# Trends in Energy, Power and Thermal Efficiency of HPC systems

Andrea Bartolini, PhD
Università di Bologna & ETH Zurich

http://www-micrel.deis.unibo.it/sitonew/
a.bartolini@unibo.it

http://www.iis.ee.ethz.ch/
barandre@iis.ee.ethz.ch

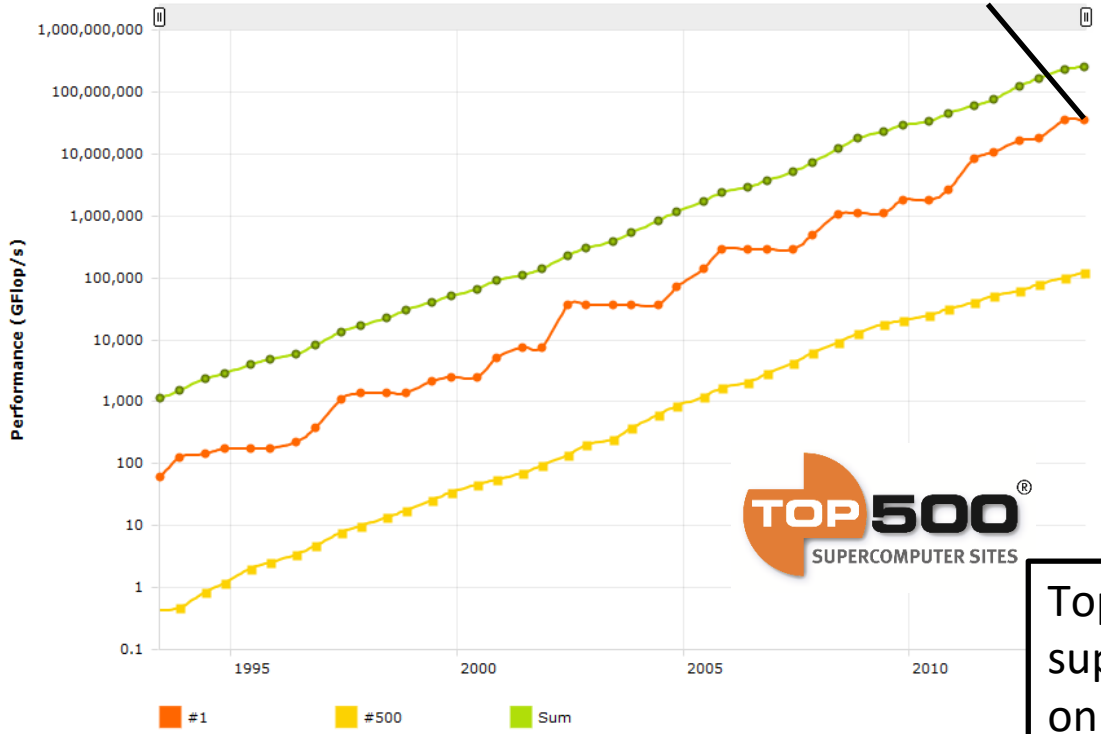
# Supercomputers Trends

- **Performance request increases**
- **Scaling is facing technological walls**
- **Top500 list**

Power, Thermal, Utilizzation Walls

#1 Tihane-2
33.2PFlops @17.8MW
(24MW with cooling)
1.9GFLOPS/W

$R_{max}$ and $R_{peak}$ values are in GFlops. For more details about other fields, che




TOP500 SUPERCOMPUTER SITES


CRAY

FUJITSU

Like 3,258 people like this. Be the first of your friends.

**TOP10 November 2013**

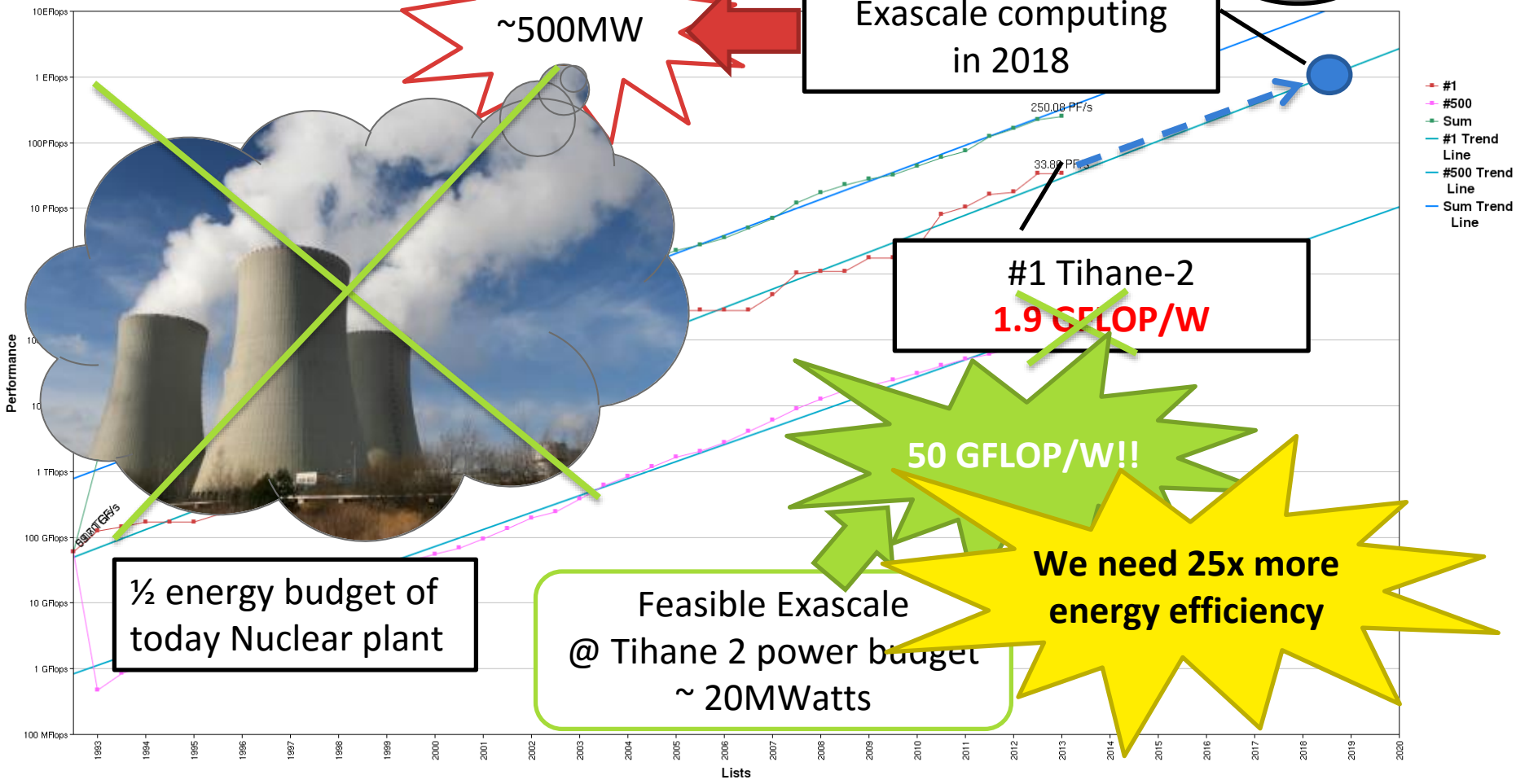
1 Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT

Top500 ranks the new supercomputers by FLOPS on Linpack Benchmark


2

# Supercomputers Trends

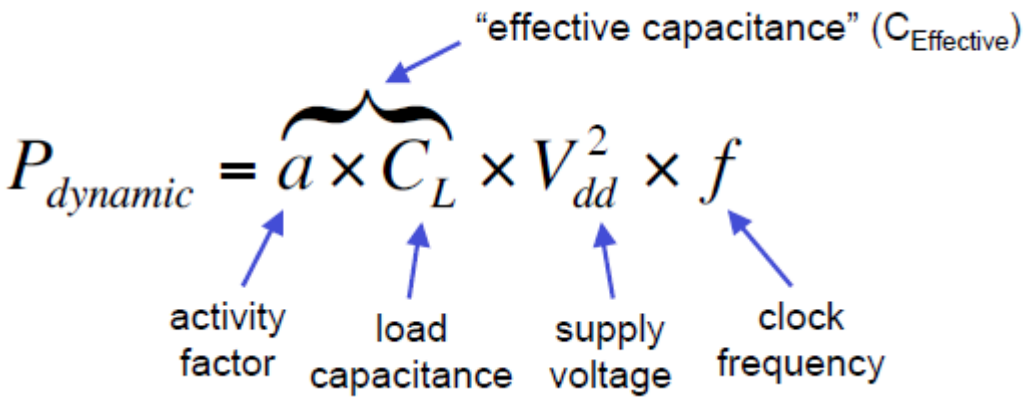
- **TOP500 projections**

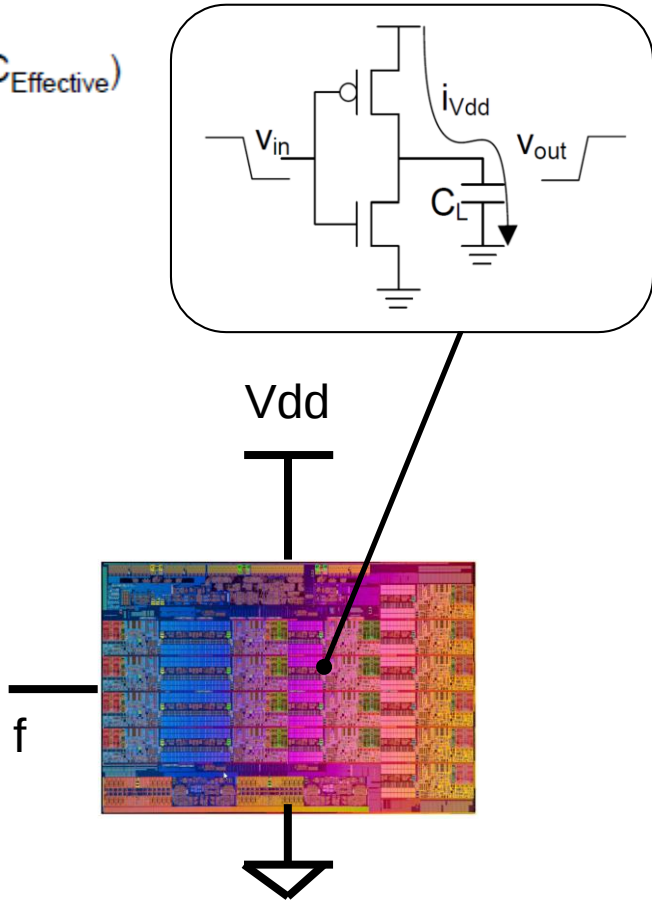


# Outline

- **Power in digital systems**
- Dynamic Power Management
- Power Management & Heterogeneity
- Characterization of thermal effects

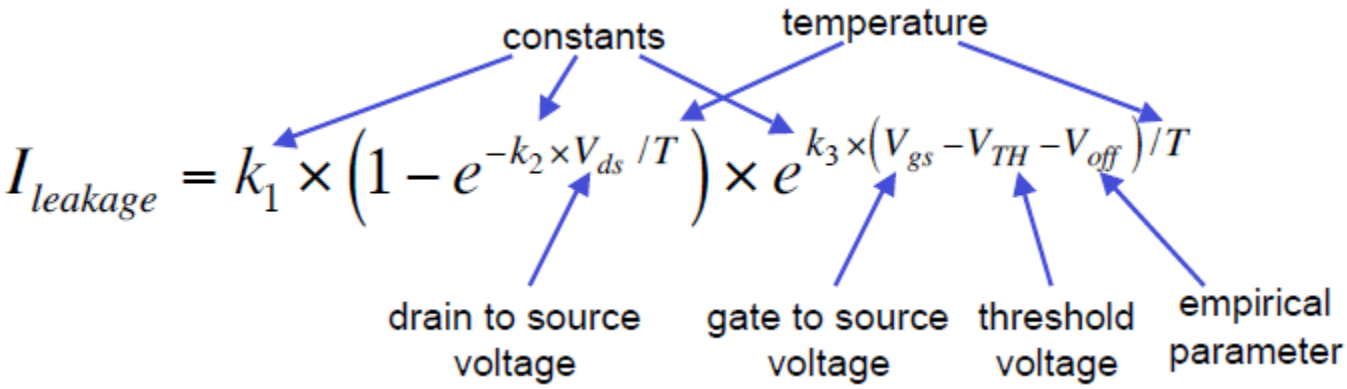
# Dynamic Power




- Linear ↓ with ↓ $C_{Effective}$
- Linear ↓ with ↓ f
- Quadratic ↓ with ↓ $V_{dd}$
- Cubic ↓ with ↓ both $V_{dd}$ and f

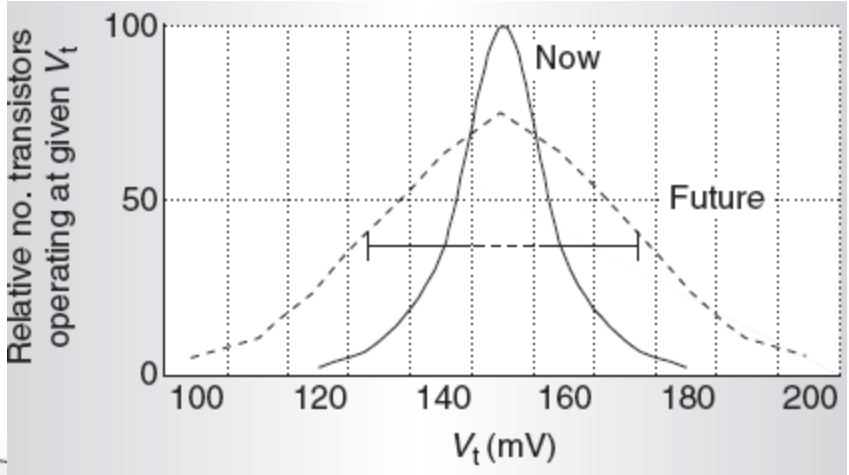
# Sub-threshold Leakage Current

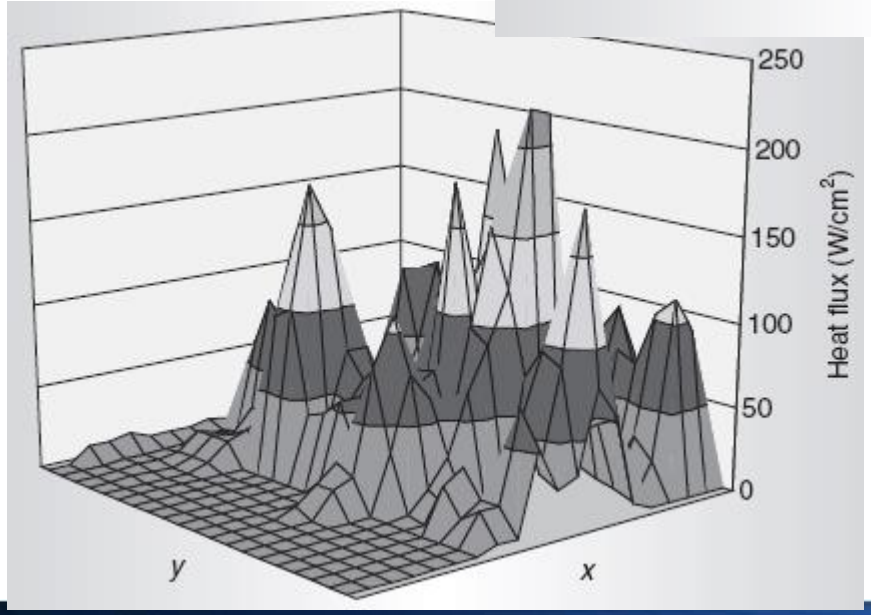


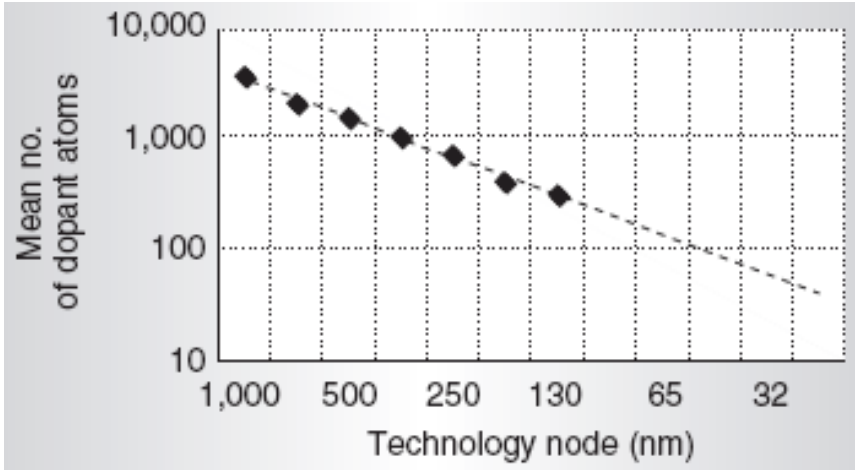

- **Exponential ↓ with ↓ $V_{ds}$**
- **Exponential ↓ with ↑ $V_{TH}$**
- **Exponential ↓ with ↓ T**

# Is the same for all the die?

- Variability:



# Alpha-Power Thermal Model

Delay:

$$D_p = \frac{C_{out}\, V_{dd}}{I_{ON}} = \frac{C_{out}\, V_{dd}}{\mu(T)[V_{dd} - V_{th}(T)]^{\alpha}}$$

Carrier Mobility:

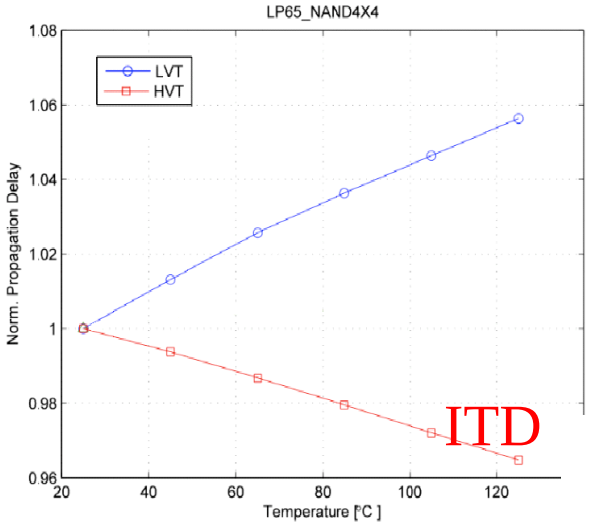
$$\mu(T) = \mu(T_o)\left(\frac{T_o}{T}\right)^m$$

Threshold Voltage:

$$V_{th} = V_{th}(T_0) - k(T - T_0)$$

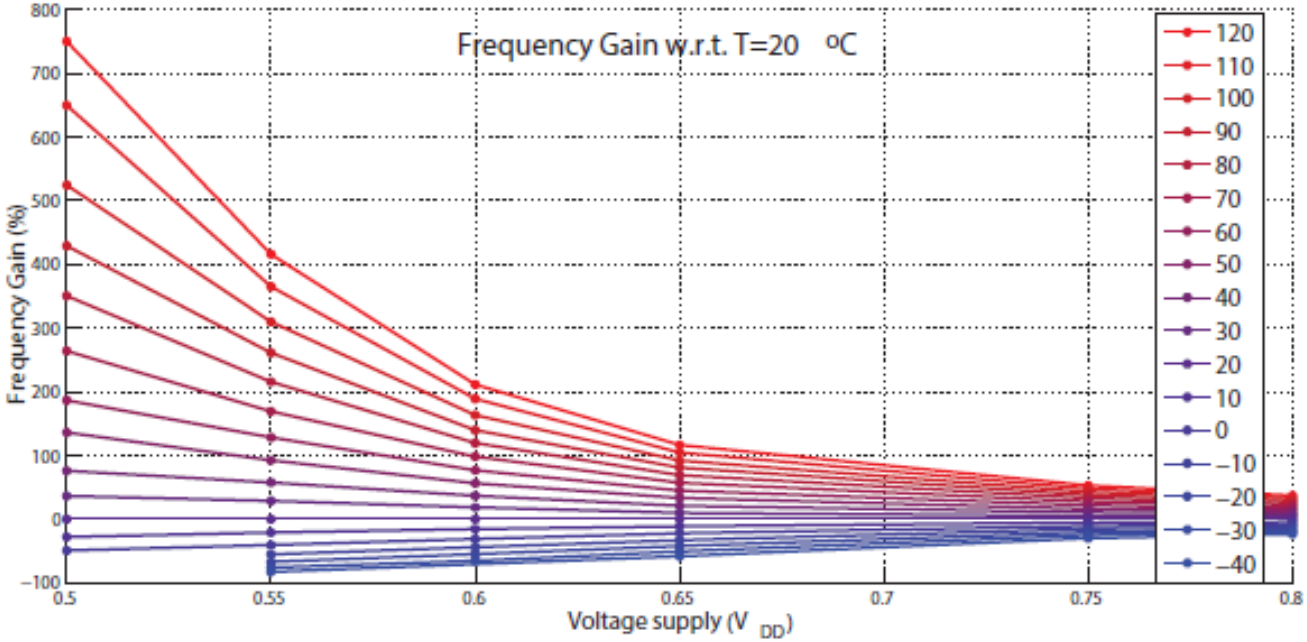
| | | |
|---|---|---|
| $T \uparrow$ | $\mu \downarrow$ | $V_{th} \downarrow$ |

# Thermal Behavior of CMOS gates




8x variation @Low Vdd

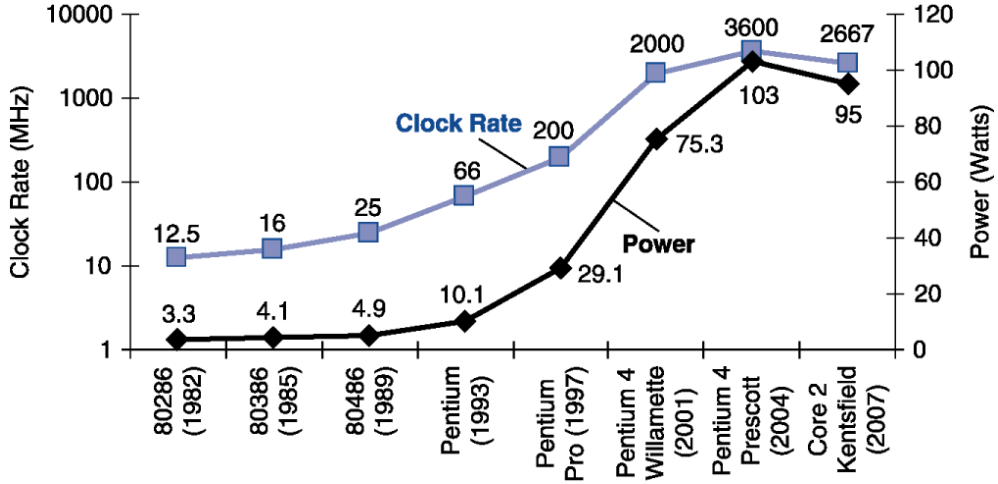
28nm FD-SOI test chip

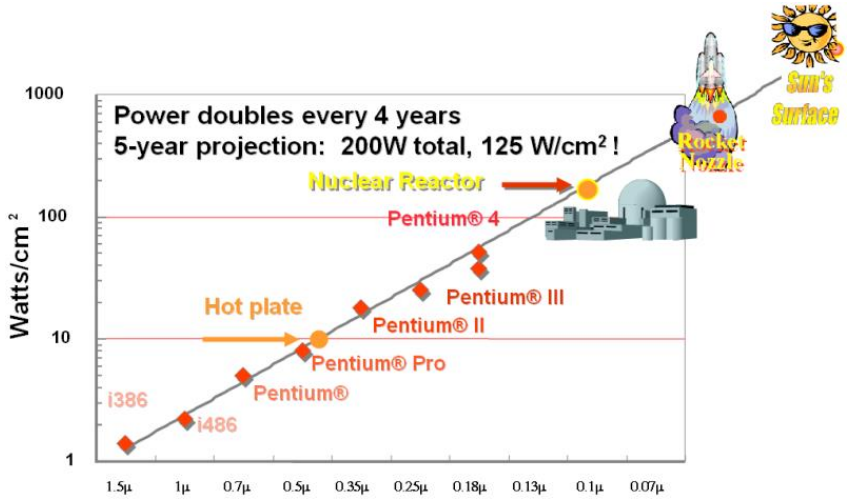


# Power Wall

**Here is a Clue to the Problem**

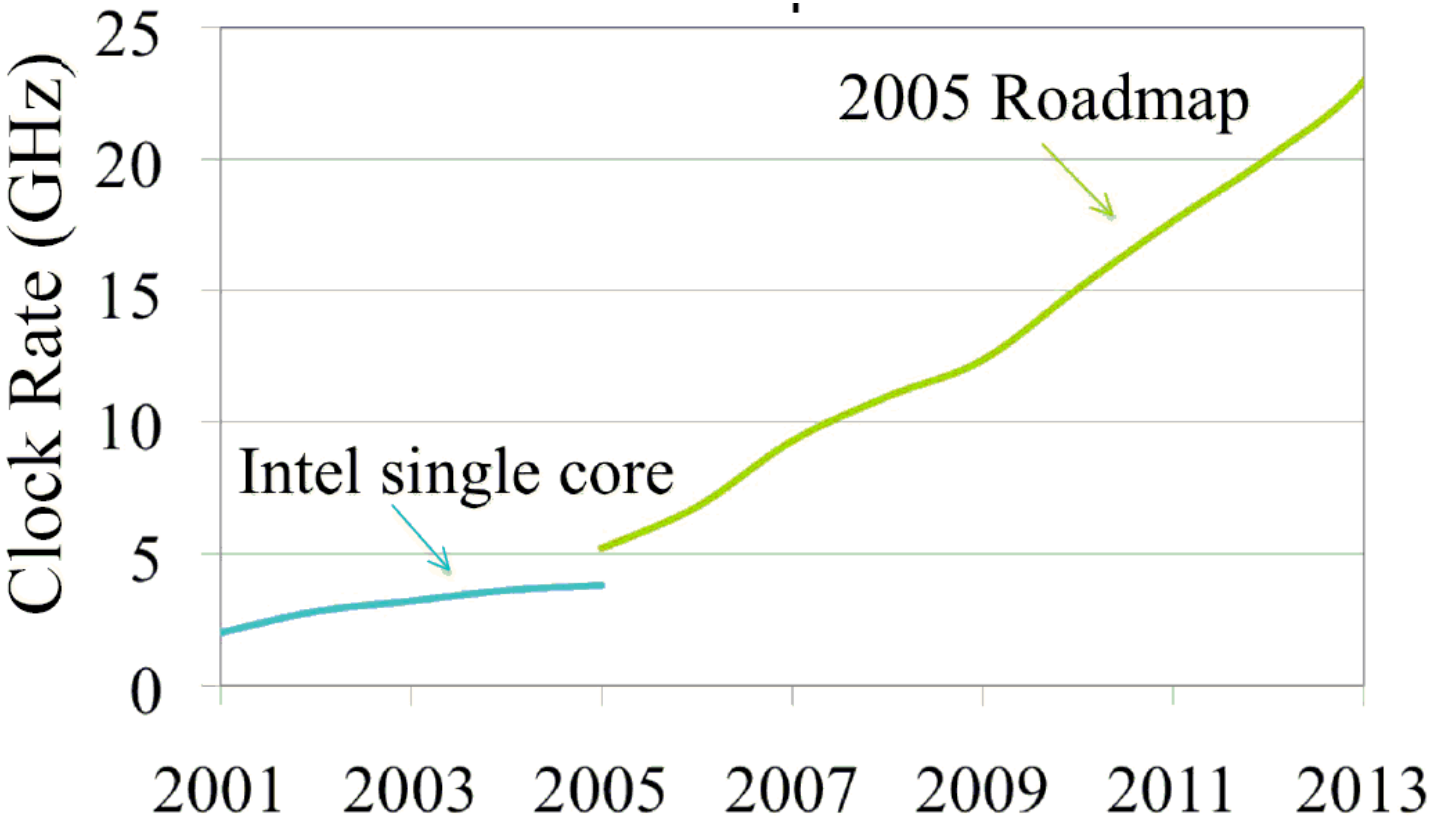
The problem is now called “the Power Wall”. It is illustrated in this figure, taken from Patterson & Hennessy.



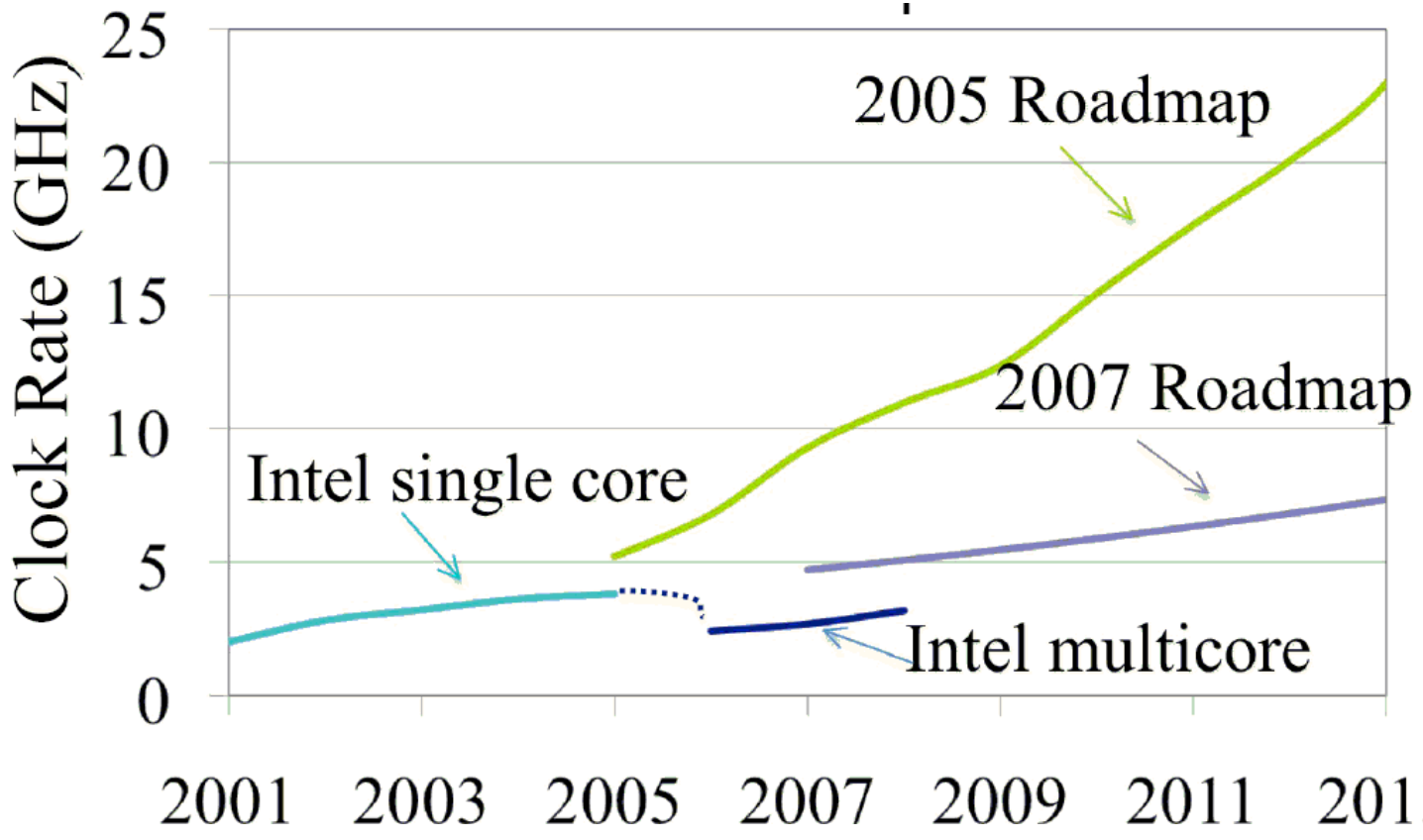

- The design goal for the late 1990’s and early 2000’s was to drive the clock rate up. This was done by adding more transistors to a smaller chip.
- Unfortunately, this increased the power dissipation of the CPU chip beyond the capacity of inexpensive cooling techniques

# Roadmap for CPU Clock Speed: Around 2005




Here is the result of the best thought in 2005. By 2015, the clock speed of the top “hot chip” would be in the 12 – 15 GHz range.

# The CPU Clock Speed Roadmap (A Few Revisions Later)

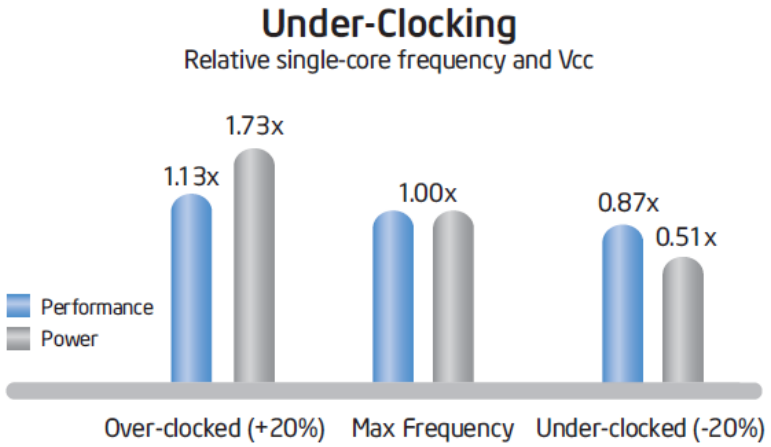


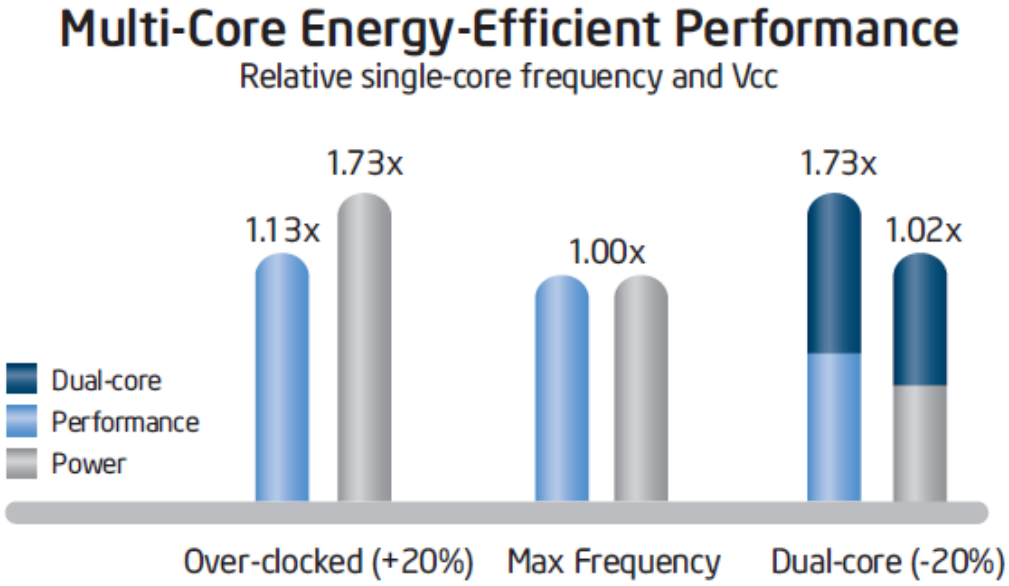

This reflects the practical experience gained with dense chips that were literally "hot"; they radiated considerable thermal power and were difficult to cool.
Law of Physics: All electrical power consumed is eventually radiated as heat.

# The MultiCore Approach

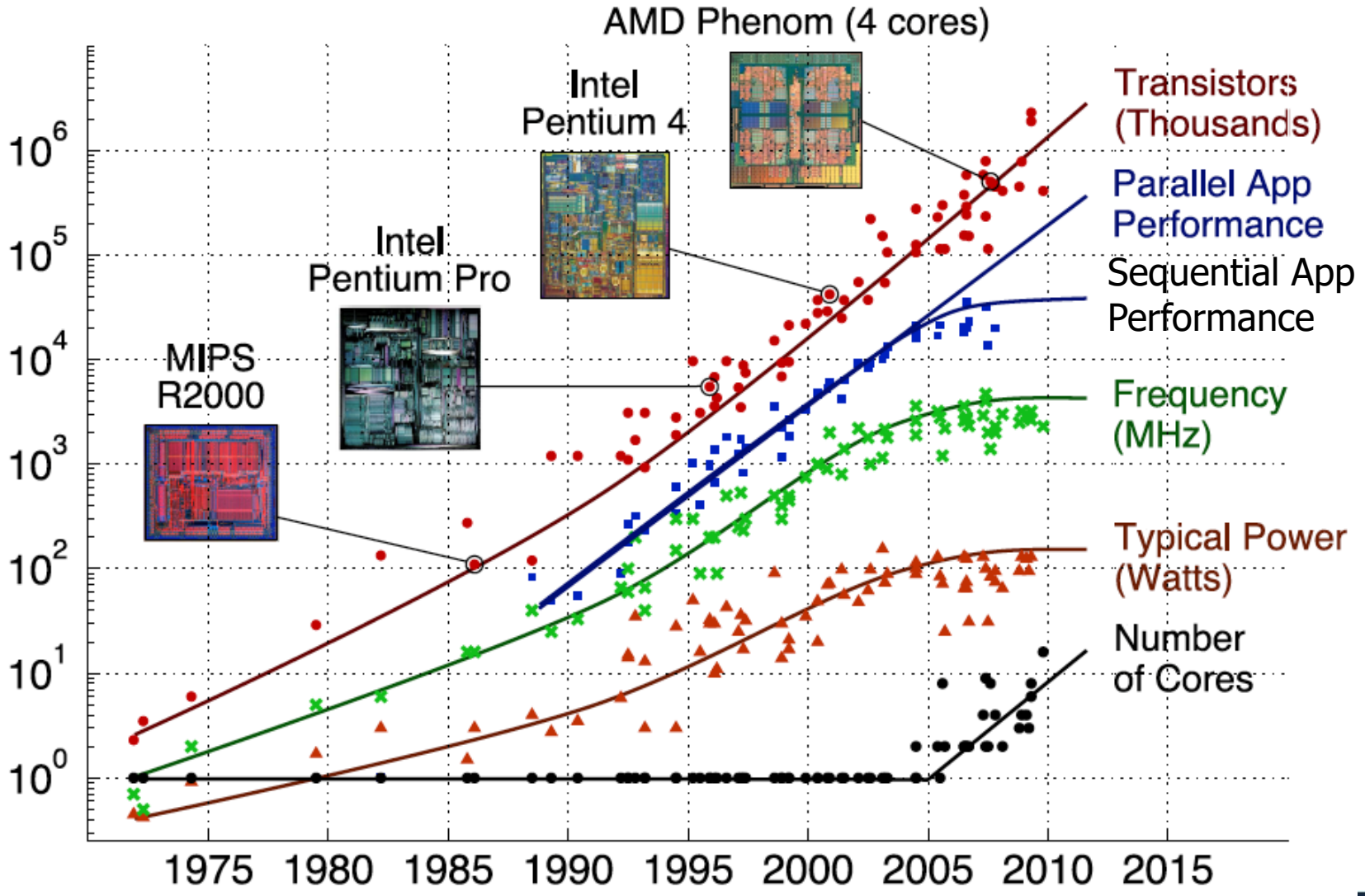
Multiple cores on the same chip

- Simpler
- Slower
- Less power demanding



# Transition to Multicore



# The Utilization Wall

- Scaling theory
  - Transistor and power budgets no longer balanced
  - Exponentially increasing problem
- Observations in the wild
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio

## Classical scaling

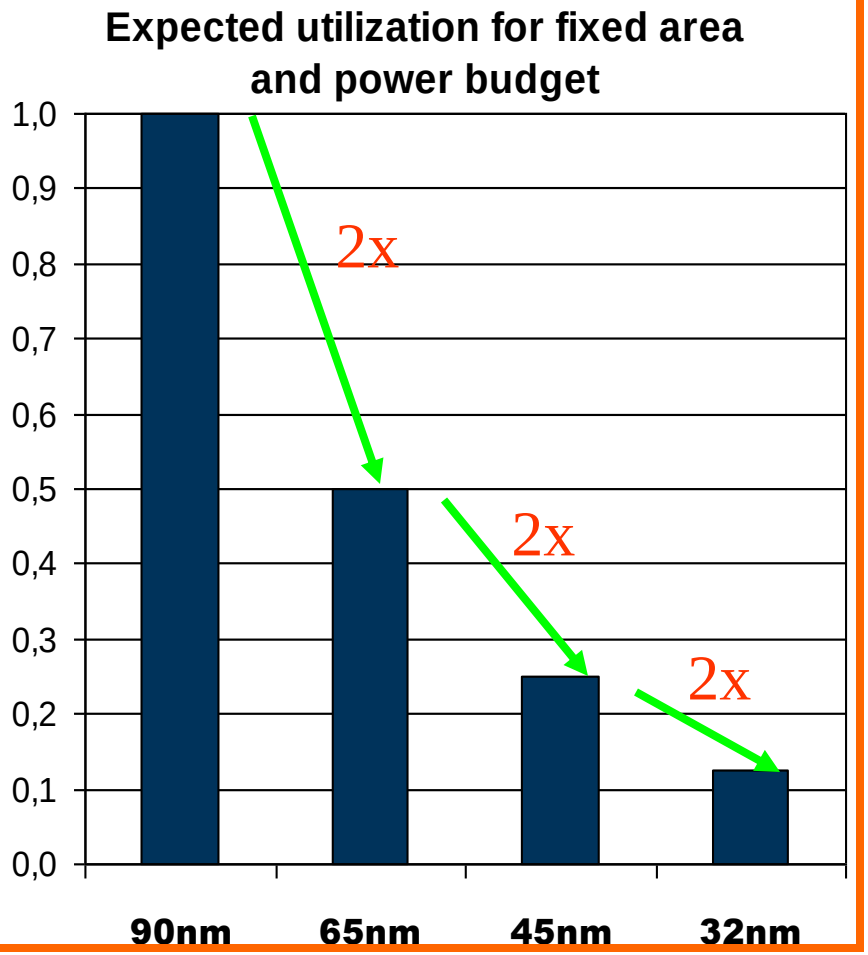
| | |
|---|---|
| Device count | $S^2$ |
| Device frequency | S |
| Device power (cap) | 1/S |
| Device power ($V_{dd}$) | $1/S^2$ |
| **Utilization** | **1** |

## Leakage limited scaling

| | |
|---|---|
| Device count | $S^2$ |
| Device frequency | S |
| Device power (cap) | 1/S |
| **Device power ($V_{dd}$)** | **~1** |
| **Utilization** | **$1/S^2$** |

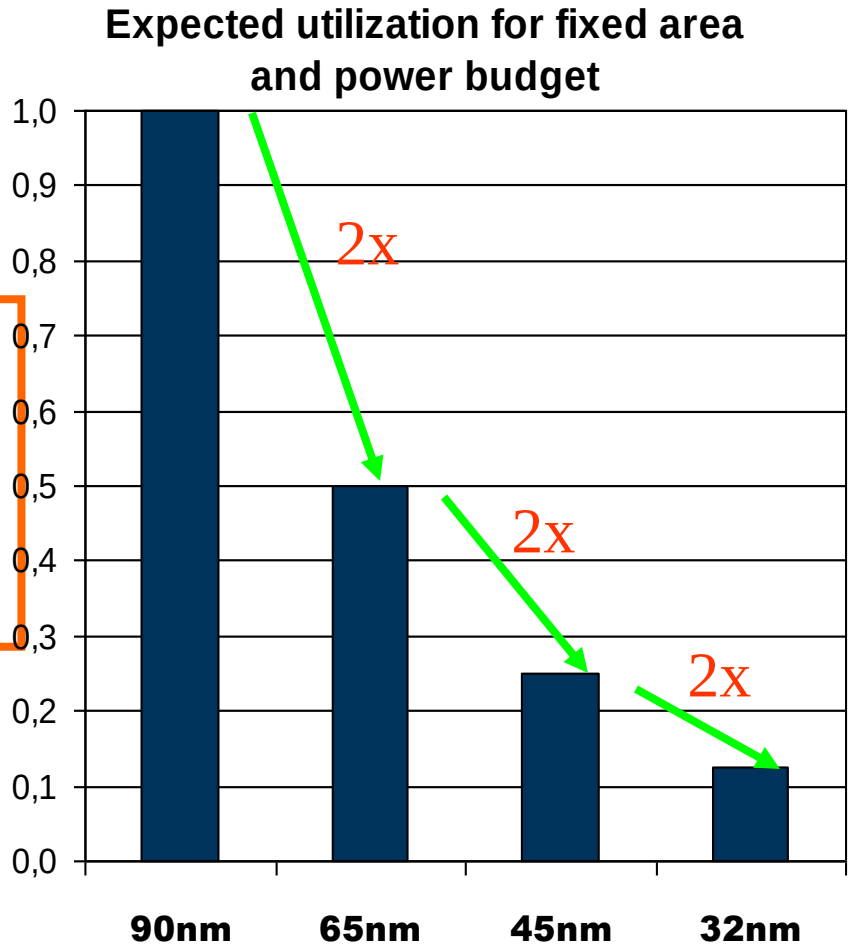
# The Utilization Wall

- Scaling theory
  - Transistor and power budgets no longer balanced
  - Exponentially increasing problem
- Observations in the wild
  - Flat frequency curve
  - “Turbo Mode”
  - Increasing cache/processor ratio



# The Utilization Wall

- Scaling theory
  - Transistor and power budgets no longer balanced
  - Exponentially increasing problem!
- Observations in the wild
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio



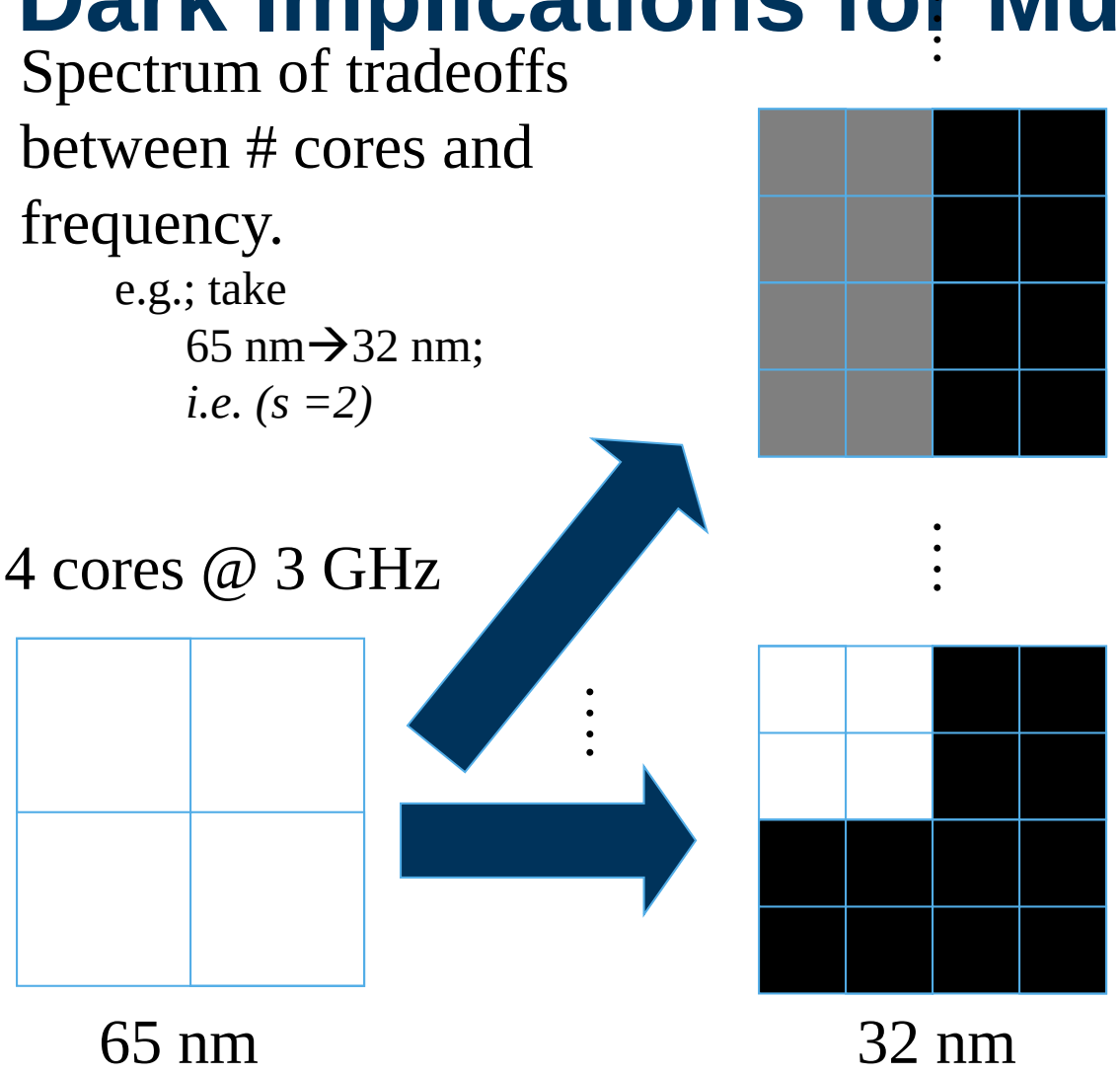
# Utilization Wall: Dark Implications for Multicore

Spectrum of tradeoffs between # cores and frequency.

e.g.; take

65 nm→32 nm;

*i.e. (s =2)*



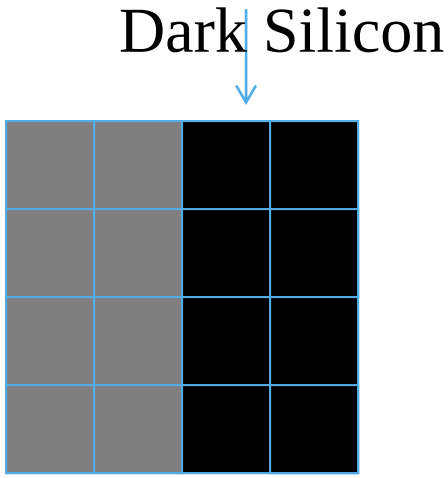

2x4 cores @ 3 GHz
(8 cores dark)
***(Industry's Choice)***

4 cores @ 2x3 GHz
(12 cores dark)



22

# What do we do with Dark Silicon?




- Insights:
  - Power is now more expensive than area
  - Specialized logic has been shown as an effective way to improve energy efficiency (10-1000x)
- Possible Approach:
  - Fill dark silicon with specialized cores to save energy on common apps
  - Near-threshold Computing
  - Turbo mode

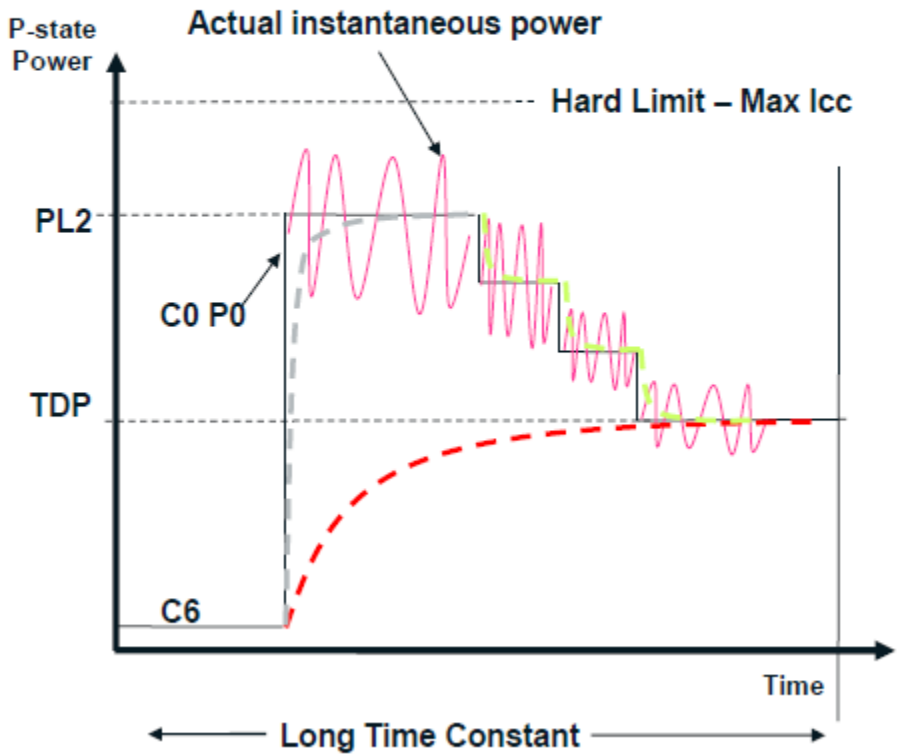
# HW – Turbo Mode Controller



## Intel® Turbo Boost Technology Behavior

Haswell Intel® Turbo Boost Technology uses transient headroom:

- Can briefly exceed TDP for maximum performance.
- Temperatures ramp more quickly, but no impact to "steady state" condition.
- Transient power limited by power delivery capacities of platform (PL2).




*[Haswell_Turbo_Boost_Technology_v0_7_02-1]*

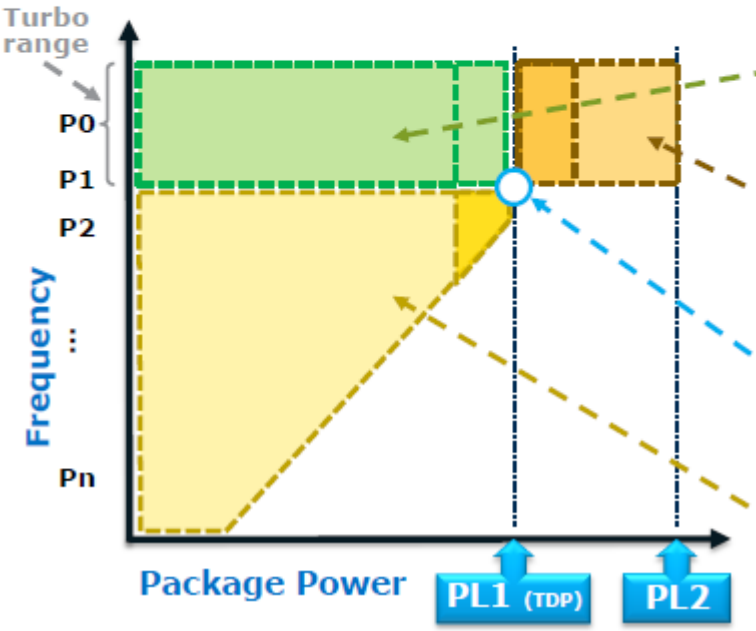
# HW – Turbo Mode Controller



## Intel® Turbo Boost Technology Power vs. Frequency

Reducing PL2 down to TDP (PL1) does not equal turbo disablement:

- You can still enjoy extended turbo performance with most applications.
- Max turbo frequency depends on number of active cores.
- Will reduce performance.




**Sustainable Turbo Region:**
- Workloads in turbo with power <PL1 can continue forever.
- PL1 can only limit frequency down to P1 if clamp bit is not set.

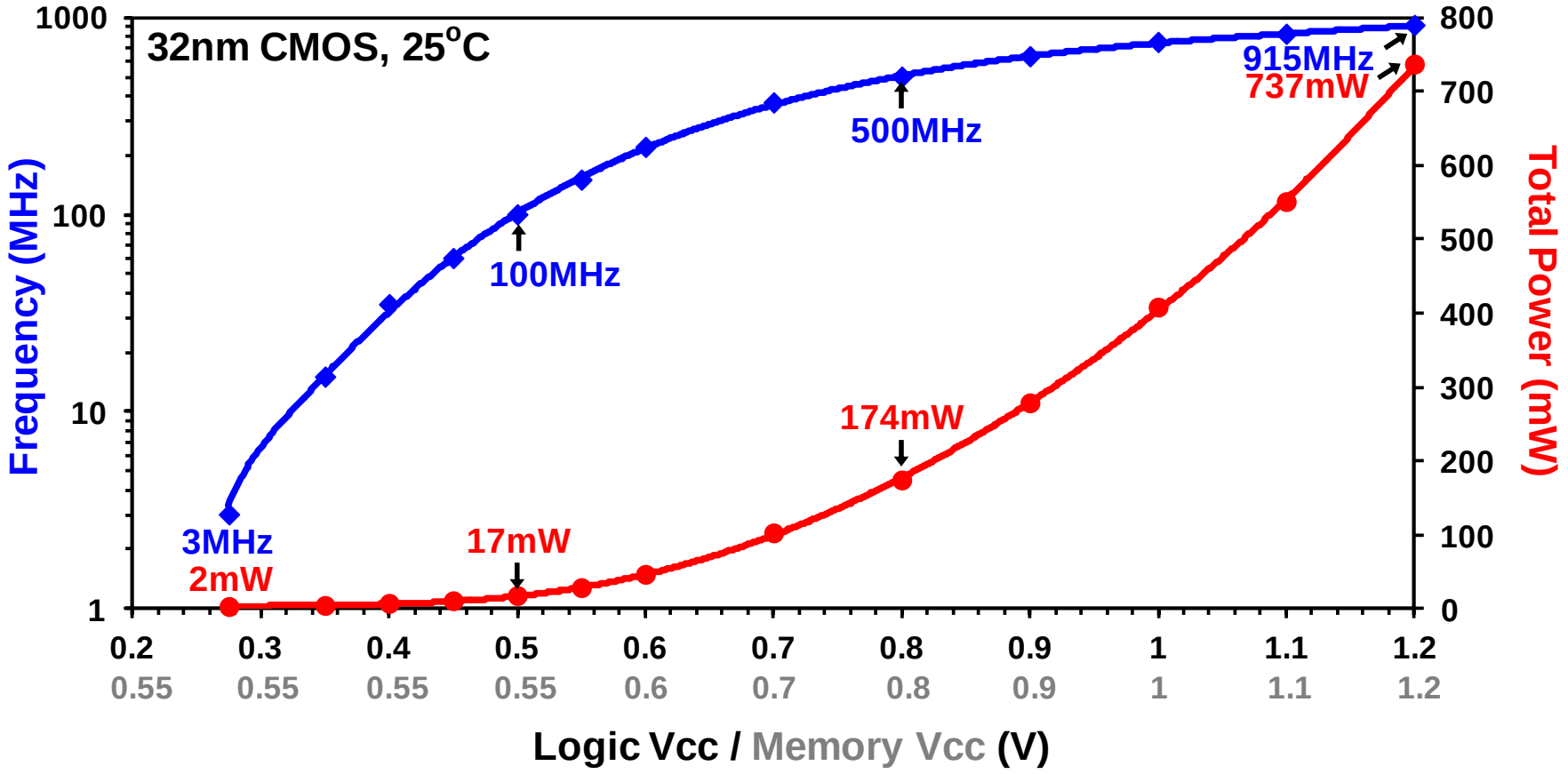
**Haswell Turbo Region:**
- Only area where PL2 has effect.
- Power in this area should not drive thermal design.

**Nominal TDP Point:**
- TDP power can occur without turbo.
- TDP will solely drive thermal design.

**Sustainable Non-Turbo Region:**
- Turbo does not exist, P-state by OS.

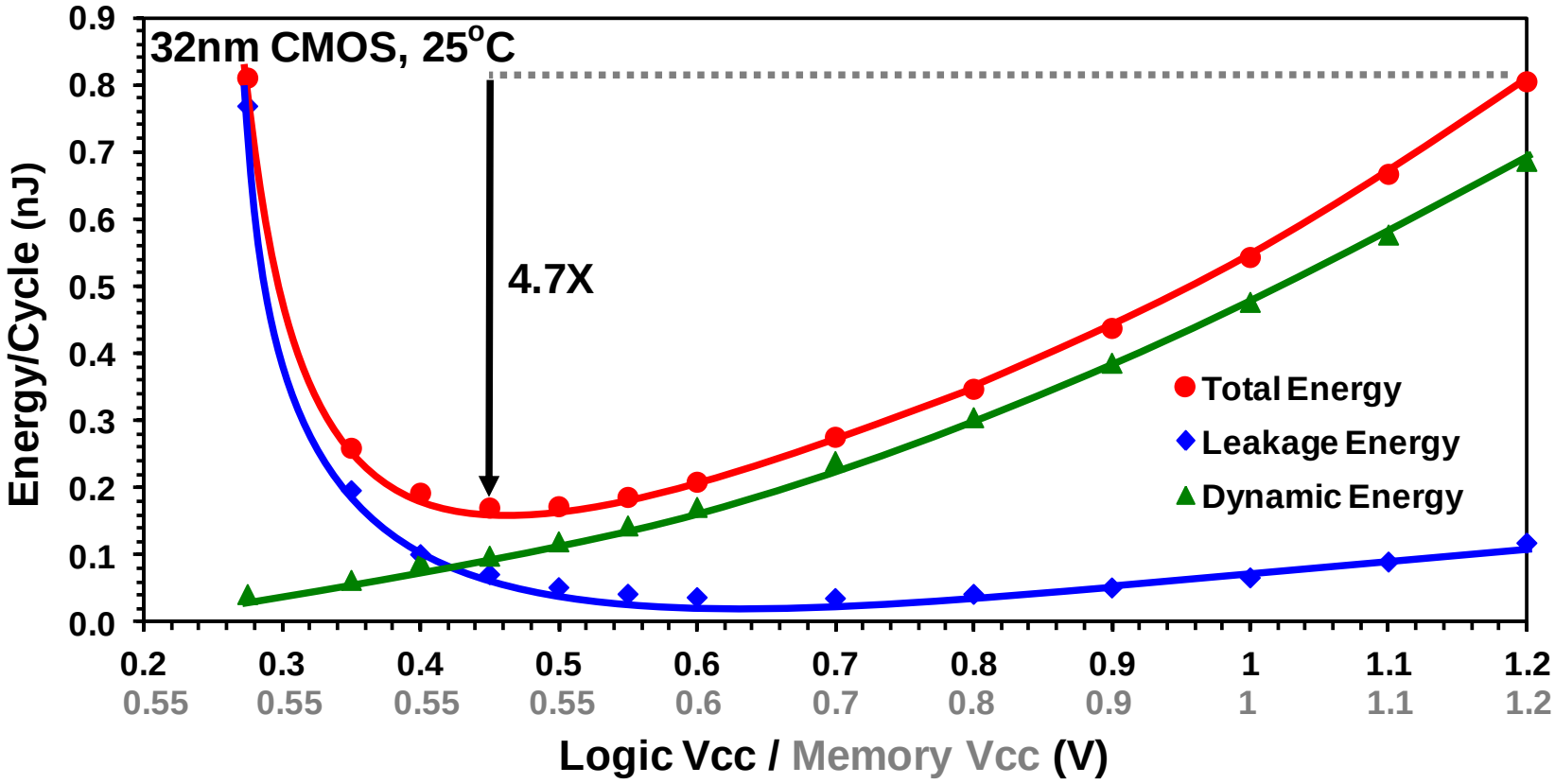
**We will evaluate the impact of turbo logic in terms of energy efficiency**

# Near-Threshold Computing




Vivek De Intel Lab

# Near-Threshold Computing




- **Used in Ultra-low power devices, induces high variation**
- **Still not suitable for HPC systems due to the performance loss**

Vivek De Intel Lab

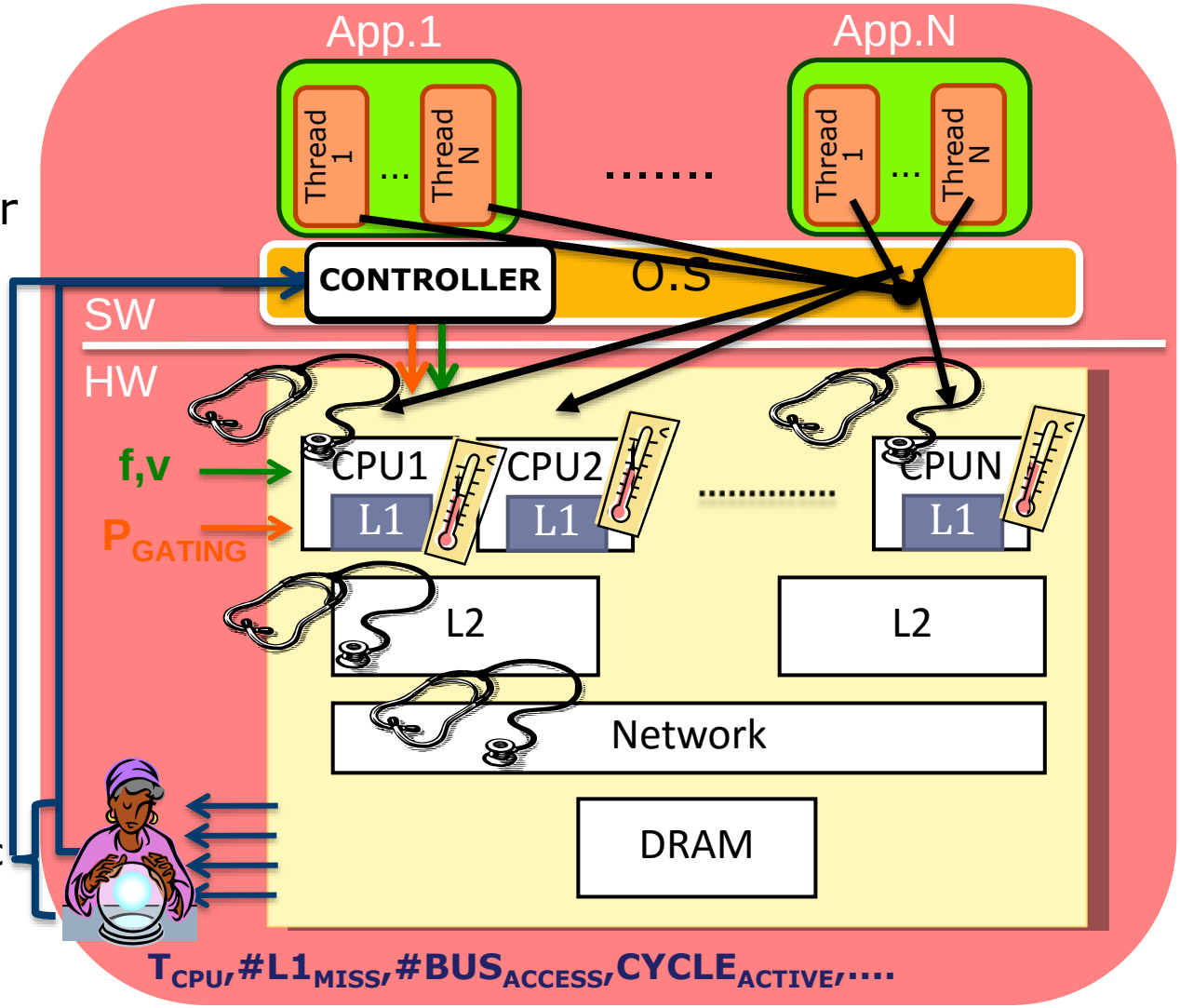
# Outline

- Power in digital systems
- **Dynamic Power Management**
- Power Management & Heterogeneity
- Characterization of thermal effects

# General Architecture



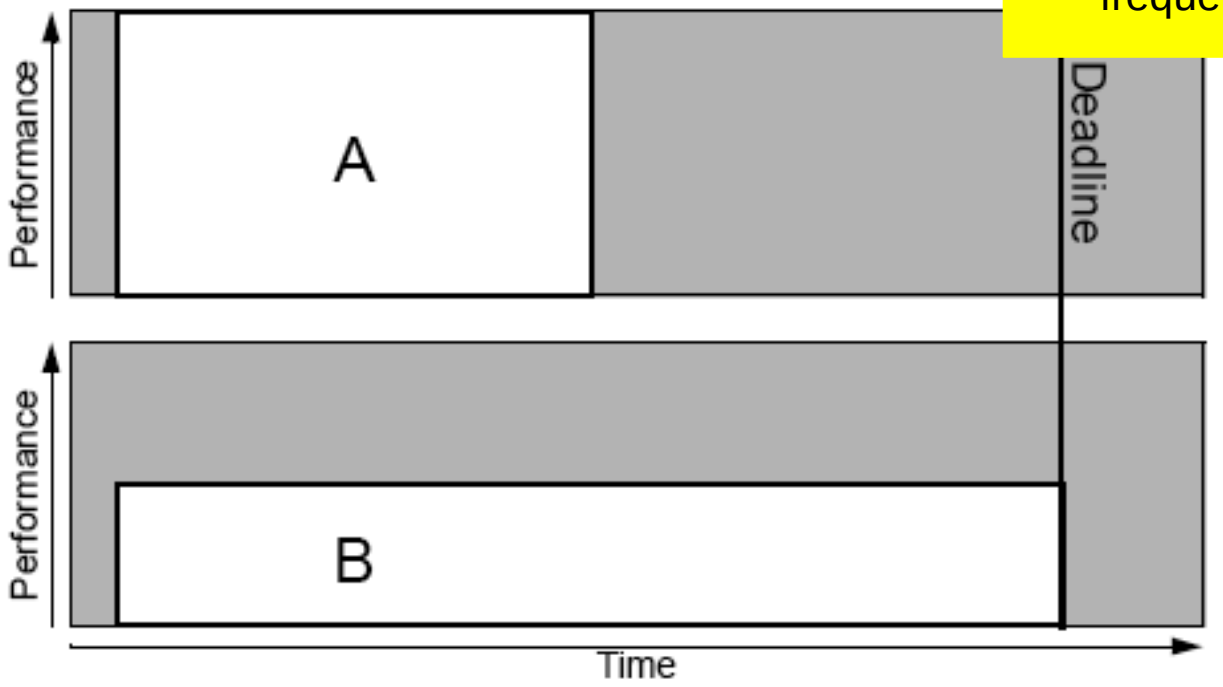
- System
- Sensors
  - Performance counter - PMU
  - Core temperature
- Actuator - Knobs
  - ACPI states
    - P-State → DVFS
    - C-State → $P_{GATING}$
  - Task allocation
- Controller
  - Reactive
    - Threshold/Heuristic
    - Controller theory
  - Proactive
    - Predictors



# DVFS – with deadline or “on-demand governor”

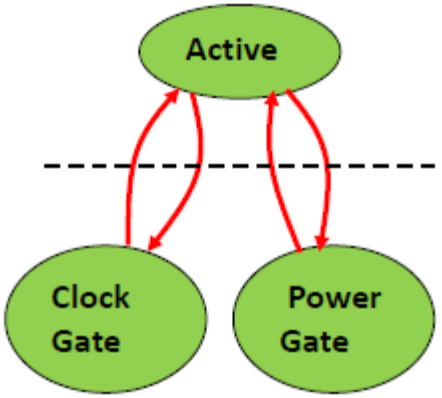
## *Key idea*: Exploit *slack* by scaling V & f to run evenly across a time quantum

Linux on-demand governor:
frequency ~ cpu-load



# Run Fast and Stop (RTFS) vs DVFS

- **Run Fast Then Stop (RFTS) is a technique where the processor runs at the highest frequency until the job is finished, then it stops.**
- **DVFS runs "low and slow" to reduce dynamic power by V^2.**
  - **Active Power Gate Clock Gate**
- **RFTS:**
  1. **Clock Gating – core continues to leak.**
  2. **Power Gating – core is powered off and doesn't leak.**



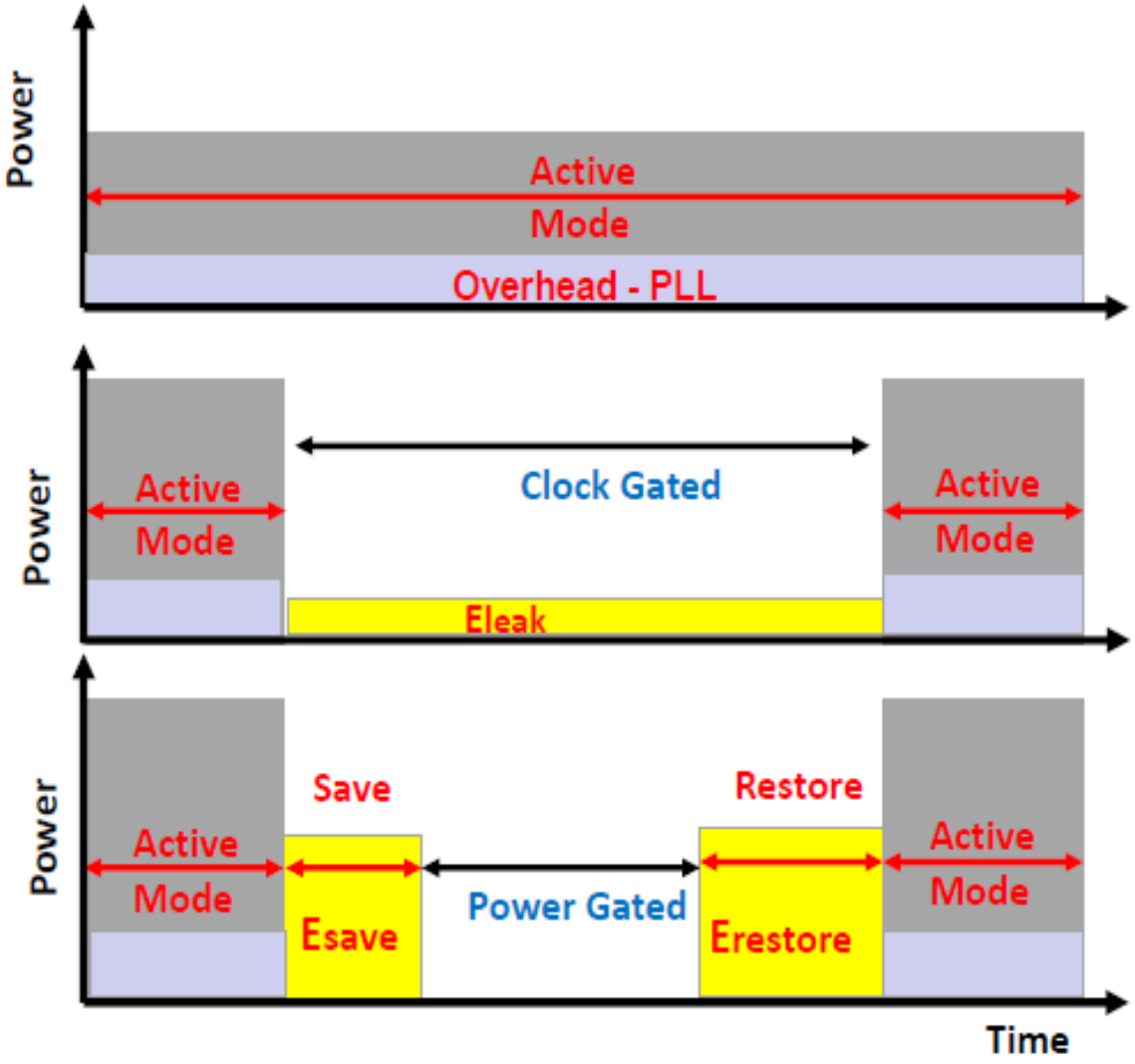
# Run Fast and Stop (RFTS) vs DVFS - II

**What is best?**

1. **DVFS**

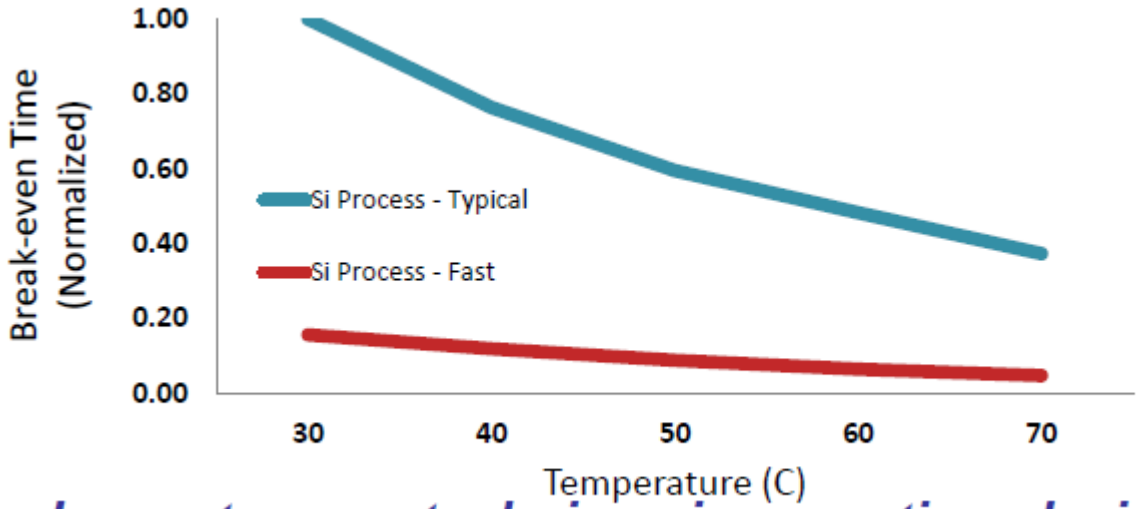
2. **RFTS:**
   - **clock gate**

3. **RFTS:**
   - **power gate**



# Run Fast and Stop vs DVFS - III

- **"Break-even time" is defined as the time that the core needs to be powered off to compensate for save and restore energy.**
- **In high leakage situations, the power gating benefit is realized in a shorter time.**



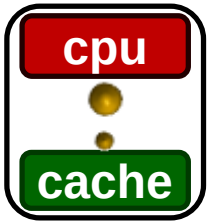

***Lowest power technique is a run-time decision.***

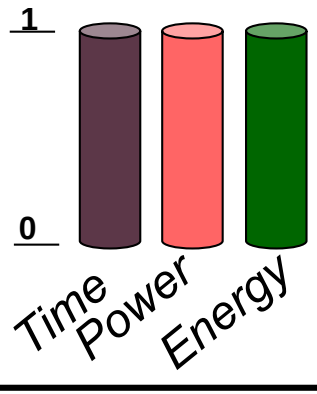
# .. depends on workload



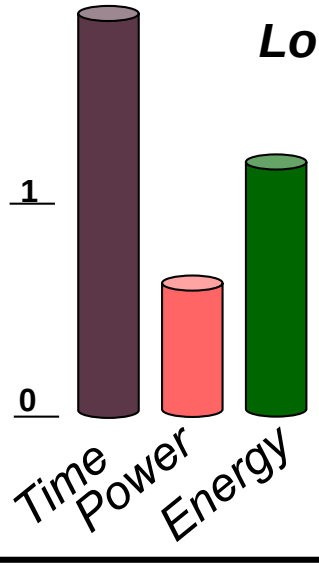
**CPU BOUND TASK**

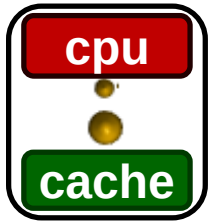
*High Frequency*



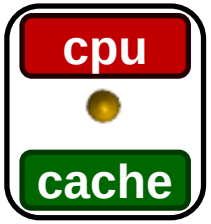



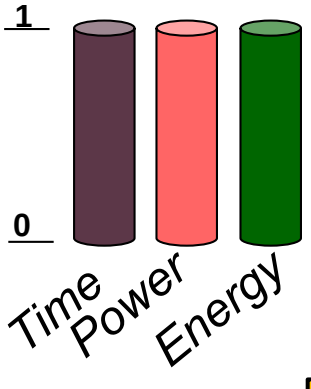

- Performance Loss
- Power reduction
- Energy Efficiency Loss!



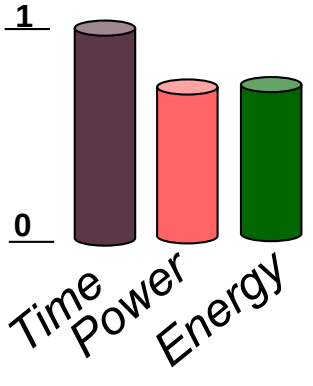



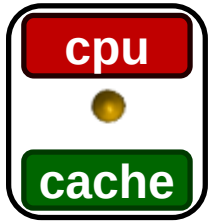

*Low Frequency*




- Same Performance
- Power reduction
- Energy Efficiency Gain!





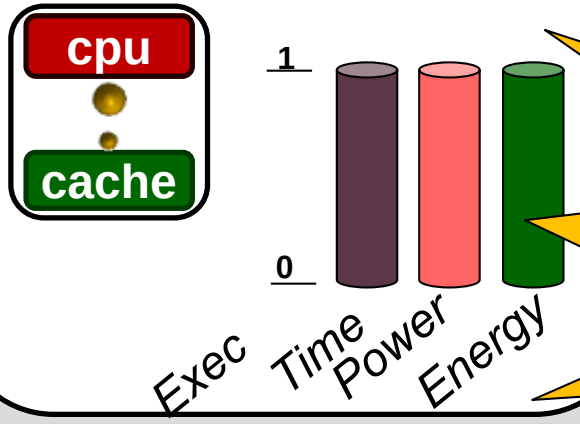

**MEMORY BOUND TASK**

*High Frequency*

*Low Frequency*

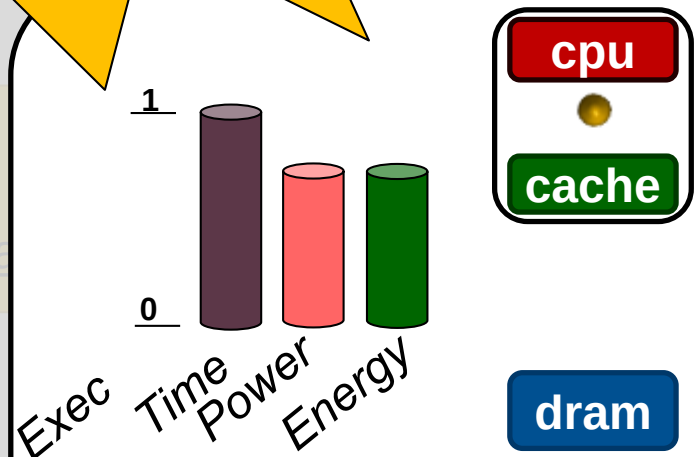
## CPU BOUND TASK

### High Frequency




**Best Policy**

- **Power Saving**
- **No performance Loss**
- **Higher Energy Efficiency**



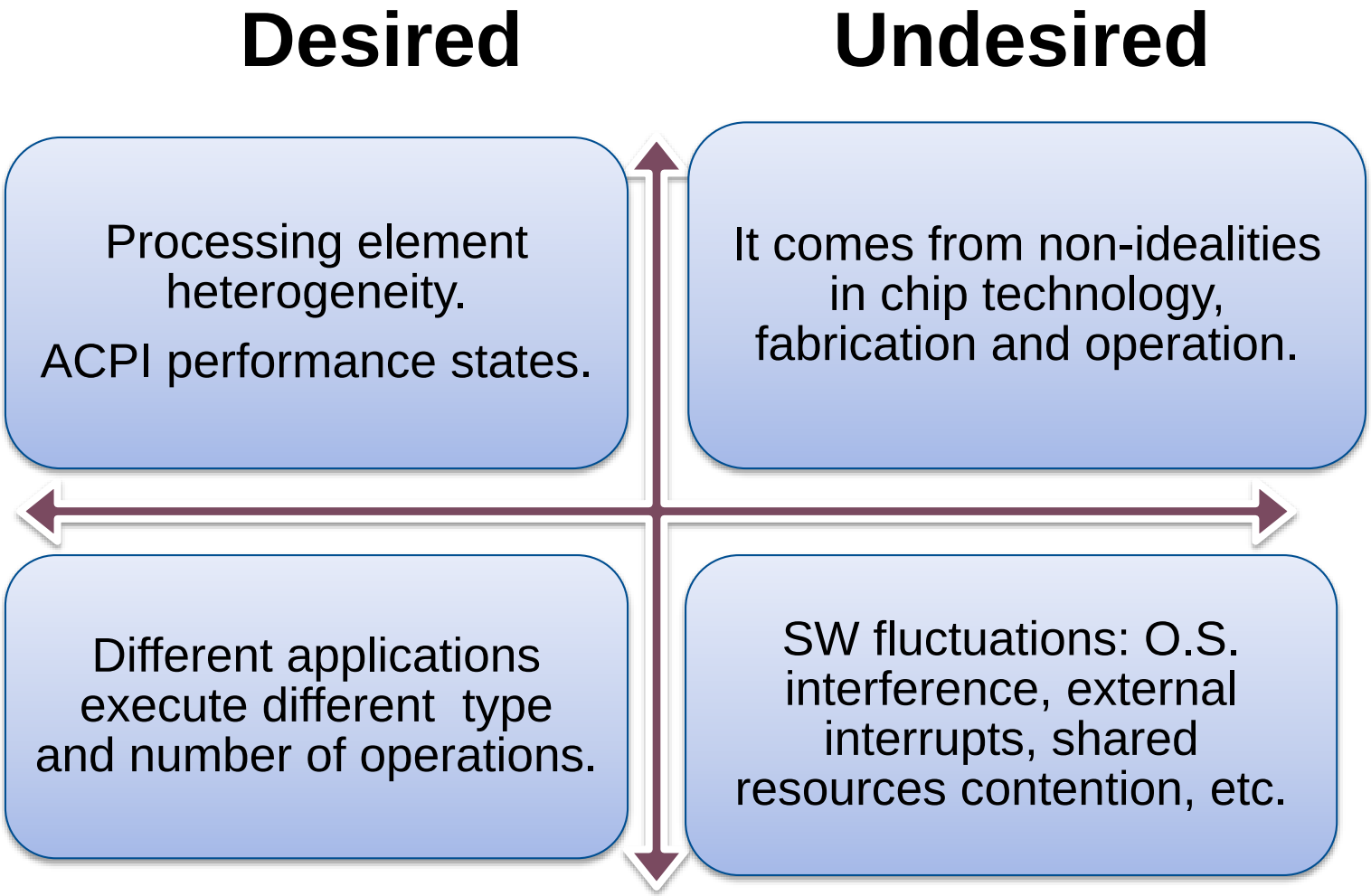

### Low Frequency

## MEMORY BOUND TASK

# Outline

- Power in digital systems
- Dynamic Power Management
- **Power Management & Heterogeneity**
- Characterization of thermal effects

# Heterogeneity in Supercomputers

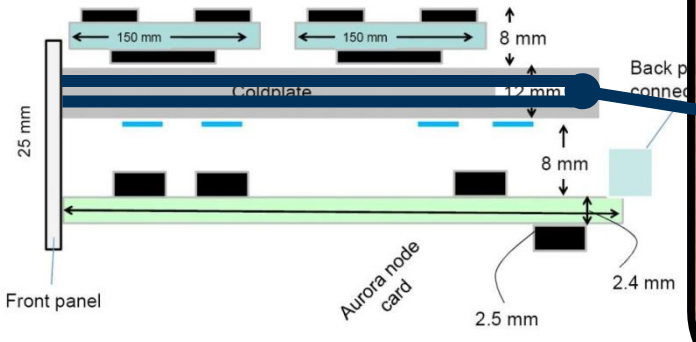


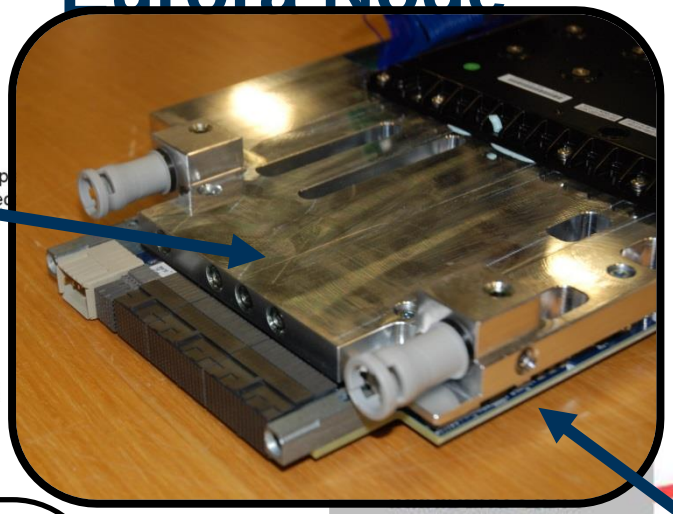
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Eurora Node

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

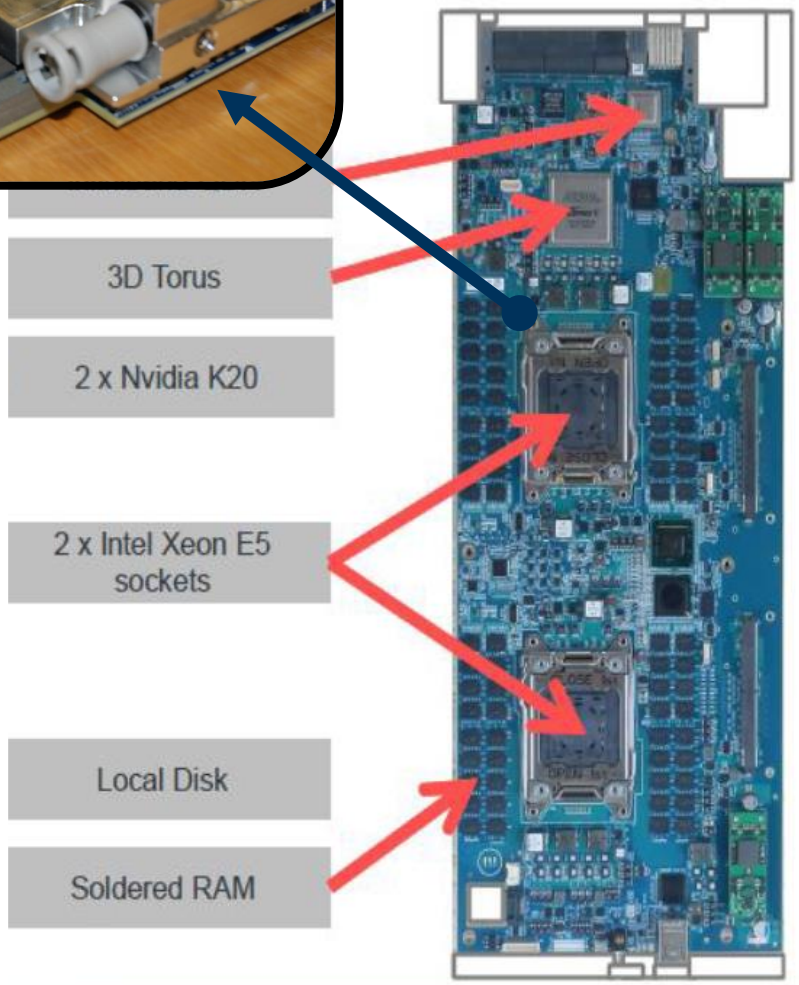
EUROTECH
Imagine. Build. Succeed.



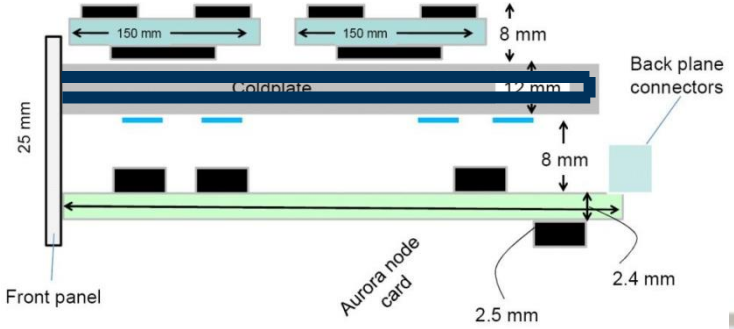
Each Node:

- 2 x Intel Xeon E5 Series sockets
- 160GB SSD
- EEC DDR3 1.6GHz
- Altera Stratix V
- Infiniband Controller
- 2 x Accelerator Expansion slots
- Cooling Plate



# Eurora Node






Xeon PHI

K20

Cooling Plate

Infiniband QDR

3D Torus

2 x Nvidia K20

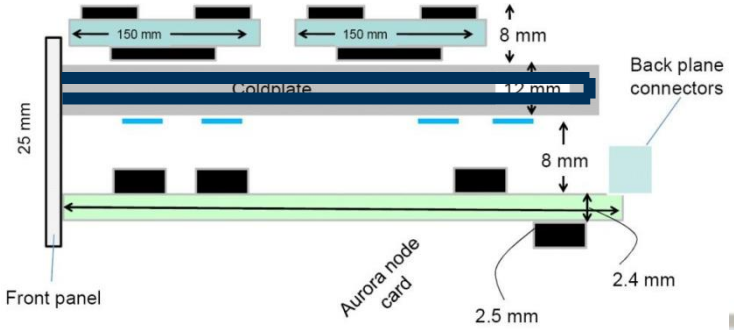
2 x Intel Xeon E5 sockets

Local Disk

Soldered RAM

# Eurora Node






K20

Xeon PHI

EUROTECH

**Eurora System - 64 nodes**

**CPUs:**

- 2 CPU - E5-2658 (node 1-32)
  - 8 cores @ 2GHz
  - TDP 95 W
  - 2.8 Turbo max freq.
- 2 CPU - E5-2687W (node 33-64)
  - 8 cores @ 3.1GHz
  - TDP 150 W
  - 3.8 Turbo max freq.

**Accelerator:**

- 2x Nvidia Kepler K20 card (node 33-64)
  - 32 GB of GDDR5
  - Peak 2 TFLOP DP @ 250W
- 2x Intel Xeon Phi (node 1-32)
  - 16GB GDDR5
  - Peak 1.4 TFFLOP DP @ 245W

**Software:**

- SMP CentOS Linux
- On demand power governor

# Workload Design

Benchmarks and tests performed :

- ***SYNT CPU:*** synthetic parallel benchmark. It emulates a **CPU bound** application.
- ***SYNT Mem:*** synthetic parallel benchmark. It emulates a **memory bound** task execution.
- ***QE:*** Quantum ESPRESSO is a freely available integrated suite of computer codes for electronic-structure calculations.

To generate the data-set used for characterizing the variability sources in Eurora:

- We designed a PBS script that set the frequencies and runs the benchmarks.
- We save the initial time and end time.
- Off-line the log files are used to navigate the traces of the Eurora monitoring framework.

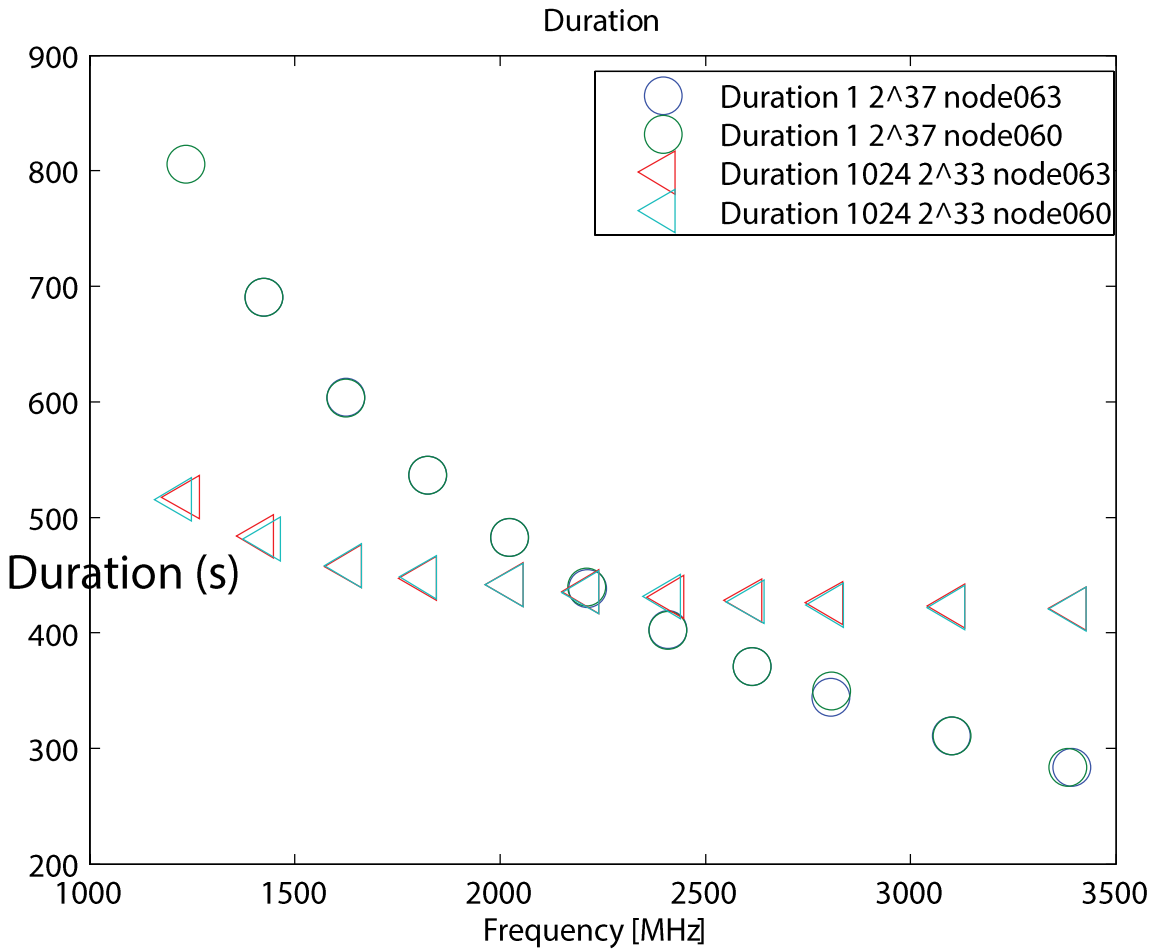
# DVFS to Energy Efficiency - Eurora

CPU bound

- Time ~ freq

Mem bound

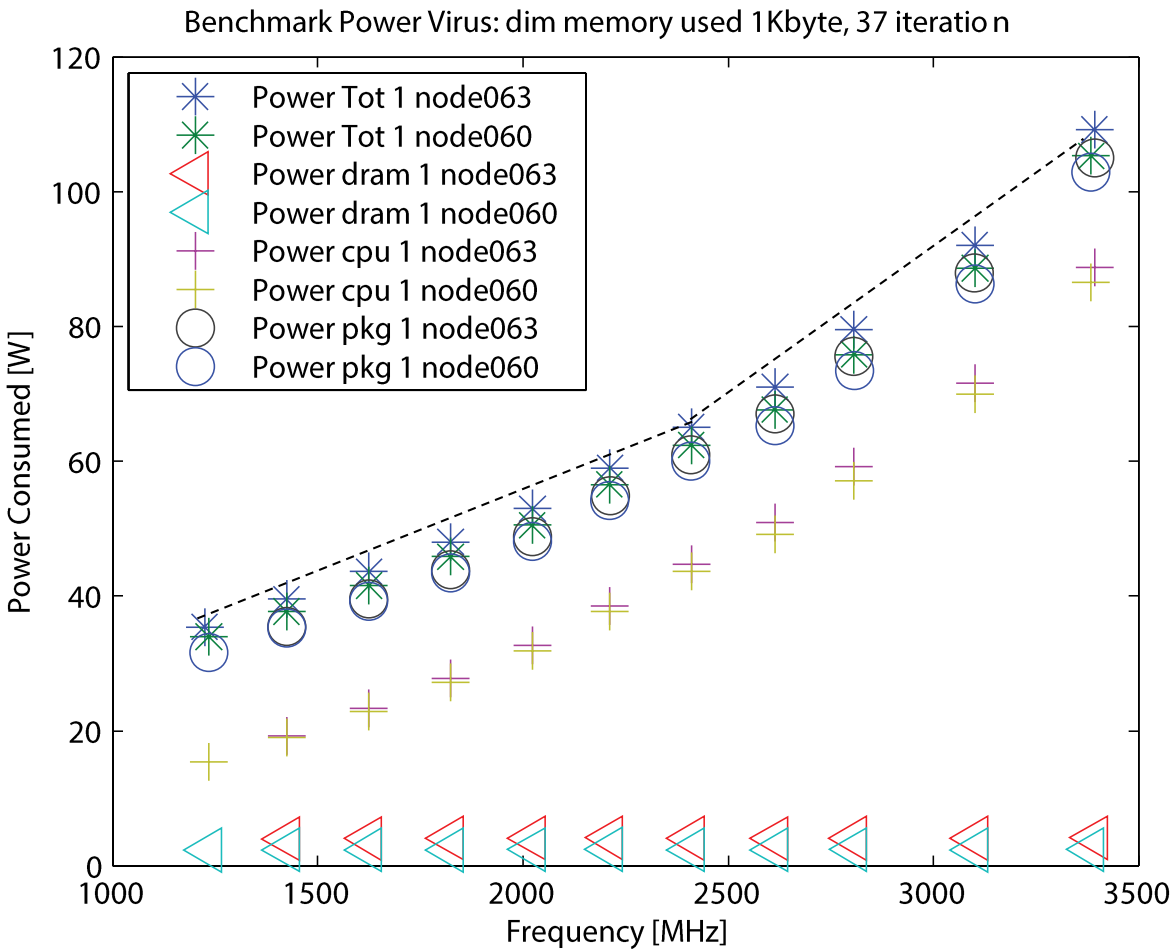
- Time ~~*~~ freq



# DVFS to Energy Efficiency - Eurora

## CPU bound

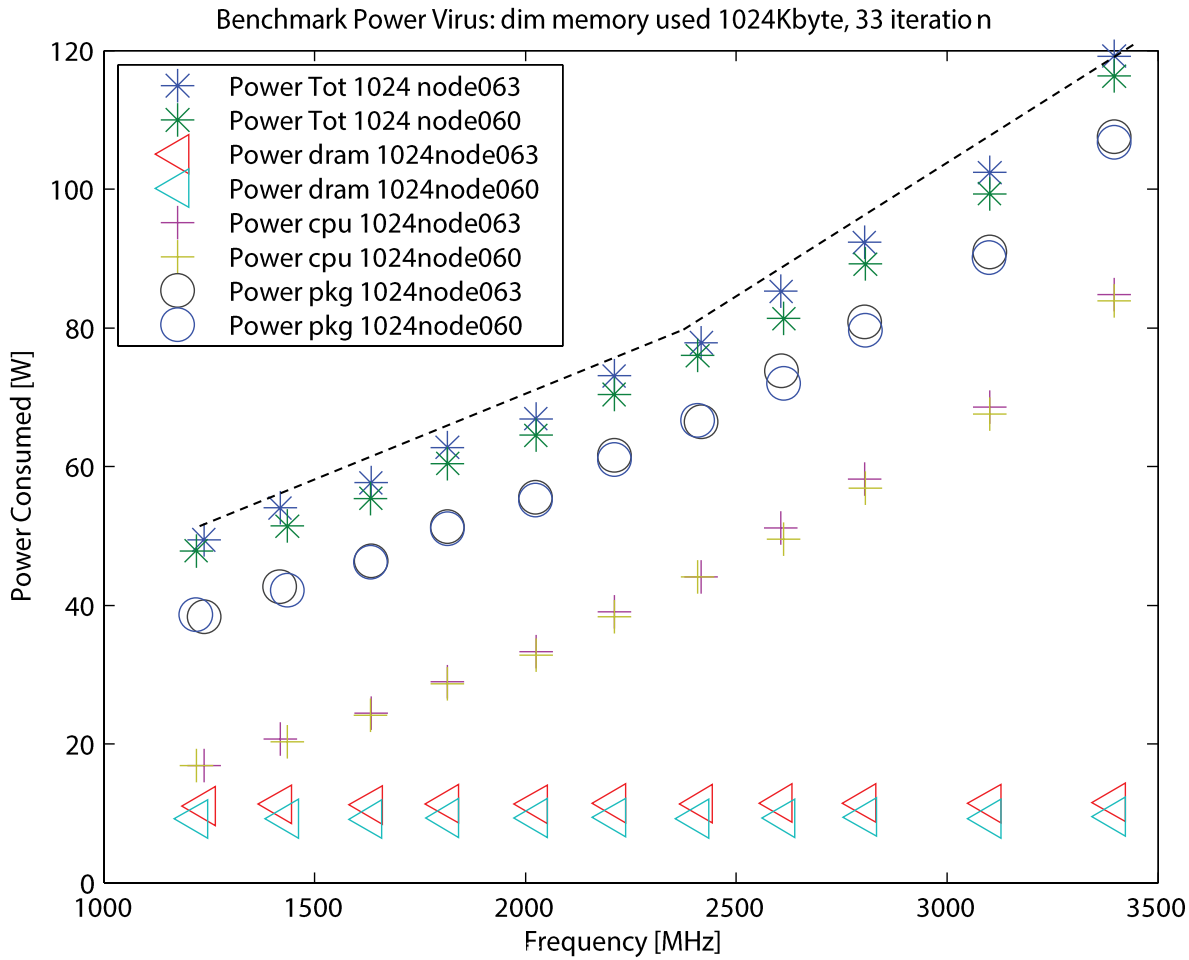
- Pdram constant
  ~ 5W
- Ppkg
  ~ 20W higher Pcpu
- Sligthly different power in between nodes ~ 5W
- Power ~ Ppkg



# DVFS to Energy Efficiency - Eurora

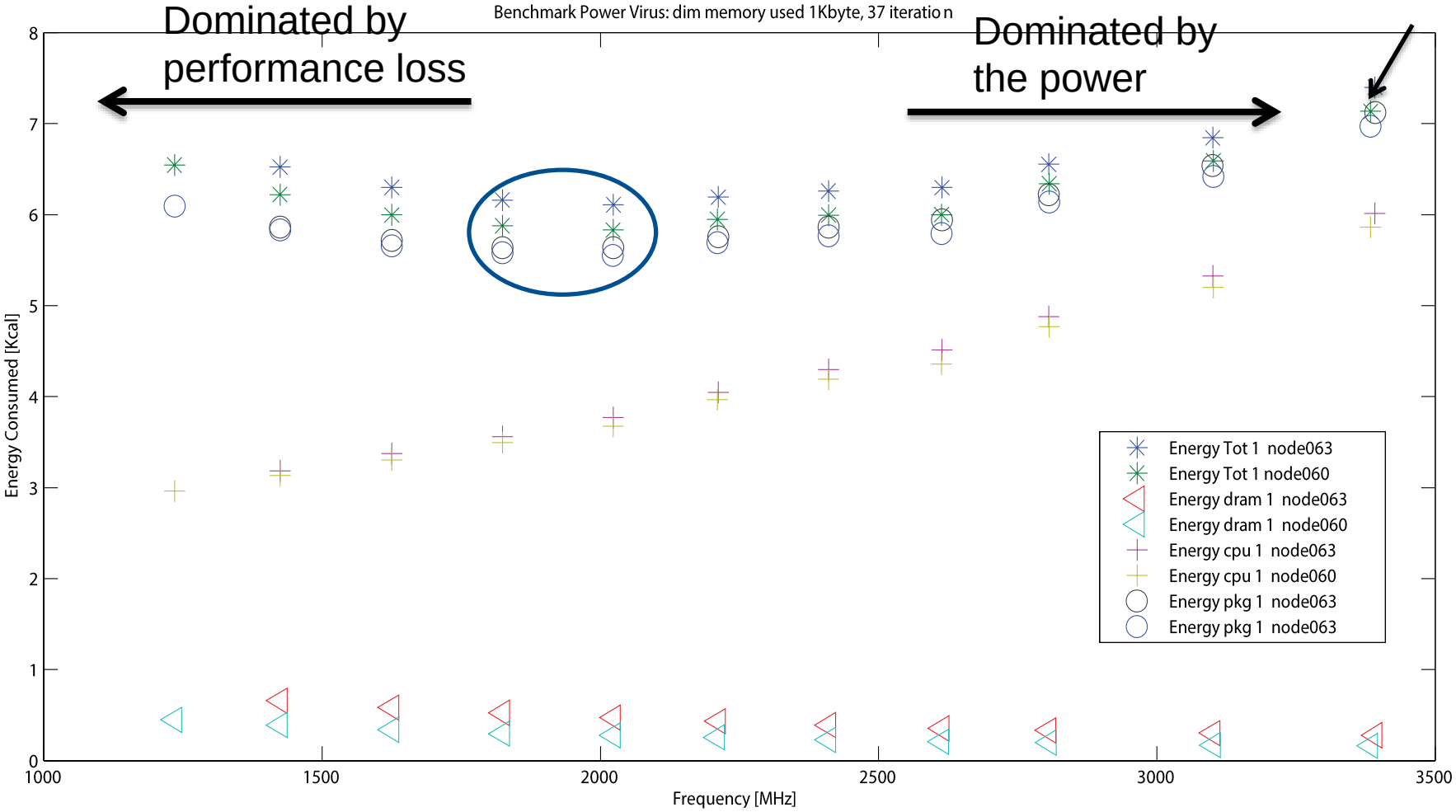
## MEM bound

- Pdram constant
  ~ 10W
- Ppkg
  ~ 25W higher Pcpu
- Sligthly different power in between nodes ~ 5W



# DVFS to Energy Efficiency - Eurora

CPU bound – energy minimum @ 1.8-2GHz

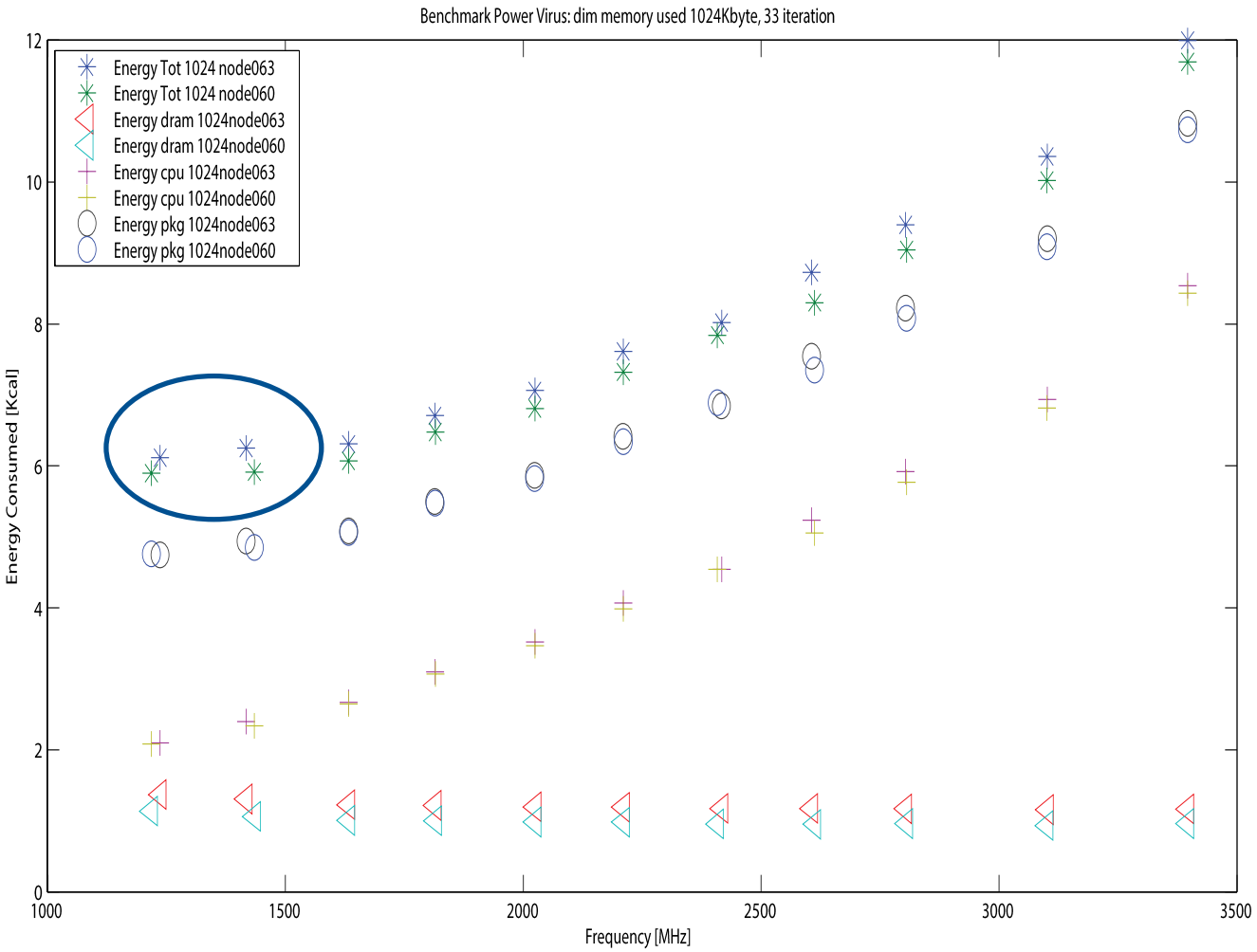


# DVFS to Energy Efficiency - Eurora

MEM bound

- Minimum E @ min freq.
- Energy gain ↓ ↓ frequenza
- @ 2^37 Iterations
  - @Turbo => 380Kcal

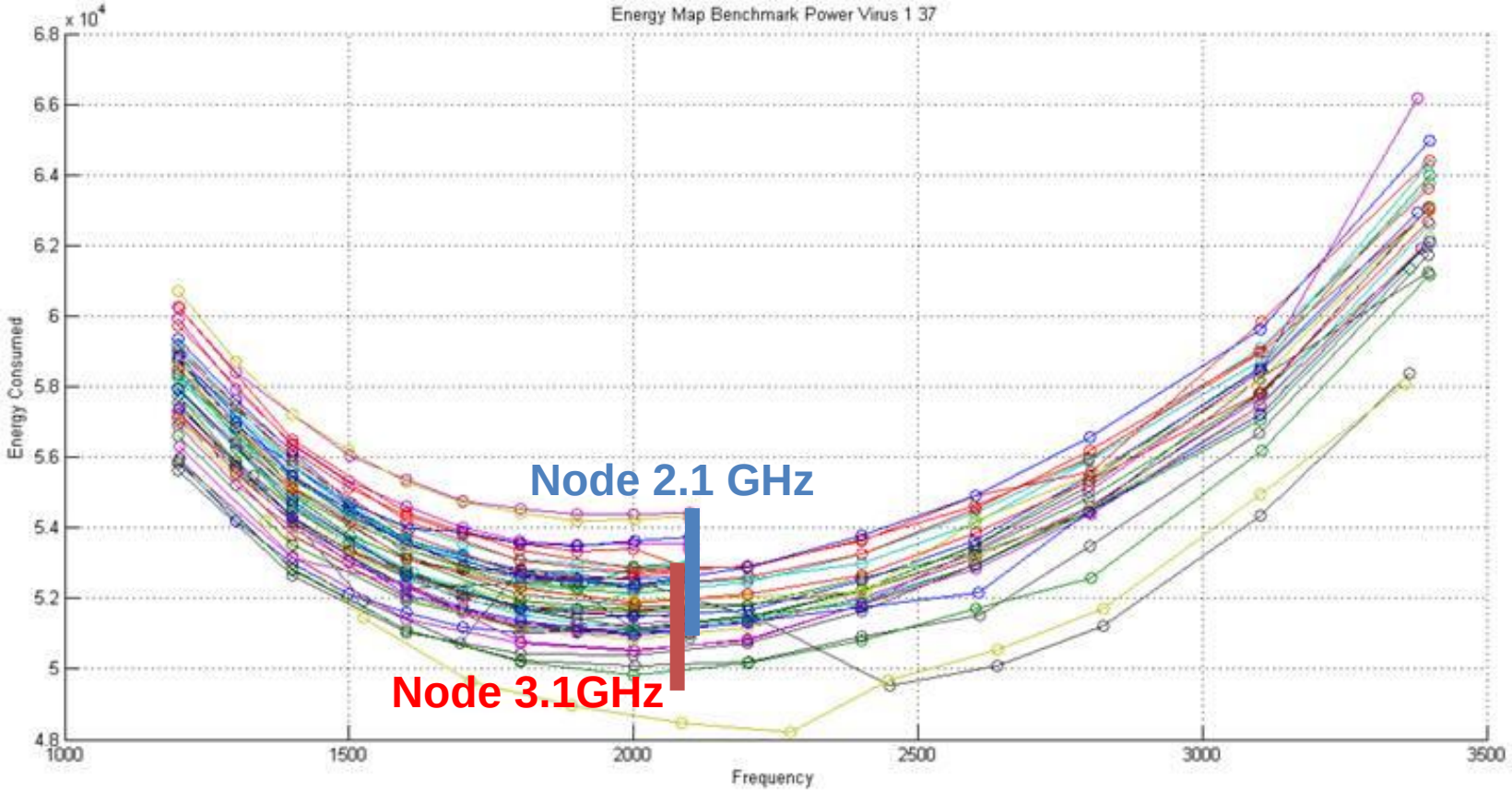
  - @ 1.2GHz =>190Kcal



# Energy Heterogeneity



- Current System (Eurora)
  - Intra node variability ~ 10%



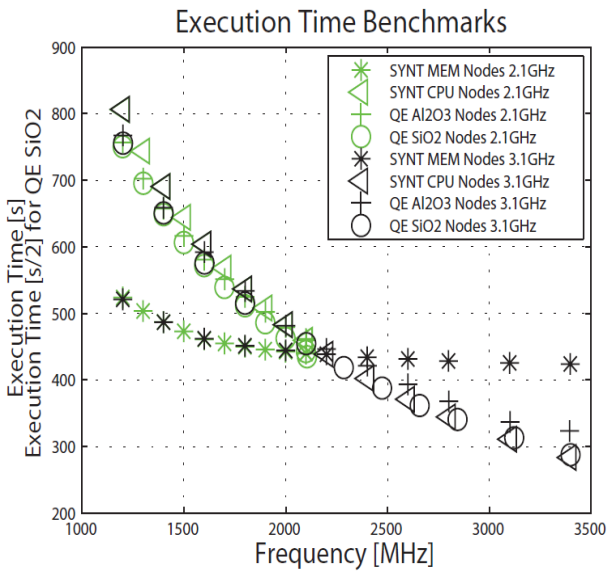
# Energy Heterogeneity

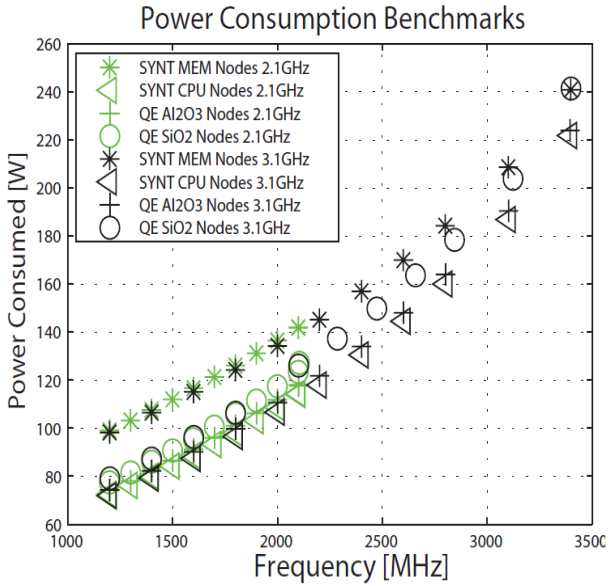


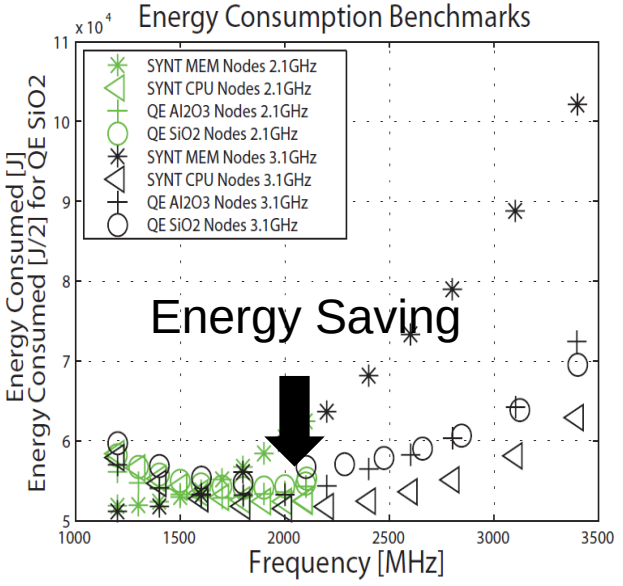
- Current System (Eurora)
  - Intra node variability ~ 10%
  - Operating point sensitivity
    - Max perf. not Min Energy
  - HW Accelerators

DVFS vs. RFTS

| Nodes | Optimal [MHz] Frequency | Ex Time [%] Overhead | Energy [%] Saving | EDP [%] Saving |
|---|---|---|---|---|
| Benchmark SYNT CPU | | | | |
| 2.1GHz | 1900 (2101) | -11 (0) | +2 (0) | -11 (0) |
| 3.1GHz | 2000 (3101) | -70 (0) | +18 (0) | -39 (0) |
| Benchmark SYNT Mem | | | | |
| 2.1GHz | 1200 (1600) | -18 (-5) | +18 (+8) | +2 (+11) |
| 3.1GHz | 1200 (1600) | -23 (-9) | +50 (+48) | +38 (+43) |
| Benchmark QE $Al^2O^3$ | | | | |
| 2.1GHz | 1700 (2101) | -20 (0) | +3 (0) | -17 (0) |
| 3.1GHz | 1800 (3100) | -65 (-4) | +27 (+11) | -21 (+8) |
| Benchmark QE-$SiO^2$ | | | | |
| 2.1GHz | 1800 (2101) | -18 (0) | +3 (0) | -15 (0) |
| 3.1GHz | 1800 (3100) | -79 (-9) | +21 (+8) | -40 (+1) |



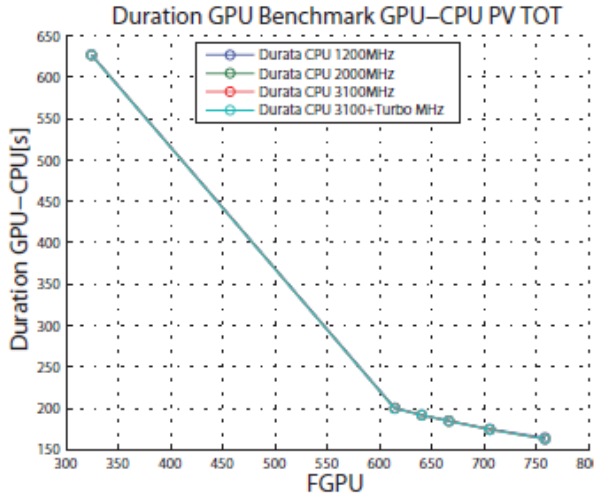



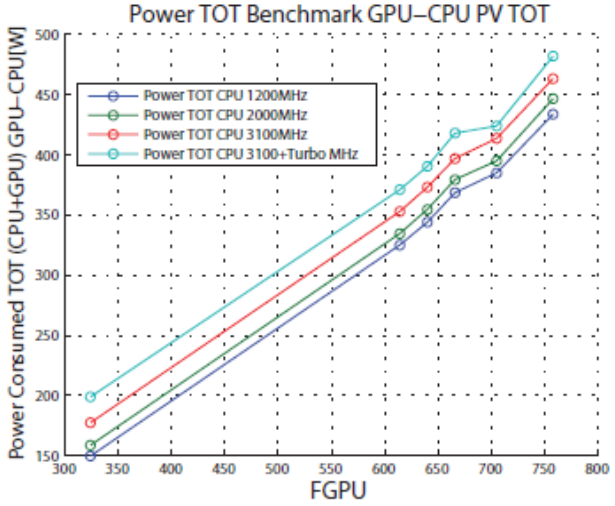



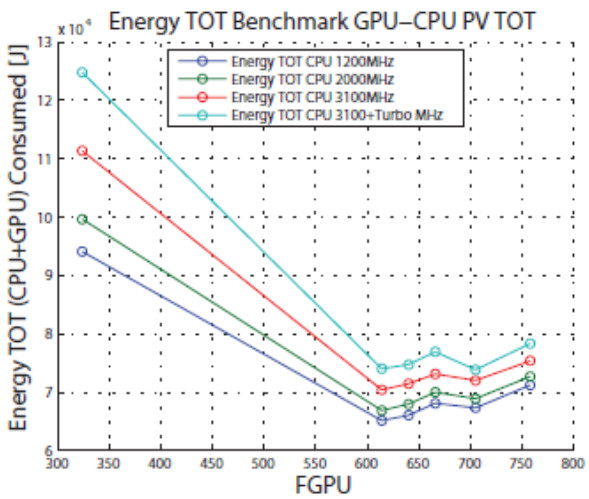
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# HW Accelerators

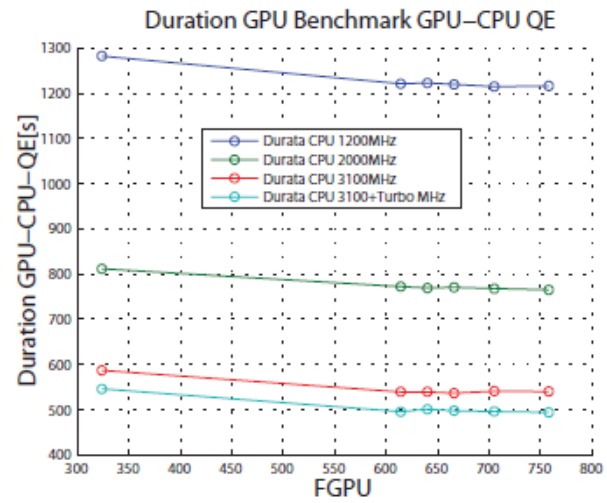
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

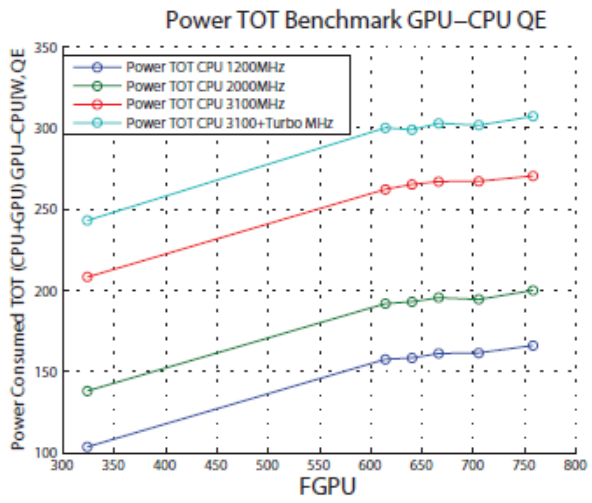


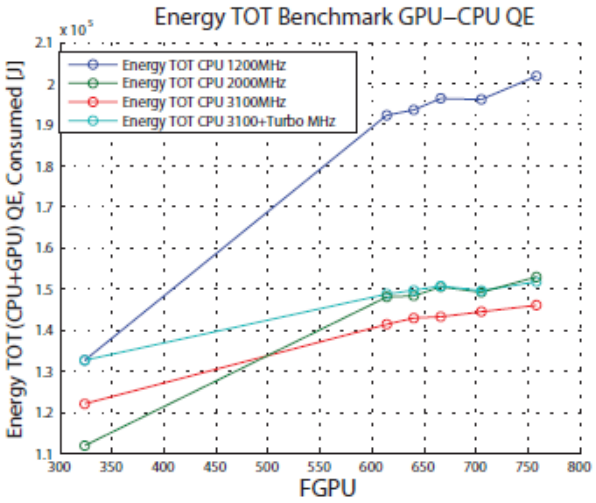




PV: 4x Power saving,5% Energy Gain (GPU DVFS) + 10% Energy Gain (CPU Min Freq)




QE: 2-3x Power saving,15% Energy Gain (GPU DVFS) + 5% Energy Gain (CPU DVFS)

# Outline

- **Power in digital systems**
- Dynamic Power Management
- Power Management & Heterogeneity
- Characterization of thermal effects

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Thermal Heterogenity Direct Liquid Cooling

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

System level

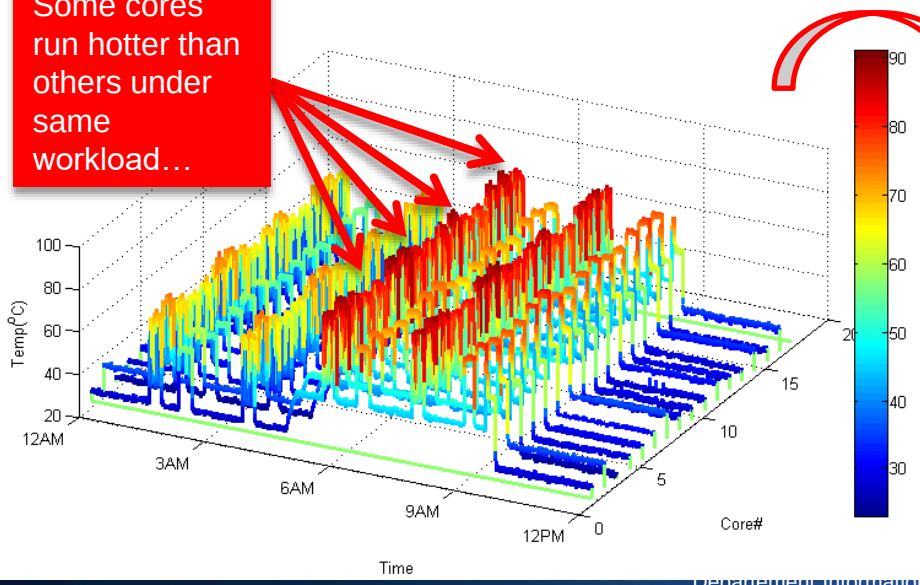
Board level

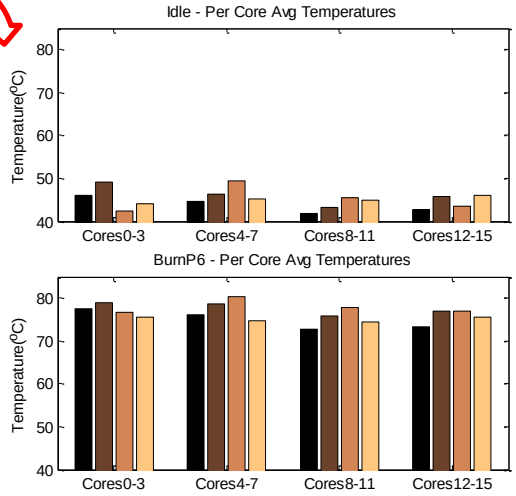
| Node n. | $R_{thCPU1}$ [°C/W] | $R_{thCPU2}$ [°C/W] | $R_{thGPU1}$ [°C/W] | $R_{thGPU2}$ [°C/W] | $\overline{T}$ [°C] |
|---|---|---|---|---|---|
| 33 | 0.3160 | 0.3081 | 0.1260 | 0.1242 | 40.1 |
| 34 | 0.3256 | 0.3080 | 0.1868 | 0.1823 | 39.8 |
| 35 | 0.2982 | 0.3083 | 0.1157 | 0.1203 | 39.7 |
| 36 | 0.3047 | 0.2985 | 0.1228 | 0.1213 | 40.1 |
| 37 | 0.3253 | 0.3162 | 0.1086 | 0.117 | 40.0 |

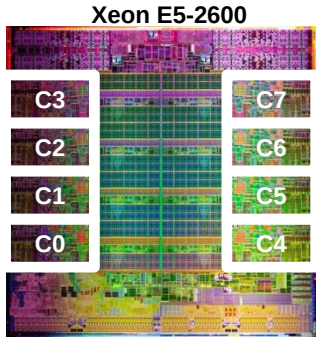
Thermal aware job schedulers can effectively allocate tasks to mitigate thermal gradients, thermal hazard, and average temperature.

Chip level

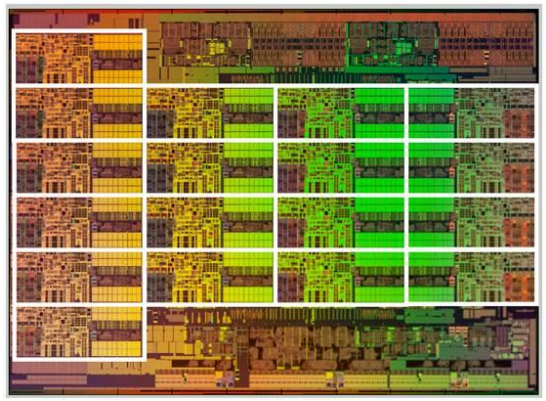
Some cores run hotter than others under same workload…

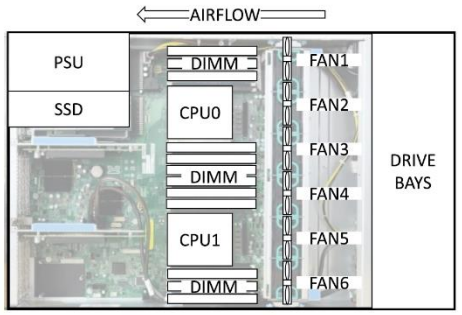


# Haswell – E5-2699 v3 Air Cooled

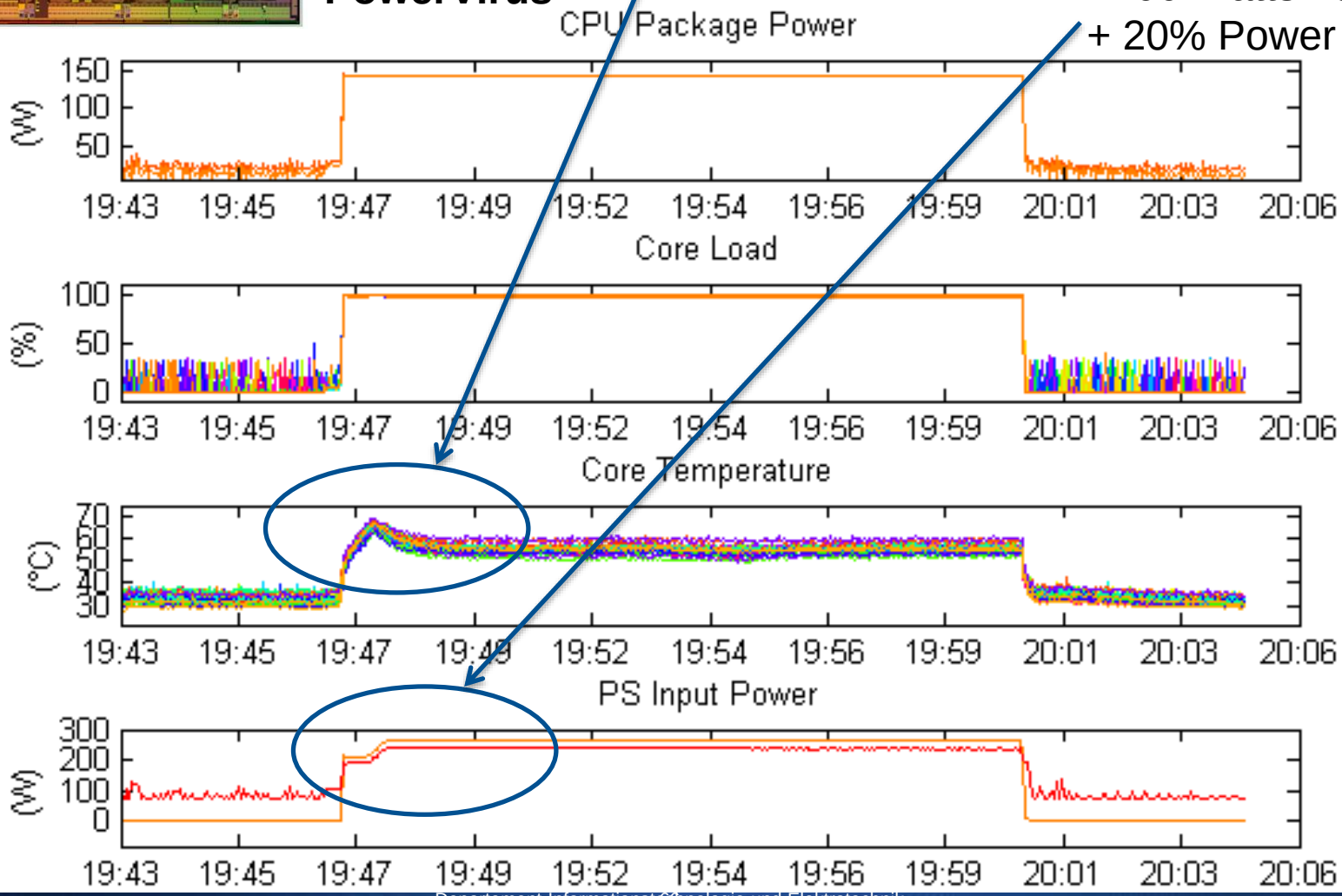




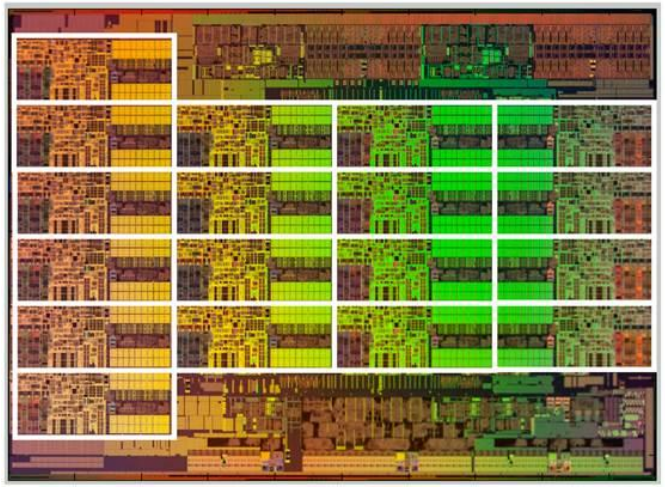

**18 cores PowerVirus**

~10°C Spatial variation

+100 Watts Fan + 20% Power

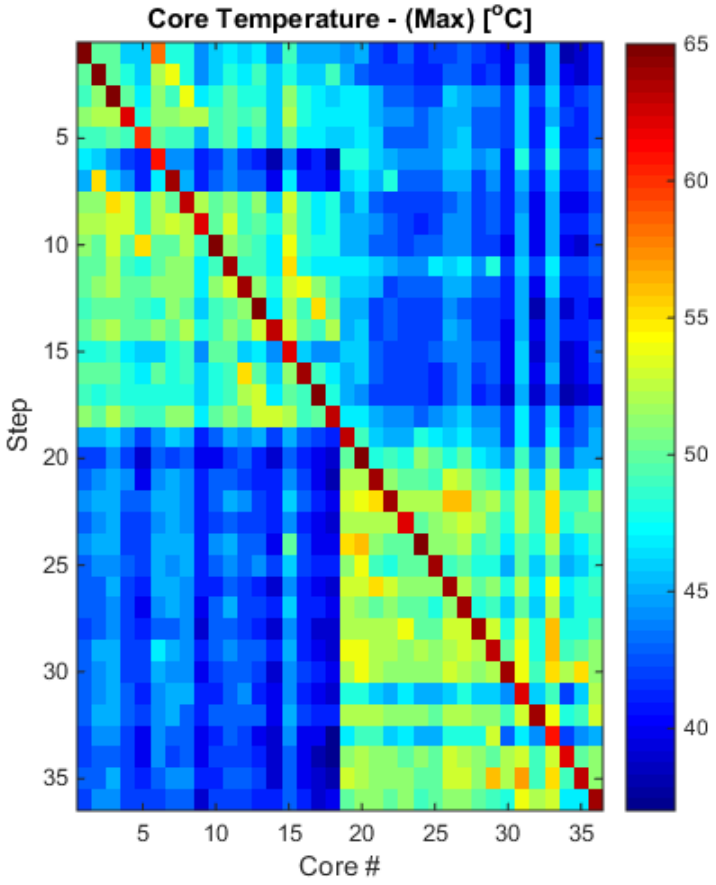


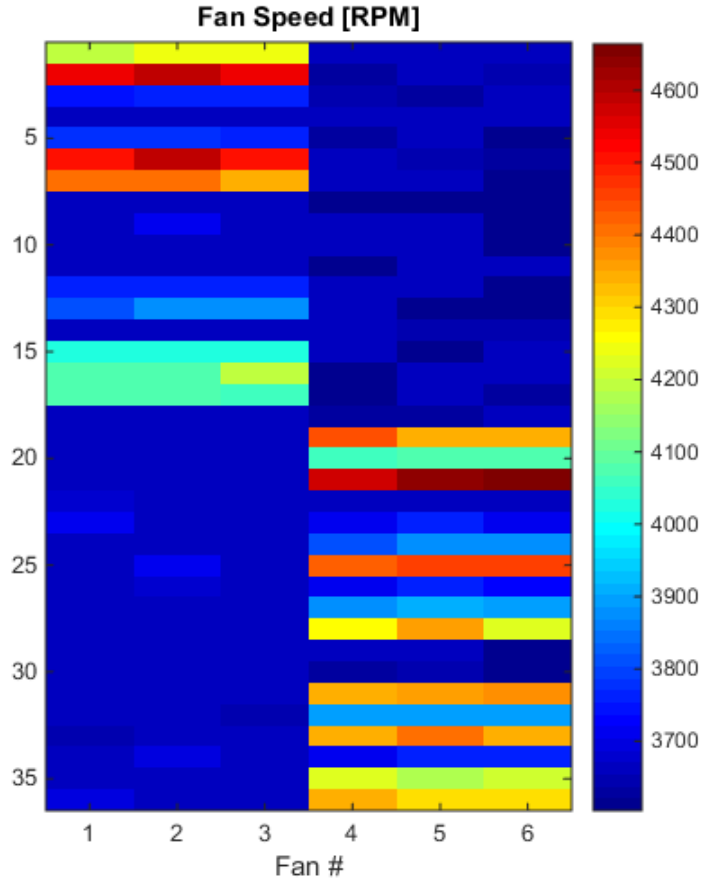
# Thermal Impact



## #1 Core Rotating Power Virus

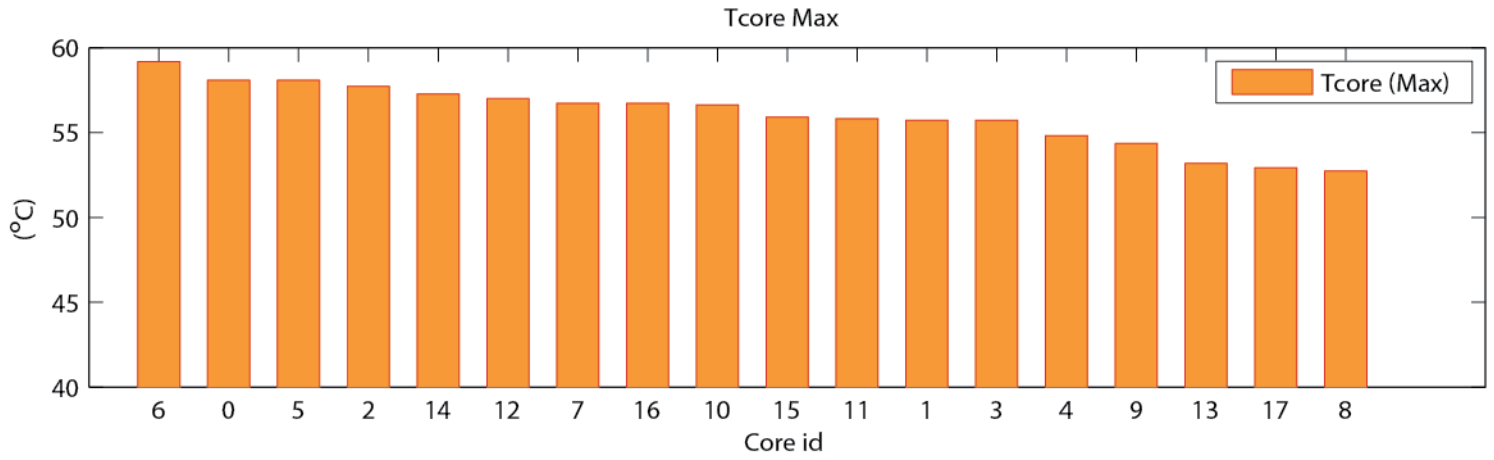
Up to 20 C Temperature difference on DIE
~ 30 C Temperature difference in between sockets
- Thermal neighbours exists!



# Hot & Cold Cores – Haswell – E5-2699 v3

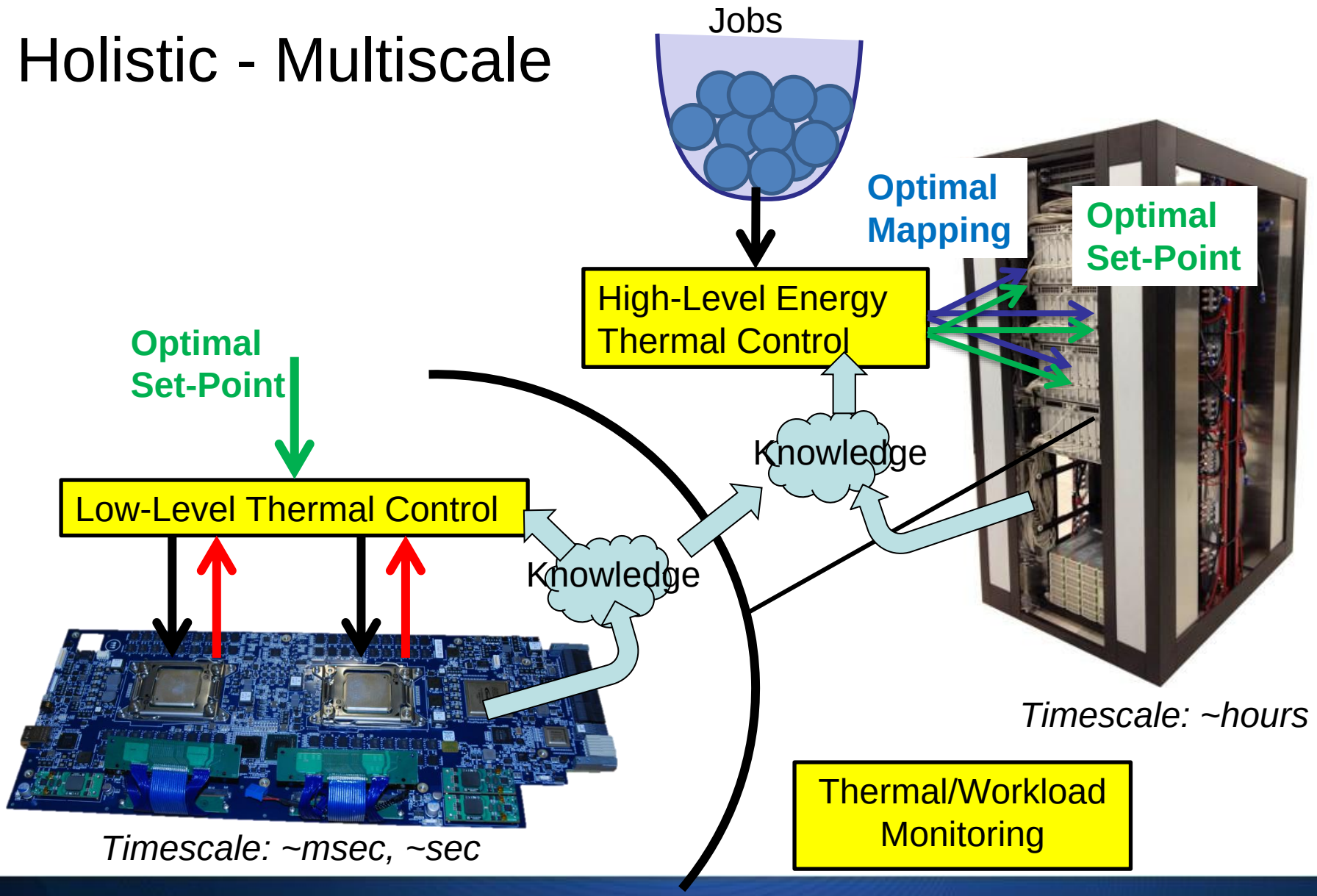
- Single core maximum temperature - ranking



# Supercomputers



## Holistic - Multiscale



# Biblio

*[TC 2014] Beneventi, Francesco, et al. "An effective gray-box identification procedure for multicore thermal modeling." Computers, IEEE Transactions on 63.5 (2014): 1097-1110.*

*[DATE 2014] Bartolini, Andrea, et al. "Unveiling eurora-thermal and power characterization of the most energy-efficient supercomputer in the world." Proceedings of the conference on Design, Automation & Test in Europe. European Design and Automation Association, 2014.*

[ISLPED 2014] Fraternali, Francesco, et al. "Quantifying the Impact of Variability on the Energy Efficiency for a Next-generation Ultra-green Supercomputer." *Proceedings of the 2014 international symposium on Low power electronics and design*. ACM, 2014.