

Making Sense of Recent Research in Temperature-Aware Design

Kevin Skadron

**Univ. of Virginia
LAVA Lab / HotSpot Group**

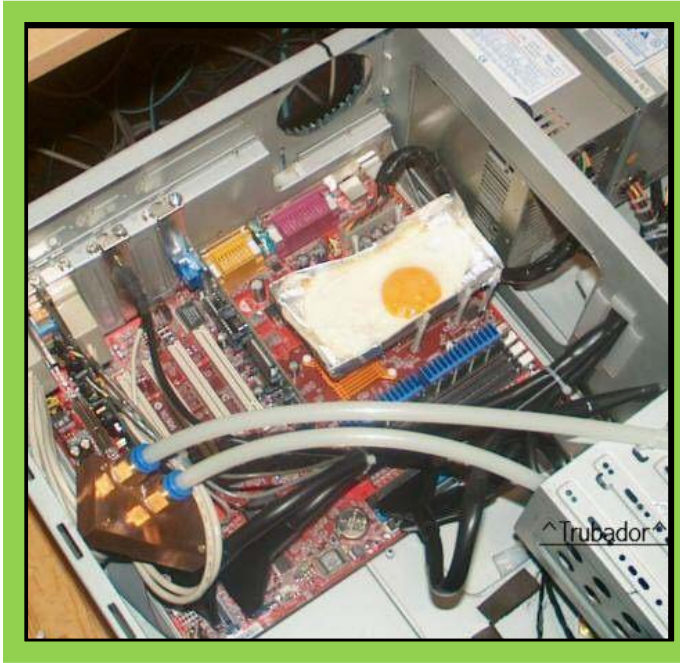


Temperature-Aware Design = Marathon

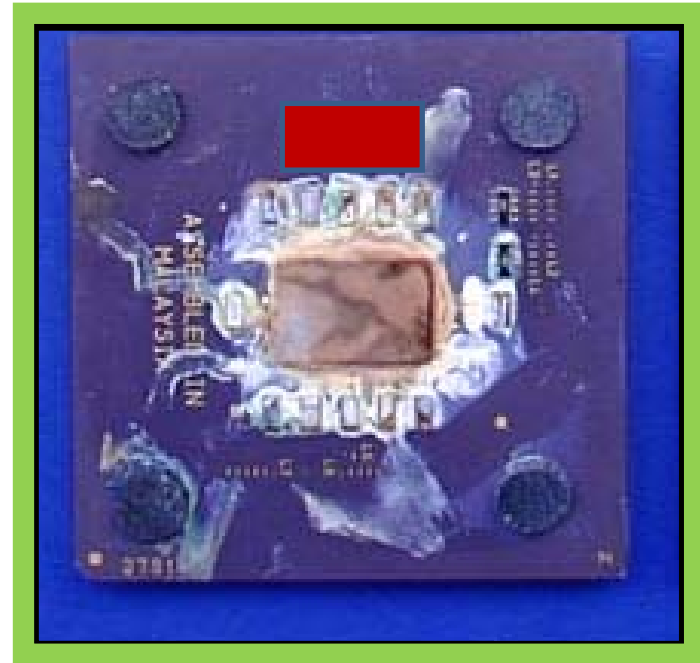
- **Marathoners pace themselves**
 - Keep an even pace for their target time
 - Too slow => not competitive
 - Too fast => they wear out
 - In the Greek legend, Pheidippides died!
 - Heat kills!
- **Multicore chips are like marathoners**



The Marathon Chip



Source: Trubador



Source: Tom's Hardware Guide
<http://www6.tomshardware.com/cpu/01q3/010917/heatvideo-01.html>

- **Speed => heat**
- **Don't want to be thermally limited**
 - **Wasted design and manufacturing costs**
- **Don't want to be Pheidippides**
 - **Angry customers!**

Key Differences: Power vs. Thermal

- **Energy ($P \times t$)**
 - Reclaim slack; want $E \propto$ work
 - Most benefit when system isn't working hard
 - Best effort
- **Power ($P \propto CV^2f$, $f \propto V$, so $P \propto CV^3$)**
 - Avoid voltage droops or damage to power supply
 - Can't exceed limits
 - Short timescales (nanoseconds—microseconds)
 - Must provision for worst-case expected workload

➤ Most important at high load
➤ Control sacrifices

performance

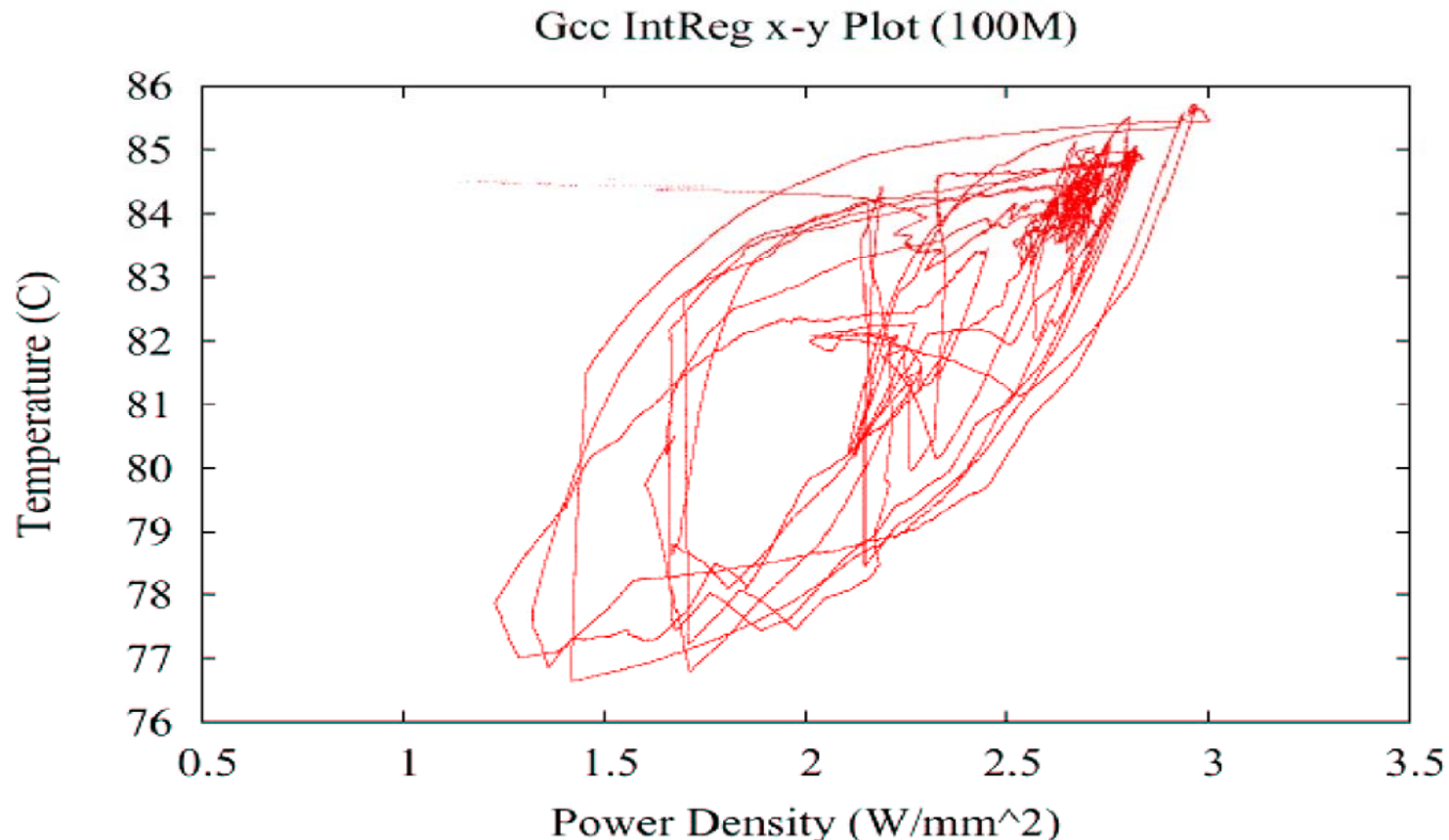
- **Thermal**
 - Can't exceed max temperature (often $\approx 100^\circ\text{C}$)
 - Actually, this is debatable
 - Long timescales (milliseconds or more)
 - Must provision for worst-case expected workload

Why P/E-aware \neq T-aware

- Thermal control operates over different time scales
- Lateral thermal coupling among units
- Lower P may mean higher T!
 - Turning off structures to reduce leakage or switched capacitance may increase power density in remaining area
- Saving *energy* can often be accomplished without affecting performance, due to slack
- *Thermal* throttling usually incurs a performance loss
- But the same hardware mechanisms may be used for all these objectives, eg. lowering voltage and frequency
 - It's still power dissipation that we're controlling
 - It's the control that matters

Thermal Modeling: P vs. T

- **Power metrics are an unacceptable proxy (IEEE Micro 2003)**
 - **Chip-wide average won't capture hot spots**
 - **Localized average won't capture lateral coupling**
 - **Different functional units have different power densities**



Thermal consequences

Temperature affects:

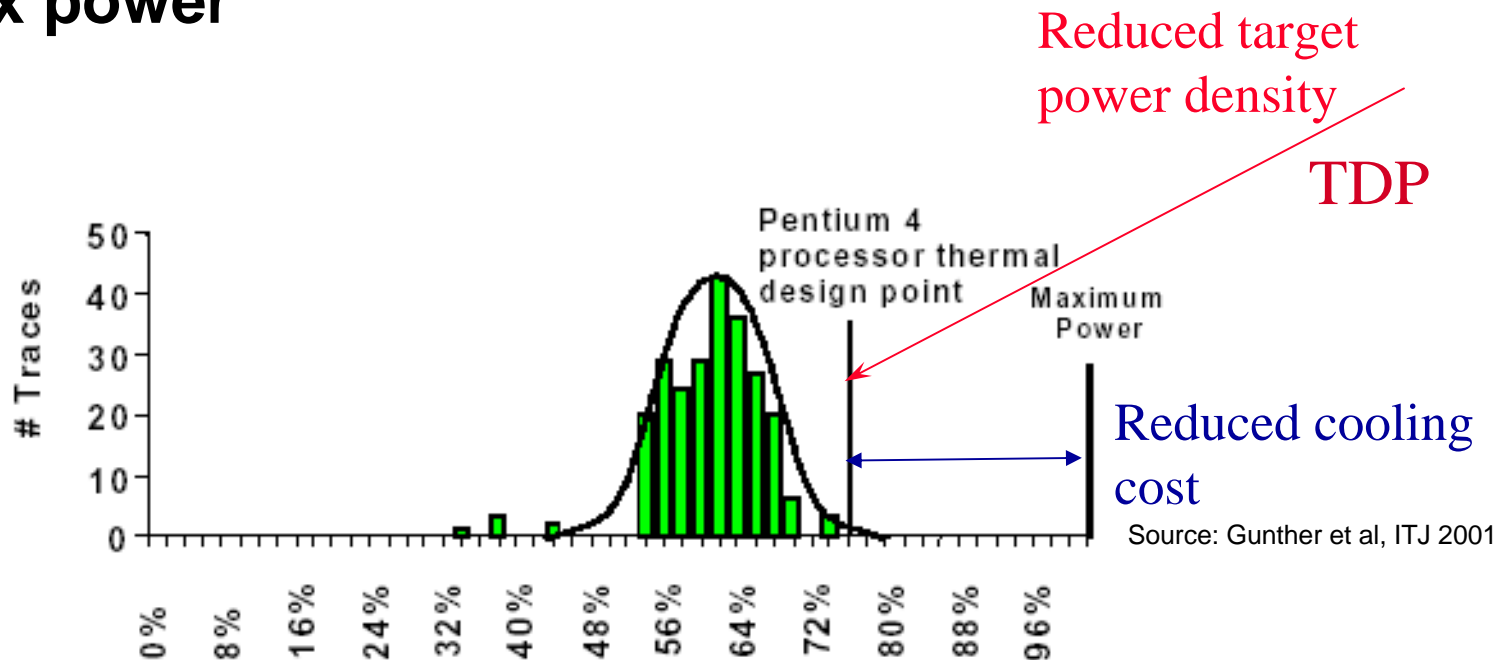
- **Circuit performance (possible timing errors)**
- **Circuit power (leakage)**
 - Exponential with temperature
- **IC failure rates**
 - Exponential with temperature
- **IC and system packaging/cooling cost**
 - Superlinear with power
- **Acoustics**
 - For PCs, this may be the real limit on cooling
- **Environment**
 - Rule of thumb: every 1 W of power in IC => 1 W of power spent on cooling

Outline

- **Single-core thermal management**
 - Design for TDP and throttling
- **Implications of multicore**
 - Why scaling will hit a power wall
 - Implications of asymmetric architectures
- **Reliability considerations**
 - How should we really be controlling temperature?
- **Pre-RTL compact thermal modeling for temperature aware architecture**
- **Lessons and research needs**

Design for TDP

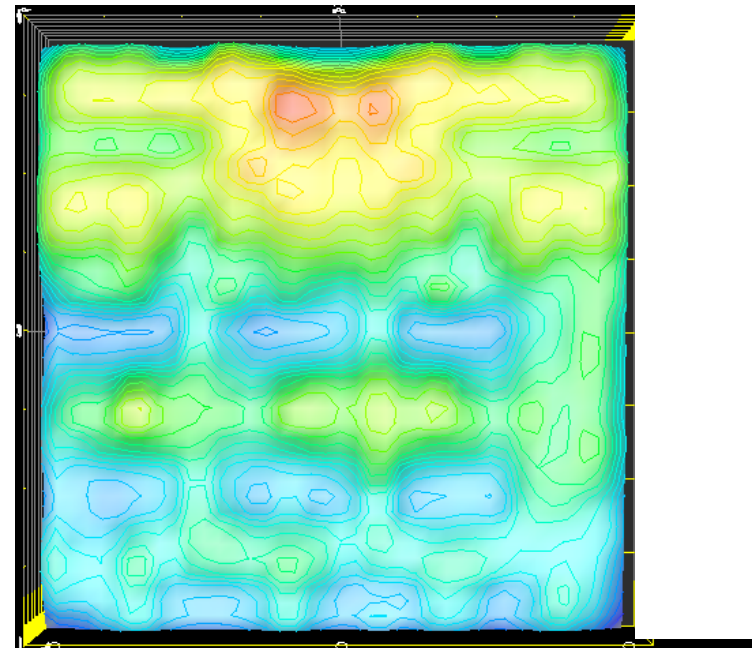
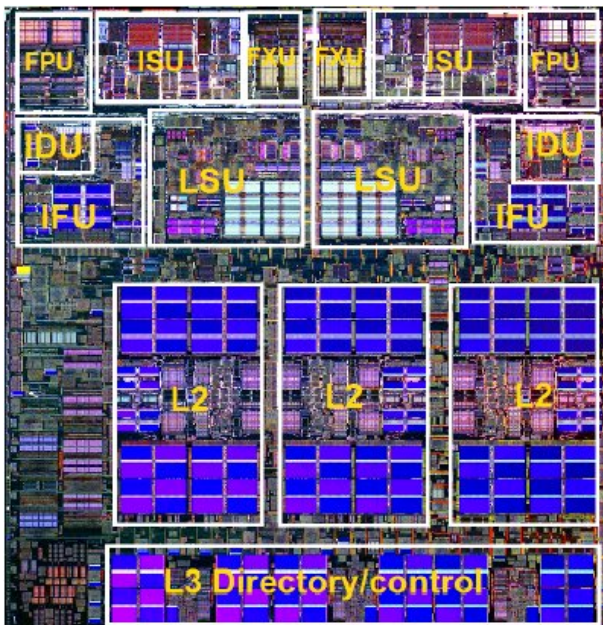
- Low-hanging fruit: don't design for a rare worst case
- Design for worst expected workload, not theoretical max power



- Throttle in rare cases where the chip overheats
 - Assumes an ill-behaved application
 - Or an ill-behaved user

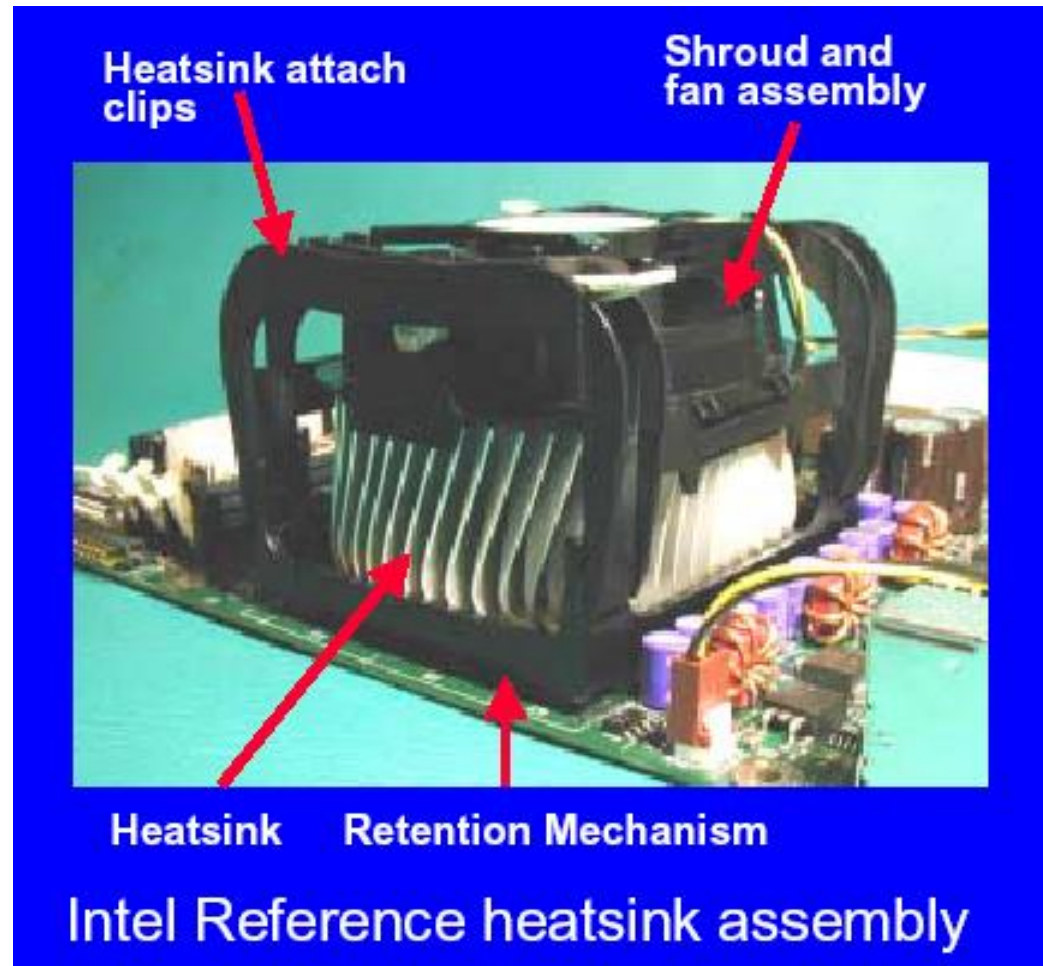
Cooling Dictated by Hotspots

- High cooling capacity “wasted” on most of the chip’s area



IBM POWER5

Intel Pentium 4 packaging



Source: Intel web site

Graphics Cards

- **Nvidia GeForce 5900 card**



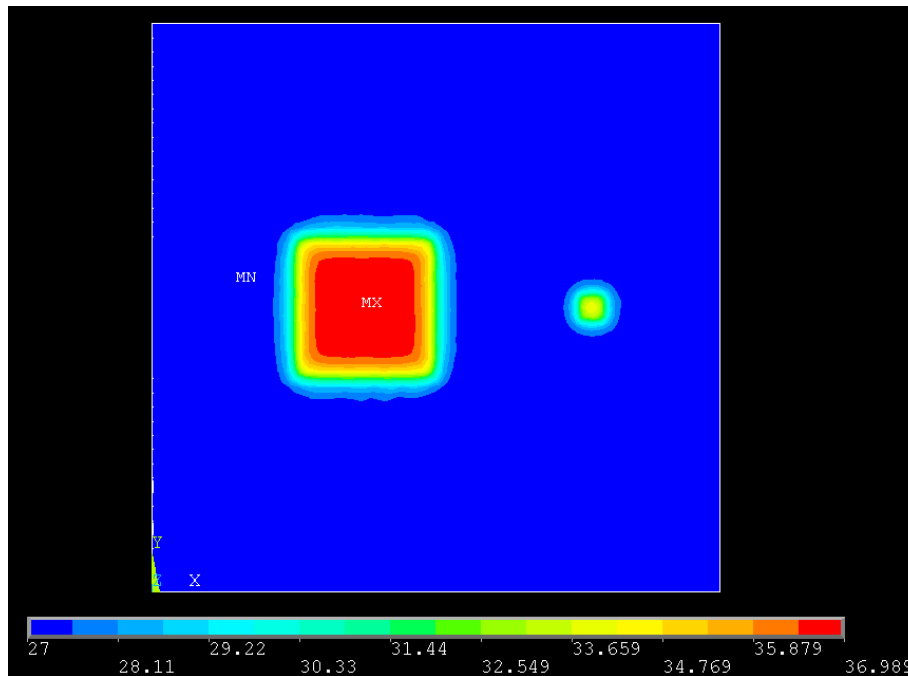
Source: Tech-Report.com

Traditional Throttling

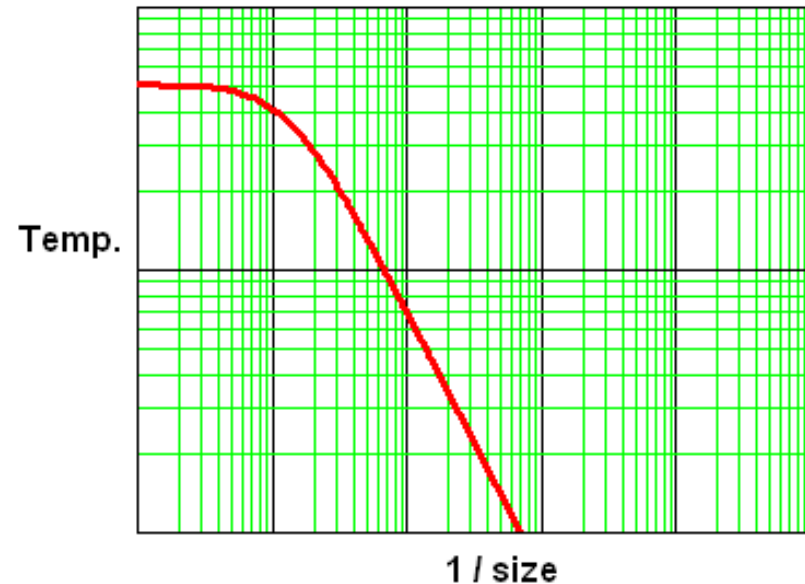
- **Dynamic voltage and frequency scaling (DVFS)**
 - Power reduction is cubic in voltage
- **Clock/fetch gating**
 - Gating is linear but low overhead
- **Hybrid of the two works better (DATE'04)**
- **Some research looked at throttling or load shifting *within* cores**
 - Units and even cores getting too small for this to matter
 - Unless targeting an individual unit allows lower performance penalty

A Brief Aside - HotSpot Granularity

- **Thermal low-pass filtering effect**
 - **At the same power density, small heat sources produce lower peak temperatures**



same power density = $2\text{W}/\text{mm}^2$



Role of Throttling

- **Old thinking: throttling is only a rare failsafe**
 - Thermal solution should be designed for a TDP safely above workloads of interest
 - Never incur slowdown in normal operating conditions
- **Today: some chips are already thermally limited**
 - Poor scaling trends
 - *Multicore* makes it easier to dissipate high total power
 - Market constraints may limit cost of thermal solution (even if it could be fully cooled)
 - Better throttling may be preferable to a brute-force frequency limit

Throttling Considerations

- **Throttling sacrifices throughput, performance**
- **Redistributing heat in space may have lower overhead**
 - Better floorplanning
 - Scheduling of incoming tasks
 - Task migration (“core hopping”)
 - Many papers on these topics
 - But core hopping may not be possible if all cores are hot and tasks are long-running
- **Individual units within a core too small to throttle individually**
 - Throttling should happen at granularity of cores
 - Unless finer-grained throttling reduces perf penalty
 - Even per-core throttling will become ineffective
 - Throttling of “core groups” has not been studied



Source: http://www.guy-sports.com/fun_pictures/computerStrangle.jpg

Why Not Better Cooling?

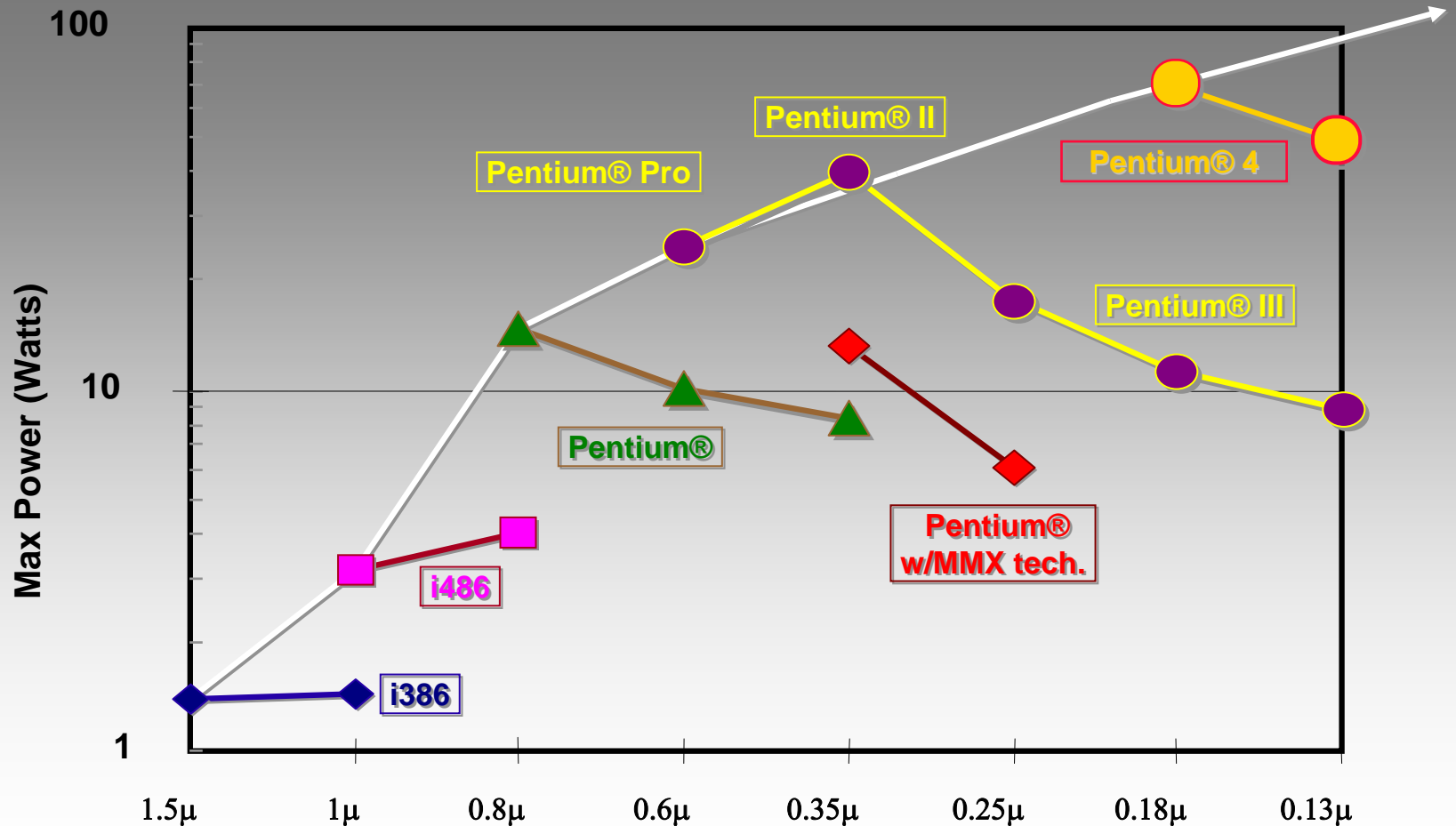
- **Cost**
 - \$10-15 seems to be the limit even for the high end (Borkar, Intel)
 - Low-cost market segments will have even lower budgets
 - Often cheaper to make one high-end design and scale it down for lower market segments
 - Scaled down chips may not “fit” their cooling
- **At the high end, we are on the verge of exceeding air-cooling capabilities (and acoustic limits)**
 - **Need new cooling with both:**
 - Manufacturing economies of scale
 - Design economies of scale
- **Also, single-core chips couldn't benefit well enough**
 - Which is how we got to multicore

Outline

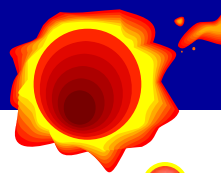
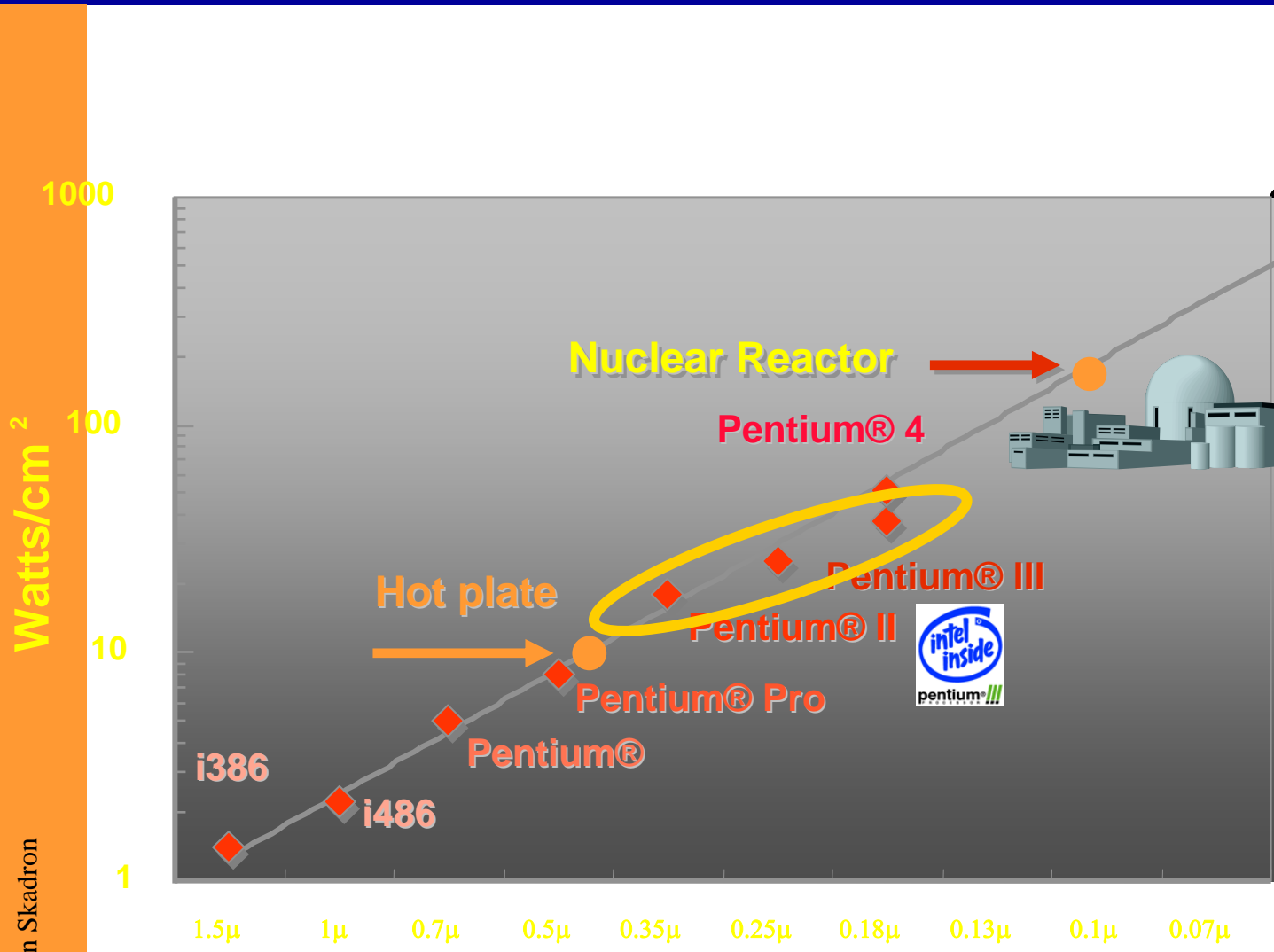
- **Single-core thermal management**
 - Design for TDP and throttling
- **Implications of multicore**
 - Why scaling will hit a power wall
 - Implications of asymmetric architectures
- **Reliability considerations**
 - How should we really be controlling temperature?
- **Pre-RTL compact thermal modeling for temperature aware architecture**
- **Lessons and research needs**

The Old Power Wall

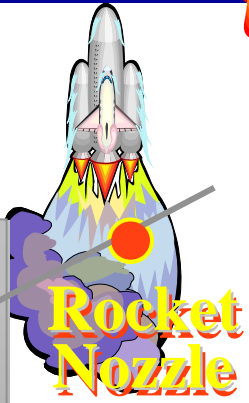
- Power density due to core microarchitecture
 - Highly ported structures, massive speculation



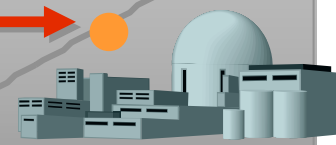
Trends in Power Density



Sun's Surface



Rocket Nozzle



Nuclear Reactor

Hot plate



* "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies" – Fred Pollack, Intel Corp. Micro32 conference key note - 1999.

Solutions to the Old Power Wall

- **Process technology, esp. Vdd scaling**
- **Clock gating**
- **Power-aware circuit design**
- **Leakage-aware SRAM design**
- **Reduced speculation**
- **Less aggressive microarchitectures**

Some good news and some bad news...

Why that Power Wall is Old

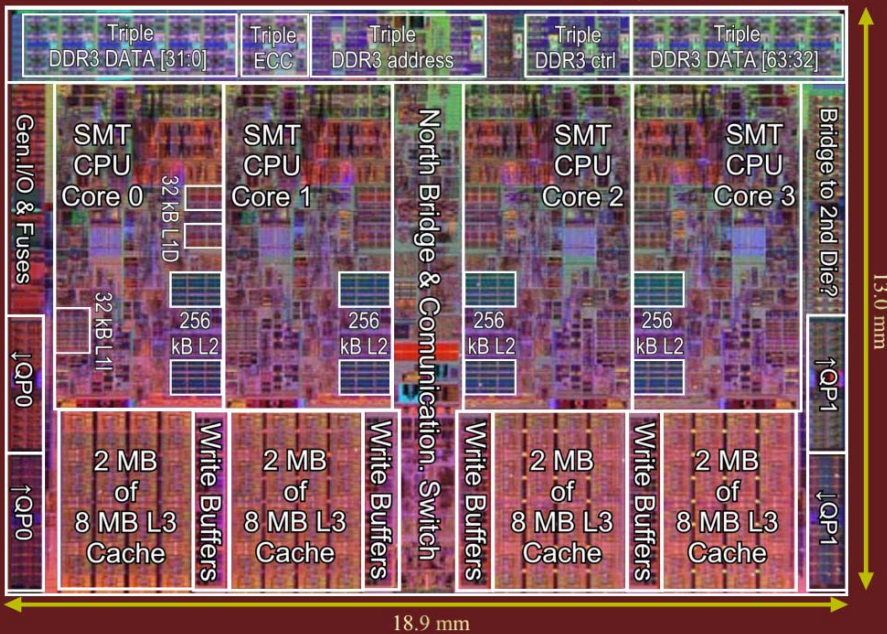
- **Individual cores not likely to get much more aggressive**
- **Combination of “ILP wall,” “frequency wall,” and “power wall”**
 - **ILP wall: can’t figure out how to keep increasing ILP without unreasonable area and power costs**
 - **Out of order issue, etc. ran out of steam**
 - **Frequency wall: can’t figure out how to increase clock frequency without unreasonable increase in power**
 - **Limited to 20-30%/generation from semiconductor scaling**
 - **Power wall: air cooling capped at ~150W (thermal design power - TDP)**
- **Moore’s Law is providing area that a single thread can’t economically use**
 - **How much cache does one core need?**
- **How to maintain ASPs?**
- **Area not spent on ILP can be spent on more cores !**
- **Small simplifications in core complexity yield large reductions in power**

Multicores

Intel Quad Core Nehalem

731 million transistors --- 8 MB L3 plus 4 x 256 kB L2 --- 3x64bit DDR3 bus
 2x Quick path I/O --- Single core size: ~24.4 mm² (excl L2)
 L2 cache tiles: 7.1 mm² / MB, L3 cache tiles: 5.7 mm² / MB (excl.tags)

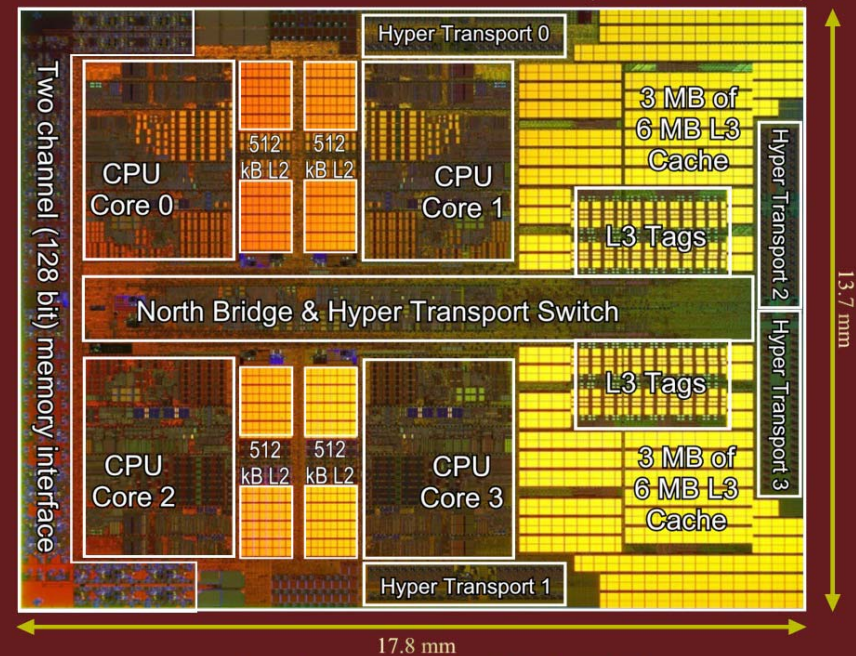
Die size 246 mm² (incl. test circ.265 mm²)



AMD Quad Core Shanghai

~705 million transistors --- 6 MB L3 plus 4 x 512 kB L2 --- 128 bit DDR2/3 bus
 4x HyperTransport I/O --- Single core size: ~15.3 mm² (excl L2)
 L2 cache tiles: 7.5 mm² / MB, L3 cache tiles: 7.5 mm² / MB (excl.tags)

Die size 243 mm² (incl. test circ.263 mm²)



www.chip-architect.com --- Rev.2 March-17, 2008

source: chip-architect.com/news/Shanghai_Nehalem.jpg

source: chip-architect.com/news/Shanghai_Nehalem.jpg

Moore's Law and Dennard Scaling

The way things should work...

- **Moore's Law: transistor density doubles every N years (currently $N \sim 2$)**
- **Dennard Scaling (constant electric field)**
 - Shrink feature size by k (typ. 0.7), hold electric field constant
 - Area scales by k^2 ($1/2$), C , V , delay reduce by k
 - $P \cong CV^2f \Rightarrow P$ goes down by k^2
 - Power density = $P/A = 1$

The Real Power Wall

- **Vdd scaling is coming to a halt**
 - **Currently 0.9-1.0V, scaling only ~2.5%/gen [ITRS'06]**
- **Vdd reductions were stopped by leakage**
- **Lower Vdd => Vth must be lower**
- **Leakage is exponential in Vth**
- **Leakage is also exponential in T**

The Real Power Wall

- Even if we generously assume C scales and frequency is flat
 - $P \cong CV^2f \Rightarrow 0.7 (0.975^2) (1) = 0.66$
- Power *density* goes up
 - $P/A = 0.66/0.5 = 1.33$
 - And this is very optimistic, because C probably scales more like 0.8 or 0.9, and we want frequency to go up, so a more likely number is **1.5-1.75X**
- If we keep %-area dedicated to all the cores the same -- total power goes up by same factor
- But max TDP for air cooling is expected to stay flat
 - Around 200-250 W total and around 1.5 W/mm²
 - Viable, *affordable* alternatives not yet apparent
- Multicore allows power to scale linearly with # cores
- The shift to multicore ***does not eliminate the wall***

Low-Fat Cores???



PClaes Oldenburg, *Apple Core – Autumn*
<http://www.greenwicharts.org/pastshows.asp>

Not Many Options

- **Almost all power savings are one-offs**
- **We need to come up with a stream of these**
 - **Fine-grained clock gating – done**
 - **Fine-grained power gating – happening**
 - **Better power-aware circuit synthesis – happening**
 - **Simplified cores – happening**
 - **SIMD/vector organizations help**
 - **Multiple voltage domains – starting**
 - **GALS (reduce clock tree) – on the horizon**
 - **Reduce margins, maybe run at ragged edge of reliability and recover, allowing lower Vdd – ???**
- **Running out of opportunities**

Where We are Today - Multicore

Programmability wall

Power wall



Classic architectures

Implications of Multicore

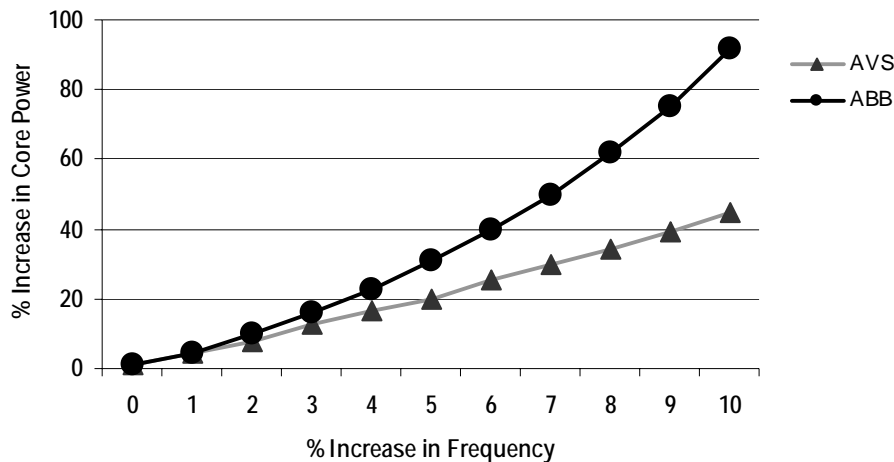
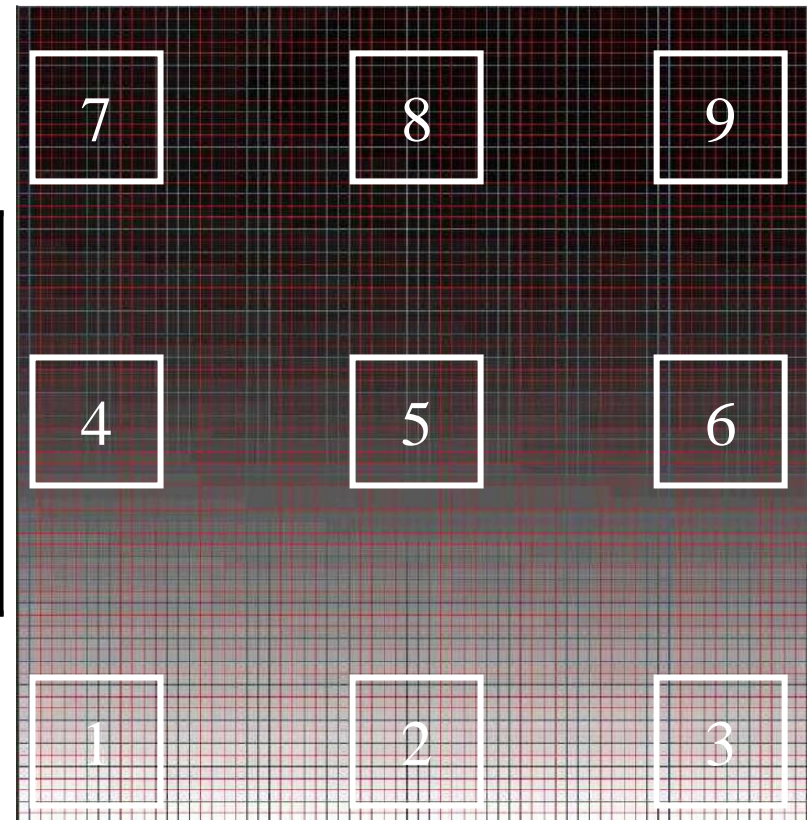
- **New opportunities**
 - Thermal-aware load balancing
 - Overdrive (eg, Core i7)
- **New problems**
 - Network-on-chip a major power dissipator (Shang et al., MICRO 2004)
 - Cores with low activity could become hot if neighboring router becomes hot
 - Throttle/re-route capability requires an NOC with alternate routes (e.g., not rings)
 - Parallel workloads may have uniform work distributions
 - Especially true with deeply multithreaded cores
 - So all cores are hot!
 - ~~At least multicore makes the power dissipation more uniform, right?~~
 - Core-to-core parameter variations

Process Variations

- **Process variations manifest themselves in a variety of ways**
 - **Within-Die (WID)**
 - Delay of critical path limited by slowest device
 - **Die-to-Die (D2D), Wafer-to-Wafer (W2W)**
 - Distribution of performance, leakage across chips
 - **Core-to-Core (C2C) – DATE'07**
 - Due to “process tilt”
- **Identical cores on the same chip will have different performance and leakage**
 - **Compensation exacerbates thermal challenges or leaves performance on the table**
 - **Requires scheduler to be aware of performance and thermal heterogeneity**
 - **Why not design with heterogeneous cores?**

Pre-compensation C2C variation

	Mean norm. freq	Mean norm. power
Row 1 (Cores 7-9)	.995 ± .005	1.00 ± .002
Row 2 (Cores 4-6)	.952 ± .004	.950 ± .004
Row 3 (Cores 1-3)	.826 ± .002	.814 ± .002



Asymmetric Organizations

- **Small number of aggressive cores for threads that don't parallelize**
- **Large number of simple cores for throughput**
 - **Power roughly linear in core size**
 - **Performance \propto square root of core size (Pollack's rule)**
 - **With sufficient threads, smaller cores boost performance/W and performance/mm²**
- **Use some area for coprocessors**
 - **GPU, media processor, perhaps crypto**
 - **Trade off flexibility for power efficiency**

The Manycore Orchestra?

The New York Times

Faster Chips Are Leaving Programmers in Their Dust

By John Markoff

Published: December 17, 2007

[Mundie] envisions modern chips that will increasingly resemble musical orchestras. Rather than having tiled arrays of identical processors, the microprocessor of the future will include many different computing cores, each built to solve a specific type of problem. A.M.D. has already announced its intent to blend both graphics and traditional processing units onto a single piece of silicon.

Thermal Implications of Asymmetry

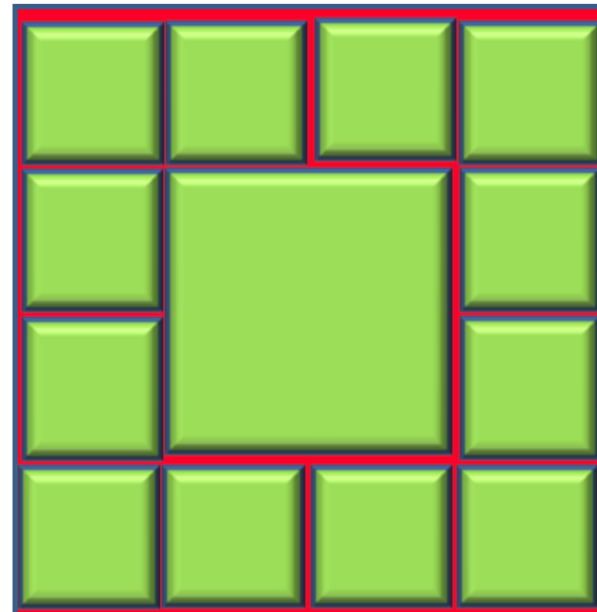
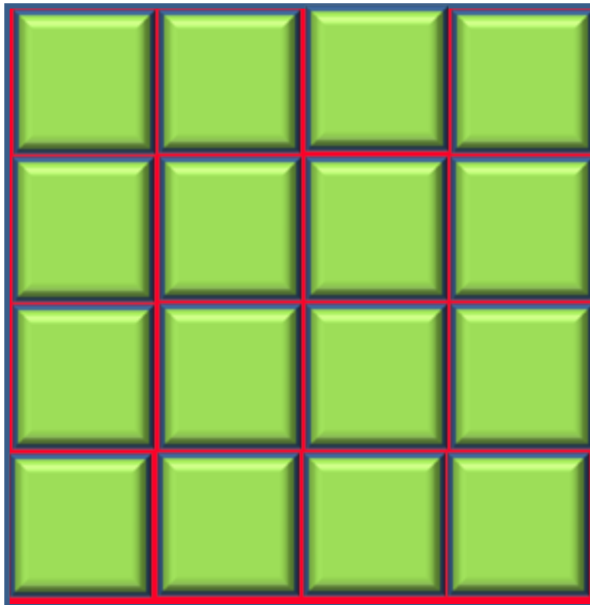
Marty and Hill, IEEE Computer 2008

$$\text{Speedup} = \frac{1.0}{(1-f) + \frac{f}{N}}$$

f is fraction of parallelized workload
(from 0 to 1.0)

Serial part

Parallel part



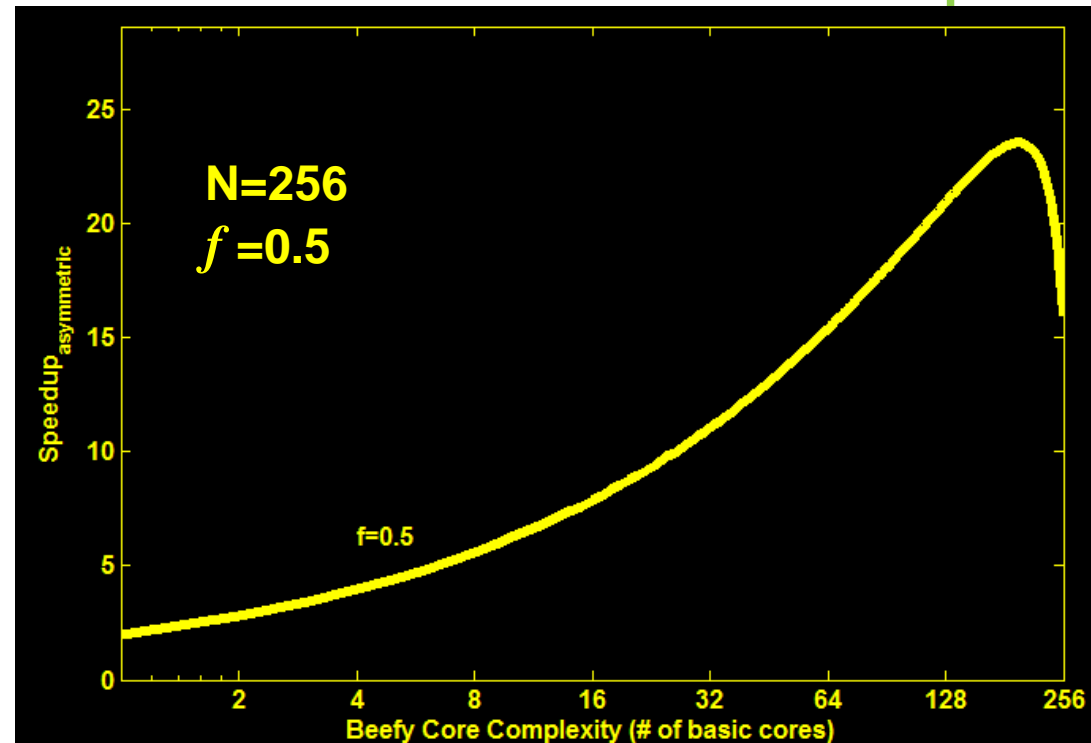
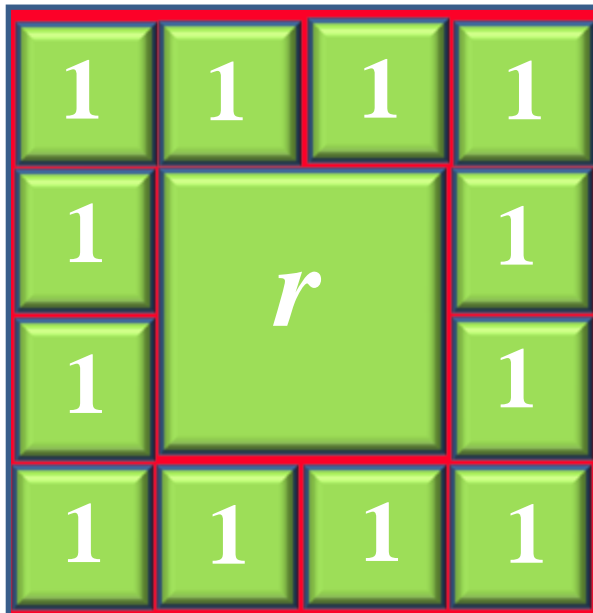
Thermal Implications of Asymmetry

$$Perf_{serial} = \text{sqrt}(r) \quad (\text{Pollack's Rule})$$

$$Speedup(f, N, r) = \frac{1.0}{\frac{(1-f)}{perf_{serial}} + \frac{f}{perf_{serial} + (N-r)}}$$

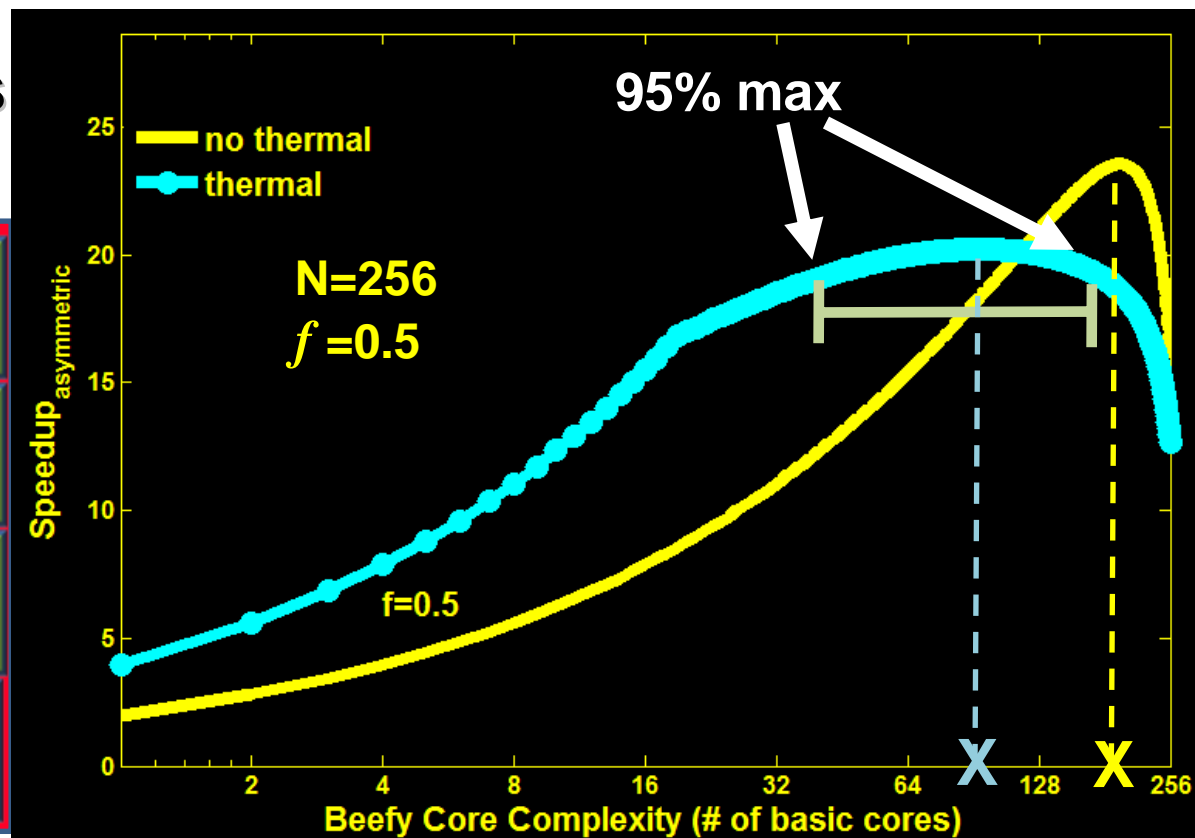
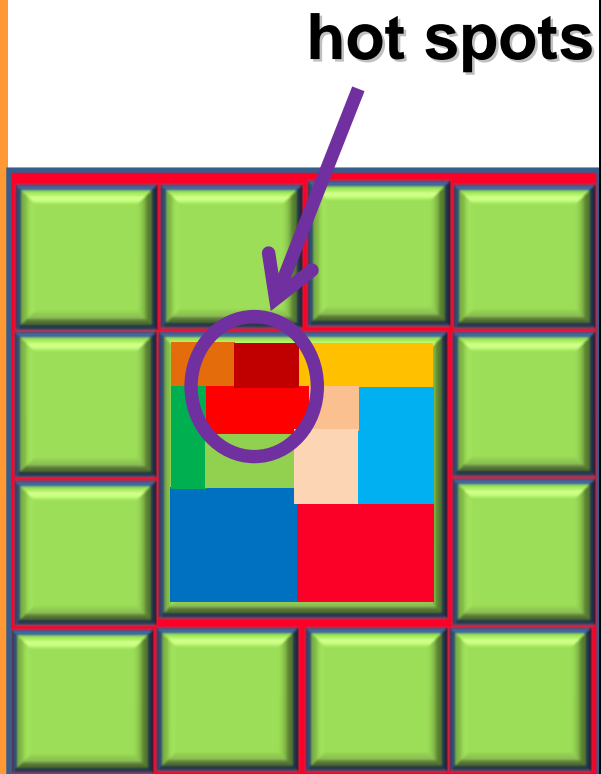
Serial part

Parallel part



Thermal Implications of Asymmetry

- Thermal limits reduce performance
- But simplify design
 - Larger cores don't help as much



3D

- **Short-term appeal is high-bandwidth memory**
- **Long-term appeal is to deal with scaling limits**
 - **Die size (reticle limit)**
 - **End to Moore's Law**
 - **3D would allow scaling within a single socket**
 - **Fast inter-layer connections**
 - **The new Moore's Law?**
 - **Many papers starting to appear on this topic**
 - **Huge thermal challenges**
 - **Surface cooling no longer sufficient**
 - **Need inter-layer cooling**

Outline

- **Single-core thermal management**
 - Design for TDP and throttling
- **Implications of multicore**
 - Why scaling will hit a power wall
 - Implications of asymmetric architectures
- **Reliability considerations**
 - How should we really be controlling temperature?
- **Pre-RTL compact thermal modeling for temperature aware architecture**
- **Lessons and research needs**

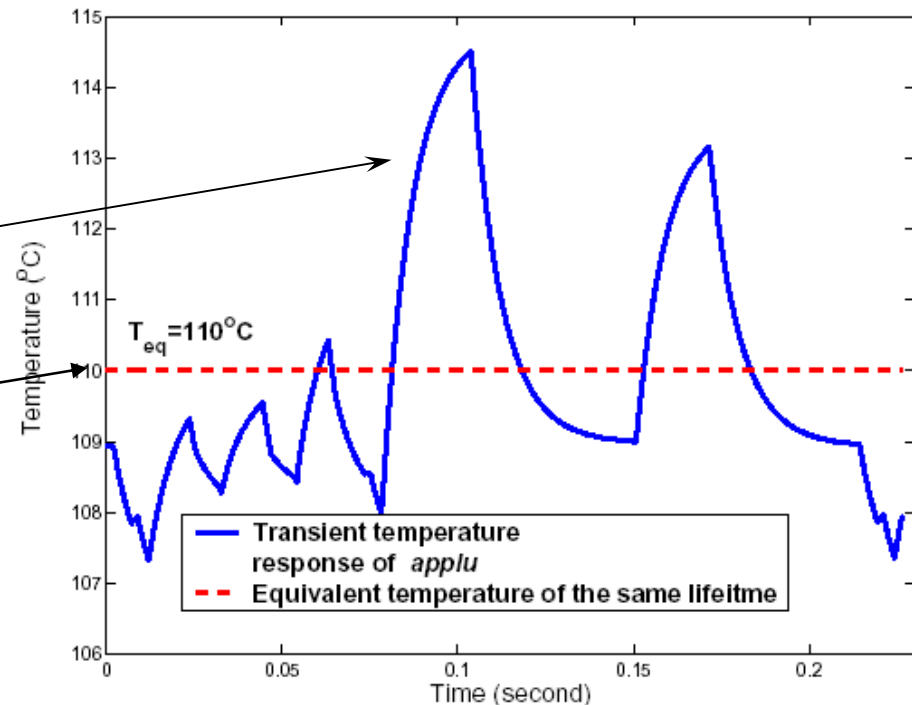
Reliability

- **Are strict temperature limits necessary?**
 - Timing errors
 - Aging
 - Thermo-mechanical stress
- **Architects don't know what to design for**
 - Timing errors => throttle to meet timing, let temperature exceed threshold
 - Aging => reliability “banking”
 - Thermo-mechanical
 - What are the time constants (how long can throttling be delayed?)
 - Do we also need to be worried about cycling?
- **This matters to chip design**
 - Even if we target the same temperature

Aging as $f(T)$

- Reliability criteria (e.g., DTM thresholds) are typically based on worst-case assumptions
- But actual behavior is often not worst case
- So aging occurs more slowly
- This means the DTM design is over-engineered!
- We can exploit this, e.g. for DTM or frequency (IEEE Micro 2005)

Spend
Bank



Sensing

- **When throttling is a failsafe, sensing can be simple and imprecise**
 - **Use generous offsets**
- **In a thermally limited era, sensing must be precise**
 - **How many sensors, where to put them?**
 - **Many papers on this topic**
 - **Need a sensor at every candidate hotspot**
 - **Process variations, TIM variations add significant complications**
 - **Every chip could have different hotspots**
 - **TIM variations could create hotspots in areas with lower power density!**
 - **Activity becomes a poor predictor of temp**

Outline

- **Single-core thermal management**
 - Design for TDP and throttling
- **Implications of multicore**
 - Why scaling will hit a power wall
 - Implications of asymmetric architectures
- **Reliability considerations**
 - How should we really be controlling temperature?
- **Pre-RTL compact thermal modeling for temperature aware architecture**
- **Lessons and research needs**

Pre-RTL Thermal Modeling

- **Want a fine-grained, dynamic model of *temperature***
 - At a granularity architects can reason about
 - That accounts for adjacency and package
 - For early design exploration
 - That is fast enough for practical use
- **HotSpot - compact model**
 - Parameterized to automatically derive a model based on various
 - Architectures
 - Power models
 - Floorplans
 - Thermal Packages
 - Downloaded 1800+ times, preparing ver. 5
 - Latest improvements described in ISPASS'09, IEEE Trans. Computers 2008

Architectural Compact Modeling

Electrical-thermal duality

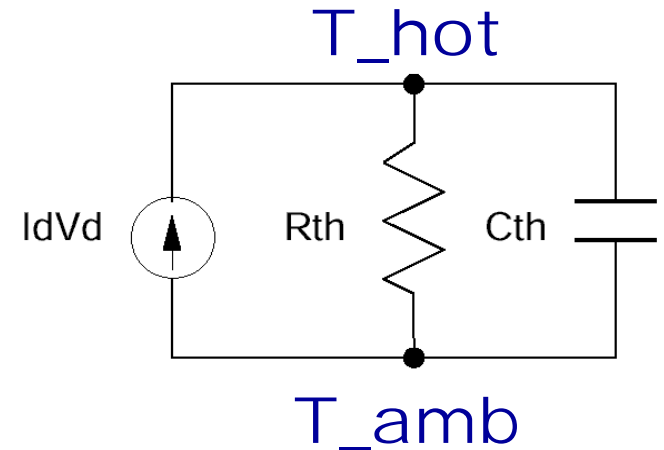
$V \cong \text{temp (T)}$

$I \cong \text{power (P)}$

$R \cong \text{thermal resistance (Rth)}$

$C \cong \text{thermal capacitance (Cth)}$

RC time constant (Rth Cth)



Kirchoff Current Law

differential eq. $I = C \cdot dV/dt + V/R$

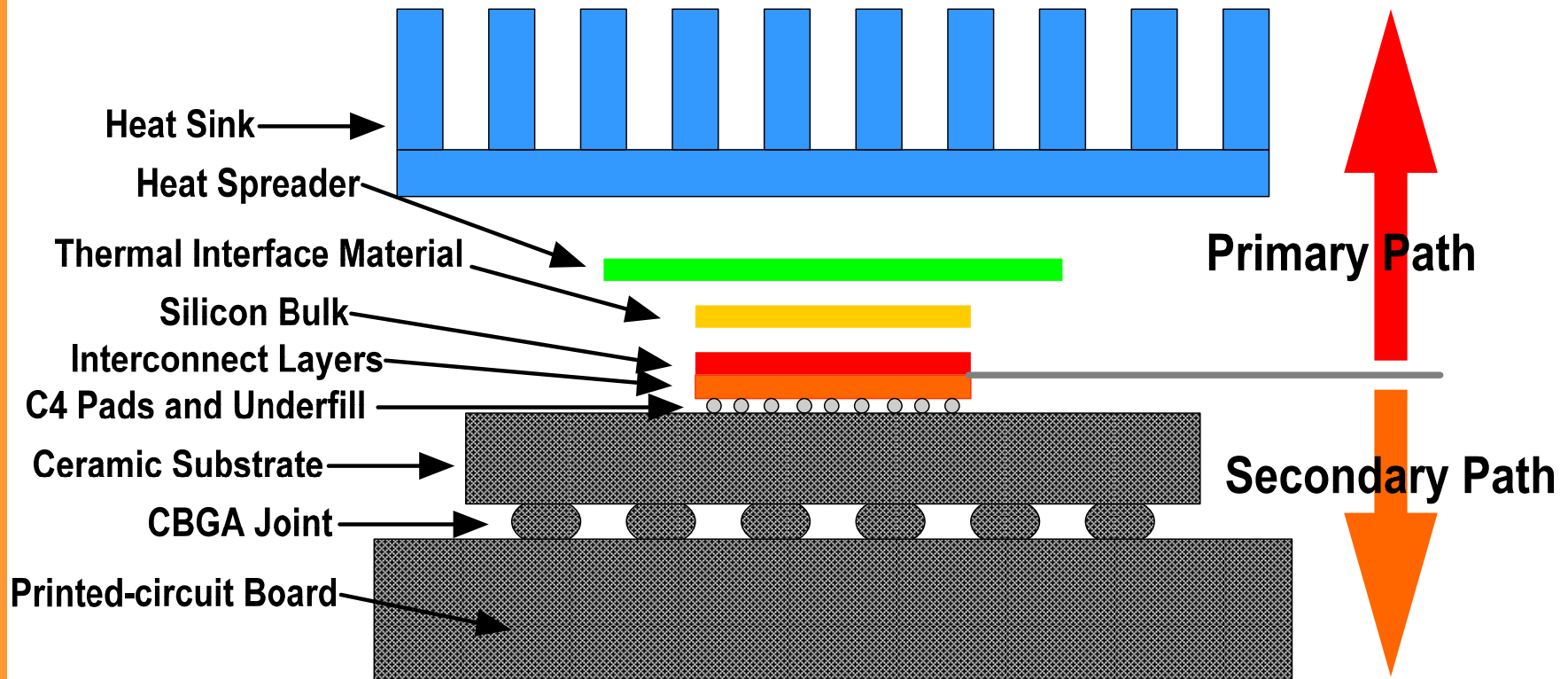
thermal domain $P = C_{th} \cdot dT/dt + T/R_{th}$

where $T = T_{hot} - T_{amb}$

At higher granularities of P, Rth, Cth

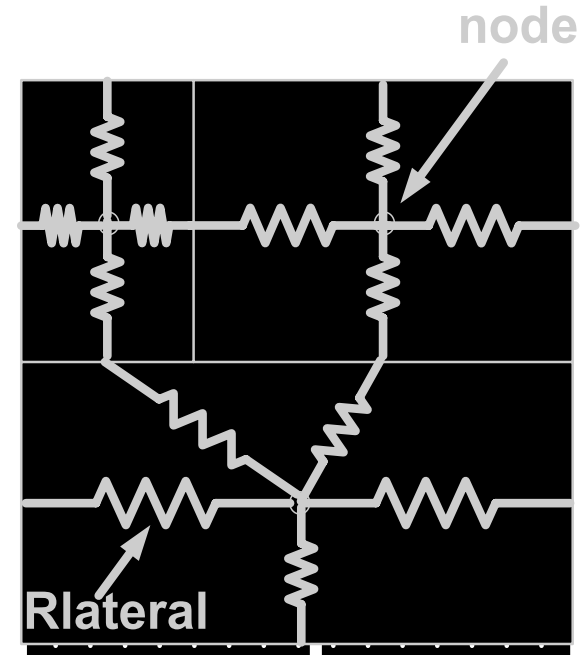
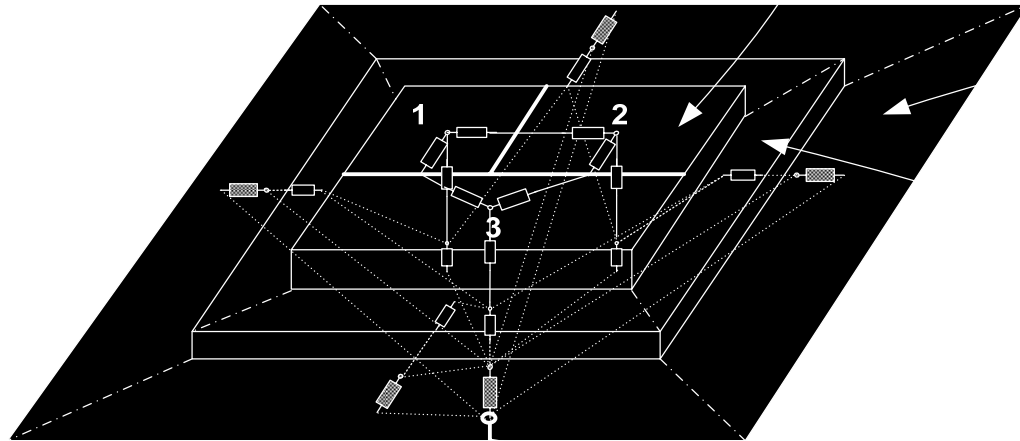
P, T are vectors and Rth, Cth are circuit matrices

HotSpot Structure



- **Multiple layers**
- **Both silicon and package**
- **Primary and secondary paths**
- **Can add more layers for 3D chips**

HotSpot Structure



- Vertical and lateral thermal resistors
- Capacitor at each node for transient
- Irregular blocks or regular grid cells
- Can model multiple Si layers in 3D chips
- Validation: FEM tools, test chip, sensors, infrared

Outline

- **Single-core thermal management**
 - Design for TDP and throttling
- **Implications of multicore**
 - Why scaling will hit a power wall
 - Implications of asymmetric architectures
- **Reliability considerations**
 - How should we really be controlling temperature?
- **Pre-RTL compact thermal modeling for temperature aware architecture**
- **Lessons and research needs**

Lessons

- **Power/energy management differ from thermal management**
- **Runtime temperature management becoming more important**
 - Try to distribute power in space, not time
 - But throttling, which is dynamic, may still be better than static limits
 - e.g., Limiting frequency in thermally limited parts
- **Controlling individual units within cores probably not useful**
 - Controlling individual cores may not even be enough
- **Asymmetric organizations mean that hotspots will remain a problem**
- **Ideal semiconductor scaling breaking down re: power density**
 - And we are running out of architectural/circuit techniques to save power
- **Workloads exhibit considerable variation within and across programs**

Research Needs

- **New, affordable, economical cooling solutions**
 - Acoustic improvements to air cooling would help too
- **Still need localized cooling (but hotspots may vary from die to die)**
- **3D-friendly cooling**
- **Sensing remains a challenge**
- **Better guidance on reliability management**
 - Is a single, strict max temp too simple?
- **Connect thermal design to architecture and workload behavior**
 - Need a way to make tradeoffs in \$\$\$
 - Person-years, risk, marginal costs, etc.

Summary

- **Temperature-aware design only becoming more important (and more difficult)**
- **Need more collaboration across reliability, thermal engineering, architecture, circuit fields**

Backup Slides

Layout Considerations

- **Multicore layout and “spatial filtering” give you an extra lever (DAC’08, to appear)**
 - The smaller a power dissipator, the more effectively it spreads its heat [IEEE Trans. Computers, to appear]
 - Ex: 2x2 grid vs. 21x21 grid: 188W TDP vs. 220 W (17%) – DAC 2008
 - Increase core density
 - Or raise Vdd, Vth, etc.
 - Thinner dies, better packaging boost this effect
- **Seek architectures that minimize area of high power density, maximize area in between, *and can be easily partitioned***

