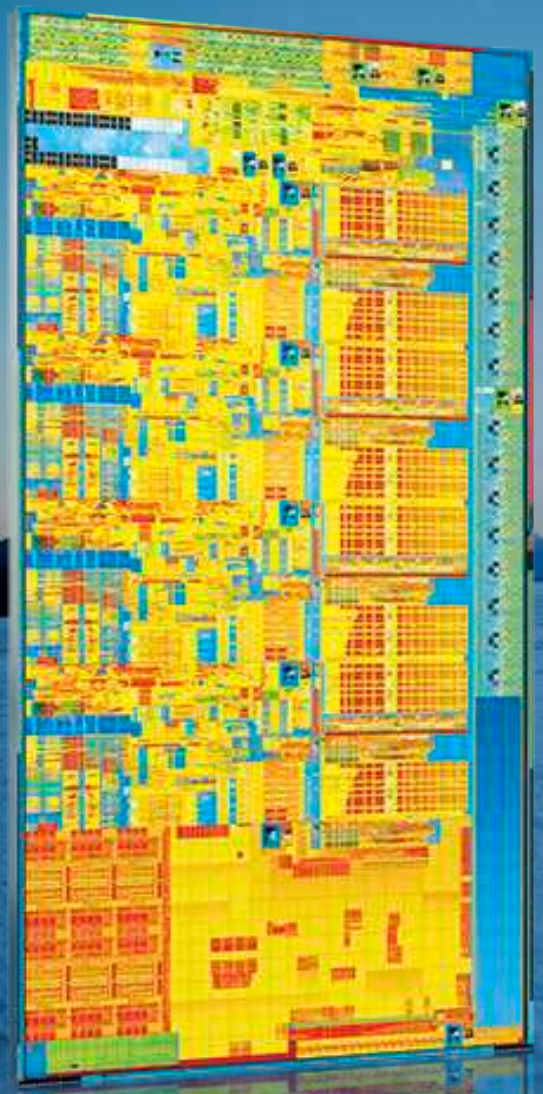
Power management architecture of the 2nd generation Intel® Core™ microarchitecture, formerly codenamed Sandy Bridge



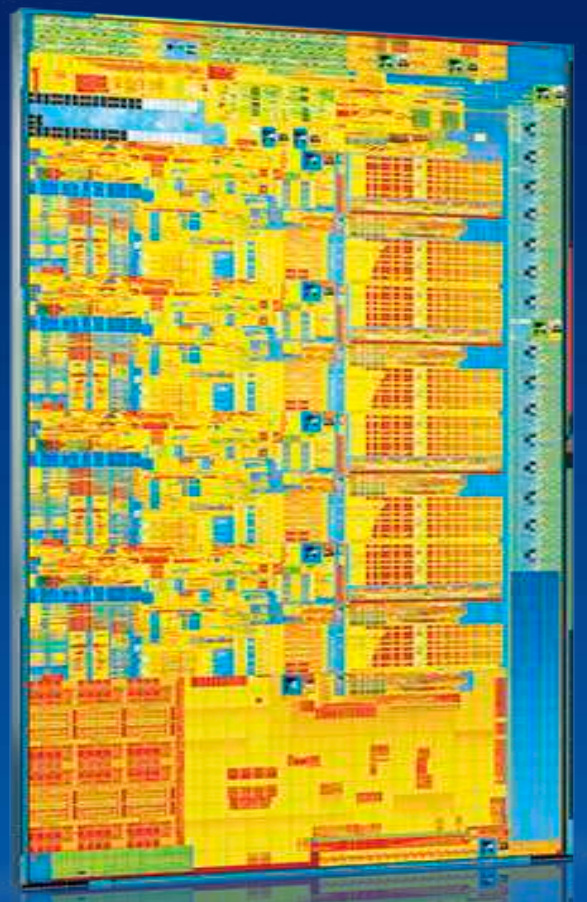
Efi Rotem - Sandy Bridge power architect

Alon Naveh, Doron Rajwan,
Avinash Ananthakrishnan, Eli Weissmann

Hot Chips Aug-2011

Agenda

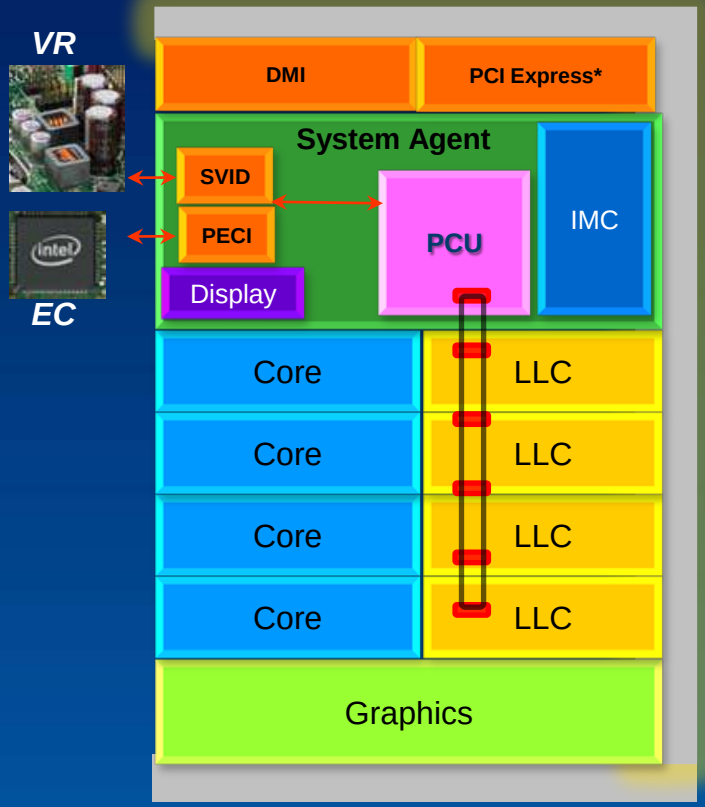
- Power management overview
- Intel® Turbo Boost Technology 2.0
- Thermal management
- Energy efficiency
- Average power management
- Platform view
- Summary



High CPU and PG performance
Power and energy efficiency

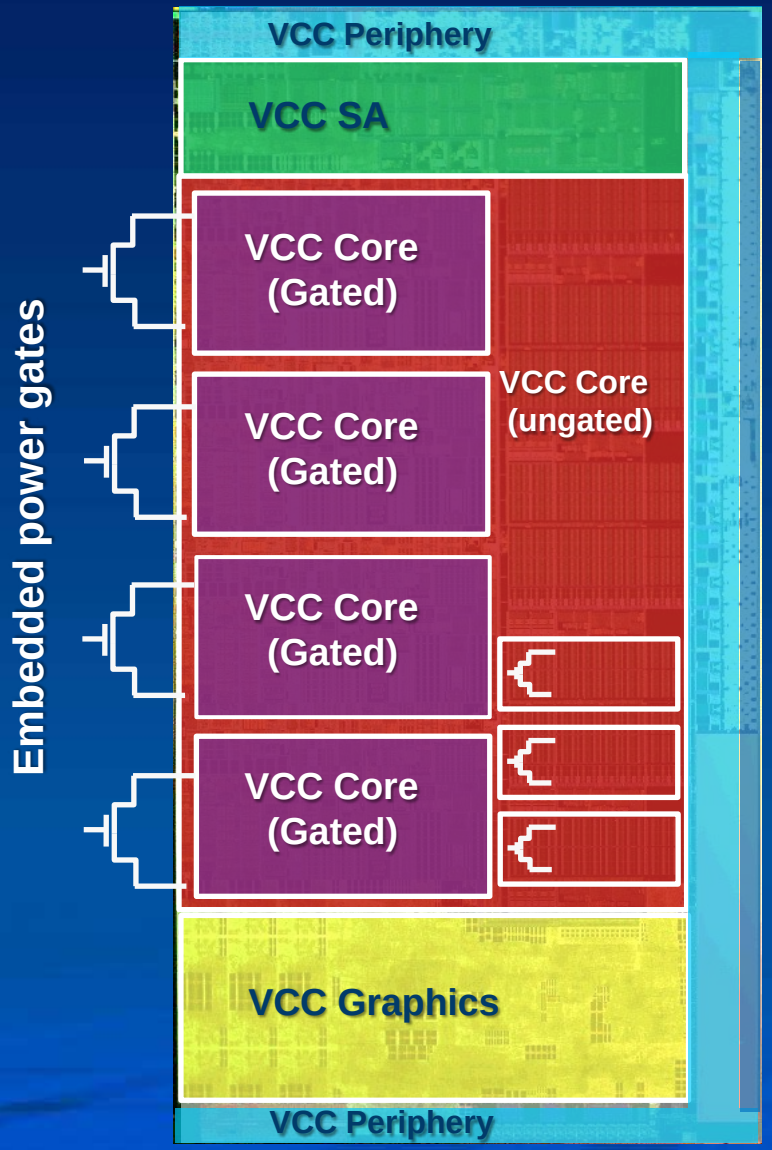
Power management overview

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



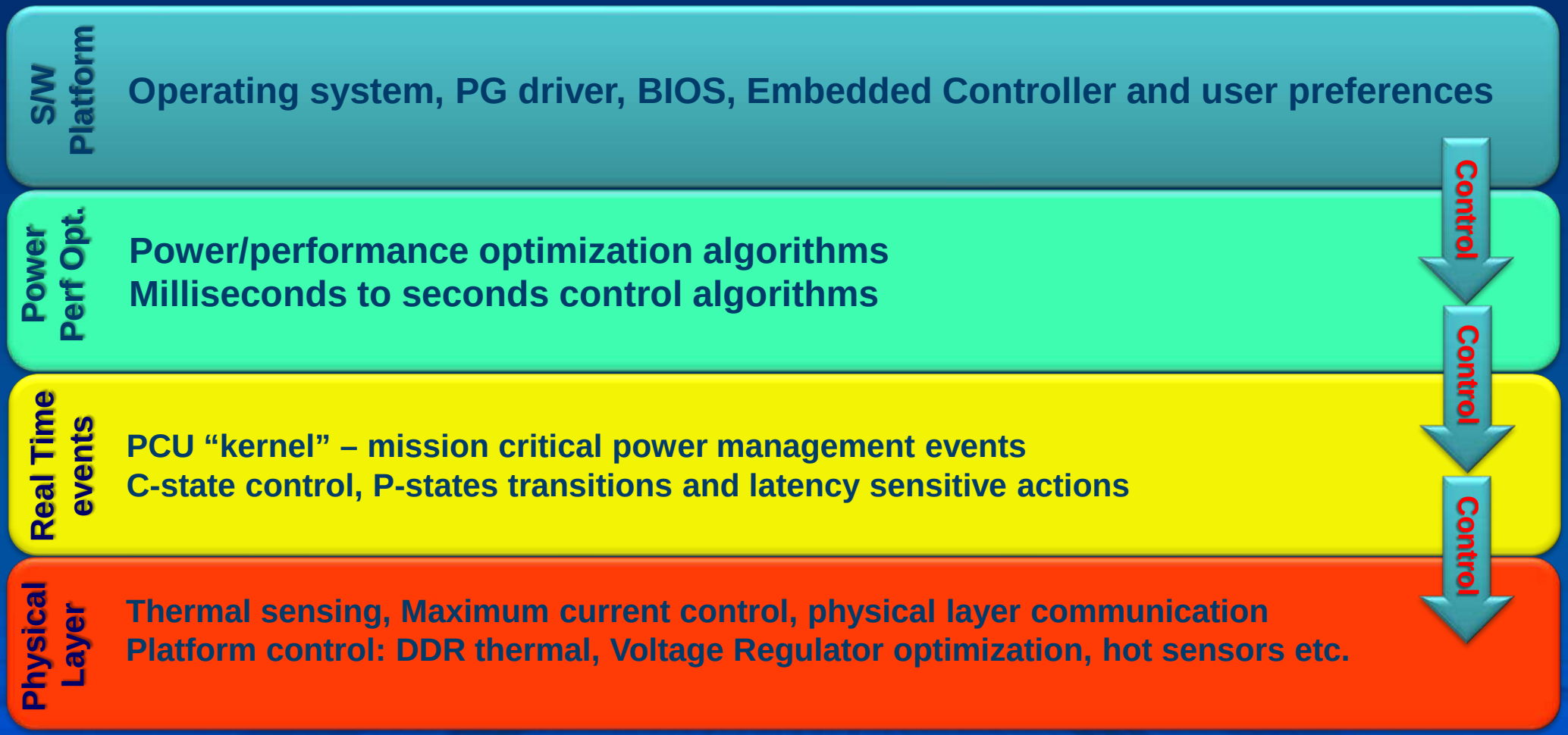
- **Two Independent Variable Power Planes:**
  - CPU cores, ring and LLC
    - Embedded power gates - Each core can be turned off individually
    - Cache power gating - Turn off portions or all cache at deeper sleep states
  - Graphics processor
    - Can be varied or turned off when not active
- **Shared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

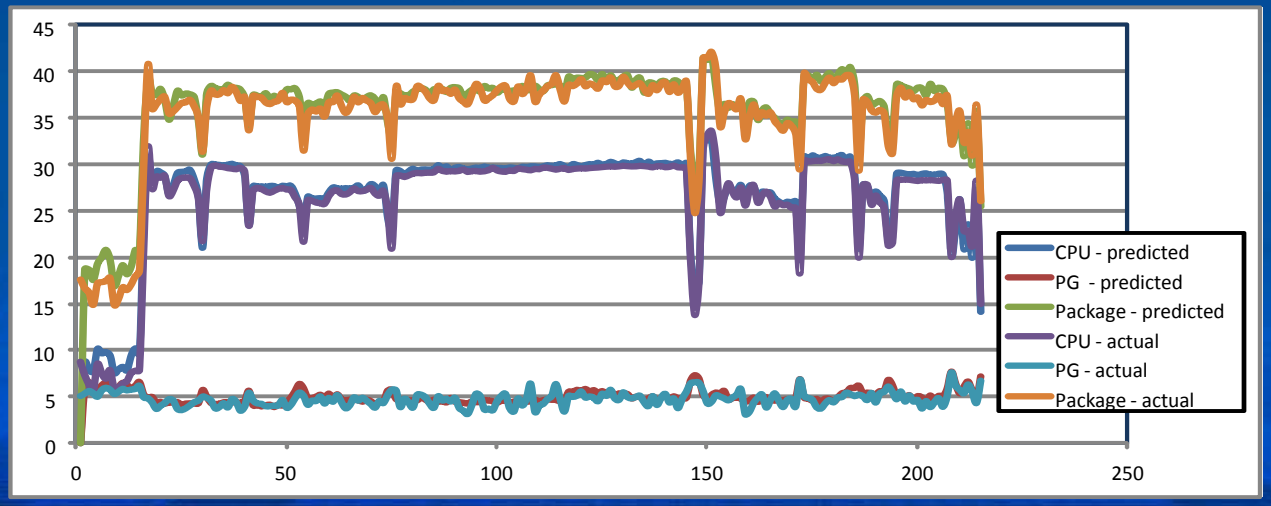
Power management features topology



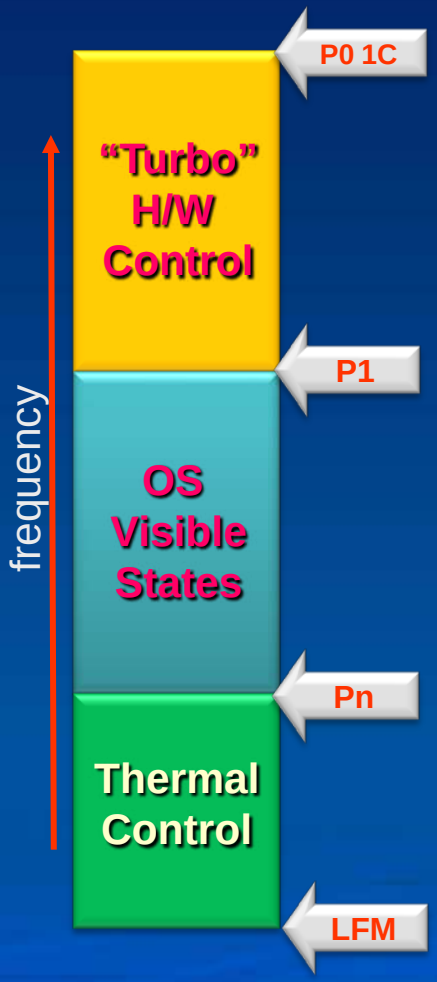
Intel® Turbo Boost Technology 2.0

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

CPU P1 P0 PG P1 P0 TDP total package sustained power

Table 5-1. TDP Specifications

Segment	State	CPU Core Frequency	Processor Graphics Core frequency	Thermal Design Power	Units	Notes
Extreme Edition (XE)	HFM	2.5 GHz up to 3.5 GHz	650 MHz up to 1300 MHz	55	W	1, 2, 7
	LFM	800 MHz	650 MHz up to 1300 MHz	36		
Quad Core SV	HFM	2.2 GHz up to 3.4 GHz	650 MHz up to 1300 MHz	45	W	1, 2, 7
	LFM	800 MHz	650 MHz up to 1300 MHz	33		
Dual Core SV	HFM	2.5 GHz up to 3.4 GHz	650 MHz up to 1300 MHz	35	W	1, 2, 7
	LFM	800 MHz	650 MHz up to 1300 MHz	26		
Low Voltage	HFM	2.1 GHz up to 3.2 GHz	500 MHz up to 1100 MHz	25	W	1, 2, 7
	LFM	800 MHz	500 MHz up to 1100 MHz	12		
Ultra Low Voltage	HFM	1.4 GHz up to 2.7 GHz	350 MHz up to 1000 MHz	17	W	1, 2, 7
	LFM	800 MHz	350 MHz up to 1000 MHz	10		

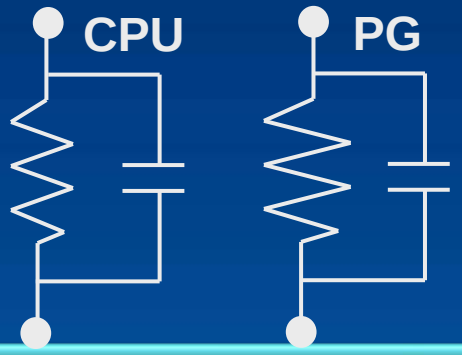
Source: http://www.intel.com/Assets/PDF/datasheet/324692.pdf

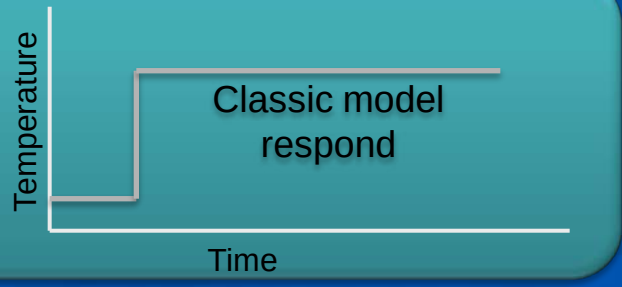
New concept: thermal capacitance

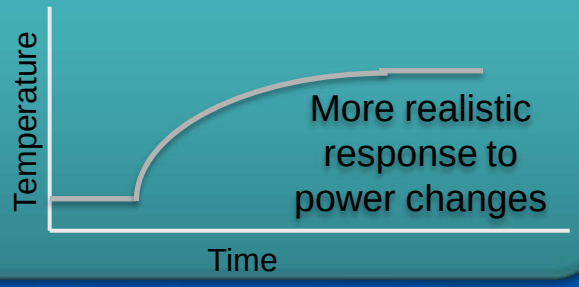
Classic Model
Steady-State Thermal Resistance
Design guide for steady state

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

Example:
Cp_Al ~ 0.9 J/(gr*'K)
100gr heat sink heated by
35W CPU → 100Sec







New concept: thermal capacitance

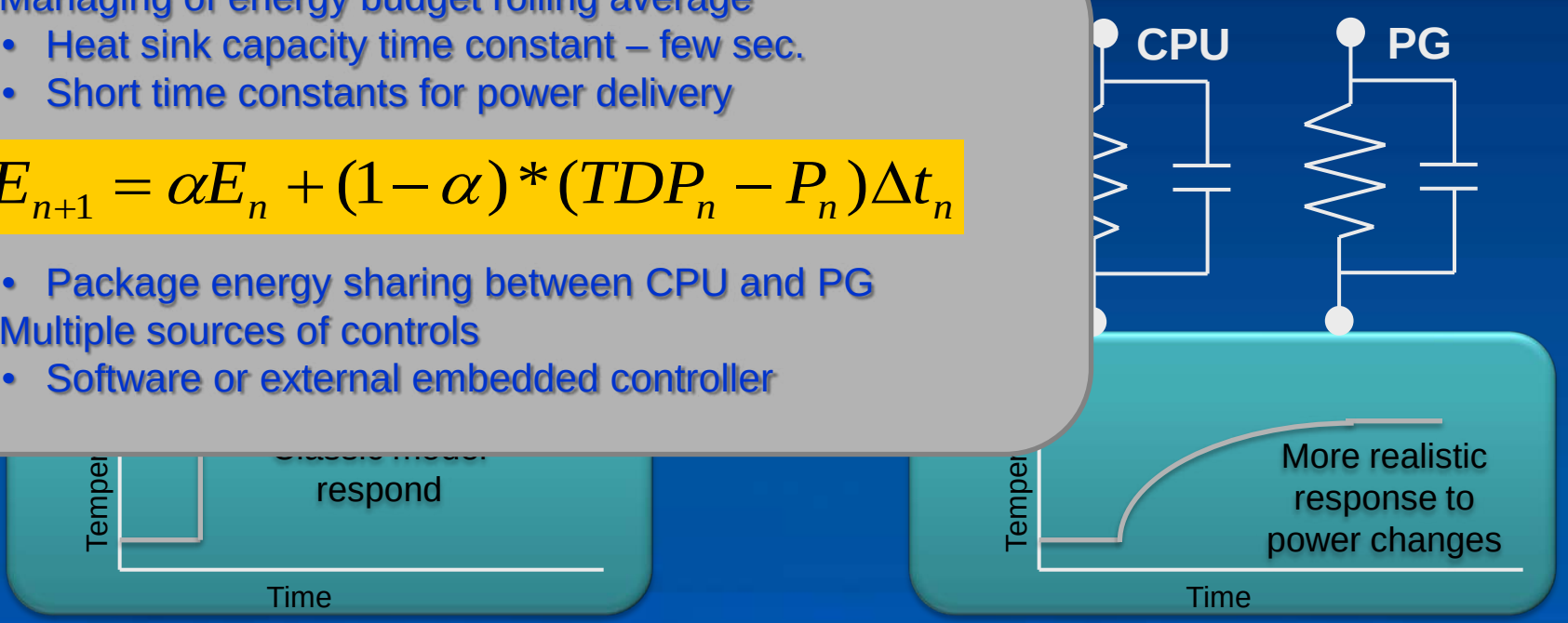
Classic Model
Steady-State Thermal Resistance
Design guide for steady state

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

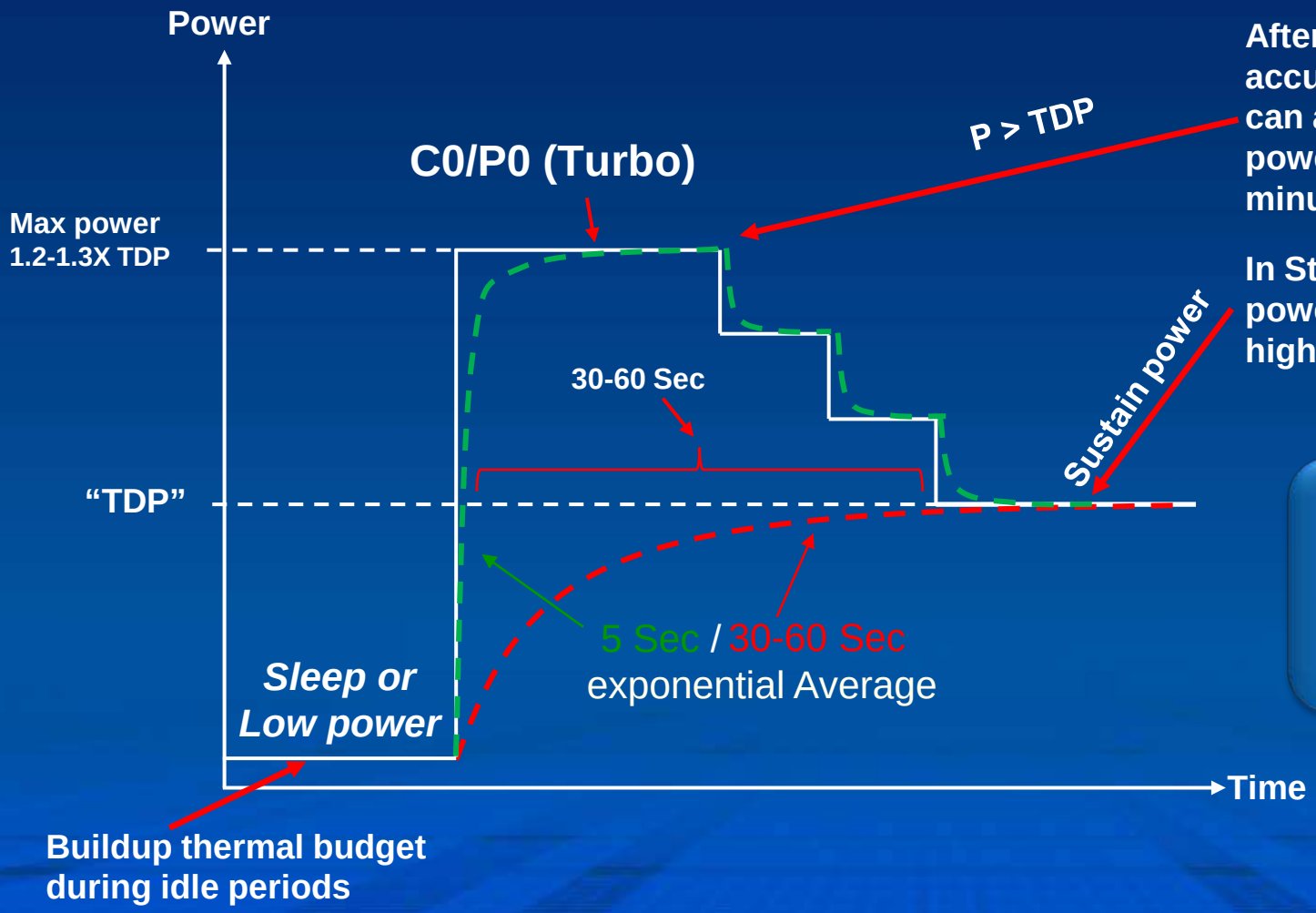
$$E_{n+1} = \alpha E_n + (1-\alpha) * (TDP_n - P_n)\Delta t_n$$

  - 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

Intel® Turbo Boost Technology 2.0 - Dynamic



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 then 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

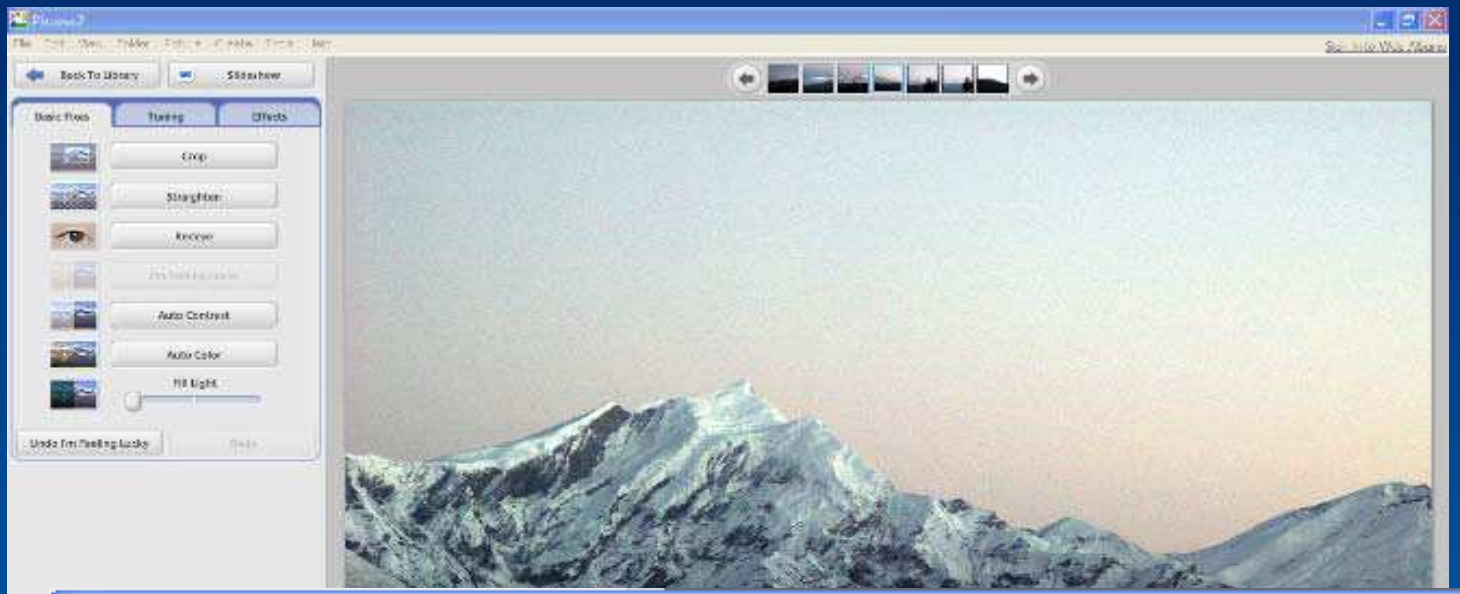
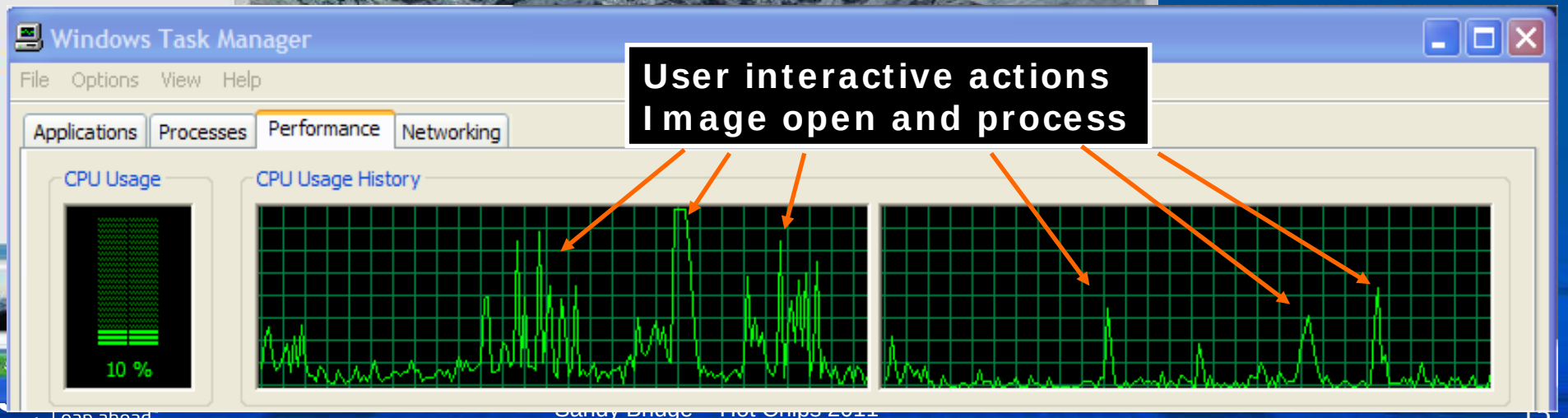
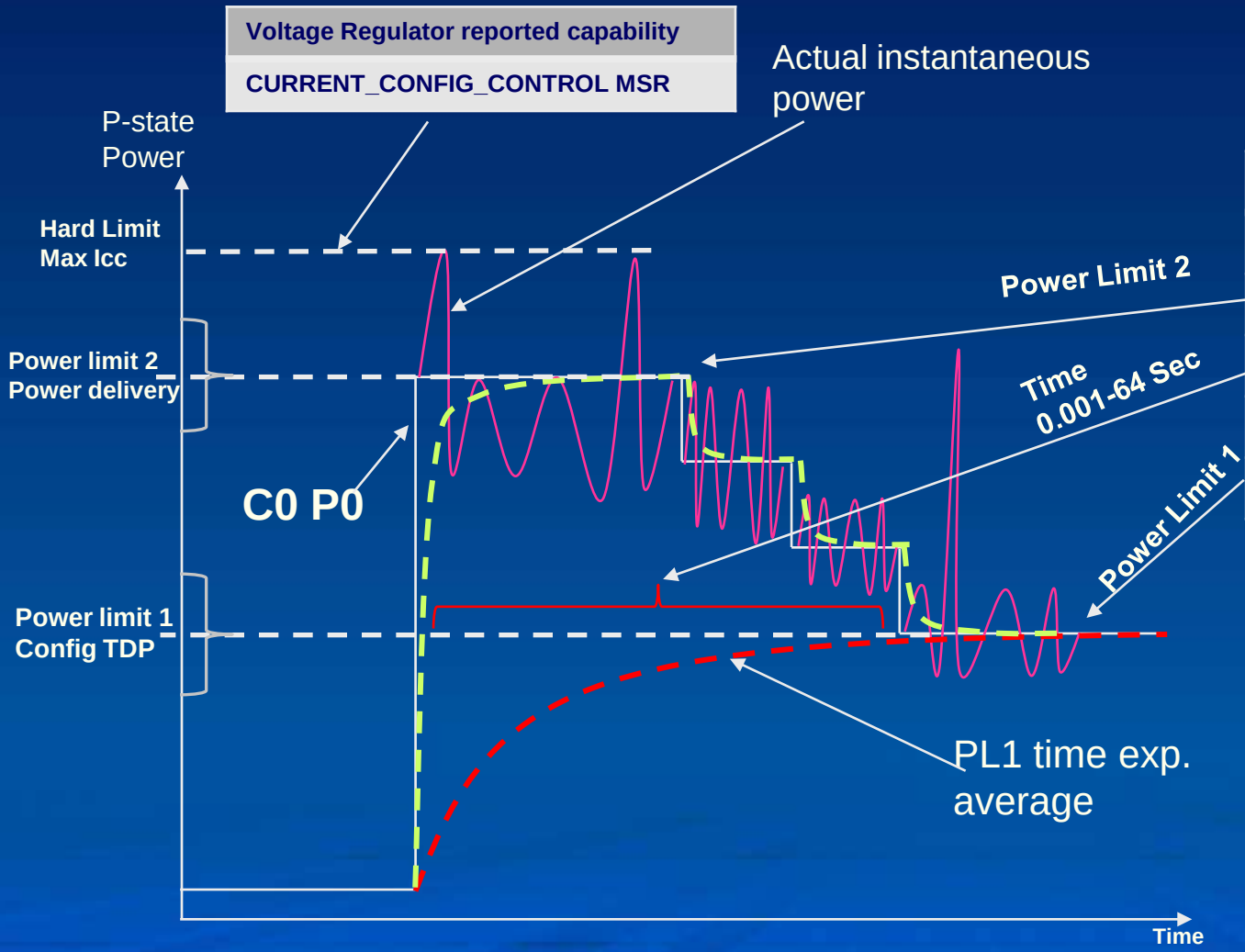


Photo editing

- Open image
- Process
- View
- Balance colors
- Red eye removal
- Contrast
- Filters
- Etc.



Turbo controls in action



Allow programmability

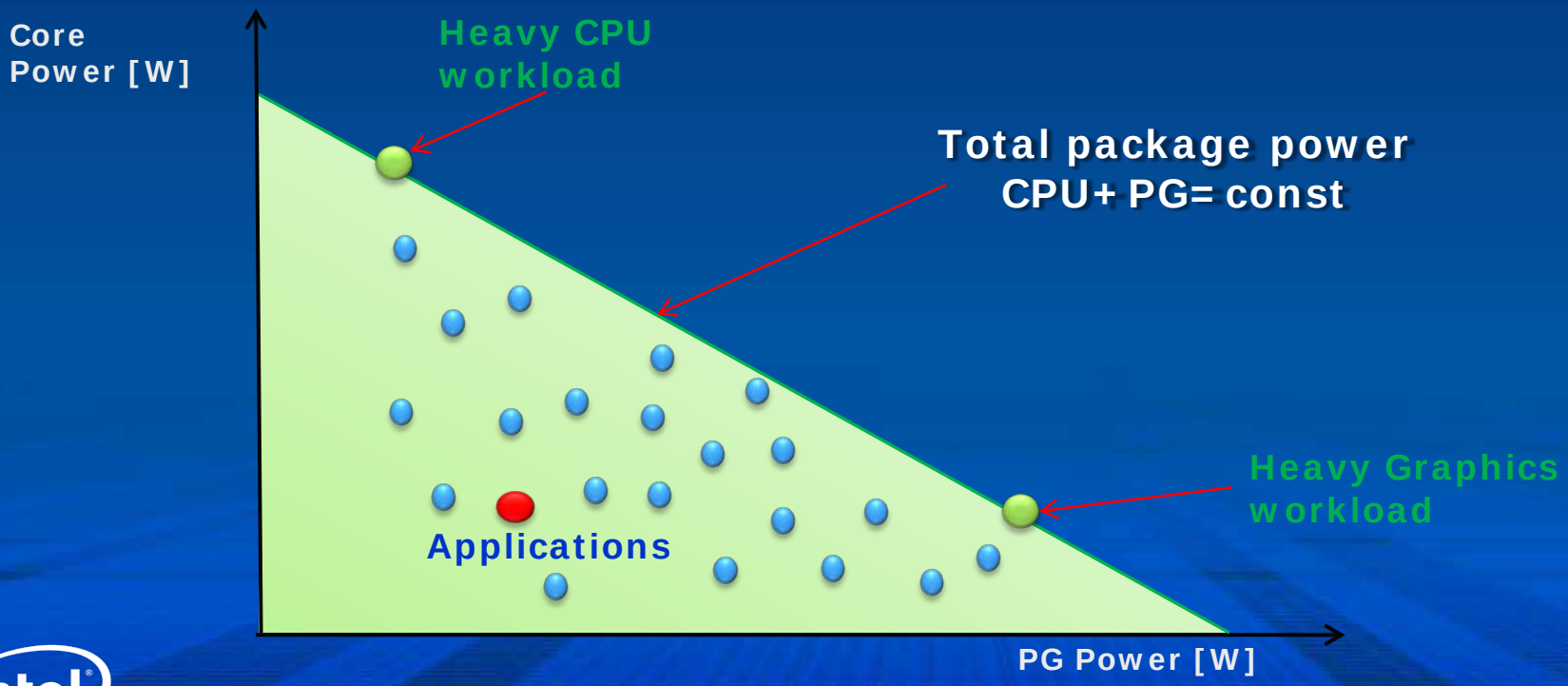
TURBO_POWER_LIMIT Control MSR
Enables and locks
Package Power limit 2 – Instantaneous
Package Power limit 1 Time interval
Package Power limit 1 clamp bit
Package Power limit 1 - power

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

User / OEM / OS preference

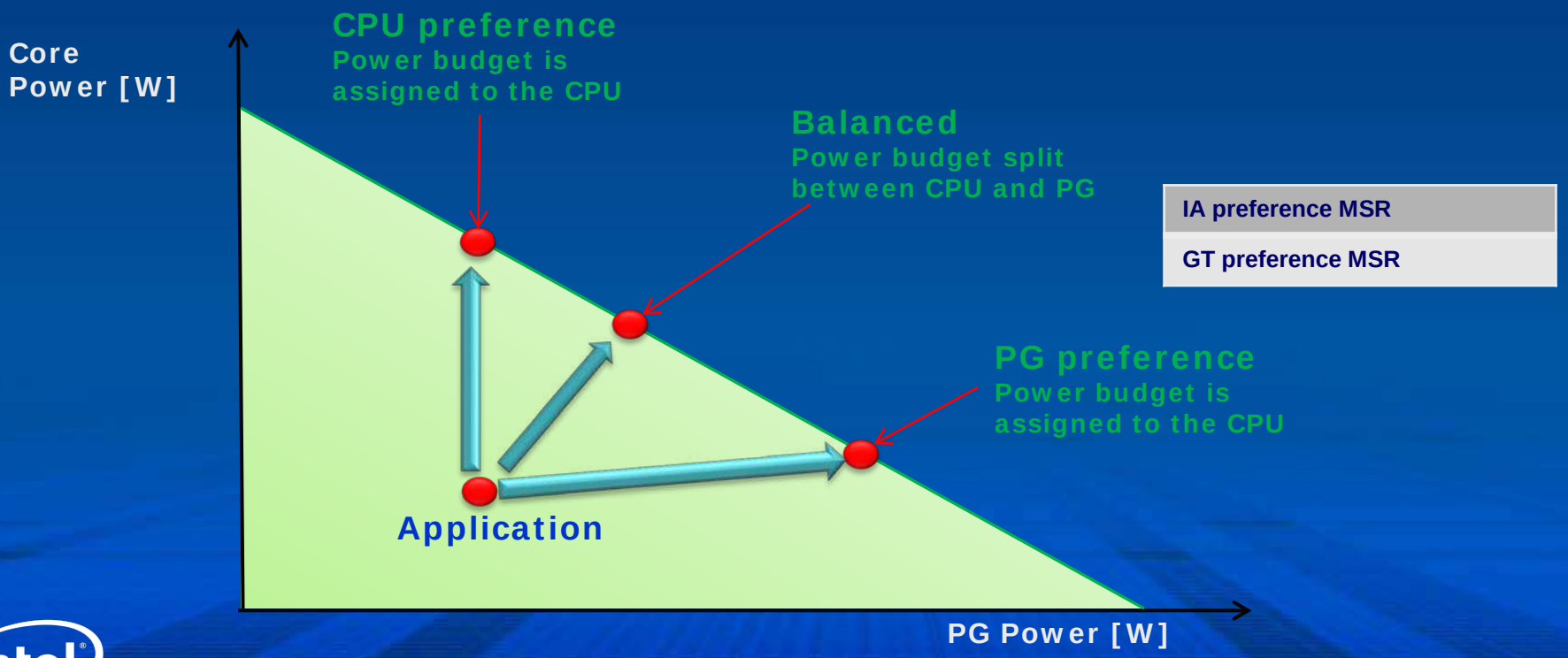
Intel® Turbo Boost Technology 2.0 - Package

- **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



Intel® Turbo Boost Technology 2.0 - Package

- **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 spit dynamically according to user preference**
  - Control algorithm translates energy headroom to turbo bins

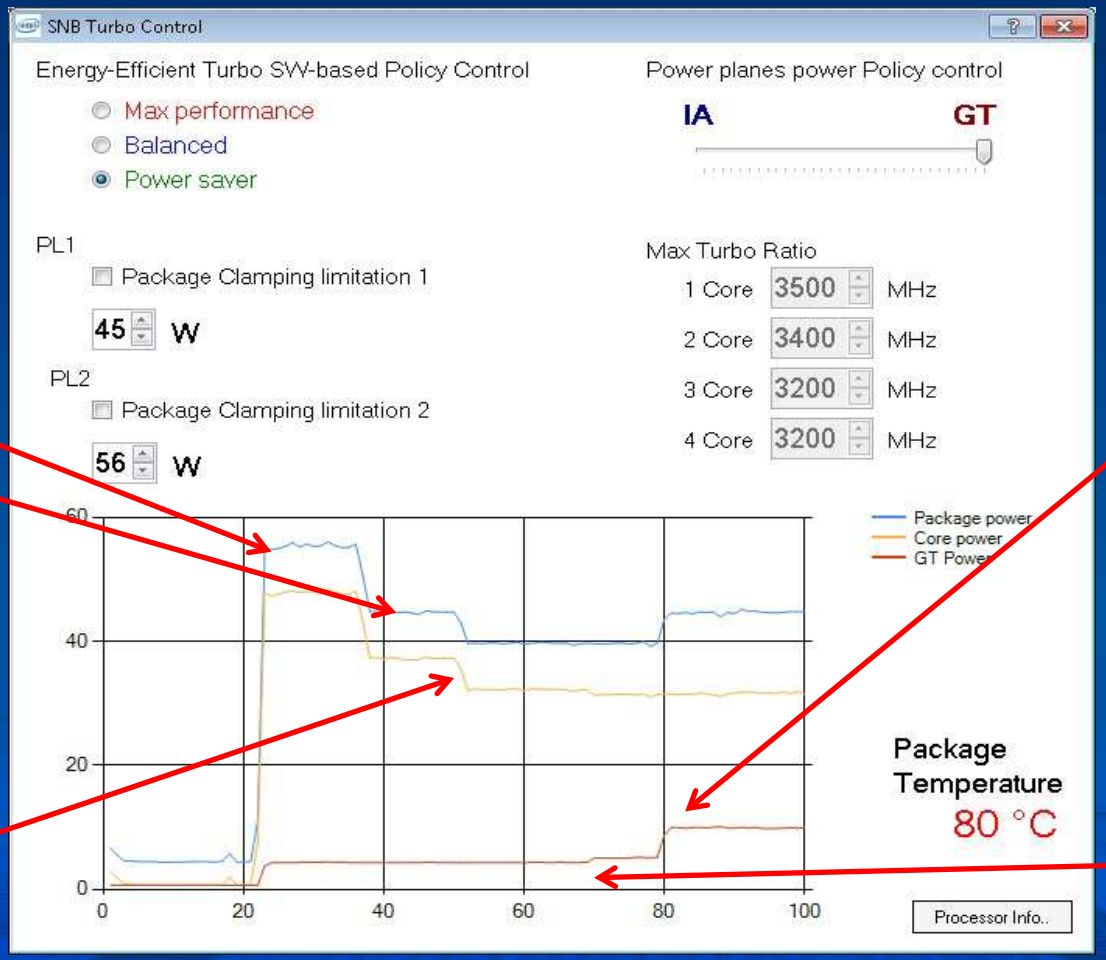


Turbo in action – measurements

- 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

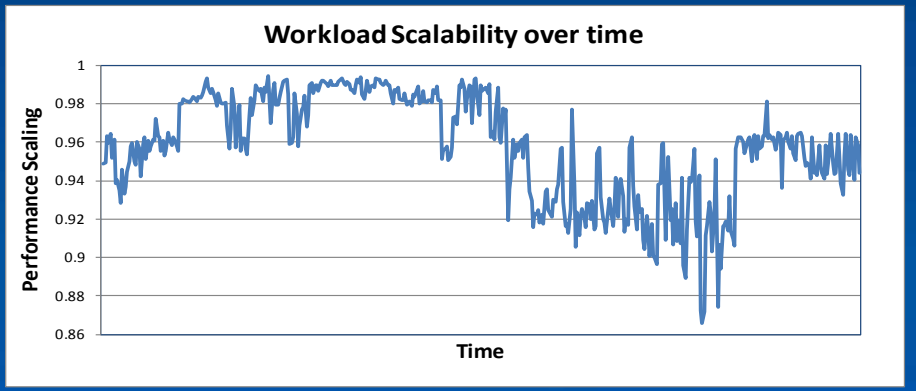
After idle period turbo to 56W for

~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

System-Agent

Pop-up: DDR-Self refresh

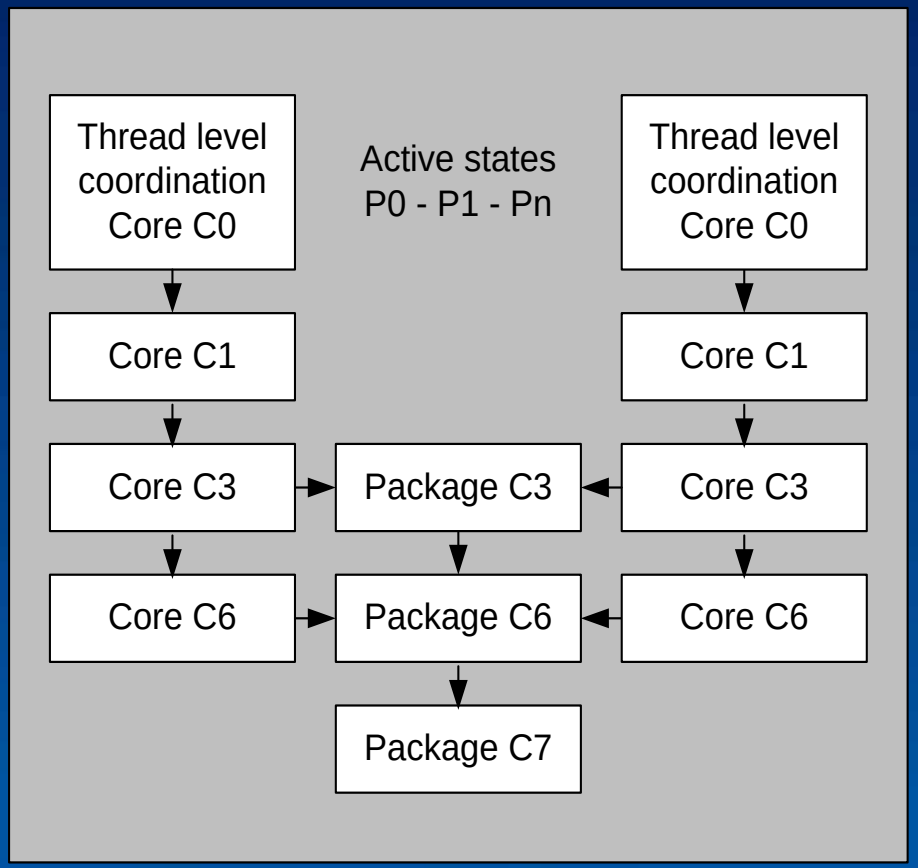
C3/6/7:
DDR clock off, IO clock off
Display-Engine in energy efficient screen refresh mode

Ring + LLC

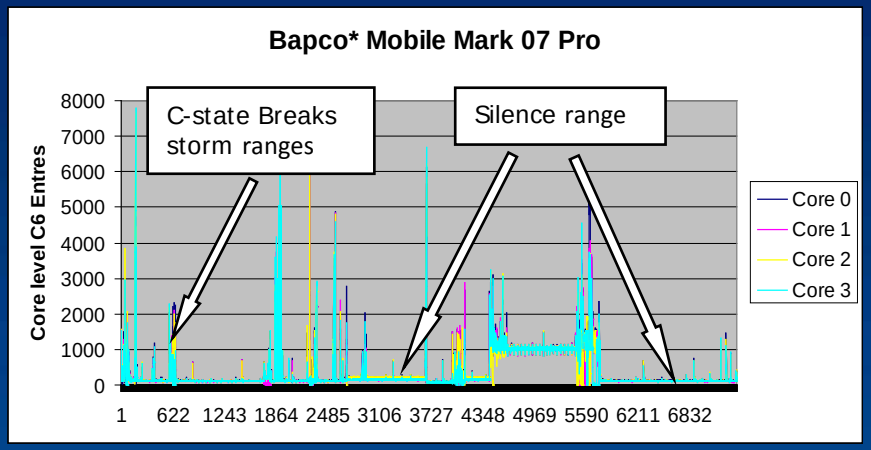
HW coordinated Clock off + low-VCC

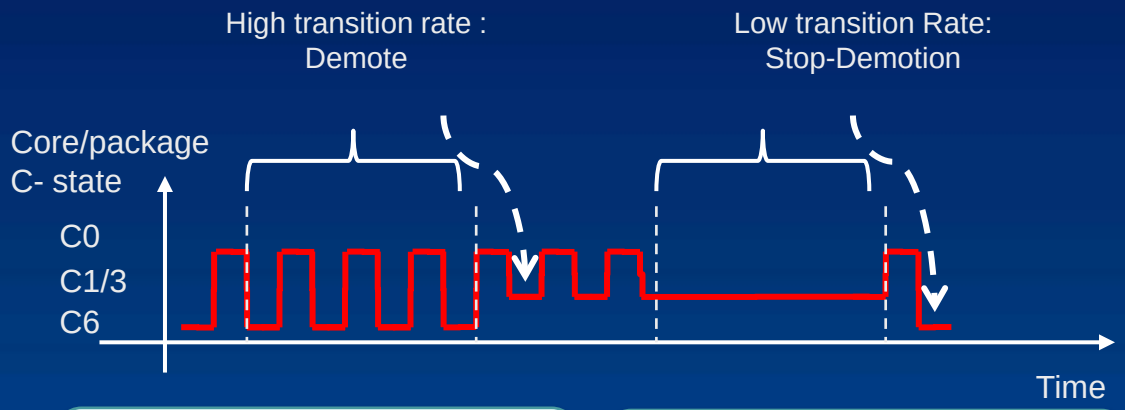
Retention voltage

LLC Flushed
Usage based close/open algorithms



Improved C-state Latency and energy efficiency





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

Auto-un-demotion: Aggressive Demotion-enable!

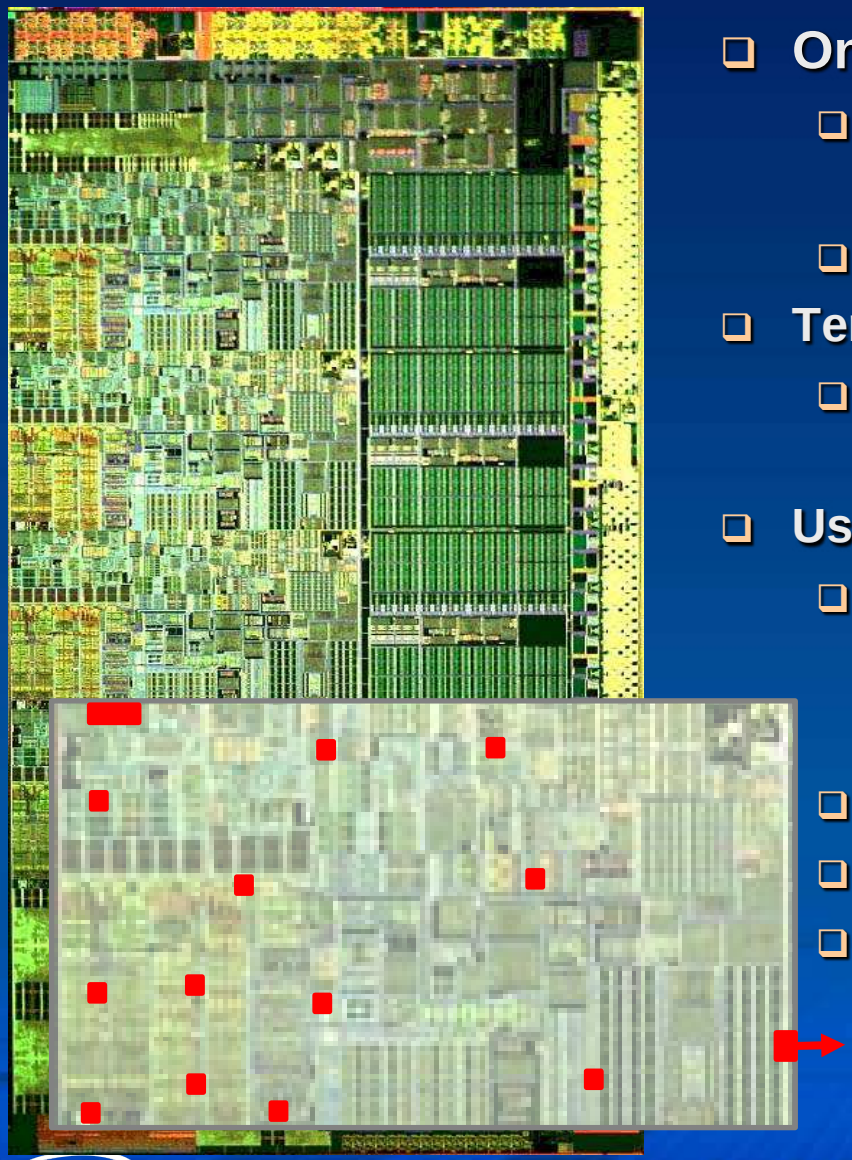
45-200mW power savings measured on Sysmark and media applications

"Interrupt storms" seen on real systems

- Performance Impact
  - Entry and exit latency
- Energy Impact
  - transition power and energy overhead

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


intel®
Leap ahead™

Turbo roadmap evolution

Mobile Desktop	Merom/Penryn (Mobile only)	Nehalem/Westmere		Sandy Bridge
		Clarksfield Lynnfield/Clarkdale	Arrandale	
Control	• CPU Core C-state •Digital power meter	• CPU Core C-states • CPU Power - Platform iMon	• CPU Core C-states • CPU Power- Platform iMon • PG Power- Platform iMon • Package Power	• CPU Core C-states • CPU/ PG/ Package power • Built-in power monitoring •Power Budget Management •Platform Control (EC / VR)
Key New Capabilities	• 1-2 turbo bin when other core is asleep	• Turbo controlled within power limit • Multi-core turbo • More turbo if cores are asleep	• PG dynamic frequency • Driver controlled power sharing between CPU and PG (Mobile)	• HW controlled power sharing between CPU - PG • Brief turbo above TDP → dynamic Turbo •More platform control via PECI 3.0 and SVID
Turbo Behavior Illustrative only. Does not represent actual number of turbo bins.