Theme Article

Cascade Lake: Next Generation Intel Xeon Scalable Processor

Mohamed Arafa, Bahaa Fahim, Sailesh Kottapalli, Akhilesh Kumar, Lily P. Looi, Sreenivas Mandava, Andy Rudoff, Ian M. Steiner, Bob Valentine, Geetha Vedaraman, and Sujal Vora
Intel Corporation

***Abstract*—This paper introduces advances in the performance of AI and deep learning inference application on the next generation Intel Xeon Scalable processor, code-named Cascade Lake, which also includes support for Intel Optane DC persistent memory, a breakthrough nonvolatile memory technology that bridges the gap between DRAM and storage.**

The next-generation Intel Xeon Scalable processor extends processor and system level innovations introduced with first generation Intel Xeon Scalable processor code-named Skylake-SP.[1] This paper provides an overview of the Skylake-SP that forms the foundation for Cascade Lake and then focuses on two significant improvements in the areas of AI and deep learning inference and support for persistent memory.

INTEL XEON SCALABLE (SKYLAKE-SP) CPU OVERVIEW

Skylake-SP included Intel's most advanced core microarchitecture at the time of its introduction providing a boost in instructions per cycle. Support for Intel AVX-512 instructions with two 512-bit wide floating-point multiply accumulate (FMA) units with each core could sustain 32 double precision floating point or 64 single precision floating point operation every clock. Increase in compute throughput was supported by doubling of L1-D cache bandwidth and quadrupling of unified L2 cache capacity to 1 MiB per core. The shared and distributed last-level on-chip cache was changed to a noninclusive cache to make more effective use of on-chip cache capacity. A new Intel Mesh interconnect was designed to move data between core, last-level cache, memory controller, IO controller, and inter-processor interconnect, which reduced number of clock cycles required to service L2 cache misses in most cases and supported higher sustained core, memory, and IO bandwidth.

Digital Object Identifier 10.1109/MM.2019.2899330
Date of publication 13 February 2019; date of current version 15 March 2019.

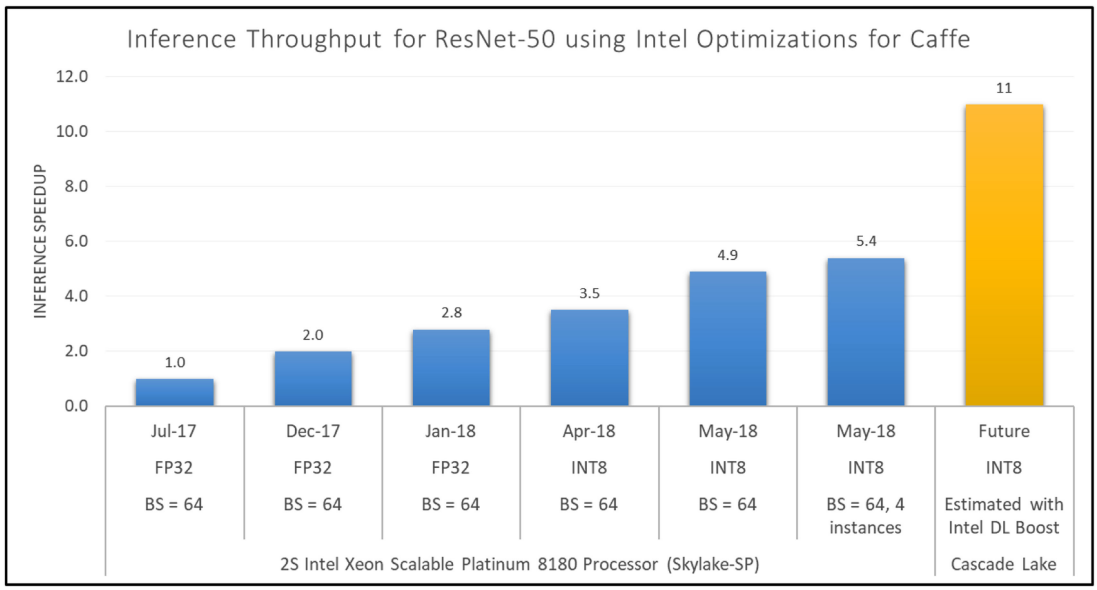

Figure 1. Performance of ResNet-50 image recognition workload on 2-socket systems with different software optimizations and data types. (Performance measured on 2 socket Intel Xeon Platinum 8180 processor with 376 GB of DDR4-2666 memory running CentOS Linux 7.3.1611 and using Intel MKL-DNN. Performance results are based on testing as of July 2017 to May 2018 and may not reflect all publicly available security updates. Results for Cascade Lake have been estimated using Intel DL Boost instructions and provided for informational purposes. Any differences in system hardware, software, or configuration may affect actual performance.

The memory and IO subsystems of the processor were significantly revamped as well. The number of memory channels was increased to 6 channels of DDR4 supporting data rates of 2666 MT/s. The IO subsystem was redesigned to support 48 lanes of PCI Express Rev 3.0 at data rates of 8 GT/s along with several integrated devices (DMA engine, volume management device, and nontransparent bridge) making Skylake-SP a very capable and efficient processor for data processing in a networked environment.

Cascade Lake Enhancements

Cascade Lake processor is compatible with first generation Intel Xeon Scalable platform. It maintains the same core count, cache size, and IO interfaces as Skylake-SP and is built on Intel's continually-refined 14-nm process technology. Design and process technology improvements result in higher core frequency within the same thermal design power. Changes in Cascade Lake processor are targeted towards improved AI and deep-learning inference performance and support for Intel Optane DC persistent memory. Cascade Lake processor also includes changes to address some variants of side-channel exploits.

IMPROVING DEEP-LEARNING ON GENERAL-PURPOSE PROCESSORS

Deep-learning is used across many domains from image recognition, natural language processing, predictive analytics, threat detection, and others. Support for Intel AVX-512 on Skylake-SP provided a great foundation for deep-learning applications on a general-purpose processor. Intel is putting significant focus on optimizing deep-learning software to take full advantage of capabilities available in Intel Xeon Scalable processors. Since the introduction of Skylake-SP, changes in Intel optimizations for Caffe have yielded 2.8 times improvement on ResNet-50 for FP32 data type as shown in Figure 1.

Compared to FP32, using a more compact data format with INT8 reduces memory bandwidth to 1/4th, improves compute throughput by 1.3 times, resulting in overall performance improvement close to 1.7 times. Further optimization with multiple instances that partition the system into multiple computing units to work on a separate batch of input data reduces the number of concurrent threads within an instance, thus reducing the synchronization and communication across cooperating threads within an instance.

The data in Figure 1 was gathered in 10 months following the launch of Skylake-SP in July 2017, and these software optimizations have resulted in 5.4 times improved performance for ResNet-50 image recognition inference workload on the same platform.

Neural machine translation is another example of a different deep learning inference workload. For this workload, an optimized software stack and refactored workload can achieve up to 14× performance improvement over off-the-shelf non-AVX-512 software stack.[2]

Intel is enabling the open source software community to take full advantage of the capabilities of Intel Xeon Scalable processors. Intel MKL-DNN[3] project was created to deliver an optimized math library for deep learning applications and frameworks, and Intel is actively involved in optimizations of a number of the open-source AI frameworks.[4]

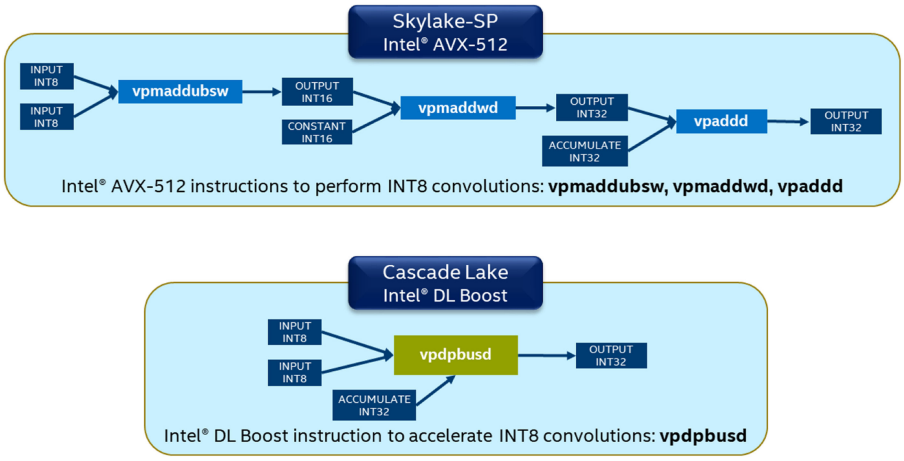

Figure 2. Instructions in the inner loop of a convolution operation for INT8 data type using Intel AVX-512 and Intel DL Boost instructions.

Intel DL Boost Instructions in Cascade Lake

As mentioned earlier, using smaller data types such as INT8 and INT16 can have significant performance benefits without compromising the quality of result for some deep learning applications. Smaller data types reduce memory bandwidth and cache footprint and allow higher compute throughput per cycle. One of the bottlenecks in processing smaller data types on the Skylake-SP generation is the number of instructions required in the inner loop of the convolution operation. In Cascade Lake, new Intel DL Boost instructions are introduced that reduce the number of instructions required for INT8 and INT16 data types on such operations.

As shown in Figure 2 for INT8 data type, the Intel AVX-512 instructions used in the convolution inner loop on Skylake-SP uses three instructions to take two 8-bit inputs to produce one 32-bit output. With the addition of new 8-bit Intel DL Boost instruction in Cascade Lake, those three instructions are replaced by a single instruction.

Note that fusing multiple instructions in Intel DL Boost also results in improved overflow behavior since the FMA provides wider internal data paths to handle saturation. Due to better overflow behavior of these fused instructions, better accuracy is expected when Intel DL Boost instructions are used with the smaller data types than Intel AVX-512 instructions.

The last bar in the chart in Figure 1 shows the estimated throughput for the same image recognition workload that was mentioned earlier—ResNet-50 using Intel optimization for Caffe. Additional two times improvement is expected in overall throughput to ResNet-50 using Intel DL Boost instructions bringing the total speedup to about 11 times from introduction of Skylake-SP to Cascade Lake through a combination of software optimization and hardware improvements.

REIMAGINING DATACENTER MEMORY HIERARCHY

A typical memory and storage hierarchy on a datacenter system consists of processor on-chip cache, DRAM for main memory, solid-state storage device for warm storage, and local or networked disk for archival storage. The performance, capacity, and cost of different layers require this four-level hierarchy to get the best utilization of the resources, meet response time expectations, and improve total cost of ownership or return on investment. For data-intensive applications, a large memory tier provides the best performance but the cost of a large memory pool gets in the way.

It is well recognized that the gap between different layers of memory and storage hierarchy has been growing in terms of latency, bandwidth, and cost per bit. A hard boundary between memory and storage with a dramatic change in performance and cost is becoming even more prominent. Introduction of an additional layer in between main memory and disk storage using NAND flash memory and NVMe interface helps bridge the latency and bandwidth gap, but the use of traditional storage hardware and software interfaces is still a limiter.

It is well recognized that the gap between different layers of memory and storage hierarchy has been growing in terms of latency, bandwidth, and cost per bit. A hard boundary between memory and storage with a dramatic change in performance and cost is becoming even more prominent.

The latency, bandwidth, and capacity/cost metrics between memory and SSD show a large gap of about 1000 times on latency, about ten times on bandwidth, and about ten times on cost/bit which makes applications with large data sets with random access patterns suffer from either prohibitive cost or suboptimal performance.

Intel is addressing this problem in multiple ways using Intel 3D XPoint memory media jointly developed with Micron. The performance characteristics of Intel 3D XPoint memory media and its persistence allows usage both as high-performance storage and higher capacity memory. Intel Optane SSDs using Intel 3D XPoint media with a standard NVMe interface have been available for some time and can be used on existing platforms without any additional support from the processor. Intel recently announced shipments of Intel Optane DC persistent memory modules that brings this new storage-class memory to the processor's memory interface via conventional DIMM slots.

Intel Optane DC persistent memory has attributes that address the growing memory and storage gap. It supports much higher capacities with memory module sizes of 128 GB, 256 GB, and 512 GB at affordable per-gigabyte cost compared to large-capacity DRAM, is DDR4 pin compatible, supports load/store accesses at 64B granularity, has latency and bandwidth much closer to memory, is persistent if used in persistent mode, and includes security and reliability features. These attributes create an opportunity for unique usages leveraging support from the processor, which is being enabled with the Cascade Lake.

Hardware Interface to Persistent Memory

Intel Optane DC persistent memory uses DDR4 electrical and mechanical interfaces with proprietary protocol extensions. The memory channel can be shared between DDR4 DIMMs and Intel Optane DC persistent memory modules. With a memory module capacity of up to 512 GB, it allows systems to be configured with greater than 3 TB of memory per processor. The memory module is accessed in 64B cache line granularity similar to DDR4, and has latency characteristics that are close to DDR4.

The Intel Optane DC persistent memory module is designed to optimize the performance of the 3D XPoint media. As such, both the data and command paths are implemented in hardware. The module has its own Power Management Integrated Circuit that is designed to generate all the needed power rails for the various components on the module. It also has an integrated controller that manages all the components on the module and configures them for optimal operation.

The other aspect of the hardware interface is the persistence boundary, the point in the data path at which stores are considered persistent. The software has no way of knowing that a store has completed its trip all the way to the persistent device, so it has to rely on support from the processor to provide the persistence domain boundary.

The minimum support needed is that updates to a cache line be pushed to write pending queue in the memory controller of the processor and platform provide support to make sure that writes waiting in the memory controller write-pending queue (WPQ) have the opportunity to drain to persistent memory if a power failure occurs. This can be done through explicit WPQ flush operations under kernel control, or by using asynchronous DRAM refresh support in the platform which maintains power long enough to allow the queue to be drained. The Intel Optane DC persistent memory modules also have enough energy store to guarantee that

any 64B cache line that reaches the module will be guaranteed to reach its 3D XPoint media.

The Cascade Lake processor also provides additional support where applications can directly manage persistent memory at a cache line granularity. This is done through instructions that allow updates to a persistent memory location to be pushed to the memory subsystem either by flushing the entire cache (through WBINVD instruction), bypassing the processor cache using nontemporal stores, or using different flavors of cache line flush (CLFLUSH, CLFLUSHOPT, and CLWB) instructions. CLFLUSHOPT is a higher performance version of cache line flush instruction that is weakly ordered with respect to other CLFLUSH/CLFLUSHOPT instructions to different cache lines. CLWB instruction allows the processor to retain ownership of a cache line in a nonmodified state while updating the persistent memory with the latest value. The software uses memory fence instructions to ensure that stores have reached the persistence domain, which is, in this case, the memory controller. Only when stores reach the memory controller, the data can be treated as persistent by the application.

In most cases, users writing applications to take advantage of persistent memory do not need to worry about these details because software libraries provide higher level interfaces that abstract these details.

The addition of Intel Optane DC persistent memory modules sharing the memory channel with DDR4 DIMMs requires quality of service considerations in usages where there are multiple workloads accessing different memories. Most of the processor resources starting from the requesting core up to the memory channel are shared between the two types of memory. However, the memory controller resources (read-pending queue, WPQ, scheduler, etc.) are independent for each memory type. It is important to ensure that a burst of accesses to one type of memory does not flood the shared processor resources such that an access to the other memory type is unable to make forward progress. The Cascade Lake processor has features that ensure that accesses to Intel Optane DC persistent memory do not unduly impact the performance of DDR4 memory and vice-versa by monitoring and managing the shared resources.

Although the access latency of Intel Optane DC persistent memory is comparable to DRAM, it is not the same. As a result, some usages may benefit from using DDR4 DRAM memory as a cache for the Intel Optane DC persistent memory. This mode of operation is referred to as MEMORY mode and it is intended to allow unmodified applications to benefit from the larger capacity and lower overall cost of this new memory technology. Since DRAM memory used as cache can contain modified data, this mode of operation is a volatile memory mode. Application benefit for using DRAM memory as cache in volatile memory mode will depend on the size of DRAM memory used as cache and the access pattern of the application. The Cascade Lake processor uses a novel cache management scheme using a combination of inclusive and noninclusive DRAM cache to reduce DRAM bandwidth overhead for writes while also eliminating the complexity of managing invalidates to processor caches on the eviction of an inclusive line from DRAM cache.

Protecting data at rest is a very important aspect of Intel Optane DC persistent memory module since the information written to the modules is persistent even if used as volatile memory. A strong encryption engine is implemented in hardware to ensure a high level of data protection.

Software Interface to Persistent Memory

The software interface for using Intel Optane DC persistent memory has been designed in collaboration with industry partners to create a unified programming model for persistent memory. The Storage Network Industry Association (SNIA) formed a technical workgroup that has published a specification of the model shown in Figure 3. This software interface is independent of any specific persistent memory technology.

The model exposes three main capabilities as shown in Figure 3. The management path allows system administrators to configure persistent memory and check its health. The path that supports the traditional storage mode supports existing applications and file systems without any change—it sees the persistent memory as very fast storage. The most exciting path exposes persistent memory through a

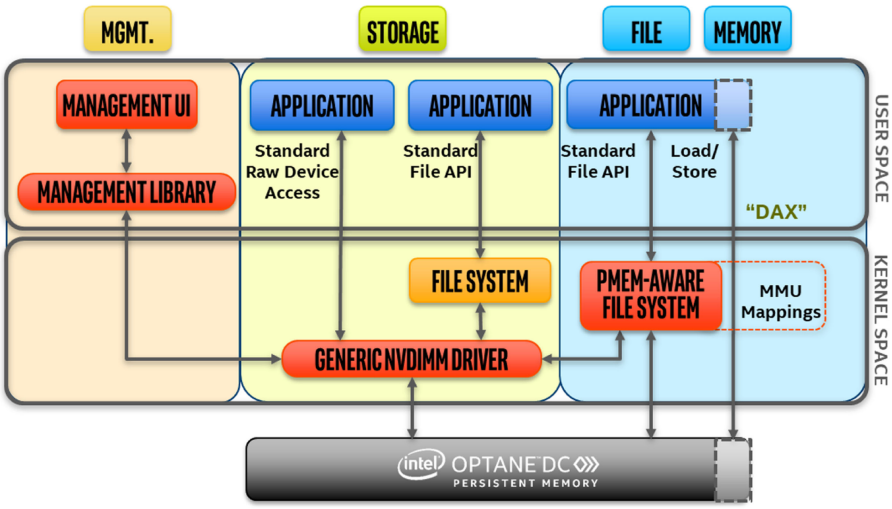

Figure 3. Persistent memory programming model from the SNIA.

pmem-aware file system so that applications have direct access using standard file APIs.

The pmem-aware file system provides direct access to the persistent memory when standard APIs are used to memory map files on that file system. This direct access does not use the page cache like traditional file systems and has been named DAX by the operating system vendors. Memory mapping calls like *mmap()* on Linux and *MapViewOfFile()* on Windows provide the DAX path shown in Figure 3.

Converting an application to map its memory into persistent memory and placing data structures in it can be a significant change. Keeping track of persistent memory allocations and making changes to data structures as transactions (to keep them consistent in the face of power failure) is complex programming that has not been required for traditional memory (because it is volatile) and is done differently for storage (since it is block-based).

The persistent memory development kit[6] provides a collection of libraries meant to make persistent memory programming easier for software developers. The developers only pull in the features they need, keeping their programs lean and fast while using persistent memory. These libraries are validated and performance tuned by Intel. They are open source and product neutral, working well on a variety of persistent memory products.

Persistent Memory Usage Examples

Some usage scenarios that illustrate the value of this new memory architecture are discussed here. Figure 4 shows the performance of an in-memory database using Apache Cassandra that compares a system with traditional memory hierarchy using DRAM and SSD with another system where SSD has been replaced with Intel Optane DC persistent memory.

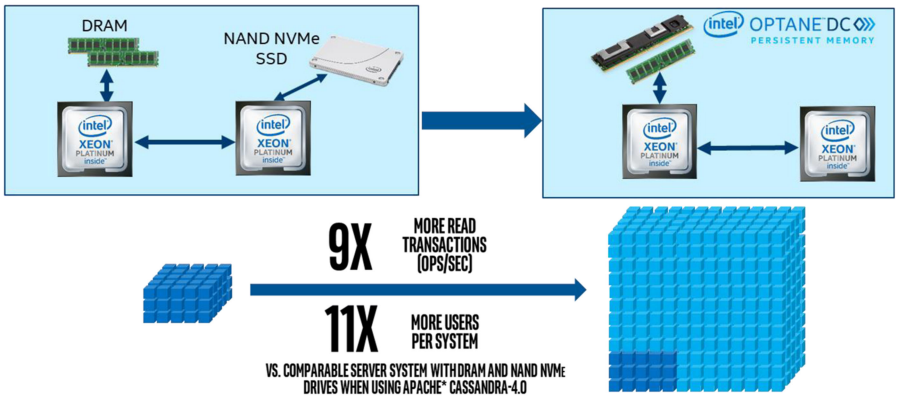

Figure 4. Performance comparison of an in-memory database with traditional memory hierarchy and a memory hierarchy with Intel Optane DC persistent memory. Performance results have been estimated based on Intel internal tests as of 29 May 2018. Performance claims are based on persistent memory-aware Apache* Cassandra-4.0 memory hierarchy workload doing 100% read versus DRAM+NAND NVMe and may not reflect all publicly available security updates.

Comparison of these two configurations shows that a system with Intel Optane DC persistent memory can support nine times more read transactions per second compared to a traditional system. This system can also support 11 times more users without violating response time expectation.

Figure 5 shows the replication of data over fabric to support higher availability

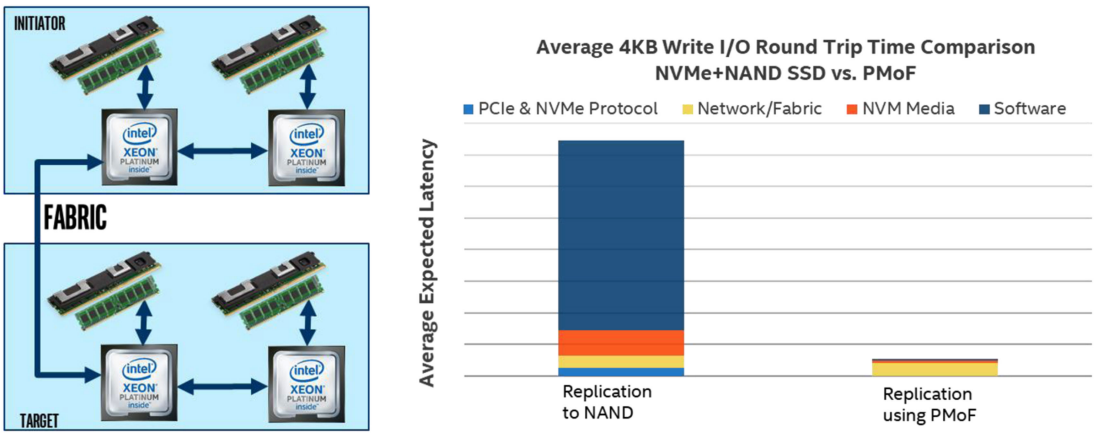

Figure 5. Comparison of performance of data replication over fabric on a system with traditional memory hierarchy and a system with Intel Optane DC persistent memory. Performance results are based on Intel internal testing as of 1 August, 2018, based on DRBD and RDMA and may not reflect all publicly available security updates.

and disaster recovery. It compares the replication time for a 4KiB block of data for a traditional system compared to a system using Intel Optane DC persistent memory. The software overhead of a storage device and media access latency is negligible with Intel Optane DC persistent memory compared to a system with a traditional memory hierarchy.

CONCLUSION

The Cascade Lake processor is compatible with first generation Intel Xeon Scalable platform and provides additional per core and aggregate performance through higher frequency and targeted improvements. It also addresses several variants of side-channel security exploits through hardware mitigations. In addition, there is a significant performance improvement for AI and deep learning inference applications using smaller data types to benefit users of general-purpose systems. It introduces support for Intel Optane DC persistent memory architecture in datacenter system memory hierarchy unleashing a new wave of innovations for data-intensive workloads.

NOTICES AND DISCLAIMERS

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

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.

REFERENCES

1. A. Kumar, D. Soltis, I. Esmer, A. Yoaz, and S. Kottapalli, "The new Intel Xeon scalable processor (formerly skylake-SP)," Hot Chips, Aug. 2017.
2. Amazing Inference Performance with Intel Xeon Scalable Processors at: https://ai.intel.com/amazing-inference-performance-with-intel-xeon-scalable-processors/
3. Intel MKL-DNN at https://01.org/mkl-dnn
4. Intel Optimizations for Deep Learning Frameworks available at: https://software.intel.com/ai-academy/frameworks

5. Intel 64 and IA-32 Architectures Software Developer's Manual, Sep. 2016.
6. SNIA NVM Programming Model at: https://www.snia.org/tech_activities/standards/curr_standards/npm
7. Intel Persistent Memory Development Kit (Intel® PMDK) at: http://pmem.io/

Mohamed Arafa is a Senior Principal Engineer with the Datacenter Architecture team, Intel, Chandler, AZ, USA. He received the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA in 1997, and is a Senior Member of the IEEE. Contact him at mohamed.arafa@intel.com.

Bahaa Fahim is a Principal Engineer with the Platform Engineering Group, Intel, Santa Clara, CA, USA. He received the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA in 2005, and is a member of the IEEE Computer Society. Contact him at bahaa.fahim@intel.com.

Sailesh Kottapalli is an Intel Senior Fellow and Chief Architect for the portfolio of Intel processors targeted for datacenter market segments. Contact him at sailesh.kottapalli@intel.com.

Akhilesh Kumar is a Principal Engineer with the Datacenter Processor Architecture team, Intel, Santa Clara, CA, USA. He received the Ph.D. degree in computer science from Texas A&M University, College Station, TX, USA in 1996, and is a member of the IEEE Computer Society. Contact him at akhilesh.kumar@intel.com.

Lily P. Looi is a Senior Principal Engineer with the Datacenter Architecture team, Intel, Hillsboro, OR, USA. She has decades of industry experience in CPU/SoC/platform architecture and performance. Contact her at lily.p.looi@intel.com.

Sreenivas Mandava is a Principal Engineer with the Platform Engineering Group, Intel, Santa Clara, CA, USA. He received the M.S. degree in computer science from SUNY Buffalo, Buffalo, NY, USA in 1997. Contact him at sreenivas.mandava@intel.com.

Andy Rudoff is a Senior Principal Engineer with Intel's nonvolatile memory software team, the Datacenter division, Boulder, CO, USA. He is a Founder of the SNIA NVM Programming Technical Workgroup and one of the creators of the Persistent Memory Development Kit. Contact him at andy.rudoff@intel.com.

Ian M. Steiner is a Principal Engineer on server CPU architecture with a specialization in power/performance optimization and debug. He has been involved in the architecture of almost all Intel server CPU's productized for the last decade. Contact him at ian.m.steiner@intel.com.

Bob Valentine is a Senior Principle Engineer with the Core and Xeon Architecture group, Intel, Haifa, Israel. He has more than thirty years of industry experience. He received the M.S. degree in computer engineering from Boston University, Boston, MA, USA in 1990. Contact him at bob.valentine@intel.com.

Geetha Vedaraman is a Platform Architect with the Artificial Intelligence Products Group, Intel, Santa Clara, CA, USA. She received the M.S. degree in computer science from University of Wisconsin-Madison, Madison, WI, USA in 1992. Contact her at geetha.vedaraman@intel.com.

Sujal Vora is a Principal Engineer with Silicon Engineering Group, Intel, Santa Clara, CA, USA. He leads Power/Thermal/Frequency modeling and projection for all server products. He received the M.S. degree in electrical engineering from Wright State University, Dayton, OH, USA in 2000, and is a member of the IEEE. Contact him at Sujal.a.vora@intel.com.