Power management architecture of the 2nd generation Intel® Core™ microarchitecture, formerly codenamed Sandy Bridge

Efi Rotem - Sandy Bridge power architect
Alon Naveh, Doron Rajwan, Avinash Ananthakrishnan, Eli Weissmann

Hot Chips Aug-2011
Agenda

- Power management overview
- Intel® Turbo Boost Technology 2.0
- Thermal management
- Energy efficiency
- Average power management
- Platform view
- Summary

High CPU and PG performance
Power and energy efficiency
Power management overview
Sandy Bridge power mgmt ID card

- **Sandy Bridge is:**
  - 1-4 CPU cores + PG
  - Integrated System Agent (SA)
  - Sliced LLC shared by all cores/PG
  - Ring interconnect + power management link

- **Package Control Unit (PCU):**
  - On chip logic and embedded controller running power management firmware
  - Communicates internally with cores, ring and SA
  - Monitors physical conditions
    - Voltage, temperature, power consumption
  - Controls power states
    - CPU and PG voltage and frequency
    - Controls voltage regulators DDR and system

- **External power management interface**
  - Accepts external inputs
    - System power management requests and limits
    - Power and temperature reading
  - MSR, MMIO and PECI system bus
Voltage and frequency domains

- **Two Independent Variable Power Planes:**
  - CPU cores, ring and LLC
  - Embedded power gates - Each core can be turned off individually
  - Cache power gating - Turn off portions or all cache at deeper sleep states
  - Graphics processor
    - Can be varied or turned off when not active

- **Shared frequency for all IA32 cores and ring**

- **Independent frequency for PG**

- **Fixed Programmable power plane for System Agent**
  - Optimize SA power consumption
  - System On Chip functionality and PCU logic
  - Periphery: DDR, PCIe, Display
Power performance fundamentals

- Maximize user experience under multiple constraints
  - User Experience (May have different preferences):
    - Throughput performance
    - Responsiveness - burst performance
    - CPU / PG performance
    - Battery life / Energy bills
    - Ergonomics (acoustic noise, heat)
  - Optimizing around Constraints to meet user preferences
    - Silicon capabilities
    - System Thermo-Mechanical capabilities
    - Power delivery capabilities
    - S/W and Operating system explicit control
    - Workload and usage

Rich set of control knobs for the user and system designer to tailor power - performance preferences
Power management features topology

**S/W Platform**
- Operating system, PG driver, BIOS, Embedded Controller and user preferences

**Power Perf Opt.**
- Power/performance optimization algorithms
- Milliseconds to seconds control algorithms

**Real Time events**
- PCU “kernel” – mission critical power management events
- C-state control, P-states transitions and latency sensitive actions

**Physical Layer**
- Thermal sensing, Maximum current control, physical layer communication
- Platform control: DDR thermal, Voltage Regulator optimization, hot sensors etc.
Intel® Turbo Boost Technology 2.0
Power metering

- Power management is based on power metering
- Sandy Bridge implements a digital power meter
  - 3rd generation of power metering in Intel® products
  - Active power - Event counters track main building blocks activities
    - 100 Micro arch. event counters - apply active energy cost to each event
    - CPU, PG, Ring, Cache, and I/O
  - Static power – Leakage and idle as a function of voltage and temperature
- Used for power management algorithms
- Architecturally exposed to software and system
  - For the use of S/W or system embedded controller
What is CPU Turbo

- **P-state**: a voltage/frequency pair (ACPI terminology)
- **P1** is guaranteed frequency
  - CPU and PG simultaneous heavy load at worst case conditions
  - Actual power has high dynamic range
- **P0** is max possible frequency
- **Pn** is the energy efficient state
  - OS control Pn-P1 range
- **P1-P0** has significant frequency range (GHz)
  - P1 to P0 range is fully H/W controlled
  - User preferences and policies
  - Single thread or lightly loaded applications
  - GFX <> CPU balancing
What is Turbo

- Turbo enabled product specifications

### Table 5-1. TDP Specifications

<table>
<thead>
<tr>
<th>Segment</th>
<th>State</th>
<th>CPU Core Frequency</th>
<th>Processor Graphics Core frequency</th>
<th>Thermal Design Power</th>
<th>Units</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extreme Edition (XE)</td>
<td>HFM</td>
<td>2.5 GHz up to 3.3 GHz</td>
<td>650 MHz up to 1300 MHz</td>
<td>55</td>
<td>W</td>
<td>1, 2, 7</td>
</tr>
<tr>
<td></td>
<td>LFM</td>
<td>800 MHz</td>
<td>650 MHz up to 1300 MHz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quad Core SV</td>
<td>HFM</td>
<td>2.2 GHz up to 3.4 GHz</td>
<td>650 MHz up to 1300 MHz</td>
<td>45</td>
<td>W</td>
<td>1, 2, 7</td>
</tr>
<tr>
<td></td>
<td>LFM</td>
<td>800 MHz</td>
<td>650 MHz up to 1300 MHz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dual Core SV</td>
<td>HFM</td>
<td>2.5 GHz up to 3.4 GHz</td>
<td>650 MHz up to 1300 MHz</td>
<td>35</td>
<td>W</td>
<td>1, 2, 7</td>
</tr>
<tr>
<td></td>
<td>LFM</td>
<td>800 MHz</td>
<td>650 MHz up to 1300 MHz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Voltage</td>
<td>HFM</td>
<td>2.1 GHz up to 3.2 GHz</td>
<td>500 MHz up to 1100 MHz</td>
<td>25</td>
<td>W</td>
<td>1, 2, 7</td>
</tr>
<tr>
<td></td>
<td>LFM</td>
<td>800 MHz</td>
<td>500 MHz up to 1100 MHz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ultra Low Voltage</td>
<td>HFM</td>
<td>1.4 GHz up to 2.7 GHz</td>
<td>350 MHz up to 1000 MHz</td>
<td>17</td>
<td>W</td>
<td>1, 2, 7</td>
</tr>
<tr>
<td></td>
<td>LFM</td>
<td>800 MHz</td>
<td>350 MHz up to 1000 MHz</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

New concept: thermal capacitance

**Classic Model**
Steady-State Thermal Resistance
Design guide for steady state

**New Model**
Steady-State Thermal Resistance
PG and CPU sharing
AND
Dynamic Thermal Capacitance

Example:
Cp_{Al} \sim 0.9 \text{ J/(g \cdot \degree K)}
100\text{gr} heat sink heated by
35W CPU \rightarrow 100\text{Sec}

More realistic response to power changes
New concept: thermal capacitance

**Classic Model**
Steady-State Thermal Resistance
Design guide for steady state

**New Model**
Steady-State Thermal Resistance
PG and CPU sharing
AND
Dynamic Thermal Capacitance

- Managing of energy budget rolling average
- Heat sink capacity time constant – few sec.
- Short time constants for power delivery

\[ E_{n+1} = \alpha E_n + (1 - \alpha) \times (TDP_n - P_n) \Delta t_n \]

- Package energy sharing between CPU and PG
- Multiple sources of controls
- Software or external embedded controller

PCU manages energy budgets over multiple time constants
Accumulated energy during idle period used when needed
After idle periods, the system accumulates “energy budget” and can accommodate high power/performance for up to a minute.

In Steady State conditions the power stabilizes on TDP, possibly at higher than nominal frequency.

Use accumulated energy budget to enhance user experience.
Usage Scenario: Responsive Behavior

- Interactive work benefits from Intel® Turbo Boost 2.0
- Idle periods intermixed with user actions

Photo editing
- Open image
- Process
- View
- Balance colors
- Red eye removal
- Contrast
- Filters
- Etc.
Turbo controls in action

Actual instantaneous power

Voltage Regulator reported capability

CURRENT_CONFIG_CONTROL MSR

P-state
Power

Hard Limit
Max Icc

Power limit 2
Power delivery

Power limit 1
Config TDP

C0 P0

Power Limit 2

Time
0.001-64 Sec

Power Limit 1

Time

PL1 time exp. average

Allow programmability

TURBO_POWER_LIMIT Control MSR

- Enables and locks
- Package Power limit 2 – Instantaneous
- Package Power limit 1 Time interval
- Package Power limit 1 clamp bit
- Package Power limit 1 - power

- Also:
  - Individual power controls available
  - Explicit frequency control

User / OEM / OS preference
Power specification is defined for the entire package
- Monolithic die – power budget shared by CPU and PG
- Sum of component power at or below specifications
Power specification is defined for the entire package
- Monolithic die – power budget shared by CPU and PG
- Sum of component power at or below specifications

Energy budget split dynamically according to user preference
- Control algorithm translates energy headroom to turbo bins

<table>
<thead>
<tr>
<th>Application</th>
<th>Core Power [W]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU preference</td>
<td>Power budget is assigned to the CPU</td>
</tr>
<tr>
<td>PG preference</td>
<td>Power budget is assigned to the CPU</td>
</tr>
<tr>
<td>Balanced</td>
<td>Power budget split between CPU and PG</td>
</tr>
</tbody>
</table>

IA preference MSR
GT preference MSR
Turbo in action – measurements

- Four core 45W 2.2 up to 3.5 GHz Sandy Bridge example
- Running CPU and PG simultaneous workloads
- Control power management knobs on the fly using a control utility

After idle period turbo to 56W for ~20Sec - stabilize at TDP = 45W
Frequency varies
Energy Efficient P-State - optimizing MIPS / Watt

- Frequency voltage scaling up is not energy efficient
  - Cubic increase in power for linear increase in frequency and performance
  - Used to get raw performance at the cost of increased energy consumption

- Not all workloads gain performance from frequency
  - For example – many memory accesses → poor performance scalability
  - “Wait slowly” → lower frequency at memory bound intervals
    - Save energy to be used for core bounded phases
    - Or just save energy with minimal performance impact

- Continuously generate “scalability” metric
  - Drop frequency if scalability is low

- User preference control
  - Max performance – ignore energy cost
  - Balanced – lower frequency at memory-bound intervals
  - Max energy savings – limited turbo

Impacts active energy - Small impact on battery life
Average Power Management
Sandy Bridge average power control

**Core Level**
- HW coordinated per-thread interface
- Only snoops supported
- Core caches flushed Vcc-gated

**Ring + LLC**
- HW coordinated Clock off + low-VCC
- Retention voltage
- LLC Flushed Usage based close/open algorithms

**System-Agent**
- Pop-up: DDR-Self refresh
- C3/6/7: DDR clock off, IO clock off Display-Engine in energy efficient screen refresh mode

**Thread level coordination**
- Core C0
  - Core C1
  - Core C3
  - Core C6
- Package C3
- Package C6
- Package C7

**Active states**
- P0 - P1 - Pn

**Thread level coordination**
- Core C0
  - Core C1
  - Core C3
  - Core C6
- Package C3
- Core C6
- Package C7
**Improved C-state Latency and energy efficiency**

“Interrupt storms” seen on real systems

- **Performance Impact**
  - Entry and exit latency

- **Energy Impact**
  - transition power and energy overhead

**Auto-demotion:**
8-15% perf (MM07, Sysmark)

**Auto-un-demotion:**
Aggressive Demotion-enable!

45-200mW power savings measured on Sysmark and media applications
Thermal management
Package thermal management

- On die thermal sensors
  - 12 sensors on each CPU core + PG, ring and SA
  - Operating range 50-100°C
- Temperature reporting
  - Maximum reading of each functional block and maximum reading of the total chip
- Used for:
  - Critical thermal protection
    - Notification, throttle and shutdown
    - Programmable throttle temperature
  - Leakage calculation of power meter
  - PCU optimization algorithms
  - External system controls (e.g. Fan control)
System thermal management

- Digital DDR power meter for thermal prediction
  - Count DDR read and write and calculate power
  - Maximum bandwidth control to prevent critical heating
  - Initiates double DDR refresh rate at high temperature

- Supports DDR thermal sensor
  - For a more accurate DDR temperature reading

- Voltage Regulator thermal sensing
  - Hot and critical conditions using in and out of band communication

- Digital package temperature reporting
  - Used by external agent for system fan control
Power efficient memory controller

- **DDR power management**
  - Aggressive DDR power savings policies, configurable by PCU
    - Normal power down
    - Pre-charge Power down
    - PLL off

- **Self Refresh**
  - Configurable policies for entering Self Refresh, based on package power states, controlled by PCU

- **IO clock controls – power down**
Platform power management
Platform power management - SVID

- SVID – Serial Voltage ID
  - New serial bus to control external Voltage Regulators
  - Three wires serial bus – control multiple VRs
  - Control VR voltage – continues fine grain optimization
    - Optimize voltage for changing conditions
- Optimize VR power savings mode - minimize power losses
  - Power States to optimize VR efficiency
  - A function of current consumption and sleep states
- Read VR parameters for PCU algorithms use
  - Load line resistance, max Icc and temperature
Platform power management – PECI

- PECI – A new platform control interface
  - Connects the PCU to external embedded controller
  - Report - PCU communicates out to the embedded controller:
    - Individual component and max package temperature
    - Individual and total package energy consumption
    - Other power management status information
    - Used for fan control and plat
  - Control:
    - Package power – instantaneous and sustain (PL1-PL2)
    - Other power management settings and preferences
    - Used by Embedded Controller to manage total system power management
Summary and conclusions

Sandy Bridge is built to maximize user experience under constraints

- Throughput performance – Turbo over long time window
- Responsiveness – Turbo dynamically for short duration
- User guided CPU / PG performance balancing
- Battery life / Energy bills – Tight control of active and idle power states
- Rich set of control available for S/W, operating system and system embedded controller allow:
  - User preferences where tradeoff exists
  - Enables small form factor platforms
  - Improved ergonomics - lower acoustic noise and heat
## Turbo roadmap evolution

<table>
<thead>
<tr>
<th>Mobile Desktop</th>
<th>Merom/Penryn (Mobile only)</th>
<th>Nehalem/Westmere</th>
<th>Arrandale</th>
<th>Sandy Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control</strong></td>
<td>• CPU Core C-state</td>
<td>• CPU Core C-states</td>
<td>• CPU Core C-states</td>
<td>• CPU Core C-states</td>
</tr>
<tr>
<td></td>
<td>• Digital power meter</td>
<td>• CPU Power - Platform iMon</td>
<td>• CPU Power - Platform iMon</td>
<td>• CPU/ PG/ Package power</td>
</tr>
<tr>
<td><strong>Key New Capabilities</strong></td>
<td>• 1-2 turbo bin when other core is asleep</td>
<td>• Turbo controlled within power limit</td>
<td>• PG dynamic frequency</td>
<td>• HW controlled power sharing between CPU - PG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Multi-core turbo</td>
<td>• Driver controlled power sharing between CPU and PG (Mobile)</td>
<td>• Brief turbo above TDP → dynamic Turbo</td>
</tr>
</tbody>
</table>

### Turbo Behavior

- **Quad Core Die**
  - Single Core Turbo
  - Dual Core Turbo
  - Quad Core Turbo

- **Dual Core Die**
  - Single Core Turbo
  - Dual Core Turbo
  - Graphics Turbo

- **Dual Core Die**
  - Quad Core Die

*Illustrative only. Does not represent actual number of turbo bins.*