LHCb Usage of HPC Centers

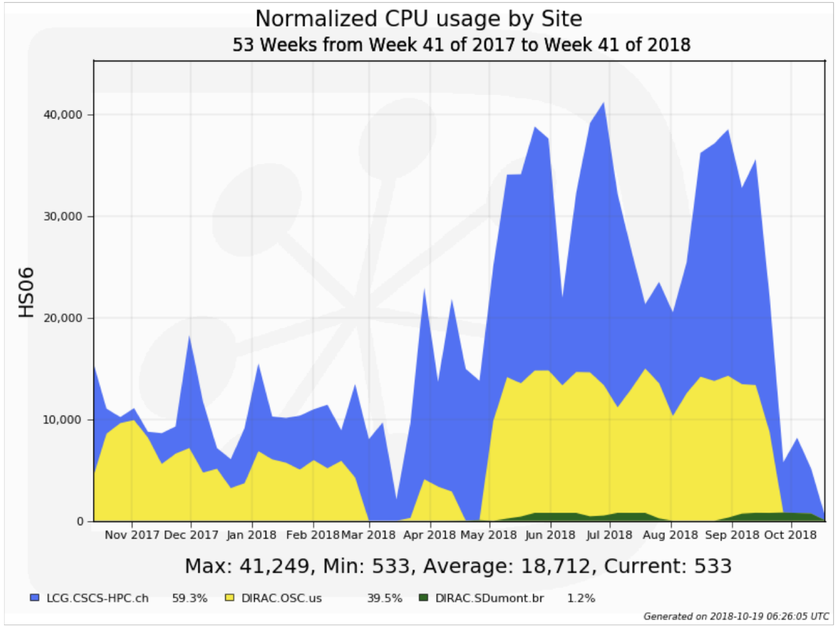
Stefan Roiser

PRACE Workshop

22 October 2018

Current Usage

- LHCb is using HPC centers in Switzerland (CSCS) and US (OSC)
  - Expansion planned, e.g. Italy (Cineca) and Brazil (Santos Dumont)
  - Use “standard” intel xeon processors
  - Worker nodes equipped with ”CVMFS” files system
  - Whenever possible, access of resources via WLCG interfaces

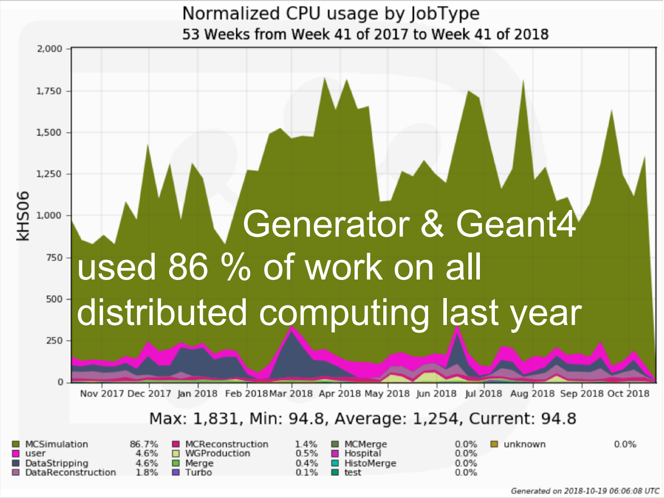


- All LHCb distributed computing resources, including HPCs, are used via the same “LHCbDIRAC” tool for workload and data management

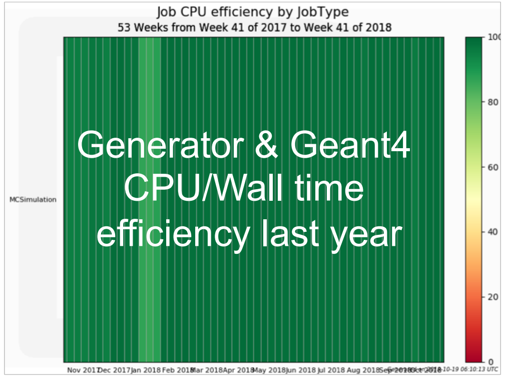
LHCb workflow(s) to deploy on HPCs

- Monte Carlo Simulation Generator & Geant4
  - i.e. particle collision and detector response
  - 80 – 90 % of work on distributed computing resources spent for Generator & Geant4
  - Simulation can be interrupted by signal

- Generator & Geant4 very simple workflow
  - No input data needed
  - Write output file O(100MB) to “close” storage site every ~ 6 hours
  - High CPU efficiency on intel CPUs



Generator & Geant4 used 86 % of work on all distributed computing last year



Generator & Geant4 CPU/Wall time efficiency last year

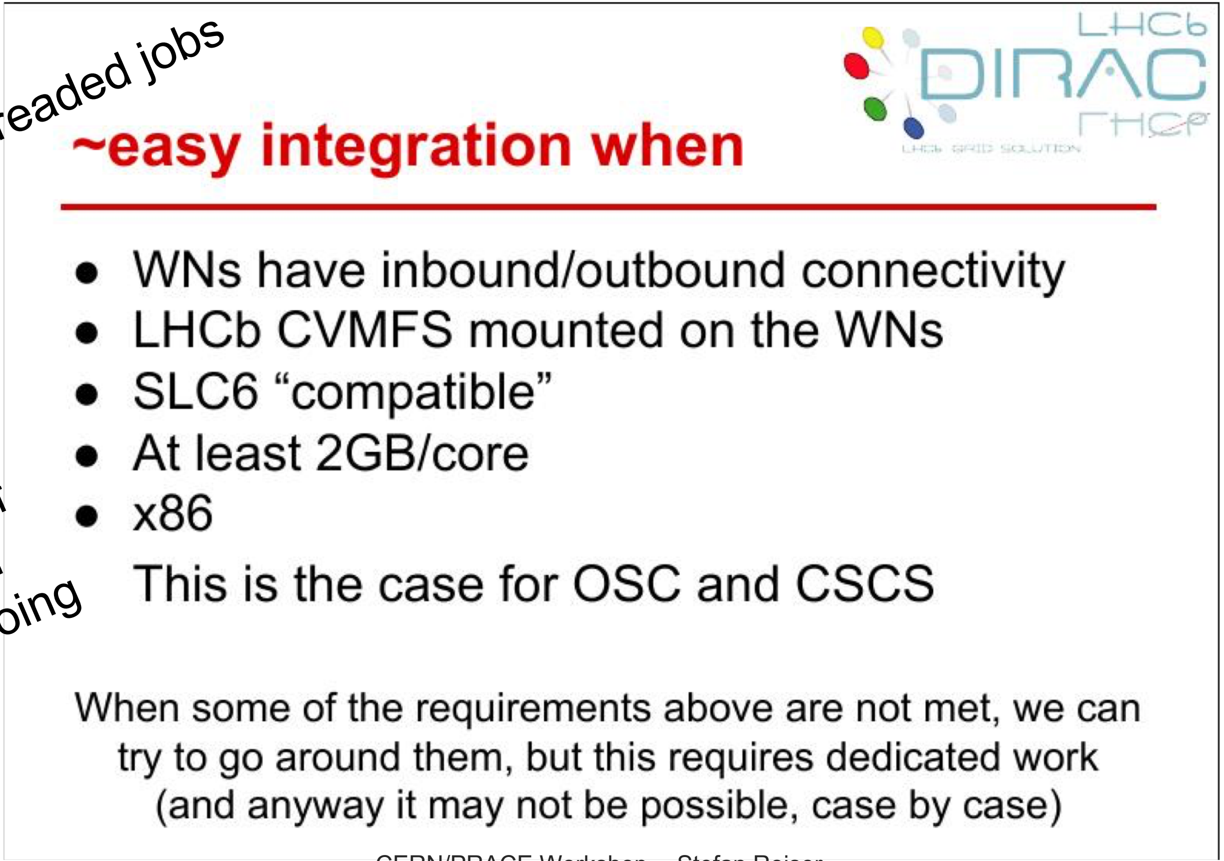
Access to Resources

F. Stagni, Nov 2017
10th LHCb Computing WS

Mostly single-threaded jobs

Multi-process version available

Development of multi-threaded software ongoing


LHCb
DIRAC
LHCb GRID SOLUTION

~easy integration when

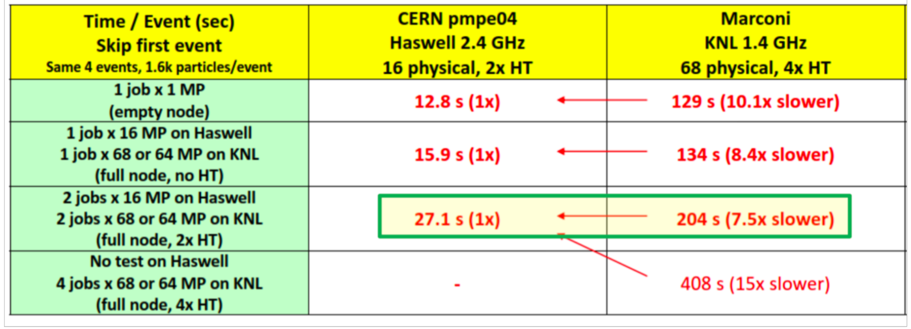
- WNs have inbound/outbound connectivity
- LHCb CVMFS mounted on the WNs
- SLC6 “compatible”
- At least 2GB/core
- x86

This is the case for OSC and CSCS

When some of the requirements above are not met, we can try to go around them, but this requires dedicated work (and anyway it may not be possible, case by case)

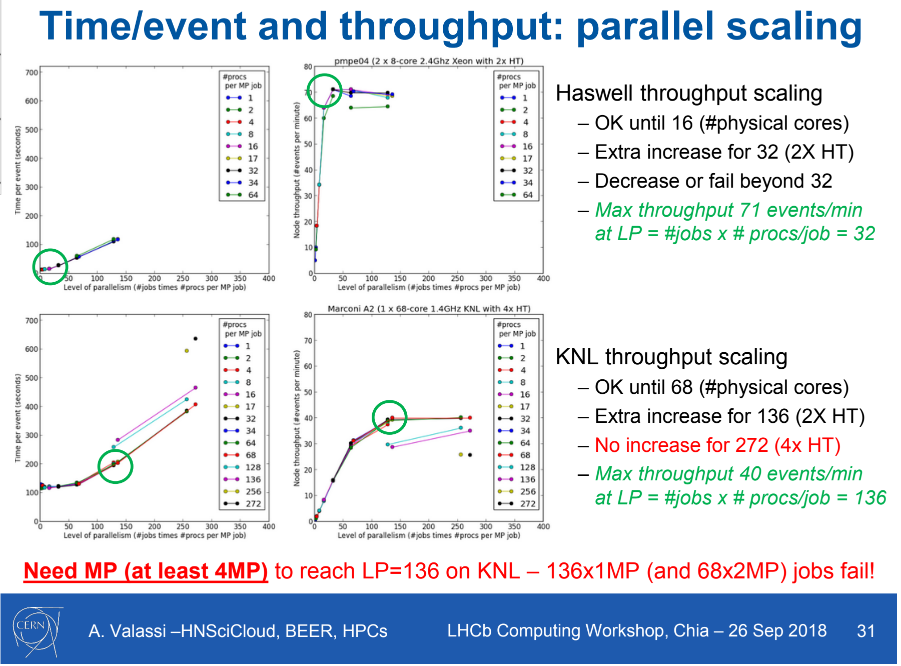
Example: Efficiency on Xeon phi

A. Valassi, Sep 2017
11th LHCb Computing WS



Time / Event (sec) Skip first event Same 4 events, 1.6k particles/event	CERN pmpe04 Haswell 2.4 GHz 16 physical, 2x HT	Marconi KNL 1.4 GHz 68 physical, 4x HT
1 job x 1 MP (empty node)	12.8 s (1x)	129 s (10.1x slower)
1 job x 16 MP on Haswell 1 job x 68 or 64 MP on KNL (full node, no HT)	15.9 s (1x)	134 s (8.4x slower)
2 jobs x 16 MP on Haswell 2 jobs x 68 or 64 MP on KNL (full node, 2x HT)	27.1 s (1x)	204 s (7.5x slower)
No test on Haswell 4 jobs x 68 or 64 MP on KNL (full node, 4x HT)	-	408 s (15x slower)

- Work to understand performance on offered Xeon phi resources
  - Running multi-process simulation
- Time / Event on fully loaded machine factor 7.5 slower
  - Not explainable only by slower core speed



Time/event and throughput: parallel scaling

Haswell throughput scaling

- OK until 16 (#physical cores)
- Extra increase for 32 (2X HT)
- Decrease or fail beyond 32
- *Max throughput 71 events/min at LP = #jobs x # procs/job = 32*

KNL throughput scaling

- OK until 68 (#physical cores)
- Extra increase for 136 (2X HT)
- No increase for 272 (4x HT)
- *Max throughput 40 events/min at LP = #jobs x # procs/job = 136*

Need MP (at least 4MP) to reach LP=136 on KNL – 136x1MP (and 68x2MP) jobs fail!

A. Valassi –HNSciCloud, BEER, HPCs LHCb Computing Workshop, Chia – 26 Sep 2018 31

Future perspectives

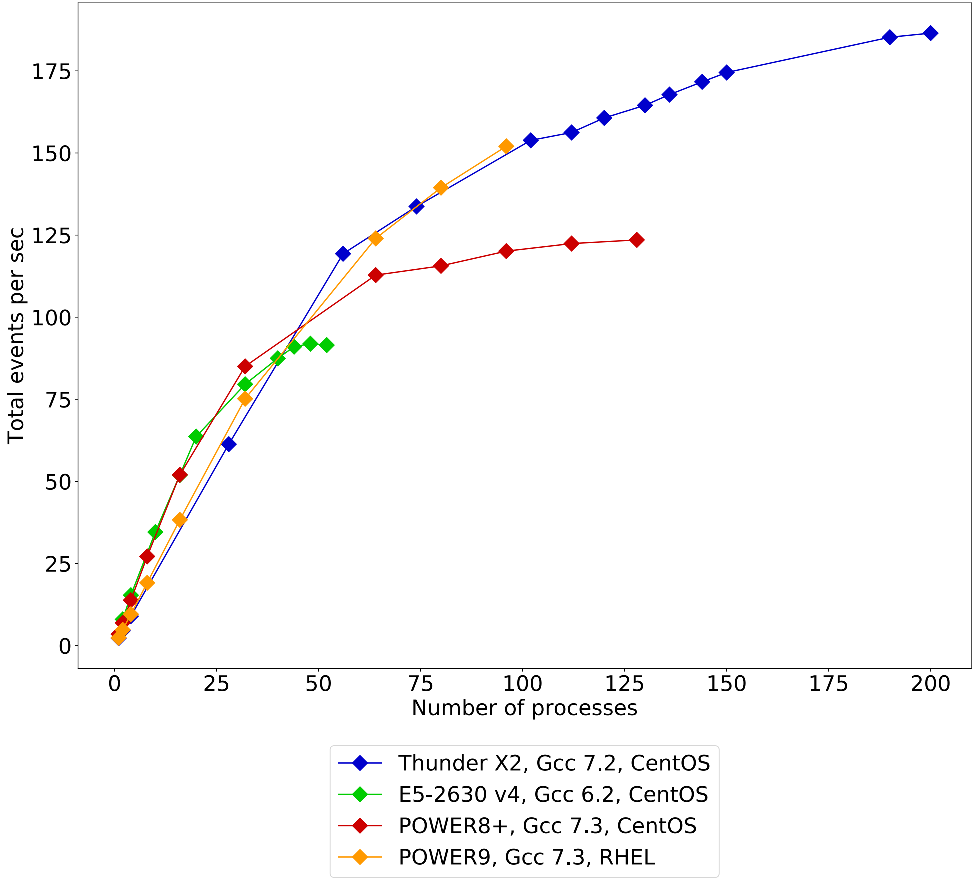
- In LHCb work ongoing to port application framework to multi-threaded
  - Huge reduction in memory consumption
  - Will help on deploying workflows on many core intel friendly architectures

- Porting of software to ARM & Openpower ongoing
  - First versions available. Some tweaking especially for vectorization needed

- Usage of non intel architectures for LHCb workflows is unclear
  - Especially in view of simulation will stay the dominant workflow for LHCb

ARM & Power performance

L. Promberger
CHEP 2018



Summary

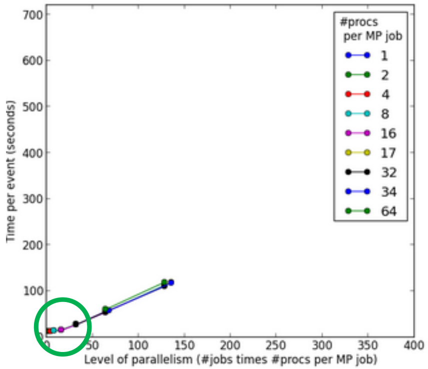
- LHCb is using and plans to extend usage of HPC centers further
  - Predominantly will deploy ”simple” simulation workflow

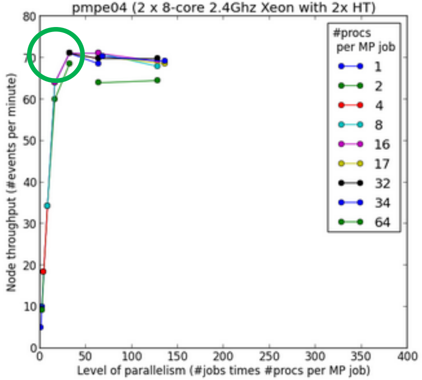
- Usage of intel compatible resources via standard interfaces and environment is straight forward
  - Usage of alternative architectures unclear
  - Slowdown in time to start exploiting resource experienced for non standard interfaces

- Future work of the experiment includes port to multi-threaded software stack

Backup

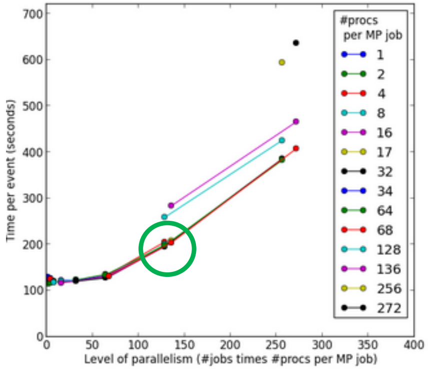
Time/event and throughput: parallel scaling

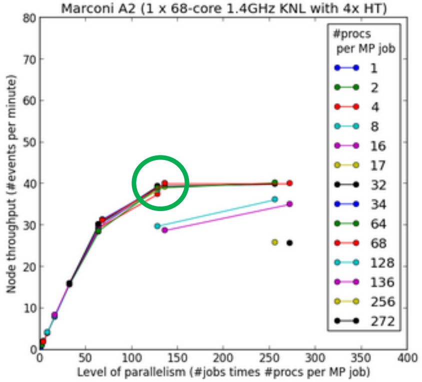




Haswell throughput scaling

- OK until 16 (#physical cores)
- Extra increase for 32 (2X HT)
- Decrease or fail beyond 32
- *Max throughput 71 events/min at LP = #jobs x # procs/job = 32*



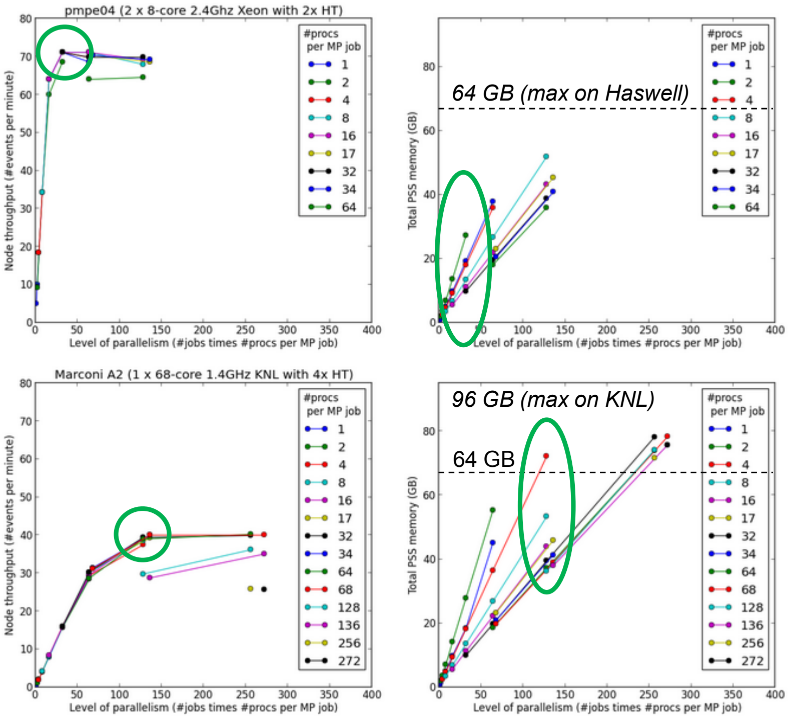


KNL throughput scaling

- OK until 68 (#physical cores)
- Extra increase for 136 (2X HT)
- No increase for 272 (4x HT)
- *Max throughput 40 events/min at LP = #jobs x # procs/job = 136*

Need MP (at least 4MP) to reach LP=136 on KNL – 136x1MP (and 68x2MP) jobs fail!

Total PSS memory



- No significant difference between Haswell and KNL in memory used
  - Difference is max total memory and max parallelism (and throughput…)

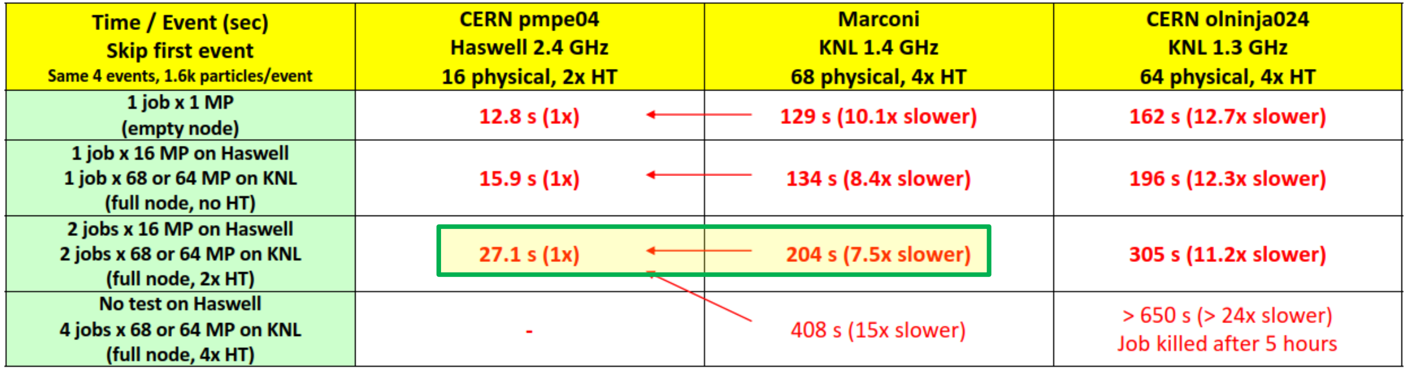
- For a given nMP procs/job, memory ~proportional to #jobs (straight line)

- For a given level of parallelism LP = *#jobs x #procs/job (~ #procs) :*
  - 1MP take less memory (0.7 GB/LP) than 2MP (0.9 GB/LP) – expected
  - Memory then decreases from 2MP (0.9 GB/LP) to 17MP (0.33 GB/LP) down to >= 68MP (0.27 GB/LP)

Need MP (at least 4MP) to reach LP=136 on KNL – 136x1MP (and 68x2MP) jobs fail!

Optimal memory at optimal throughput (40 events/min @LP=136) is for 17MP to 68MP

Summary of timing numbers



Time / Event (sec) Skip first event Same 4 events, 1.6k particles/event	CERN pmpe04 Haswell 2.4 GHz 16 physical, 2x HT	Marconi KNL 1.4 GHz 68 physical, 4x HT	CERN olninja024 KNL 1.3 GHz 64 physical, 4x HT
1 job x 1 MP (empty node)	12.8 s (1x)	129 s (10.1x slower)	162 s (12.7x slower)
1 job x 16 MP on Haswell 1 job x 68 or 64 MP on KNL (full node, no HT)	15.9 s (1x)	134 s (8.4x slower)	196 s (12.3x slower)
2 jobs x 16 MP on Haswell 2 jobs x 68 or 64 MP on KNL (full node, 2x HT)	27.1 s (1x)	204 s (7.5x slower)	305 s (11.2x slower)
No test on Haswell 4 jobs x 68 or 64 MP on KNL (full node, 4x HT)	-	408 s (15x slower)	> 650 s (> 24x slower) Job killed after 5 hours

- Timings for maximum throughput configurations:
  - –Haswell (2x 8-core 2xHT): use LP=32 *(32x single-process Gauss jobs)*
  - –KNL (1x 68-core 2xHT): use LP=136 *(e.g. 8x 17MP GaussMP jobs)*
  - –Haswell 27s/evt (71 evts/min) vs. KNL 204s/evt (40 evts/min)
  - –*KNL 7.5x slower than Haswell (CPU + Turbo speed is ~2x-3x slower)*
    - *Extra slowdown ~3x on KNL (due to memory access? to be understood)*

- For reference: *20M core-hours on Marconi* (68-core) is 300k node-hours
  - –This is 33 KNL nodes for one year (1y = 9k h) [i.e. 4.5k SP KNL slots]
  - –Equivalent to 33x40/71=18.6 Haswell [or 4.5k/7.5 = 600 SP Haswell slots]
  - –Haswell has 32 slots → *equivalent to 600 SP Haswell slots for one year*

Performance - The machines

	ThunderX2	E5-2630 v4	Power8+	Power9
Architecture	ARM	Intel	PowerPc	PowerPc
Platform	aarch64	x86_64	ppc64le	ppc64le
Compiler	GCC 7.2	GCC 6.2	GCC 7.3	GCC 7.3
Number logical cores	224	40	128	176
Threads per core	4	2	8	4
Cores per socket	28	10	8	22
Sockets/NUMA nodes	2	2	2	2
RAM (GB)	256	64	256	128
Largest intrinsic set	NEON	AVX2	Altivec	Altivec
CPU performance	top-notch	cost-efficient		
	high-tier	mid-tier		