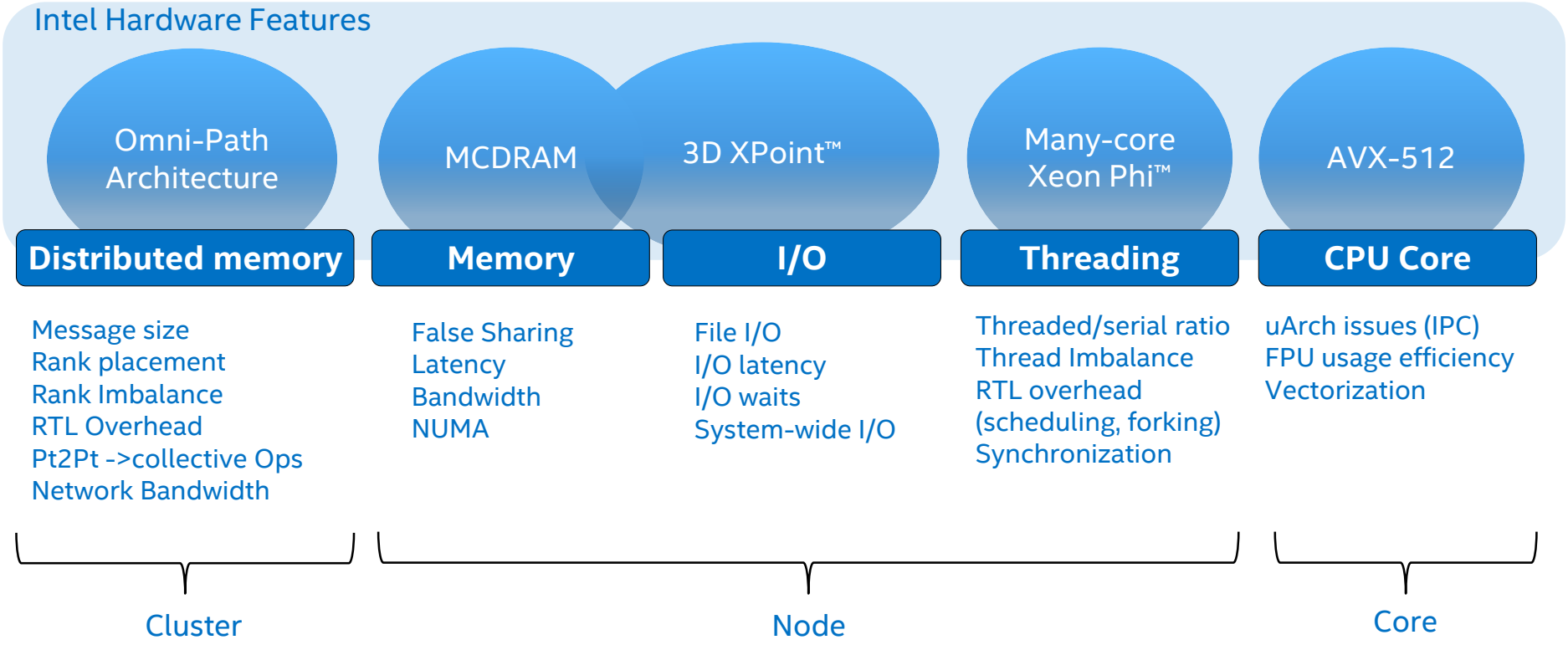
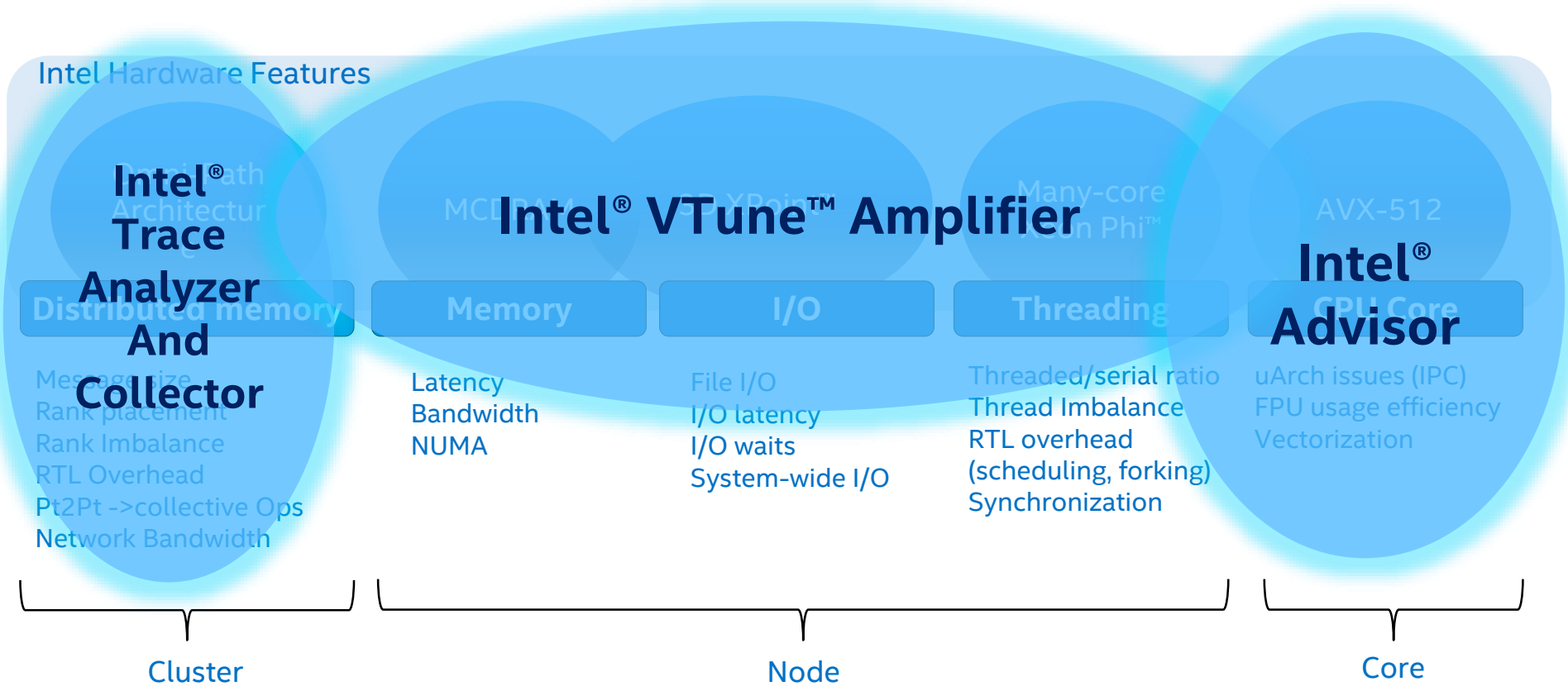
CONFIGURATION-FREE PROFILING AT SCALE

Bei Wang, Dmitry Prohorov and Carlos Rosales

Aspects of Application Performance



Intel® Tools Covering the Aspects



Before diving Into a particular tool …

- How to assess that I have **potential in performance** tuning?
- **Which tool** should I use first?
- What to use at **large scale** avoiding being overwhelmed with huge trace size, post processing time and collection overhead?
- How to **quickly** evaluate environment settings or incremental code changes?

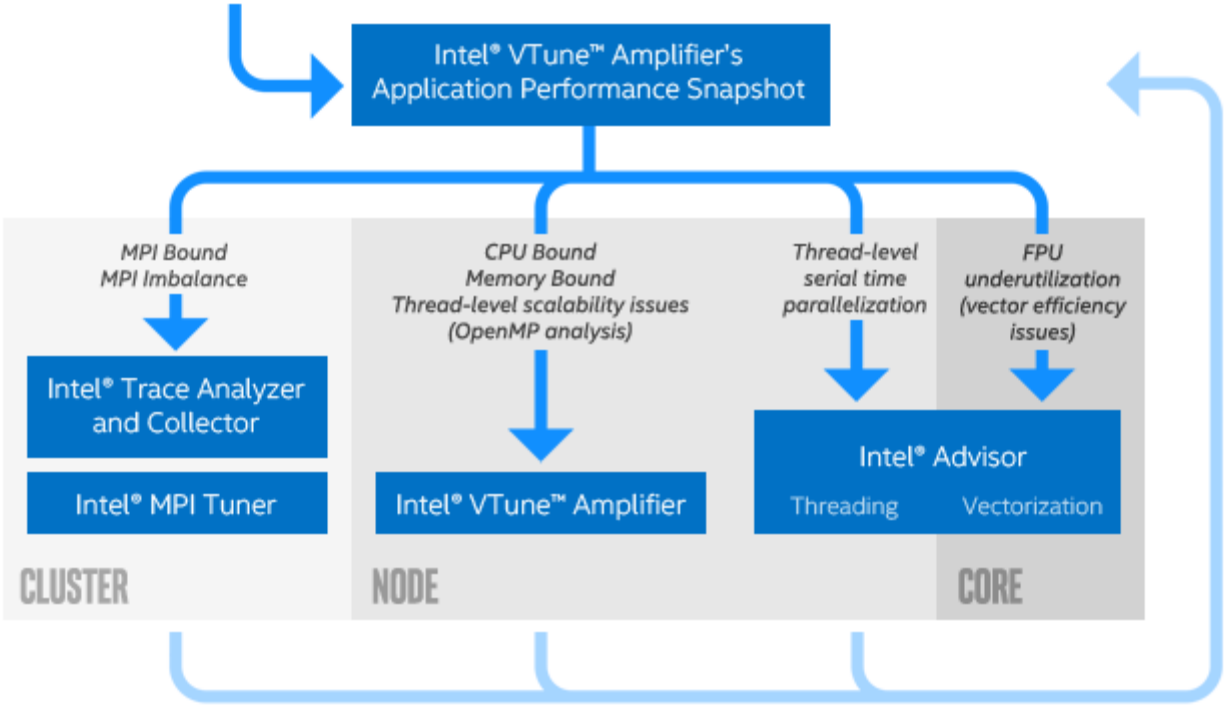
VTune™ Amplifier's Application Performance Snapshot

Application Performance Snapshot (APS)

- High-level **overview** of application performance
- Identify primary optimization areas and **next steps** in analysis
- **Easy** to install, run, explore results with CL or HTML reports
- **Scales** to large jobs
- Multiple ways to get started:
  - Part of Intel® Parallel Studio XE or VTune™ Amplifier distributions
  - Separate **free** download (~100MB) from APS page

  https://software.intel.com/sites/products/snapshots/application-snapshot/

Performance Optimization Workflow based on APS



APS Usage

Setup Environment

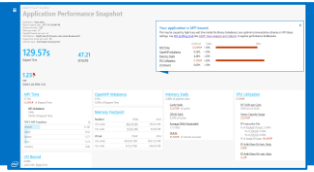
- >source <APS_Install_dir>/apsvars.sh

Run Application

- >**aps** <application and args>
- MPI: >mpirun <mpi options> **aps** <application and args>

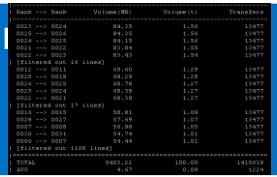
Generate Report on Result Folder

- >**aps** –report <result folder>

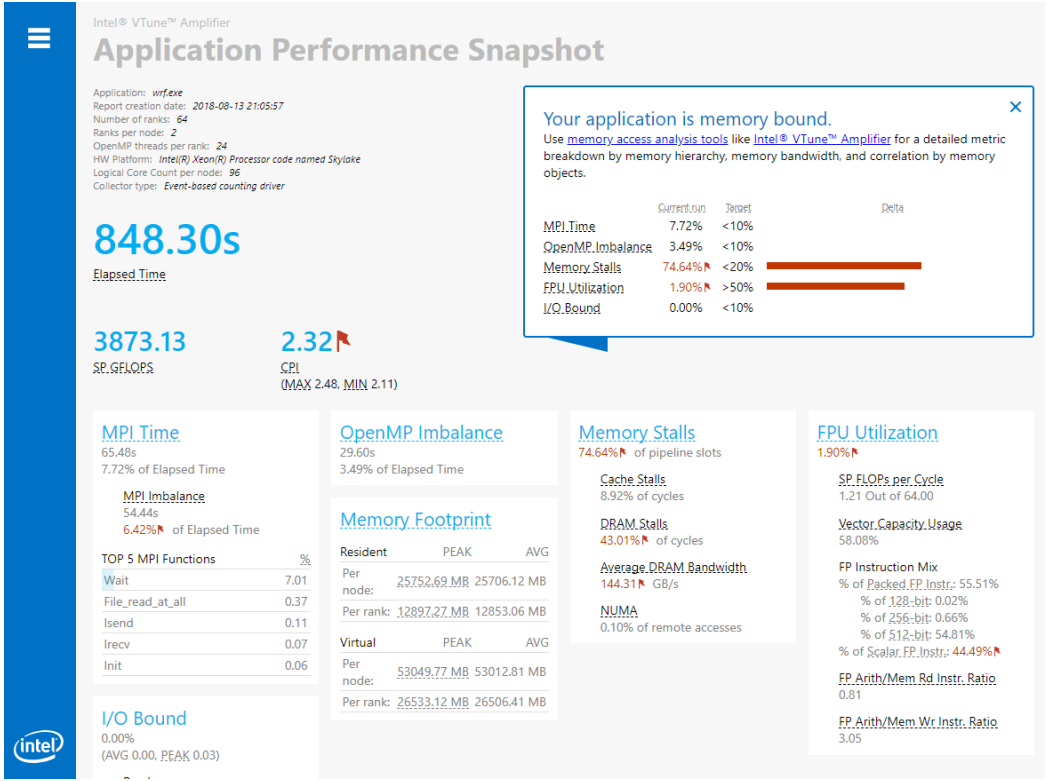


Generate CL reports with detailed MPI statistics o… Result Folder

- **aps-report** –<option> <result folder>



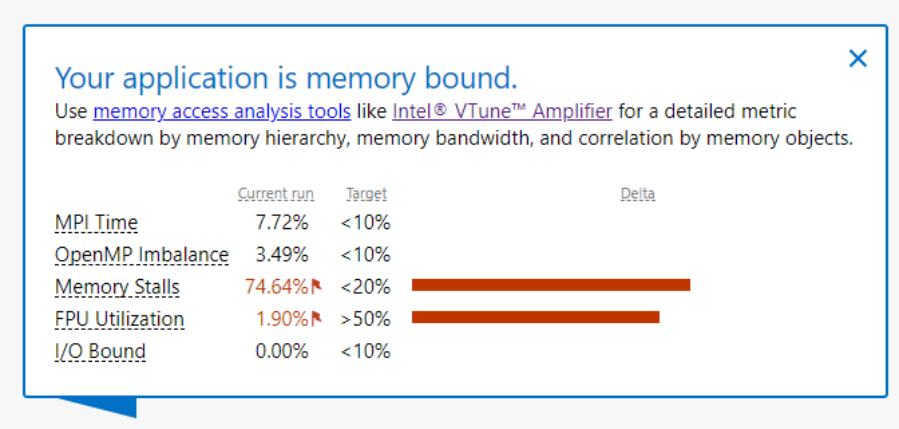
APS HTML Report



Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

APS HTML Report Breakdown - Overview

- Overview shows all areas and relative impact on code performance
- Provides recommendation for next step in performance analysis
- “X” collapses the summary, removing the flags (objective numbers only)



Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

APS HTML Report Breakdown – Parallel Runtimes

MPI Time

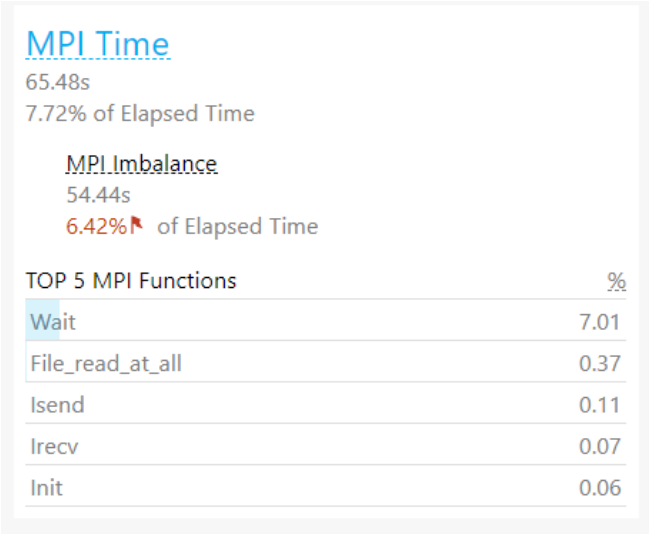
- How much time was spent in MPI calls
- Averaged by ranks with % of Elapsed time
- Available for MPICH-based libraries

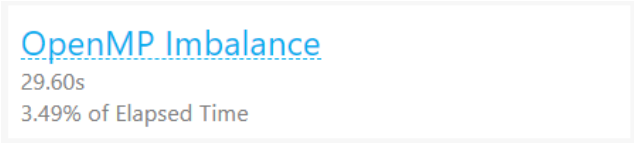
MPI Imbalance

- Unproductive time spent in MPI library waiting for data
- Available for Intel® MPI

OpenMP* Imbalance

- Time spent at OpenMP* Barriers normalized by number of threads
- Available for Intel® OpenMP*





Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

APS HTML Report Breakdown – Memory Access

- Memory stalls measurement with breakdown by cache and DRAM
- Average DRAM Bandwidth(1)
- NUMA ratio
- Intel® Xeon Phi™:
  - Back-end stalls with L2-demand access efficiency
  - Average DRAM AND MCDRAM Bandwidth (1)

(1) Average DRAM and MCDRAM bandwidth collection is available with Intel driver or perf system wide monitoring enabled on a system

Memory Stalls
74.64% of pipeline slots

Cache Stalls
8.92% of cycles

DRAM Stalls
43.01% of cycles

Average DRAM Bandwidth
144.31 GB/s

NUMA
0.10% of remote accesses

Back-End Stalls
35.00% of pipeline slots

L2 Hit Bound
13.85% of cycles

L2 Miss Bound
27.28% of cycles

Average DRAM Bandwidth
3.59 GB/s

Average MCDRAM Bandwidth
102.69 GB/s

Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

APS HTML Report Breakdown – vectorization

FPU Utilization based on HW-event statistics with

- Breakdown by vector/scalar instructions
- Floating point vs memory instruction ratio

FPU Utilization
1.90%

SP FLOPs per Cycle
1.21 Out of 64.00

Vector Capacity Usage
58.08%

FP Instruction Mix
% of Packed FP Instr.: 55.51%
% of 128-bit: 0.02%
% of 256-bit: 0.66%
% of 512-bit: 54.81%
% of Scalar FP Instr.: 44.49%

FP Arith/Mem Rd Instr. Ratio
0.81

FP Arith/Mem Wr Instr. Ratio
3.05

Intel® Xeon Phi™: SIMD Instr. per Cycle

- Scalar vs. vectorized instructions

SIMD Instr. per Cycle
0.13

FP Instruction Mix
% of Packed SIMD Instr.: 64.60%
% of Scalar SIMD Instr.: 35.40%

Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

TESTING WITH GTC-P AND WRF

About WRF

Weather **R**esearch and **F**orecasting Model

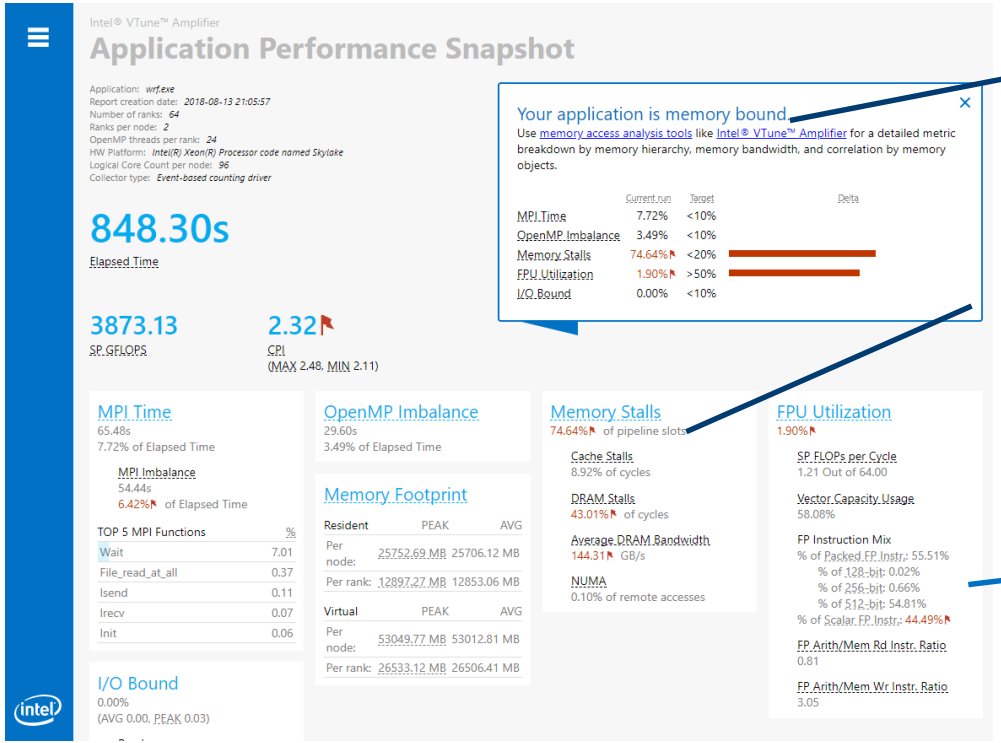
- Mesoscale numerical weather prediction using Finite Differences
- Operational model at National Centers for Environmental Prediction (NCEP)
- Parallelized using MPI and OpenMP*

Tested configurations

- ASRC test case, limited to 10 minute evolution.
- Pnetcdf build without tiling.
- Timings from “Timing for main” in rsl output files, discarding input read at step 1.
- Hybrid build of version 3.8.1 Using Intel® Compiler 2018.2 and Intel® MPI Library 2018.2

https://www.mmm.ucar.edu/weather-research-and-forecasting-model

Typical HTML Report for WRF



Main bottleneck identified and next steps suggested

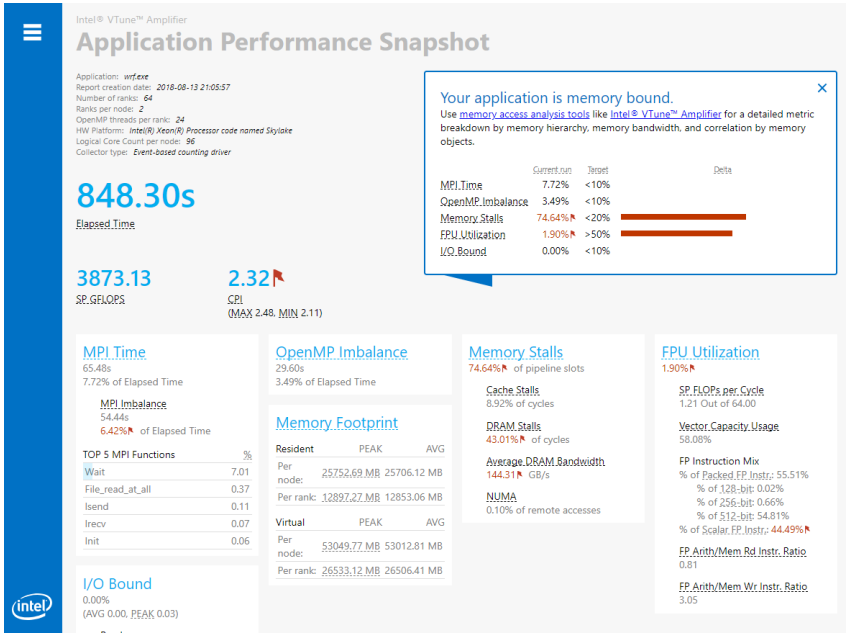
High BW use and high stalls - Bad

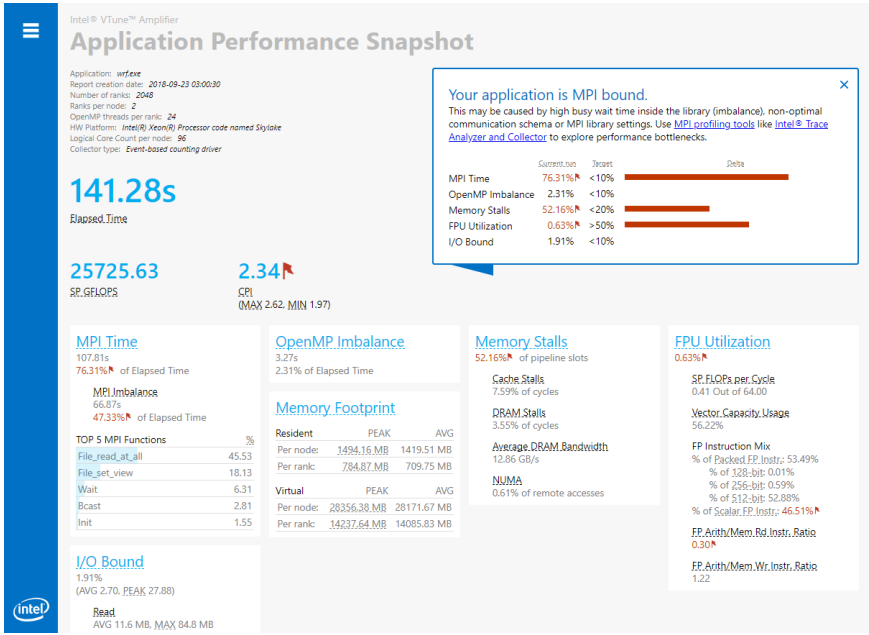
Negligible remote accesses - Good

Only 50% SIMD - Secondary area of improvement

Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

Characteristics Change at Scale





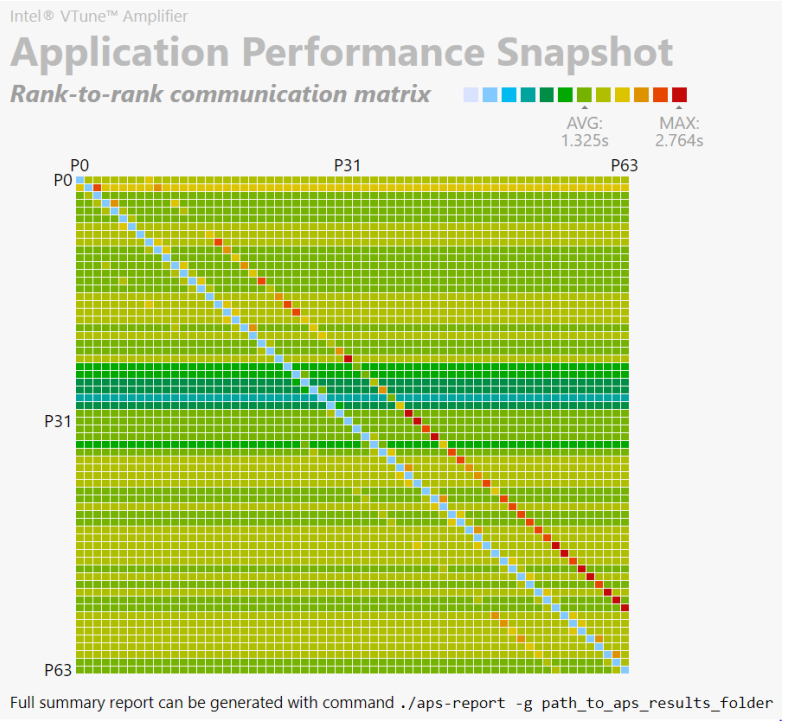
Configuration Information - See Configuration Details

Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.

Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

Rank to Rank Communication Report (WRF)

New in 2019



A rank-to-rank communication heatmap based on aggregated communication time can be generated

Many other reports are available in post-processing from the command line

- Message Size Summary
- MPI Time per Rank
- Collective Communications

Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

About GTC-P

Gyrokinetic **T**oroidal **C**ode - **P**rinceton

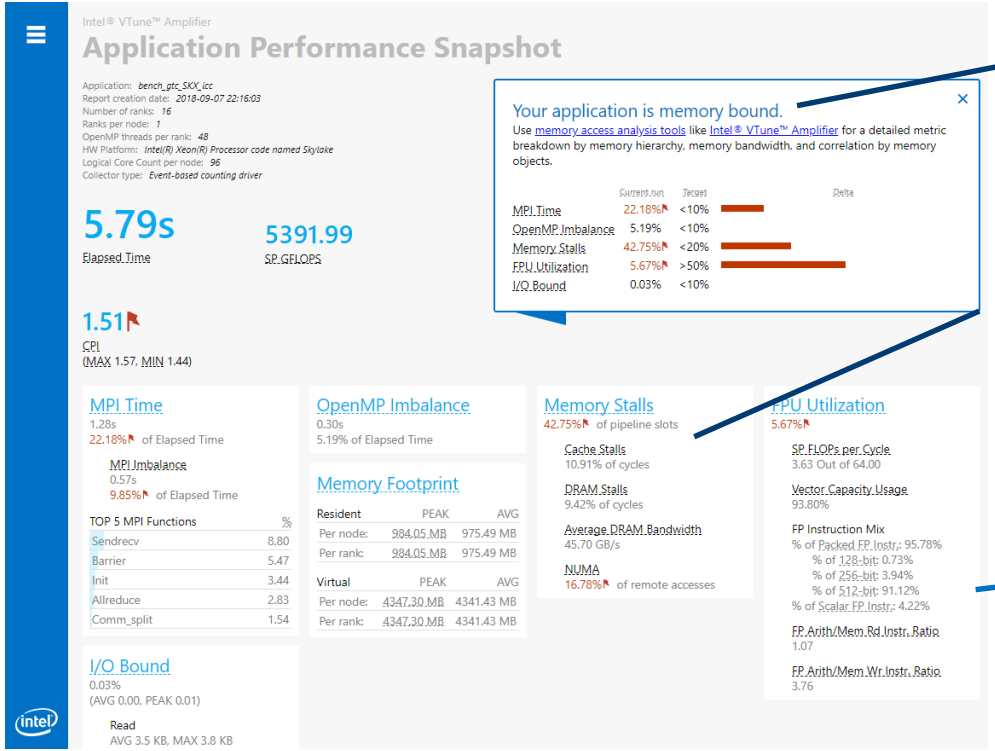
- Highly scalable Particle-In-Cell code solving 5D Vlasov-Poisson
- Run on many TOP 10 systems
- Parallelized with MPI and OpenMP*
- No external library dependencies
- Designed with architectural flexibility in mind

Tested configurations

- Standard benchmark workloads
  - 16 Nodes - A.txt 100 16
  - 64 Nodes - B.txt 100 16
  - 256 Nodes - C.txt 100 16
  - 1024 Nodes - D.txt 100 16
- Timings from “Total_time” in standard output.
- Hybrid build using Intel® Compiler 2018.2 and Intel® MPI Library 2018.2.

https://extremescaleglobalpic.princeton.edu/gtcp

Typical HTML Report for GTC-P



Main bottleneck identified and next steps suggested

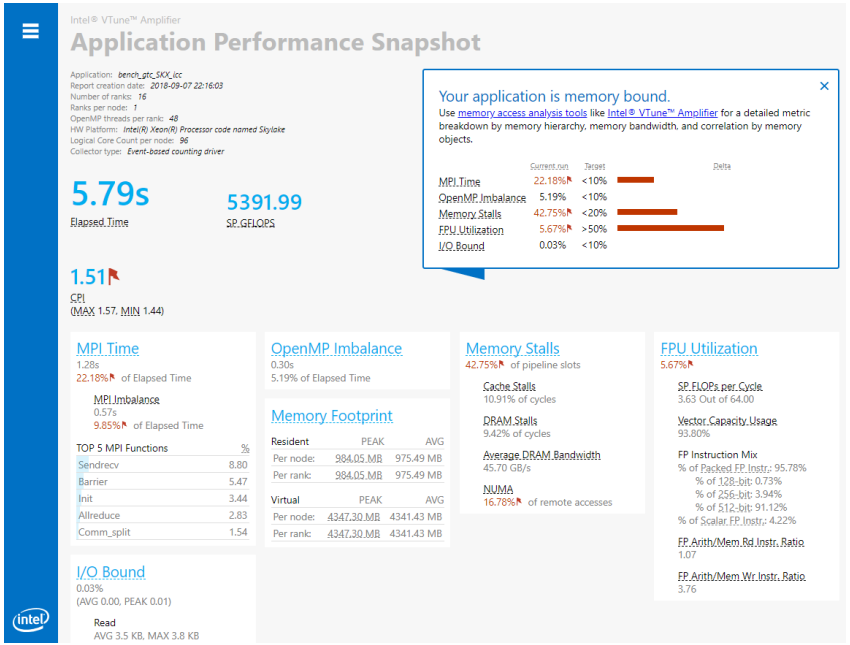
Medium BW use but high stalls - Bad

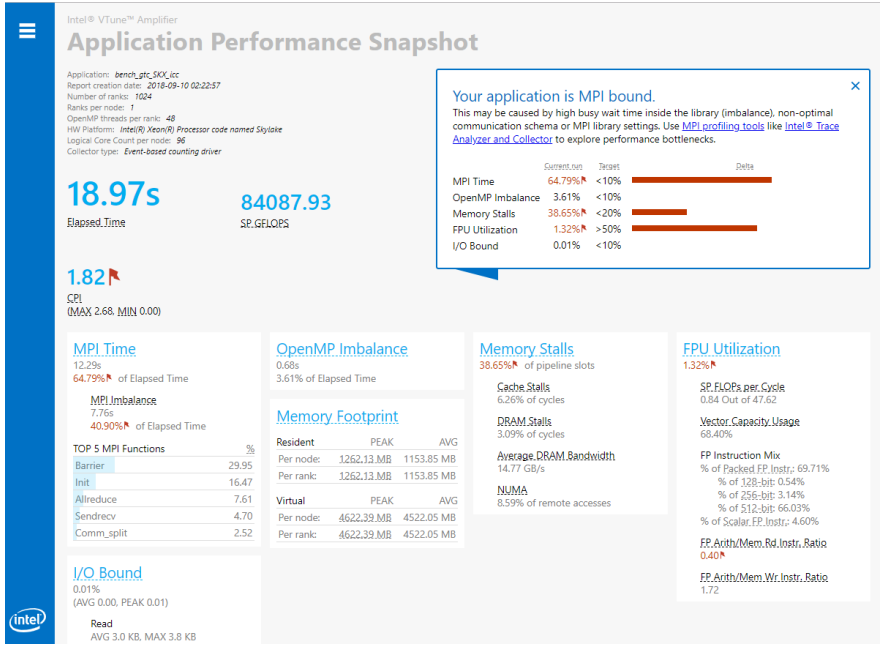
Some remote accesses - Not an immediate concern

Excellent vectorization

Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

Characteristics Change at Scale



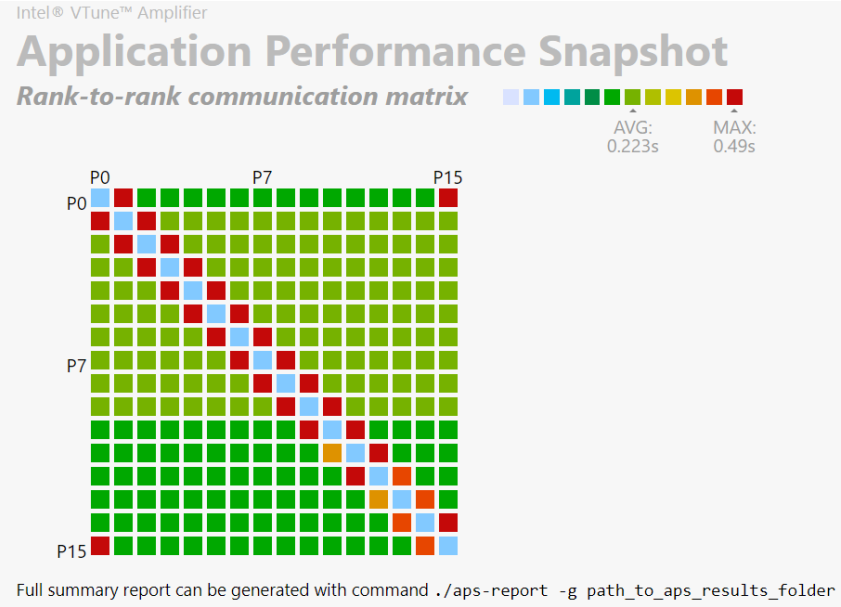


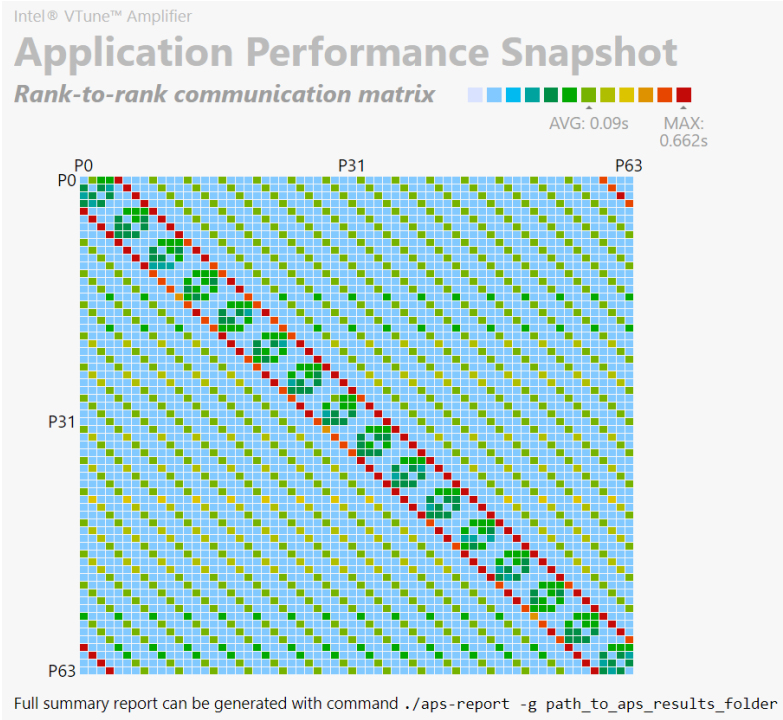
Configuration Information - See Configuration Details

Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.

Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

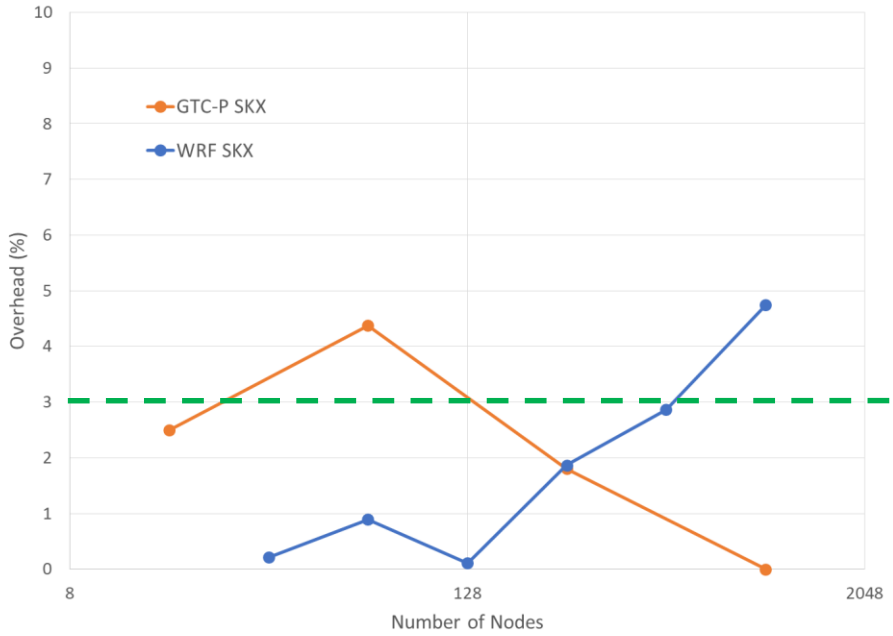
Rank to Rank Communication Report (GTC-P)





Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

Overhead for GTC-P and WRF on Intel® Xeon® Scalable processor (Skylake)

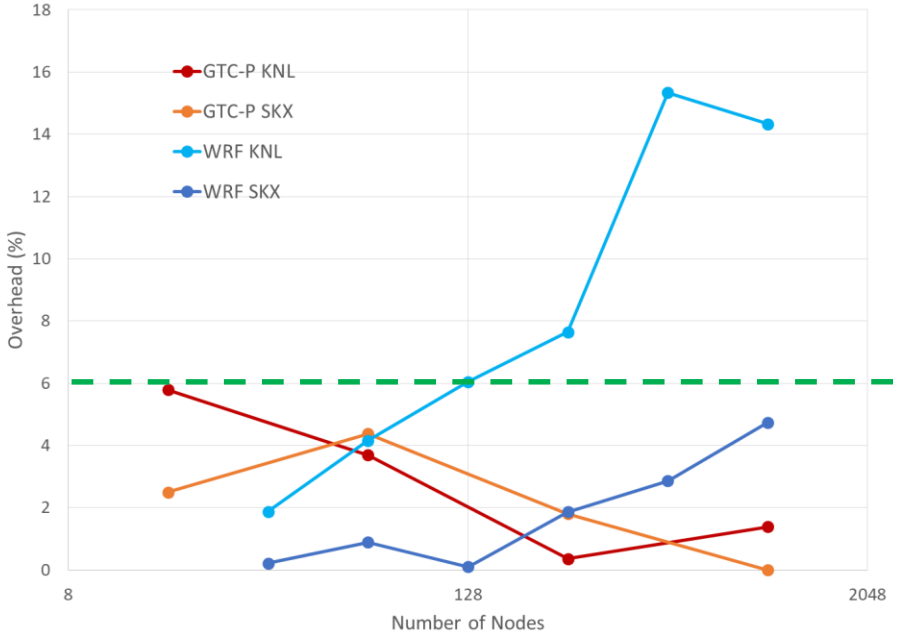


Most collections show average overhead below 3%

Suspect IO related issues on large scale WRF runs influence overhead

Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; intel® Fortran Compiler XE 18.0.2.199 for Linux*

Overhead for GTC-P and WRF on Intel® Xeon® Scalable processor (Skylake) and Intel® Xeon Phi™ (KNL)



Most collections show average overhead below 6%

Suspect IO related issues on large scale WRF runs influence overhead

Configuration Information - See Configuration Details
Hardware: Intel® Xeon Phi™ CPU 7250 @ 1.40GHz; 96 GB DDR4 RAM, configured in Cache-Quadrant mode (KNL). Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; intel® Fortran Compiler XE 18.0.2.199 for Linux*

Collection Control API (2018 Update 2 and Newer)

To measure a particular application phase or exclude initialization/finalization the following control API calls may be used

MPI applications:

- Pause: MPI_Pcontrol(0)
- Resume: MPI_Pcontrol(1)

Non-MPI applications:

- Pause: __itt_pause()
- Resume: __itt_resume()

Use aps “-start-paused” option to start application without profiling and skip initialization phase

Summary

- Application Performance Snapshot provides overall program performance characteristics with moderate overhead and no configuration requirements
- Simple presentation provides simple means for tracking performance changes with scale and code modifications.
- Overhead for production workloads on Intel® Xeon® Scalable processors (Skylake) is typically below 3% and on Intel® Xeon Phi™ (KNL) typically below 6%
- Instances of high overhead (over 5% on Skylake, over 15% on KNL) have been observed and required further investigation
- Also, I need larger workloads to run at 1k nodes…

Configuration Information - See Configuration Details
Hardware: Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake/SKX). Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.
Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

Configuration Details

Hardware:
Intel® Xeon Phi™ CPU 7250 @ 1.40GHz; 96 GB DDR4 RAM, configured in Cache-Quadrant mode (KNL).
Intel® Xeon® Platinum 8160 @ 2.1 GHz; 192 GB DDR4 RAM (Skylake).
Intel® Omni-Path Host Fabric Interface, fat tree topology with 28/20 oversubscription.

Software: CentOS 7.4; Intel® Fabric Software 10.6.1.0.2; MPI Library 2018 Update 2; Intel® C++ Compiler XE 18.0.2.199 for Linux*; Intel® Fortran Compiler XE 18.0.2.199 for Linux*

Runtime (WRF):
- 2 MPI ranks / Node and 24 threads per rank (Skylake)
- 4 MPI Ranks / Node and 16 threads per rank (KNL)

Runtime (GTC-P):
- 1 MPI rank / node and 48 threads per rank (Skylake)
- 1 MPI rank / node and 128 threads per rank (KNL)

Legal Disclaimer & Optimization Notice

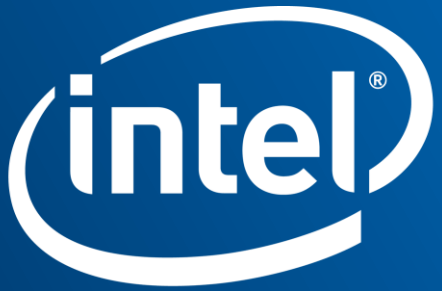
The benchmark results reported above may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.

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. For more complete information visit www.intel.com/benchmarks.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

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



Software