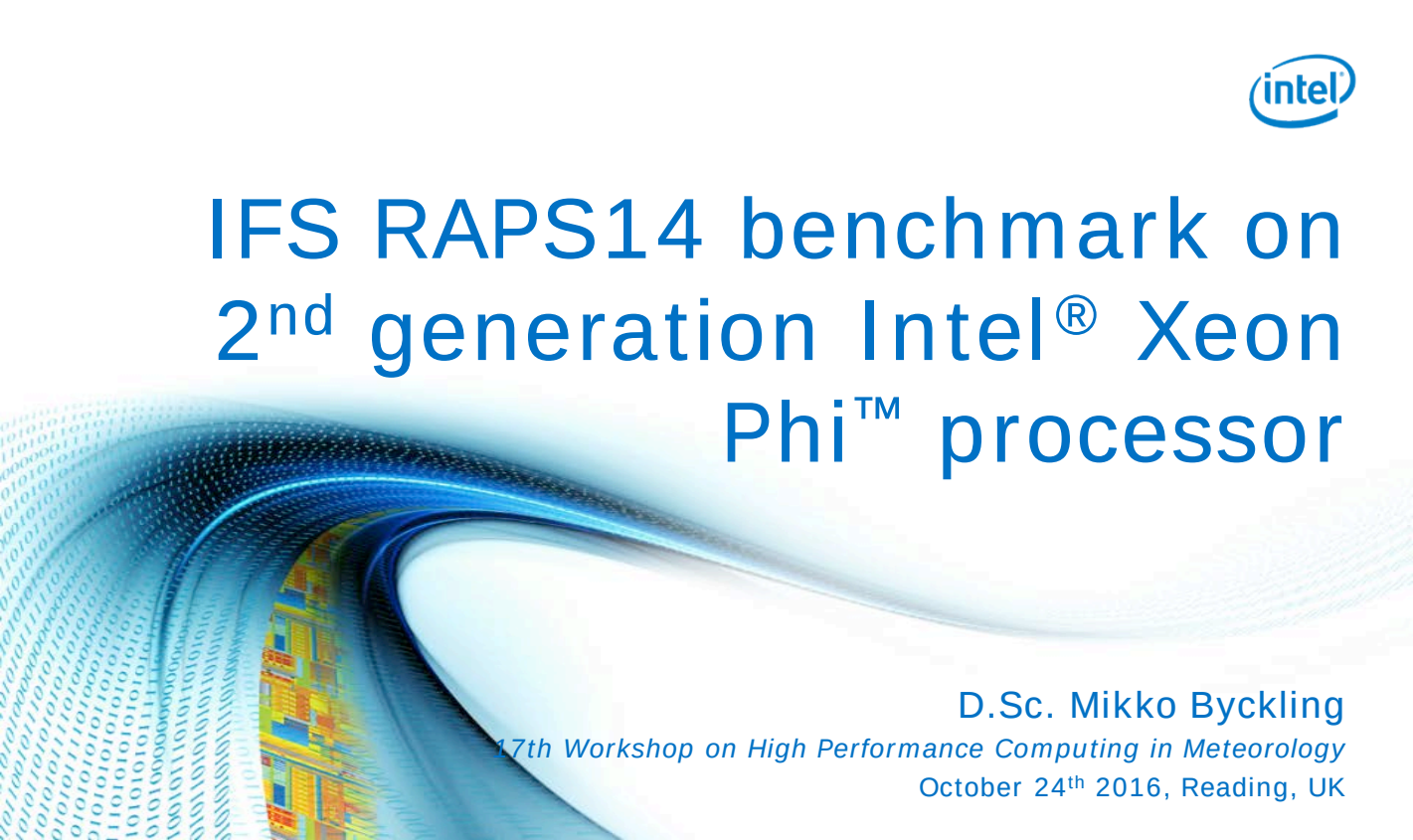
IFS RAPS14 benchmark on 2nd generation Intel® Xeon Phi™ processor



D.Sc. Mikko Byckling

17th Workshop on High Performance Computing in Meteorology

October 24th 2016, Reading, UK

Legal Disclaimer & Optimization Notice

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Optimization Notice

Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Contents

- Intel® Xeon Phi™ processor codenamed “Knights Landing” (KNL) overview
  - Architecture, cluster and memory modes
- IFS RAPS14 benchmark overview
- IFS RAPS14/TL159 benchmarks
  - Notes on optimization effort
  - Time to solution, application profile, energy to solution

Intel® Xeon Phi™ processor overview

Intel® Xeon Phi™ architecture

Instruction set architecture

Intel® Xeon® Processor compatible, adds Intel® AVX-512

On-package memory

16GB, up to 490 GB/s STREAM TRIAD

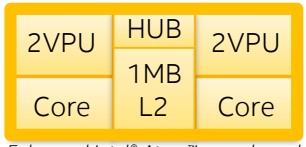
Platform Memory

Up to 384GB (6ch DDR4 2400)

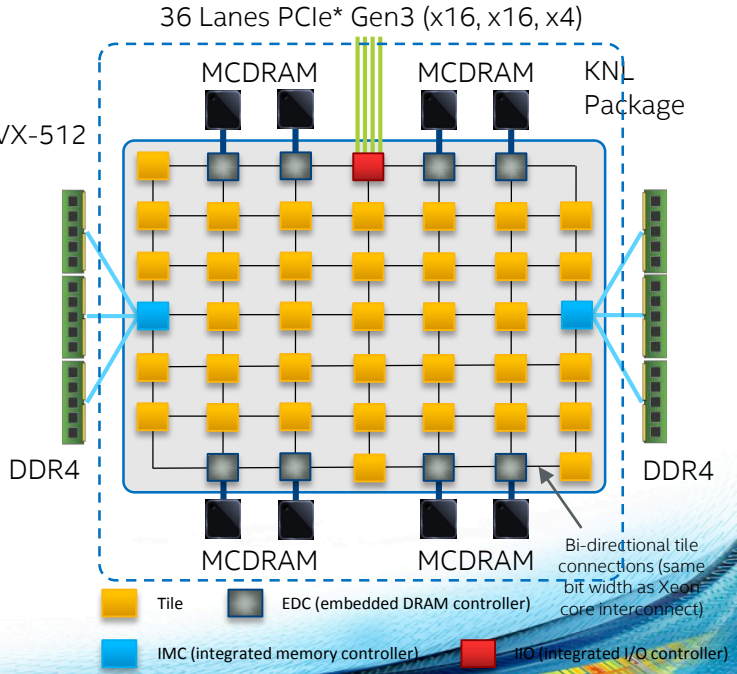
Fixed Bottlenecks

- ✓ 2D Mesh Architecture
- ✓ Out-of-Order Cores
- ✓ 3X single-thread vs. KNC

TILE:
(up to 36)

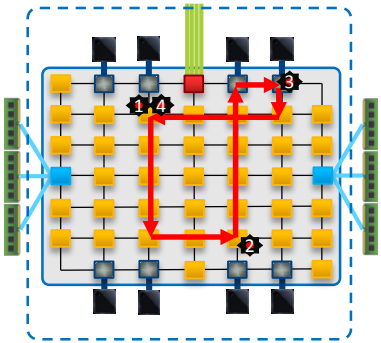
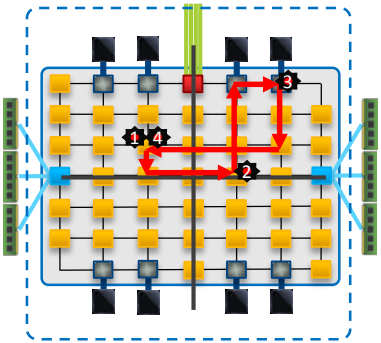
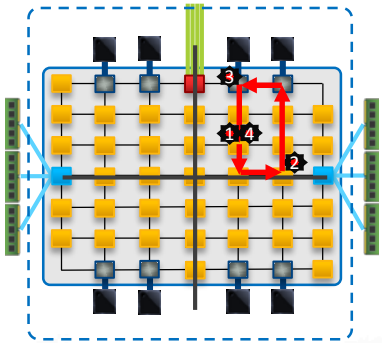


Enhanced Intel® Atom™ cores based on Silvermont Microarchitecture



SOFTWARE AND SERVICES

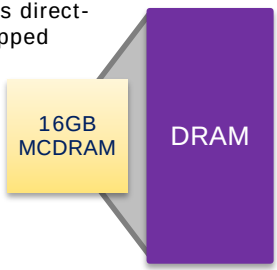
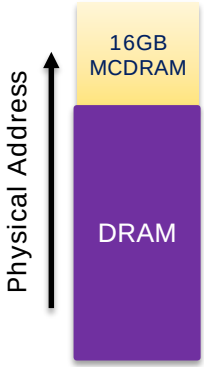
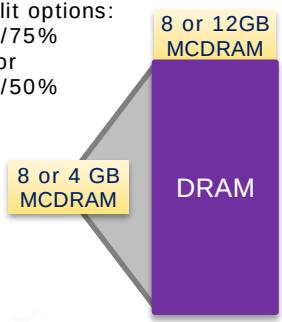
KNL cluster modes

All-2-All	Quadrant	Sub-NUMA clustering
		
No affinity between *tile*, *directory* and *memory*	Chip divided into 4 virtual *software transparent* quadrants **Affinity** between *directory* and *memory*	Chip visible to the OS as a 4S Xeon **Affinity** between *tile*, *directory* and *memory*

SOFTWARE AND SERVICES

Cluster modes are BIOS-selectable

KNL on-package memory modes

Cache mode	Flat mode	Hybrid mode
64B cache lines direct-mapped	Physical Address	Split options: 25/75% or 50/50%
		
MCDRAM as a “L3 cache” between CPU and DDR (HW managed)	Application manages the use of MCDRAM and DDR	MCDRAM both as cache and application managed memory

SOFTWARE AND SERVICES

Memory modes are BIOS-selectable

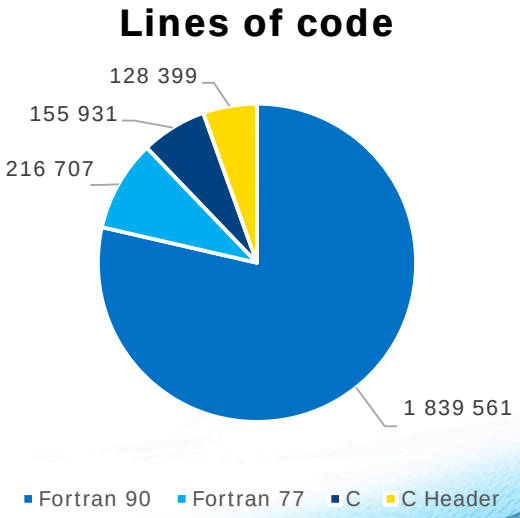
IFS RAPS14 benchmark

IFS RAPS14 benchmark

- In development since the early 90's
  - Better performance measure than a Linpack run for ECMWF's numerical weather prediction applications
- IFS RAPS14 is based on IFS cycle 41R2
  - Includes TL159, TCO639 and TCO1279 models
- Bitwise reproducible results expected
  - With RAPS14, issues either with the compiler or MKL seemed to prevent reproducibility (even on a Xeon) → Opted to get a performance baseline instead

IFS RAPS14 benchmark: statistics

- 2.4M lines of code
  - Nearly flat profile
- Well parallelized with MPI and OpenMP
  - RAPS14/TL159: ~6% in MPI library, ~90% in OpenMP parallel regions
  - RAPS14/TL159: **48.7 Gflops/sec** (~4% of peak of a dual Intel® Xeon® 2697v4)



IFS RAPS14/TL159 benchmarks

KNL runtime configuration

- Optimal runtime configuration found with a search through the parameter space of MPI ranks, OpenMP threads, **`NPROMA`** and **`NRPROMA`**
  - Optimal parameters for KNL rather different from the optimal parameters for Xeon
- 2MB pages and **`tbbmalloc_proxy`** library beneficial for both Xeon and KNL
  - For KNL the performance impact more pronounced, up to ~15-20%

KNL code optimization effort

- AVX-512 vectorization enabled with **`-xMIC-AVX512`** compiler flag
  - With **`–O3`** the compiler too aggressive on optimizations for some routines, switched to **`–O2`** instead
  - In some cases **`–vec-threshold0`** flag used to change compiler heuristics and ensure vectorization
- Due to *assumed* dependencies the compiler failed to vectorize some of the key hotspots
  - Added **`!DIR$ IVDEP`** to ~10 routines, one loop rewritten
- *Less than* **100 lines of code modified** in total!

Benchmark test systems*

Intel® Xeon® 2697v4

- 2 sockets, 18 cores/socket, 36 cores, 72 threads, 2.3Ghz
- DDR4 64GB 2400Mhz
- TDP 145W/socket, 290W in total

Intel® Xeon Phi™ 7250

- 1 socket, 68 cores, 272 threads, 1.4Ghz
- DDR4 96GB 2400Mhz
- 16GB of MCDRAM
- TDP 215W

SOFTWARE AND SERVICES

* For a full list of configuration options, see Single node configuration

IFS RAPS14 runtime configuration*

Intel® Xeon® 2697v4

- 18 MPI tasks
- 2 threads per task
- **NPROMA=16**
- **NRPROMA=4**

Intel® Xeon Phi™ 7250

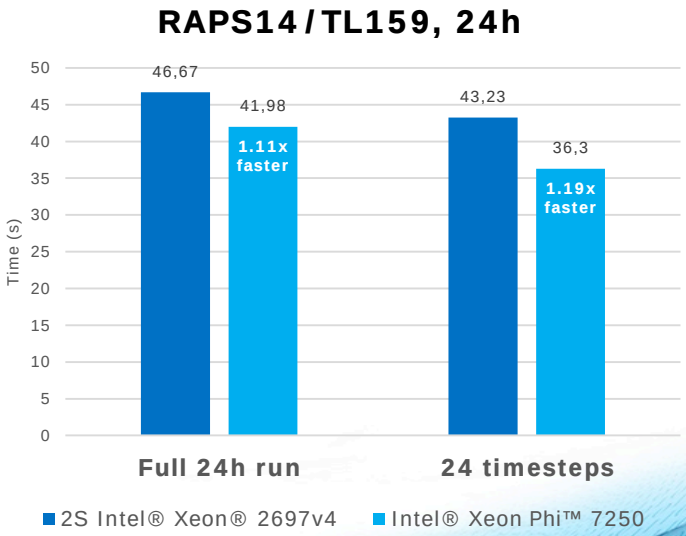
- 34 MPI tasks
- 4 threads per task
- **NPROMA=48**
- **NRPROMA=8**
- Quadrant cluster mode, cache memory mode

SOFTWARE AND SERVICES

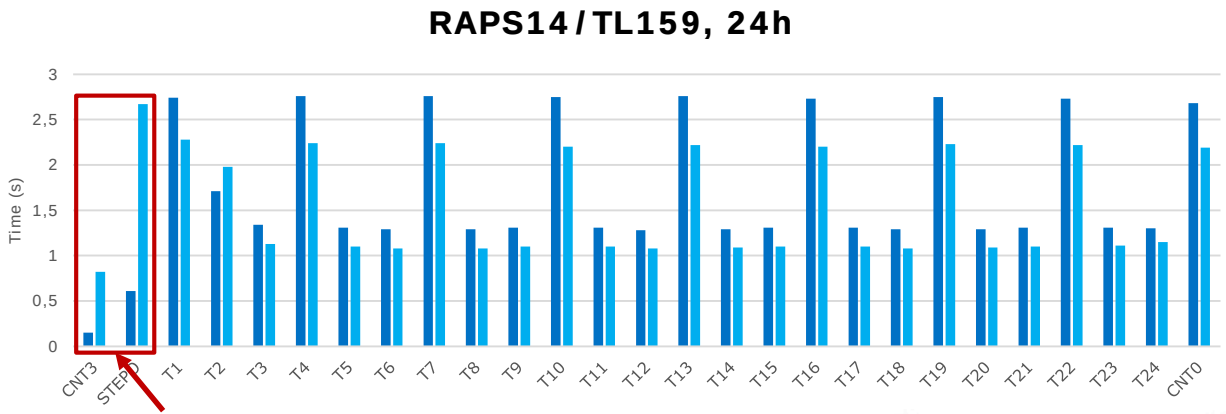
* For a full list of configuration options, see Single node configuration

Time to solution

- Best known compiler / runtime settings and *the same source code* used for both systems
  - **24h run**: wall-clock time for the whole run as given by IFS RAPS14
  - **24 timesteps**: total wall-clock time for 16 regular time steps and 8 radiation time steps



Time to solution, time step breakdown



SOFTWARE AND SERVICES

Time to solution, main hotspots

Intel® Xeon® 2697v4

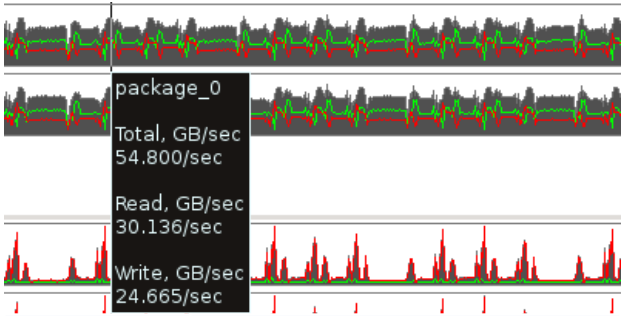
Function	CPU time	CPU time / thread	Instr. retired	CPI
cloudsc	101,82	2,83	4,02E+11	0,68
radlswr	90,12	2,50	1,56E+11	1,57
[libiomp5.so]	73,37	2,04	2,07E+11	0,98
srtm_spcvrt_mcica	69,04	1,92	3,42E+11	0,54
cpg	65,30	1,81	4,75E+10	3,73
[libmkl_avx2.so]	47,38	1,32	1,98E+11	0,64
laitri	46,59	1,29	7,22E+10	1,75
laitli	43,96	1,22	4,63E+10	2,54
rrtm_rtrn1a_140gp_mcica	39,98	1,11	1,12E+11	0,96
__intel_avx_rep_memset	39,86	1,11	6,06E+10	1,76

Intel® Xeon Phi™ 7250

Function	CPU time	CPU time / thread	Instr. retired	CPI
cloudsc	407,72	3,00	2,10E+11	2,86
[libmpi.so.12.0]	298,86	2,20	2,40E+11	1,86
srtm_spcvrt_mcica	246,48	1,81	1,41E+11	2,61
radlswr	181,38	1,33	8,35E+10	3,25
[libmkl_avx512_mic.so]	143,68	1,06	6,90E+10	3,00
srtm_reftra	124,26	0,91	7,32E+10	2,56
__intel_mic_avx512f_memcpy	119,63	0,88	4,10E+10	4,42
cloudvar	111,65	0,82	8,29E+10	1,99
laitri	108,33	0,80	3,40E+10	4,71
mcica_cld_gen	108,11	0,79	2,99E+10	5,45

Time to solution, memory bandwidth

Intel® Xeon® 2697v4



	Bandwidth (GB/sec)
Peak	110-120
Average	80-90

Intel® Xeon Phi™ 7250

	Bandwidth (GB/sec)
Peak	340-360
Average	200-250

SOFTWARE AND SERVICES

Energy to solution

Intel® Xeon® 2697v4

Average power consumption: **420W**

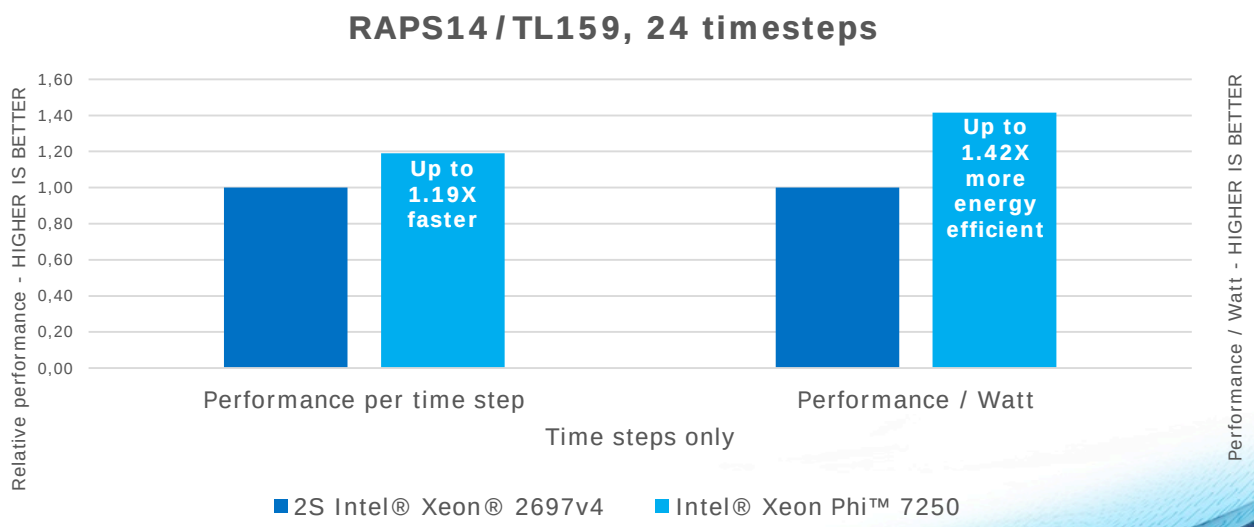
Intel® Xeon Phi™ 7250

Average power consumption: **353W**

SOFTWARE AND SERVICES

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For configuration details, see single node configuration.

KNL performance summary



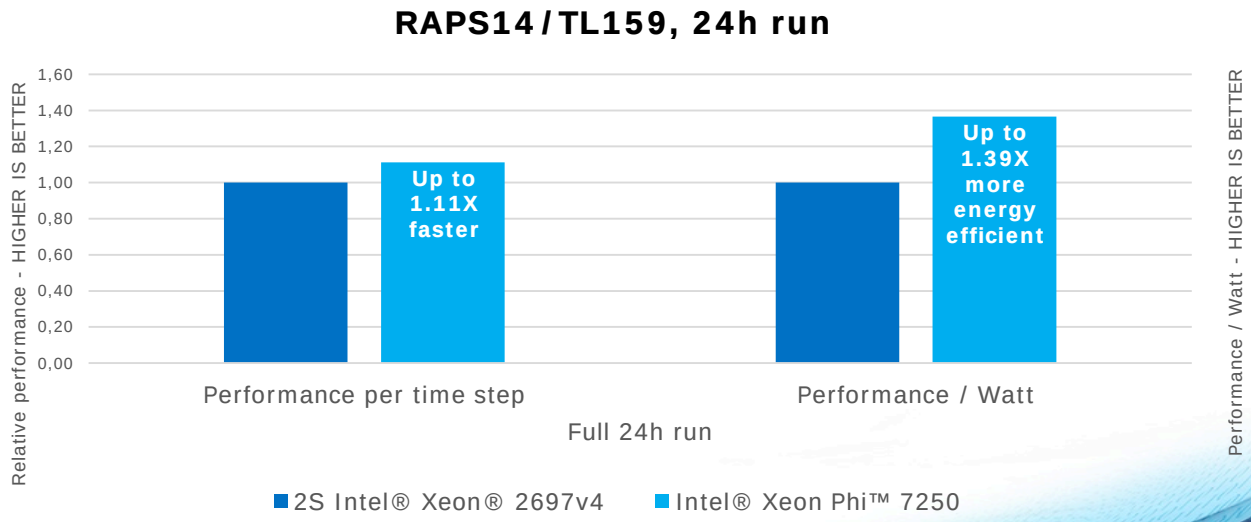
SOFTWARE AND SERVICES

Conclusions

- Intel® Xeon Phi™ processor offers better **performance** and **energy efficiency** while maintaining an established codebase
  - Performing serial IO or scalar operations should be avoided
  - Code optimizations benefit Intel® Xeon® processors as well
- Future work: bit reproducible results, further multi-node experiments with TCO639

Backup

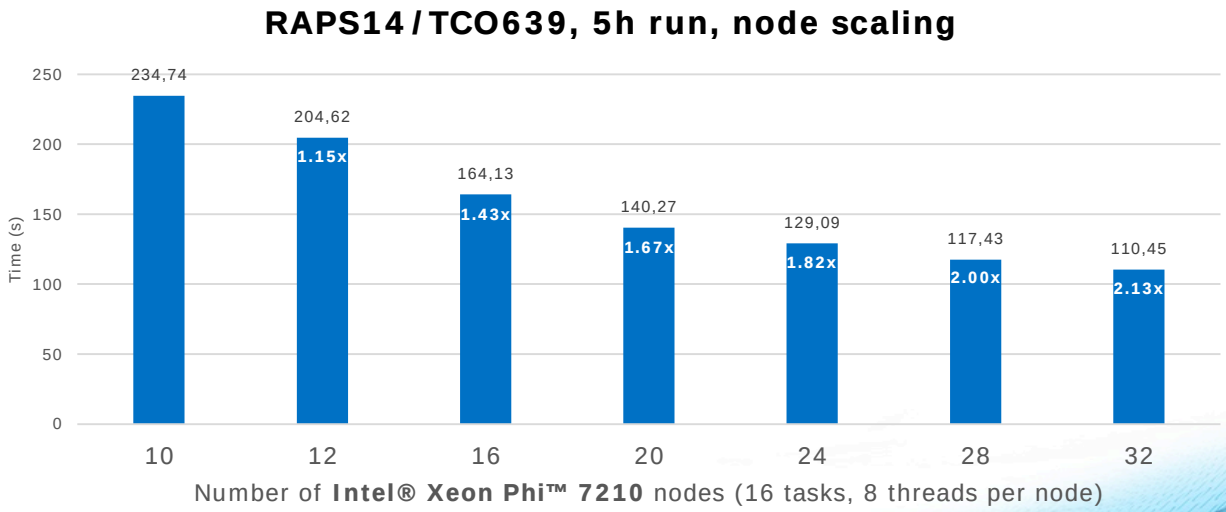
KNL performance summary



SOFTWARE AND SERVICES

Preliminary: TCO639 node scaling

Results computed on ECMWF's Cray* XC40* KNL partition



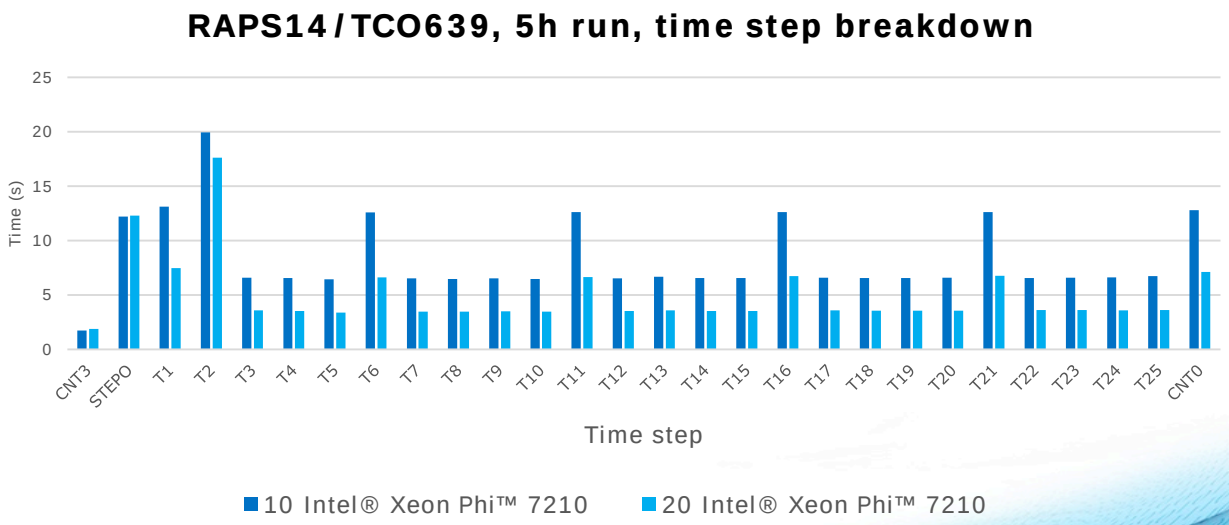
SOFTWARE AND SERVICES

*Other names and brands may be claimed as the property of others.

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For configuration details, see multi-node configuration.

Preliminary: TCO639 node scaling

Results computed on ECMWF's Cray* XC40* KNL partition



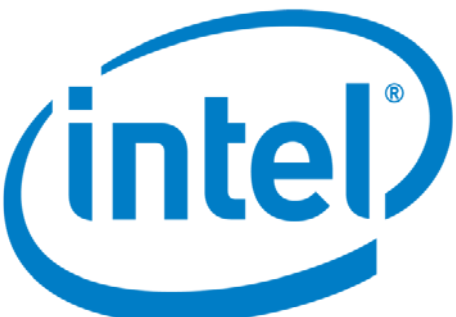
SOFTWARE AND SERVICES

Configuration details: single node

Intel® Xeon® processor E5-2697 v4: Dual Intel® Xeon® processor E5-2697 v4 2.3Ghz, 18 cores/socket, 36 cores, 72 threads (HT and Turbo ON), BIOS GRRFSDP1.86B0271.R00.1510301446, DDR4 64 GB, 2400 MHz, RHEL 7.2, 1.0 TB SATA drive WD1003FZEX-00MK2A0, /proc/sys/vm/nr_hugepages=8000, Intel Compiler 2017, tbbmalloc_proxy
Intel® Xeon® settings: 18 MPI tasks, 2 OpenMP threads per task, NRPROMA=-4, NPROMA=-16. Environment variables: OMP_STACKSIZE=48M, KMP_AFFINITY=compact, , KMP_BLOCKTIME=12, I_MPI_FABRICS=shm, I_MPI_PIN_DOMAIN=omp, I_MPI_PIN_PROCESSOR_LIST=allcores, I_MPI_PIN_ORDER=bunch, TBB_MALLOC_USE_HUGE_PAGES=1
Intel® Xeon Phi™ processor 7250: Intel® Xeon Phi™ processor 7250, 68 cores, 272 threads, 1.4 GHz base core freq., Turbo ON, 1.7 GHz uncore freq., MCDRAM 16 GB 7.2 GT/s, BIOS GVPRCRB1.86B.0011.R02.1608040407, DDR4 96 GB 2400MHz, Quadrant cluster mode, MCDRAM cache memory mode, RHEL 7.2, XPPSL 1.4.1, 1.0 TB SATA drive WD1003FZEX-00MK2A0, /proc/sys/vm/nr_hugepages=24000, Intel Compiler 2017, tbbmalloc_proxy
Intel® Xeon Phi™ settings: 34 MPI tasks, 4 OpenMP threads per task, NRPROMA=-8, NPROMA=-48. Environment variables: OMP_STACKSIZE=48M, KMP_AFFINITY=compact, KMP_BLOCKTIME=12, KMP_HW_SUBSET=2t, I_MPI_FABRICS=shm, I_MPI_SHM_LMT=direct, I_MPI_PIN_ORDER=scatter, TBB_MALLOC_USE_HUGE_PAGES=1
Compiler settings: Vectorization flags for Intel® Xeon®: -xCORE-AVX2. Vectorization flags for Intel® Xeon Phi™: –xMIC-AVX512.
Recipe: IFS RAPS14 is available under a license from ECMWF. A full list of code modifications and compiler settings used has been delivered and is available to licensed developers from ECMWF. The same improved source code was used for testing both Intel® Xeon® and Intel® Xeon Phi™.
Power Data: Total system wall power is measured out-of-band over iPMI interface, polling the BMC chip every 0.1 seconds. Energy usage is matched to the average of internally timed code segments to arrive at performance per Watt estimate.
Average time step length: Average time step length computed by averaging the timings for 16 normal and 8 radiation time steps in a 24h forecast run.
Average energy consumption: Dual Intel® Xeon® processor E5-2697 v4 2.3Ghz, 18 cores/socket 420W (418W 24h run),
Intel® Xeon Phi™ processor 7250, 68 cores (272 threads), 1.4 GHz 352W (340W 24h run).

Configuration details: multi-node

Multi-node benchmarks computed on Cray* XC40* partition at ECMWF with Intel® Xeon Phi™ processor 7210.
Intel® Xeon Phi™ processor 7210: Intel® Xeon Phi™ processor 7210, 64 cores, 256 threads, 1.3 GHz base core freq., Turbo ON, 1.6 GHz uncore freq., MCDRAM 16 GB 7.2 GT/s, BIOS GVPRCRB1.86B.0011.R02.1608040407, DDR4 96 GB 2400MHz, Quadrant cluster mode, MCDRAM cache memory mode, Cray CCE 8.5.3, Intel Compiler 2017, tbbmalloc_proxy
Intel® Xeon Phi™ settings, Cray XC40, TL159: 32 MPI tasks, 4 OpenMP threads per task, NRPROMA=-8, NPROMA=-48. Environment variables: OMP_STACKSIZE=96M, KMP_AFFINITY=compact, KMP_BLOCKTIME=12, TBB_MALLOC_USE_HUGE_PAGES=1
Intel® Xeon Phi™ settings, Cray XC40, TCO639: 16 MPI tasks, 8 OpenMP threads per task, NRPROMA=-8, NPROMA=-48. Environment variables: OMP_STACKSIZE=64M, KMP_AFFINITY=compact, KMP_BLOCKTIME=12, TBB_MALLOC_USE_HUGE_PAGES=1
Compiler settings: Vectorization flags for Intel® Xeon®: -xCORE-AVX2. Vectorization flags for Intel® Xeon Phi™: –xMIC-AVX512.
Recipe: IFS RAPS14 is available under a license from ECMWF. A full list of code modifications and compiler settings used has been delivered and is available to licensed developers from ECMWF. The same improved source code was used for testing both Intel® Xeon® and Intel® Xeon Phi™.
Cray* XC40*, TL159 ALPS settings: -m1500h -d 4 -j 2 -N 32 -cc depth -r 1 --p-state=1301000
Cray* XC40*, TCO639 ALPS settings: -m2500h -d 8 -j 2 -N 16 -cc depth -r 1 --p-state=1301000


intel®