW: 1.88
- KNL-specific optimizations will likely yield better throughput!
- Speedup and improved reconstruction efficiency can be ported into production code

SIMD Kalman Filter

- Kalman Filters are a major contributor to overall HLT execution time (~60%).
- Used throughout the HLT for prediction, filtering and smoothing



Scalability of Kalman Filter fit and smoother
on Intel(R) Xeon Phi(TM) 7210 @ 1.30 GHz



Background image: Shutterstock

SIMD Kalman Filter

- Well optimized and vectorized code offers > 6x speedup over production code.
- KNL shows ~2x speedup over 2 socket HSW system.

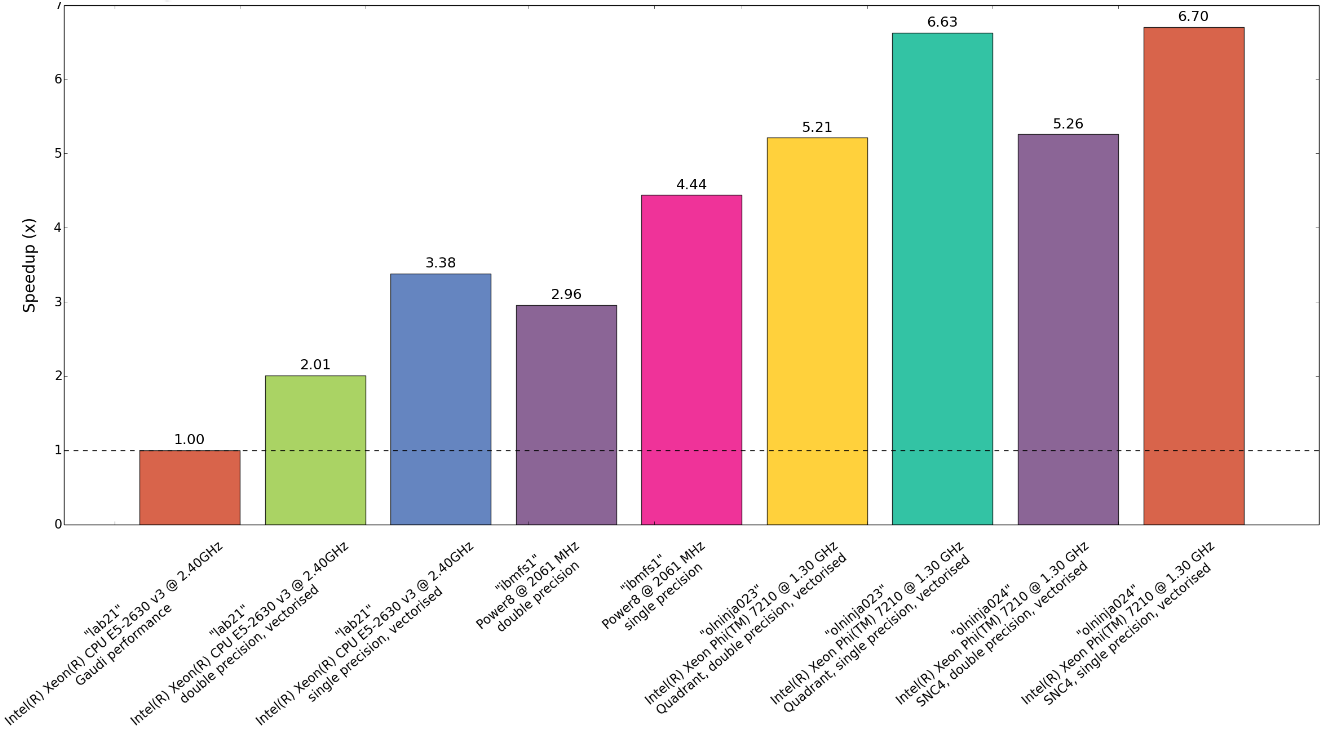


image: Shutterstock