POWER7: IBM's Next Generation Server Processor

Ronald Kalla
POWER7 Chief Engineer

Balaram Sinharoy
POWER7 Chief Core Architect

Acknowledgment: This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002
Outline

- POWER Processor History
- POWER7™ Motivation
- Design Overview
- Summary
20+ Years of POWER Processors

Major POWER® Innovation
- 1990 RISC Architecture
- 1994 SMP
- 1995 Out of Order Execution
- 1996 64 Bit Enterprise Architecture
- 1997 Hardware Multi-Threading
- 2001 Dual Core Processors
- 2001 Large System Scaling
- 2001 Shared Caches
- 2003 On Chip Memory Control
- 2003 SMT
- 2006 Ultra High Frequency
- 2006 Dual Scope Coherence Mgmt
- 2006 Decimal Float/VSX
- 2006 Processor Recovery/Sparing
- 2009 Balanced Multi-core Processor
- 2009 On Chip EDRAM

* Dates represent approximate processor power-on dates, not system availability
POWER7 Processor Chip

- 567mm² Technology: 45nm lithography, Cu, SOI, eDRAM
- 1.2B transistors
  - Equivalent function of 2.7B
  - eDRAM efficiency
- Eight processor cores
  - 12 execution units per core
  - 4 Way SMT per core
  - 32 Threads per chip
  - 256KB L2 per core
- 32MB on chip eDRAM shared L3
- Dual DDR3 Memory Controllers
  - 100GB/s Memory bandwidth per chip sustained
- Scalability up to 32 Sockets
  - 360GB/s SMP bandwidth/chip
  - 20,000 coherent operations in flight
- Advanced pre-fetching Data and Instruction
- Binary Compatibility with POWER6

* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.*
POWER7 Design Principles:

Multiple optimization Points

- Balanced Design
  - Multiple optimization points
  - Improved energy efficiency
  - RAS improvements

- Improved Thread Performance
  - Dynamic allocation of resources
  - Shared L3

- Increased Core parallelism
  - 4 Way SMT
  - Aggressive out of order execution

- Extreme Increase in Socket Throughput
  - Continued growth in socket bandwidth
  - Balanced core, cache, memory improvements

- System
  - Scalable interconnect
  - Reduced coherence traffic

* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

Graphs for illustration purposes only (Not actual data)
POWER7 Design Principles:

**Flexibility and Adaptability**

- **Cores:**
  - 8, 6, and 4-core offerings with up to 32MB of L3 Cache
  - Dynamically turn cores on and off, reallocating energy
  - Dynamically vary individual core frequencies, reallocating energy
  - Dynamically enable and disable up to 4 threads per core

- **Memory Subsystem:**
  - Full 8 channel or reduced 4 channel configurations

- **System Topologies:**
  - Standard, half-width, and double-width SMP busses supported

- **Multiple System Packages**

  **2/4s Blades and Racks**
  - Single Chip Organic
  - 1 Memory Controller
  - 3 4B local links

  **High-End and Mid-Range**
  - Single Chip Glass Ceramic
  - 2 Memory Controllers
  - 3 8B local links
  - 2 8B Remote links

  **Compute Intensive**
  - Quad-chip MCM
  - 8 Memory Controllers
  - 3 16B local links (on MCM)

* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.
POWER7: Core

- **Execution Units**
  - 2 Fixed point units
  - 2 Load store units
  - 4 Double precision floating point
  - 1 Vector unit
  - 1 Branch
  - 1 Condition register
  - 1 Decimal floating point unit
  - 6 Wide dispatch/8 Wide Issue
- **Recovery Function Distributed**
- **1,2,4 Way SMT Support**
- **Out of Order Execution**
- **32KB I-Cache**
- **32KB D-Cache**
- **256KB L2**
  - Tightly coupled to core
POWER7: Performance Estimates

POWER7 Continues Tradition of Excellent Scalability

- Core performance increased by:
  - Re-pipelined execution units
  - Reduced L1 cache latency
  - Tightly coupled L2 cache
  - Additional execution units
  - More flexible execution units
  - Increased pipeline utilization with SMT4 and aggressive out of order execution

- Chip Performance Improved Greater than 4X:
  - High performance on chip interconnect
  - Improved cache utilization
  - Dual high speed integrated memory controllers

- System
  - Advanced SMP links will provide near linear scaling for larger POWER7 systems.

* Performance estimates relate to processor only and should not be used to estimate projected server performance.
Energy Management: Architected Idle Modes

Two Design Points Chosen for Technology

- **Nap** (optimized for wake-up time)
  - Turn off clocks to execution units
  - Reduce frequency to core
  - Caches and TLB remain coherent
  - Fast wake-Up

- **Sleep** (optimized for power reduction)
  - Purge caches and TLB
  - Turn off clocks to full core and caches
  - Reduce voltage to V-retention
    - Leakage current reduced substantially
  - Voltage ramps-up on wake up
  - No core re-initialization required
Adaptive Energy Management: Energy Scale™

- Chip FO4 Tuned for Optimal Performance/Watt in Technology

- DVFS (Dynamic Voltage and Frequency Slewing)
  - -50% to +10% frequency slew independent per core
  - Frequency and voltage adjusted based on:
    - Work load and utilization.
    - On board activity monitors

- Turbo-Mode
  - Up to 10% frequency boost
  - Leverages excess energy capacity from:
    - Non worst case work loads
    - Idle cores

- Processor and Memory Energy Usage can be independently Balanced.
  - Real time hardware performance monitors used.
  - On board power proxy logic estimates power

- Power Capping Support
  - Allows budgeting of power to different parts of system

![SPEC Power vs Load Level Graph](image)
POWER7: Reliability and Availability Features

- **Dynamic Oscillator Failover**
  - OSC0
  - OSC1

- **Fabric Bus Interface to other Chips and Nodes**
  - ECC protected
  - Node hot add /repair

- **Core Recovery**
  - Leverage speculative execution resources to enable recovery
  - Error detected in GPRs FPRs VSR, flushed and retried
  - Stacked latches to improve SER

- **Alternate Processor Recovery**
  - Partition isolation for core checkstops

- **L3 eDRAM**
  - ECC protected
  - SUE handling
  - Line delete
  - Spare rows and columns

- **GX IO Bus**
  - ECC protected
  - Hot add

- **InfiniBand® Interface**
  - Redundant paths

- **64 Byte ECC on Memory**
  - Corrects full chip kill on X8 dimms
  - Spare X8 devices implemented

- **Dual memory chip failures do not cause outage**
  - Selective memory mirror capability to recover partition from dimm failures
  - HW assisted scrubbing
  - SUE handling
  - Dynamic sparing on channel interface
  - PowerVM Hypervisor protected from full dimm failures

* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.*
Summary

Power Systems™ continue strong

- 7th Generation Power chip:
  - Balanced Multi-Core design
  - EDRAM technology
  - SMT4
- Greater then 4X performance in same power envelope as previous generation
- Scales to 32 socket, 1024 threads balanced system
- Building block for peta-scale PERCS project

POWER7 Systems Running in Lab

- AIX®, IBM i, Linux® all operational

* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.