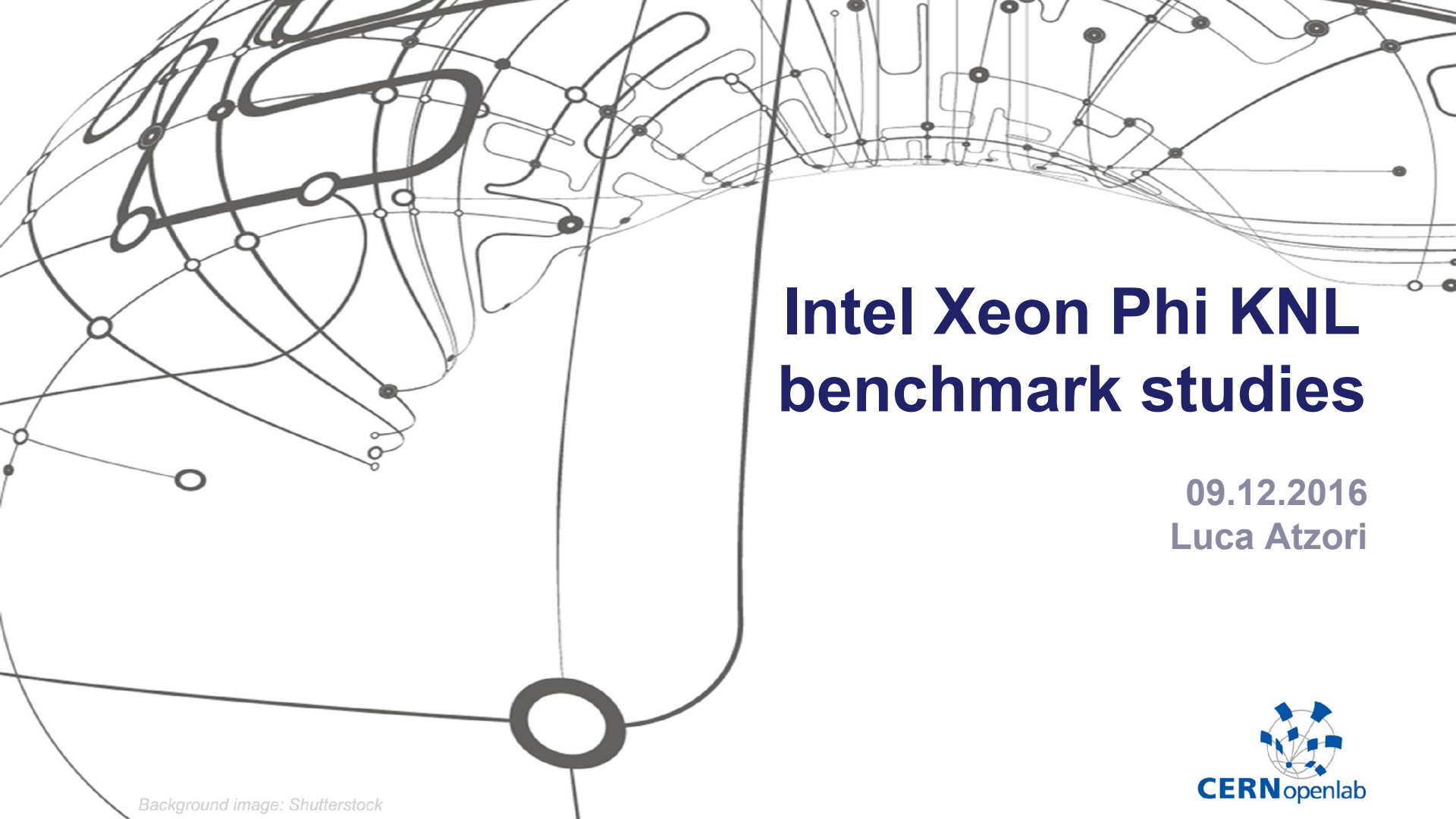


# Intel Xeon Phi KNL benchmark studies

09.12.2016

Luca Atzori

CERN openlab

Background image: Shutterstock

# Intel Xeon Phi: Knights Landing

**High-throughput successor to Knights Corner (KNC)**

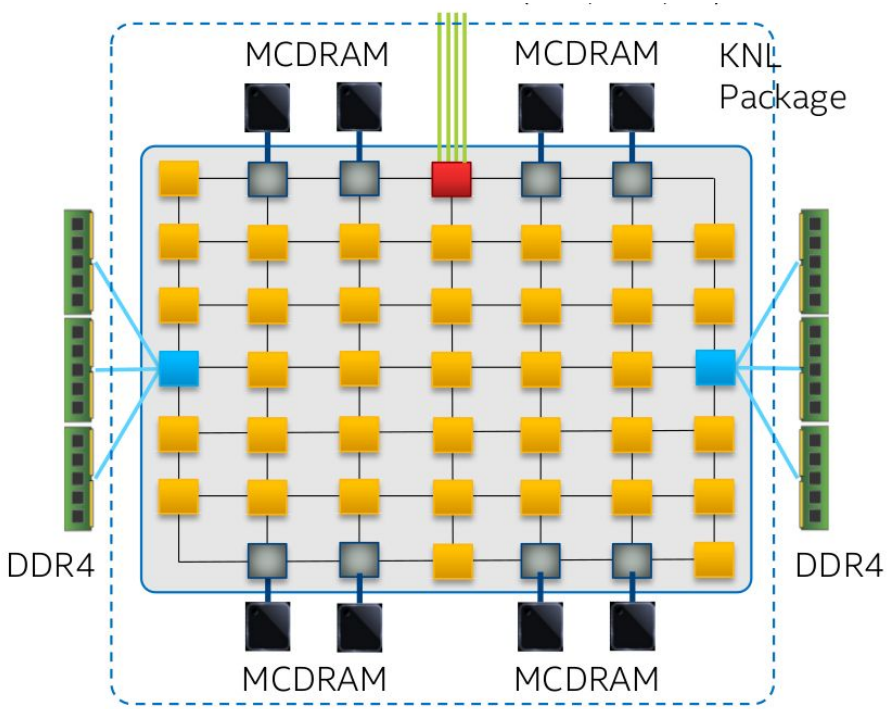
- **Bootable processor**
- **Up to 72 cores**
- **Very high bandwidth, flexible MCDRAM**
- **Improved single-thread performance**
- **Estimated peak performance of 3+ TFLOPs**


Source: Intel

# KNL Architecture Overview

- 2D mesh architecture
- Out-of-order cores
- Up to 384GB platform DDR4 memory

**Tile:**


Source: Intel

# KNL: Cluster modes

## All-to-All:

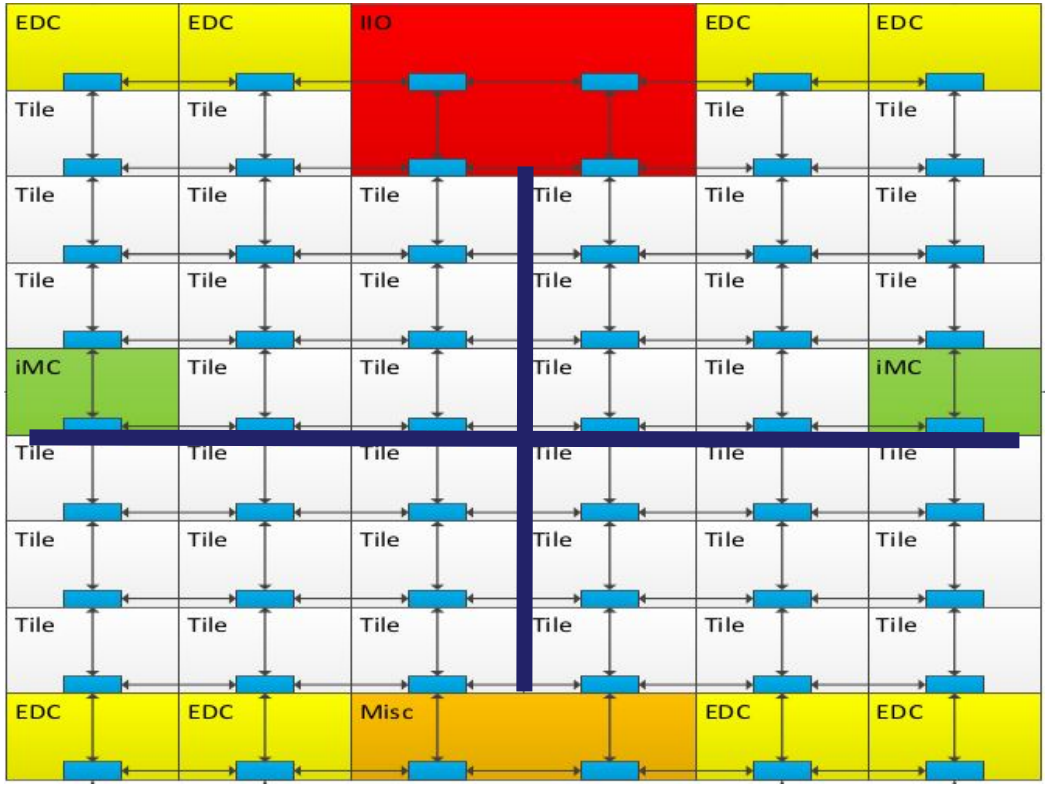
- Addresses uniformly distributed across all the Tag Directories

## Quadrant / Hemisphere:

- Die logically divided into 4 or 2 parts
- Addresses requested from a controller in a certain quadrant are mapped only to TDs of that quadrant

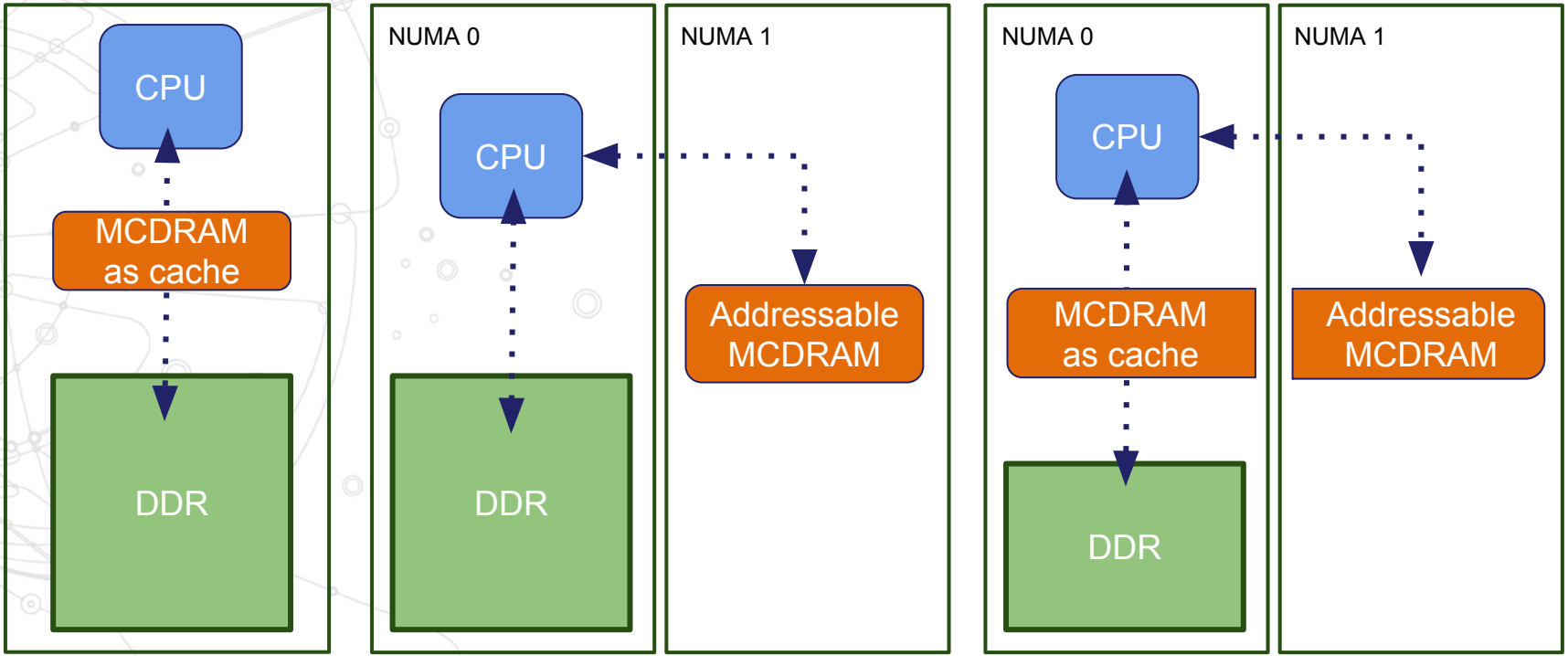
## Sub-NUMA Clustering:

- Parts are also NUMA nodes
- Possible to bind software threads and accessing NUMA-local memory




Source: Intel

# KNL: Memory modes




**Cache mode** **Flat mode** **Hybrid mode**

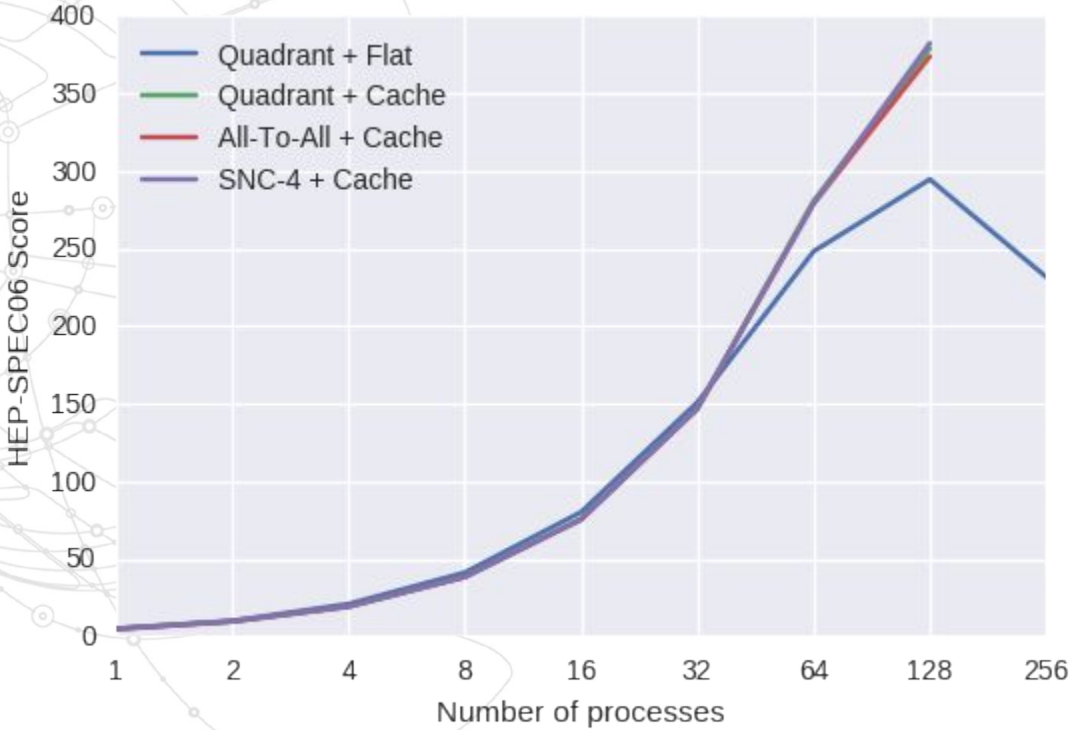
# System setup

**Developer Edition of Intel® Xeon Phi™ Processor 7210:**

- 64 cores @ 1.30 GHz
- 16GB MCDRAM
- 96 GiB RAM DDR4

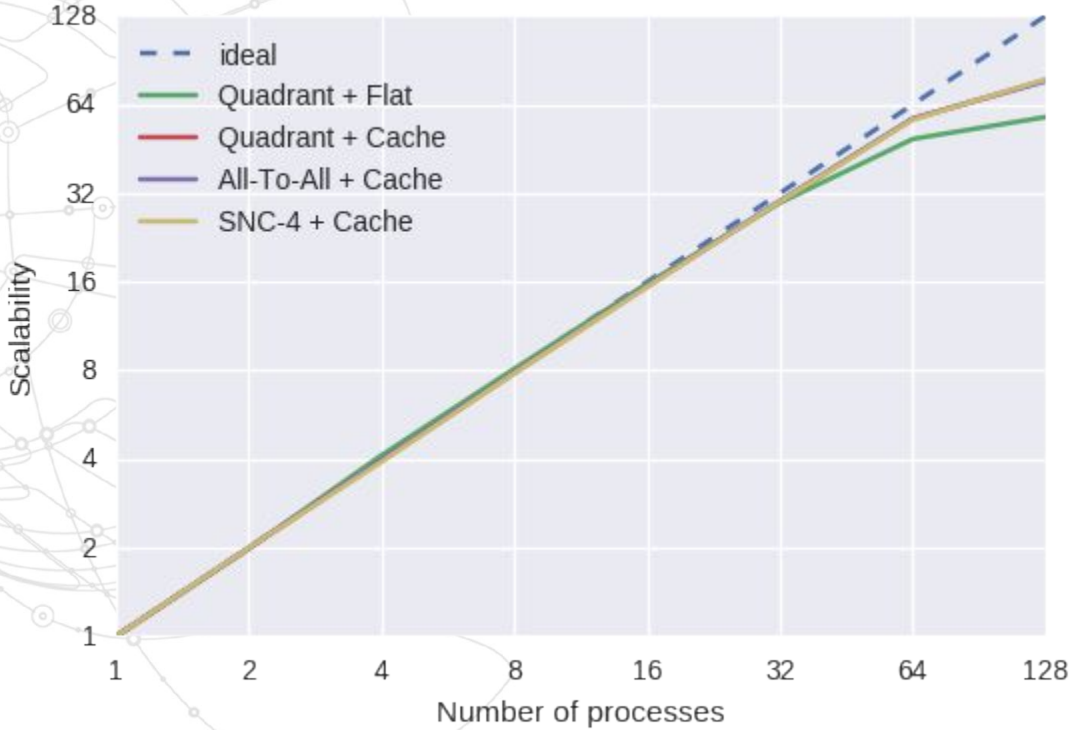
**CentOS Linux release 7.2.1511 running kernel 3.10.0-327.36.1.el7.x86_64**

# HEP-SPEC06



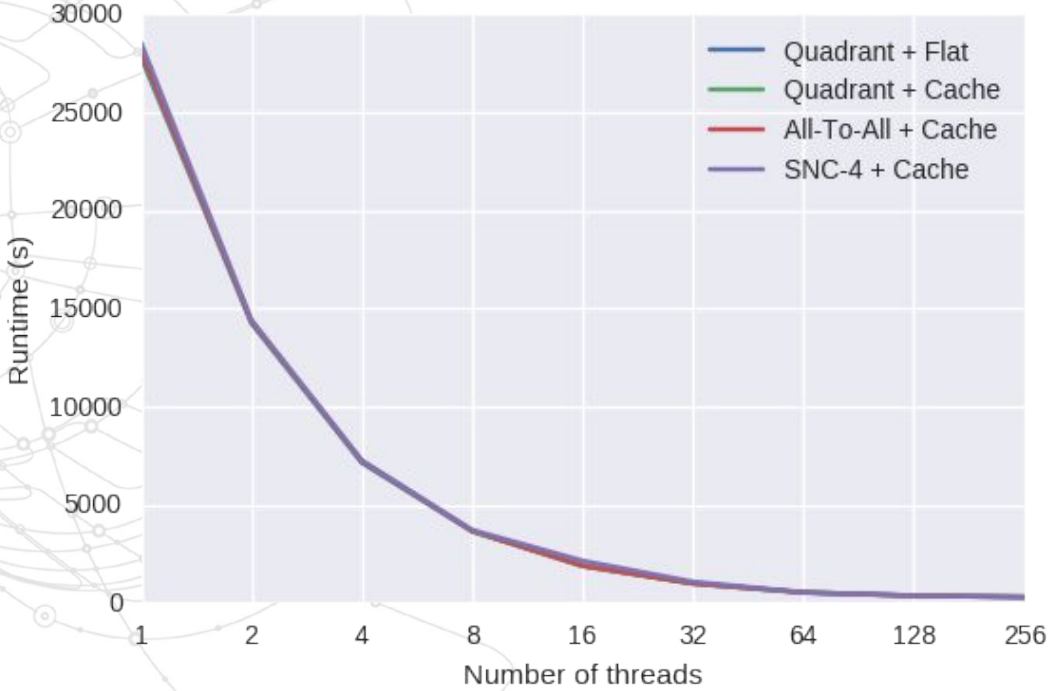

- **Weak scaling**
- **High memory footprint!**
  - Not possible to bind to MCDRAM
  - Starts swapping with 256 instances (memory usage is more than 90GB)

# HEP-SPEC06



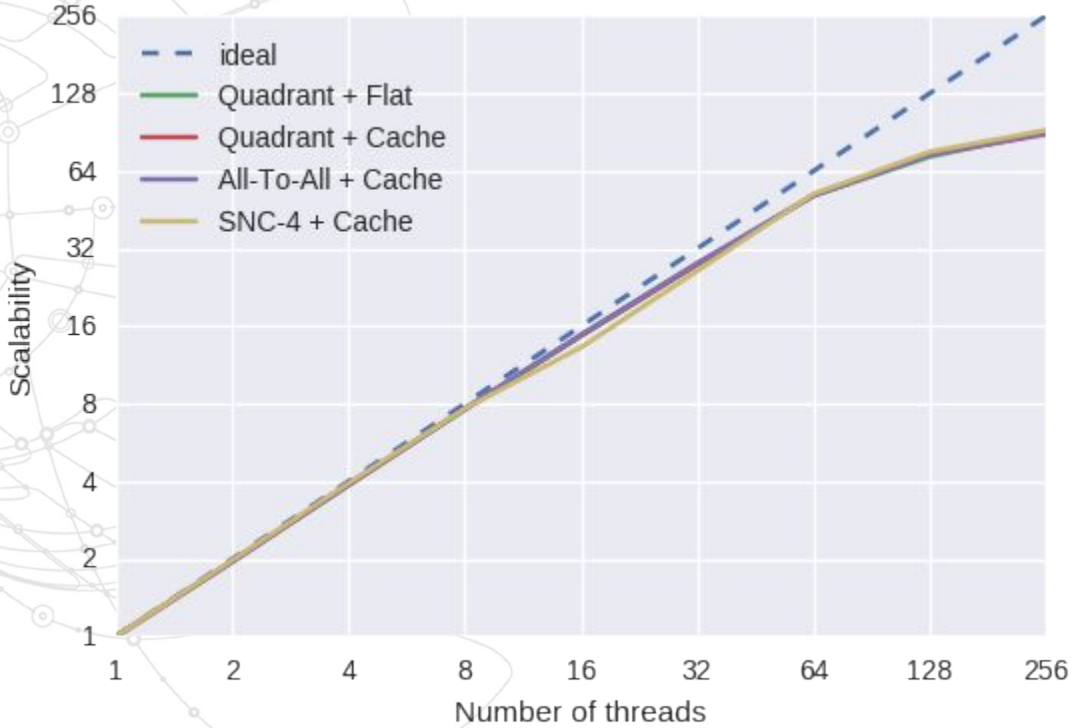

- **Flat outperformed by the Cache configurations**
- **H-S06 contains benchmarks that benefit from memory locality**
- **Scalability:**
  - ~ 90% with 64 processes
  - ~ 61% with 128 processes

# ParFullCMS



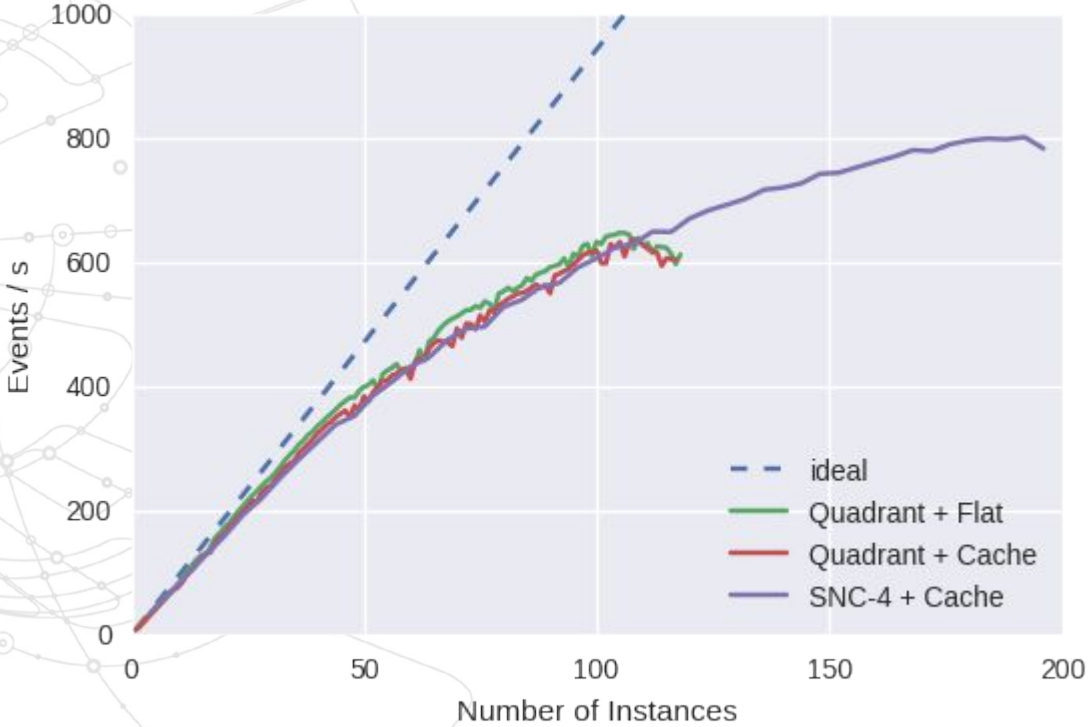

- **Strong scaling**
- **Memory footprint allows the binding on MCDRAM (3.6GB with 256 threads)**
- **Latency bound: no benefit from Flat configuration**

# ParFullCMS




- **Almost same performance for every clustering mode**
- **Scalability**
  - ~ 81% with 64 threads
  - ~ 36% with 256 threads

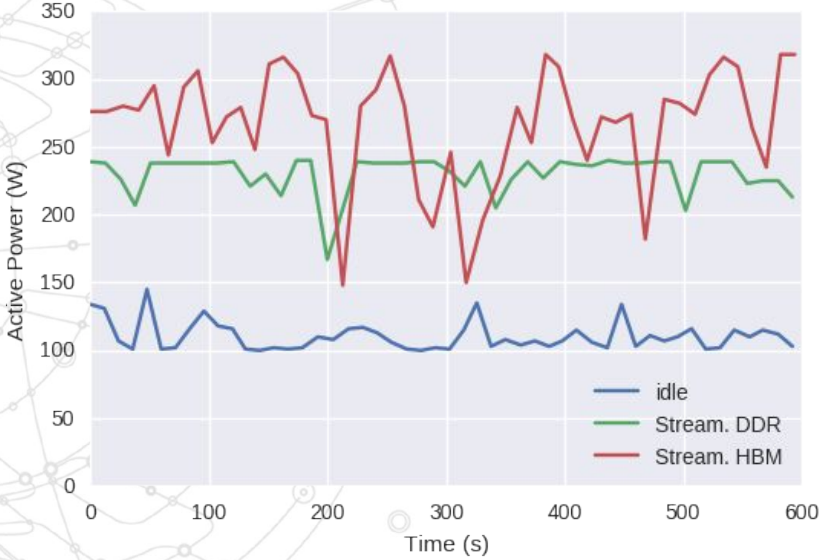
# LHCb high-level trigger



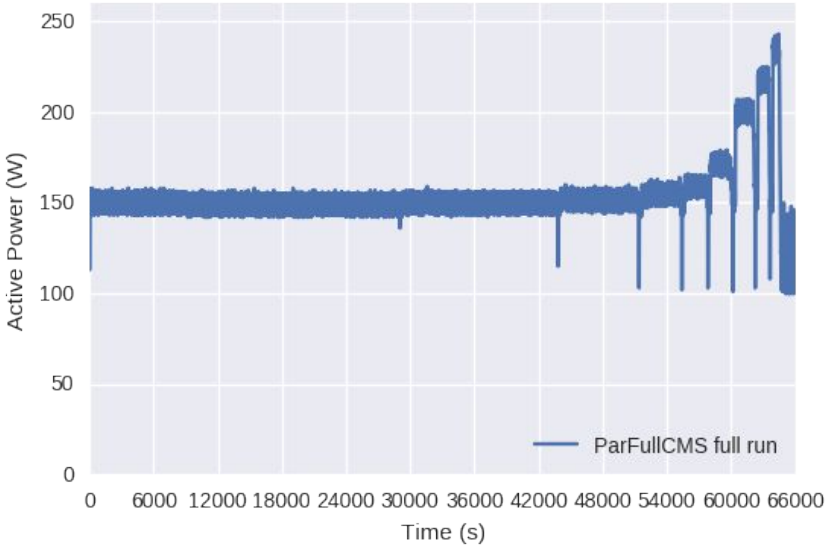

- Similar to production software in LHCb
- Quadrant configurations run with DDR4 memory pinning
- Flat is ~ 4% faster than Cache

Thanks to Rainer Schwemmer

# Power benchmarks

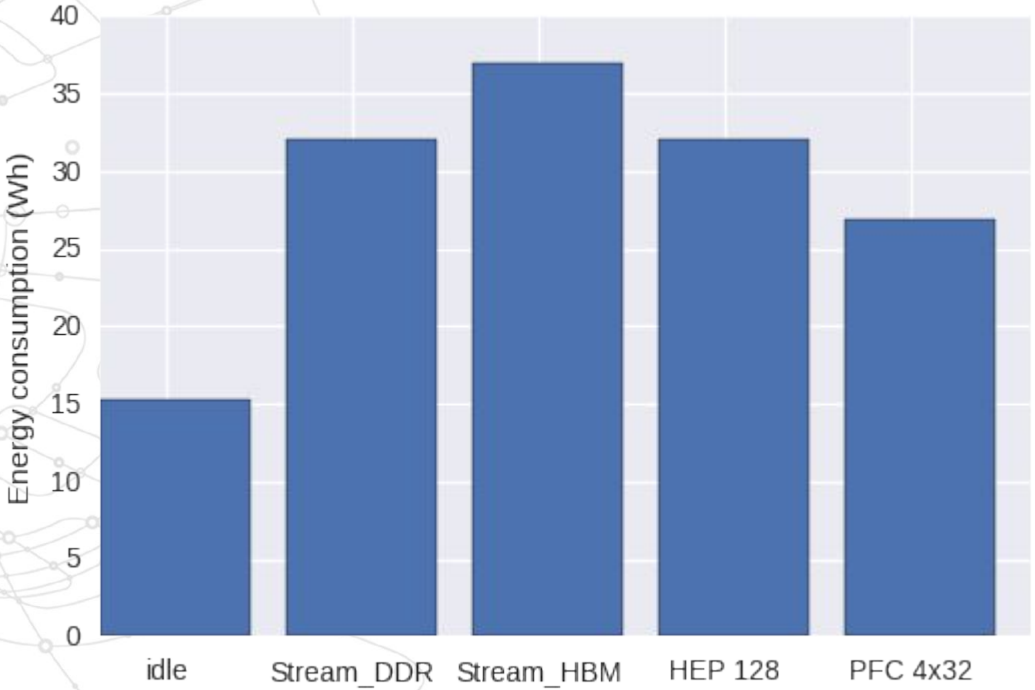



**Streaming from MCDRAM requires more power than DDR4**




**Power consumption of an entire run of ParFullCMS**

# Power benchmarks




**Comparison of the energy consumption between five analysed configurations.**

CERN openlab

# Thank you!

**Full technical report:**
https://luatzori.web.cern.ch/knl/KNL_Report.html