McCormick

Northwestern Engineering

When Core Multiplicity Doesn't Add Up

Keynote ISPDC 2010

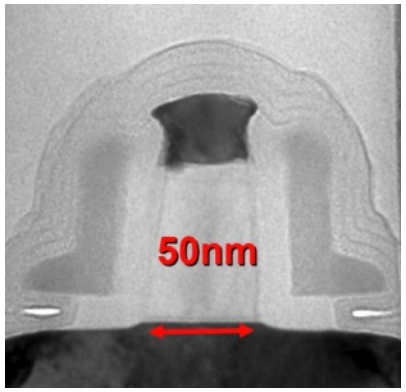
Nikos Hardavellas

PARAG@N – Parallel Architecture Group

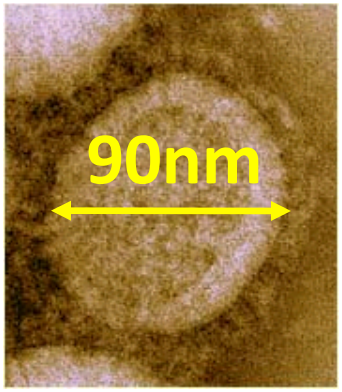
Northwestern University

Collaborators: M. Ferdman, B. Falsafi, A. Ailamaki

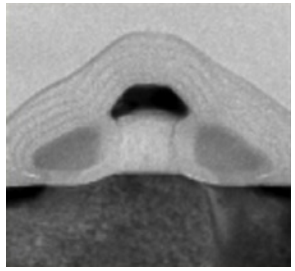
Moore's Law Is Alive And Well



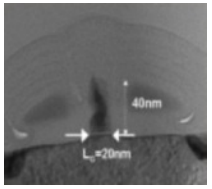
90nm transistor
(Intel, 2005)



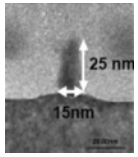
Swine Flu A/H1N1
(CDC)



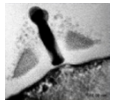
65nm
2007



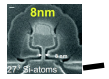
45nm
2010



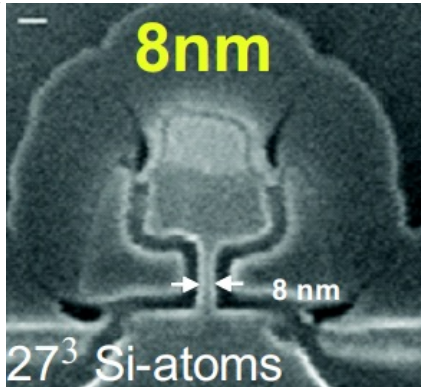
32nm
2013



22nm
2016

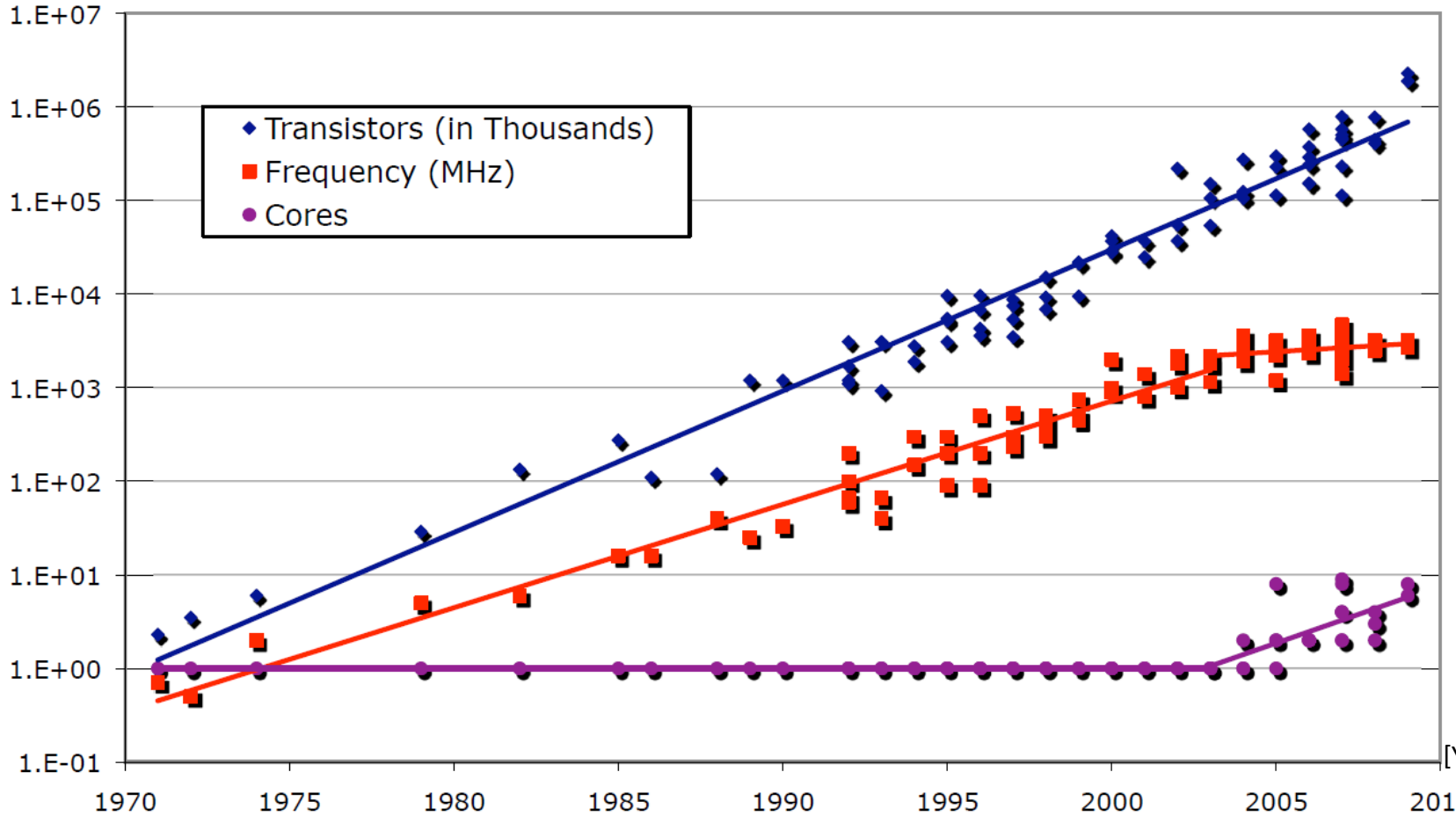


16nm
2019



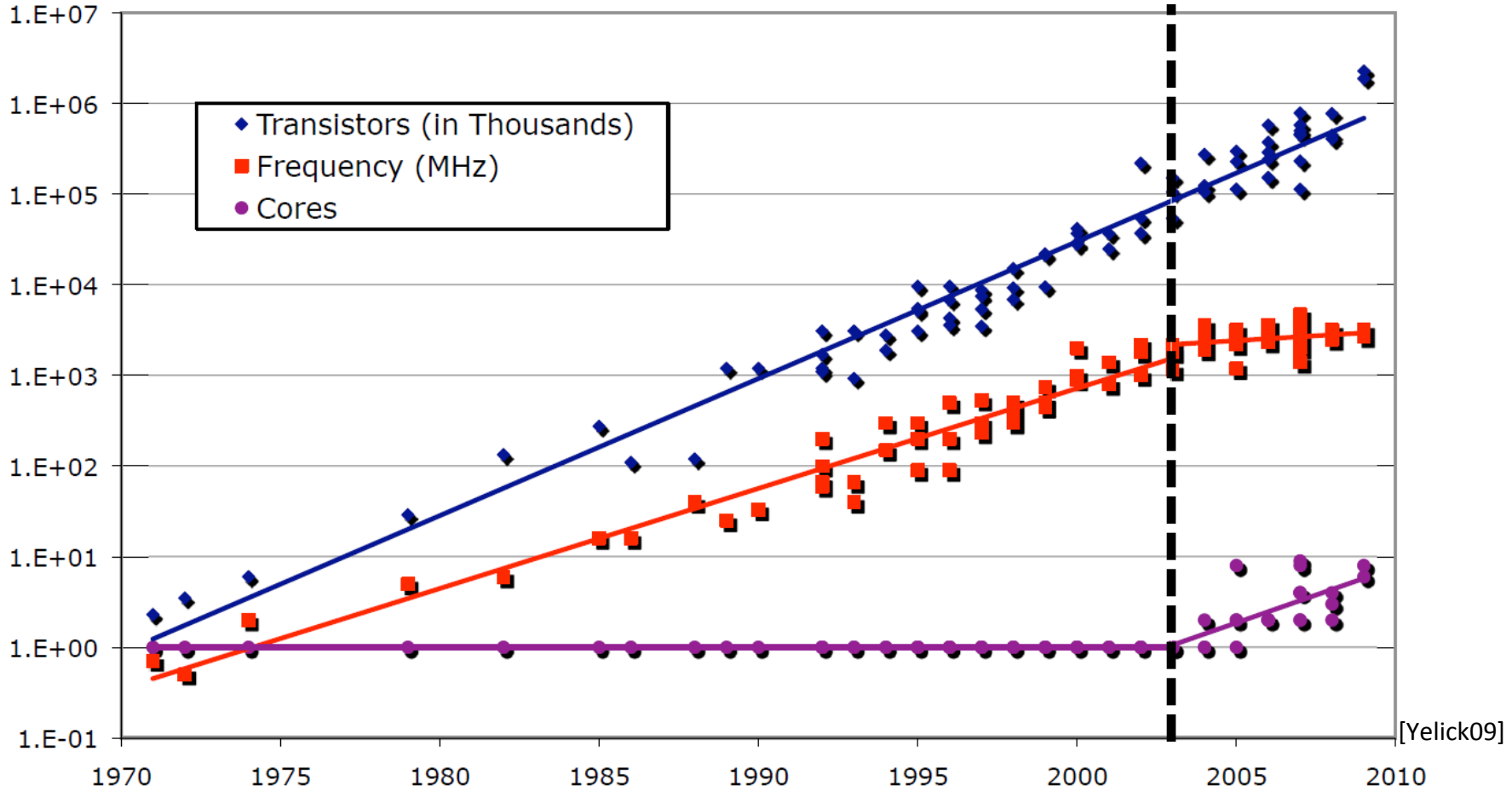
- Device scaling continues for at least another 10 years

Moore's Law Is Alive And Well



[Yelick09]

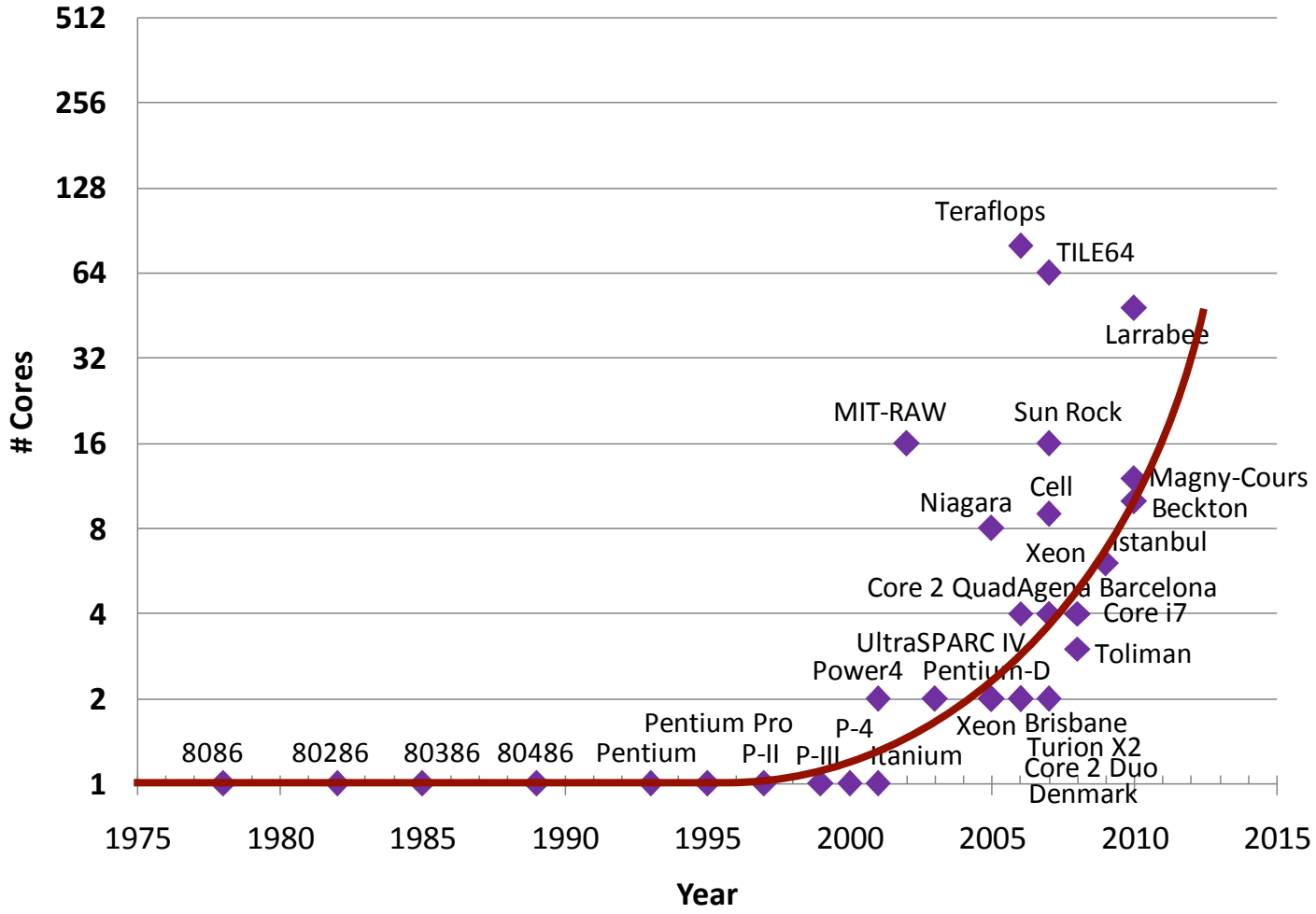
Good Days Ended Nov. 2002



[Yelick09]

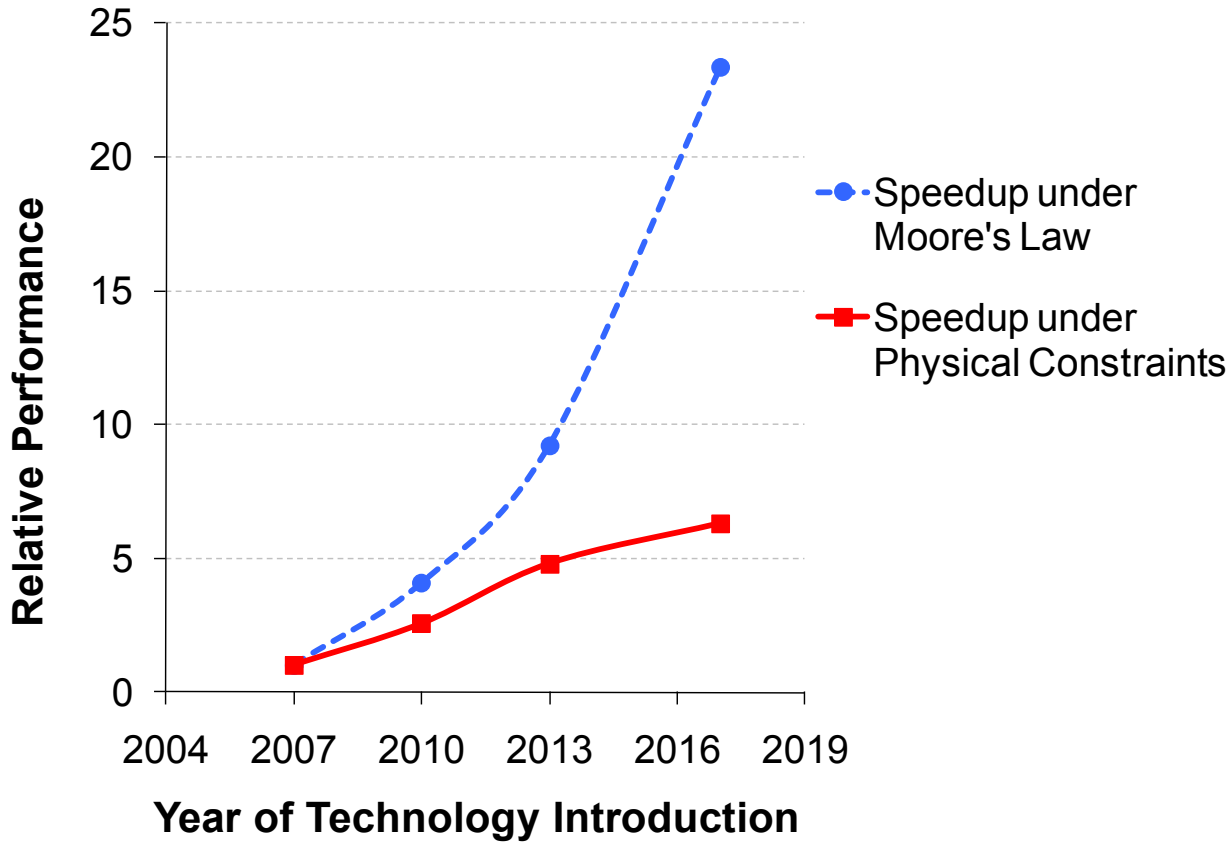
- "New" Moore's Law: 2x cores with every generation

"New" Moore's Law



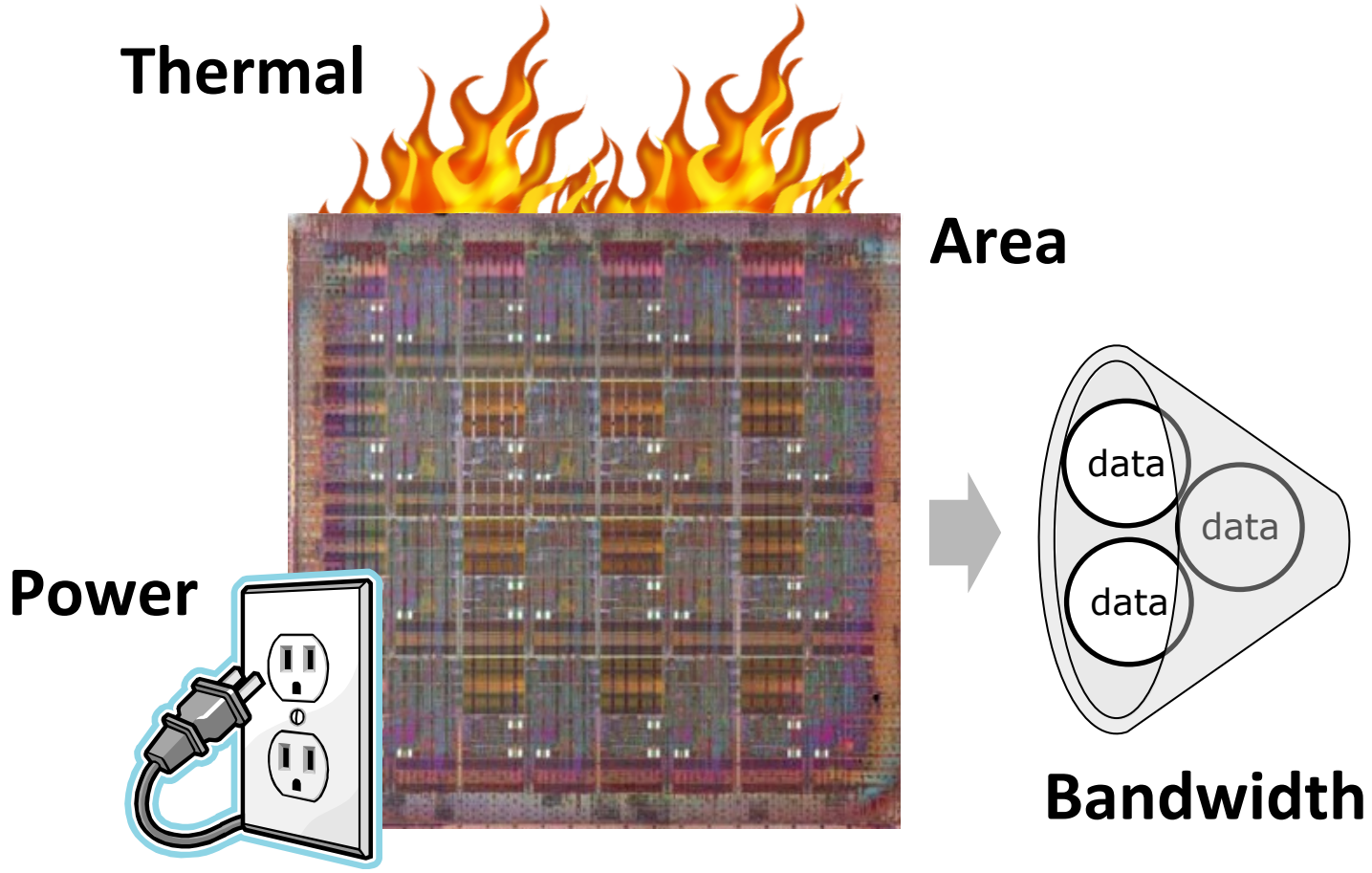
So, are 1000-core chips a viable architecture?

Performance Expectations vs. Reality



➡ Physical constraints limit speedup

Physical Constraints Hamper Performance



- What are the “best” designs given physical constraints?

First-Order Analytical Modeling

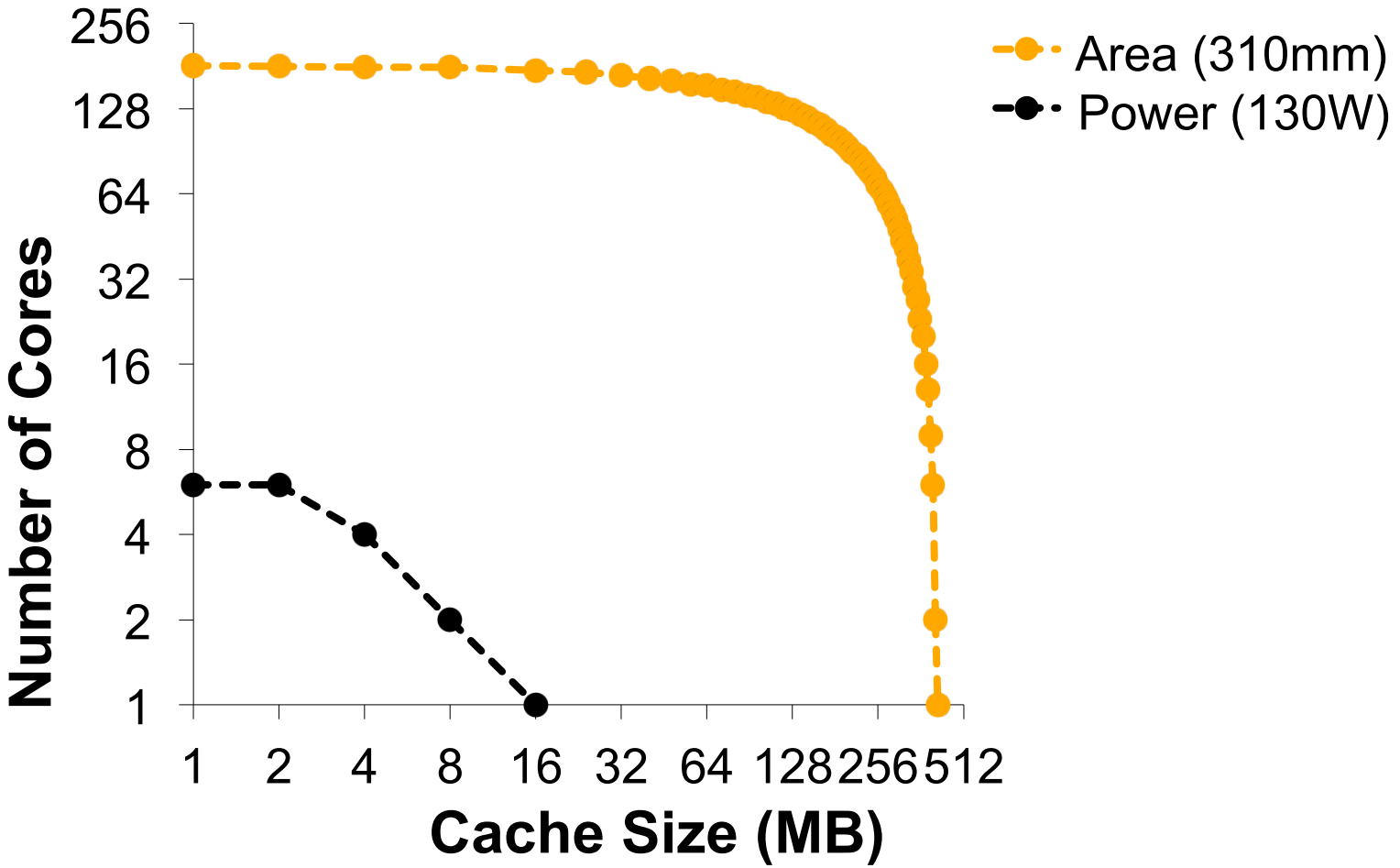
Physical characteristics modeled after UltraSPARC T2, ARM11

- **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

Caveats

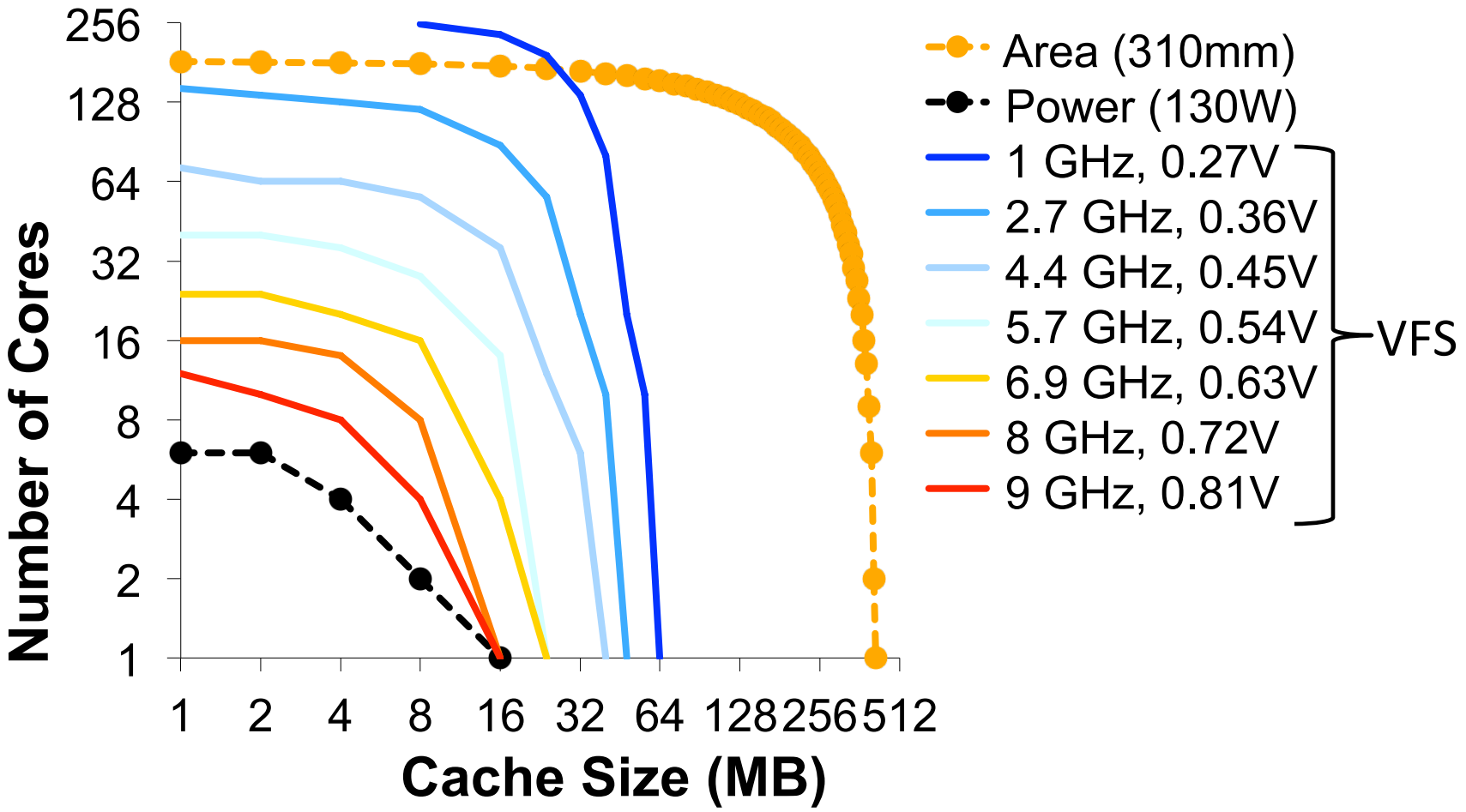
- First-order model
  - The intent is to uncover trends relating the effects of technology-driven physical constraints to the performance of commercial workloads running on multicores
  - The intent is NOT to offer absolute numbers
- Performance model works well for workloads with low MLP
  - Database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

Area vs. Power Envelope



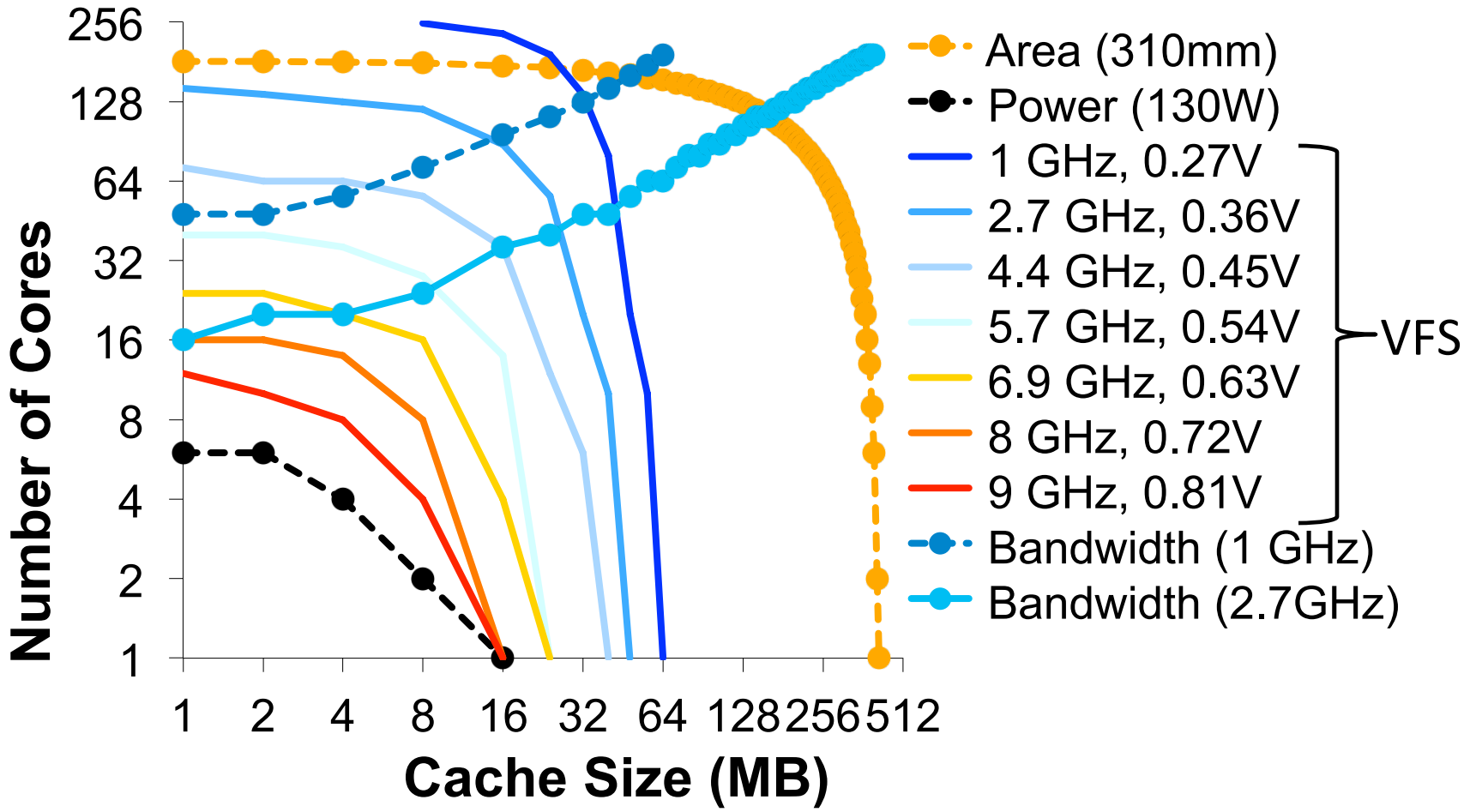
Good news: can fit 100's cores. Bad news: cannot power them all

Pack More Slower Cores, Cheaper Cache



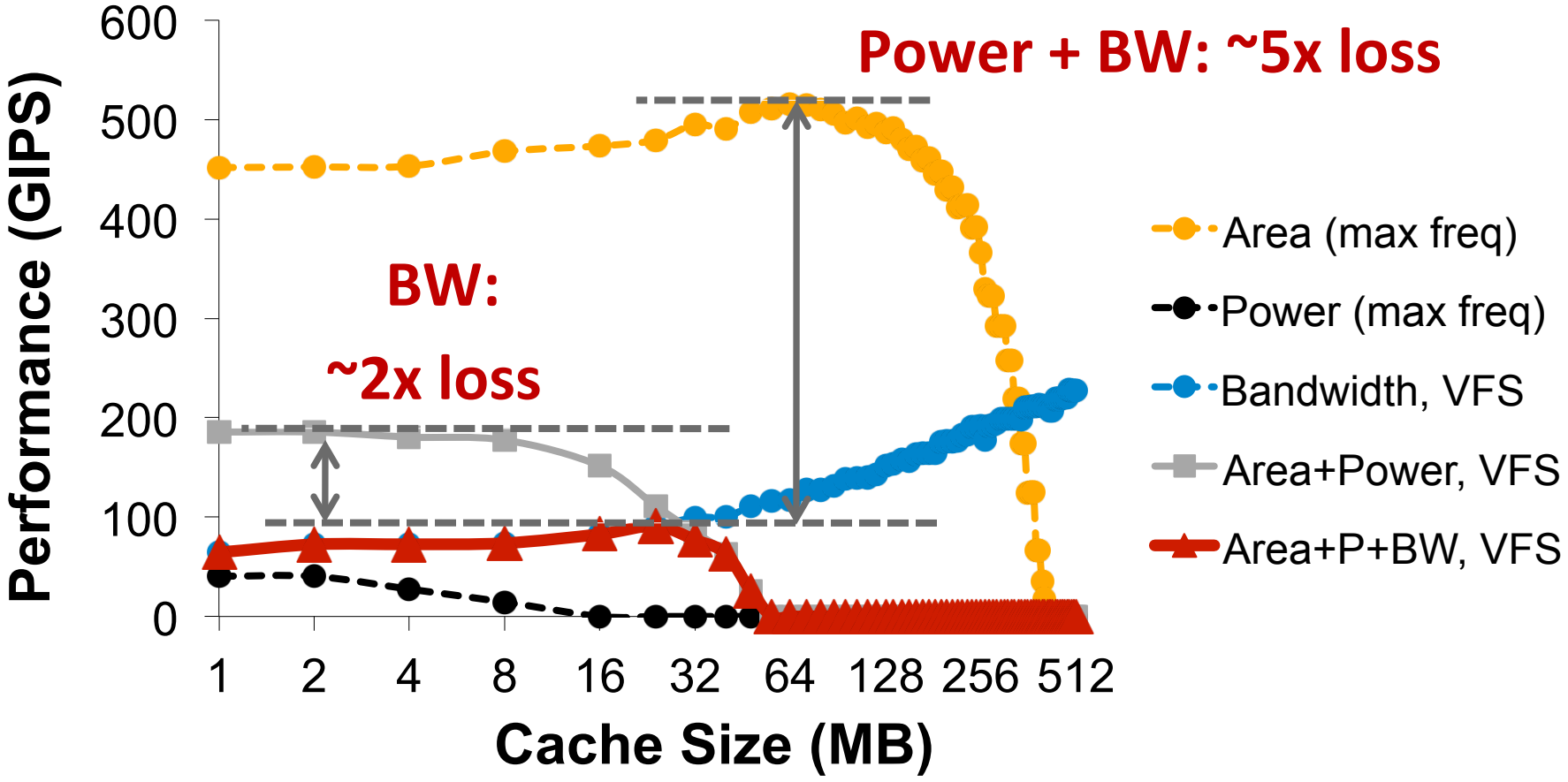
- The reality of The Power Wall: a power-performance trade-off

Pin Bandwidth Constraint



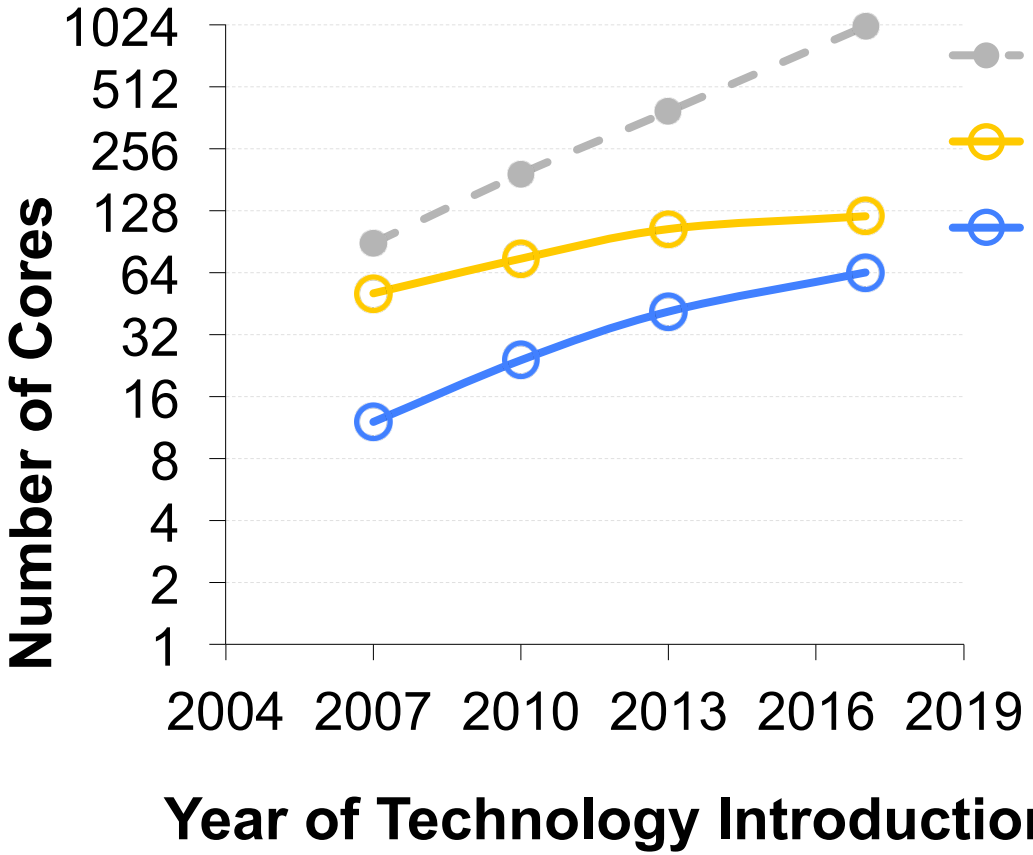
- Bandwidth constraint favors fewer + slower cores, more cache

Example of Optimization Results



- First bandwidth-constrained, then power-constrained
- Fewer + slower cores, lots of cache

Core Counts for Peak-Performance Designs

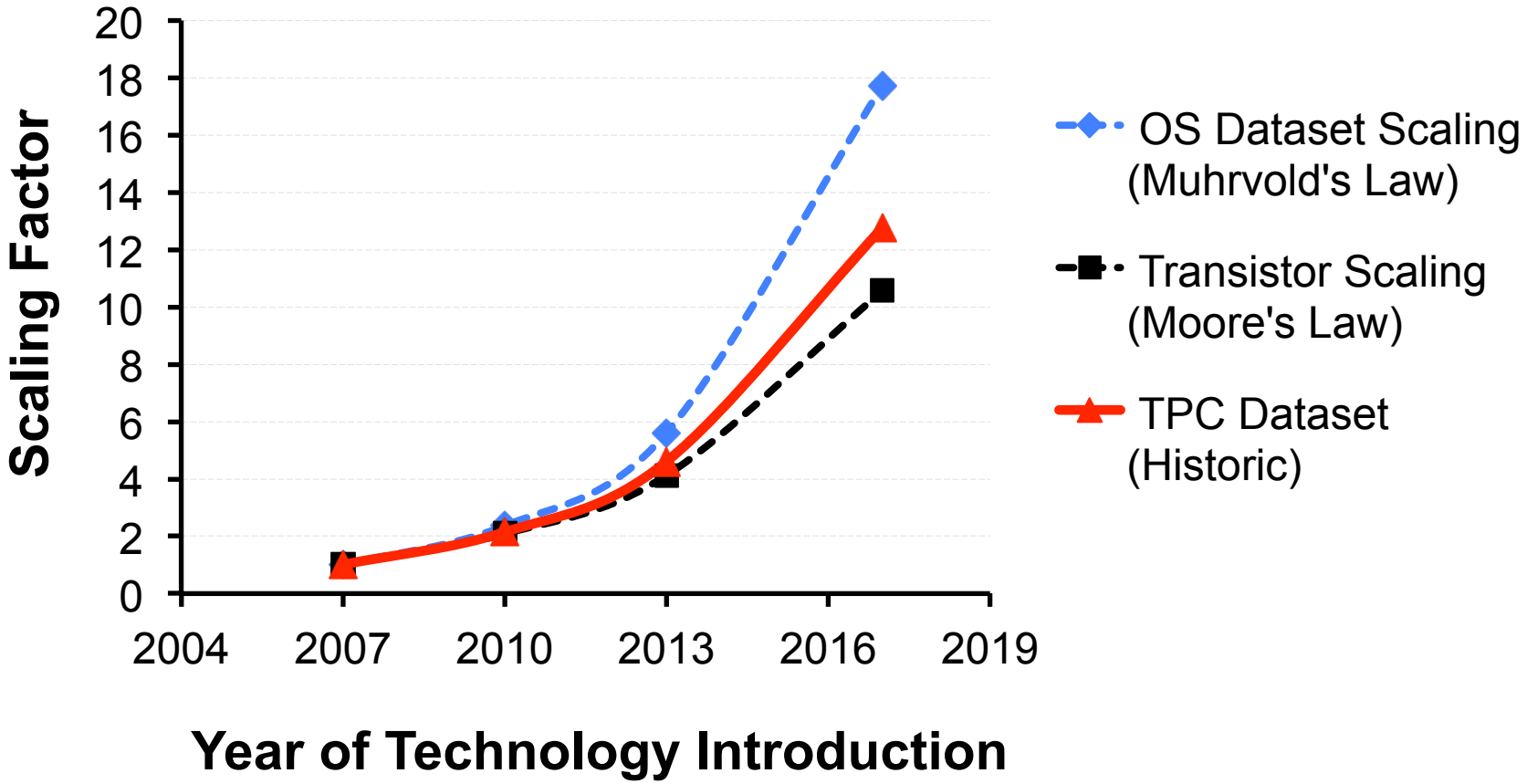


Physical characteristics modeled after

- UltraSPARC (GPP)
- ARM11 (EMB)

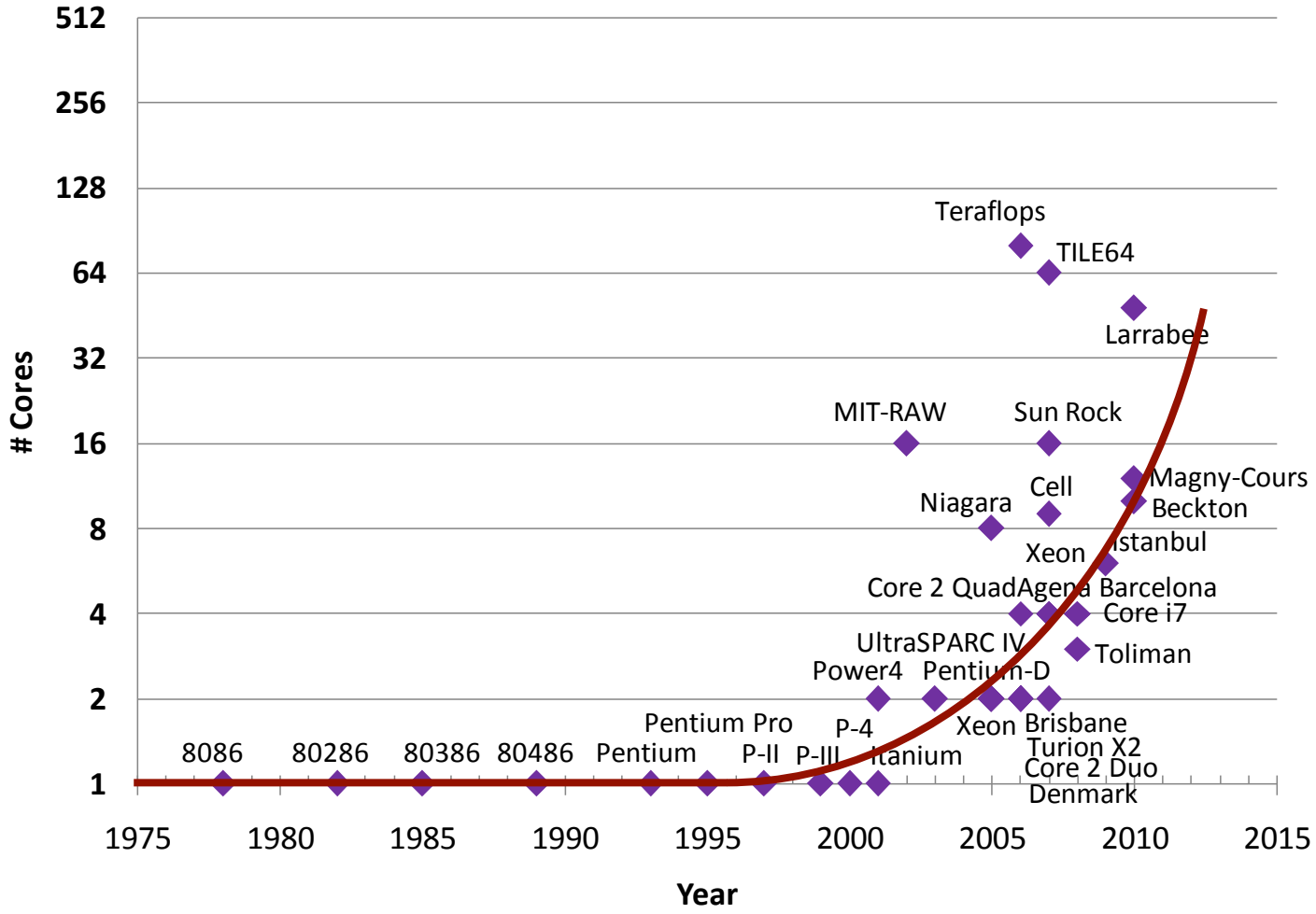
- Designs > 64-120 cores impractical for server workloads
- Pin B/W and power envelopes + dataset scaling limit core counts

Datasets Scale Faster than Moore's Law



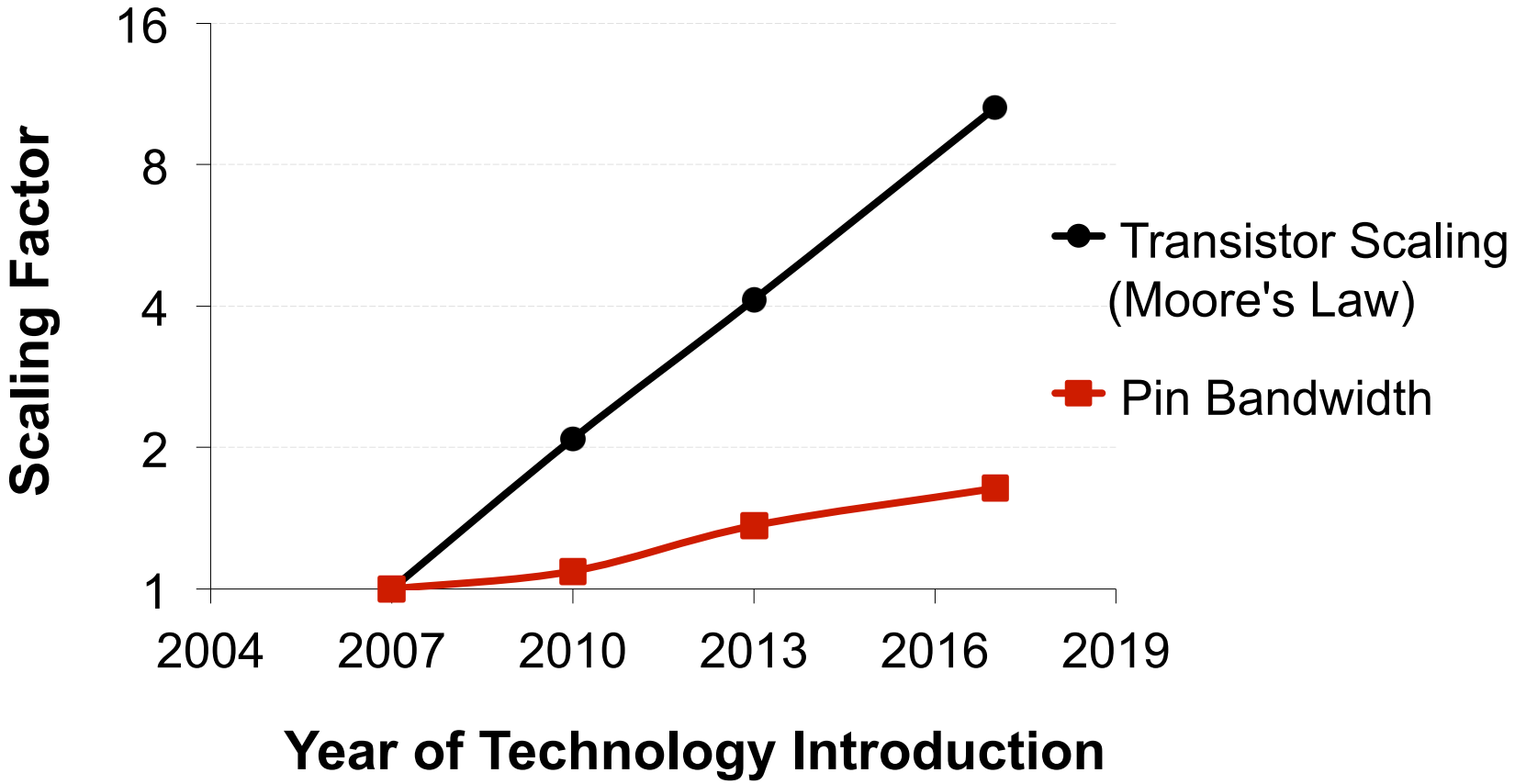
➡ Need more off-chip bandwidth

Core Counts Increase Fast (thus far…)



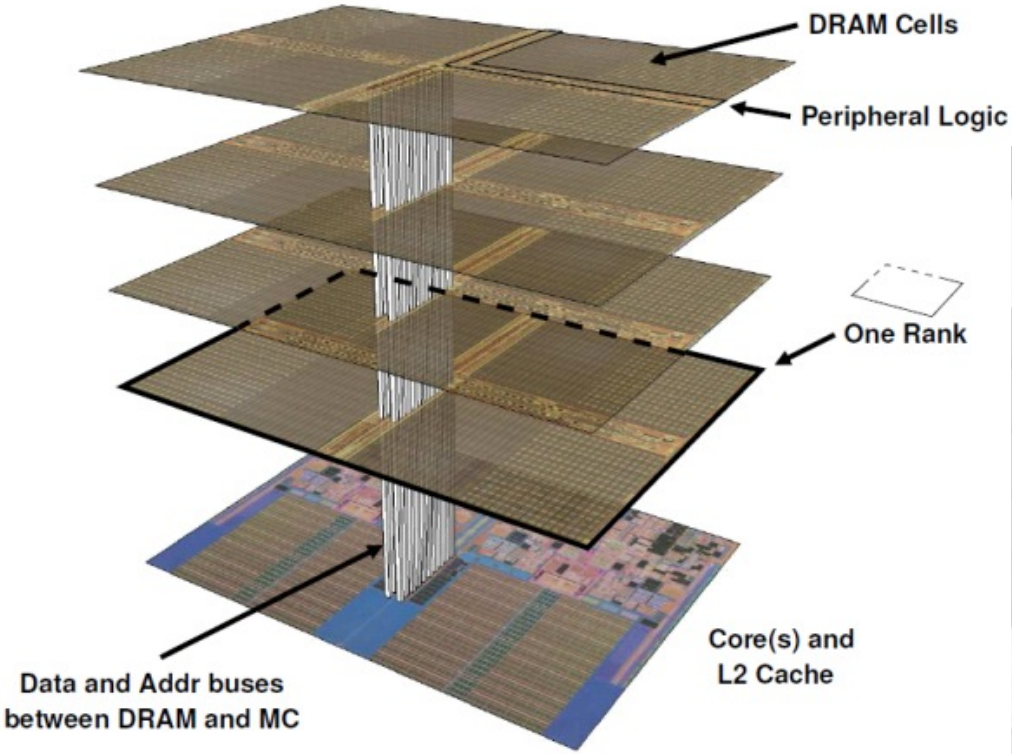
➡ Need more off-chip bandwidth

But, Off-Chip Bandwidth Scales Slowly

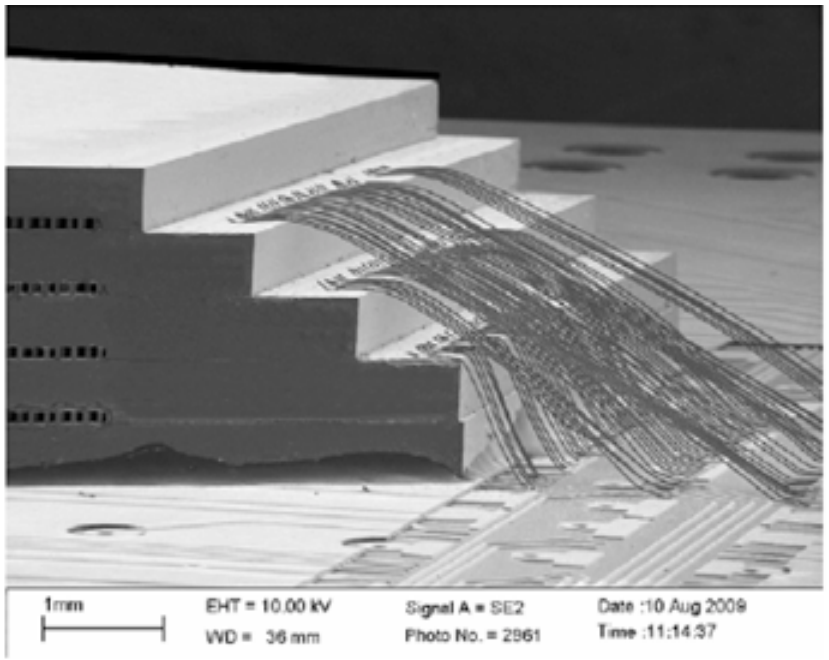


➡ Limited by #pins, off-chip clock → meet The Bandwidth Wall!

Breaking the Bandwidth Wall: 3D-die stacking



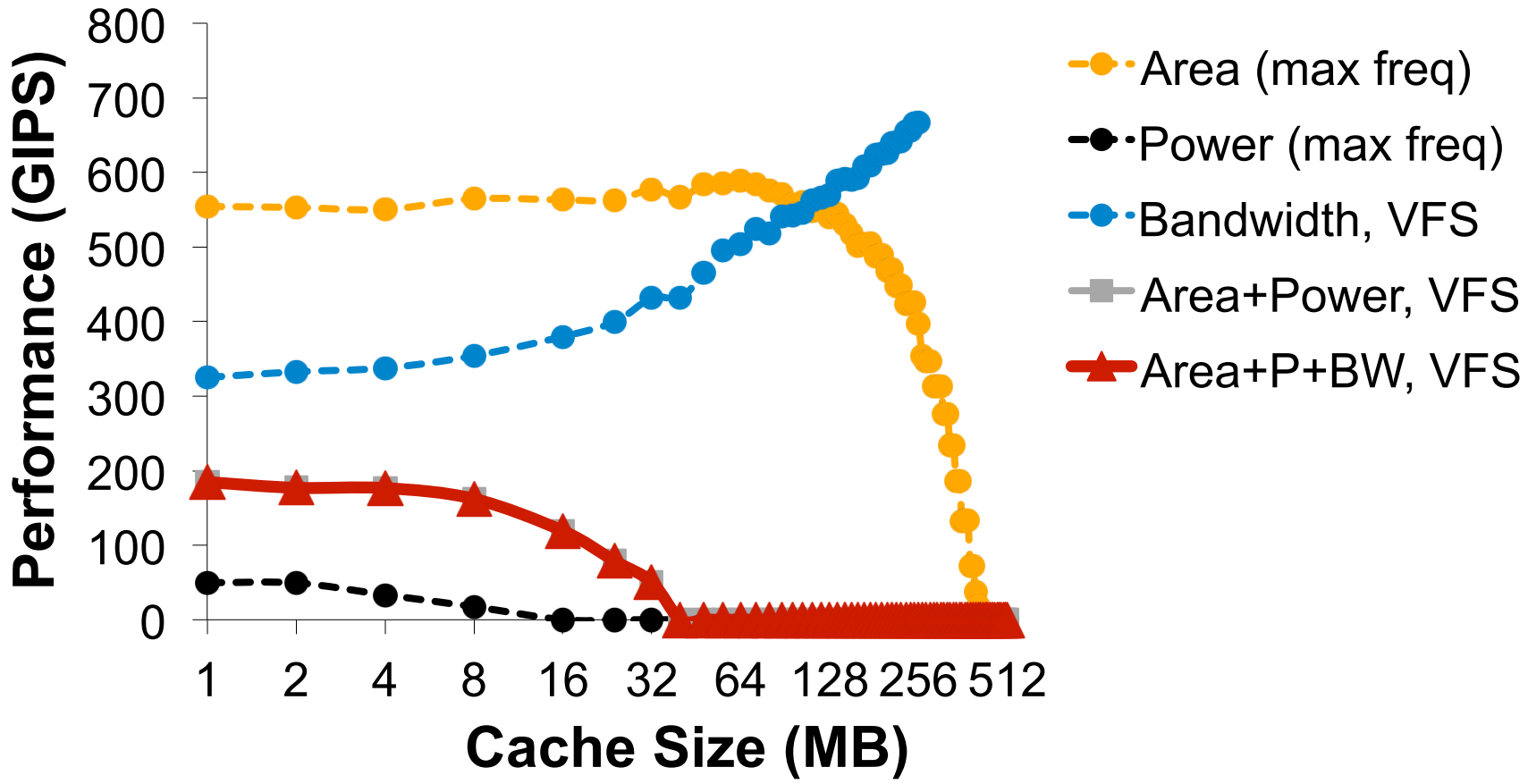
[Loh et al., ISCA'08]



[IBM]

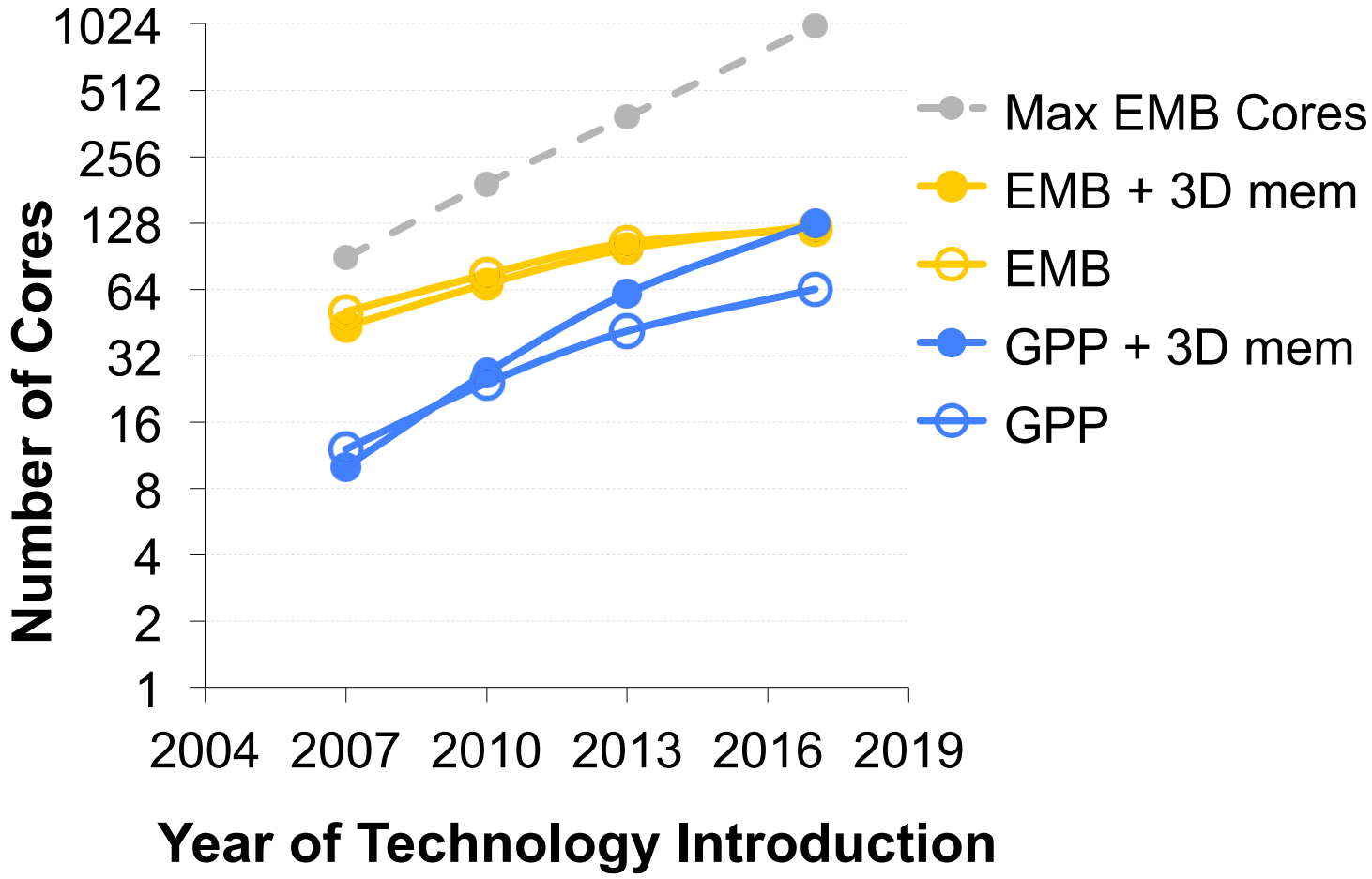
- Delivers TB/sec of bandwidth to "in-package" DRAM (use as cache)

Performance Analysis of 3D-Stacked Multicores



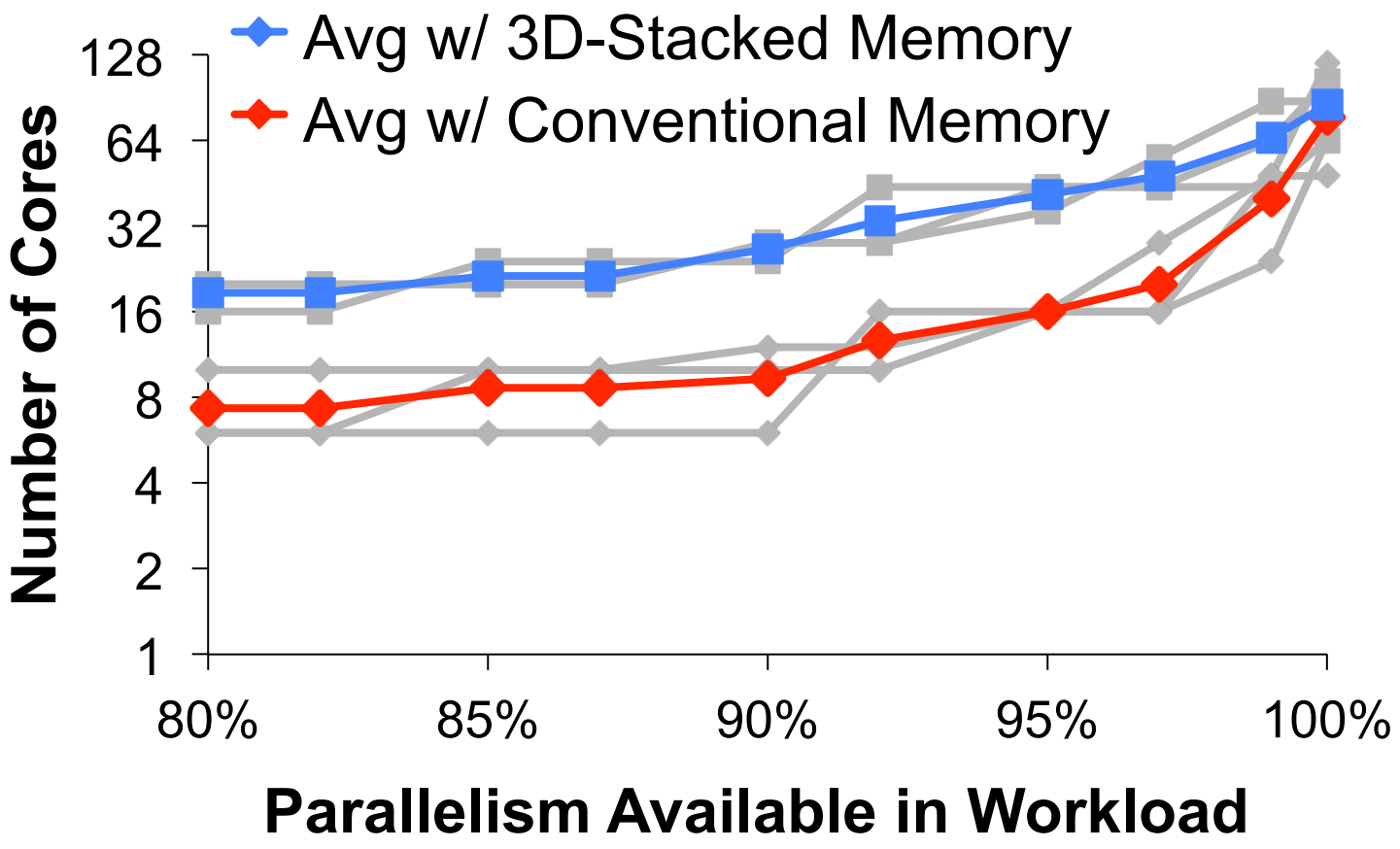
➧ Chip becomes fully power-constrained

Peak-Performance 3D-Stacked Multicore Designs



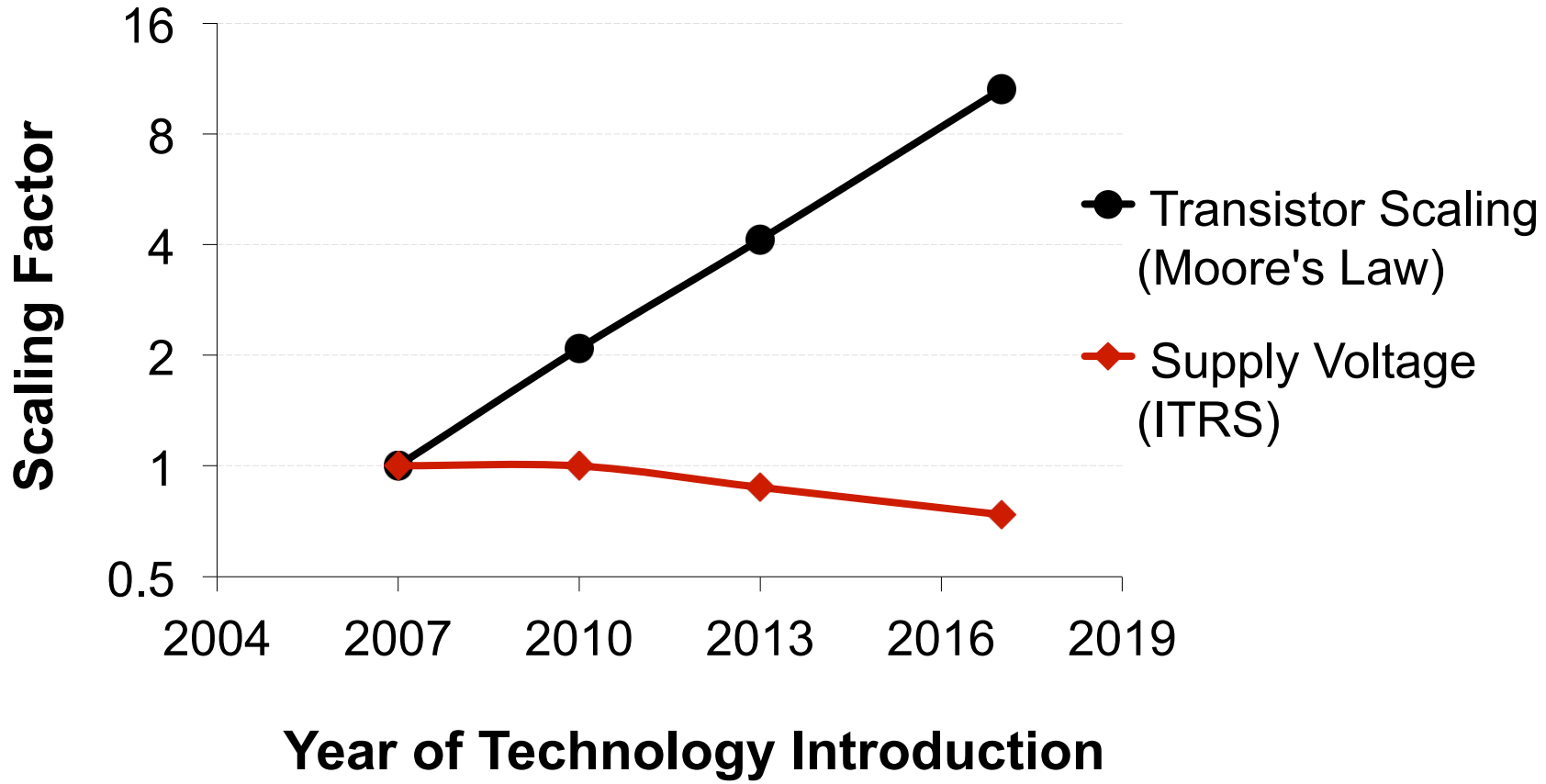
➡ Power envelope + Amdahl's Law limit the core count

Impact of Amdhal's Law



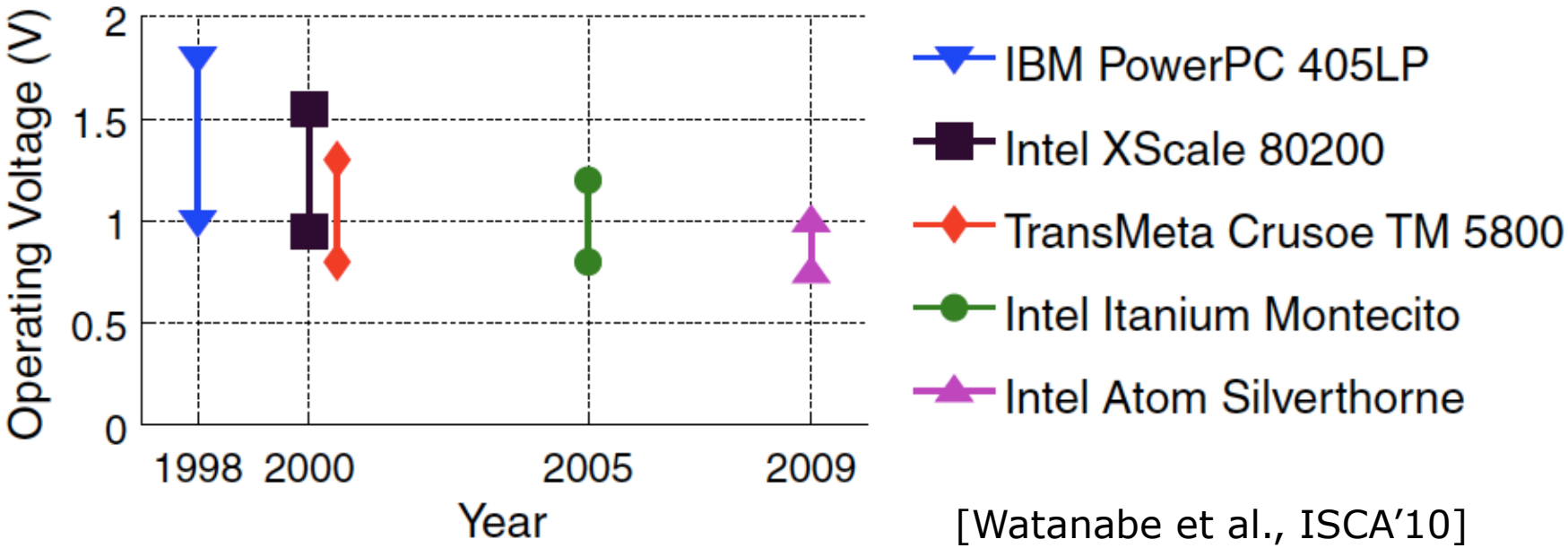
- Even 100% parallel workloads with 3D-memory are limited
  - So, the real limiter is Power!

Voltage Scales Slower Than Moore's Law



➡ Need exponentially more power per chip

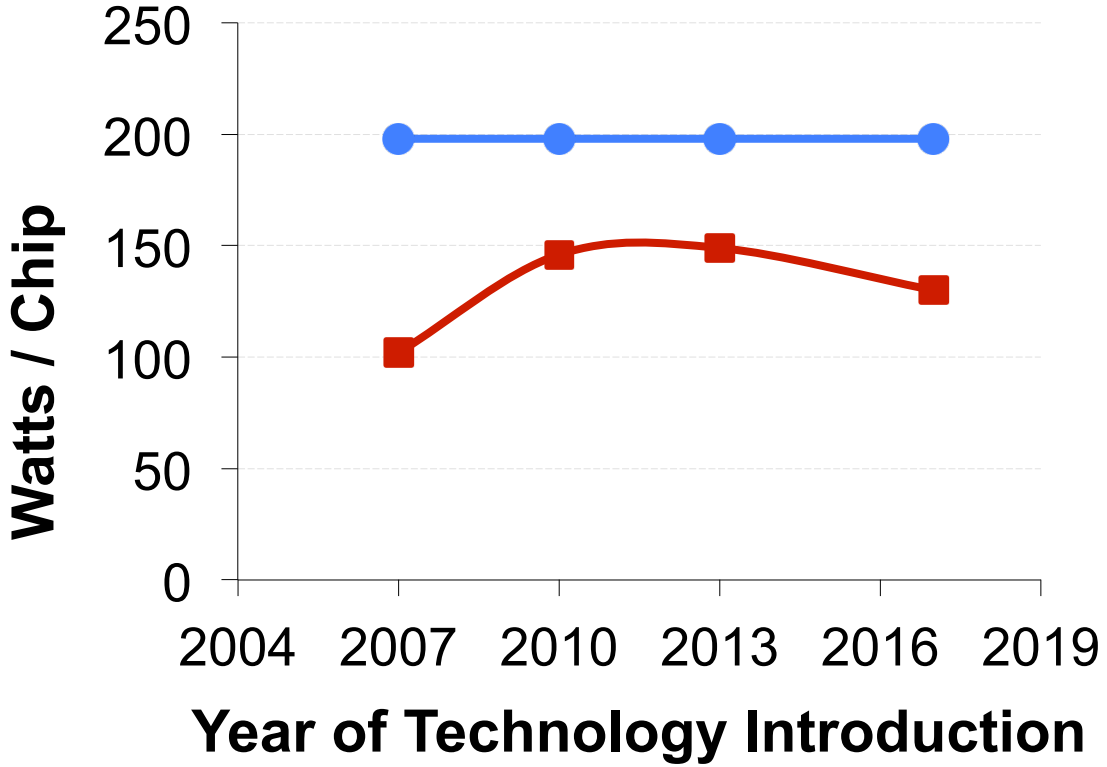
Conventional Power-Reducing Techniques Are Inadequate



[Watanabe et al., ISCA'10]

- Shrinking range of operational voltage hampers voltage-freq. scaling
  - Traditional techniques cannot reduce power requirements

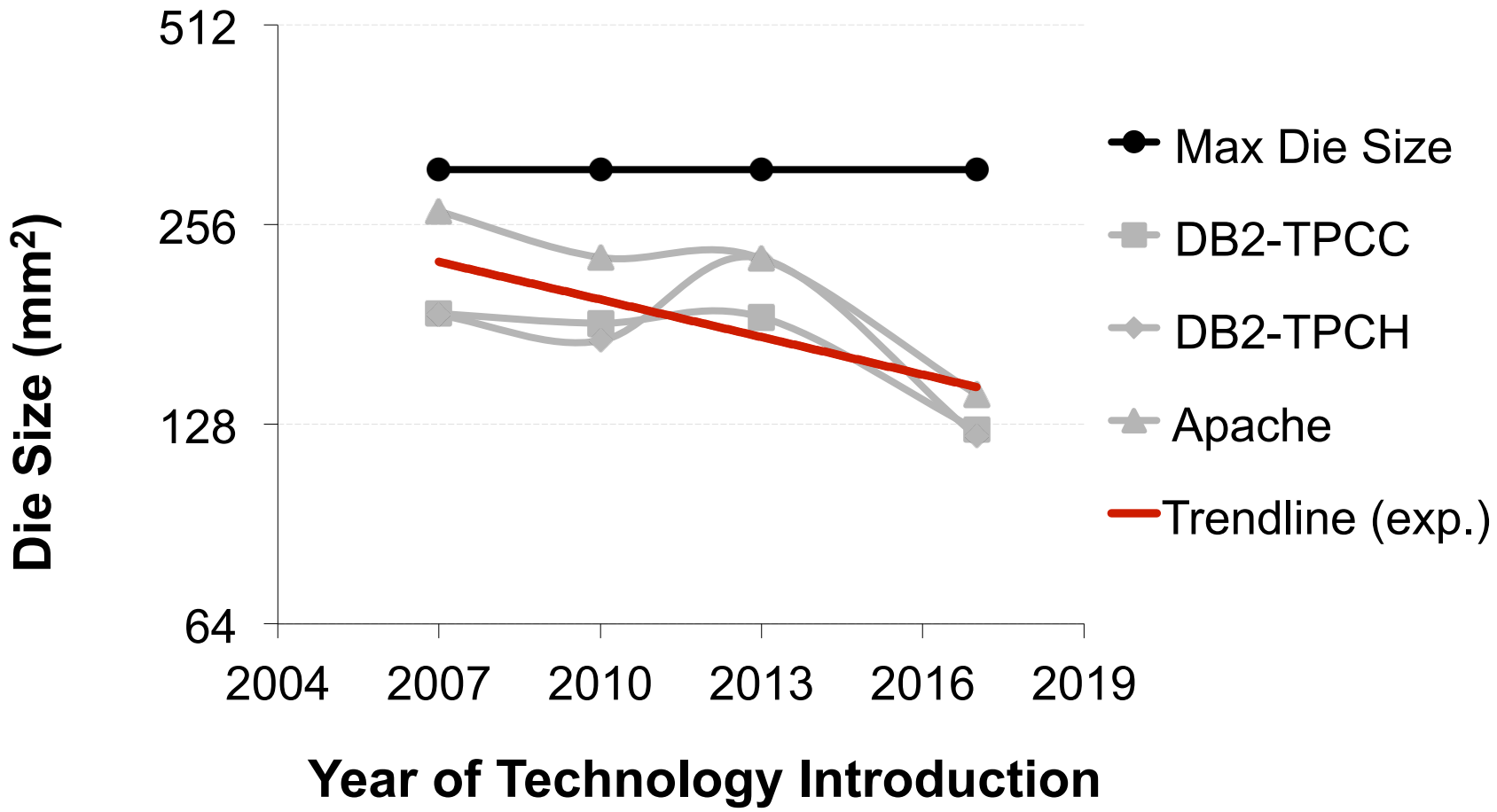
But Chip Power Does Not Scale



Note: even if cooling is solved, power delivery may be the new constraint

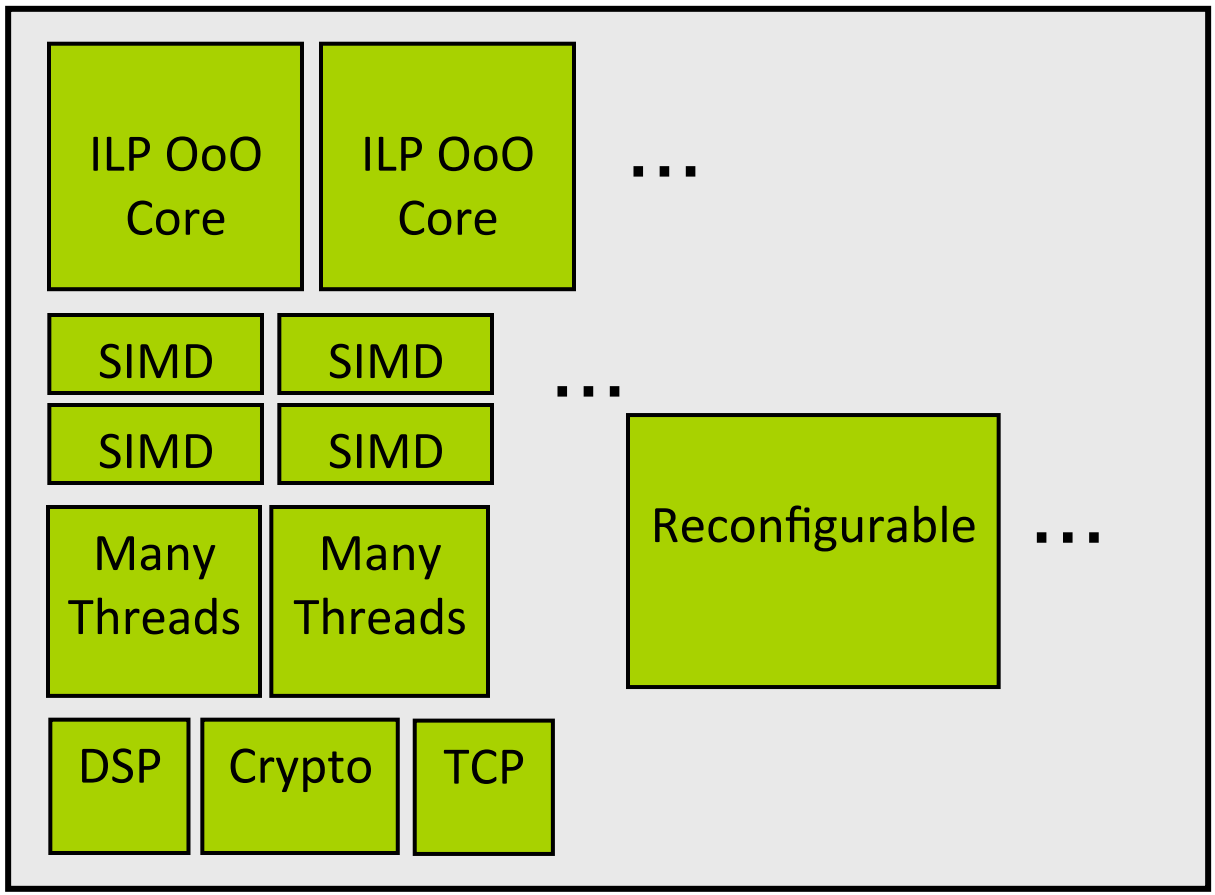
- Chip power does not scale, but more transistors need more power
- Cannot power all silicon simultaneously! Large die area left unused!

Exponentially Large Die Area Left Unutilized



➡ Exploit unutilized area to build specialized cores

Example of a Specialized Multicore Chip



- Many custom cores on chip; power only the most useful ones

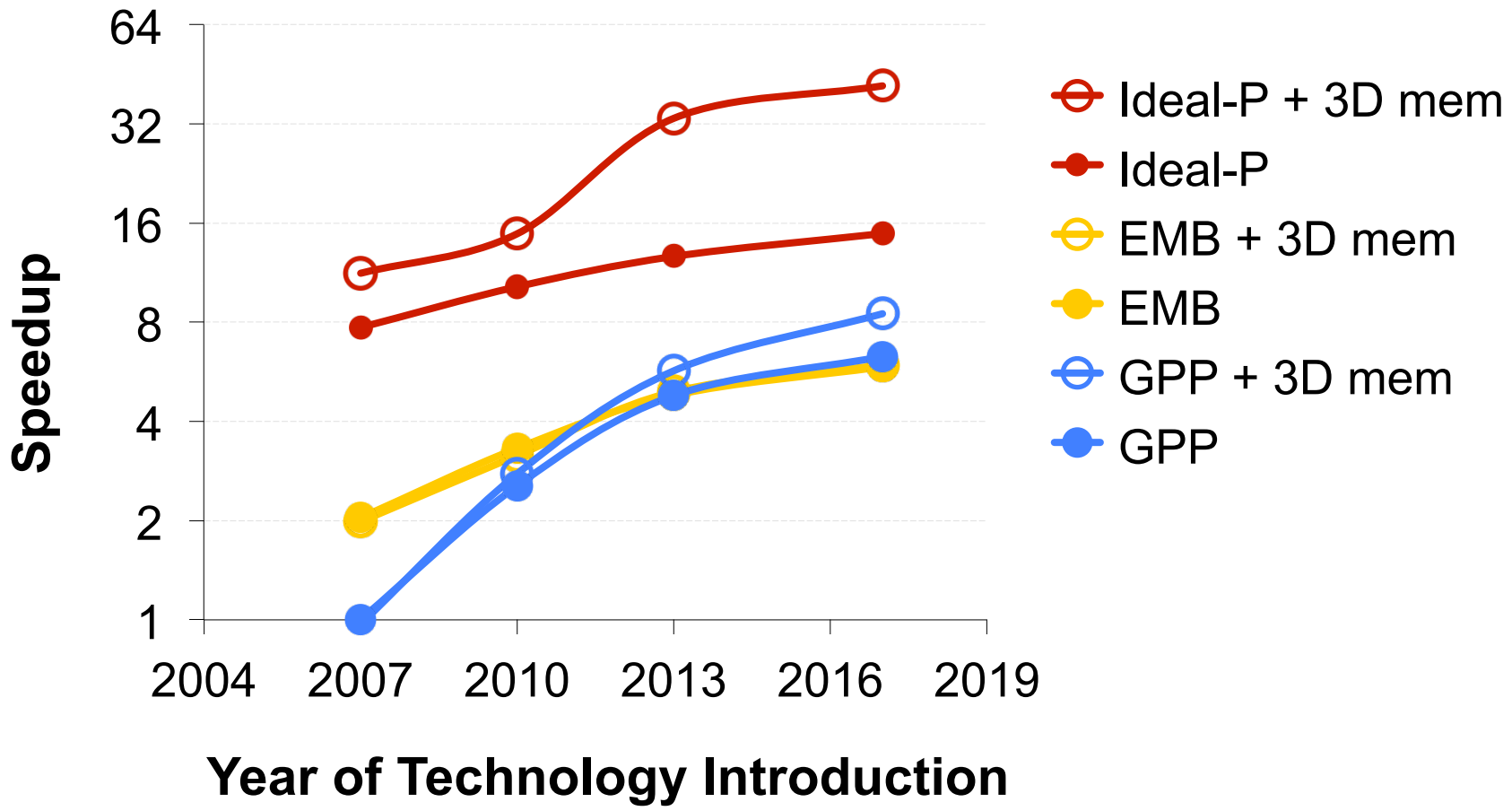
First-Order Core Specialization Model

- 720p HD H.264 encoder (high-definition video encoder)
- Several optimized implementations exist
  - Commercial ASICs, FPGAs, CMP software
- Wide range of computational motifs

		Frames per sec	Energy per frame (mJ)	Performance gap with ASIC	Energy gap with 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

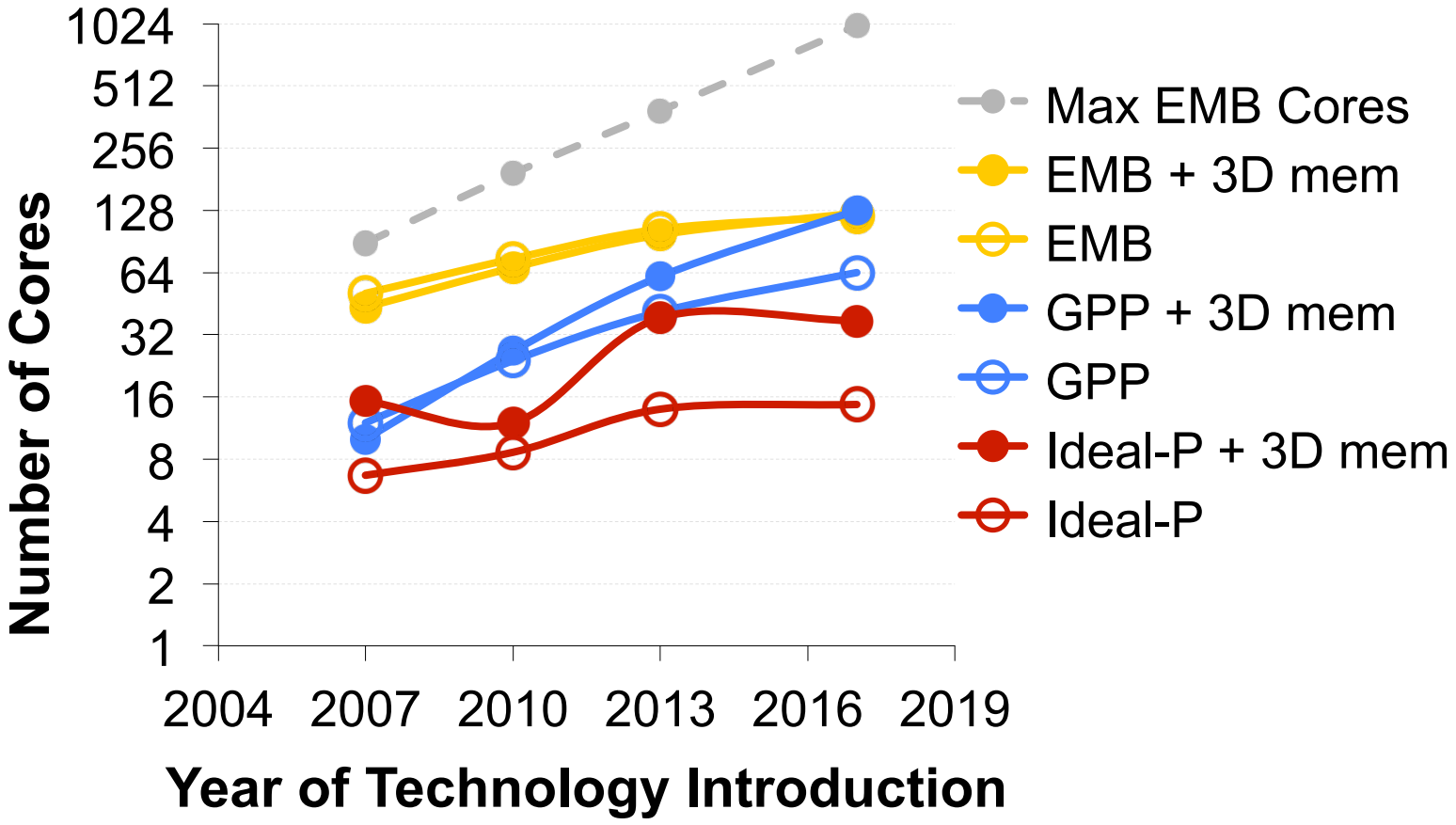
[Hameed et al., ISCA'10]

Performance of Specialized Multicores



➡ Specialized multicores deliver 2x-12x higher performance

Core Counts for Specialized Multicores



- Only few cores need to run at a time
- Vast unused die area will allow the implementation of many cores

Core Specialization

- Could use existing general designs
  - OoO for ILP, in-order-CMT for memory-latency-bound, SIMD for data-parallel, systolic arrays, GPUs-on-chip, etc
- Could use customizable cores
  - Tensilica Xtensa (custom ISA and datapath, operation fusion)
- Could even add reconfigurable logic
- Generality of implemented operations
  - Target specific application
  - Common macro-operations
  - General ISA
- Trade-offs in performance, power, programmability, generality

➡ Wide range of "heterogeneity" and "specialization" meanings

Take-Home Message

- Physical constraints and software pragmatics limit core counts
  - …and performance
- Emerging/exotic technologies may solve some problems
  - …but silicon area will be wasted unless we act on it!
- What should we do? reduce wasted energy per unit of work
  - Heterogeneity, core specialization
  - Use underutilized die area to implement specialized cores
  - Only power the few cores needed
  - The rest of the chip remains off to conserve energy
- Need to innovate across software/hardware stack
  - Programmability, tools are a great challenge

Thank You!

"Multicore: This is the one which will have the biggest impact on us. We have never had a problem to solve like this. A breakthrough is needed in how applications are done on multicore devices."

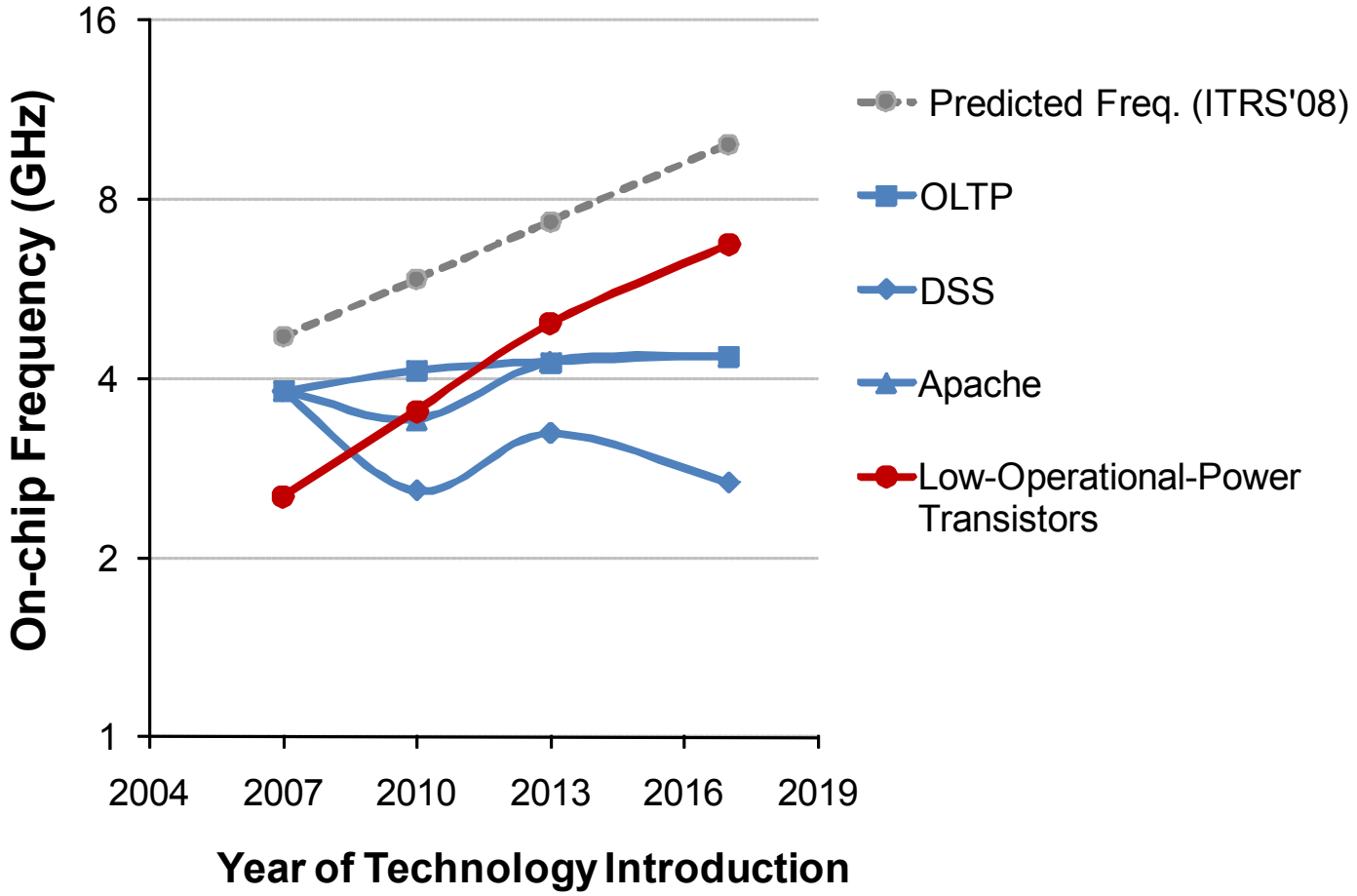
– *Bill Gates*

"It's time we rethink some of the basics of computing. It's scary and lots of fun at the same time."

– *Burton Smith*

Backup

Static Power: Exploit Clock Scaling



- Cores run slow, within range of LOP transistors
- 20x less leakage, no performance hit, 25% higher perf./Watt