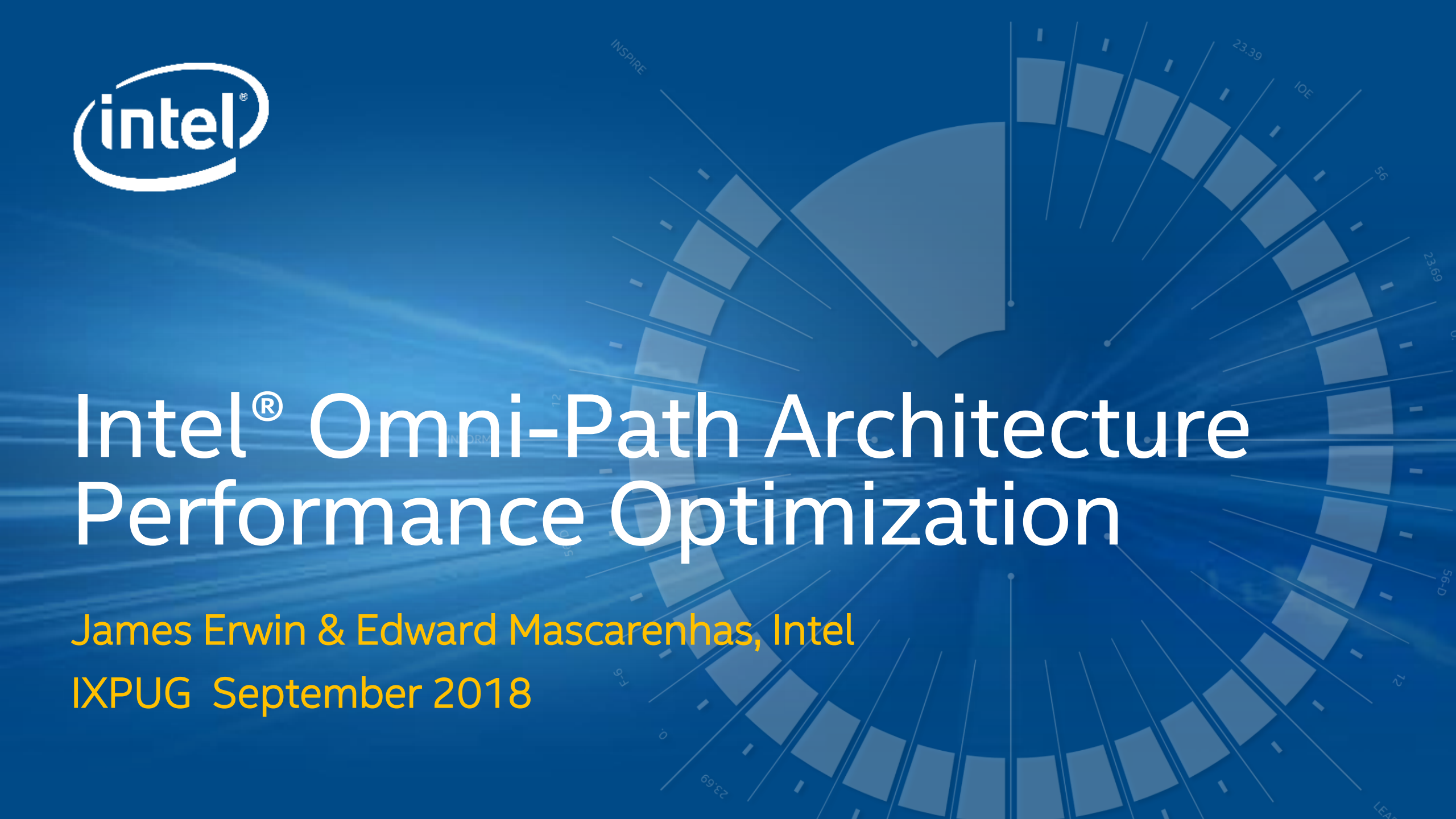


Intel® Omni-Path Architecture Performance Optimization

James Erwin & Edward Mascarenhas, Intel

IXPUG September 2018

Outline

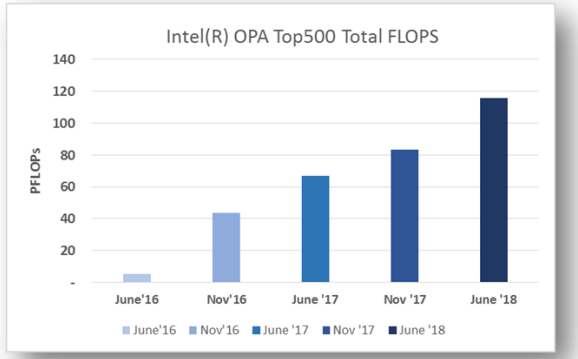
Intel® OPA performance at a glance

Tuning Intel® OPA performance

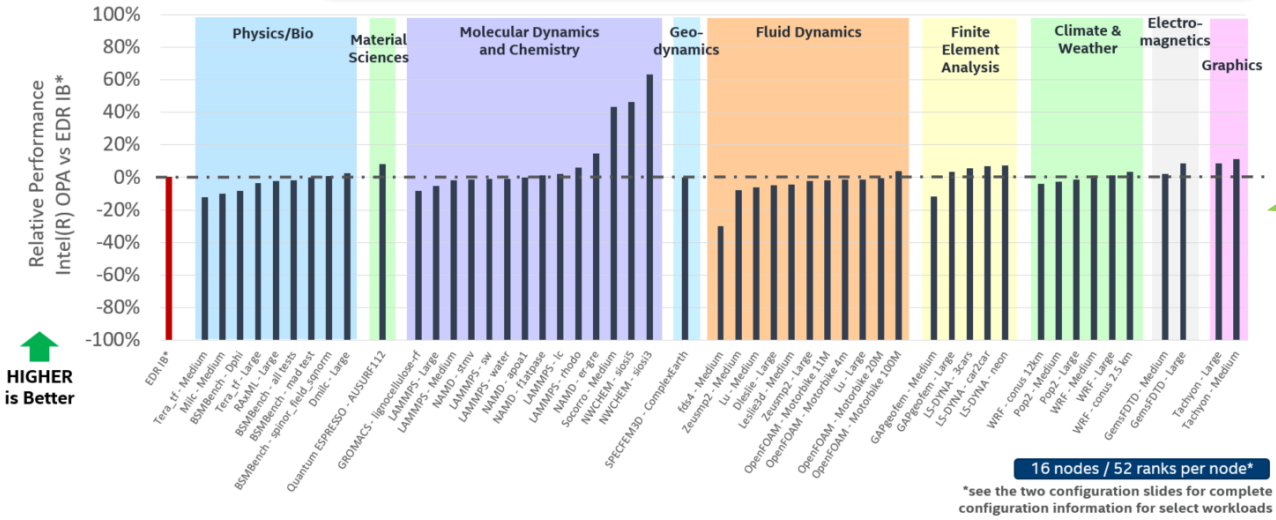
Next Steps & Where to get more information

Intel® OPA - Performance at a glance

High performance and reliable scalability at major supercomputing sites
39% increase in total FLOPs since Nov'17



TOP 500 The List.
www.top500.org



Compelling **HPC application performance**[1] versus 100Gb InfiniBand*

Better **Total Cost of Ownership**[2] vs 100Gb InfiniBand*, allowing for more compute and storage hardware within a fixed budget

intel OPTANE™ DC PERSISTENT MEMORY

1. Performance results are based on testing as of April 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. See full configuration at end of presentation.
2. See configurations at end of presentation

Performance Tuning Guide overview

BIOS Settings

OS Settings

OPA Driver settings

MPI

Verbs

IPoFabric/IP Router

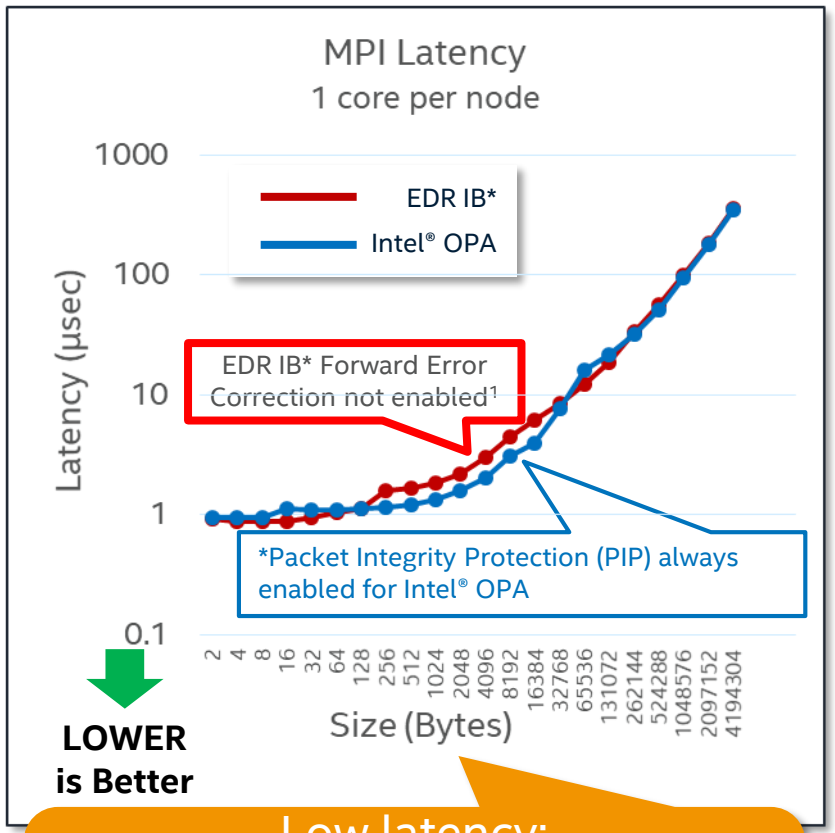
IRQ affinity/irqbalance

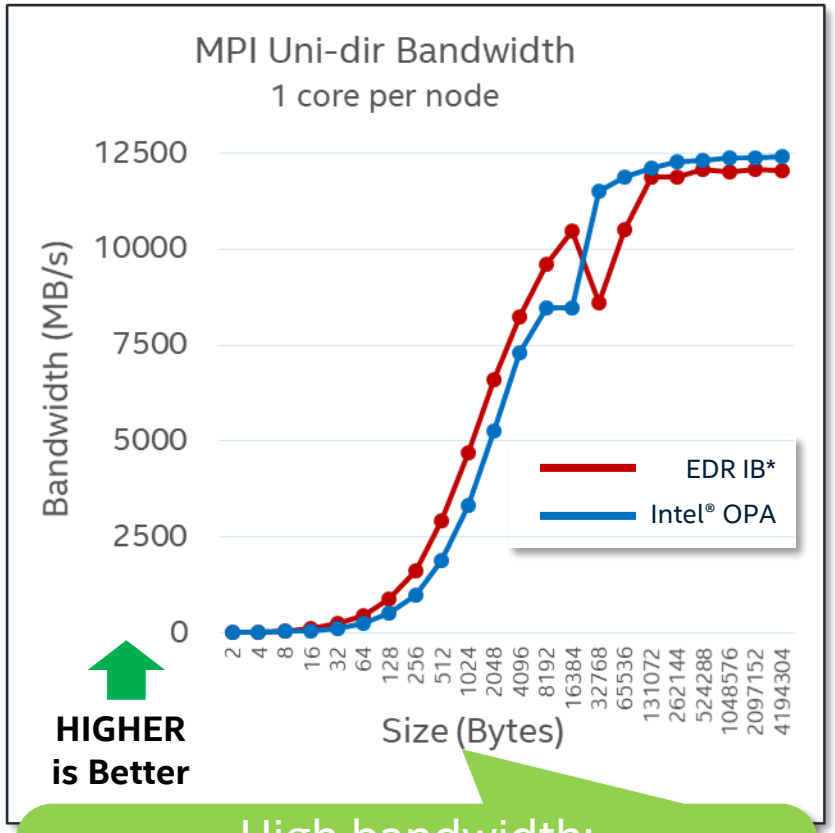
Please consult the Tuning Guide for detail beyond this short talk … many times optimal performance is achieved “out of the box”

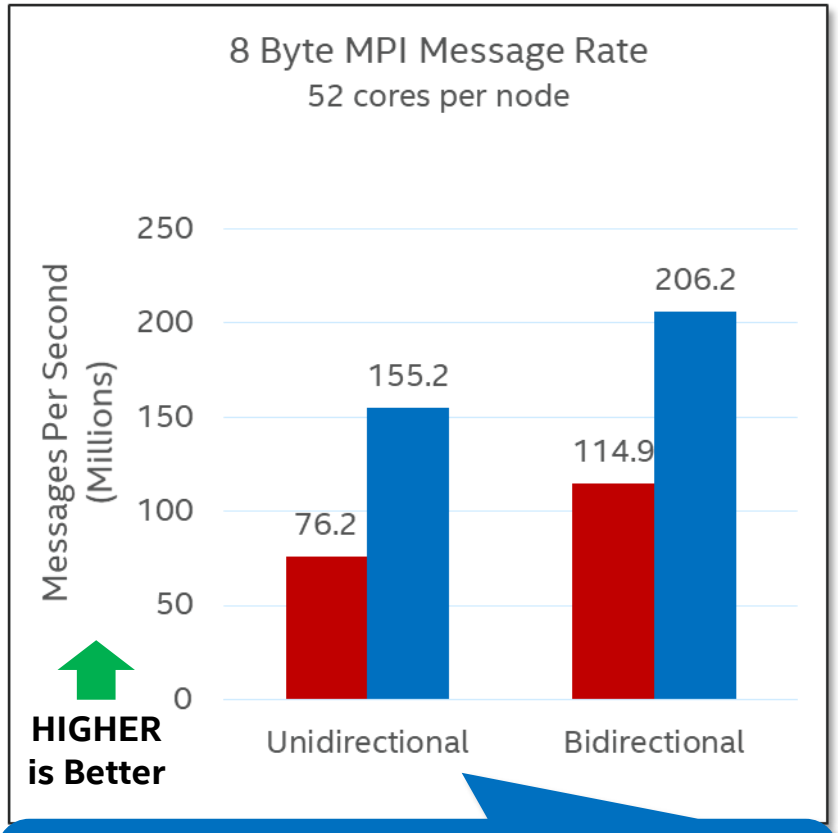
Tuning Intel® OPA Performance

- BIOS/OS
- Fabric health
- Intel® OPA Host Fabric Interface(HFI) driver
- Enhanced Fabric Features (routing, Quality of Service, …)
- Run time tunings, workload dependent

MPI Latency, Bandwidth, and Message Rate







Low latency:

- **BIOS:** CPU Power & Perf Policy: Performance or Balanced Performance
- **OS:** Performance governor with either ACPI or Intel Pstate frequency driver

High bandwidth:

- **Platform/BIOS:** Make sure HFI is enumerated as PCIe* Gen3 / x16 (lspci)
- **Fabric Links:** Check for degraded (1x, 2x, or 3x) links (opareport -o slowlinks)

High message rate:

- Follow tips for latency and bandwidth
- Use sequential core MPI rank pinning with PSM (use OFI/libfabric with Intel® MPI Library instead of TMI)

Always run your MPI applications over Intel® OPA with Performance Scaled Messaging (PSM)

Performance results are based on testing as of July 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
1. See configuration slide for complete test details.

Major Intel® OPA Tunings

For Applications

- Always use **Performance Scaled Messaging (PSM)** and not OFA/Verbs
- Use latest **Intel® MPI Library** for optimized collective performance
- Intel® Omni-Path Host Fabric Interface (HFI) driver
  - Increase **"Receive Header Queue Count" - rcvhdrcnt** - from 2048 (default) to 4096 or 8192 (tuned)
- Usually the **default HFI and PSM** settings are best

For Storage

- Use connected mode for highest IPoFabric and IP router performance[1]
  - Watch for **IPoFabric enhancements** rolling into IFS 10.8 and 10.9 software releases
- Use **Lustre version 2.10** or later for OPA enhancements
- Consult tuning guide for **GPFS-specific tunings**

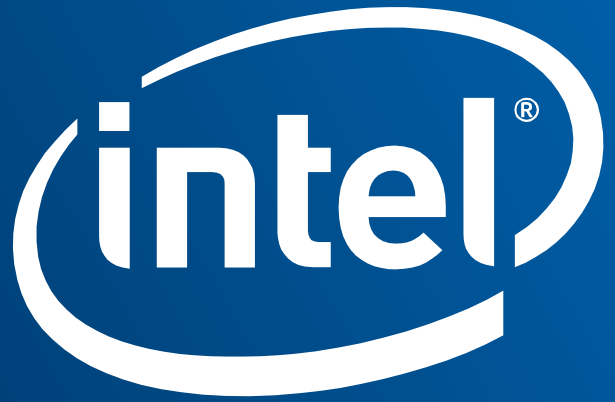
For Large Clusters

- Analyze impact of **Adaptive/dispersive routing** on cluster workloads
- Separating low priority storage from HPC applications with **Traffic Flow Optimization**
- Reduction of **system jitter** (tickless kernel, idle=halt), unwanted system processes

1. Consult Performance Tuning Guide for complete details, exceptions apply

Next Steps

- Consult Performance Tuning Guide for extensive tuning help
- Open a ticket with fabricsupport@intel.com for any performance concerns - we are here to help you!
- More helpful resources:
  - www.intel.com/omnipath - product information and customer sponsored articles
  - https://itpeernetwork.intel.com/ - performance focused articles



Thank you!

Configuration for Application Performance - Intel® Xeon® Platinum 8170 processors
Page 1 of 2

General/common configurations	EDR IB*	Intel® OPA
Processor and memory	Dual socket Intel® Xeon® Platinum 8170 processor nodes, 192 GB 2666 MHz DDR4 memory per node. Intel® Turbo Boost Technology and Hyperthreading Technology enabled. Unless otherwise noted, one MPI rank per physical CPU core is used.	
Kernel & CPU microcode	**3.10.0-693.21.1.el7.x86_64, 0x2000043. Variant 1, 2, and 3 mitigated.**	
Network software	MLNX_OFED_LINUX-4.3-1.0.1.0	Intel Fabric Suite 10.6.1
Network hardware	Intel® OPA: Intel Corporation Device 24f0 - Series 100 Host Fabric Interface (HFI). Series 100 Edge switch - 48 port. EDR IB: Mellanox Technologies MT27800 Family [ConnectX-5]. Mellanox MSB7800-36 port EDR InfiniBand switch.	
Operating System	Red Hat Enterprise Linux* Server Release 7.4	
Message Passing Interface (MPI) Library & Compilers	Intel® MPI 2018 Update 1. Intel parallel_studio_xe_2018.1.038	
Communication library	Intel® OPA: The better performing of either I_MPI_FABRICS=[shm:tmi or tmi] EDR IB*: The better performing of I_MPI_FABRICS=[shm:ofa, ofa, shm:dapl, or dapl]. I_MPI_DAPL_TRANSLATION_CACHE=1, I_MPI_DAPL_UD_TRANSLATION_CACHE=1, I_MPI_OFA_TRANSLATION_CACHE=1	
Storage	All input/output performed with NFSv3 with 1GbE to Intel SSDSC2BB48 drives. IO is not heavily used for any of the workloads.	

Application-specific configurations:

- BSMBench - An HPC Benchmark for BSM Lattice Physics Version 1.0. 32 ranks per node. Parameters: global size is 64x32x32x32, proc grid is 8x4x4x4. Machine config build file: cluster.cfg
- FDS (Fire Dynamics Simulator) version 6.5.3. strong_scaling_test, a General purpose input file to test FDS timings. 50 MPI ranks per node, 800 total MPI ranks.
- GROMACS version 2016.2. http://www.prace-ri.eu/UEABS/GROMACS/1.2/GROMACS_TestCaseB.tar.gz lignocellulose-rf benchmark. -g -static-intel. CC=mpicc CXX=mpicxx -DBUILD_SHARED_LIBS=OFF -DGMX_FFT_LIBRARY=mkl -DGMX_MPI=ON -DGMX_OPENMP=ON -DGMX_CYCLE_SUBCOUNTERS=ON -DGMX_GPU=OFF -DGMX_BUILD_HELP=OFF -DGMX_HWLOC=OFF -DGMX_SIMD=AVX512 GMX_OPENMP_MAX_THREADS=256. Run detail: gmx_mpi mdrun -s run.tpr -gcom 20 -resethway -noconfout
- LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) Feb 16, 2016 stable version release. Official Git Mirror for LAMMPS (http://lammps.sandia.gov/download.html) ls, rhodo, sw, and water benchmark. 52 ranks per node and 2 OMP threads per rank. Common parameters: I_MPI_PIN_DOMAIN=core Run detail: Number of time steps=100, warm up time steps=10 (not timed) Number of copies of the simulation box in each dimension: 8x8x4 and problem size: 8x8x4x32k = 8,192k atoms Build parameters: Modules: yes-asphere yes-class2 yes-kspace yes-manybody yes-misc yes-molecule yes-mpiio yes-opt yes-replica yes-rigid yes-user-omp yes-user-intel. Binary to be built: lmp_intel_cpu. . Runtime lammps parameters: -pk intel 0 -sf intel -v n 1

CONTINUE ON NEXT PAGE

Performance results are based on testing as of April 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. C

Configuration for Application Performance - Intel® Xeon® Platinum 8170 processors
Page 2 of 2

- LS-DYNA, A Program for Nonlinear Dynamic Analysis of Structures in Three Dimensions Example pfile: gen { nodump nobeamout dboutonly } dir { global one_global_dir local /tmp/3cars }. Higher performance shown with mpp s R8.1.0 Revision 105896 or mpp s R9.1.0 Revision: 113698
- NAMD version 2.10b2, stmv and apoa1 benchmark. Build detail: CHARM 6.6.1. FFTW 3.3.4. Relevant build flags: ./config Linux-x86_64-icc --charm-arch mpi-linux-x86_64-ifort-smp-mpicxx --cxx icpc --cc icc --with-fftw3.
- NWCHEM release 6.6. Binary: nwchem_armci-mpi_intel-mpi_mkl with MPI-PR run over MPI-1. Workload: siosi3 and siosi5. http://www.nwchem-sw.org/index.php/Main_Page. -genv CSP_VERBOSE 1 -genv CSP_NG 1 -genv LD_PRELOAD libcasper.so
- OpenFOAM is a free, open source CFD software package developed primarily by [OpenCFD](http://www.openfoam.com) . Version v1606+ . Gcc version 4.8.5 for Intel MPI. All default make options.
- Quantum ESPRESSO is an integrated suite of Open-Source computer codes for electronic-structure calculations and materials modeling at the nanoscale. It is based on density-functional theory, plane waves, and pseudopotentials. http://www.quantum-espresso.org/ ./configure --enable-openmp --enable-parallel. BLAS_LIBS= -lmkl_intel_lp64 -lmkl_intel_thread -lmkl_core ELPA_LIBS_SWITCH = enabled SCALAPACK_LIBS = $(TOPDIR)/ELPA/libelpa.a -lmkl_scalapack_lp64 -lmkl_blacs_openmpi_lp64 DFLAGS= -D__INTEL -D__FFTW -D__MPI -D__PARA -D__SCALAPACK -D__ELPA -D__OPENMP $(MANUAL_DFLAGS) AUSURF112 benchmark, all default options
- SPECFEM3D_GLOBE simulates the three-dimensional global and regional seismic wave propagation based upon the spectral-element method (SEM). It is a time-step algorithm which simulates the propagation of earth waves given the initial conditions, mesh coordinates/ details of the earth crust. small_benchmark_run_to_test_more_complex_Earth benchmark, default input settings. specfem3d_globe-7.0.0. FC=mpiifort CC=mpiicc MPIFC=mpiifort FCFLAGS=-g -xCORE_AVX2 CFLAGS=-g -O2 -xCORE_AVX2. run_this_example.sh and run_mesher_solver.sh, NCHUNKS=6, NEX_XI=NEX_ETA=80, NPROC_XI=NPROC_ETA=10. 600 cores used, 52 cores per node
- Spec MPI2007, https://www.spec.org/mpi/. *Intel Internal measurements marked estimates until published. Applications listed with “-Large” or “-Medium” in the name were part of the spec MPI suite. Compiler options: -O3 -xCORE-AVX2 -no-prec-div. Intel MPI: mpiicc, mpiifort, mpiicpc. Open MPI: mpicc, mpifort, mpicxx. Run detail: mref and lref suites, 3 iterations. 121.pop2: CPORTABILITY=-DSPEC_MPI_CASE_FLAG. 126.lammps: CXXPORTABILITY = -DMPICH_IGNORE_CXX_SEEK. 127.wrf2: CPORTABILITY = -DSPEC_MPI_CASE_FLAG -DSPEC_MPI_LINUX. 129.tera_tf=default=default=default: srcalt=add_rank_support 130.socorro=default=default=default: srcalt=nullify_ptrs FPORTABILITY = -assume nostd_intent_in CPORTABILITY = -DSPEC_EIGHT_BYTE_LONG CPORTABILITY = -DSPEC_SINGLE_UNDERSCORE.
- WRF - Weather Research & Forecasting Model (http://www.wrf-model.org/index.php) version 3.5.1. -xCORE_AVX2 -O3 . Net CDF 4.4.1.1 built with icc. Net CDF-fortran version 4.4.4 built with icc.

Performance results are based on testing as of April 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Configurations & Disclaimers

"Latency, Bandwidth, and Message Rate" slide: Intel® Xeon® Platinum 8170 processors, dual socket servers. Intel® Turbo Boost Technology enabled, Intel® Hyper-Threading Technology enabled. 2666 MHz DDR4, 192GB per node. One Intel® OPA 48-port Edge switch hop, using 2M copper cables. Intel MPI Benchmarks 2017 Update 1, PingPong, Uniband, and Biband. Intel® MPI Library 2018.1.163, icc 18.0.1. Red Hat Enterprise Linux Server release 7.4 (Maipo). Intel Fabric Suite (IFS) 10.6.1.0.2. 3.10.0-693.11.6.el7.x86_64, 0x2000043 microcode. Variants 1,2,3 mitigated. EDR IB*: MLNX_OFED_LINUX-4.3-1.0.1.0. Mellanox Technologies MT27800 Family [ConnectX-5]. Mellanox MSB7800-36 port EDR InfiniBand switch. One switch hop, 2M copper cables. Open MPI 3.1.0 and HPC-X 2.1.0. 1. SwitchIB-FW-11_1200_0102-release_nodes.pdf: "Removed out-of-the-box FEC, reaching 90ns latency, on Mellanox GA level copper cables equal to or shorter than 2m."

"Intel® OPA - Performance at a glance slide: TCO claim": Configurations: Intel® Omni-Path Architecture: Configuration assumes a 750-node cluster, and number of switch chips required is based on a full bisectional bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of director switches and edge switches. Includes hardware acquisition costs (server and fabric), 24x7 3-year support (Mellanox Gold support), and 3-year power and cooling costs. Mellanox and Intel® OPA component pricing from www.kernelsoftware.com, with prices as of March 20, 2018. Mellanox power data based on Mellanox CS7500 Director Switch, Mellanox SB7700/SB7790 Edge switch, and Mellanox ConnectX-5 VPI adapter card product briefs posted on www.mellanox.com as of August 15, 2017. Intel® OPA power data based on product briefs posted on www.intel.com as of April 4, 2017. Power and cooling costs based on $0.1071 per kWh, and assumes server power costs and server cooling cost are equal and additive.

Performance results are based on testing as of July 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

THIS SLIDE MUST BE USED WITH ANY SLIDES REMOVED FROM THIS PRESENTATION

NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

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 http://www.intel.com/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

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.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as property of others.