McCormick

Northwestern Engineering

The Rise of Dark Silicon

Nikos Hardavellas

Northwestern University, EECS

Energy is Shaping the IT Industry

#1 of Grand Challenges for Humanity in the Next 50 Years

[Smalley Institute for Nanoscale Research and Technology, Rice U.]

- A 1,000m^2 datacenter is 1.5MW!
- Datacenter energy consumption in US **>100 TWh** in 2011 [EPA]
  - 2.5% of domestic power generation, $7.4B
- Global computing consumed ~**408 TWh** in 2010 [Gartner]

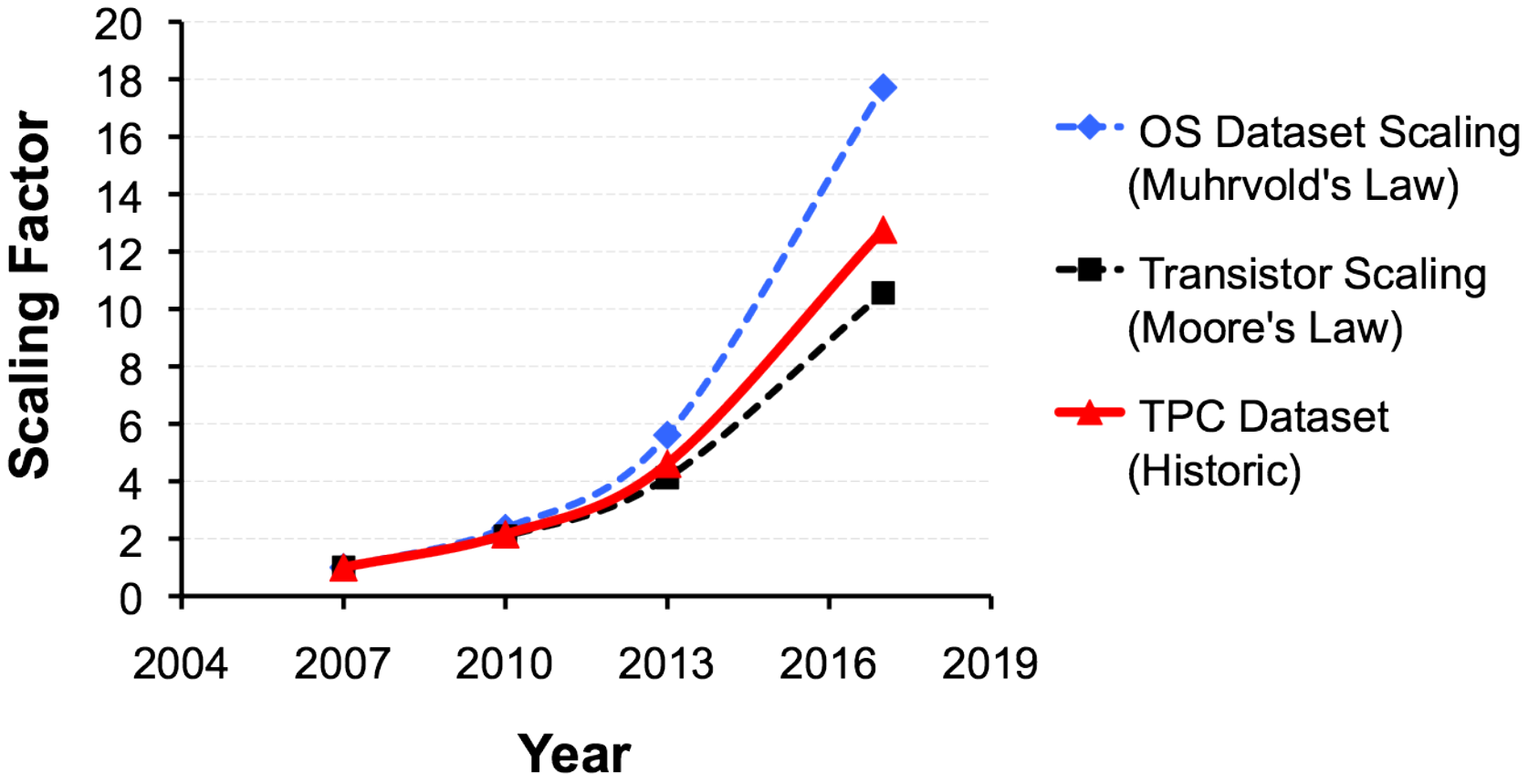
- Carbon footprint of world's data centers ≈ **Czech Republic**
- CO_2-equiv. emissions of US datacenters ≈ **Airline Industry (2%)**

- **10% annual growth** on installed computers worldwide [Gartner]

➡ Exponential increase in energy consumption

© Hardavellas

Application Dataset Scaling



➧ Application datasets scale faster than Moore's Law!

© Hardavellas

Datasets Grow Exponentially

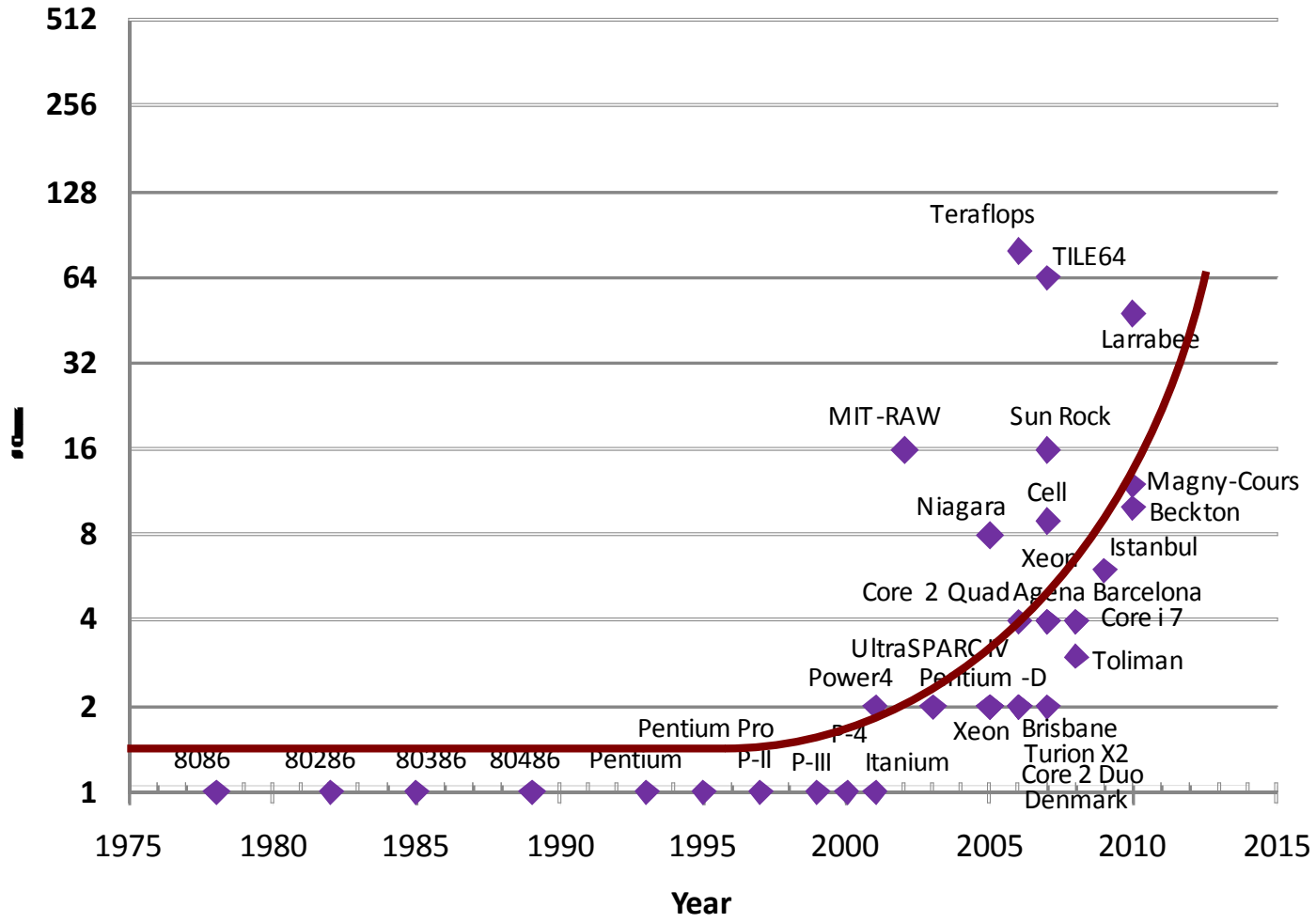
- SPEC and TPC datasets grow faster than Moore's Law
- Large Hadron Collider
  - March 2011: 1.6PB data produced and transferred to Tier-1
- Large Synoptic Survey Telescope
  - Produces 30 TB/night
  - Roughly equivalent to 2 Sloan Digital Sky Surveys daily
    - Sloan produced more data than entire history of astronomy before it

- Massive data require massive computations to process them

➡ Exponential increase in energy consumption

© Hardavellas

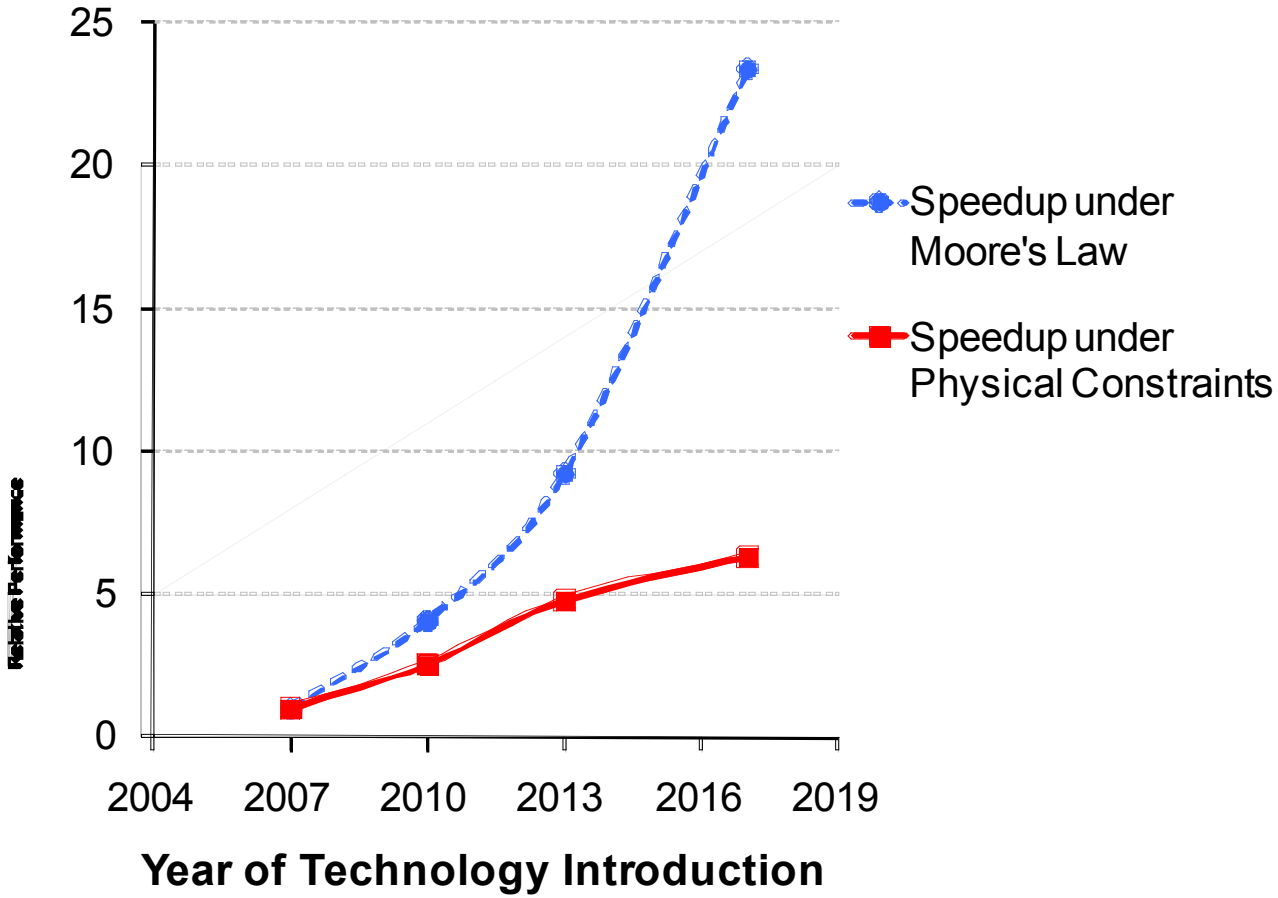
Exponential Growth of Core Counts



Does performance follow same curve?

© Hardavellas

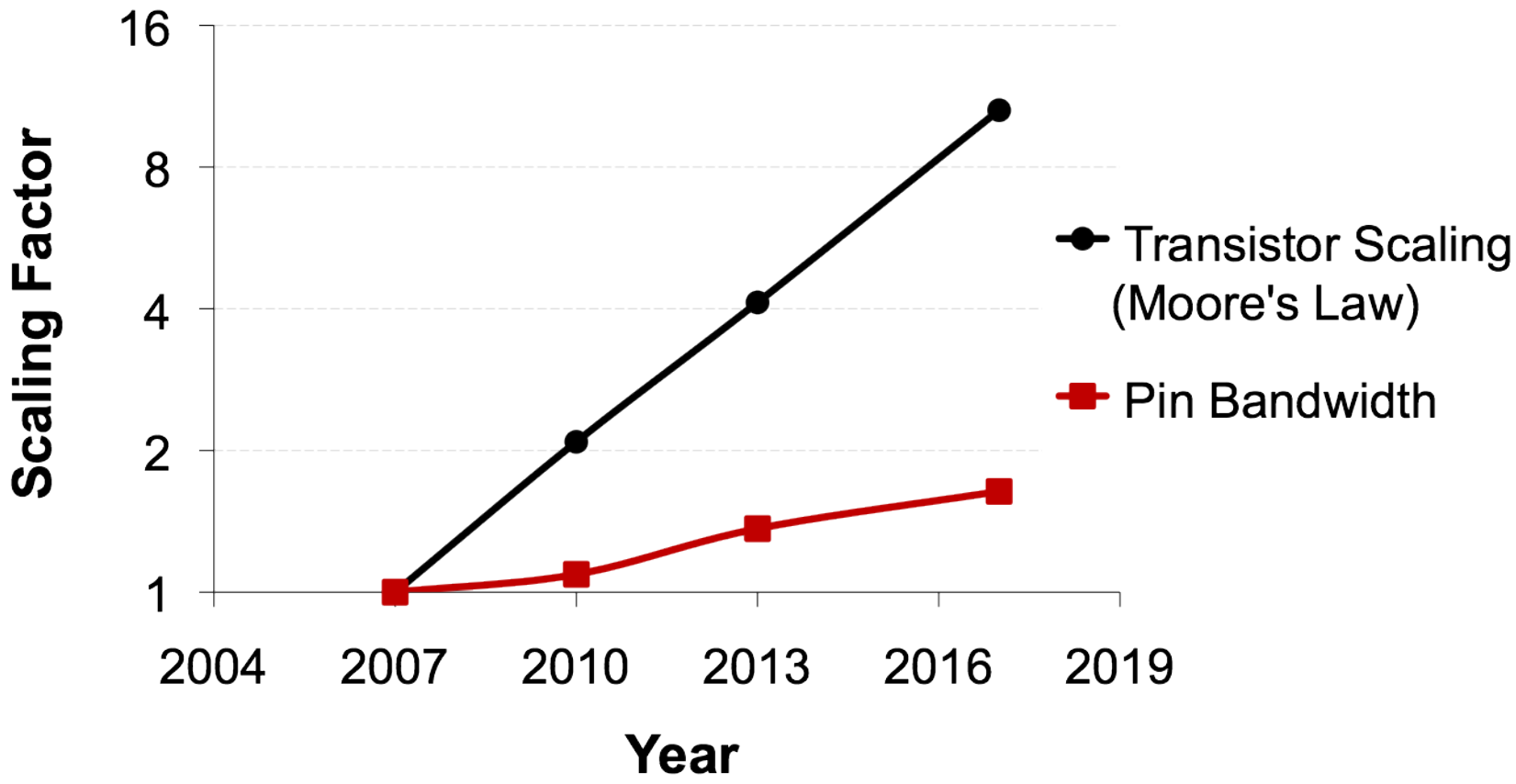
Performance Expectations vs. Reality



Physical constraints limit speedup

© Hardavellas

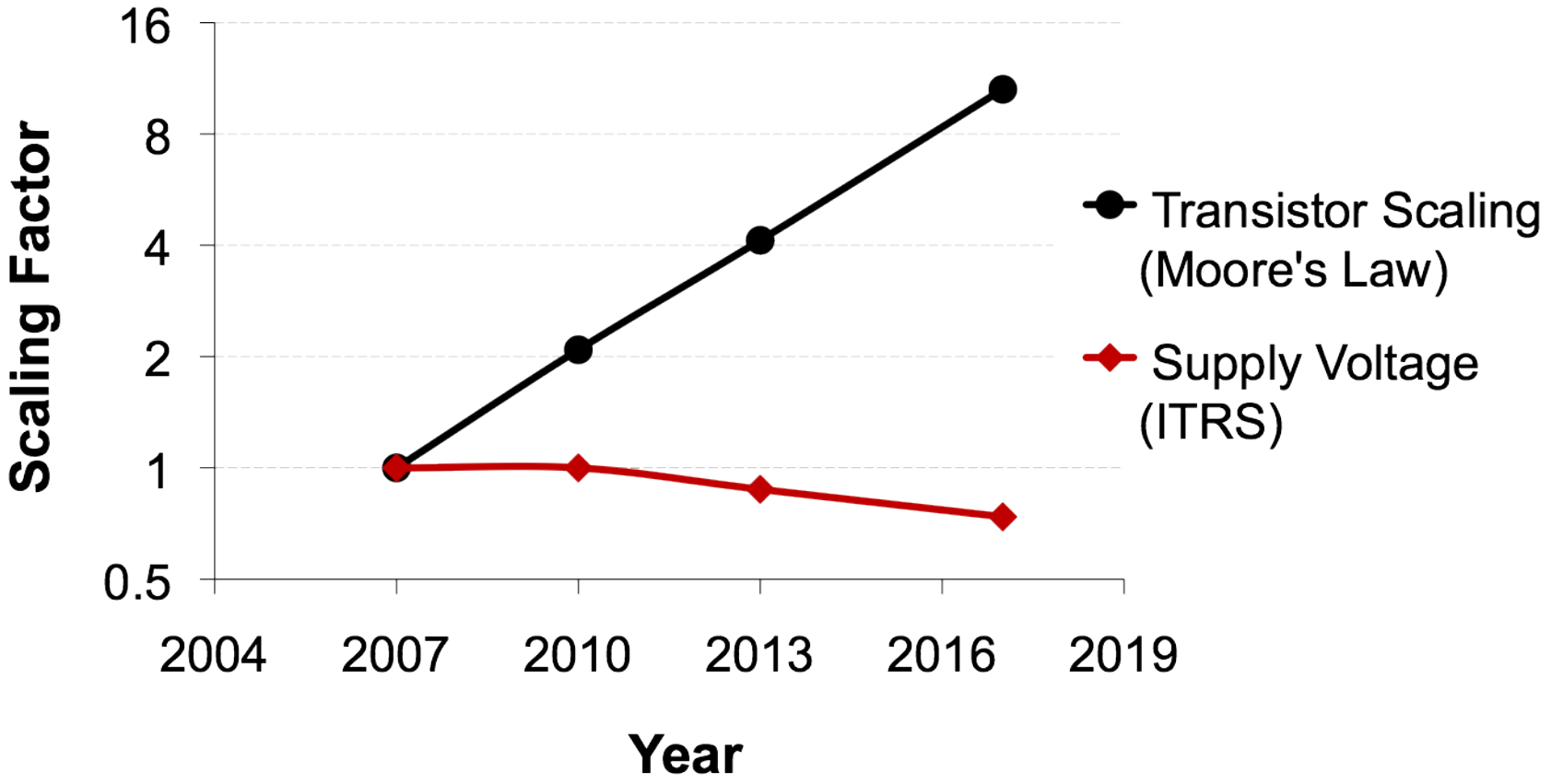
Pin Bandwidth Scaling



Cannot feed cores with data fast enough

© Hardavellas

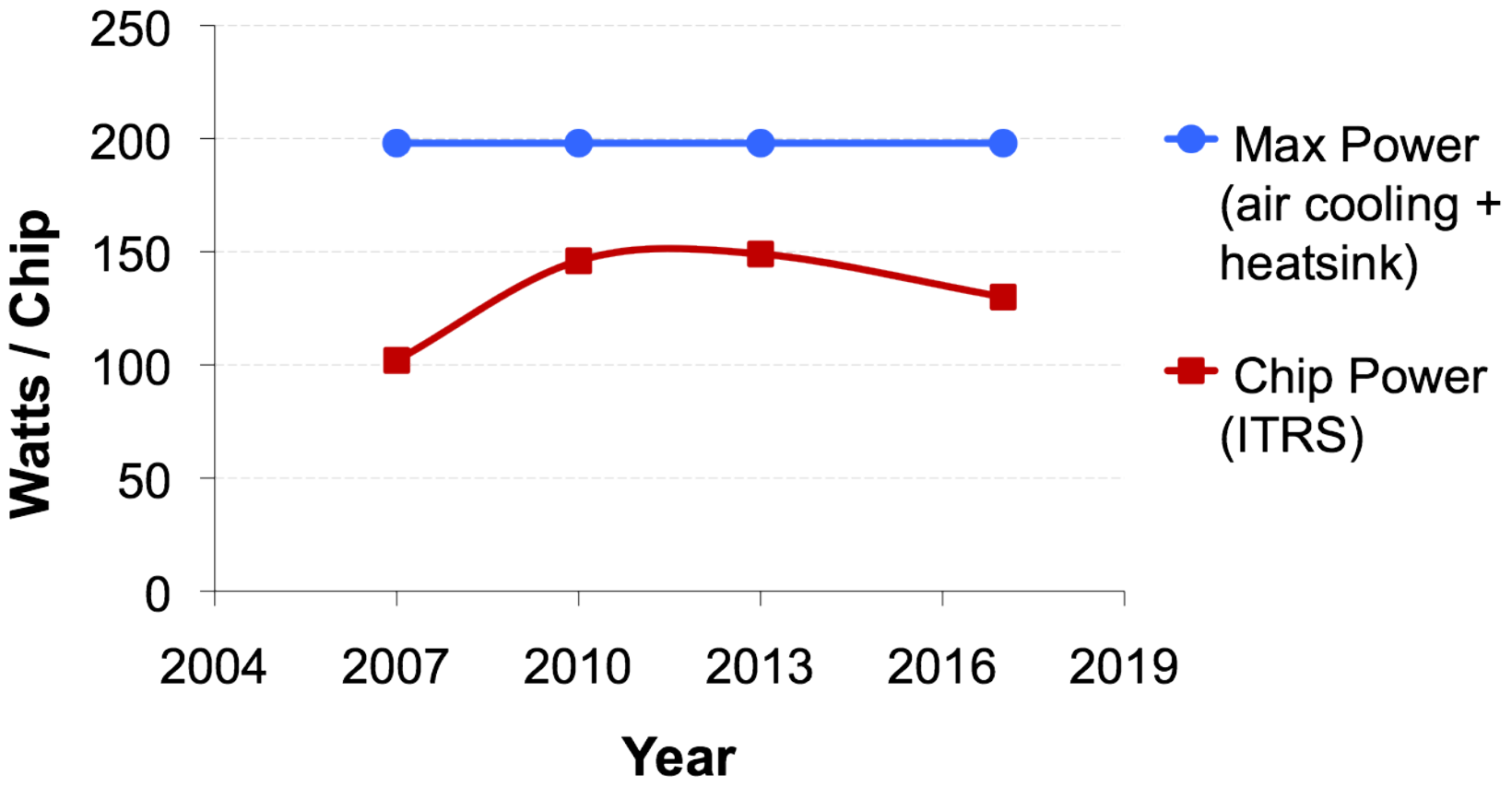
Supply Voltage Scaling



➡ Cannot power up all transistors simultaneously → Dark Silicon

© Hardavellas

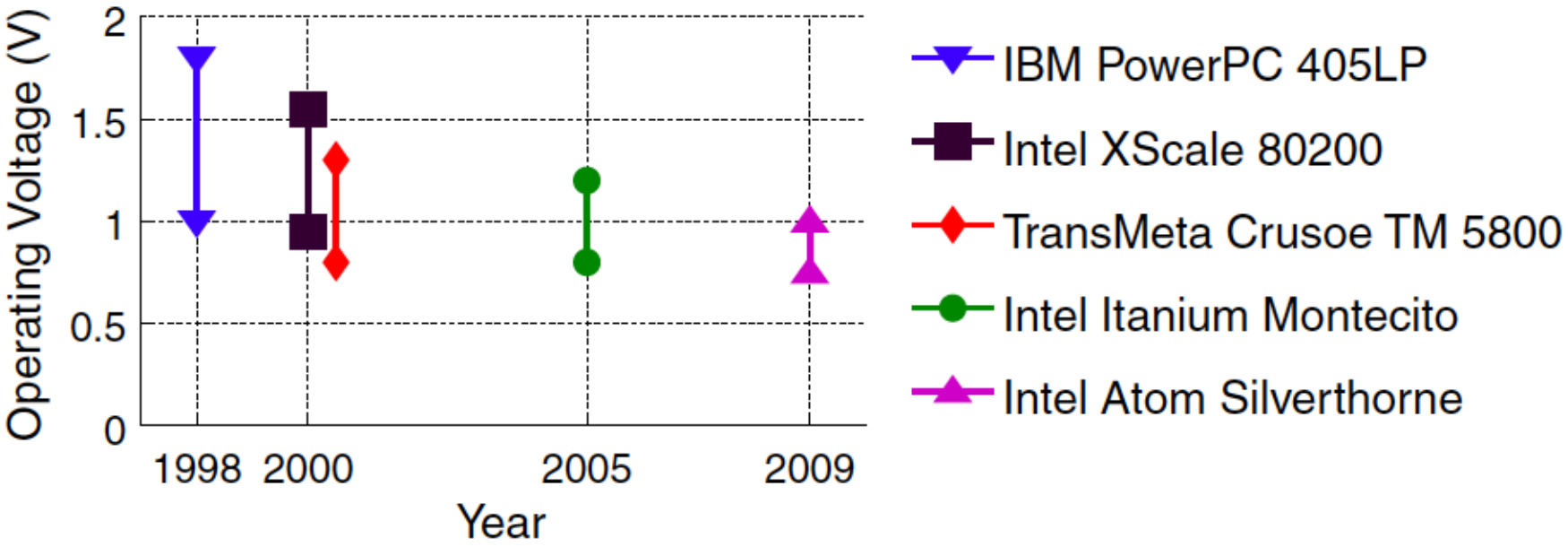
Chip Power Scaling



➧ Cooling does not scale!

© Hardavellas

Range of Operational Voltage



[Watanabe et al., ISCA'10]

➡ Shrinking range of operational voltage hampers voltage-freq. scaling

© Hardavellas

Where Does Server Energy Go?

Many sources of power consumption:

- Server only [Fan, ISCA'07]
  - **Processor chips (37%)**
  - Memory (17%)
  - Peripherals (29%)
  - …
- Infrastructure (another 50%)
  - Cooling
  - Power distribution

© Hardavellas

A Study of Server Chip Scalability

- Actual server workloads today
  - Easily parallelizable (performance-scalable)

- Actual physical char. of processors/memory
  - ITRS projections for technology nodes
  - Modeled power/performance across nodes

- For server chips
  - Bandwidth is near-term limiter

  → **Energy is the ultimate limiter**

© Hardavellas

First-Order Analytical Modeling

[Hardavellas, IEEE Micro 2011]

Physical characteristics modeled after UltraSPARC T2, ARM11

- **Area:** Cores + caches = 72% die, scaled across technologies
- **Power:** ITRS projections of V_{dd}, V_{th}, C_{gate}, I_{sub}, W_{gate}, S_0
  - Active: cores=f(GHz), cache=f(access rate), NoC=f(hops)
  - Leakage: f(area), f(devices), 66^{o}C
  - Devices/ITRS: Bulk Planar CMOS, UTB-FD SOI, FinFETs, HP/LOP
- **Bandwidth:**
  - ITRS projections on I/O pins, off-chip clock, f(miss, GHz)
- **Performance:** CPI model based on miss rate
  - Parameters from real server workloads (DB2, Oracle, Apache)
  - Cache miss rate model (validated), Amdahl & Myhrvold Laws

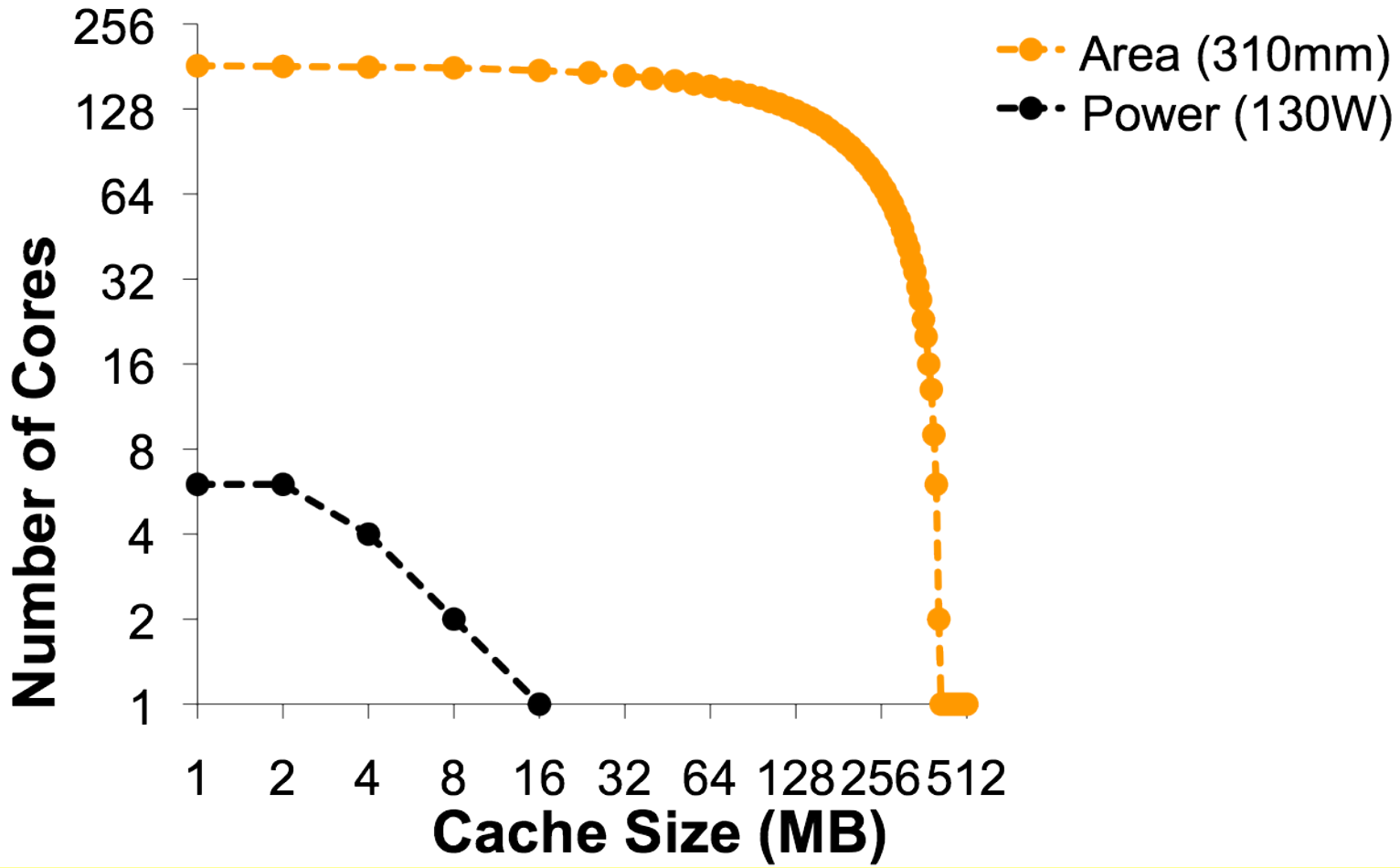
© Hardavellas

Caveats

- First-order model
  - The intent is to uncover trends relating the effects of technology-driven physical constraints to the performance of commercial workloads running on multicores
  - The intent is NOT to offer absolute numbers
- Performance model works well for workloads with low MLP
  - Database (OLTP, DSS) and web workloads are mostly memory-latency-bound
- Workloads are assumed parallel
  - Scaling server workloads is reasonable

© Hardavellas

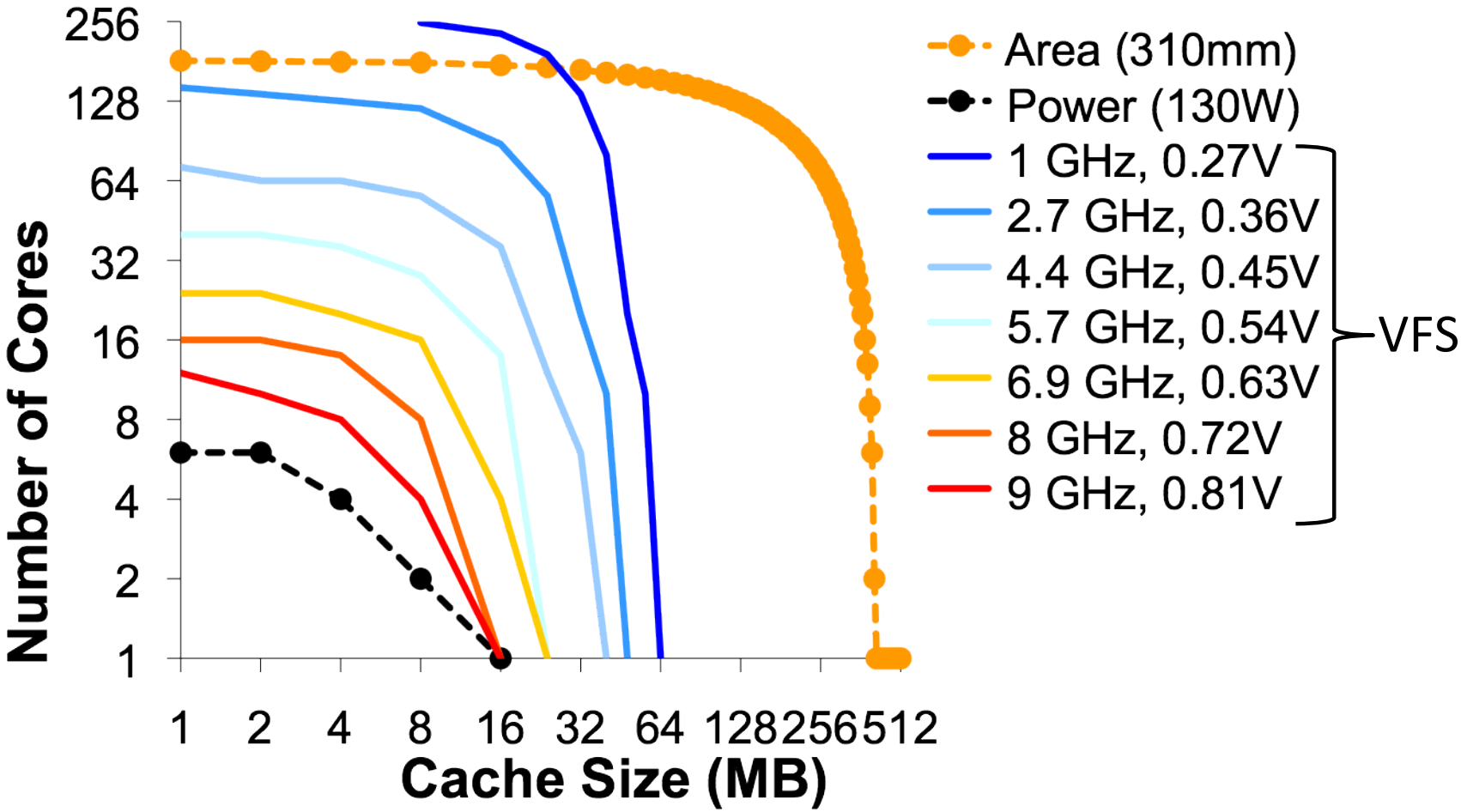
Area vs. Power Envelope



Good news: can fit 100's cores. Bad news: cannot power them all

© Hardavellas

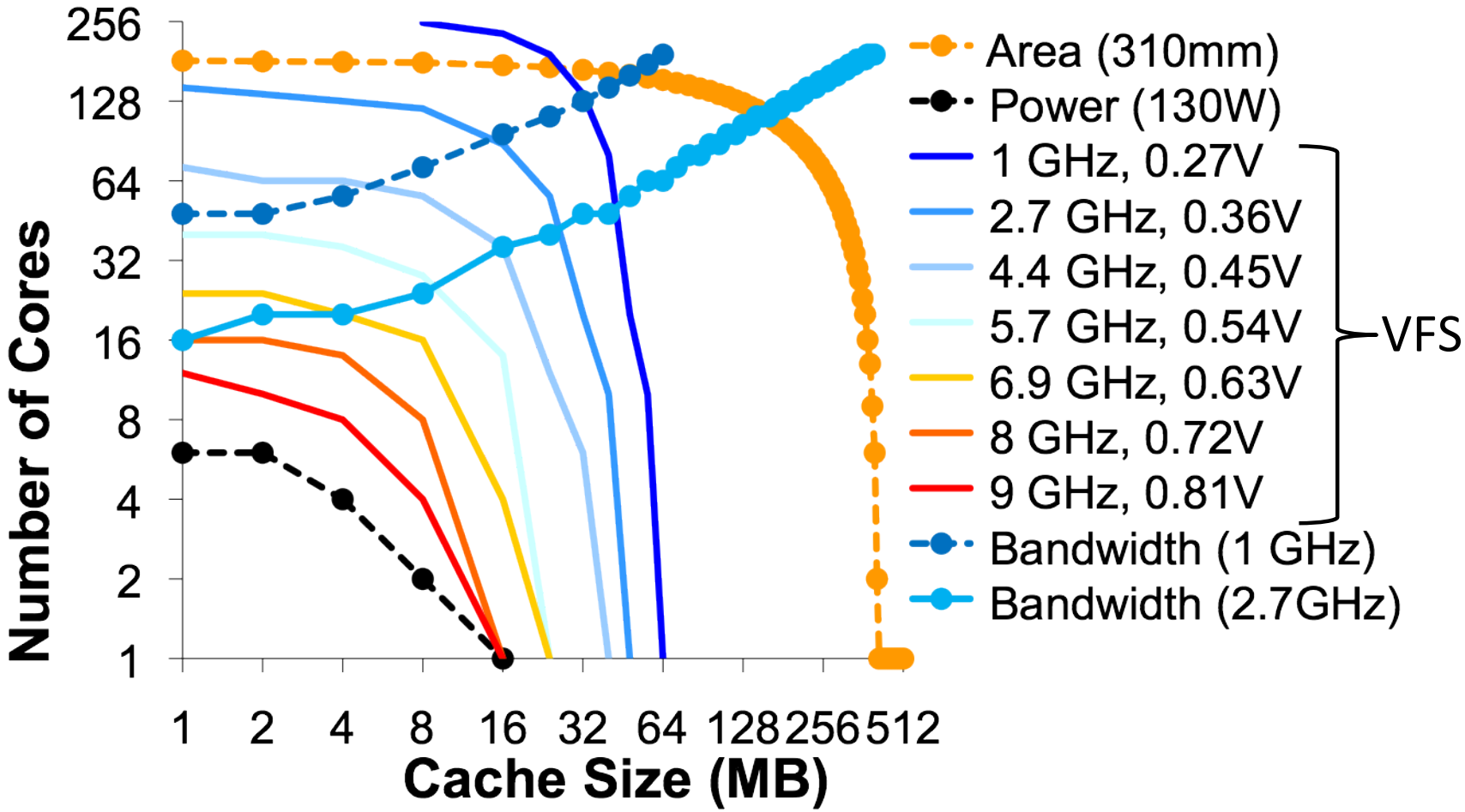
Pack More Slower Cores, Cheaper Cache



- The reality of The Power Wall: a power-performance trade-off

© Hardavellas

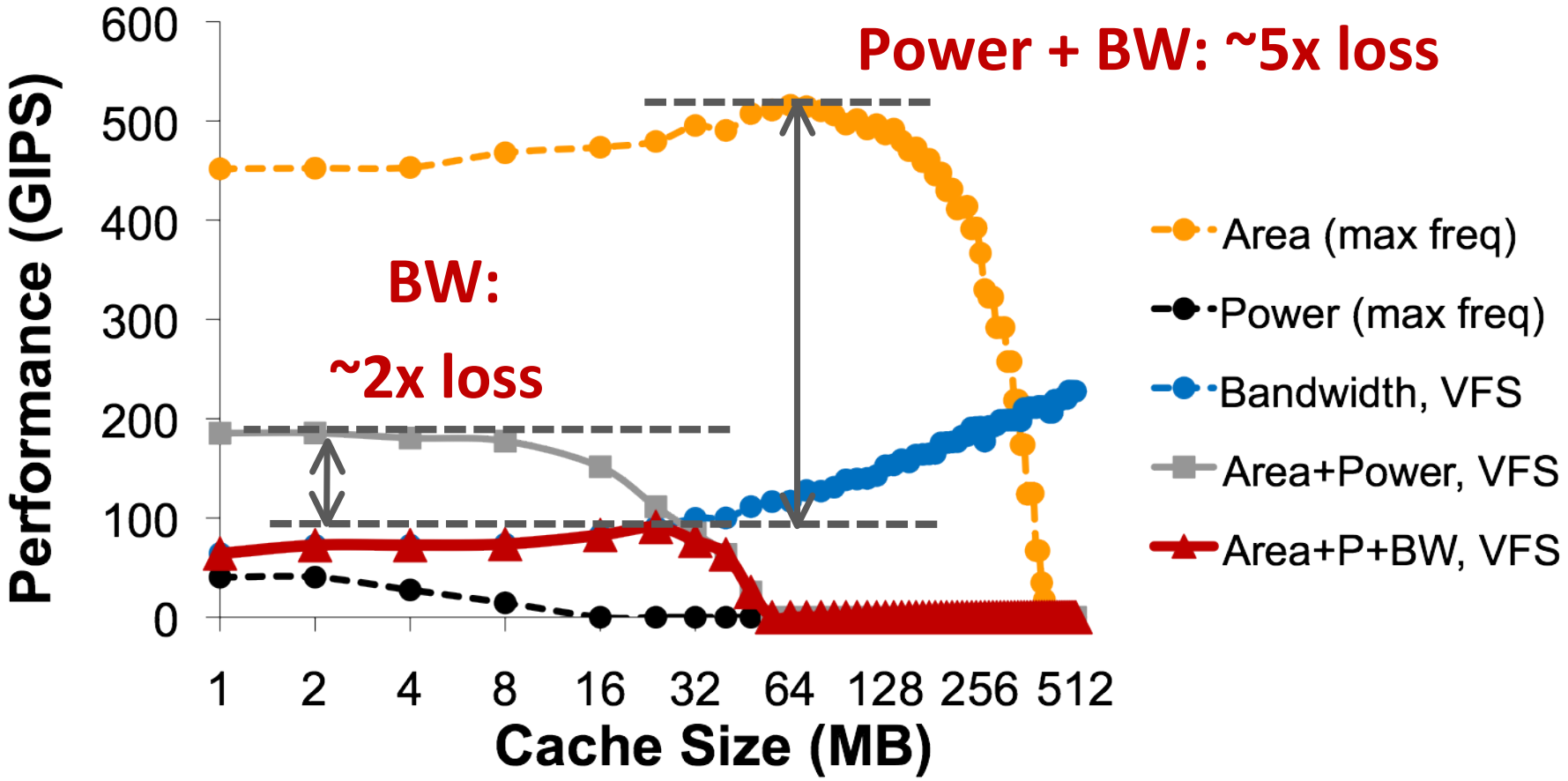
Pin Bandwidth Constraint



- Bandwidth constraint favors fewer + slower cores, more cache

© Hardavellas

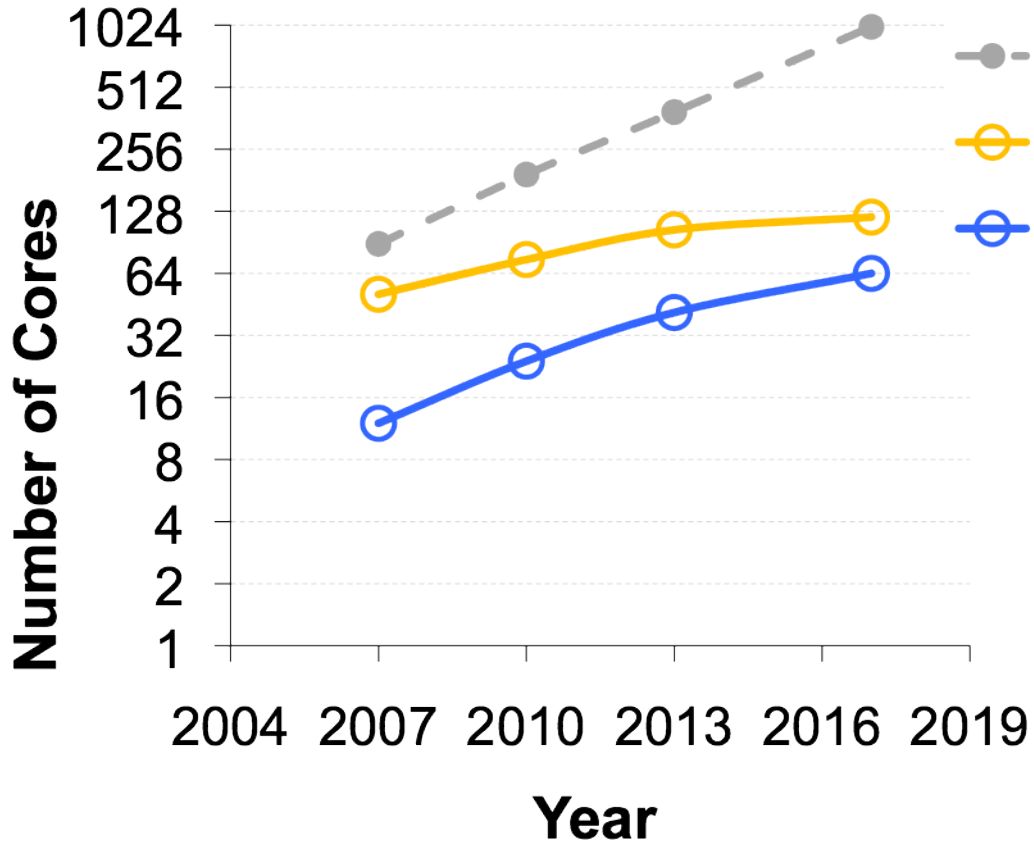
Example of Optimization Results



- Jointly optimize parameters, subject to constraints, SW trends
- Design is first bandwidth-constrained, then power-constrained

© Hardavellas

Core Counts for Peak-Performance Designs



Physical characteristics modeled after

- UltraSPARC T2 (GPP)
- ARM11 (EMB)

- Designs > 120 cores impractical for general-purpose server apps
- B/W and power envelopes + dataset scaling limit core counts

© Hardavellas

Short-Term Scaling Implications

Caches are getting huge

- Need cache architectures to deal with >> MB

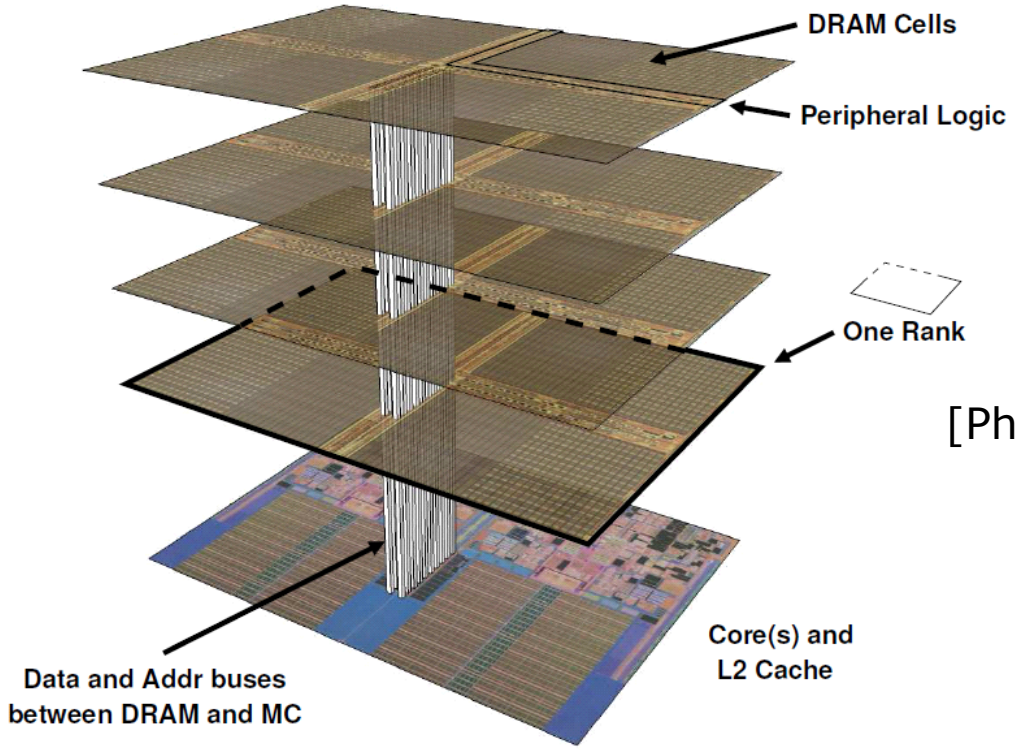
→Elastic Caches

- Adapt behavior to executing workload to minimize transfers
- Reactive NUCA [Hardavellas, ISCA 2009][Hardavellas, IEEE Micro 2010]
- Dynamic Directories [Das, NUTR 2010, in submission]
  - …but that's another talk…

➧ Need to push back the bandwidth wall!!!

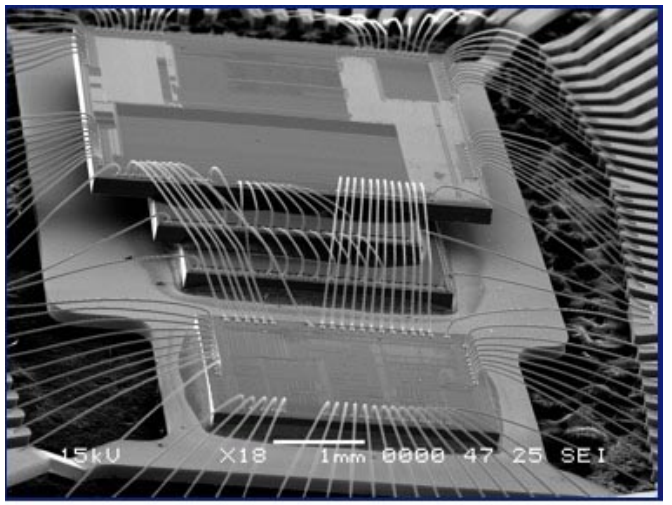
© Hardavellas

Mitigating Bandwidth Limitations: 3D-stacking

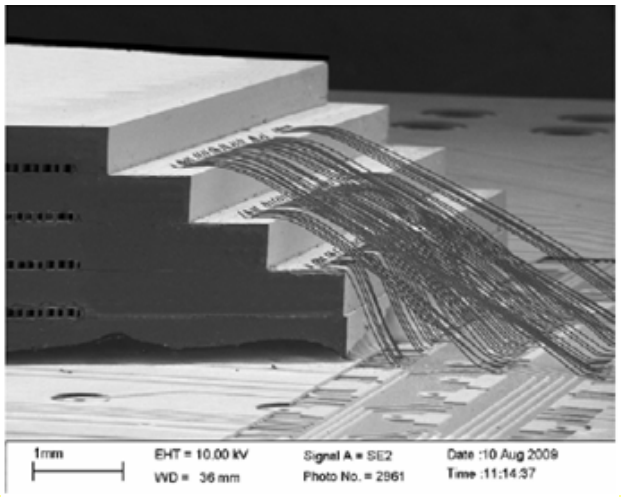


[Loh et al., ISCA'08]

[Philips]



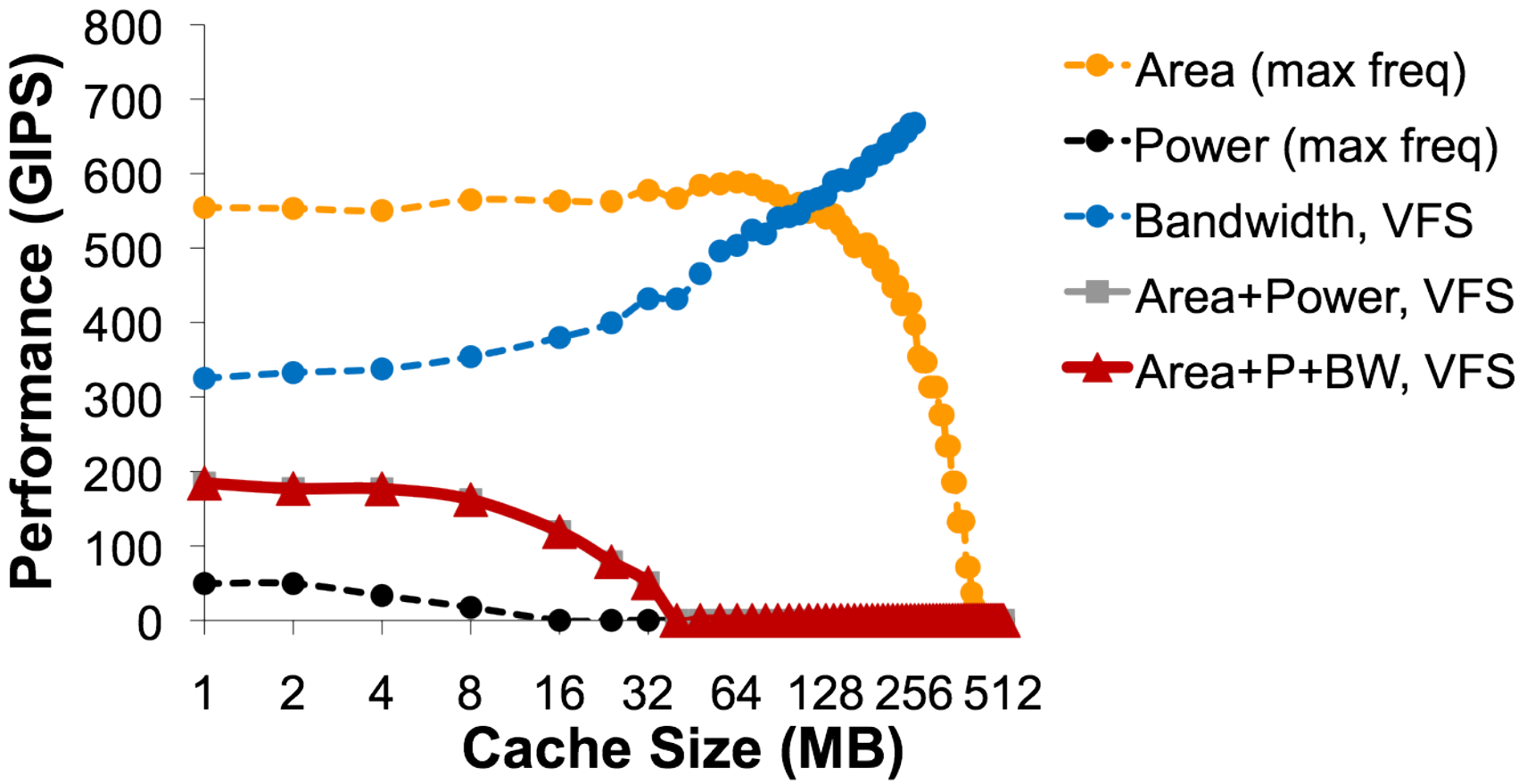
[IBM]



➡ Delivers TB/sec of bandwidth; use as large "in-package" cache

© Hardavellas

Performance Analysis of 3D-Stacked Multicores



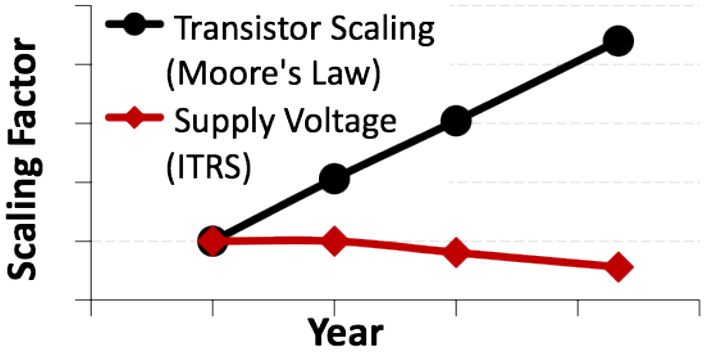
➡ Chip becomes power-constrained

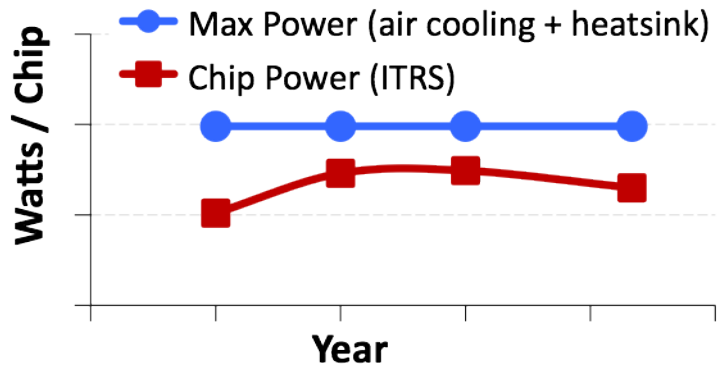
© Hardavellas

The Rise of Dark Silicon

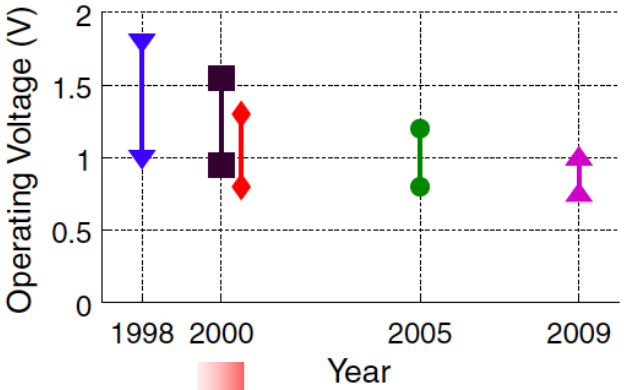
Transistor counts increase exponentially, but…

Can no longer power the entire chip (voltages, cooling do not scale)





Traditional HW power-aware techniques inadequate (e.g., voltage-freq. scaling)

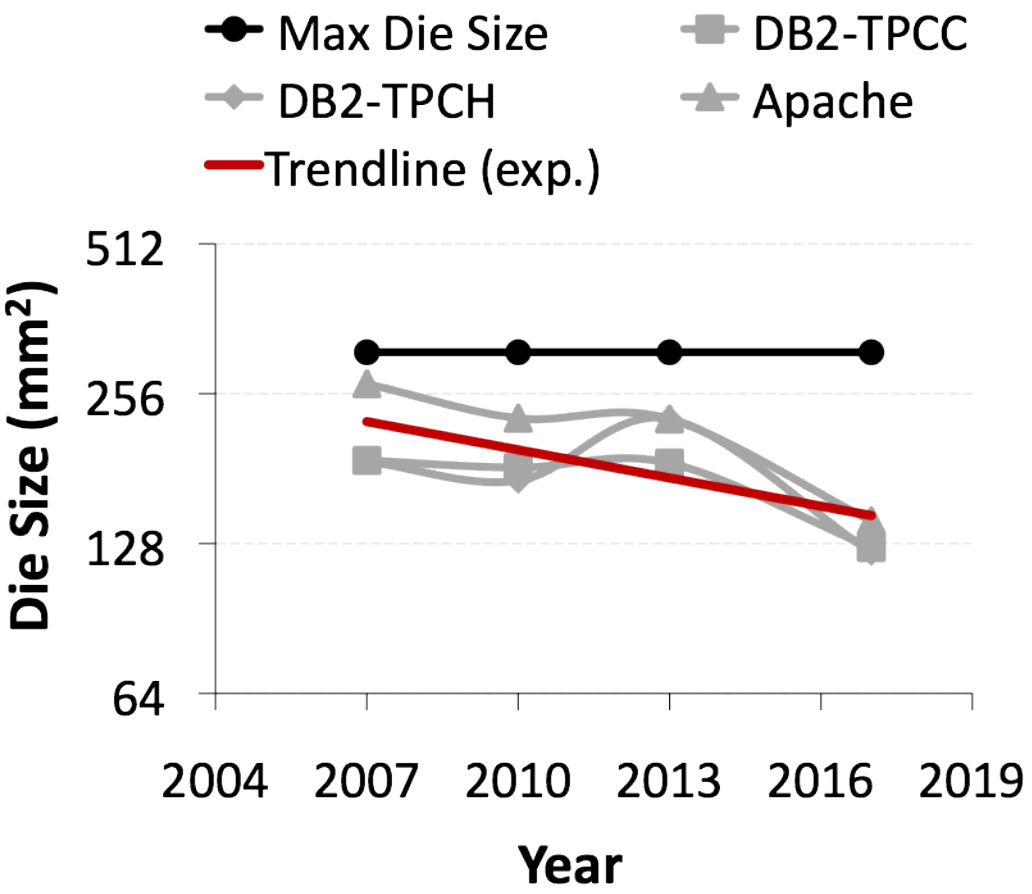


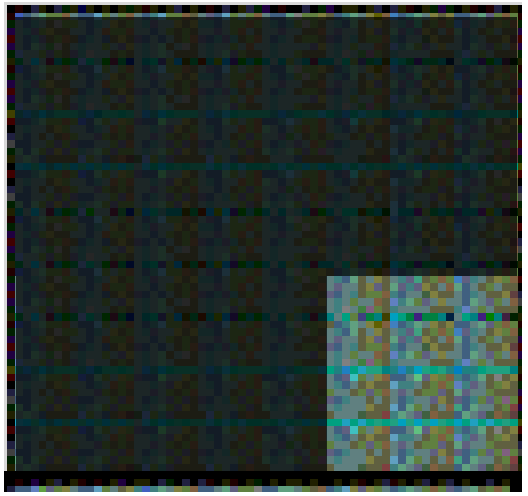
[Watanabe et al., ISCA'10]

Dark Silicon !!!

© Hardavellas

Exponentially-Large Area Left Unutilized





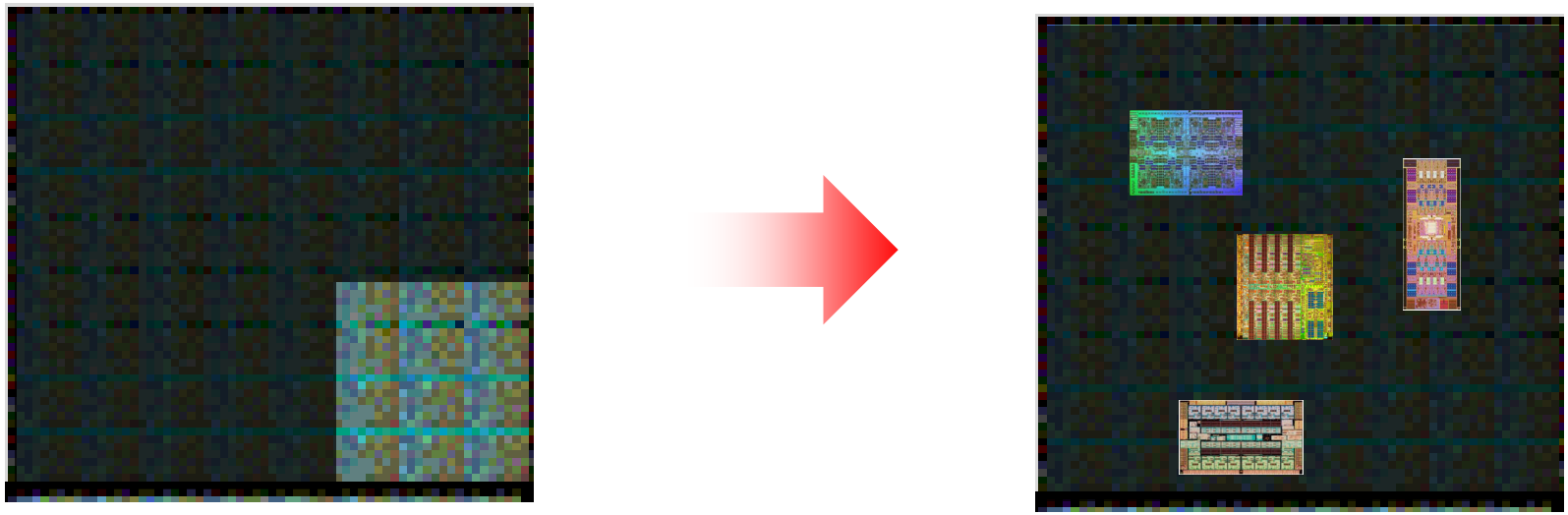
Should we waste it?

© Hardavellas

Repurpose Dark Silicon for Specialized Cores

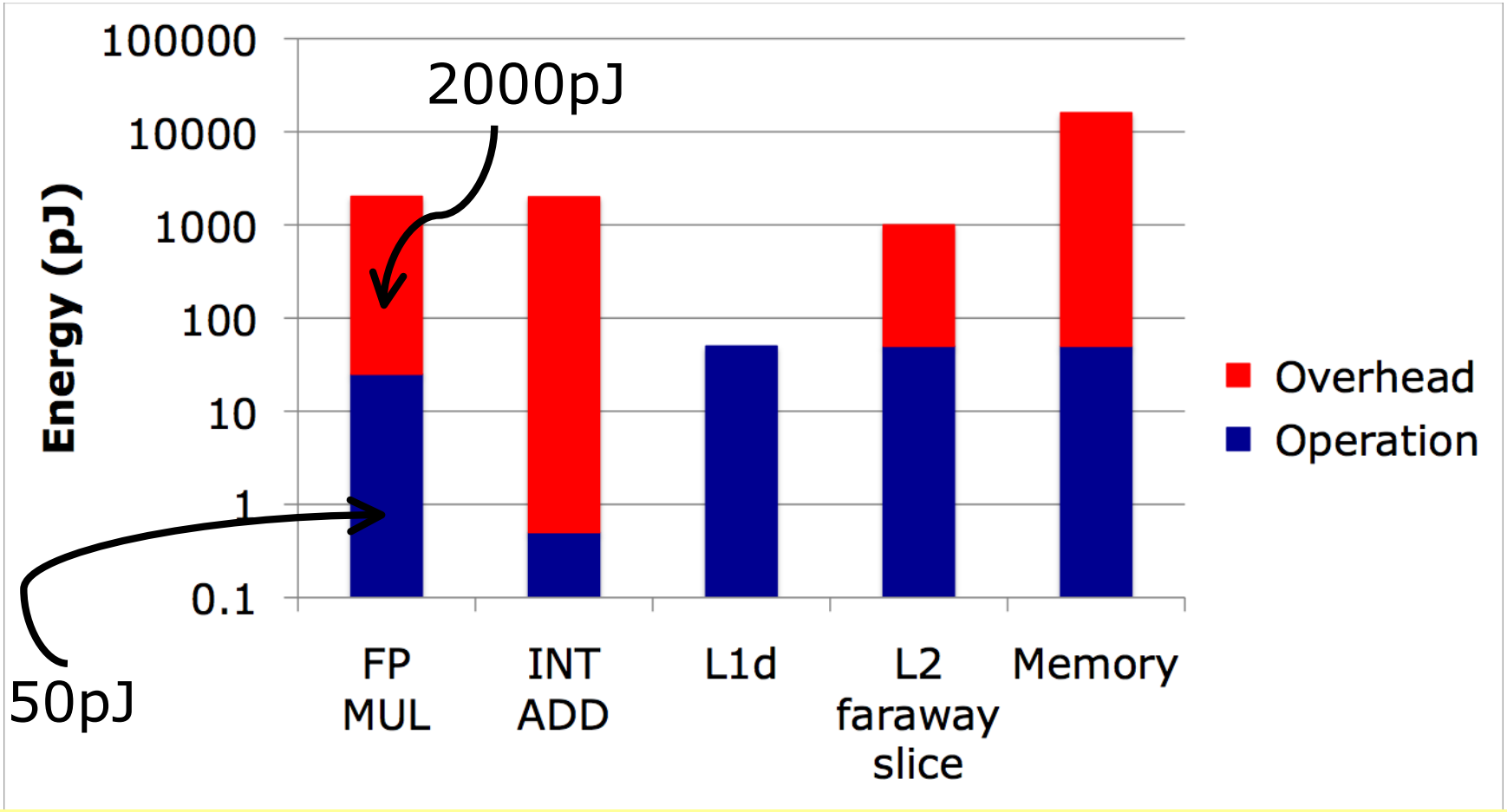
[Hardavellas, IEEE Micro 2011]

- Don't waste it; harness it instead!
  - Use dark silicon to implement specialized cores
- Applications cherry-pick few cores, rest of chip is powered off
- Vast unused area → many cores → likely to find good matches



© Hardavellas

Overheads of General-Purpose Processors



- Core specialization will minimize most overheads
- ASICs ~100-700x more efficient than general-purpose cores

© Hardavellas

First-Order Core Specialization Model

- Modeling of physically-constrained CMPs across technologies
- Model of specialized cores based on ASIC implementation of H.264:
  - Implementations on custom HW (ASICs), FPGAs, multicores (CMP)
  - Wide range of computational motifs, extensively studied

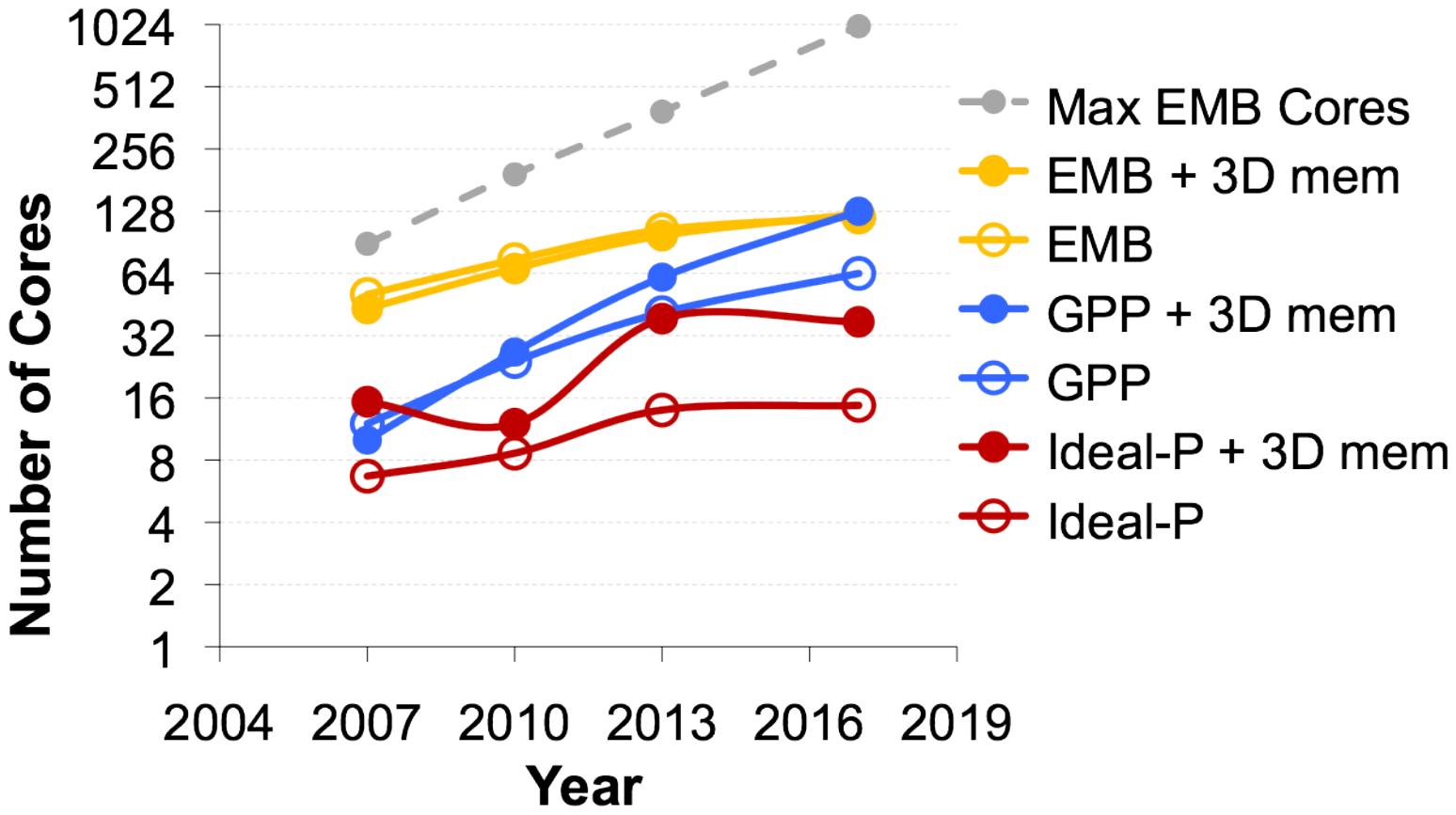
		Frames per sec	Energy per frame (mJ)	Performance gap of CMP vs. ASIC	Energy gap of CMP vs. ASIC
ASIC		30	4		
CMP	IME	0.06	1179	525x	707x
	FME	0.08	921	342x	468x
	Intra	0.48	137	63x	157x
	CABAC	1.82	39	17x	261x

[Hameed et al., ISCA 2010]

➧ ***12x LOWER ENERGY*** compared to best conventional alternative

© Hardavellas

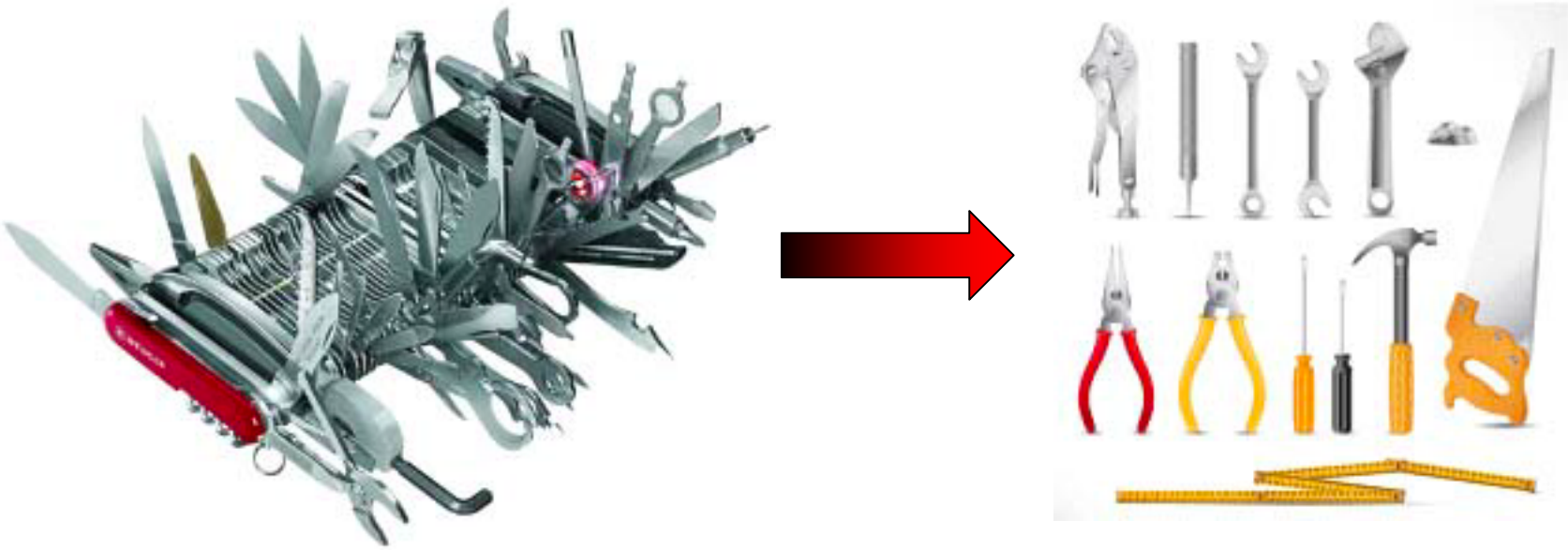
Specialized Multicores: Power Only Few Cores



- Only few cores need to run at a time for max speedup
- Vast unused die area will allow many cores

© Hardavellas

The New Core Design

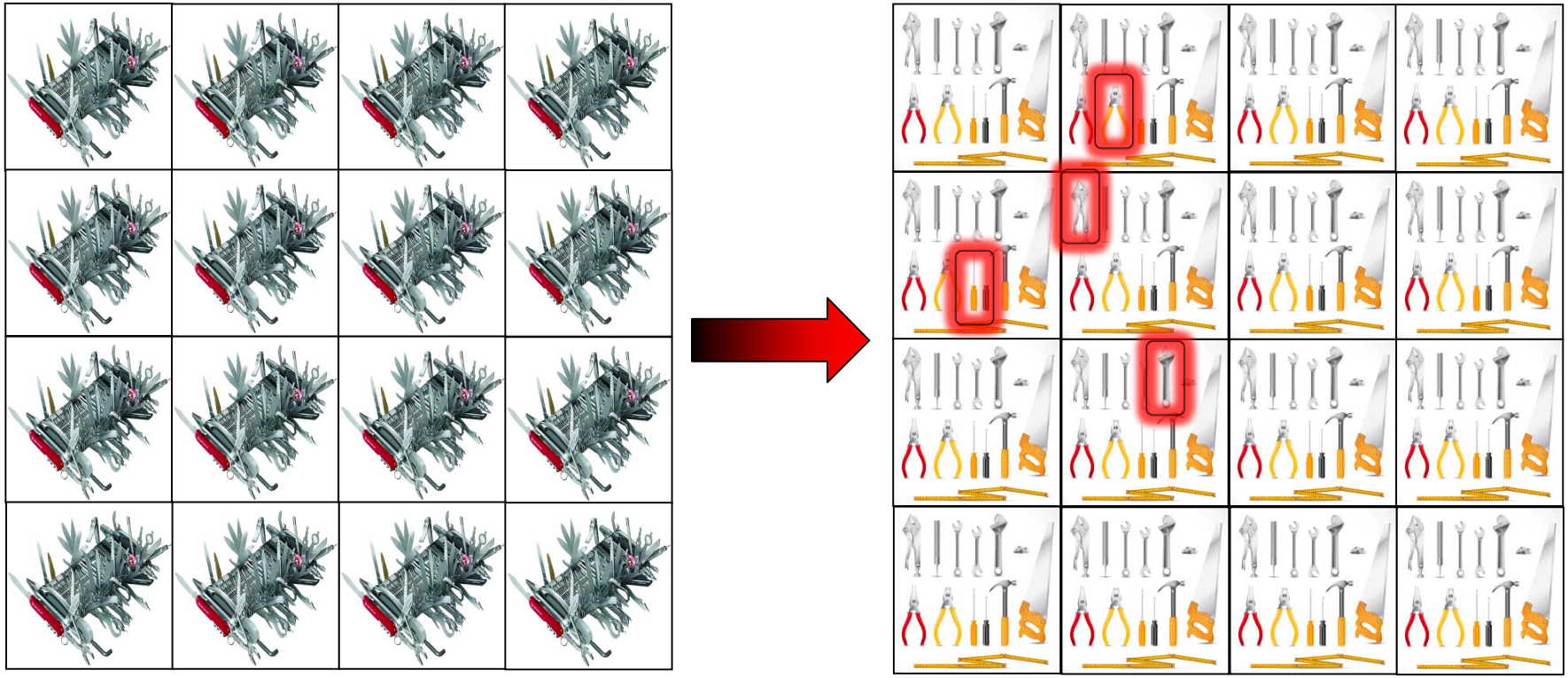


[analogy by A. Chien]

From fat conventional cores, to a sea of specialized cores

© Hardavellas

The New Multicore



Power up only what you need

© Hardavellas

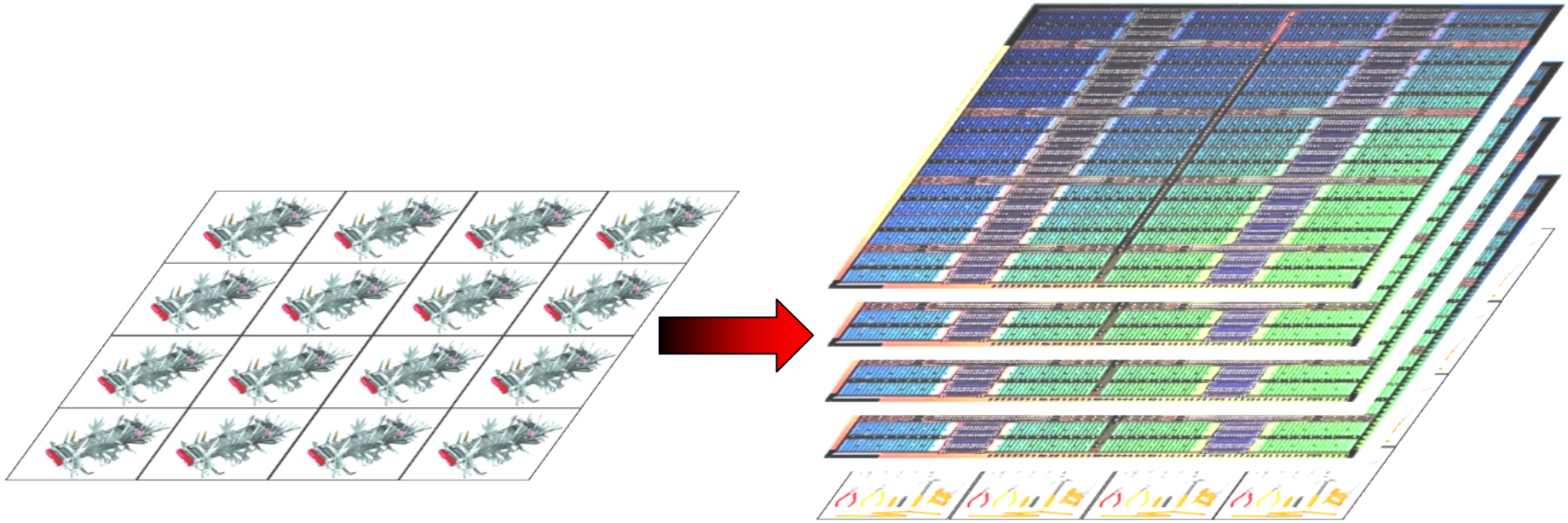
Design for Dark Silicon: Many Open Questions

To get 12x lower energy (12x performance for same power budget):

- Which candidates are best for off-loading to specialized cores?
- What should these cores look like?
  - Exploit commonalities to avoid core over-specialization
  - Can we classify all computations into 10 bins?
- What are the appropriate language/compiler/runtime techniques to drive execution migration?
  - Impact on scheduler?
- How to restructure software/algorithms for heterogeneity?

© Hardavellas

The New Multicore Node



Can push further with more exotic technologies
(e.g., nanophotonics) …but that's another talk

➧ Specialized cores + 3D-die memory stacking

© Hardavellas

Thank You!

Questions?

References:

•N. Hardavellas, M. Ferdman, B. Falsafi, and A. Ailamaki. Toward Dark Silicon in Servers. IEEE Micro, Vol. 31, No. 4, July/August 2011.

•N. Hardavellas. Chip multiprocessors for server workloads. PhD thesis, Carnegie Mellon University, Dept. of Computer Science, August 2009.

•N. Hardavellas, M. Ferdman, A. Ailamaki, and B. Falsafi. Power scaling: the ultimate obstacle to 1K-core chips. Technical Report NWU-EECS-10-05, Northwestern University, Evanston, IL, March 2010.

•R. Hameed, W. Qadeer, M. Wachs, O. Azizi, A. Solomatnikov, B. C. Lee, S. Richardson, C. Kozyrakis, and M. Horowitz. Understanding sources of inefficiency in general-purpose chips. In Proc. of ISCA, June 2010.

If you want to know more about my “other talks” come find me!