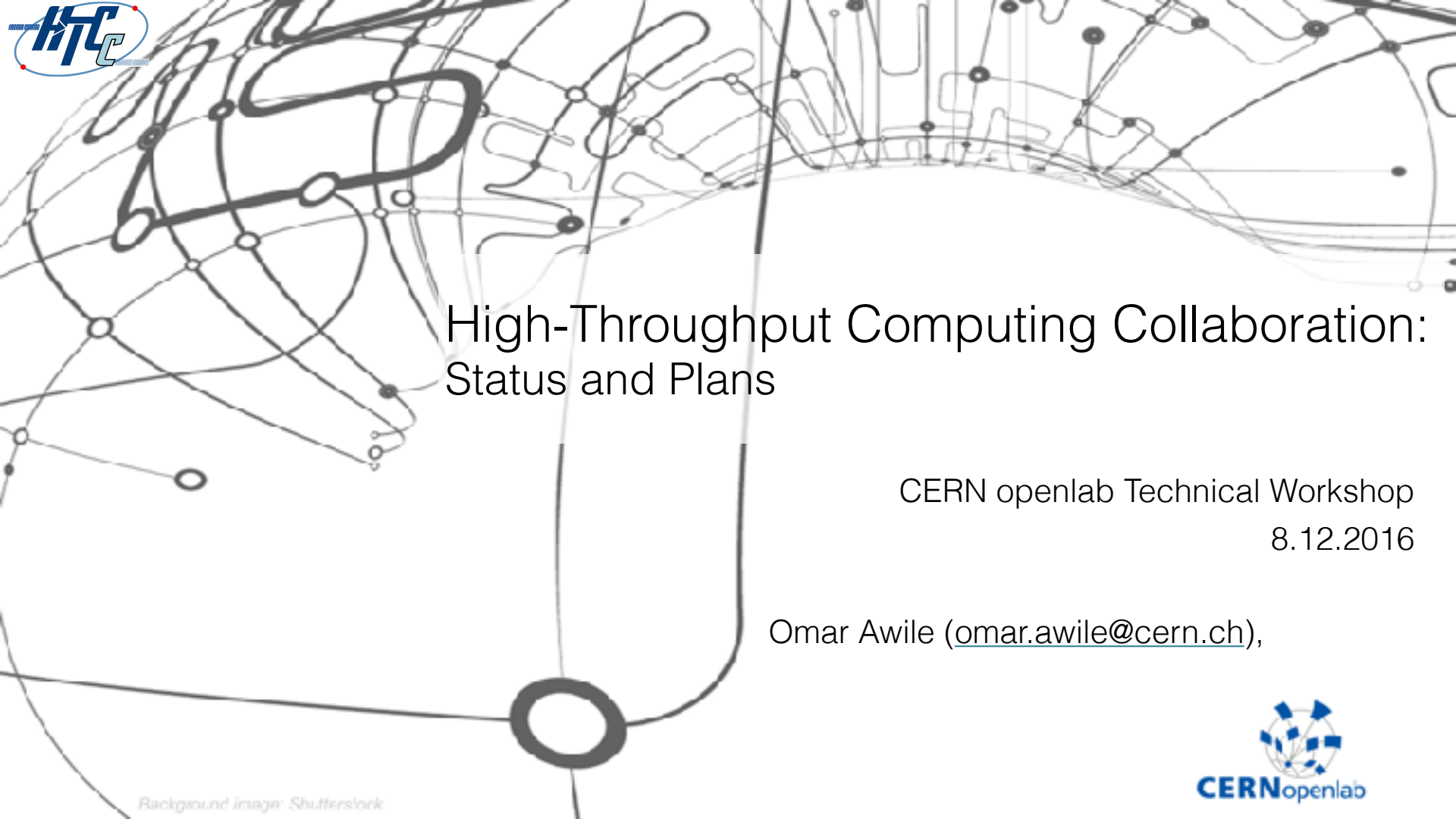


High-Throughput Computing Collaboration:
Status and Plans

CERN openlab Technical Workshop

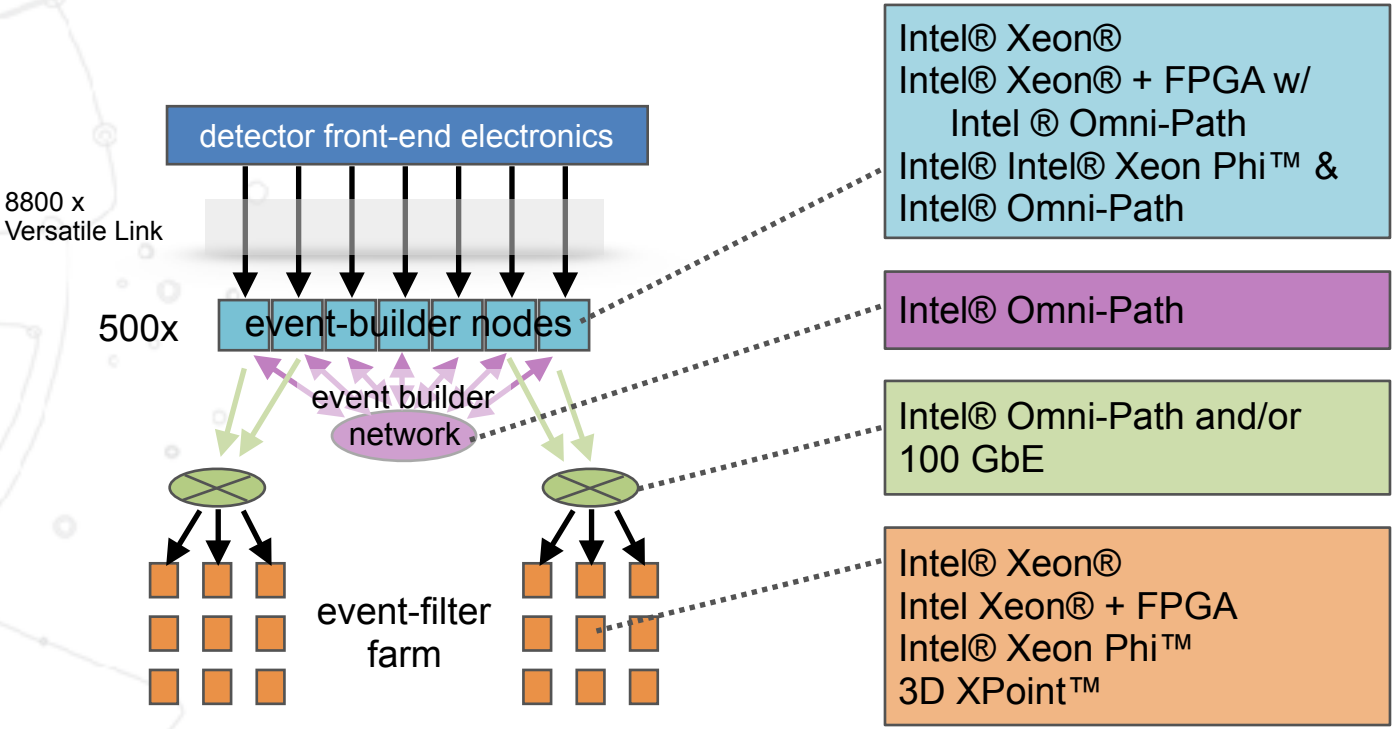
8.12.2016

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

HTCC in a nutshell

- Apply upcoming Intel technologies in an “Online” computing context at the Large Hadron Collider
  - Level 1 Trigger (L1)
  - Data Acquisition (DAQ) and event-building
  - Accelerator-assisted processing for High Level Trigger (HLT)
- Use LHCb as an example, but applicable and useful for all Trigger & DAQ (TDAQ) systems

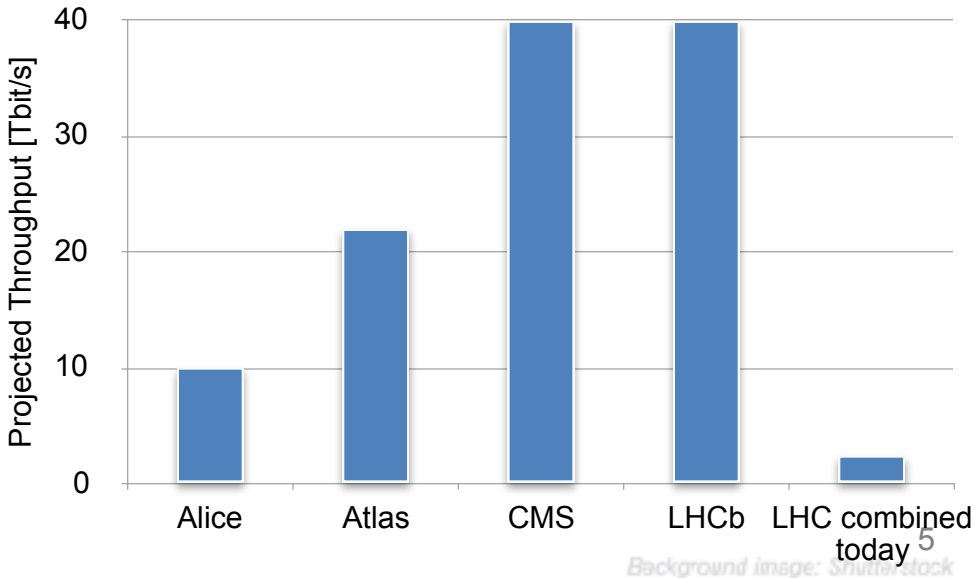
LHCb TDAQ Architecture Using Intel



Omni-Path

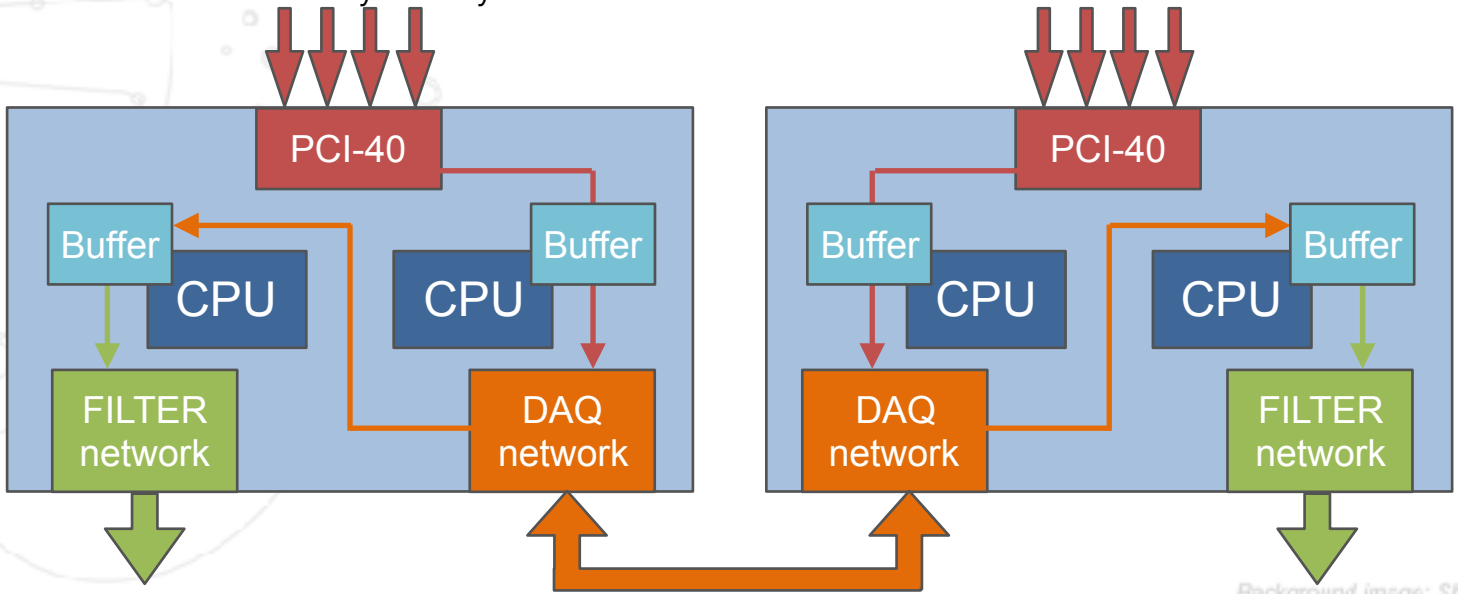
DAQ Challenges

- 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



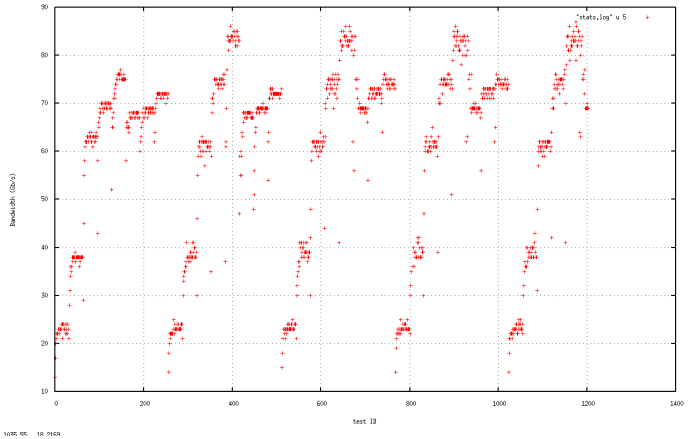
DAQ Data-collector & builder units

- 400 Gigabit/s total I/O
  - 100 Gigabit/s **input from detector** (PCI-40)
  - 100 Gigabit/s **in- and output to/from DAQ network**
  - 100 Gigabit/s **output to Filter farm**
- Stresses also the memory sub-system



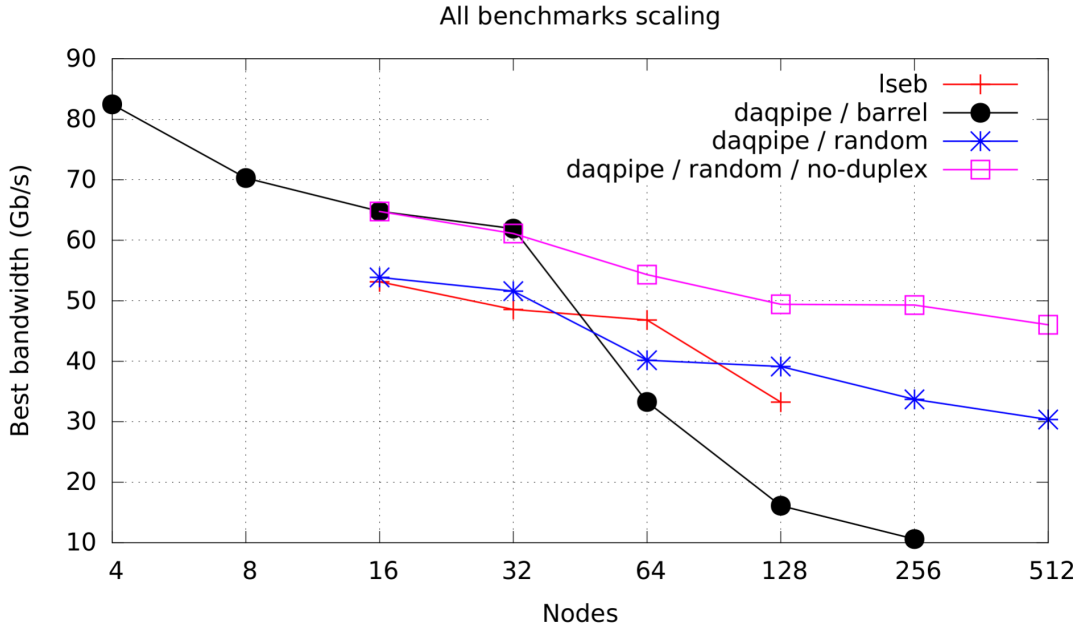
How to evaluate a 40 Tbit/s interconnect (without buying one)

- 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…)



Scaling on Marconi (Omni-Path) cluster

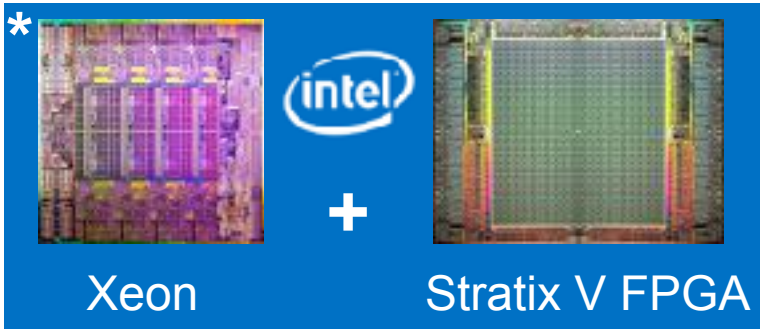
- Still a bit fragile w.r.t. traffic pattern
- Performance w.r.t. to EDR not yet exactly known
- Eagerly waiting for integrated version



FPGA & Xeon

New (and old) challenges on FPGA

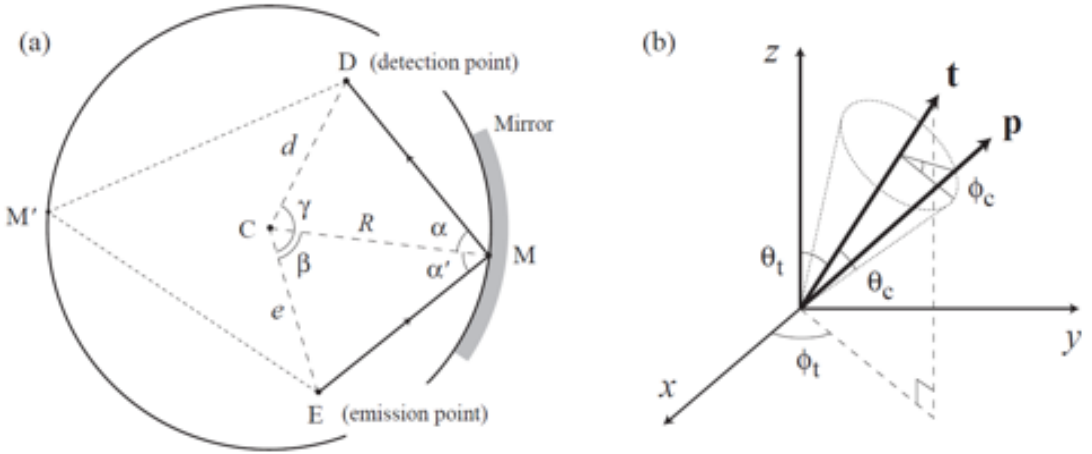
- Sophisticated algorithms need more time, bigger FPGAs, more data
- Long-term maintenance issues with custom hardware and low-level firmware
  - Upgrades usually mean replacing all the hardware
- Exact reproducibility of results without the custom hardware challenging and/or computationally intensive
- HTCC question: Can we build something similar with the integrated FPGA on Xeon® Platform?
- Filter algorithms contain very expensive calculations that may be well suited for offloading onto FPGAs.
- Cache-coherent FPGA to provide “custom extensions”



* Xeon+FPGA is currently still Beta HW!

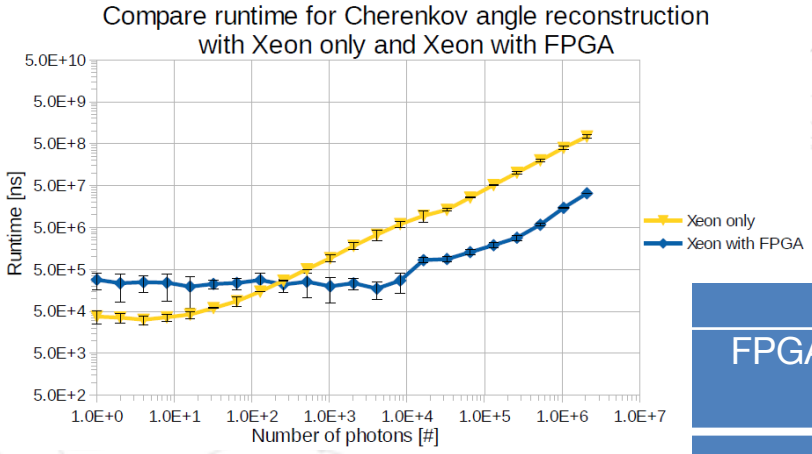
Test case: RICH PID algorithm

- Calculate Cherenkov angle Θ_c for each track t and detection point D, inverse ray-tracing, hyperbolic functions, etc…
- Currently not processed for every event, because it is too costly!



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

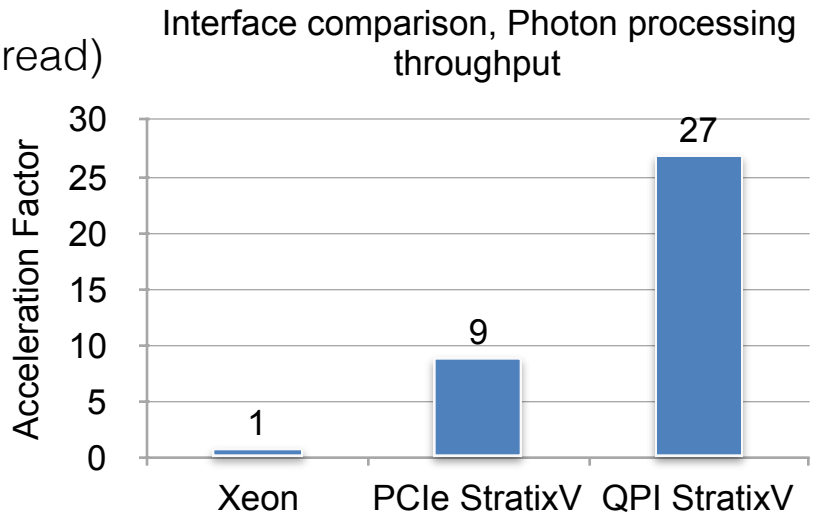


- Acceleration of factor up to 35 (26 using OpenCL) with Intel® Xeon/FPGA
- Theoretical limit of photon pipeline: a factor 64 with respect to single Intel® Xeon® thread
- Bottleneck: Data transfer bandwidth to FPGA

	Verilog RTL	OpenCL
FPGA Resource Type	FPGA Resources used [%]	FPGA Resources used [%]
ALMs	88	63
DSPs	67	82
Registers	48	24

The case for QPI

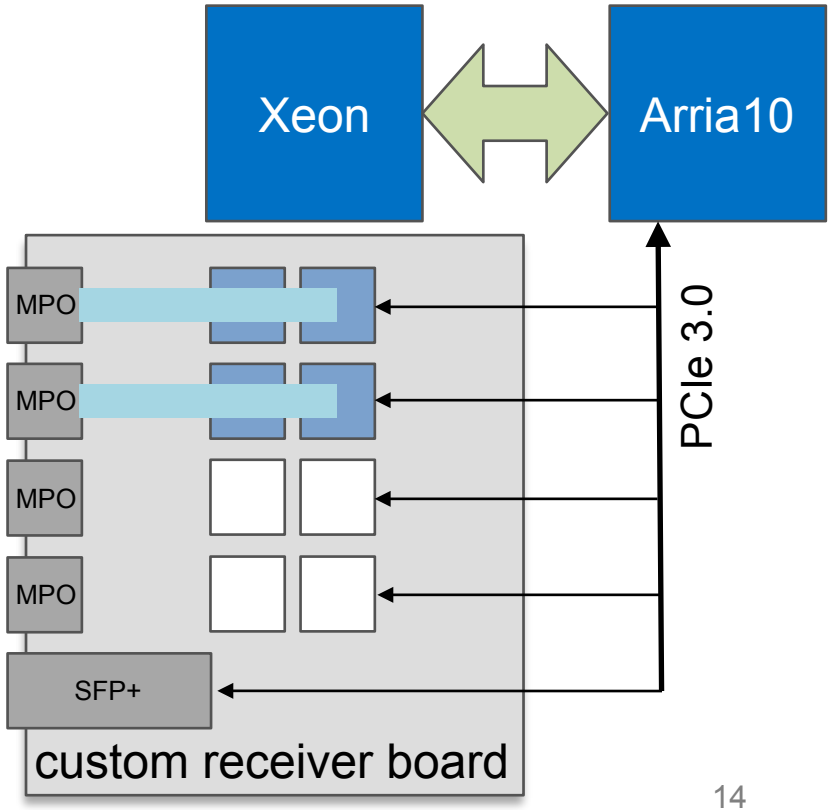
- StratixV programmed in OpenCL
- Compared to vectorized E2630v2 (single-thread)
- Still room for improvement (pipeline could do at least 2x more)



- 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.

A simplified PCI-40 card

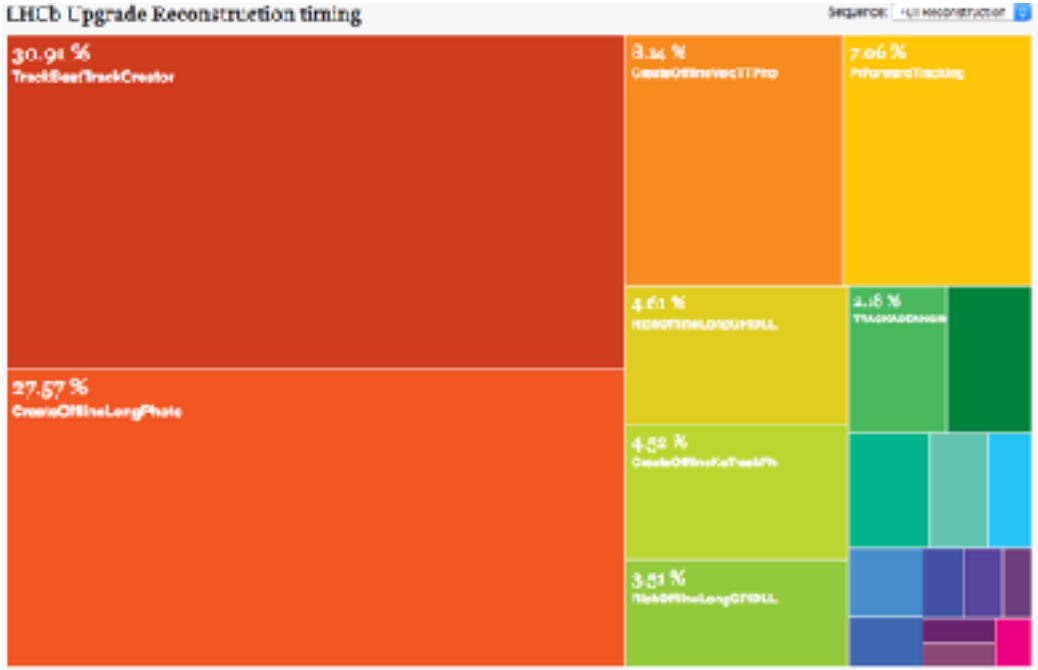
- Simplified receiver card without an FPGA.
- Use Xeon+FPGA for data processing
- Currently being discussed with CMS and Alice.



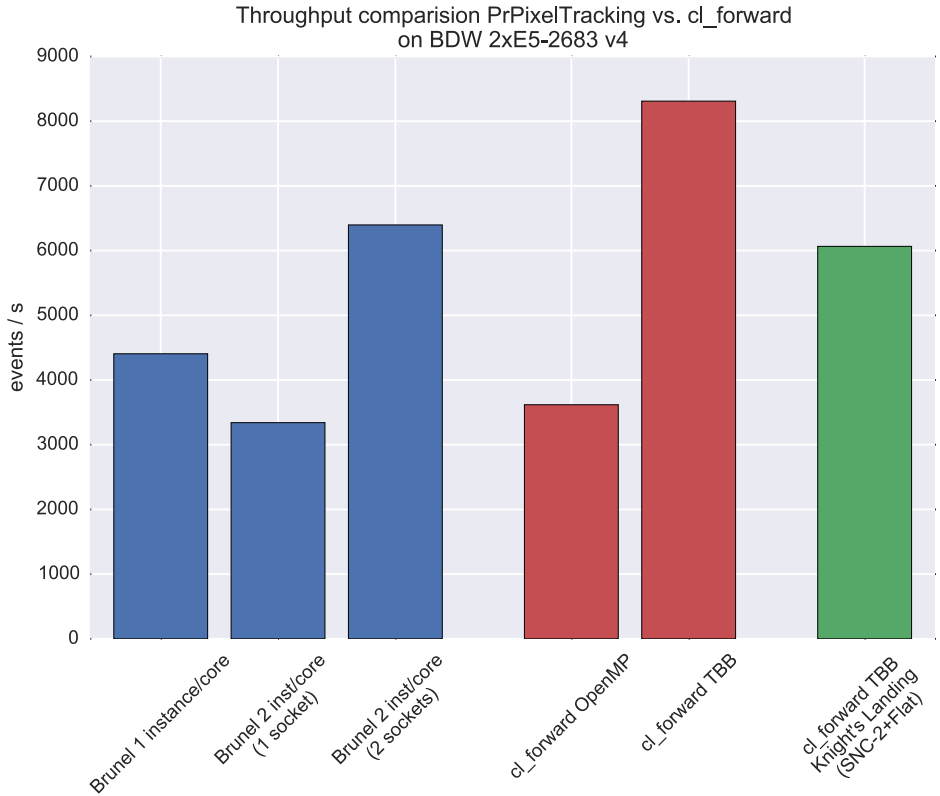
Xeon Phi & Xeon

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)



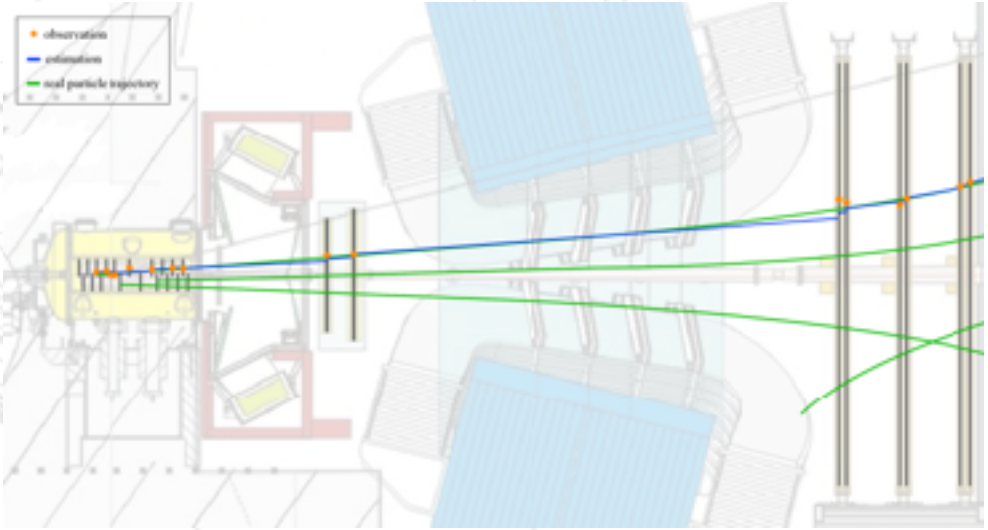
TBB for accelerating track-reconstruction

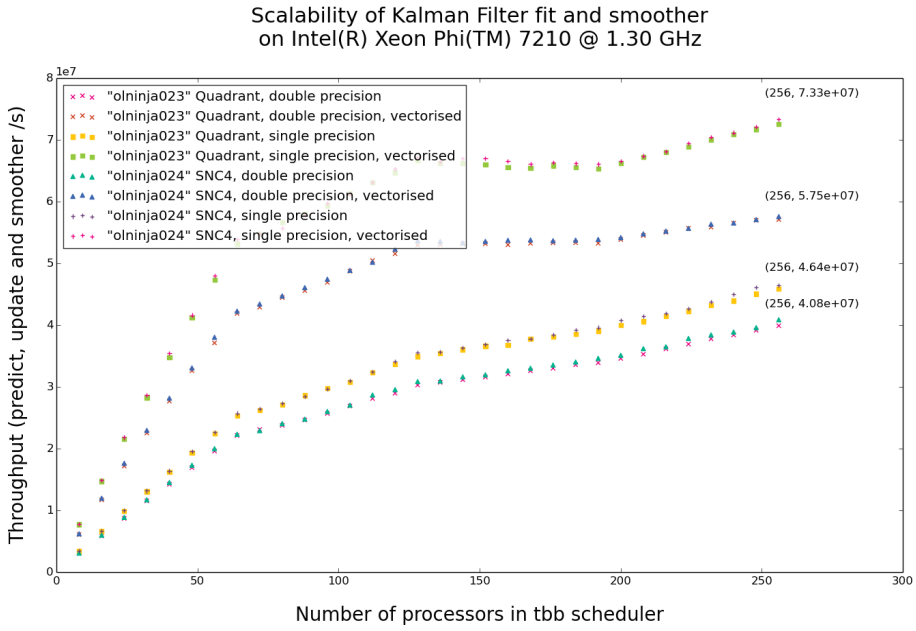


- Straight-line track reconstruction in the velopixel subdetector.
- Comparing TBB with official reconstruction (run in multiple instances without HT)
- tbbPixel speedup on BDW: 1.88
- KNL-specific optimizations will likely yield better throughput!
- Speedup and improved reconstruction efficiency can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

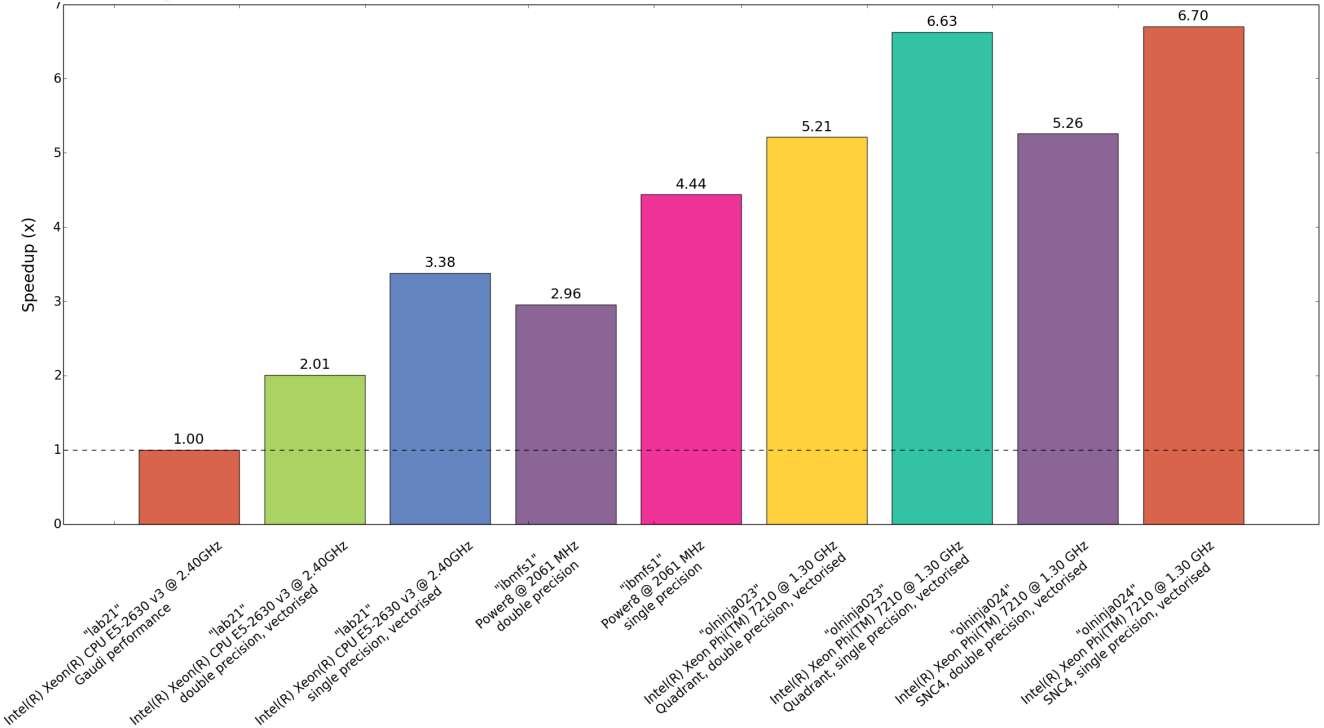




Background image: Shutterstock

SIMD Kalman Filter

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



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.

HTCC conclusions

- 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?

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, Karel Hà, Jon Machen (Intel), Rainer Schwemmer, Sébastien Valat, Balázs Vőneki