POWER9

Jeff Stuecheli

POWER Systems, IBM Systems
Recent and Future POWER Processor Roadmap

- **POWER7+**
  - 32 nm
  - 2012
- **POWER7**
  - 45 nm
  - 2010
- **POWER8 Family**
  - 22 nm
  - 2014 – 2016
- **POWER9 Family**
  - 14 nm
  - 2H17 – 2H18+
- **POWER10 Family**
  - 2020+

Power Systems Technology and Architecture
Leveraging the economics of the New Era
POWER9 Processor – Common Features

New Core Microarchitecture
- Stronger thread performance
- Efficient agile pipeline
- POWER ISA v3.0

Enhanced Cache Hierarchy
- 120MB NUCA L3 architecture
- 12 x 20-way associative regions
- Advanced replacement policies
- Fed by 7 TB/s on-chip bandwidth

Cloud + Virtualization Innovation
- Quality of service assists
- New interrupt architecture
- Energy Scale (Workload optimized frequency)

Leadership Hardware Acceleration Platform
- Enhanced on-chip acceleration
- Nvidia NVLink 2.0: High bandwidth and advanced new features (25G)
- CAPI 2.0: Coherent accelerator and storage attach (PCIe G4)
- OpenCAPI: Improved latency and bandwidth, open interface (25G)

State of the Art I/O Subsystem
- PCIe Gen4 – 48 lanes

High Bandwidth Signaling Technology
- 16 GT/s interface
  - Local SMP
- 25 GT/s Common Link interface
  - Accelerator, remote SMP

14nm finFET Semiconductor Process
- Improved device performance and reduced energy
- 17 layer metal stack and eDRAM
- 8.0 billion transistors
New POWER9 Cores

Optimized for Stronger Thread Performance and Efficiency

- Increased execution bandwidth efficiency for a range of workloads including commercial, cognitive and analytics
- Sophisticated instruction scheduling and branch prediction for unoptimized applications and interpretive languages
- Adaptive features for improved efficiency and performance especially in lower memory bandwidth systems

Available with SMT8 or SMT4 Cores

8 or 4 threaded core built from modular execution slices

POWER9 SMT8 Core
- PowerVM Ecosystem Continuity
- Strongest Thread
- Optimized for Large Partitions

POWER9 SMT4 Core
- Linux Ecosystem Focus
- Core Count / Socket
- Virtualization Granularity
POWER9 – Dual Memory Subsystems

Scale Out
Direct Attach Memory

- 8 Direct DDR4 Ports
  - Up to 120 GB/s of sustained bandwidth
  - Low latency access
  - Commodity packaging form factor
  - Adaptive 64B / 128B reads

Scale Up
Buffered Memory

- 8 Buffered Channels
  - Up to 230GB/s of sustained bandwidth
  - Extreme capacity – up to 8TB / socket
  - Superior RAS with chip kill and lane sparing
  - Compatible with POWER8 system memory
  - Agnostic interface for alternate memory innovations
## Four targeted implementations

### SMP scalability / Memory subsystem

<table>
<thead>
<tr>
<th>Scale-Out – 2 Socket Optimized</th>
<th>Scale-Up – Multi-Socket Optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust 2 socket SMP system</td>
<td>Scalable System Topology / Capacity</td>
</tr>
<tr>
<td>Direct Memory Attach</td>
<td>• Large multi-socket</td>
</tr>
<tr>
<td></td>
<td>Buffered Memory Attach</td>
</tr>
<tr>
<td></td>
<td>• 8 Buffered channels</td>
</tr>
</tbody>
</table>

### Core Count / Size

- **SMT4 Core**
  - 24 SMT4 Cores / Chip
  - Linux Ecosystem Optimized

- **SMT8 Core**
  - 12 SMT8 Cores / Chip
  - PowerVM Ecosystem Continuity
Modular Constructs $\rightarrow$ High-speed 25 GT/s Signaling

Utilize Best-of-Breed 25 GT/s Optical-Style Signaling Technology

Flexible & Modular Packaging Infrastructure

- Multi-Drawer SMP Interconnect
- NVLINK 2 GPU Accelerator Attach
- OpenCAPI Accelerator Attach
POWER9 – Premier Acceleration Platform

- Extreme Processor / Accelerator Bandwidth and Reduced Latency
- Coherent Memory and Virtual Addressing Capability for all Accelerators
- OpenPOWER Community Enablement – Robust Accelerated Compute Options

State of the Art I/O and Acceleration Attachment Signaling
- PCIe Gen 4 x 48 lanes – 192 GB/s duplex bandwidth
- 25Gb/s Common Link x 48 lanes – 300 GB/s duplex bandwidth

Robust Accelerated Compute Options with OPEN standards
- On-Chip Acceleration – Gzip x1, 842 Compression x2, AES/SHA x2
- CAPI 2.0 – 4x bandwidth of POWER8 using PCIe Gen 4
- NVLink 2.0 – Next generation of GPU/CPU bandwidth and integration using 25G (x48)
- Open CAPI 3.0 – High bandwidth, low latency and open interface using 25G (x32)
OpenCAPI – An Open heterogeneous architecture standard

1. Accelerators: The performance, virtual addressing and coherence capabilities allow FPAA and ASIC accelerators to be added as if they were integrated into a custom microprocessor.

2. Coherent Network Controller: OpenCAPI provides the bandwidth that will be needed to support rapidly increasing network speeds. Network controllers based on virtual addressing can eliminate software overhead without the programming complexity usually associated with custom networking protocols.

3. Advanced Memory: OpenCAPI allows system designers to take full advantage of emerging memory technologies to reduce the economics of the datacenter.

4. Coherent Storage Controller: OpenCAPI allows storage controllers to bypass kernel software overhead, enabling extremely IOPS performance without wasting valuable CPU cycles.

OpenCAPI specifications are downloadable from the website at www.opencapi.org
- Register
- Download
# Proposed POWER Processor Technology and I/O Roadmap

**POWER7 Architecture**
- 2010 POWER7
  - 8 cores
  - 45nm
  - New Micro-Architecture
  - New Process Technology

**POWER8 Architecture**
- 2012 POWER7+
  - 8 cores
  - 32nm
  - Enhanced Micro-Architecture
  - New Process Technology

**POWER9 Architecture**
- 2014 POWER8
  - 12 cores
  - 22nm
  - New Micro-Architecture

**POWER8 Architecture w/ NVLink**
- 2016 POWER8
  - 12 cores
  - 22nm
  - Enhanced Micro-Architecture
  - NVLink

**POWER9 Architecture w/ Adv. I/O**
- 2017 P9 SO
  - 12/24 cores
  - 14nm
  - New Micro-Architecture
  - Direct attach memory
  - Buffered Memory

**POWER10**
- 2018 P9 SU
  - 12/24 cores
  - 14nm
  - Enhanced Micro-Architecture
  - Buffered Memory
  - New Memory Subsystem

## Sustained Memory Bandwidth
- **POWER7 Architecture**
  - Up To 65 GB/s
- **POWER8 Architecture**
  - Up To 210 GB/s
- **POWER9 Architecture**
  - Up To 210 GB/s

## Standard I/O Interconnect
- **POWER7 Architecture**
  - PCIe Gen2
- **POWER8 Architecture**
  - PCIe Gen3
- **POWER9 Architecture**
  - PCIe Gen4 x48

## Advanced I/O Signaling
- **POWER7 Architecture**
  - N/A
- **POWER8 Architecture**
  - N/A
- **POWER9 Architecture**
  - 20 GT/s 160GB/s

## Advanced I/O Architecture
- **POWER7 Architecture**
  - N/A
- **POWER8 Architecture**
  - CAPI 1.0, NVLink 1.0
- **POWER9 Architecture**
  - CAPI 2.0, OpenCAPI3.0, NVLink2.0

## Advanced I/O Architecture
- **POWER7 Architecture**
  - N/A
- **POWER8 Architecture**
  - N/A
- **POWER9 Architecture**
  - CAPI 2.0, OpenCAPI4.0, NVLink3.0

Statement of Direction, Subject to Change
Future Evolution of System Architecture

OpenCAPI Northbound

Yesterday’s Plumbing
Tomorrow’s Differentiation

Cores & Caches
768 GB/s System Bus

OpenCAPI & PCI

Yesterday’s Plumbing
Tomorrow’s Differentiation

PCI Gen X

OpenCAPI / NVLink

CPU/ Accelerator Bandwidth

CPU

Accelerator

GPU

1x

2x

7-10x

Accelerator

GPU

NVIDIA GPU

5x

System bottleneck
CORAL – IBM Delivers Summit and Sierra

- Deployments beginning with full acceptance in 2018
- Significant application performance over Titan (AMD/NVIDIA)
  - Achieved with ¼ the servers

3+EFLOPS
Tensor Ops

10X
Perf Over Titan

5-10X
Application Perf Over Titan
CORAL ORNL (Summit) 200PF System

POWER9:
- 22 Cores
- 4 Threads/core
- 0.54 DP TF/s
- 3.07 GHz

Volta:
- 7.0 DP TF/s
- 16GB @ 0.9 TB/s

POWER9 2 Socket Server
- 2 P9 + 6 Volta GPU
- 512 GiB SMP Memory (32 GiB DDR4 RDIMMs)
- 96 GiB GPU Memory (HBM stacks)
- 1.6 TB NVMe

System

Compute Rack:
- 18 nodes
- 775 TF/s
- 10.7 TiB
- 59 KW max

ESS Building Block

SSC (4 ESS GL4):
- 8 servers, 16 JBOD
- 16.8 PB (gross)
- 38 KW max

Mellanox IB4X EDR Switch IB-2

Mellanox IB4X EDR
- 648p Directors
- Full bisection

*The Spectrum Scale Installation at ORNL will be the world’s largest HPC File System at 250PB in 1H 2018
CORAL LLNL (Sierra) 125PF System

**POWER9:**
- 22 Cores
- 4 Threads/core
- 0.54 DP TF/s
- 3.07 GHz

**Volta:**
- 7.0 DP TF/s
- 16GB @ 0.9 TB/s

**POWER9 2 Socket Server**
- 2 P9 + 4 Volta GPU
- 256 GiB SMP Memory (16 GiB DDR4 RDIMMs)
- 64 GiB GPU Memory (HBM stacks)
- 1.6 TB NVMe

**Standard 2U 19in. Rack mount Chassis**

**System**
- Compute Rack: 18 nodes
  - 523 TF/s
  - 5.6 TiB
  - 45 KW max

- ESS Building Block
  - SSC (4 ESS GL4):
    - 8 servers, 16 JBOD
    - 16.8 PB (gross)
    - 38 KW max

- Floor plan rack concept
  - Compute (240)
  - Switch (9)
  - Storage (24)
  - Infrastructure (4)
Enhanced Core and Chip Architecture for Emerging Workloads
• New Core Optimized for Emerging Algorithms to Interpret and Reason
• Bandwidth, Scale, and Capacity, to Ingest and Analyze

Processor Family with Scale-Out and Scale-Up Optimized Silicon
• Enabling a Range of Platform Optimizations – from HSDC Clusters to Enterprise Class Systems
• Extreme Virtualization Capabilities for the Cloud

Premier Acceleration Platform
• Heterogeneous Compute Options to Enable New Application Paradigms
• State of the Art I/O
• Engineered to be Open
Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquiries, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.

Revised September 26, 2006
Special notices (continued)

IBM, the IBM logo, ibm.com AIX, AIX (logo), IBM Watson, DB2 Universal Database, POWER, PowerLinux, PowerVM, PowerVM (logo), PowerHA, Power Architecture, Power Family, POWER Hypervisor, Power Systems, Power Systems (logo), POWER2, POWER3, POWER4, POWER4+, POWER5, POWER5+, POWER6, POWER6+, POWER7, POWER7+, and POWER8 are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries.

A full list of U.S. trademarks owned by IBM may be found at: http://www.ibm.com/legal/copytrade.shtml.

NVIDIA, the NVIDIA logo, and NVLink are trademarks or registered trademarks of NVIDIA Corporation in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries or both.

PowerLinux™ uses the registered trademark Linux® pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the Linux® mark on a world-wide basis.

The Power Architecture and Power.org wordmarks and the Power and Power.org logos and related marks are trademarks and service marks licensed by Power.org.

The OpenPOWER word mark and the OpenPOWER Logo mark, and related marks, are trademarks and service marks licensed by OpenPOWER.

Other company, product and service names may be trademarks or service marks of others.