Microprocessor Power Impacts

Mandy Pant
May 2010
Contents

Trends in power consumption
Utilization and breakdown of power usage
Initial efforts to control power with thermal management
Processor power and performance states
Enhanced processor power control features
System interaction of processor power features
Future directions
Summary
Contents

Trends in power consumption
Utilization and breakdown of power usage
Initial efforts to control power with thermal management
Processor power and performance states
Enhanced processor power control features
System interaction of processor power features
Future directions
Summary
Power Density vs. Critical Dimension

W/cm² vs. CD (µm)

Rocket Nozzle
Nuclear Reactor
Hot Plate


May 2010
Energy Star

Thin Client Operational Mode Power Requirements

<table>
<thead>
<tr>
<th>Mode</th>
<th>Power Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Off Mode</td>
<td>≤ 2 W</td>
</tr>
<tr>
<td>Sleep Mode (if applicable)</td>
<td>≤ 2 W</td>
</tr>
<tr>
<td>Idle State</td>
<td></td>
</tr>
<tr>
<td>Category A:</td>
<td>≤ 12.0 W</td>
</tr>
<tr>
<td>Category B:</td>
<td>≤ 15.0 W</td>
</tr>
</tbody>
</table>

Source: Intel Energy Star Web Site
May 2010
Data Center Trends

Limiting Growth Factor

- Available Power: 31%
- Cooling Capacity: 28%
- There are no limiting factors: 17%
- Floor Space: 10%
- Funding: 9%
- Other: 5%
- Floor Loading (weight): 0%

Power & cooling capacity limits growth

Source: D. Filani, et al., Intel Technology Journal, Q1 2008
Power Density vs. Critical Dimension

Contents

Trends in power consumption

Utilization and breakdown of power usage

Initial efforts to control power with thermal management

Processor power and performance states

Enhanced processor power control features

System interaction of processor power features

Future directions

Summary
System power levels under Energy Efficient Performance 2.0 workload

- Simulates employee Sysmark 2007 based workload

Sleep + Idle + <10% load totals 71.7% of the 9 hour day!

System power levels during a 9-hour workday following the EEP 2.0 Model demonstrate that 75 percent of time is at or very near idle utilization.

Source: P. Zagacki, et al., Intel Technology Journal, Q4 2008
System Idle Power Breakdown

Approximate component power consumption (including losses) for a 45-W AC idle platform.

Source: P. Zagacki, et al., Intel Technology Journal, Q4 2008
Server Utilization is Low

Source: Intel IT Brief on Server Rightsizing, July 2006
Server Power Breakdown

Server focus is on processors, memory, power delivery, and power removal

Power Control Message

Focus on sleep, idle, and low utilization conditions
Focus on CPU, chipset, memory, and power delivery losses
Contents

Trends in power consumption
Utilization and breakdown of power usage
Initial efforts to control power with thermal management
Processor power and performance states
Enhanced processor power control features
System interaction of processor power features
Future directions
Summary
Power Optimization

Small gain in performance as maximum power is approached

Motivates having a throttle to constrain the design

Source: R. Viswanath, et al., Intel Technology Journal, Q3 2000
Initial Thrust – Temperature Throttling

Intel Thermal Monitor 1 (TM1)
- If processor silicon reaches its maximum junction temperature or the power reduction is requested
- Thermal Control Circuit (TCC) activates and core clocks are started and stopped to reduce power and temperature
- Bus traffic is snooped in the normal manner, and interrupt requests are latched and serviced during the time that the clocks are on
- After the silicon cools or power reduction request ends (with hysteresis), clock modulation ends

Source: Intel Datasheets; US Patent 7275012
Throttle States

T-state clock modulation duty cycle is in eighths
Lowest setting normally used
• Heavy handed – major reduction in power and performance

<table>
<thead>
<tr>
<th>Duty Cycle Code</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x0</td>
<td>Undefined</td>
</tr>
<tr>
<td>0x1</td>
<td>12.5% clocks on / 87.5% clocks off</td>
</tr>
<tr>
<td>0x2</td>
<td>25% clocks on / 75% clocks off</td>
</tr>
<tr>
<td>0x3</td>
<td>37.5% clocks on / 62.5% clocks off</td>
</tr>
<tr>
<td>0x4</td>
<td>50% clocks on / 50% clocks off</td>
</tr>
<tr>
<td>0x5</td>
<td>62.5% clocks on / 37.5% clocks off</td>
</tr>
<tr>
<td>0x6</td>
<td>75% clocks on / 25% clocks off</td>
</tr>
<tr>
<td>0x7</td>
<td>87.5% clocks on / 12.5% clocks off</td>
</tr>
</tbody>
</table>

Future products could add ratio selectability

Source: Intel Datasheets
Frequency and Voltage Transitions

Intel Thermal Monitor 2 (TM2)
- A more gentle and elegant mechanism than TM1
- If processor silicon reaches its maximum junction temperature
- Core voltage and clock frequency are stepped down just enough to reduce temperature to safe operating levels
- After the silicon cools (with hysteresis), voltage and clocks are stepped back up

Both TM1 and TM2 can co-exist
- TM2 is activated 1st, and TM1 called only if TM2 is not sufficient

Source: Intel Thermal/Mechanical Specifications and/or Datasheets
# Thermal Throttling Summary

Temperature above activation limit – TM2 V/f reduction, prochot asserted
Temperature persists above limit – TM1 clock modulation, prochot asserted
Temperature rises above safe limit – catastrophic shutdown, thermtrip asserted
System requests power reduction by asserting ForcePR (servers) or Prochot (DT/Mobile) – TM2 V/f reduction

<table>
<thead>
<tr>
<th>Item</th>
<th>Processor Input</th>
<th>Processor Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM1/TM2</td>
<td>Core $X \geq$ TCC Activation Temperature</td>
<td>All Cores TCC Activation</td>
</tr>
<tr>
<td>PROCHOT#</td>
<td>Core $X \geq$ TCC Activation Temperature</td>
<td>PROCHOT# Asserted</td>
</tr>
<tr>
<td>THERMTRIP#</td>
<td>Core $X \geq$ THERMTRIP # Assertion Temperature</td>
<td>THERMTRIP# Asserted, all cores shut down</td>
</tr>
<tr>
<td>FORCEPR#</td>
<td>FORCEPR# Asserted</td>
<td>All Cores TCC Activation</td>
</tr>
</tbody>
</table>

Source: Intel Datasheets

May 2010
Contents

Trends in power consumption
Utilization and breakdown of power usage
Initial efforts to control power with thermal management

Processor power and performance states
Enhanced processor power control features
System interaction of processor power features
Future directions
Summary
Power Optimization

Source: R. Viswanath, et al., Intel Technology Journal, Q3 2000
Advanced Configuration and Power Interface States

When the processor is not executing code, it is idle
A processor low-power idle state is defined by ACPI as a C-state
More power savings actions are taken for numerically higher C-states
In general, lower power C-states have longer entry and exit latencies

<table>
<thead>
<tr>
<th>Global (G) State</th>
<th>Sleep (S) State</th>
<th>Processor Core (C) State</th>
<th>Processor State</th>
<th>System Clocks</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0</td>
<td>S0</td>
<td>C0</td>
<td>Full On</td>
<td>On</td>
<td>Full On</td>
</tr>
</tbody>
</table>
When the processor is not executing code, it is idle

A processor low-power idle state is defined by ACPI as a C-state

More power savings actions are taken for numerically higher C-states

In general, lower power C-states have longer entry and exit latencies

<table>
<thead>
<tr>
<th>Global (G) State</th>
<th>Sleep (S) State</th>
<th>Processor Core (C) State</th>
<th>Processor State</th>
<th>System Clocks</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0</td>
<td>S0</td>
<td>C0</td>
<td>Full On</td>
<td>On</td>
<td>Full On</td>
</tr>
<tr>
<td>G0</td>
<td>S0</td>
<td>C1/C1E</td>
<td>Auto-Halt</td>
<td>On</td>
<td>Auto-Halt</td>
</tr>
<tr>
<td>G0</td>
<td>S0</td>
<td>C3</td>
<td>Deep Sleep</td>
<td>On</td>
<td>Deep Sleep</td>
</tr>
<tr>
<td>G0</td>
<td>S0</td>
<td>C6</td>
<td>Deep Power Down</td>
<td>On</td>
<td>Deep Power Down</td>
</tr>
<tr>
<td>G1</td>
<td>S3</td>
<td>Power off</td>
<td>Power off</td>
<td>Off, except RTC</td>
<td>Suspend to RAM</td>
</tr>
<tr>
<td>G1</td>
<td>S4</td>
<td>Power off</td>
<td>Power off</td>
<td>Off, except RTC</td>
<td>Suspend to Disk</td>
</tr>
<tr>
<td>G2</td>
<td>S5</td>
<td>Power off</td>
<td>Power off</td>
<td>Off, except RTC</td>
<td>Soft Off</td>
</tr>
<tr>
<td>G3</td>
<td>NA</td>
<td>Power off</td>
<td>Power off</td>
<td>Power off</td>
<td>Hard off</td>
</tr>
</tbody>
</table>

Source: Intel Datasheets
Resolution of C-states occurs at thread, core, and package levels

- A core is at the lowest C-state of any of its threads, a package at the lowest C-state of its cores

A core transitions to C0 state when an interrupt occurs or when there is an access to the monitored address (if the state was entered using an MWAIT)

For core C1/C1E and C3, an interrupt directed toward a single thread wakes only that thread but the core resolves to C0

For core C6, an interrupt in either thread wakes both into C0 state

Any interrupt coming into the processor package may wake any core

Source: Intel Datasheets
May 2010
Each frequency/voltage operating point is defined as a P-state

Inflection point occurs where minimum operating voltage is reached
• Only less power-efficient frequency scaling below that point

Desire a wider dynamic range above the inflection point (lower Vmin)
• Better system power control, more turbo boost

Source: A. Naveh, et al., Intel Technology Journal, Q2 2006
Processor Domains

8-core Nehalem (core i7) server example

- Has a PLL per core and a power gate per core
- Has a PLL for the un-core and large caches
- Has PLL’s for the QPI
- Has PLL’s for the SMI

Source: S. Rusu, et al., JSSCC, Jan. 2010
Minimizing Power in Disabled Blocks

Disabled cores ➤ Power gated

Active/ Shut-off

Core

Active

Shut-off

0.85V

Virtual VCC

Leakage Reduction

40x

0V

Disabled cache slices ➤ All major arrays in shut-off

Active/ Shut-off

SRAM array

Active

Sleep/ Shut-off

0.9V

Virtual VCC

Leakage Reduction

35%

83%

0V

Source: S. Rusu, et al., JSSCC, Jan. 2010
Infrared Image of Power Gate Shutoff

Only the temperature sensor (TS) remains on in the shut-off cores (upper left and upper right)

Source: S. Rusu, et al., JSSCC, Jan. 2010
Intel® Turbo Boost Technology is a feature that allows the processor core to opportunistically and automatically run faster than its rated operating frequency if it is operating below power, temperature, and current limits.

- Maximum frequency is dependant on the product, SKU, and the number of active cores.
- No special hardware support is necessary.
- BIOS and the operating system can enable or disable Intel Turbo Boost Technology.

Allows work to complete more quickly and then go to C6, saving overall energy.

Source: Intel 5500 Series Animated Brief, Intel Datasheets
Contents

Trends in power consumption
Utilization and breakdown of power usage
Initial efforts to control power with thermal management
Processor power and performance states
Enhanced processor power control features
System interaction of processor power features
Future directions
Summary
Power Control Unit Connectivity to the Processor Cores

Source: S. Rusu, et al., JSSCC, Jan. 2010
Thermal Sensors

9 temperature sensors
- One in each core hot spot
- One in the die center
- Temperature information is available through the Platform Environment Control Interface (PECI) bus for system fan management

Source: S. Rusu, et al., JSSCC, Jan. 2010
Fan Speed Control

Systems management controller reads processor temperature
Adjusts fan speed as needed to keep components cool
• Minimizes energy usage and fan noise

Source: Intel Thermal/Mechanical Specifications
36 May 2010
Improving power efficiency at light loads on VRD 11.1 controllers.
Load Adaptive Voltage Regulation

Core VR 5 phases
Cache VR 3 phases
Nehalem-EX Processor

Full Load Mode
All VR phases are enabled
Maximum VR efficiency

Core VR 1 phase
Cache VR 2 phases
Nehalem-EX Processor

Idle Mode
Turn off 4 core and 1 cache phases
Maximum VR efficiency

Nehalem-EX extends the VR phase shut-off to the cache supply
About 2W power reduction per socket in idle mode

Source: S. Rusu, et al., JSSCC, Jan. 2010
CPU and Memory Working Together

In many of today’s systems, especially servers and workstations, memory is a major energy consumer.

Growing opportunity for memory power management.

Processors with integrated memory controller have added features to reduce memory power:

- Either based on link low power state (L1)
- Or, based on processor C-state

<table>
<thead>
<tr>
<th>Mode</th>
<th>Memory State with External Graphics</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0, C1, C1E</td>
<td>Dynamic memory rank power down based on idle conditions.</td>
</tr>
<tr>
<td>C3, C6</td>
<td>Dynamic memory rank power down based on idle conditions If there are no memory requests, then enter self-refresh. Otherwise, use dynamic memory rank power down based on idle conditions.</td>
</tr>
<tr>
<td>S3</td>
<td>Self Refresh Mode</td>
</tr>
<tr>
<td>S4</td>
<td>Memory power down (contents lost)</td>
</tr>
</tbody>
</table>

Source: Intel Datasheets
Residency Examples

Top graph shows platform C-state residency under idle conditions

- C4 state support added for 2008 platform netting major improvements
- Interrupts, driver wake-ups, and USB polling prevent 100% C4

Bottom chart shows residency improvement with mouse/keyboard USB polling selectively suspended

- Approaching 100% C4

Source: P. Zagacki, et al., Intel Technology Journal, Q4 2008
Contents

Trends in power consumption
Utilization and breakdown of power usage
Initial efforts to control power with thermal management
Processor power and performance states
Enhanced processor power control features
System interaction of processor power features

Future directions

Summary
Core Count Trend

The increasing core count trend is

- Adding pressure to fit more cores within a given power budget
- Creating opportunities for power/performance optimization for code with low thread counts and/or low utilization

Source: S. Rusu, et al., JSSCC, Jan. 2010
Data Center Management Systems

Policy Directives
(e.g. Limit platform power to 250W)

Results, Notifications & Alerts,

Power Sensors

Temperature Sensors

Utilization ‘sensors’
(workload behavior)

Decide
Policy Engine

Monitor

Action

Closed-loop control Algorithms

Power Management architecture

Source: D. Filani, et al., Intel Technology Journal, Q1 2008
Power Optimization

Managing scaled performance is going to be the next step.
Contents

Trends in power consumption
Utilization and breakdown of power usage
Initial efforts to control power with thermal management
Processor power and performance states
Enhanced processor power control features
System interaction of processor power features
Future directions

Summary
Summary

Processor power and power density were trending sharply up
Have taken a hard right-hand turn the last few years holding the line or even reducing processor power
• Added temperature throttling (T-states)
• Added processor low power states (C-states)
• Added processor performance levels (P-states)
Have enhanced processors with a number of advanced power features
• Clock and voltage domains with power gate shut-off
• Digital temperature sensors with PECI read-outs for fan control
• Load adaptive voltage regulator capability
• Memory rank power down and self-refresh support
Working towards system and data center power management architecture
Expect continued innovation to support future Moore’s Law trajectory
ACKNOWLEDGEMENTS

I’d like to thank Bill Bowhill, Stefan Rusu, Dave Ayers for helping me with the material