IBM's POWER10 Processor

William J. Starke, Brian W. Thompto, Jeff A. Stuecheli, José E. Moreira, *International Business Machines Corporation, Armonk, NY, 10504, USA*

The IBM POWER10 processor represents the 10th generation of the POWER family of enterprise computing engines. It is built on a balance of computation and bandwidth, delivered by powerful processing cores and intrachip interconnect, respectively. Multiple system interconnect infrastructures support configurations with up to 16 processor chips and up to 1920 simultaneous threads of execution, as well as an expansive memory system with up to 2 Petabytes of addressing space. Cross-system memory sharing and coherent accelerator attach are also supported. The POWER10 processing core has been significantly enhanced over its POWER9 predecessor, including the addition of an all-new matrix math engine. Throughput gains from POWER9 to POWER10 average 30% at the core level and three-fold at the socket level. Those gains can reach ten- or twenty-fold at the socket level for matrix-intensive computations.

The IBM POWER10 processor delivers significant gains in capacity and capability over its immediate POWER9 predecessor[1,2]: an average 20% single-thread performance boost, and 30% gain in core throughput over a wide range of applications. Combined with a two-and-a-half increase in the number of cores per package, these improvements result in three times or better per socket throughput on popular integer, floating-point, and commercial workloads, and 2–4 times increased memory bandwidth, depending on memory technology. For matrix math, the gains in performance can reach 10 or 20 times through a new computational engine.

Additional breakthroughs include: a new PowerAXON system interconnect with 1 TB/s of bandwidth per POWER10 chip and support for cross-system memory clustering; a new Open Memory Interface (OMI) that supports multiple industry-standard memory technologies on the same processor chip; a modular building block die that enables systems with up to 1920 simultaneous threads of execution; hardware-enforced security to protect sensitive code and data from attacks; and AI-optimized machine instructions to address the increased computing demands of modern machine learning/deep learning business applications.

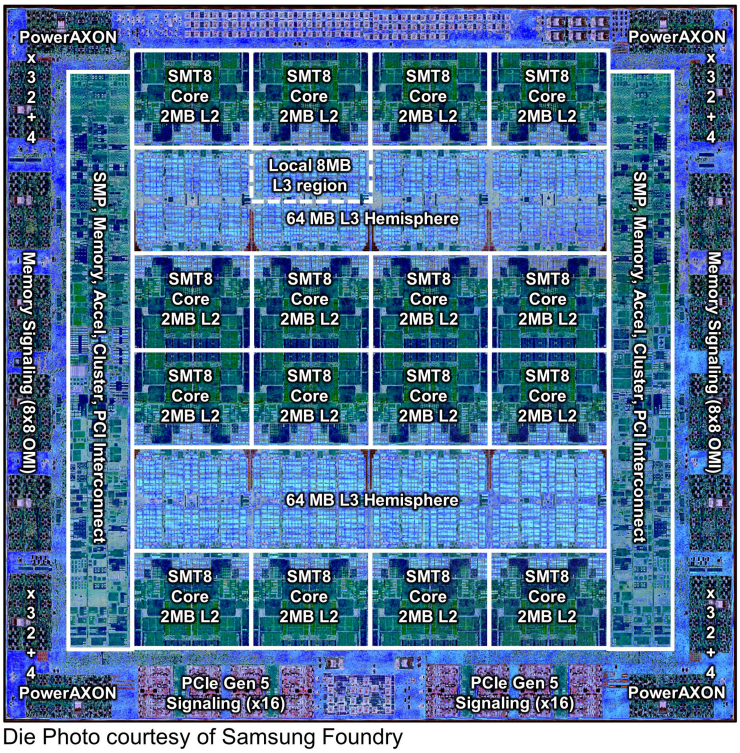
0272-1732 © 2021 IEEE
Digital Object Identifier 10.1109/MM.2021.3058632
Date of publication 10 February 2021; date of current version 26 March 2021.

POWER10 PROCESSOR CHIP

A POWER10 processor die (see Figure 1) consists of 18 billion transistors in 602 mm^2 of silicon, compared to 8 billion transistors in POWER9, and is built in Samsung's 7-nm technology with 18 metal layers. The central part of the die, approximately 300 mm^2, is occupied by 16 enterprise-grade cores, each capable of running eight simultaneous threads of execution (SMT8), and their associated 2- and 8-MB levels 2 and 3 cache regions, respectively. To better match the supply and demand of processor chips with the maximum number of cores, we cap the number of active cores in a die to 15, keeping one core as a manufacturing spare. This results in up to 120 simultaneous threads of execution, backed by 120 MB of level-3 cache.

The remaining half of the POWER10 processor chip area is dedicated to the system interconnect, including the two protocol spines to the left and right of the core/cache complex, supporting the various interconnect protocols for memory, multiple processors, accelerators, clusters, and I/O. The periphery of the die is filled with high bandwidth, power efficient signaling circuits that implement the PowerAXON,[3] OMI,[3] and PCI Gen5 I/O infrastructures.

Not visible in Figure 1 are the large numbers of communication trunk lines, which run horizontally over the two L3 hemispheres and vertically over the protocol spines. The placement of the L3 hemispheres, the protocol spines, the trunk lines, the intercept of these vertical and horizontal trunk lines, and the location of the protocol spines next to the



Die Photo courtesy of Samsung Foundry

FIGURE 1. POWER10 processor chip. Approximately half the die area is dedicated to cores and caches. The other half is for the various system interconnects, including memory interfaces, SMP, accelerators, clustering, and I/O.

signaling infrastructure are the result of rearchitecting the chip floorplan around a computation-to-bandwidth balance.

POWER10 processor chips can be packaged in either single- or dual-chip modules (SCM/DCM). The SCM configuration is optimized for scale-up systems and maximizes power, interconnect bandwidth, and memory capacity delivered to each core. It also supports more flexible topologies, allowing configurations with up to 16 processor chips. The DCM configuration is optimized for scale-out systems and maximizes computational and I/O density while trading off the power and memory per core compared with the SCM. It limits configurations to a maximum of four DCMs (eight processor chips).

The POWER10 chip introduces new security features for cloud paradigms that extend trusted virtualization environments to include protected containers and include in-line memory encryption and application level protections against attacks.

POWER10 SYSTEMS

POWER10 systems are built around the three interconnect infrastructures shown in Figure 2: the OMI, for connecting processors to memory; the PowerAXON interface, for interconnecting processor chips to other processors and accelerators and for implementing cross-system memory clustering; and PCI Gen 5 for I/O and other system interconnect.

PowerAXON and OMI signaling runs at rates of up to 32 GT/s. With a combined 256 bidirectional lanes, this results in up to 2 TB/s of total bandwidth on a processor chip, with 128 lanes and 1 TB/s for each interface. These are shown to the left and right of the POWER10 chip in Figure 2, respectively. (See Figure 1 for physical placement of the interfaces. Each PowerAXON corner has 32 lanes plus 4 spares.)

THE POWER10 CHIP INTRODUCES NEW SECURITY FEATURES FOR CLOUD PARADIGMS THAT EXTEND TRUSTED VIRTUALIZATION ENVIRONMENTS TO INCLUDE PROTECTED CONTAINERS AND INCLUDE IN-LINE MEMORY ENCRYPTION AND APPLICATION LEVEL PROTECTIONS AGAINST ATTACKS.

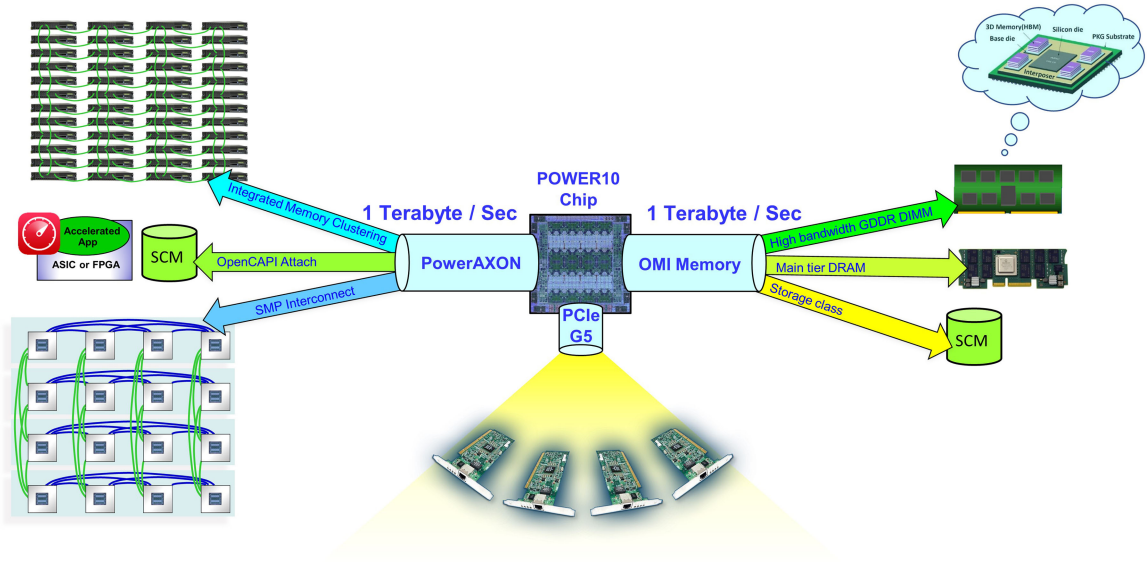


FIGURE 2. POWER10 system interconnect. OMI is used for attaching memory to the processor. PowerAXON provides the SMP, clustering, and accelerator interfaces. PCIe Gen5 is used for I/O and other interconnect.

OMI is a technology agnostic memory interface based on open standards. Memory is attached to the processor chip through an OMI-compliant buffer chip, which encapsulates technology specific requirements as first introduced in IBM's POWER8 processor and its companion Centaur memory buffer chip.[4] POWER10 systems will initially use DDR4 memory, through a buffer chip built by Microchip.[3] This buffer chip implements 25.6-GT/s signaling over an 8-bit interface, which matches its 8-byte DDR4 channel operating at 3.2 GT/s. With 16 channels in use, up to 410 GB/s of peak DDR4 bandwidth can be achieved per POWER10 chip, with a latency that is only 10 ns over traditional DDR4 DIMMs. DDR5 memory DIMMs can be supported later through a future buffer chip.

Alternative memory technologies can also be deployed with POWER10 processors using OMI. As shown in the right half of Figure 2, both high-bandwidth GDDR memory and high-capacity nonvolatile storage-class memory can be connected to the same OMI channels through corresponding buffer chips and using standard OMI DIMM form factors. A fully populated 16-channel GDDR configuration would achieve over 800 GB/s of memory bandwidth to a single POWER10 processor. This approaches the bandwidth achieved with high-bandwidth memory (HBM), but at higher capacities and lower cost. Alternatively, a storage-class (nonvolatile) memory solution could achieve capacities in the Terabytes per DIMM range.

The PowerAXON infrastructure is used for system scaling, including multiprocessor interconnect, device attach, and memory clustering. The largest single-system configuration consists of 16 SCMs, interconnected as shown in the lower left corner of Figure 2. Each module is at most two hops away from any other module. A system with this configuration can run up to 1920 simultaneous threads of execution and contain up to 256 OMI DIMMs (16 DIMMs attached to each of 16 SCMs), with a maximum capacity of 1 Petabyte of memory. (POWER10 processors have a 2-Petabyte physical memory address space.)

PowerAXON can also be used to support OpenCAPI, an open, asymmetric protocol for coherently attaching compute accelerators, memory devices, network interfaces, and storage controllers, either in a device slot or a cabled external enclosure. Since its introduction with POWER9, a variety of vendors have provided OpenCAPI-attached devices that expand and enhance the functionality of POWER systems. POWER10 OpenCAPI provides a new level of performance and functionality over the prior version.

The third and final functionality of PowerAXON in POWER10 that we discuss in this article is memory clustering, shown in the top left corner of Figure 2. This new feature of POWER10, which is called *memory inception*, delivers the long-sought functionality of server disaggregation. Memory inception enables systems to directly share each other's main memory. The latency through memory inception is 50–100 ns over that of a remote (2-hop) socket within a server, and it is still low enough to be used as main memory. The protocol for memory inception is built on top of the OpenCAPI protocol and

different from the SMP protocol used to build (up to) 16-socket systems. Memory inception does not implement a cache coherence scheme and is not meant to enable larger single system image configurations. Rather, the goal is to allow one server to map its address space to the physical memory of another server.

As a scenario for using memory inception, consider the case of a cluster of homogeneous servers, each with enough memory for the average workload. By borrowing memory from other machines, a hosting system can run large memory workloads that go beyond the capacity of any single server. Another scenario is a hub-and-spoke configuration, in which a very large central server has a big pool of memory, distributed as needed across a large set of much smaller machines. This combines the cost efficiency of small machines with the memory capacity of a much larger server.

Memory inception can also be used as the message layer for a large cluster of POWER10 servers. Combined with the processor's 2-Petabyte address space, memory inception can use the address translation facilities in each server to create a multihop interconnect, with messages delivered simply by writing to the target memory. A robust, fully hardware managed end-to-end message capability is possible in clusters with thousands of nodes, delivering high bandwidth, low latency, and flexible topologies.

The final component of the POWER10 system interconnect, shown in the bottom of Figure 2, are the PCIe Gen 5 interfaces. PCIe is central to the I/O infrastructure of POWER10 systems, and up to 64 lanes are available in a DCM (32 lanes per chip). With a signaling rate of 32 GT/s, a single DCM can achieve 252 GB/s of I/O bandwidth in each direction.

POWER10 CORE

The POWER10 core is the processing engine that runs both system and user software, responsible for the computational capacity and capability of POWER10 systems. There are two focus areas in the design of the POWER10 core: performance strength and power efficiency. A 30% average increase in core throughput while cutting power consumption in half combine to deliver a 2.6-fold average increase in energy efficiency for computations. The increased energy efficiency has allowed the implementation of DCMs with up to 30 SMT8 cores, and up to a three-fold throughput over current POWER9 modules with similar power consumption. The POWER10 core retains the modular architecture from POWER9 that provides a second variant of the chip with twice as many SMT4 cores per chip (up to 60 per DCM).

The microarchitecture of the POWER10 core, together with key factors affecting its performance and power efficiency, are shown in Figure 3. The block diagram shows those microarchitecture resources available for the execution of 1 to 4 simultaneous threads, corresponding to half of the total resources in an SMT8 core. POWER10 core components colored in green were somewhat improved in capacity over the predecessor POWER9 core. Those colored in blue had their capacity at least doubled and those in red had their capacity at least quadrupled. These additional resources along with various other improvements in latency and microarchitecture are responsible for the 30% average increase in core throughput and a much higher boost in performance in some cases.

Each POWER10 SMT8 core has an associated 2 MiB L2 cache that provides both instructions and data and is four times the capacity of POWER9. For each half of the core, instructions are fetched at a sustained rate of up 32-bytes per cycle and predecoded before being installed in a 48-KiB instruction cache (50% more capacity than POWER9). During the predecode stage, select pairs of instructions can be identified for fusion into a single internal operation of the microarchitecture, which leads to a faster and more efficient execution of those instructions. The new 64-bit prefix instructions in Power ISA 3.1[5] are also identified in that stage. POWER10 then decodes and dispatches to the execution slices up to 8 instructions per cycle per thread, or 16 instructions per cycle per SMT8 core. This represents a 33% increase in dispatch rate when compared to POWER9. Over a thousand instructions can be in-flight, from dispatch to commit, in a POWER10 SMT8 core, representing a doubling of the out-of-order execution capabilities over POWER9. The translation lookaside buffer (TLB) has been increased four-fold with 8192 entries per SMT8 core, while at the same time reducing the latency and increasing throughput over POWER9.

The four execution slices of POWER9 have been widened to 128 bits each. This has resulted in a doubling of the general SIMD rate of execution, to a maximum of four SIMD instructions per cycle per thread or up to 8 SIMD instruction per cycle per SMT8 core. Crypto processing in the execution slices have also been enhanced, with an overall four-fold gain in throughput from POWER9 to POWER10 core.

A single-thread of execution can load up to two 32-byte data chunks per cycle from the L1 cache,

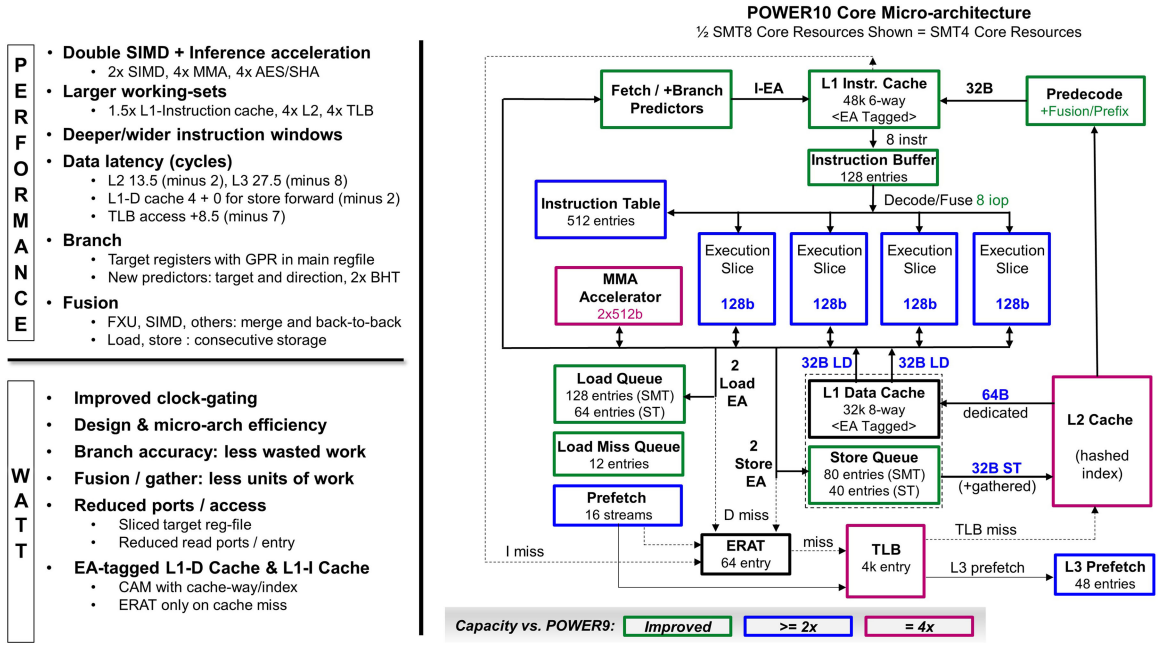


FIGURE 3. POWER10 core microarchitecture. The boxes on the left show the improvements over POWER9 on performance and power efficiency, respectively. The latency numbers include both absolute values and improvements over POWER9.

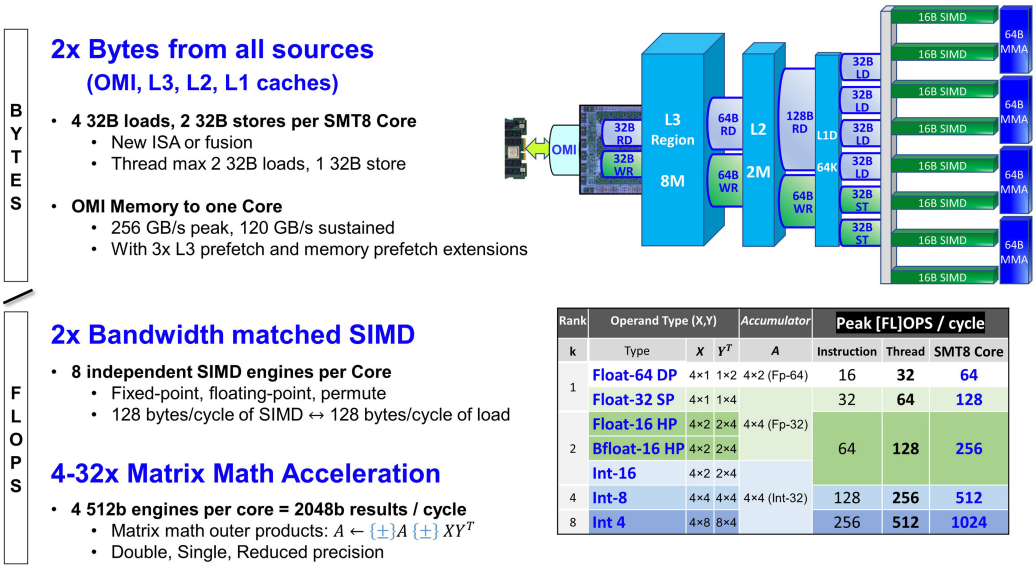
with a total SMT8 core load bandwidth of 128 bytes per cycle. (The same bandwidth can also be achieved from the L2 cache.) A single thread of execution can store up to four instructions per cycle by gathering from up to two store queue entries when each entry includes a fused store operation. Stores always target the L2 cache and maximum bandwidth is 32 bytes per cycle per thread or 64 bytes per cycle per SMT8 core.

Complementing the four general purpose execution slices, POWER10 introduces a new matrix math accelerator (MMA) unit, optimized for the execution of new matrix instructions in Power ISA 3.1. The instructions perform BLAS2- and BLAS3-class operations on eight 512-bit accumulator registers that are added to the architecture. The instructions use either two or three 128-bit vector-scalar registers to perform rank-1, -2, -4 or -8 updates on either a 4×2 or 4×4 matrix stored in an accumulator. Each input vector-scalar register contains either a 2×1 vector of double-precision elements, a 4×1 vector of single-precision elements, a 4×2 matrix of 16-bit elements (half-precision floating-point,[6] bfloat16,[7] or signed integer), a 4×4 matrix of 8-bit elements (signed/unsigned integer), or a 4×8 matrix of 4-bit elements (signed integer).

The MMA microarchitecture reduces data switching by storing the accumulators locally in the unit itself, significantly reducing the total data movement (bits $\times$ distance) when compared to an equivalent 512-bit SIMD operation. The result is improved power efficiency and higher frequency, enabling the POWER10 core to achieve a four-fold increase in matrix math throughput compared to the POWER9 core.

In addition, a focus on power efficiency dominated many other elements of the POWER10 core microarchitecture and design. When compared to the POWER9 core, there is more use of clock-gating and an emphasis on reducing data switching. The branch prediction accuracy has been improved, which results in less wasted work and improves thread latency. Instruction fusion also helps with both performance and power efficiency, by combining multiple instructions in fewer operations. POWER10 supports both independent and dependent forms of fusion. Dependent fusion combines the execution of two instructions that share a register dependence into a single operation (with no dependent latency) or a latency optimized pair of operations, whereas independent fusion enables the combining of loads or stores to adjacent memory locations into a single wider access reducing resource consumption and conflicts.

The register file for the general-purpose and vector-scalar registers requires four times fewer write-



Rank	Operand Type (X,Y)			*Accumulator*	Peak [FL]OPS / cycle		
k	Type	*X*	Y^T	*A*	Instruction	Thread	SMT8 Core
1	Float-64 DP	4×1	1×2	4×2 (Fp-64)	16	32	64
	Float-32 SP	4×1	1×4		32	64	128
2	Float-16 HP	4×2	2×4	4×4 (Fp-32)	64	128	256
	Bfloat-16 HP	4×2	2×4				
	Int-16	4×2	2×4				
4	Int-8	4×4	4×4	4×4 (Int-32)	128	256	512
8	Int 4	4×8	8×4		256	512	1024

FIGURE 4. POWER10 core speeds and feeds. Load/store and SIMD bandwidth have been doubled over POWER9, matching SIMD and load throughputs. The matrix math unit offers increased throughput in computational-intensive operations.

ports per entry, when compared to POWER9, without compromising performance. The L1 data and instruction caches have been converted to use effective-address tags, which means that address translation from effective- to real-address only needs to be performed on L1 cache misses, as opposed to on every load or store, as with POWER9. Cache latency has also been optimized reducing time in flight and pipeline hazards including the addition of zero-cycle store forwarding latency for loads and reduced latency for L2 and L3 caches.

The L1 data cache is a write-through cache and stores commit to L2, which is physical-address indexed and tagged, and inclusive of L1. L3 is a victim cache for evictions from its local L2 and is also physical-address indexed and tagged. The L3 region associated with a core can also be populated with cast-outs from other cores depending on the state of dynamic sharing policies. A miss from a load or store by a core can be sourced by any L2 or L3, either on the same or another chip in the system.

The speeds and feeds of the POWER10 core are summarized in Figure 4. The SMT8 core attaches to a multilayered memory hierarchy, which provides it with up to 32 bytes/cycle of simultaneous read and write bandwidth to main memory. Read and write bandwidth to the local 8 MiB L3 cache is 64 bytes/cycle for an SMT8 core. Read and write bandwidth to the level-2 and level-1 cache is 128 and 64 bytes/cycle for an SMT8 core, respectively.

THE IBM POWER10 PROCESSOR COMBINES INNOVATIONS IN SILICON TECHNOLOGY, SYSTEM INTERCONNECTS, MEMORY SYSTEMS, AND PROCESSING CORE TO DELIVER A COMPUTE ENGINE THAT IS OPTIMIZED FOR THE MODERN DAY DEMANDS OF THE ENTERPRISE ON PREMISES AND IN THE CLOUD.

The table in the lower right corner of Figure 4 shows the computation rates that can be achieved with the MMA instructions. The columns show the maximum rate of operations per cycle for a single instruction, a single-thread (which can issue up to two instructions per cycle) and per SMT8 core. The operation rate increases as the input data type gets shorter, and varies from a maximum of 64 (floating-point) operations per cycle per SMT8 core when the inputs are double-precision floating-point elements to a maximum of 1024 (integer) operations per cycle per SMT8 core when the inputs are 4-bit integers.

POWER10 PERFORMANCE

Comparisons of performance between POWER9 and POWER10 are shown in Figure 5. The comparisons are

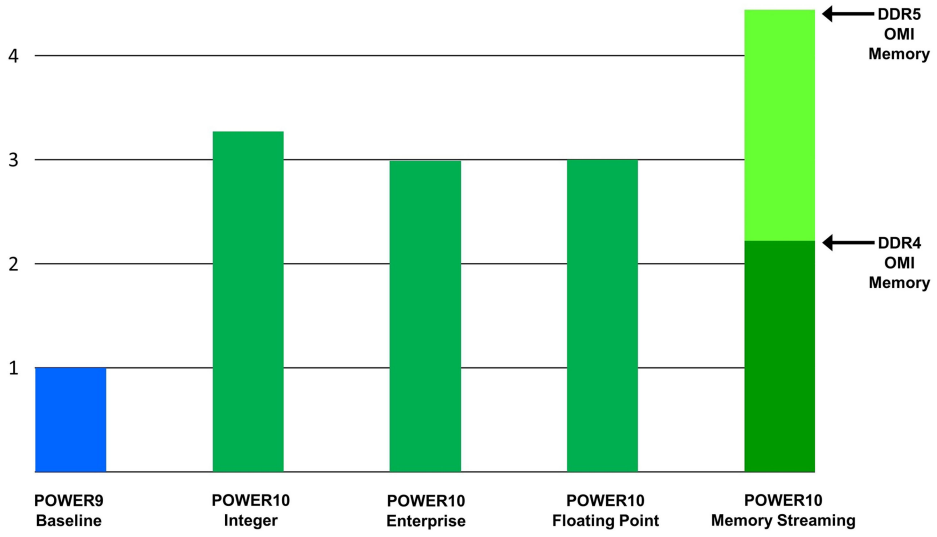


FIGURE 5. POWER10 general purpose socket performance gains. The three-fold improvement in performance comes from a combination of increased number of cores and more powerful cores. DDR5 will double memory bandwidth in the future.

derived from presilicon simulations and have been correlated against first-pass silicon. We do not yet have the final version of the chips and the results reflect the projected frequency of operation for production POWER10 parts. The figures are for a dual-socket POWER10 system relative to a dual-socket POWER9 S924 server. We observe a three-fold improvement in performance across integer (SPECint2017_rate), floating-point (SPECfp2017_rate), and commercial benchmarks. For memory streaming benchmarks, the POWER10 gains over POWER9 range from two- to four-fold, using DDR4 and DDR5 memory, respectively.

For computations that are heavy on matrix math, the gains from POWER9 to POWER10 are even more substantial, as shown in Figure 6. LINPACK is expected to run ten times faster in POWER10 than POWER9, when compared socket-to-socket. The same is expected for single-precision floating-point implementation of the Resnet-50 benchmark. When some of the new mixed-precision math features of POWER10 are taken into account, our evaluation shows that Resnet-50 will execute up to 15 (with bfloat16 data type) or 20 (with 8-bit integer data type) times faster than the standard single-precision Resnet-50 in POWER9.

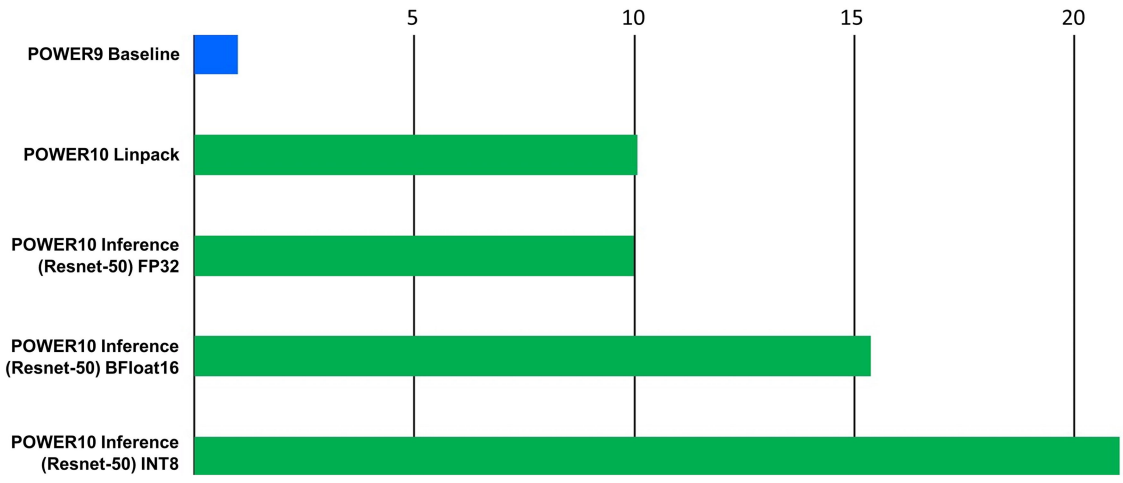


FIGURE 6. POWER10 SIMD/AI socket performance gains. The matrix math accelerator delivers four times the throughput of POWER9 SIMD. Combined with two-and-a-half times the number of cores, it results in a ten-fold improvement in socket throughput for computational-intensive operations. Further gains are possible with the new reduced-precision operations.

CONCLUSION

The IBM POWER10 processor combines innovations in silicon technology, system interconnects, memory systems, and processing core to deliver a compute engine that is optimized for the modern day demands of the Enterprise on premises and in the cloud. Innovative memory interfaces and cross-system memory functionality enable new architectures for cloud computing, while improved performance and power efficiency of the computational cores support a three-fold improvement in comparable system throughput over a broad spectrum of applications. For matrix-intensive computations in the fields of machine- and deep-learning, the gains can be more than ten-fold.

REFERENCES

1. S. K. Sadasivam, B. W. Thompto, R. Kalla, and W. J. Starke, "IBM power9 processor architecture," *IEEE Micro*, vol. 37, no. 2, pp. 40–51, Mar./Apr. 2017.
2. W. A. Hanson, "The CORAL supercomputer systems," *IBM J. Res. Develop.*, vol. 64, no. 3/4, pp. 1:1–1:10, May/Jul. 2020.
3. J. Stuecheli, S. Willenborg, and W. Starke, "IBM's next generation POWER processor," in *Proc. IEEE Hot Chips 31 Symp.*, Cupertino, CA, USA, 2019, pp. 1–19.
4. W. J. Starke *et al.*, "The cache and memory subsystems of the IBM POWER8 processor," *IBM J. Res. Develop.*, vol. 59, no. 1, pp. 3:1–3:13, Jan./Feb. 2015.
5. IBM Corporation, "Power ISA version 3.1," May 2020.
6. "*IEEE Standard for Floating-Point Arithmetic—Redline*," in IEEE Std 754-2019 (Revision of IEEE 754-2008), pp. 1–148, Jul. 2019.
7. G. Tagliavini, S. Mach, D. Rossi, A. Marongiu, and L. Benin, "A transprecision floating-point platform for ultra-low power computing," *Proc. Des., Autom. Test Eur. Conf. Exhib.*, Dresden, Germany, 2018, pp. 1051–1056, doi: 10.23919/DATE.2018.8342167.

WILLIAM J. STARKE is an IBM Distinguished Engineer and the Chief Architect and Engineer for the IBM POWER10 processor. He is responsible for shaping the processor cache hierarchy, symmetric multiprocessor (SMP) interconnect, cache coherence, memory, and I/O controllers, accelerators, and logical system structures for Power Systems. He is an IBM Master Inventor, authoring roughly 300 patents. Starke received a B.S. in computer science from Michigan Technological University. Contact him at wstarke@us.ibm.com.

BRIAN W. THOMPTO is an IBM Distinguished Engineer and the Chief Architect for the IBM POWER10 core. He has led global development teams across ten generations of IBM POWER and IBM System z processors. He has been recognized by two IBM corporate awards and as an IBM Master Inventor. Thompto received a B.S. in electrical engineering and computer science from the University of Wisconsin Madison. Contact him at bthompto@us.ibm.com.

JEFF A. STUECHELI is a Senior Technical Staff Member in the IBM Power Systems Development Group. He works on server hardware architecture. His most recent work includes advanced memory architectures, cache coherence, and accelerator design. He has contributed to the development of numerous IBM products in the POWER architecture family, most recently the POWER10 design. He has been appointed an IBM Master Inventor, authoring over 190 patents. Stuecheli received a B.S., an M.S., and a Ph.D. in electrical engineering from The University of Texas Austin. Contact him at jeffas@us.ibm.com.

JOSÉ E. MOREIRA is a Distinguished Research Staff Member with the IBM Thomas J. Watson Research Center. Moreira received a B.S. in physics and a B.S. and an M.S. in electrical engineering from the University of Sao Paulo and a Ph.D. in electrical engineering from the University of Illinois at Urbana-Champaign. He is a Fellow of IEEE and a Distinguished Scientist of ACM. Contact him at jmoreira@us.ibm.com.