SOLUTION BRIEF

HPC Converged Platform

The Case for Running AI and Analytics on HPC Clusters

Creating a converged platform to run simulation and modeling, artificial intelligence (AI), and analytics workloads in a single cluster infrastructure supports breakthrough innovation while increasing the value and utilization of resources. Servers based on Intel® architecture are the ideal foundation for that convergence. This solution brief introduces the challenges and opportunities around running AI and analytics workloads on existing high-performance computing (HPC) clusters.

The exponential growth in size of enterprise, academic, and government data stores over the past decade has fueled the need for resources that can transform that data from nascent potential to actionable insight. Using analytics engines such as Apache Spark*, the basis of intelligent insights has become commonplace across industries. These solutions continue to become more sophisticated as time goes on, driving more value for a variety of organizations.

One application of AI (including machine and deep learning) is to make analytics more powerful yet, with neural networks trained to predict future events based on past inputs, for example. A study by Narrative Science reports that 61 percent of respondents are currently implementing AI, with predictive analytics as the most widely used type of AI-powered solution.[1] AI continues to become more important across a range of organizations; MarketWatch* estimates growth of the global AI market at a compound annual growth rate of 36 percent through 2024.[2]

There is a tendency in many organizations to adopt analytics platforms and AI technologies as distinct entities, rather than to focus on integrating with existing systems architectures to make business processes more effective and efficient. In many cases, dedicated new clusters may be created for deploying these capabilities, as shown in Figure 1. This approach creates data silos and the need for expensive operations related to moving and staging data. The existence of multiple clusters also sets the stage for under-utilized infrastructure, particularly for AI clusters used primarily for training deep learning networks, which tends to be sporadic.

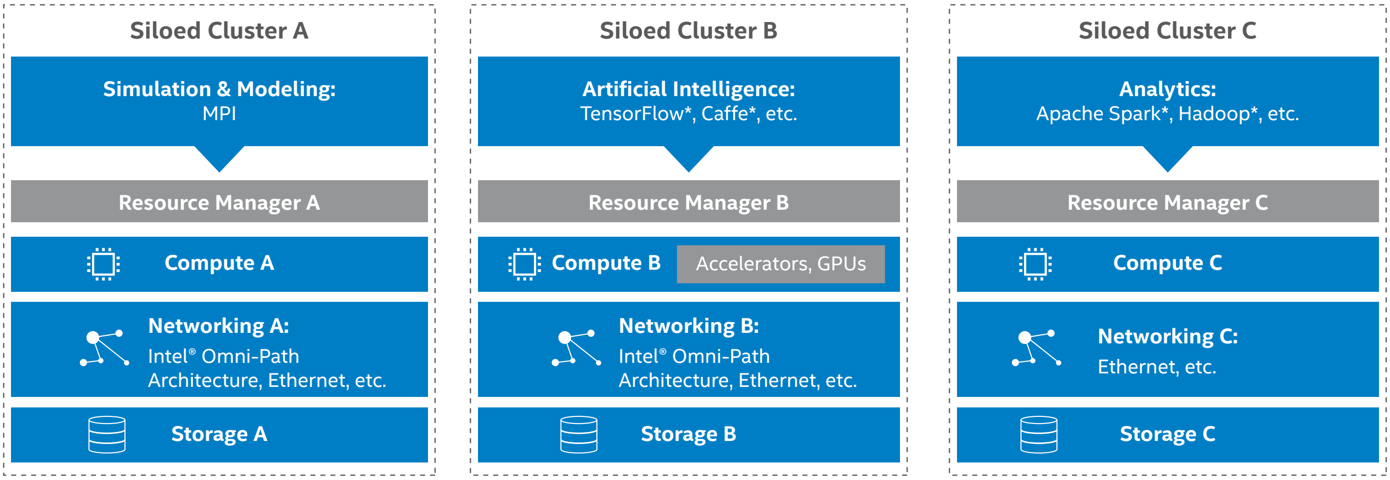


Figure 1. Workloads spread across separate, siloed clusters.

Overcoming Challenges to Realize the Value of HPC

As organizations with existing HPC infrastructures adopt AI and analytics, many cite one or more of the following challenges preventing them from running all three workloads on one system:

- **Coexistent frameworks and software stacks** for simulation and modeling, analytics, and AI must be made to work together harmoniously. Unfortunately, the software stacks and frameworks for HPC, analytics, and AI are vastly different, and each workload must have its own software stack loaded on a cluster. In particular, the resource managers responsible for parceling out resources for these workloads are typically designed very differently. Intel provides a growing set of solution architectures for this purpose, as discussed in the companion brief to this one, *Supporting Simulation and Modeling, Analytics, and AI on a Common Platform*.
- **Non-HPC-oriented hardware** such as servers equipped with accelerators and GPUs draws the interest of many organizations as a potential means to drive greater performance for AI workloads in particular. In fact, the characteristics of the systems they may already own in typical HPC infrastructures (for example, robust compute cores connected with high-performance network fabrics and to high-performance shared storage) are well-suited to the needs of analytics and AI workloads. Intel has invested heavily to increase AI performance on Intel® Xeon® Scalable Processors as well as introducing new AI instructions specific to Intel® Xeon® CPUs that make them an excellent choice for AI workloads.
- **Cultural and operational separation** between HPC teams' focus on bare metal, on-premises clusters, as opposed to the cloud orientations of many AI and analytics teams, which may extend to functional approaches such as DevOps. In addition, AI and analytics teams typically use higher-level languages such as Python*, Scala*, and Java*, whereas HPC teams are more likely to be using lower-level languages such as C/C++ and Fortran.

Overcoming these challenges and the assumptions that accompany them can be an important factor in guiding an organization on the path to efficient infrastructure for all three types of workloads. The cost benefits of bringing simulation and modeling, AI, and analytics workloads together onto a single cluster infrastructure are twofold. From a capital expenditure (CAPEX) point of view, it reduces the need for new expenditures while deriving maximum benefit from existing and future cluster investments. In terms of operating expenditure (OPEX), it reduces the cost of running and maintaining the environment by simplifying the infrastructure to run on one cluster instead of multiple clusters.

These benefits are enabled through solution stacks offered by the same server original equipment manufacturers (OEMs) that IT organizations already work with as suppliers for existing HPC infrastructure. As shown in Figure 2, converged HPC stacks are poised to drive innovation and deliver value across workflows, including conventional HPC, HPC-based AI, and HPC-based analytics. Projected forward to 2021, implementations of conventional HPC are expected to continue dominating the install base, but analytics and especially AI are forecast to exhibit far higher rates of growth.[3]

Emerging Needs for Simulation and Modeling, AI, and Analytics

As IT organizations set their architectural strategies for the next several years, many increasingly see points where AI and analytics intersect with other workloads and business processes. At the same time, the data sets that simulation and modeling jobs must handle are growing at massive rates, spurred on by current data-intensive science as well as emerging fields in cognitive computing. These trends emphasize the necessity of rich interoperability among the underlying systems, which is hampered by running simulation and modeling, AI, and analytics workloads on separate clusters. For example, consider the inefficiency of a process where simulation and modeling, data cleaning, and AI-based inference each occur on separate clusters.

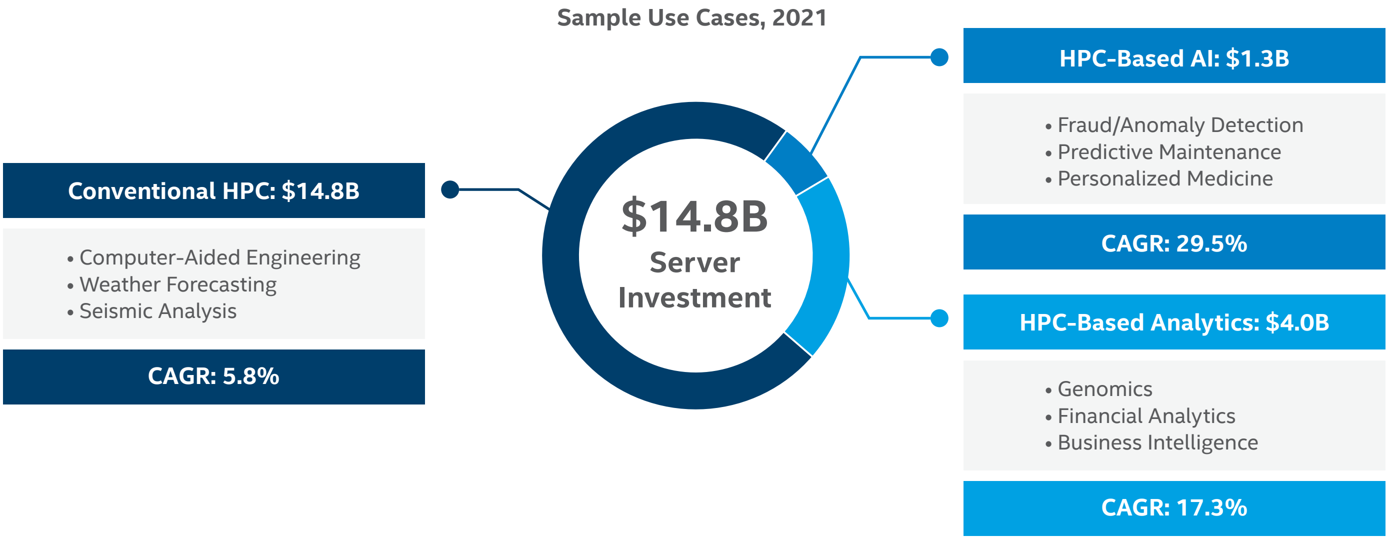


Figure 2. Investment levels and growth across sample HPC workloads through 2021.[3]

By running all three together on the same cluster, a number of advantages can be achieved. First, as mentioned above, the need to copy and stage data is dramatically reduced or eliminated. In addition, delay associated with transferring large amounts of data between clusters can be avoided, which is particularly important to flexibly support real-time workflows. This convergence will only become more important as emerging needs demand new capabilities within organizations.

For example, the buildout of IoT endpoints across industries will generate data flows that are orders of magnitude larger than what came before, in many cases. Sorting through those flows to identify data points of interest, patterns, and other insights is a potentially enormous task that is well-suited to AI-based analytics. Based on those outputs, opportunities may be identified for optimization of a business process or agent-based system, which could be investigated in all its permutations by simulation and modeling software. AI could be used to improve design and calibration of the actual simulation and modeling jobs.

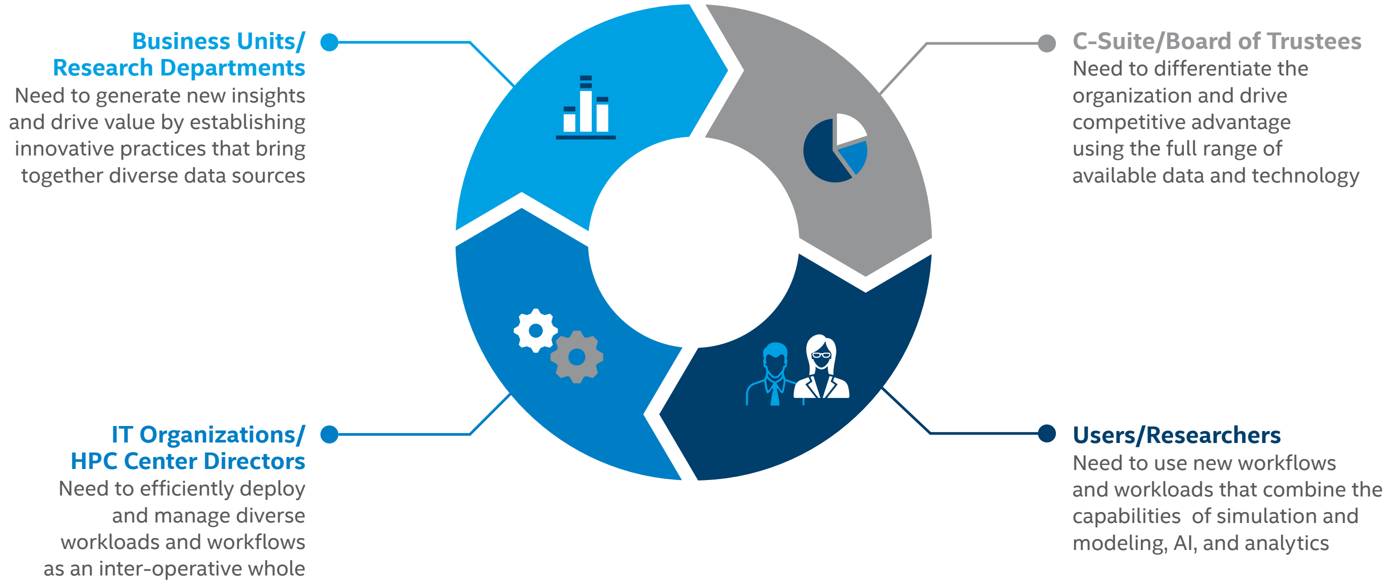
The imperative for interoperation among the systems that cooperate across this type of overall workflow is clear. In addition, the more seamlessly these systems are integrated together, the more efficient and effective the processes can be. Successfully meeting this set of interconnected challenges paves the way for an array of use cases that span across industries and research fields, ranging from data insight, fraud detection, and robotics to individualized medicine, meteorology, and climatology.

Accordingly, many companies and institutions find themselves needing to modernize their HPC infrastructures for greater flexibility. This modernization specifically must strive to meet and bring together the needs of diverse parts of the organization, as illustrated in Figure 3. Multipurpose clusters are ideal for meeting this range of interests with maximum flexibility, efficiency, and performance. Servers based on Intel architecture are the ideal foundation for this infrastructure.

An Open Systems Approach to Large-Scale Converged Infrastructure

Existing HPC clusters built using large-capacity, general-purpose servers based on Intel architecture can deliver flexible, cost-effective performance across workloads. In addition to eliminating the need to purchase, configure, manage, and support separate resources, this approach reduces complexity. As opposed to using systems that are designed for a specific function, jobs can be scheduled anywhere across the environment as needed, which allows for maximum utilization of resources and avoids bottlenecks associated with specific resources.

Intel works across the ecosystem—including contributing to open source projects as well as co-engineering with hardware and software providers—to make converged HPC clusters function optimally, as illustrated in Figure 4. This enablement streamlines adoption for end customers, helping them mitigate complexity and risk. It includes ensuring that resource managers and other software for simulation and modeling, AI, and analytics work together seamlessly and that frameworks are optimized for performance, scalability, stability, and security on Intel architecture. Intel also works with server providers to engineer and validate systems that deliver the best results possible.



Business Units/ Research Departments
Need to generate new insights and drive value by establishing innovative practices that bring together diverse data sources

C-Suite/Board of Trustees
Need to differentiate the organization and drive competitive advantage using the full range of available data and technology

IT Organizations/ HPC Center Directors
Need to efficiently deploy and manage diverse workloads and workflows as an inter-operative whole

Users/Researchers
Need to use new workflows and workloads that combine the capabilities of simulation and modeling, AI, and analytics

Figure 3. Motivation for HPC convergence across the organization.

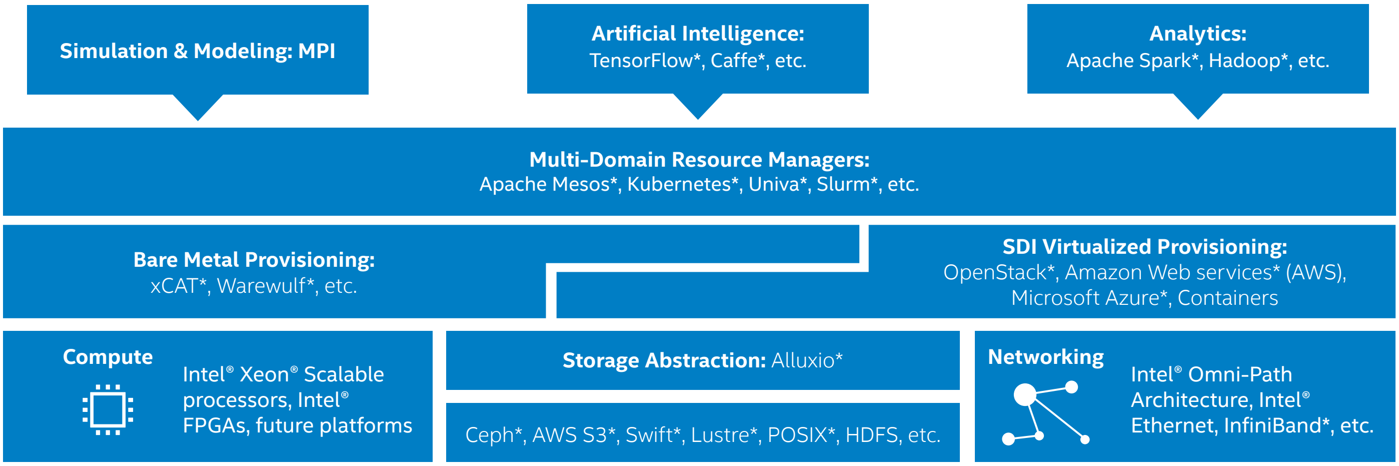

Simulation & Modeling: MPI
Artificial Intelligence: TensorFlow*, Caffe*, etc.
Analytics: Apache Spark*, Hadoop*, etc.
Multi-Domain Resource Managers: Apache Mesos*, Kubernetes*, Univa*, Slurm*, etc.
Bare Metal Provisioning: xCAT*, Warewulf*, etc.
SDI Virtualized Provisioning: OpenStack*, Amazon Web services* (AWS), Microsoft Azure*, Containers
Compute
Intel® Xeon® Scalable processors, Intel® FPGAs, future platforms
Storage Abstraction: Alluxio*
Ceph*, AWS S3*, Swift*, Lustre*, POSIX*, HDFS, etc.
Networking
Intel® Omni-Path Architecture, Intel® Ethernet, InfiniBand*, etc.

Figure 4. Unified architecture across workloads.

The architecture's foundation is a common hardware and software infrastructure based on Intel® building blocks and other industry-standard components. These include a comprehensive range of popular interconnects and object stores, as well as various approaches to storage abstraction, which pull data together from all three types of workloads, creating a single data pool that is available wherever it is needed. Therefore, both the isolation of data in silos and the need to constantly move and stage large data sets among separate clusters are eliminated.

The model's unified storage architecture is based on a distributed object store that offers a number of storage abstractions for interoperability among data-access approaches. The unified fabric architecture harmonizes the behavior of various interconnect fabrics.

The architecture described here embraces both bare-metal provisioning and the virtualized provisioning approaches of software-defined infrastructure (SDI). Bare-metal approaches are the standard method used by most HPC organizations, while SDI focuses on allowing virtual resources to be spun up on demand. That dynamic approach can be instrumental in increasing agility, spawning jobs based on the constantly changing circumstances and business or research needs associated with analytics workloads.

The resource-management layer is in many ways the crux of the converged platform, providing the efficient and robust abstraction of the resources used to execute simulation and modeling, AI, and analytics workloads on the same cluster. This capability allows for the integration of functions such as resource allocation, job scheduling, and resolution of contention issues among otherwise-incompatible workloads. Domain-specific plugins at this layer allow cluster operators to tailor the operation of the environment to their needs.

This generalized stack makes it much faster and easier for organizations to explore new implementations of simulation and modeling, AI, and analytics, especially in terms of how the capabilities of all three intersect to drive new business and research value.

Building on a Broad Intel® Technology Foundation

Rather than prescribing a rigid solution, Intel's architectural approach for simulation and modeling, AI, and analytics includes openness and flexibility as a primary design requirement. In addition to supporting hardware building blocks from the entire range of popular manufacturers, the stack draws on a vast ecosystem of software. Along with tools provided directly by Intel, many open-source and commercial third-party software packages are optimized for performance, scalability, stability, and security on Intel architecture.

Pre-validated combinations of hardware and software building blocks, tailored to specific organizational needs, are available as Intel® Select Solutions, as illustrated in Figure 5. This infrastructure helps accelerate performance while simplifying implementation and reducing risk for end customers associated with data center modernization, using systems that are available from a wide choice of popular server manufacturers.

Intel® Architecture Building Blocks

Cluster architectures draw on a variety of Intel architecture building blocks, across the hardware stack, including the following:

- **2nd Gen Intel® Xeon® Scalable processors** are the heart of robust computing clusters that power excellent results across workloads. For example, in recent testing, it achieved an average performance improvement of up to 3.7x in HPC CPU benchmarks compared to a three-year-old system.[4] It also delivered a world-class 5.8x performance improvement on LINPACK* CPU benchmarks[5] and 1.7x better floating-point performance per core compared to competing processors.[6] These processors are also built to meet requirements of the most demanding AI inference workloads. 2nd Gen Intel Xeon Scalable processors running Intel® Deep Learning Boost have been shown to improve inference throughput by up to 25x versus the Intel® Xeon® Platinum 8180 processor.[7]

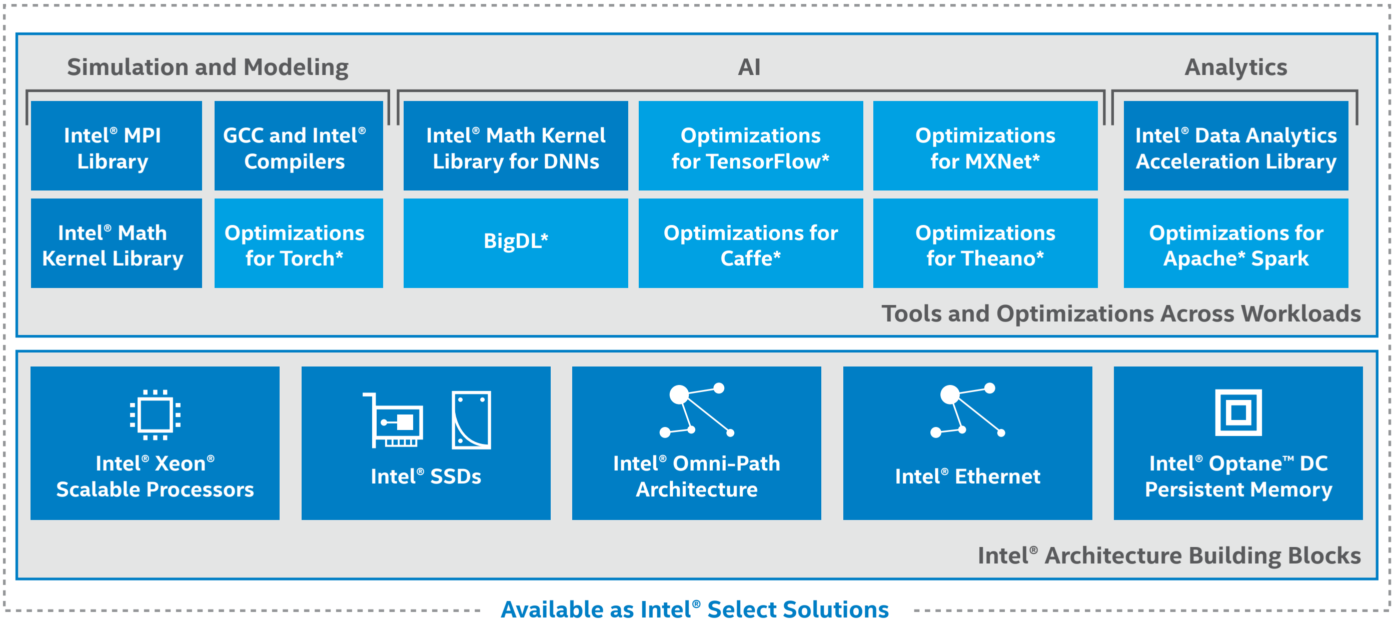

Simulation and Modeling
AI
Analytics
Intel® MPI Library
GCC and Intel® Compilers
Intel® Math Kernel Library for DNNs
Optimizations for TensorFlow*
Optimizations for MXNet*
Intel® Data Analytics Acceleration Library
Intel® Math Kernel Library
Optimizations for Torch*
BigDL*
Optimizations for Caffe*
Optimizations for Theano*
Optimizations for Apache* Spark
Tools and Optimizations Across Workloads
Intel® Xeon® Scalable Processors
Intel® SSDs
Intel® Omni-Path Architecture
Intel® Ethernet
Intel® Optane™ DC Persistent Memory
Intel® Architecture Building Blocks
Available as Intel® Select Solutions

Figure 5. Intel® architecture hardware and software stack.

- **Intel® Optane™ DC persistent memory** offers the unprecedented combination of high capacity, affordability, and persistence. By moving and maintaining larger amounts of data closer to the processor, data-intensive HPC and AI workloads can be processed quickly and on a large scale.
- **Intel® SSDs** provide a range of storage options that enable customers to create their own balance between cost and performance. Intel® Optane™ DC SSDs deliver dramatically lower and more consistent latency, high endurance and high/balanced performance breaking through NAND SSD bottlenecks to unleash system performance.
- **Intel® Fabric products** include Intel® Omni-Path Architecture (Intel® OPA), a high-bandwidth and low-latency fabric that optimizes performance and eases deployment of HPC clusters, as well as Intel® Ethernet, an established industry leader with a broad array of options for speed, cable medium, and port count. Both are built to work smoothly in clusters with other fabrics such as InfiniBand*.

Tools and Optimizations for Intel Architecture Across Workloads

Simulation and Modeling-Focused Tools and Optimizations

A few key tools and optimizations that target simulation and modeling applications and workloads include the following:

- **Intel® MPI Library** is an implementation of the open-source MPICH specification designed to create, run, test, and maintain applications optimized for Intel architecture-based clusters, supporting any of multiple fabrics chosen at runtime. It provides both a runtime environment and a software development kit.
- **GCC and Intel® Compilers** enable flexible optimization capabilities for Intel architecture; the similar behaviors of both compilers allow developers to easily switch back and forth between the two. The compilers also integrate with popular toolchains in common, further aiding interoperability.
- **Intel® Math Kernel Library (Intel® MKL)** is a collection of pre-optimized math routines that are broadly applicable for technical and scientific computing. Intel maintains the functions in Intel MKL for each platform generation, enabling end customers to take advantage of hardware advances simply by relinking and recompiling their code.
- **Torch*** is an open-source scientific computing framework with integrated support for many machine-learning algorithms. It includes a simple and robust scripting language (LuaJIT) as well as an underlying C/CUDA implementation. Intel maintains a fork of the project that is optimized for Intel Xeon processors.

Analytics-Focused Tools and Optimizations

Following is a sample of solutions (among many others) that Intel provides or enables for improved results from analytics workloads:

- **Intel® Data Analytics Acceleration Library (Intel® DAAL)** is a collection of optimized routines to accelerate big data analytics problems. It is designed to augment the performance of applications built on popular data platforms such as Hadoop*, Apache Spark, R*, and Matlab*.
- **Apache Spark** is a particular optimization focus for Intel among analytics frameworks. Originally developed at UC Berkeley, Apache Spark is a big-data processing engine with built-in modules for streaming, SQL, machine learning, and graph processing.

AI-Focused Tools and Optimizations

The ecosystem for AI frameworks and other components is fast-growing, and Intel participates by helping ensure that AI implementations run best on Intel architecture, including the following:

- **Intel® MKL for Deep Neural Networks (Intel® MKL-DNN)** is a performance library for deep learning applications that provides highly vectorized and threaded building blocks for implementing deep neural networks on Intel architecture-based systems. Learn more: https://01.org/mkl-dnn.
- **Deep learning frameworks** such as BigDL*, Caffe*, MXNet*, TensorFlow*, and Theano* are optimized to accelerate AI workflows, including simplified development of applications that benefit from fast training of deep neural networks on servers and clusters based on Intel architecture. Learn more: http://intel.ai/framework-optimizations.

Conclusion

Converging simulation and modeling, AI, and analytics platforms is a necessary evolution for enterprises and institutions that anticipate running all three types of workloads in the coming years. By supporting all three types of jobs in a single environment, many organizations will be able to reduce both CAPEX and OPEX as they embrace the full spectrum of compute capabilities that will be needed to support emerging business and research needs.

Accelerate HPC innovation:
www.intel.com/hpc

[1] *Outlook on Artificial Intelligence* in the Enterprise 2018. Narrative Science https://medium.com/@narrativesci/2018-outlook-on-artificial-intelligence-b1b63a7386f4.

[2] Press Release, *Artificial Intelligence Market to Rise at Spectacular CAGR of 36.10% During 2016-2024, Players in End-use Industries Leverage its Potential to Automate Processes* – TMR. MarketWatch, August 16, 2018, https://www.marketwatch.com/press-release/artificial-intelligence-market-to-rise-at-spectacular-cagr-of-3610-during-2016-2024-players-in-end-use-industries-leverage-its-potential-to-automate-processes---tmr-2018-08-16.

[3] Analyst Report, *The Business Value of Leading-Edge High Performance Computing*. Hyperion Research, 2017, https://hpe.lookbookhq.com/c/the-business-value-p?x=1aqP6M.

[4] Average geomean of STREAM, HPCG, HPL, WRF, OpenFOAM, LS-Dyna, VASP, NAMD, LAMMPS, Black Scholes, and Monte Carlo., Individual workload may vary. Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure. 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. OpenFOAM Disclaimer: This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark. Configurations:

3.7x avg gain w/Intel® Xeon® Platinum 9242 processor vs 3-year old server: Average geomean of STREAM, HPCG, HPL, WRF, OpenFOAM, LS-Dyna, VASP, NAMD, LAMMPS, Black Scholes, and Monte Carlo. Individual workload may vary. Intel Xeon E5-2697 v4 Processor: Intel Reference Platform with 2S Intel Xeon E5-2697 v4 processors (2.3GHz, 18C), 8x16GB DDR4-2400, 1 SSD, Cluster File System: Panasas (124 TB storage) Firmware v6.3.3.a & OPA based IEEL Lustre, BIOS: SE5C610.86B.01.01.0027.071020182329, Microcode: 0xb00002e, Oracle Linux Server release 7.6 (compatible with RHEL 7.6) on a 7.5 kernel using ksplice for security fixes, Kernel: 3.10.0-862.14.4.el7.crt1.x86_64, OFED stack: OFED OPA 10.8 on RH7.5 with Lustre v2.10.4, HBA: 100Gbps Intel OPA 1 port PCIe x16, Switch: Intel OPA Edge Switch 100 Series 48 Port. STREAM OMP 5.10, Triad, HT=ON, Turbo=OFF, 1 thread per corescore: 128.36. HPCG, Binary included MKL 2019u1, HT=ON, Turbo=OFF, 1 thread per corescore: 23.78. HPL 2.1, HT=ON, Turbo=OFF, 2 threads per corescore: 1204.64. WRF 3.9.1.1, conus-2.5km, HT=ON, SMT=ON, 1 thread per corescore: 4.54. OpenFOAM 6.0, 42M_cell_motorbike, HT=ON, Turbo=OFF, 1 thread per corescore: 3500. LS-Dyna 9.3-Explicit AVX2 binary, 3car, HT=ON, SMT=ON, 1 thread per corescore: 2814. VASP 5.4.4, CuC, HT=ON, Turbo=OFF, 1 thread per corescore: 384.99. NAMD 2.13, apoa1, HT=ON, Turbo=OFF, 2 threads per corescore: 4.4. LAMMPS version 12 Dec 2018, Water, HT=ON, Turbo=ON, 2 threads per corescore: 54.72. Black Scholes, HT=ON, Turbo=ON, 2 threads per corescore: 2573.77. Monte Carlo, HT=ON, Turbo=ON, 2 threads per corescore: 43.2. Intel Xeon 9242 Processor: Intel Reference Platform with 2S Intel Xeon 9242 processors (2.2GHz, 48C), 16x16GB DDR4-2933, 1 SSD, Cluster File System: 2.12.0-1 (server) 2.11.0-14.1 (client), BIOS: PLYXCRB1.86B.0572.D02.1901180818, Microcode: 0x4000017, CentOS 7.6, Kernel: 3.10.0-957.5.1.el7.x86_64, OFED stack: OFED OPA 10.8 on RH7.5 with Lustre v2.10.4, HBA: 100Gbps Intel OPA 1 port PCIe x16, Switch: Intel OPA Edge Switch 100 Series 48 Port. STREAM OMP 5.10, Triad, HT=ON, Turbo=OFF, 1 thread per corescore: 407. HPCG, Binary included MKL 2019u1, HT=ON, Turbo=OFF, 1 thread per corescore: 81.91. HPL 2.1, HT=ON, Turbo=OFF, 2 threads per corescore: 5314. WRF 3.9.1.1, conus-2.5km, HT=ON, SMT=ON, 1 thread per corescore: 1.44. OpenFOAM 6.0, 42M_cell_motorbike, HT=ON, Turbo=OFF, 1 thread per corescore: 1106. LS-Dyna 9.3-Explicit AVX2 binary, 3car, HT=ON, SMT=ON, 1 thread per corescore: 768. VASP 5.4.4, CuC, HT=ON, Turbo=OFF, 1 thread per corescore: 133.96. NAMD 2.13, apoa1, HT=ON, Turbo=OFF, 2 threads per corescore: 19.9. LAMMPS version 12 Dec 2018, Water, HT=ON, Turbo=ON, 2 threads per corescore: 276.1. Black Scholes, HT=ON, Turbo=ON, 2 threads per corescore: 9044.32. Monte Carlo, HT=ON, Turbo=ON, 2 threads per corescore: 227.62. OpenFOAM Disclaimer: This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark. Date of testing March 15th, 2019.

[5] OpenFOAM Disclaimer: This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark. Configurations: LINPACK: AMD EPYC 7601: Supermicro AS-2023US-TR4 with 2 AMD EPYC 7601 (2.2GHz, 32 core) processors, SMT OFF, Turbo ON, BIOS ver 1.1a, 4/26/2018, microcode: 0x8001227, 16x32GB DDR4-2666, 1 SSD, Ubuntu 18.04.1 LTS (4.17.0-041700-generic Retpoline), High Performance Linpack v2.2, compiled with Intel(R) Parallel Studio XE 2018 for Linux, Intel MPI version 18.0.0.128, AMD BLIS ver 0.4.0, Benchmark Config: Nb=232, N=168960, P=4, Q=4, Score =1095GFs, tested by Intel as of July 31, 2018. vs. 1-node, 2x Intel® Xeon® Platinum 9282 cpu on Walker Pass with 768 GB (24x 32GB 2933) total memory, ucode 0x400000A on RHEL7.6, 3.10.0-957.el7.x86_65, IC19u1, AVX512, HT off, Turbo on, score=6411, test by Intel on 2/16/2019.

[6] 1-copy SPECrate2017_fp_base* 2 socket Intel 8280 vs 2 socket AMD EPYC 7601. Intel® Xeon®-SP 8280, Intel Xeon-based Reference Platform with 2 Intel® Xeon® 8280 processors (2.7GHz, 28 core), BIOS ver SE5C620.86B.0D.01.0348.011820191451, 01/18/2019, microcode: 0x5000017, HT OFF, Turbo ON, 12x32GB DDR4-2933, 1 SSD, Red Hat EL 7.6 (3.10.0-957.1.3.el7.x86_64), 1-copy SPECrate2017_fp_rate base benchmark compiled with Intel Compiler 19.0.1.144, -xCORE-AVX512 -ipo -O, executed on 1 core using taskset and numactl on core 0. Estimated score = 9.6, as of 2/6/2019 tested by Intel with security mitigations for variants 1,2,3,3a, and L1TF.

AMD EPYC 7601, Supermicro AS-2023US-TR4 with 2S AMD EPYC 7601 with 2 AMD EPYC 7601 (2.2GHz, 32 core) processors, BIOS ver 1.1c, 10/4/2018, SMT OFF, Turbo ON, 16x32GB DDR4-2666, 1 SSD, Red Hat EL 7.6 (3.10.0-957.5.1.el7.x86_64), 1-copy SPECrate2017_fp_rate base benchmark compiled with AOCC ver 1.0 -Ofast, -march=znver1, executed on 1 core using taskset and numactl on core 0. Estimated score = 5.56, as of 2/8/2019 tested by Intel.

[7] 5x inference throughput improvement on Intel® Xeon® Platinum 9242 processor with Intel® DL Boost : Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9242(48 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS:SE5C620.86B.0D.01.0241.112020180249, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer syntheticData:3x224x224, 48 instance/2 socket, Datatype: INT8 vs. Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

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.

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

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 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. Check with your system manufacturer or retailer or learn more at intel.com.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

0419/SM/MESH 338330-001US