IBM Power9 Processor Architecture

The IBM Power9 processor has an enhanced core and chip architecture optimized for emerging workloads with superior thread performance and higher throughput to support next-generation computing. Multiple variants of silicon target the scale-out and scale-up markets. With a new core microarchitecture design, along with an innovative I/O fabric to support accelerated computing requirements, the Power9 processor meets the diverse computing needs of the cognitive era and provides a platform for accelerated computing.

Satish Kumar Sadasivam
Brian W. Thompto
Ron Kalla
William J. Starke
IBM

The Power9 processor is fabricated in 14-nm fin field-effect transistor (FinFET) process technology and contains 8 billion transistors. It uses a 17-metal-layer stack technology and continues to exploit embedded DRAM (eDRAM) for its L3 cache design. The Power9 processor (see Figure 1) can have up to 24 cores on a single chip, with each core supporting up to four hardware threads. The Power9 processor also comes in a 12-core variant, which supports up to eight hardware threads built on the same basic building block of the Power9 core microarchitecture.

The Power9 core delivers strong thread performance without compromising the high-socket throughput. This is achieved by an efficient pipeline with a new core microarchitecture and a highly efficient memory subsystem, which comes in two variants to cater to the needs of the scale-out and scale-up domains. Scale-out systems support up to two nodes and are suitable for distributed cloud computing environments. The scale-up systems support large-scale symmetric multiprocessor (SMP) configurations with huge memory capacity and bandwidth for enterprise requirements.

The cores of the Power9 processor are supported and fed by a 120-Mbyte L3 cache based on nonuniform cache architecture (NUCA). The L3 cache is designed as 12 regions of a 20-way associative cache with advanced replacement policies. The L3 cache system is, in turn, fed by an on-chip fabric that delivers up to 7 Tbytes per second (TBps) of on-chip bandwidth.

The Power9 chip includes an I/O subsystem with 48 PCI Express (PCIe) Gen4 lanes, enabling heterogeneous computing. Two interfaces of high-bandwidth signaling technology enable a large SMP and accelerator computation, as shown in Figure 2. A 16-GBps interface provides support for connecting neighboring socket(s) of an SMP system. A 25-GBps interface supports a wide range

Published by the IEEE Computer Society

of external attach capabilities for the design of heterogeneous computing systems as well as providing support for SMP connections in SMP systems.

Key features, targeted at cloud and virtualization domains, include a new interrupt architecture, quality-of-service assists, hardware-enforced trusted execution environments, and workload-optimized operating frequency.

Workload-Optimized Design

The Power9 processor has been engineered to target four emerging domains:

- emerging analytics, AI, and cognitive computing;
- technical and high-performance computing (HPC);
- cloud and hyperscale datacenters; and
- enterprise computing.

These four domains have many distinct requirements and highly diverse workloads. The Power9 design architecture and family of chips were designed to address the requirements of all these categories of workloads. The basic Power9 building block is a modular design that can support multiple targeted implementations and create a family of processors with state-of-the-art accelerated computing capabilities.

Table 1 summarizes the key requirements of emerging domains and select features of the Power9 processor suited for each domain.

Core Pipeline

The microprocessor's pipeline structure is subdivided into a front-end pipeline and several different execution unit pipelines. The front-end pipeline presents speculative in-order instructions to the mapping, sequencing, and dispatch units. It ensures orderly completion of the real execution path, discarding any other potential speculative results associated with mispredicted paths. The execution unit pipelines allow out-of-order issuing of both speculative and nonspeculative operations. The execution unit pipelines are composed of execution slices and progress independently from the front-end pipeline and from one another.

As Figure 3 shows, the Power9 core microarchitecture has a reduced pipeline

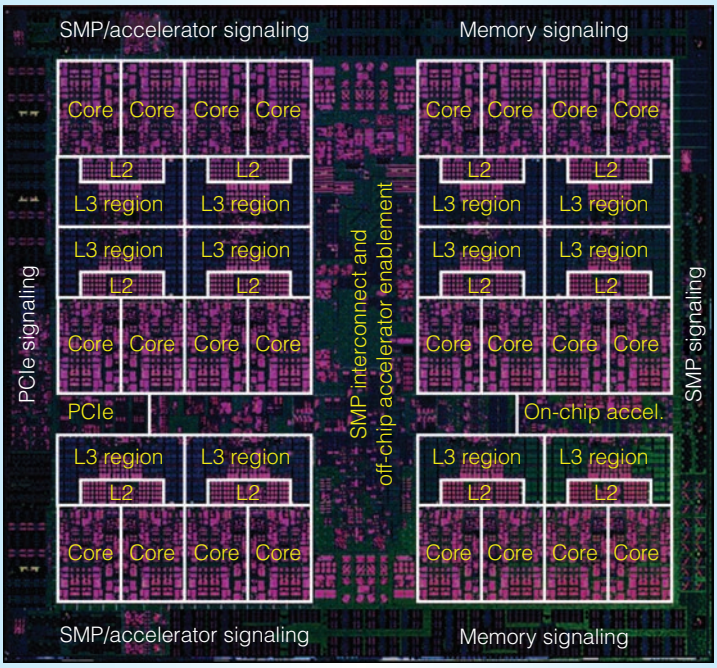

Figure 1. The Power9 chip diagram shows the number of cores, L2 and L3 cache regions, and interconnects.

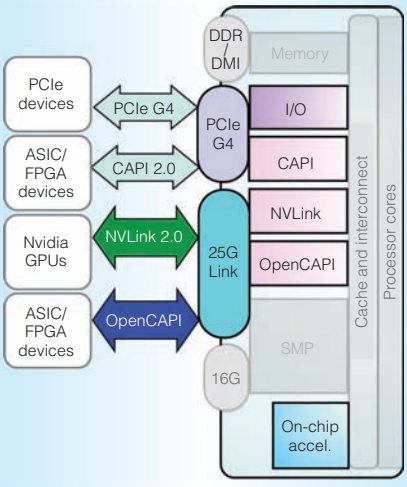

Figure 2. Power9 processor integration with external accelerating devices supporting PCIe Gen4, CAPI 2.0, NVLink2.0, and OpenCAPI.

Table 1. Key requirements of emerging domains.

Emerging domain	Support for key requirements
Emerging analytics, AI, and cognitive computing	• New core for stronger thread performance • Delivers twice the computational resources per socket • Built for acceleration and OpenPower-solution enablement
Technical and high-performance computing	• High-bandwidth interface for GPU attach • Advanced GPU/CPU interaction and memory sharing • High-bandwidth direct-attach memory
Cloud and hyperscale datacenters	• Power, packaging, and cost optimizations for a range of platforms • Superior virtualization features of security, power management, quality of service, and interrupt • State-of-the-art I/O technology for network and storage performance
Enterprise computing	• Large, flat, scale-up systems • Buffered memory for maximum capacity • Leading reliability, availability, and serviceability (RAS) • Improved caching

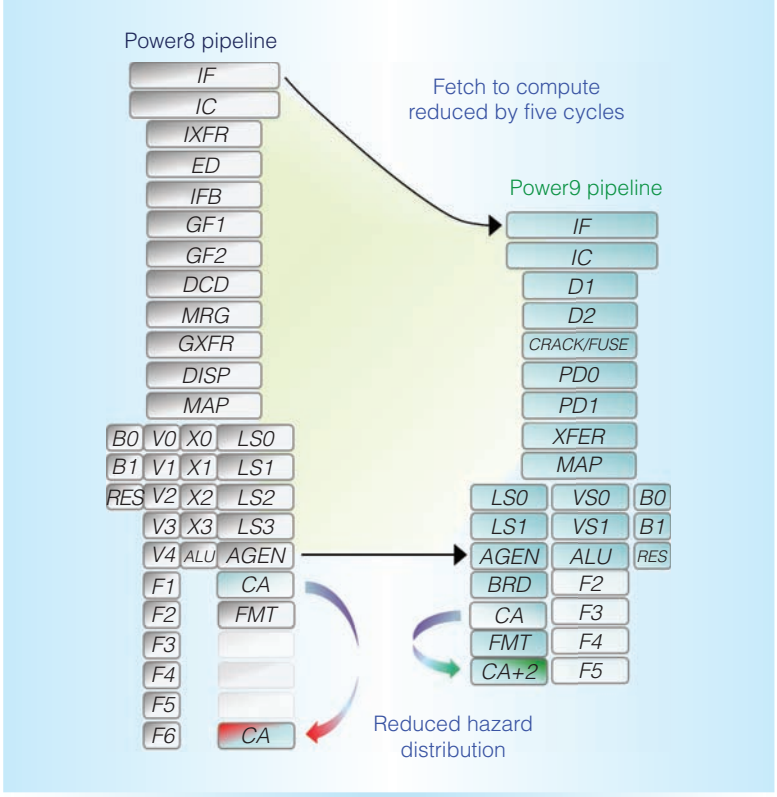

Figure 3. Pipeline diagram comparing Power8 and Power9 processor pipeline stages, also highlighting the reduction in the number of stages in the Power9 core.

length. The latency from fetch to computation is reduced by five cycles compared with the IBM Power8 design.[1] The latency from fetch to retirement is also reduced, for example, by eight cycles for floating-point operations. (The description for the various pipeline stages are provided in "Glossary" sidebar.) These improvements are enabled by microarchitectural changes while continuing to support similar cycle-time design constraints compared with IBM Power8. Central to these improvements is the adoption of an execution slice microarchitecture. Power9 also removes the instruction grouping technique from the front end of the core pipeline, one of the basic building blocks of earlier-generation IBM Power cores, enabling individual instruction allocation and retirement. Support for more robust operations within the sequencing and execution pipelines improves overall instruction efficiency and results in a reduction in dynamic instruction cracking at decode.

Variants of the core support completion of up to 128 (SMT4 core) or 256 (SMT8 core) instructions in every cycle. As a result, the core can free up the out-of-order resources quickly, which in turn supports a faster

Glossary

Here, we list some terms relating to the IBM Power8 and Power9 processor.

Power8

AGEN	Address generation
B	Branch
CA	Cache access
DCD	Decode
DISP	Dispatch to issuing resource
ED	Early decode
F	Floating point
FMT	Data formatting
GF	Dispatch group determination
GXFR	Group transfer
IC	Instruction cache access
IF	Instruction fetch
IFB	Instruction fetch buffer
IXFR	Instruction transfer
LS	Load store
MAP	Register mapping
MRG	Microcode selection and merge
RES	Branch resolution
X	Fixed point
V	Vector/float pipe

Power9

ALU	Arithmetic logic unit
BRD	Address broadcast
D	Decode
PD	Predispatch
VS	Vector scalar
XFER	Transfer

flow of instructions from the front end to the back end.

Advanced branch prediction improves the front-end efficiency and single-thread performance, resulting in a significant reduction in wasted instruction execution cycles due to branch mispredictions. The Power9 processor improves both direction prediction and target address prediction techniques to handle hard-to-predict branches.

The Power9 processor introduces new features to proactively avoid hazards in the load store unit (LSU) and improve the LSU's execution efficiency. Local pipe control of load and store operations enables better avoidance of hazards and reduces the hazard disruption with local recycling of instructions. New lock management control improves the performance of both contested and uncontested locks.

Execution Slice Microarchitecture

The Power9 core microarchitecture offers improved efficiency and is better aligned with workload requirements. To support future computing needs and enable the core microarchitecture to scale, the revamped core microarchitecture uses a modular execution slice architecture.

Essential to the modular execution slice architecture are the symmetric data-type execution engines. A symmetric data-type execution engine, called a *slice,* acts as a 64-bit computational building block and is coupled with a 64-bit load-store building block. Each computational slice supports a range of data types, including fixed-point, floating-point, 128-bit, scalar, and multiple data (SIMD) execution. This architecture enables the seamless interchange of different data types between operations. It also enables higher performance for a diverse range of workloads by providing shared computational resources and a shared datapath for all data types. This enables the core to have increased pipeline utilization of the execution resources and enables efficient management of instruction and dataflow through the machine.

Two execution slices are combined to form a super-slice, which enables 128-bit computation for both fixed-point and floating-point computations. This design enables the applications with scalar instructions to achieve greater performance, because each slice can handle various instruction types per scalar operation independently; one super-slice can handle either two scalar or one vector operation, providing robust performance for both scalar and vector codes. Two such super-slices combine to form a four-way simultaneous multithreading (SMT4) core. Four such super-slices form an SMT8 core.

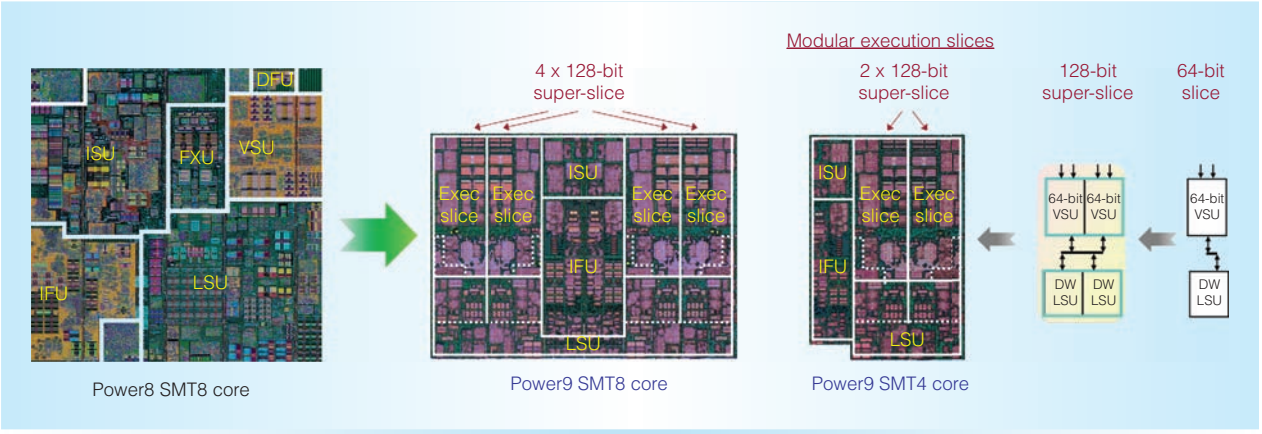

Figure 4. Modular execution slice diagram showing the slice and super-slice view of SMT4 and SMT8 Power9 core configurations.

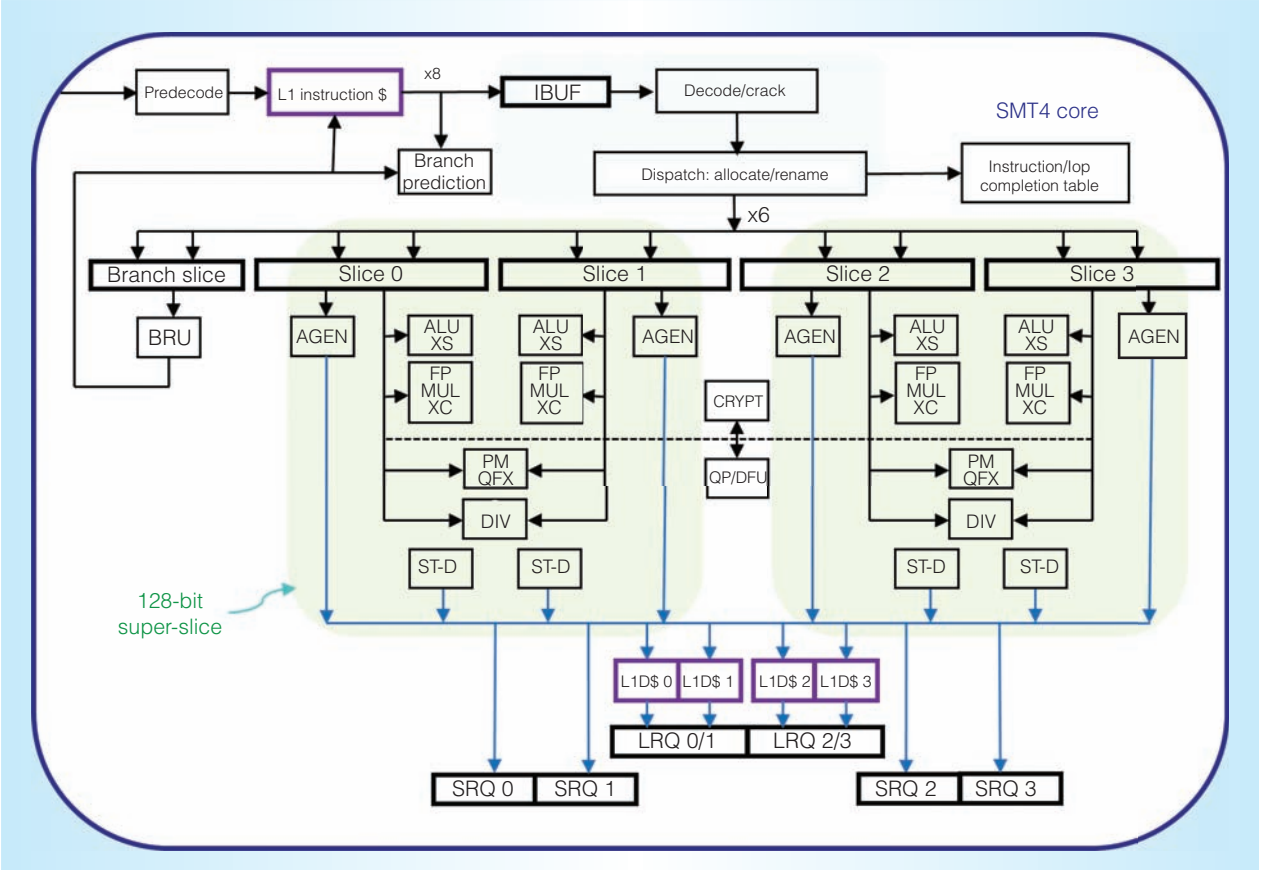

Figure 5. Power9 SMT4 core. The detailed core block diagram shows all the key components of the Power9 core.

Figure 4 shows several modular execution slices, which are the basic building blocks of the Power9 core microarchitecture.

Core Computational Capabilities

Figure 5 shows a more detailed view of a Power9 SMT4 core, along with its core

computational capabilities. The SMT8 core provides double the execution resources of the SMT4 core.

The Power9 SMT4 core includes a 32-Kbyte eight-way instruction cache. The instruction fetch unit can fetch up to eight instructions per cycle into the instruction buffer with a highly optimized branch predictor to support speculative fetching of instructions. The enhanced instruction prefetcher fetches instruction lines speculatively to reduce the instruction-cache miss occurrences. The decode unit can decode up to six instructions per cycle. The dispatch unit can dispatch up to six instruction operations per cycle to the back-end execution slices. The instructions are tracked using an instruction completion table that can track up to 256 operations out of order per SMT4 core. The Power9 core's issue capability is a maximum of nine instructions operations to the back-end execution slices.

The four execution slices in an SMT4 core can each issue any 64-bit computational/load/store operation and an address generation (AGEN) operation. The computational operations can be either fixed point or floating point (single or double precision). When issuing 128-bit operations, two execution slices are used. The Power9 core provides a separate branch slice for handling branch instructions. The execution pipes are commonly called vector and scalar unit (VSU) pipes because there is no differentiation in the execution engine for handling different data-type operations. Therefore, a single Power9 SMT4 core can handle up to four scalar 64-bit operations, including loads and stores, or two vector 128-bit operations, along with four load and store AGEN operations and one branch instruction every cycle. This makes the Power9 execution unit capable of handling any type of mixed-operation requirements with a high utilization of execution engines.

The VSU pipe of the Power9 SMT4 core can handle

- four arithmetic logic unit simple fixed-point operations,
- four floating-point or fixed-point multiply operations and complex 64-bit fixed-point operations,

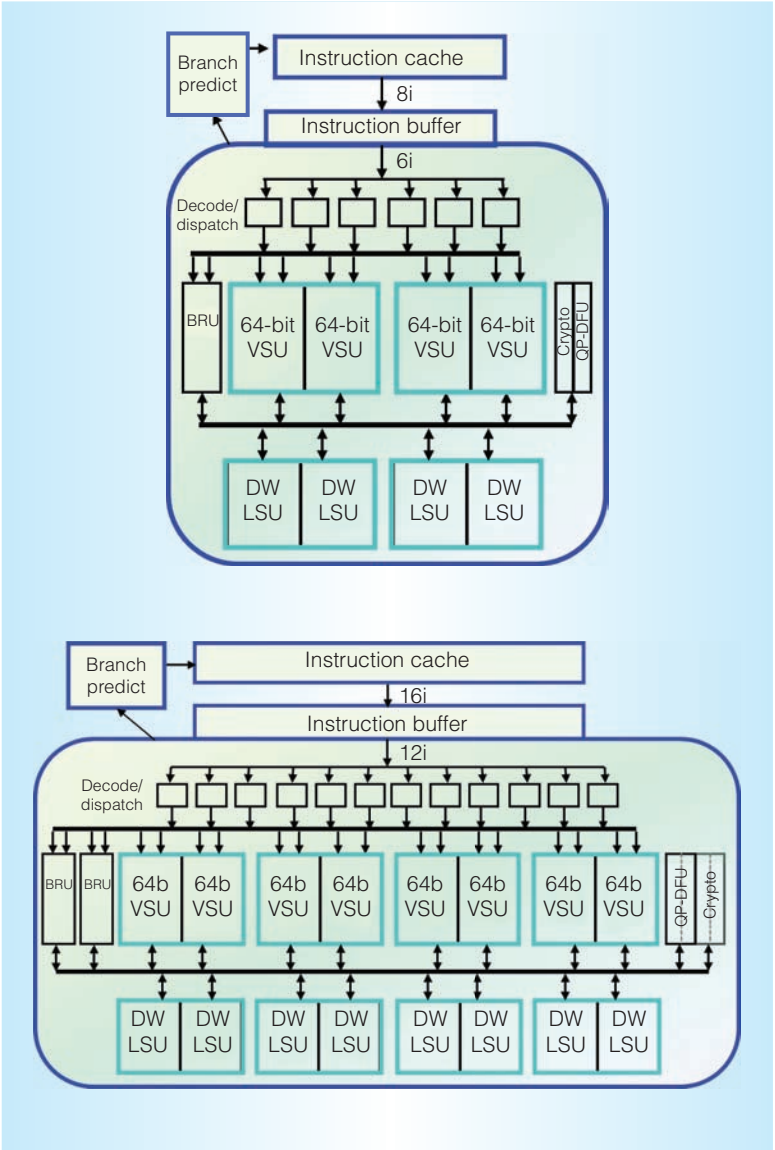

Figure 6. SMT4 and SMT8 core architecture shows the differences in fetch width, issue width, and number of slices available in each configuration.

- two 128-bit permute operations,
- two 128-bit quadword fixed-point operations,
- one 128-bit quadword floating-point operation,
- one decimal floating-point operation, and
- crypto operations.

The Power9 contains four double-precision floating-point units, one per slice. Each of these units is optimized for fully pipelined double-precision multiply-add functionality.

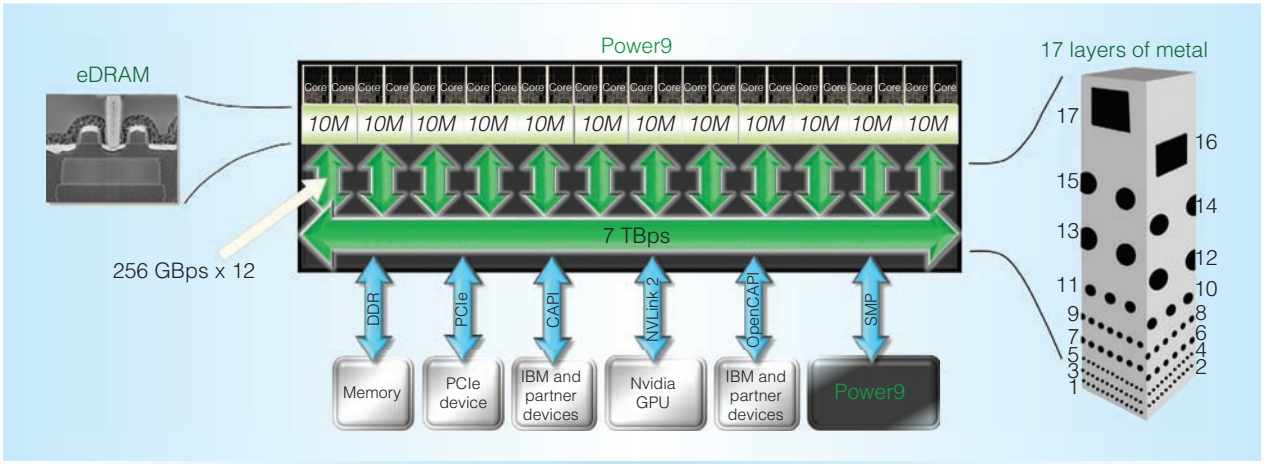

Figure 7. The SMT8 cache shows the 120-Mbyte L3 shared nonuniform cache architecture (NUCA) cache, high-throughput interconnects, and on-chip fabric.

In addition, each unit can perform the floating-point divide and square root instructions. The Power9 VSU implements the Vector Scalar Extension architecture, specifying two-way double-precision or four-way single-precision operations per cycle using one 128-bit super-slice.

LSU slices can handle up to four double-word loads or stores. Each SMT4 core has a private 32-Kbyte eight-way data cache that is accessed by the four-LSU execution slice. The L2 and L3 cache regions are shared by two SMT4 cores.

SMT4 and SMT8 Core Architecture

As Figure 6 shows, the Power9 processor offers different core variants to address different market requirements. The SMT4 core has two 128-bit super-slices, and the SMT8 core has four 128-bit super-slices.

The instruction fetch, decode, and issue capabilities of the SMT8 core are twice that of the SMT4 core. This enables the core to scale enough to support eight hardware threads and maintain good throughput for all of them. The SMT8 core can also power-gate half of the core execution resources when only one thread is active per core.

The SMT8 core provides a large shared resource pool to individual partitions, providing for efficient large partition management and enabling seamless partition mobility from Power8 processor servers. The SMT4 core provides increased resource management granularity to the hypervisor for increased granularity of core computational resources.

Data Capacity and Throughput

In this section, we discuss the Power9 cache and throughput.

Cache

The Power9 processor has a 512-Kbyte, private, eight-way set-associative L2 cache per SMT8 core. This cache also functions as a privately shared cache between two SMT4 cores. The Power9 processor has a total of 120 Mbytes of L3 NUCA cache.

As Figure 7 shows, the eDRAM-based L3 cache comprises 12 regions of 10 Mbytes each per SMT8 core. The L3 cache is 20-way associative, optimized for up to eight threads sharing the same cache region. Each of these L3 regions acts as fast local cache for each SMT8 core and is accessed by other cores in the processor using the internal fabric interconnect. The 10-Mbyte cache region acts as a shared cache between two SMT4 cores. This cache topology lets each L3 cache congruence class on the chip support up to 240 ways concurrently.

The cache capacity supports massively parallel computation and also enables a highly efficient heterogeneous interaction. The Power9 L3 cache implements new replacement policies utilizing reuse patterns and data-type awareness to improve its efficiency for data-intensive workloads. The Power9 processor also implements an

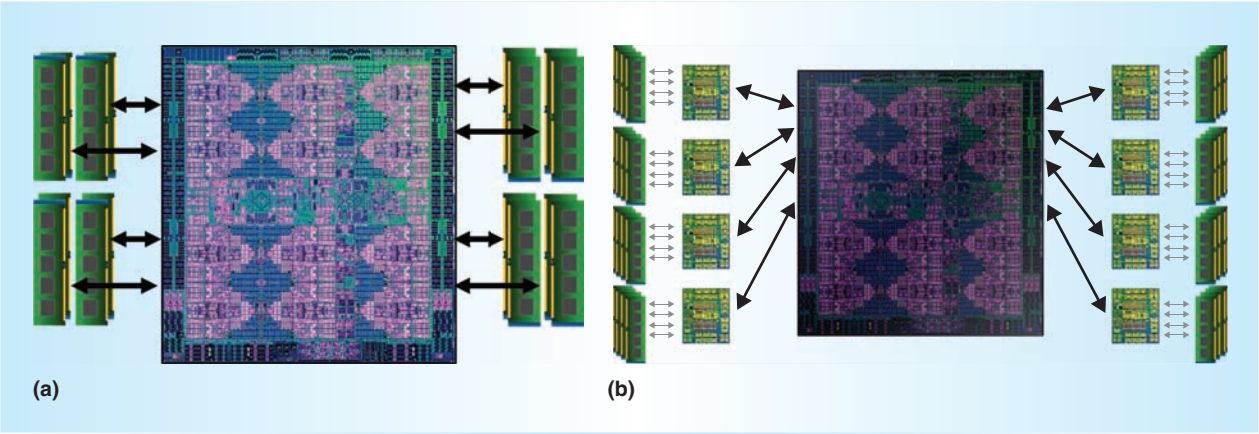

Figure 8. Two variants of the Power9 core's memory design architecture: direct attach memory and buffered memory. (a) Scale-out variant. (b) Scale-up variant.

adaptive prefetch mechanism with coordination between the processor cores, caches, and memory controllers. This mechanism optimizes prefetch aggressiveness in cases in which the consumption or utilization of prefetched data is low or the available memory bandwidth is limited.

The Power9 core and chip architecture enable large SMP scaling and heterogeneous computing using accelerators and attached devices. A key component of this design is a high-throughput on-chip interconnect fabric composed of separate command and data switching interconnects. The on-chip data switching infrastructure is built from a 2D topology of switch segments logically arranged in an 8 × 12 (96-element) structure. Each element can transfer 32 bytes at a 2.4-GHz clock rate, yielding 76.8 GBps per element. The 96 elements provide an aggregate 7.3 TBps for on-chip data switching. This enables each of the processor cores to move data in and out of the core at the rate of 256 GBps while simultaneously exchanging data with memory, attached devices, and accelerators.

Memory Subsystem

The Power9 system design supports both scale-out and scale-up domains. Although the Power9 core microarchitecture remains the same for the both domains, the memory design architecture is tailored to two variants to suit both of these domains.

The first variant, shown in Figure 8a, is targeted toward the scale-out domain. It has direct-attach double data rate type four (DDR4) memory. Each DDR4 unit is self-contained and consists of four independent ports that connect to DIMM slots. This unit is replicated twice on the Power9 processor to provide a maximum of eight ports, supporting up to 120 GBps of sustained memory bandwidth with a maximum memory capacity of up to 4 Tbytes per socket. These low-latency access ports support 64-byte or 128-byte adaptive reads from the memory.

The second variant, shown in Figure 8b, is targeted toward the scale-up domain. It has a buffered memory architecture. A socket supports eight such buffered memory channels, which can deliver a sustained bandwidth of 230 GBps, with a maximum memory capacity of up to 8 Tbytes per socket. This design variant also provides superior reliability, availability, and serviceability (RAS) capabilities with chip kill and lane sparing support. It is also compatible with Power8 system memory.[2]

Processor Family

The Power9 processor comes as a processor family that has four implementations to address different market segments. This is achieved by the Power9 processor's modular and scalable design. The execution slice microarchitecture forms the basic building block for all four targeted implementations.

The scale-out design comes in two variants. Both of these variants have similar SMP scaling support and a similar memory subsystem. The scale-out variant with the

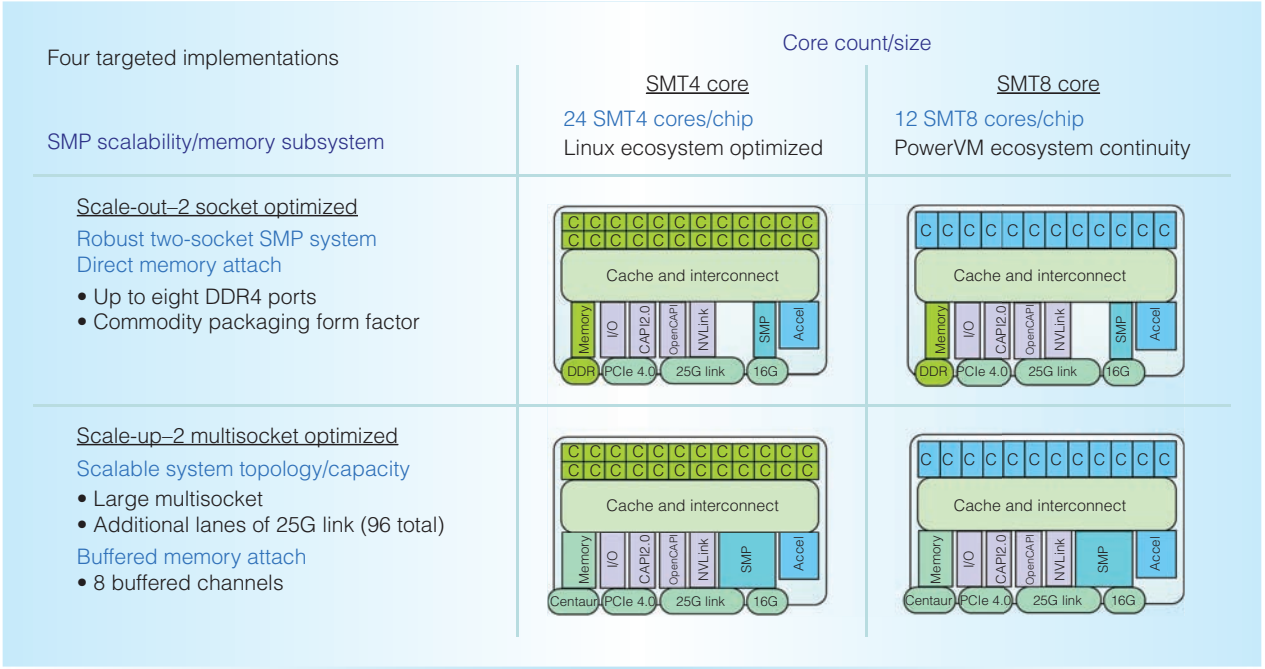

Figure 9. Power9 family of processors showing different targeted implementations for the two-socket and multisocket design.

direct-attach DDR4 memory supports one or two sockets. Two different implementations of the scale-out domain target two different ecosystems. The 24-SMT4 core implementation is targeted toward Linux ecosystem needs, and the 12-SMT8 core implementation is targeted toward a PowerVM-based ecosystem.

The scale-up design also comes in both SMT4 and SMT8 variants. These variants are optimized to support larger SMP connectivity through 96 lanes of 25G Link, along with eight buffered memory channels on each socket to support a large memory bandwidth.

Figure 9 shows all four targeted implementations of the Power9 family of processors.

Power ISA Support for Emerging Markets

The Power9 design philosophy addresses the needs of various emerging market domains. The Power instruction set architecture (ISA) version 3.0, which Power9 implements, also supports various emerging workload segments. Power9 has focused ISA support for cognitive computing, cloud, HPC, and enterprise solutions market domains. The key Power ISA 3.0 enhancements fall into the following categories:[3]

- *Broader data-type support.* The Power9 processor provides native 128-bit quadword-precision floating-point arithmetic that is IEEE compliant.
- *Support for emerging algorithms.* The Power9 processor implements a single-instruction random number generator that is certified by the National Institute of Standards and Technology. Atomic memory operations are supported for near-memory computation and include logical, arithmetic, max, min, and compare operations. Atomic operations are issued by a processor core thread but are executed at the memory controller, enabling optimization of high-scale writing in data-centric applications.
- *Cloud optimization.* To optimize for cloud environments, the Power9 processor has an interrupt architecture that automates interrupt routing to partitions to boost the performance of virtualization.
- *Enhanced accelerator virtualization.* Both on-chip and off-chip accelerators

can be addressed by user programs with virtual memory addresses. This reduces overhead and latency for communicating with the accelerators. On-chip accelerators have been expanded to include two 842 and one Gzip compression accelerator, as well as two AES/SHA accelerators.

- *Energy and frequency management.* Power9 power management supports the concept of workload-optimized frequency (WOF). Using this mechanism, the chip's performance can be pushed to the socket's thermal and current limits to maximize performance. The power-management firmware components can interlock to exploit powered-off cores to increase the frequency of the operating cores. When enabled, the WOF mechanism can also exploit cores that are running but drawing less power due to lower performance states (P-states) or lighter workloads.

These key ISA enhancements enable the Power9 processor to deliver better performance for emerging workloads. In the rest of this article, we will discuss the performance of the Power9 processor compared with its previous-generation Power8 processor.

Performance

Figure 10 shows the socket-level performance improvements over the Power8 processor across a broad range of workloads. The improvements are shown for the scale-out configuration with similar bandwidth constraints and at constant frequency to compare against the previous-generation Power systems. Emerging workloads such as scripting languages, graph analytics, and business intelligence show close to twice the improvement over the Power8 processor. Integer, floating-point arithmetic performance improved well above one and half times. Commercial workloads, which primarily cover the enterprise domain, gained between one and one-half to two times improvements compared to the Power8 processor.

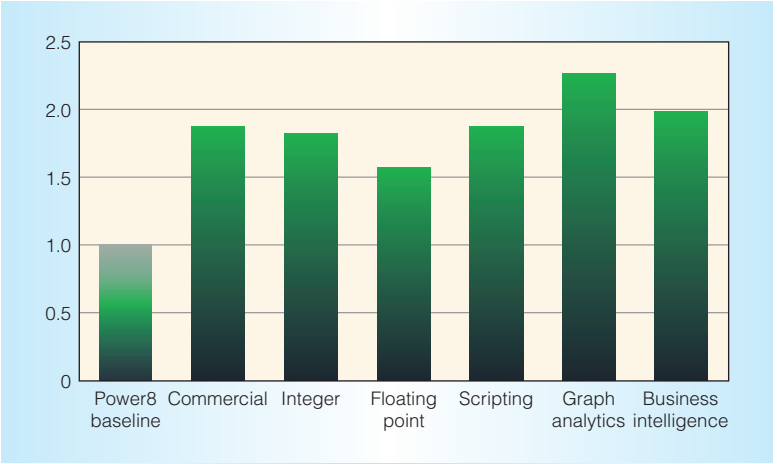

Figure 10. Power9 performance improvements. The figure shows improvements over the Power8 processor for a spectrum of workloads at constant frequency.

Heterogeneous Computing Architecture

A primary objective of the Power9 processor is to provide seamless integration of the processor with external accelerators and configurable devices to make it a truly heterogeneous computing platform. This is accomplished by two key aspects of the design.

First, the Power9 processor provides PCIe Gen4 with 48-lane connectivity to support PCIe devices.[4] This PCIe Gen4 connectivity provides a 192-GBps duplex bandwidth. Over these high-bandwidth interfaces, Power9 supports IBM Coherent Accelerator Processor Interface (CAPI)[5] 2.0 connectivity to attach ASIC and field-programmable gate array (FPGA) devices on any of the Power9-based socket configurations. Power9 CAPI 2.0 connectivity provides four times more bandwidth than its Power8 predecessor does. This enables the Power9 processor to seamlessly integrate with any external accelerating devices, including ASIC and FPGA devices with a coherent memory support with the host memory using CAPI 2.0 (see Figure 2).

The next key design aspect is the 25G Link, which delivers up to 300 GBps of bidirectional bandwidth to connected accelerators over 48 lanes. This link provides support for Nvidia NVLink 2.0 connectivity and supports OpenCAPI,[6] a low-latency and high-bandwidth coherently attached device interface with open standards. Nvidia's next-generation GPUs are designed to work with NVLink 2.0 to provide high performance along with seamless integration of the GPU

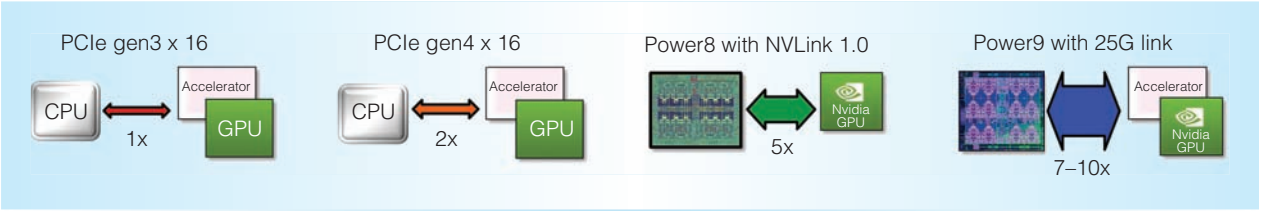

Figure 11. Power9 processor integration bandwidth with high-performing GPU or accelerators.

and the host CPU. The Power9 processor supports up to 48 lanes of 25G links for accelerator attach.

Figure 11 shows the performance capability of the Power9 processor to integrate high-performing GPUs or accelerators. Traditionally, GPUs are attached to the PCIe slots. In comparison with PCIe Gen3-based GPU attach, the PCIe Gen4-based GPU attach provides two times the bandwidth. With the support of NVLink 1.0 in Power8 with NVLink,[7] the capacity has grown to five times more bandwidth. The 25G Link on Power9 provides 7 to 10 times more bandwidth in comparison with the PCIe Gen3-based design, which attaches GPUs and other accelerators.

The Power9 processor supports seamless CPU-to-accelerator interactions with its ability to provide coherent memory sharing. This significantly reduces software and hardware overhead due to data interactions between the CPU and accelerators (including GPUs). Accelerator-attached devices are also supported by enhanced virtual address translation capabilities on the Power9 die, further reducing CPU and accelerator interaction latencies.

Both NVLink 2.0 and the OpenCAPI interface provide an efficient programming model with much less programming complexity to accelerate complex analytics and cognitive-based applications. The combination of seamless data sharing, low latency, and high-bandwidth communication between the CPU and an accelerator enables application of heterogeneous computing to a new class of applications.

The IBM Power9 processor architecture is designed to suit a wide range of platform optimizations with a family of processors designed for both scale-out and scale-up applications. Power9 introduces a new core microarchitecture and an enhanced cache and chip architecture to support high bandwidth, computational scale, and data capacity. Architectural innovations provide for enhanced virtualization capabilities targeting key market segments, including acceleration and the cloud. A state of the art I/O subsystem, engineered to be open, enables the Power9 processor to support a wide range of externally attached devices with high bandwidth, low latency, and tight coupling for the next generation of heterogeneous and accelerated computing applications. MICRO

References

1. B. Sinharoy et al., "IBM Power8 Processor Core Microarchitecture," *IBM J. Research and Development*, vol. 59, no. 1, 2015, pp. 2:1–2:21.
2. W.J. Starke et al., "The Cache and Memory Subsystems of the IBM Power8 Processor," *IBM J. Research and Development*, vol. 59, no. 1, 2015, pp. 3:1–3:13.
3. "IBM Power ISA Version 3.0—OpenPower," May 2016; http://openpowerfoundation.org/?resource_lib=power-isa-version-3-0.
4. "PCIe Gen4 Specification," 2016; http://pcisig.com.
5. "OpenCAPI," 2017; http://opencapi.org/technical.
6. J. Stuecheli et al., "CAPI: A Coherent Accelerator Processor Interface," *IBM J. Research and Development*, vol. 59, no. 1, 2015, pp. 7:1–7:7.
7. S. Gupta, "What Is NVLink? And How Will It Make the World's Fastest Computers

Possible?" blog, 14 Nov. 2014; http://blogs.nvidia.com/blog/2014/11/14/what-is-nvlink.

Satish Kumar Sadasivam is a senior engineer at IBM. His work involves next-generation Power microarchitecture concept design and performance evaluation. Sadasivam received an MS in computer science from the Madras Institute of Technology. He is an IBM master inventor. Contact him at satsadas@in.ibm.com.

Brian W. Thompto is a senior technical staff member for IBM's Power Systems processor team and a lead architect for Power9 and future Power processors. Thompto received a BS in electrical engineering and computer science from the University of Wisconsin. He is an IBM master inventor. Contact him at bthompto@us.ibm.com.

Ron Kalla is the chief engineer for IBM Power9. His work has included processors for IBM S/370, M68000, iSeries, and pSeries machines, as well as post-silicon hardware bring-up and verification. He is an IBM master inventor. Contact him at rkalla@us.ibm.com.

William J. Starke is an IBM Distinguished Engineer and chief architect for the Power processor storage hierarchy. 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. Starke received a BS in computer science from Michigan Technological University. Contact him at wstarke@us.ibm.com.