IBM’s POWER10 Processor
2020 OpenPOWER Academia Discussion Group Workshop

Brian Thompto
IBM Distinguished Engineer
IBM POWER Processor Technology Roadmap

POWER7/7+ 45/32 nm
- Multi-core Optimized
- Up to 8 cores/die (32 HW threads)
- eDRAM L3 Cache

POWER8 Family 22nm
- Up to 12 cores/die (96 HW threads)
- Agnostic Memory
- Enterprise Focus
- Big Data Optimized
- PCIe G3 / CAPI / NVLINK
- OpenPOWER

POWER9 Family 14nm
- Up to 24/12 cores/die (96 HW threads)
- Modular new Core uArch
- Direct-Attach Memory
- OMI Memory
- PowerAXON Modular Attach
- PCIe G4 / CAPI 2.0
- Coherent NVLINK / OpenCAPI
  - #1, #2 Supercomputers

POWER10 Family 7nm
- Up to 60/30 cores/socket (240 HW threads)
- Modular Building Block Die
- New Core uArch
- AI-optimized ISA
- Energy Efficiency Focus
- HW Enforced Security
- Enterprise Focus
- PowerAXON 2.0
- PCIe G5
- Memory Clustering

POWER11 Family

Under development...
IBM POWER Processor Technology Roadmap: Today’s Discussion

POWER10 Family
7nm

- Up to 60/30 cores/socket
- (240 HW threads)
- Modular Building Block Die
- New Core uArch
- AI-optimized ISA
- Energy Efficiency Focus
- HW Enforced Security
- Enterprise Focus
- PowerAXON 2.0
- PCIe G5
- Memory Clustering
POWER10 Design Focus

Data Plane Bandwidth, Capacity, Composability, Scale
   Terabyte/second sockets, Petabyte system memory capacities, 16-socket SMP \rightarrow Clusters

Powerful Enterprise Core
   New Core Architecture, Flexibility, Larger caches, Reduced Latencies

End-to-end Security
   Hardware enabled and co-optimized with PowerVM hypervisor

Energy Efficiency
   3x improvement over POWER9

AI-Infused Core
   10-20x matrix-math performance / socket compared to POWER9
POWER10 Processor Chip

**Technology and Packaging:**
- 602mm² 7nm Samsung (18B devices)
- 18 layer metal stack, enhanced device
- Single-chip or Dual-chip sockets

**Computational Capabilities:**
- Up to 15 SMT8 Cores (2 MB L2 Cache / core)
  (Up to 120 simultaneous hardware threads)
- Up to 120 MB L3 cache (low latency NUCA mgmt)
- 3x energy efficiency relative to POWER9
- Enterprise thread strength optimizations
- AI and security focused ISA additions
- 2x general, 4x matrix SIMD relative to POWER9
- EA-tagged L1 cache, 4x MMU relative to POWER9

**Open Memory Interface:**
- 16 x8 at up to 32 GT/s (1 TB/s)
- Technology agnostic support: near/main/storage tiers
- Minimal (< 10ns latency) add vs DDR direct attach

**PowerAXON Interface:**
- 16 x8 at up to 32 GT/s (1 TB/s)
- SMP interconnect for up to 16 sockets
- OpenCAPI attach for memory, accelerators, I/O
- Integrated clustering (memory semantics)

**PCle Gen 5 Interface:**
- x64 / DCM at up to 32 GT/s

Die Photo courtesy of Samsung Foundry
Socket Composability: **SCM & DCM**

**Single-Chip Module Focus:**
- 602mm\(^2\) 7nm (18B devices)
- Core/thread Strength
  - Up to 15 SMT8 Cores (4+ GHz)
- Capacity & Bandwidth / Compute
  - Memory: x128 @ 32 GT/s
  - SMP/Cluster/Accel: x128 @ 32 GT/s
  - I/O: x32 PCIe G5
- System Scale (Broad Range)
  - 1 to 16 sockets

**Dual-Chip Module Focus:**
- 1204mm\(^2\) 7nm (36B devices)
- Throughput / Socket
  - Up to 30 SMT8 Cores (3.5+ GHz)
- Compute & I/O Density
  - Memory: x128 @ 32 GT/s
  - SMP/Cluster/Accel: x192 @ 32 GT/s
  - I/O: x64 PCIe G5
  - 1 to 4 sockets

(Multi-socket configurations show processor capability only, and do not imply system product offerings)
System Composability: **PowerAXON & Open Memory Interfaces**

Multi-protocol
“Swiss-army-knife”
Flexible / Modular Interfaces

- **PowerAXON corner**
  - 4x8 @ 32 GT/s

- **PowerAXON**
  - 1 Terabyte / Sec

- **POWER10 Chip**

- **OMI Memory**
  - 1 Terabyte / Sec

- **OMI edge**
  - 8x8 @ 32 GT/s
  - 6x bandwidth / mm² compared to DDR4 signaling

**Built on best-of-breed**
- Low Power, Low Latency,
- High Bandwidth Signaling Technology

IBM POWER10
System Enterprise Scale and Bandwidth: SMP & Main Memory

Multi-protocol
“Swiss-army-knife”
Flexible / Modular Interfaces

Allocate the bandwidth however you need to use it

Build up to 16 SCM socket
Robustly Scalable
High Bisection Bandwidth
“Glueless” SMP

1 Terabyte / Sec
POWER10 Chip
OMI Memory
PowerAXON

Built on best-of-breed
Low Power, Low Latency,
High Bandwidth
Signaling Technology

Main tier DRAM

SMP Interconnect

Initial Offering:
Up to 4 TB / socket
OMI DRAM memory
410 GB/s peak bandwidth
(MicroChip DDR4 buffer)
< 10ns latency adder

DIMM swap upgradeable:
DDR5 OMI DRAM memory
with higher bandwidth
and higher capacity

(PowerAXON and OMI Memory configurations show processor capability only, and do not imply system product offerings)
Data Plane Bandwidth and Capacity: **Open Memory Interfaces**

OMI-attached GDDR DIMMs can provide low-capacity, high bandwidth alternative to HBM, without packaging rigidity & cost (Up to 800 GB/s sustained)

OMI-attached storage class memory can provide high-capacity, encrypted, persistent memory in a DIMM slot. (POWER10 systems can support 2 petabytes of addressable load/store memory)

*PowerAXON and OMI Memory configurations show processor capability only, and do not imply system product offerings*
System Heterogeneity and Data Plane Capacity: OpenCAPI

OpenCAPI attaches FPGA or ASIC-based Accelerators to POWER10 host with High Bandwidth and Low Latency

OpenCAPI-attached storage class memory can provide high-capacity, encrypted, persistent memory in a device form factor. (POWER10 systems can support 2 petabytes of addressable load/store memory)

(PowerAXON and OMI Memory configurations show processor capability only, and do not imply system product offerings)
Pod Composability: **PowerAXON Memory Clustering**

Memory Inception capability enables a system to map another system’s memory as its own. Multiple systems can be clustered, sharing each other’s memory.

(PowerAXON and OMI Memory configurations show processor capability only, and do not imply system product offerings)
Memory Clustering: Distributed Memory Disaggregation and Sharing

Use case: Share load/store memory amongst directly connected neighbors within Pod. Unlike other schemes, memory can be used:
- As low latency local memory
- As NUMA latency remote memory

Example: Pod = 8 systems each with 8TB
Workload A Rqmt: 4 TB low latency
Workload B Rqmt: 24 TB relaxed latency
Workload C Rqmt: 8 TB low latency plus 16TB relaxed latency

All Rqmts met by configuration shown

POWER10 2 Petabyte memory size enables much larger configurations

(Memory cluster configurations show processor capability only, and do not imply system product offerings)
Memory Clustering: Enterprise-Scale Memory Sharing

Pod of Large Enterprise Systems
Distributed Sharing at Petabyte Scale

Or Hub-and-spoke with memory server
and memory-less compute nodes

(Memory cluster configurations show processor capability only, and do not imply system product offerings)
Memory Clustering: **Pod-level Clustering**

Use case: Low latency, high bandwidth messaging scaling to 1000’s of nodes

Leverage 2 Petabyte addressability to create memory window into each destination for messaging mailboxes

(Memory cluster configurations show processor capability only, and do not imply system product offerings)
System Composability: PCIe Gen 5 Industry I/O Attach

(PowerAXON and OMI Memory configurations show processor capability only, and do not imply system product offerings)
POWER10 General Purpose Socket Performance Gains

(Performance assessments based upon pre-silicon engineering analysis of POWER10 dual-socket server offering vs POWER9 dual-socket server offering)
Powerful Core = Enterprise Strength + AI Infused

New Enterprise Micro-architecture
- Flexibility
  - Up to 8 threads per core / 240 per socket
- Optimized for performance and efficiency
  - +30% avg. core performance*
  - +20% avg. single thread performance*
  - 2.6x core performance efficiency* (3x @ socket)

AI Infused
- 4x matrix SIMD acceleration*
- 2x bandwidth & general SIMD*
- 4x L2 cache capacity with improved thread isolation*
- New ISA with AI data-types

* versus POWER9

1-2 POWER10 chips per socket

- Up to 30 SMT8 Cores
- Up to 60 SMT4 Cores

(Performance assessments based upon pre-silicon engineering analysis of POWER10 dual-socket server offering vs POWER9 dual-socket server offering)
Powerful Core: Enterprise Flexibility

Multiple World-class Software Stacks

Resilience and full stack integrity
- PowerVM, KVM
- AIX, IBMi, Linux on Power, OpenShift

Partition flexibility and security
- Full-core level LPAR
- Thread-based LPAR scheduling
- NEW: With PowerVM Hypervisor
  - Nested KVM + PowerVM
  - Hardware assisted container/VM isolation

Hardware Based Workload Balance
### Powerful Architecture: AI Infused and Future Ready

**POWER10 implements Power ISA v3.1**

- v3.1 was the latest open Power ISA contributed to the OpenPOWER Foundation: Royalty free and inclusive of patents for compliant designs

<table>
<thead>
<tr>
<th>POWER10 Architecture – Feature Highlights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prefix Architecture</strong></td>
</tr>
<tr>
<td><strong>New Instructions and Datatypes</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Advanced System Features and Ease of Use</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Advanced EnergyScale</strong></td>
</tr>
<tr>
<td><strong>Security for Cloud</strong></td>
</tr>
</tbody>
</table>
Security: End-to-End for the Enterprise Cloud

Cloud Workload Security

Secure Virtualization

Processor Security Foundations

Application Security

VM's
Applications
Services
Middleware

Confidential Computing

Secure Containers:
- Transparent to applications
- End-to-end encryption

Nested Virtualization - KVM on PowerVM:
- Stronger container isolation without performance penalty
- HW enabled and transparent

Crypto Performance:
- Core crypto improvements for today’s algorithms (AES, SHA3)
  and ready for the future (PQC, FHE)

Dynamic Execution Control Register (DEXCR)

Main memory encryption:
Stronger confidentiality against physical attacks

Performance enhanced side channel avoidance

Hardened container memory isolation

IBM POWER10
Powerful Core: Enterprise Strength

P10 Core Micro-architecture
½ SMT8 Core Resources Shown = SMT4 Core Equivalent

Fetch / Branch Predictors

L1 Instr. Cache
48k 6-way
<EA Tagged>

32B

Predecode
+Fusion/Prefix

Instruction Buffer
128 entries

Decode/Fuse 8 iop

Instruction Table
512 entries

Execution Slice
128b

Execution Slice
128b

Execution Slice
128b

Execution Slice
128b

MMA Accelerator
2x512b

Load Queue
128 entries (SMT)
64 entries (ST)

Load Miss Queue
12 entries

Prefetch
16 streams

L1 Data Cache
32k 8-way
<EA Tagged>

Store Queue
80 entries (SMT)
40 entries (ST)

L2 Cache
(hashed index)

L2 Cache
(dedicated)

L2 Cache
(2-way)

2 Load EA

2 Store EA

32B LD

32B LD

D miss

I miss

ERAT
64 entry

TLB
4k entry

TLB
64 entry

TLB miss

L3 Prefetch
48 entries

Capacity vs. POWER9: Improved

>= 2x

= 4x

IBM POWER10
**Powerful Core: Enterprise Strength**

- Double SIMD + Inference acceleration
  - 2x SIMD, 4x MMA, 4x AES/SHA

---

**P10 Core Micro-architecture**

1/2 SMT8 Core Resources Shown = SMT4 Core Equivalent

- Fetch/Branch Predictors
- Instruction Table 512 entries
- Instruction Buffer 128 entries
- Decode/Fuse 8 iop
- MMA Accelerator 2x512b
- Execution Slice 128b
- Execution Slice 128b
- Execution Slice 128b
- Execution Slice 128b
- L1 Instr. Cache 48k 6-way <EA Tagged>
- 8 instr
- L1 Data Cache 32k 8-way <EA Tagged>
- L2 Cache (hashed index)
- L3 Prefetch 48 entries
- Load Queue 128 entries (SMT) 64 entries (ST)
- Load Miss Queue 12 entries
- Store Queue 80 entries (SMT) 40 entries (ST)
- Prefetch 16 streams
- D miss
- TLB 4k entry
- TLB miss
- L3 prefetch
- ERAT 64 entry
- miss
- Store Queue (gathered)
- TLB 64B dedicated
- 32B ST
- 32B LD
- 32B LD
- 32B LD
- 2 Load EA
- 2 Store EA
- I miss
- Capacity vs. POWER9: Improved
  - >= 2x
  - = 4x

---

IBM POWER10
Powerful Core: Enterprise Strength

- Double SIMD + Inference acceleration
  - 2x SIMD, 4x MMA, 4x AES/SHA
- Larger working-sets
  - 1.5x L1-Instruction cache, 4x L2, 4x TLB

IBM POWER10
Powerful Core: Enterprise Strength

- Double SIMD + Inference acceleration
  - 2x SIMD, 4x MMA, 4x AES/SHA
- Larger working-sets
  - 1.5x L1-Instruction cache, 4x L2, 4x TLB
- Deeper/wider instruction windows
**Powerful Core: Enterprise Strength**

- **Double SIMD + Inference acceleration**
  - 2x SIMD, 4x MMA, 4x AES/SHA
- **Larger working-sets**
  - 1.5x L1-instruction cache, 4x L2, 4x TLB
- **Deeper/wider instruction windows**
- **Data latency (cycles)**
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4 +0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
**Powerful Core: Enterprise Strength**

- **Double SIMD + Inference acceleration**
  - 2x SIMD, 4x MMA, 4x AES/SHA
- **Larger working-sets**
  - 1.5x L1-instruction cache, 4x L2, 4x TLB
- **Deeper/wider instruction windows**
- **Data latency (cycles)**
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4 +0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
- **Branch**
  - Target registers with GPR in main regfile
  - New predictors: target and direction, 2x BHT

---

**P10 Core Micro-architecture**

1/8 SMT8 Core Resources Shown = SMT4 Core Equivalent
Powerful Core: Enterprise Strength

- Double SIMD + Inference acceleration
  - 2x SIMD, 4x MMA, 4x AES/SHA
- Larger working-sets
  - 1.5x L1-Instruction cache, 4x L2, 4x TLB
- Deeper/wider instruction windows
- Data latency (cycles)
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4 +0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
- Branch
  - Target registers with GPR in main regfile
  - New predictors: target and direction, 2x BHT
- Fusion
  - Fixed, SIMD, other: merge and back to back
  - Load, store: consecutive storage
Powerful Core: Energy Efficient

- **Double SIMD + Inference acceleration**
  - 2x SIMD, 4x MMA, 4x AES/SHA
- **Larger working-sets**
  - 1.5x L1-Instruction cache, 4x L2, 4x TLB
- **Deeper/wider instruction windows**
- **Data latency (cycles)**
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4 +0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
- **Branch**
  - Target registers with GPR in main regfile
  - New predictors: target and direction, 2x BHT
- **Fusion**
  - Fixed, SIMD, other: merge and back to back
  - Load, store: consecutive storage
- **Improved clock-gating**
- **Design & micro-arch efficiency**
Powerful Core: Energy Efficient

- Double SIMD + Inference acceleration
  - 2x SIMD, 4x MMA, 4x AES/SHA
- Larger working-sets
  - 1.5x L1-instruction cache, 4x L2, 4x TLB
- Deeper/wider instruction windows
- Data latency (cycles)
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4+0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
- Branch
  - Target registers with GPR in main regfile
  - New predictors: target and direction, 2x BHT
- Fusion
  - Fixed, SIMD, other: merge and back to back
  - Load, store: consecutive storage

Improved clock-gating
Design & micro-arch efficiency
Branch accuracy: less wasted work
Fusion / gather: less units of work
Reduced ports / access
  - Sliced target reg-file
  - Reduced read ports / entry

IBM POWER10
Powerful Core: Energy Efficient

- **Double SIMD + Inference acceleration**
  - 2x SIMD, 4x MMA, 4x AES/SHA
- **Larger working-sets**
  - 1.5x L1-Instruction cache, 4x L2, 4x TLB
- **Deeper/wider instruction windows**
- **Data latency (cycles)**
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4 +0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
- **Branch**
  - Target registers with GPR in main regfile
  - New predictors: target and direction, 2x BHT
- **Fusion**
  - Fixed, SIMD, other: merge and back to back
  - Load, store: consecutive storage

- **Improved clock-gating**
- **Design & micro-arch efficiency**
- **Branch accuracy: less wasted work**
- **Fusion / gather: less units of work**
- **Reduced ports / access**
  - Sliced target reg-file
  - Reduced read ports / entry
- **EA-tagged L1-D Cache & L1-I Cache**
  - CAM with cache-way/index
  - ERAT only on cache miss

---

IBM POWER10
**Powerful Core: Strength & Efficiency**

- **Double SIMD + Inference acceleration**
  - 2x SIMD, 4x MMA, 4x AES/SHA
- **Larger working-sets**
  - 1.5x L1-Instruction cache, 4x L2, 4x TLB
- **Deeper/wider instruction windows**
- **Data latency (cycles)**
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4 +0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
- **Branch**
  - Target registers with GPR in main regfile
  - New predictors: target and direction, 2x BHT
- **Fusion**
  - Fixed, SIMD, other: merge and back to back
  - Load, store: consecutive storage

---

**P10 Core Micro-architecture**

½ SMT8 Core Resources Shown = SMT4 Core Equivalent

- **Fetch / Branch Predictors**
  - I-EA
  - L1 Instr. Cache 48k 6-way <EA Tagged>
  - 8 instr
  - Decode/Fuse 8 iop
  - Instruction Buffer 128 entries

- **Instruction Table**
  - 512 entries

- **MMA Accelerator**
  - 2x512b

- **Execution Slice**
  - 128b

- **Load Queue**
  - 128 entries (SMT)
  - 64 entries (ST)

- **Load Miss Queue**
  - 12 entries

- **Prefetch**
  - 16 streams

- **L1 Data Cache**
  - 32k 8-way <EA Tagged>

- **Store Queue**
  - 80 entries (SMT)
  - 40 entries (ST)

- **ERAT**
  - 64 entry

- **TLB**
  - 4k entry

- **L2 Cache**
  - (hashed index)
  - 64B dedicated
  - 32B ST (+gathered)

- **L3 Prefetch**
  - 48 entries

---

**Capacity vs. POWER9:**

- Improved
- >= 2x
- = 4x

---

**IBM POWER10**
Powerful Core: Strength & Efficiency

- Double SIMD + Inference acceleration
  - 2x SIMD, 4x MMA, 4x AES/SHA
- Larger working-sets
  - 1.5x L1-Instruction cache, 4x L2, 4x TLB
- Deeper/wider instruction windows
- Data latency (cycles)
  - L2 13.5 (minus 2), L3 27.5 (minus 8)
  - L1-D cache 4 +0 for Store forward (minus 2)
  - TLB access +8.5 (minus 7)
- Branch
  - Target registers with GPR in main regfile
  - New predictors: target and direction, 2x BHT
- Fusion
  - Fixed, SIMD, other: merge and back to back
  - Load, store: consecutive storage

1.3x

0.5x

= 2.6x performance / watt

POWER10 vs. POWER9 Core

IBM POWER10
Powerful Core: AI Infused Bandwidth and Compute

2x Bytes from all sources (OMI, L3, L2, L1 caches*)

* versus POWER9

IBM POWER10
Powerful Core: AI Infused Bandwidth and Compute

2x Bytes from all sources (OMI, L3, L2, L1 caches*)

- 4 32B loads, 2 32B stores per SMT8 Core
  - New ISA or fusion
  - Thread max 2 32B loads, 1 32B store

* versus POWER9
Powerful Core: AI Infused Bandwidth and Compute

2x Bytes from all sources
(OMI, L3, L2, L1 caches*)

- 4 32B loads, 2 32B stores per SMT8 Core
  - New ISA or fusion
  - Thread max 2 32B loads, 1 32B store

- OMI Memory to one Core
  - 256 GB/s peak, 120 GB/s sustained
  - With 3x L3 prefetch and memory prefetch extensions

* versus POWER9
Powerful Core: AI Infused Bandwidth and Compute

2x Bytes from all sources (OMI, L3, L2, L1 caches*)

- 4 32B loads, 2 32B stores per SMT8 Core
  - New ISA or fusion
  - Thread max 2 32B loads, 1 32B store
- OMI Memory to one Core
  - 256 GB/s peak, 120 GB/s sustained
  - With 3x L3 prefetch and memory prefetch extensions

2x Bandwidth matched SIMD*

- 8 independent SIMD engines per Core
  - Fixed, float, permute

* versus POWER9
2x Bytes from all sources
(OMI, L3, L2, L1 caches*)

- 4 32B loads, 2 32B stores per SMT8 Core
  - New ISA or fusion
  - Thread max 2 32B loads, 1 32B store

- OMI Memory to one Core
  - 256 GB/s peak, 120 GB/s sustained
  - With 3x L3 prefetch and memory prefetch extensions

2x Bandwidth matched SIMD*

- 8 independent SIMD engines per Core
  - Fixed, float, permute

4-32x Matrix Math Acceleration*

- 4 512b engines per core = 2048b results / cycle
  - Matrix math outer products: $A \leftarrow \{\pm\} A \{\pm\} XY^T$
  - Double, Single, Reduced precision

* versus POWER9
2x Bytes from all sources
(OMI, L3, L2, L1 caches*)

- 4 32B loads, 2 32B stores per SMT8 Core
  - New ISA or fusion
  - Thread max 2 32B loads, 1 32B store

- OMI Memory to one Core
  - 256 GB/s peak, 120 GB/s sustained
  - With 3x L3 prefetch and memory prefetch extensions

2x Bandwidth matched SIMD*

- 8 independent SIMD engines per Core
  - Fixed, float, permute

4-32x Matrix Math Acceleration*

- 4 512b engines per core = 2048b results / cycle
  - Matrix math outer products: $A \leftarrow (\pm)A (\pm)XY^T$
  - Double, Single, Reduced precision

* versus POWER9
AI Infused Core: **Inference Acceleration**

- **4x+ per core throughput**
- **3x → 6x thread latency reduction (SP, int8)**

**POWER10 Matrix Math Assist (MMA) instructions**
- 8 512b architected Accumulator (ACC) Registers
- 4 parallel units per SMT8 core

**Consistent VSR 128b register architecture**
- Minimal SW ecosystem disruption – no new register state
- Application performance via updated library (OpenBLAS, etc.)
- POWER10 aliases 512b ACC to 4 128b VSR’s
  - Architecture allows redefinition of ACC

**Dense-Math-Engine microarchitecture**
- Built for data re-use algorithms
- Includes separate physical register file (ACC)
- 2x efficiency vs. traditional SIMD for MMA

---

* versus POWER9

---

**Matrix Optimized / High Efficiency**

Result data remains local to compute

**Inference Accelerator dataflow** (2 per SMT8 core)
POWER10 SIMD / AI Socket Performance Gains

(Performance assessments based upon pre-silicon engineering analysis of POWER10 dual-socket server offering vs POWER9 dual-socket server offering)
Thank You from the POWER10 Development Team!

POWER10: Triple-Optimized for HPC

- 10x Linpack / socket vs POWER9
- Up to 800 GB/s memory bandwidth / socket
- Integrated Clustering (1000's of Nodes)

OMI-attached GDDR DIMMs can provide low-capacity, high bandwidth alternative to HBM, without packaging rigidity & cost (Up to 800 GB/s sustained)

1 Terabyte / Sec
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, IBM (logo), AIX, AIX (logo), EnergyScale, IBM i, i for business (logo), Power, POWER, PowerVM, PowerVM (logo), PowerLinux, PowerLinux (logo), Power Architecture, Power ISA, POWER9, and POWER10 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.

Red Hat, OpenShift, and the OpenShift logo are registered trademarks of Red Hat, Inc. in the United States and other countries.
The OpenPOWER word mark and the OpenPOWER logo mark, and related marks, are trademarks and service marks licensed by OpenPOWER Foundation.
OpenCAPI and the OpenCAPI logo are trademarks of the OpenCAPI Consortium.
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.

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