Bridging High Performance and Low Power in the era of Big Data and Heterogeneous Computing

Ruchir Puri
Emrah Acar, Minsik Cho, Mihir Choudhury,
Haifeng Qian, Matthew Ziegler
IBM Thomas J Watson Research Center, Yorktown Hts, NY
Lessons from History on giving advice

• I will try to avoid giving advice during my remarks.
  As the little school girl wrote, "Socrates was a wise Greek philosopher who walked around giving advice to people. They poisoned him."
Outline

- Big Data optimized system design
  - Power8: A high performance system backbone
  - Power Management & Reduction
  - Design methodology to bridge power performance gap

- What's Next?
  - SW driven HW Acceleration in the era of heterogeneous computing
  - Commercial Workload Case studies
# Recent POWER History

<table>
<thead>
<tr>
<th>Technology</th>
<th>POWER5 2004</th>
<th>POWER6 2007</th>
<th>POWER7 2010</th>
<th>POWER7+ 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cores</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Threads</td>
<td>SMT2</td>
<td>SMT2</td>
<td>SMT4</td>
<td>SMT4</td>
</tr>
<tr>
<td>Caching</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On-chip</td>
<td>1.9MB</td>
<td>8MB</td>
<td>2 + 32MB</td>
<td>2 + 80MB</td>
</tr>
<tr>
<td>Off-chip</td>
<td>36MB</td>
<td>32MB</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Bandwidth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sust. Mem.</td>
<td>15GB/s</td>
<td>30GB/s</td>
<td>100GB/s</td>
<td>100GB/s</td>
</tr>
<tr>
<td>Peak I/O</td>
<td>6GB/s</td>
<td>20GB/s</td>
<td>40GB/s</td>
<td>40GB/s</td>
</tr>
</tbody>
</table>

**Technology**
- 130nm SOI
- 65nm SOI
- 45nm SOI eDRAM
- 32nm SOI eDRAM
- 22nm SOI eDRAM

**Bandwidth**
- 12 SMT8
- 6 + 96MB
- 128MB
- 230GB/s
- 64GB/s
**POWER8 Chip Overview**

- Up to 2.5x socket perf vs. P7+
- 649mm² die size, 4.2B transistors
- 12 high-performance cores
- Large Caches
  - L2: 512KB private SRAM per core
  - L3: 96MB shared eDRAM w/ 8MB “fast access” partition per core
  - L4: Up to 128MB, located on memory buffer chip
- 4 High Speed I/O interfaces
  - Memory, On-Node SMP, Off-Node SMP, PCIe Gen3
- CAPI: open infrastructure for off-chip, memory-coherent accelerators
**POWER8 Technology**

- 22nm SOI
- 15 layer BEOL:
  - 5-1x, 2-2x, 3-4x, 3-8x, 2-UTM
- 3-Vt thin-oxide logic transistors for power optimization
- Multiple thick-oxide transistors (for I/O and analog support)

- 3 app-optimized SRAM cells:
  - 0.160um² 6T perf-oriented
  - 0.144um² 6T perf-density balance for directories/L2
  - 0.192um² 8T multi-port
- Technology eDRAM cell: 0.026um²
**POWER8 Core:** Back bone of big data computing system

**Enhanced Micro Architecture**
- Increased Execution Bandwidth, +4 units
- SMT 8
- 64KB L1 D-Cache, 32KB 8-way I-Cache
- 64B Cache Reload
- 4KB TLB
- Transactional Memory

**Arrays/Register Files**
- 2 CAM & 6 SRAM Topologies
- 31 Multi-ported Register Files for Queuing & Architected Registers

**Power Management**
- Power Gating & Voltage Regulation in 5 columns
- 1 Thermal Diode
- 3 Digital Thermal Sensors
- 3 Critical Path Monitors
Combined I/O Bandwidth = 7.6Tb/s

Putting it all together with the memory links, on- and off-node SMP links as well as PCIe, at 7.6Tb/s of chip I/O bandwidth
**SRAM Power Savings**

- Global bitline restored to reduced voltage $V_{DD} - V_T$
  - 20% AC power savings

- Smart way select prediction to reduce restore power

- Early and late wordline gating features

- Wordline driver header devices
  - 16% DC power savings

- Output socket buffer concept: driver size tuned load of each instance
Resonant clocking reduced chip power by 4%, as well as improving clock jitter in those meshes, which translates into a significant frequency boost.
Power Regulation and Reduction: Exploiting processor inactivity

- Power consumption varies at every time scale
- Key = sense and act in time
**POWER8 On-Chip Controller (OCC)**

- Allows for fast, scalable monitoring and response (ns timescale)
  - Independent of Hypervisor or Guest OS(s)
  - OR
  - In conjunction with Hypervisor interaction with Guest OS(s)

![Diagram of POWER8 On-Chip Controller (OCC)](image)

- Guest OS
- Guest OS
- Guest OS
- Guest OS

Hypervisor

- C0
- C1
- C0
- C1

MC NEST I/O

OCC

POWER8

Real-time Updates

Policy/Budget Control

Not Real-time
OCC = full POWERPC 405 core with 512KB private memory
- Uses continuous running, real-time OS
- Monitors workload activity, chip temperature and current
- Adjusts frequency and voltage to optimize performance within system power and thermal constraints
• Small Analog-$V_{DD}$ domains not shown
• Precision voltage reference not shown
• Core/L2 and L3 regions divide the same input power into independent power-gated regions
On-Die Per Core Power gating

Die Per Core Power gating

L3 mixed VDD & VCS

Core/L2 mixed VDD & VCS

Pwr Mgmt Controller

Charge pump

1000X leakage reduction

<<1% di/dt noise when powering up a core.
On-Die Per Core Voltage Regulation

- Each of the 12 core/cache partitions can adapt voltage to optimize power vs. performance demands

![Diagram of voltage regulation](image)
On Chip Voltage Regulation Benefit

- DVFS results vs DFS: ~33% power savings @ 62% freq

![Graph showing relative power vs relative performance for Bypass and Regulated Modes with Vin=1.1v]
**Power analysis & improvements**

- Design effort = 17% savings of chip total power
Large Block Structured Synthesis

- Enhanced process which included:
  - Structured dataflow
  - Congestion-aware stdcell placement
  - Embedded “hard” IP (e.g. arrays, regfiles, complex custom cells)

- 30% fewer unique blocks vs. POWER7

- Improvements in block power and total design area
  - 15% area reduction

- Gate-level design TAT sign-off improvement of 3-10x
Design Efficiency for Power/Performance

Push for more synthesis & Structured Synthesis

Custom productivity innovations

Push for larger macros & Large Block Synthesis
Automated Datapath techniques achieve significant wirelength and timing improvement over conventional synthesis.

Improvements over conventional synthesis

<table>
<thead>
<tr>
<th>Design Version</th>
<th>Timing</th>
<th>Wire Length</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>b) Designer latch preplacement</td>
<td>28%</td>
<td>30%</td>
<td>5%</td>
</tr>
<tr>
<td>c) Automated latch placement</td>
<td>16%</td>
<td>27%</td>
<td>2%</td>
</tr>
</tbody>
</table>
The tailored macro power optimization methodology applies high exploration effort early during the synthesis step as well as final exploration during post-route tuning. Figure from [2].

I. Order macros based on expected power savings ROI

II. Process macros starting from highest ROI, continue down the macro list as design effort allows or ROI becomes unattractive

1) Synthesis
   X scenarios explore the design space both in parallel and iteratively

2) Routing
   Y top synthesis scenarios are routed

3) Tuning
   Z top routed scenarios are tuned

4) Top tuned scenario selected for RIT

Design methodology to bridge the high performance and low power gap
High Performance: IFU and VSU as LBSS

IFU:  580K gates, 628K nets
  - 37 embedded array/register files

VSU:  697K gates, 723K nets
  - 20 embedded array/register files
  - hierarchical embedded synthesis

Preplaced IP
Structured dataflow
P8 Core: A finely tuned power performance compute engine
What's Next?

- Technology trends are motivating increasing focus on acceleration and specialization as more impactful means to increase system value.

- Targeted specialization can result in dramatic improvements – *10X and more in both performance and power efficiency*.

- A broad understanding of workloads, system structures, and algorithms is needed to determine what to accelerate / specialize, and how:
  - Via SW; via HW; via SW+HW
  - Many choices; co-optimization necessary

- A methodology for software and system co-optimization, based on inventing new software algorithms, that have strong affinity to hardware acceleration.

  A new dimension to algorithm effectiveness: hardware mapping efficiency.
Concluding Remarks

- Life as usual will continue, only with more sweat and blood.
  - Technology becomes much harder
  - Design effort enormous.. Marching onwards.. P9..
  - Power increasingly becoming first order metrics at system level and percolates down to chip and then to core, and finally design methodology

- Specialization will become increasingly relevant esp. as power efficiency becomes more important.
  - For commercial workloads, must contend with massive scaling of the CPUs and algorithmic paradigms at both SMP and cloud computing level.

- Where POWER is critical and performance a key requirement, then specialization will be indispensable.
I hope you enjoyed the talk and if you did not, I hope you had a good nap.
END