Low-Power, High-Performance Architecture of the PWRficient Processor Family

Presented by Tse-Yu Yeh
Director, Architecture & Verification

Hot Chips 18
Aug 21, 2006
Outline

- Design paradigm
- Family
- Core
- Performance and power
- Summary
Low-Power Design Paradigm

- Design goal — 970-class performance at less than 7 watts per core
- Power dissipation is a primary consideration at all levels
  - Circuit and Process
    - Voltage/frequency scaling
    - Multiple power planes for optimal voltage selection per region
  - Microarchitecture and Logic
    - Clock gating to reduce power of idle circuits
    - Active and pre-charge standby modes in external DRAM array
    - Sizing and hierarchy of cache structures
  - Architecture
    - Integration to reduce I/O interface power
    - Internal and external power-saving modes — CPU, memory controller, PCIe
  - Verification
    - Monitor power consumption against budget throughout the design process
  - Software
    - Power management of I/O devices and CPU
Power-Influenced μ/Architectural Choices

- Extensive fine-grained clock gating used throughout core and SOC
- If a function can be performed sequentially without performance loss, why build power-hungry parallel mechanisms?
- The smaller, the better
  - Design employs smaller subsections that are powered up for frequent accesses
  - Much attention to clocking only those elements that are performing work
- Each major block’s design criteria included goals for power in addition to the traditional area, complexity & timing budgets
  - Energy per memory access at each level of cache vs. DRAM
  - Leakage, voltage, and execution frequency
  - Overall execution time saving
- Speculation has to be effective or it becomes a power sink
Fine-Grain Clock Gating Reduces Dynamic Power

![Graph showing the comparison between fine-grain clock gating and coarse-grain clock gating. The graph illustrates the percentage of flops clocked over time during different conditions: Reset and Flop Initialization, Normal Operation, and Thermal Virus. The graph shows a significant reduction in the percentage of flops clocked during coarse-grain clock gating compared to fine-grain clock gating.]
Device-Specific $V_{dd}$ Reduces Static and Dynamic Power

**Conventional approach**
- Operate at 1.1V across entire process range
- Fast parts tend to be very leaky

**P.A. Semi approach**
- Operate at device-specific optimal $V_{dd}$
- Partition power plane for optimal voltage selection per region
- Enables full process range for power yield
PWRficient Family

▶ PA6T core — the CPU
  ▶ Power Architecture compliant, 64-bit, high-performance FPU and VMX
  ▶ 7W @ 2GHz worst-case power dissipation

▶ CONEXIUM™ — the on-chip coherent interconnect
  ▶ Scalable cross-bar interconnect
  ▶ 1–8 SMP cores
  ▶ 1 or 2 L2 caches, sized 512KB–8MB
  ▶ 1–4 64-bit DDR2 memory controllers

▶ ENVOI™ — the I/O system
  ▶ SERDES I/O—PCI Express®, XAUI, SGMII
  ▶ Offload engines—TCP/IP, iSCSI, cryptography, and RAID
  ▶ Support I/O—Boot bus, UARTs, SMBus, GPIOs
PWRficient PA6T-1682M Block Diagram

CONEXIUM™ Interchange

ENVOI™ Intelligent I/O

*Transaction trace memory †Peripheral trace memory
The PA6T Core
PA6T Core Features

- Fully compliant Power Architecture implementation
  - Power Architecture version 2.04
  - Full FPU and VMX SIMD capabilities
  - 64-bit with 32-bit capability

- Super-scalar, out-of-order design
  - L1 instruction cache
    - Fetch four instructions per cycle into 64-entry scheduler
    - Issue up to 3 per cycle into 6 functional units
  - Branch predictors
  - Strongly ordered memory model
    - Issue out of order, retire in order

- Hypervisor and virtualization support

- High-performance memory hierarchy @ 2GHz
  - L1 data—32GB/s read or write
  - L2 data—16GB/s read plus 16GB/s write
  - DDR2-1067—16GB/s read or write
  - 16 transactions in flight

- CONEXIUM Interchange
  - 1G transactions per second
  - 64GB/s peak data rate
  - MOESI coherency protocol
PA6T Low-Power Features

► Improved branch prediction
  ▶ Minimize incorrect speculation to avoid wasting power

► Efficient superscalar design
  ▶ Index-based, out-of-order execution engine minimizes the use of CAMs and avoids unnecessary replays

► Energy-efficient memory pipelines
  ▶ Hierarchical address translation to balance power consumption and performance
  ▶ Highly integrated memory pipeline that supports out-of-order execution and sequential consistency with minimal data movement

► Coherent memory subsystem
  ▶ Coherent memory subsystem designed for high throughput and low latency while minimizing the energy used per reference

► Extensive power-management capabilities
  ▶ Doze, nap, and sleep power-down modes trade varying degrees of power savings with recovery time
Processor Pipeline

Instruction Fetch

Instruction Decode & Map

Instruction Issue

Integer

Floating Point

VMX

Load/Store
Front End

- **I-cache**
  - 64KB
  - 2-way associative
  - 2-cycle
  - 4 instructions/fetch
  - 4 fetches to CONEXIUM
  - Next-line pre-fetch after a miss

- **Instruction buffer**
  - 4 fetch groups (16 instructions)

- **Decode**
  - 4 µ ops/dispatch

- **Map**
  - 4 µ ops renamed/cycle
  - 64 rename registers
Branch Processing

- **Branch prediction**
  - 0-cycle bubble: 16-entry next-fetch prediction

- **4-cycle bubble (taken prediction)**
  - Path: 16K bi-mode taken/not taken (4 predicted)
  - 16-entry return stack
  - 64-entry history-assisted target predictor
  - IP-relative

- **Early verification**

- **Branch execution**
  - 2-cycle branch execution on integer P1

- **Branch misprediction**
  - 13-cycle path mispredict
  - 13-cycle target mispredict

Back to main diagram
Scheduler and Execution Pipes

- **Schedule & Issue**
  - 64-entry schedule buffer
  - 64×64 dependency
  - 3 pickers
  - Pre-slotted
  - Replay from scheduler
  - Replay prevention

- **Retire**
  - Commit to architecture registers
  - Precise exception
  - 4 µ ops/cycle

- **3 pickers, 5 execution pipes + one load/store pipe**
  - Integer (P0&1): 2-cycle
  - FP (P1): 8-cycle
  - VMX (P1): 8-cycle FP or 6-cycle complex integer
  - VMX (P0): 2-cycle permute or simple integer
  - Load/store (P2)
Latency — Load to Use
- L1 cache 4
- L2 cache 24
- Open page 100
- Closed page 120
- Remote L1 42

Loads and Stores
- Issued out of order
- Strongly-ordered stores

32-Entry Load/Store Queue
- Re-ordered for consistency
- Checking and retire
- Store-to-load forwarding

16 blocks in flight
12-entry hardware prefetcher

Load/Store Processing
Address Translation

- I-address translation
  - 4-entry direct μIERAT
  - 64-entry 2-way IERAT
- D-address translation
  - 128-entry 2-way DERAT
- SLB — 64-entry fully-assoc
- TLB — 1024-entry 4-way
- H/W Page table walk
  - 4-walks in flight
  - Prefetch of next PTE after a TLB miss

EA 64-bit effective address
VA 65-bit virtual address
RA 44-bit real (physical) address

ERAT Effective-to-real address translation
SLB Segment look-aside buffer
HTW Hardware page-table walk
CONEXIUM Interchange

- **Transaction initiators**: cores & I/O bridge
  - All devices respond
  - MOESI-style protocol
    - Minimize copy-back to memory and L2-cache to save power
- **Address bus cycles at half the core frequency**—1G address/sec
- **Address arbitration** enforces strong ordering
- **Data connected as crossbar**
  - Scales with number of agents
- **Each port provides 16-byte dual-simplex connections**
Memory Hierarchy

- On-chip memories are power efficient
  - RAM structures have low power density due to low inherent activity
  - Only a few of many bit cells accessed per cycle
  - On-chip RAMs save power by avoiding chip-to-chip bus structures

- Most on-chip memory is devoted to caches
  - Caches have diminishing (logarithmic) performance return vs. size
Power Saving Modes

- **Doze (entry time immediate, wakeup time immediate)**
  - Cores idle at reduced frequency; continue snooping on the bus
  - Wake up immediately—no state reloading needed

- **Nap (entry time 2–16µs*, wakeup time <0.5 ms)**
  - Core clock stopped and voltage lowered to reduce leakage
  - D-cache modified data is flushed by hardware
  - All architecture state is retained
  - SRAM remains power on, value retained
    - Branch predictors: state retained
    - TLB, I-cache: invalidate if snooped

- **Sleep (entry time 2–16µs*, wakeup time <1 ms†)**
  - Core powered off (either or both cores)
  - D-cache modified data is flushed by hardware
  - Some architecture state must be saved by software
  - On wakeup, core goes through power-on-reset sequence

* Depends on number of modified L1 data cache lines
† Wakeup time is dominated by power regulator restart time
PWRfcient Performance & Power
High Performance at Low Power Across a Range of Metrics

PWRficient 1682M PROVIDES MAINSTREAM PERFORMANCE AT LOW POWER

▶ General-purpose computing
  ▶ SPECint®2000 >1000 per core *

▶ Floating-point performance
  ▶ SPECfp®2000 >2000 per core *
  ▶ Imaging
    ▶ FFT 24 GFlops/sec (total) †

▶ System bandwidth
  ▶ Sustained block copy
    ▶ 10 Gigabytes/sec *
  ▶ High-speed SERDES I/O
    ▶ 104Gbps aggregate peak bandwidth

▶ Application offloads
  ▶ TCP/IP termination
    ▶ > 20Gbps $§
  ▶ Encryption
    ▶ 10Gbps IPSec/SSL encryption and authentication $§
    ▶ 3,000 public-key handshakes/sec in software *
  ▶ Storage
    ▶ 2.0GB/s RAID5 (data+parity) $§

*Estimated max sustained performance at 2GHz
†75% of estimated max sustained performance at 2GHz, 1D single-precision FFT with 2K elements
$Estimated peak performance at 2GHz
Power Efficient

POWER IS FIRST-ORDER DESIGN PRINCIPLE

- MC optimizes DRAM power
- L2 cache and I/O are coherent with core off
- Dynamic power-control hardware and software voltage/frequency management

> 15,000 gated clocks
Separate power rails for cores, I/O pads
Turn off unused I/Os
PA6T-1682M Projected Power Dissipation

- Projected power assuming full power-regulation scheme
  - Independent per-core VRM
  - Dynamic control loop based on frequency and operating conditions
  - VRM for SOC
  - Static control

<table>
<thead>
<tr>
<th></th>
<th>Max Freq</th>
<th>Typ</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA6T core only</td>
<td>2.0GHz</td>
<td>4W</td>
<td>7W</td>
</tr>
<tr>
<td>PA6T-1682M-FCN</td>
<td>2.0GHz</td>
<td>17W</td>
<td>25W</td>
</tr>
<tr>
<td>PA6T-1682M-FCG</td>
<td>1.5GHz</td>
<td>8W</td>
<td>15W</td>
</tr>
<tr>
<td>PA6T-1682M-FCD</td>
<td>1.0GHz</td>
<td>6W</td>
<td>10W</td>
</tr>
<tr>
<td>I/O coherent nap</td>
<td></td>
<td></td>
<td>2W*</td>
</tr>
</tbody>
</table>

*PA6T-1682M-FCN nap power may be higher
Summary

- Power-aware design from ground up to maximize performance/Watt
- Modular design supports rapid family deployment
- Interesting convergence of needs across a wide range of design points
  - Networking, telecom, servers, mil/aero, imaging
- Performance and power enable wide range of applications
Contact P.A. Semi

For further information, please visit P.A. Semi web site at:

www.pasemi.com

Kindly direct sales inquiries to:

pasasales@pasemi.com

Full contact information:

P.A. Semi, Inc.
3965 Freedom Circle, Floor 8
Santa Clara
CA 95054-1203 USA
Main: 408.200.4500
Fax: 408.200.4501
Thank You

The P.A. Semi name and the P.A. Semi logo and combinations thereof are trademarks of P.A. Semi, Inc. The Power name is a trademark of International Business Machines Corporation, used under license therefrom. SPECint and SPECfp are registered trademarks of the Standard Performance Evaluation Corporation (SPEC). All other trademarks are the property of their respective owners.