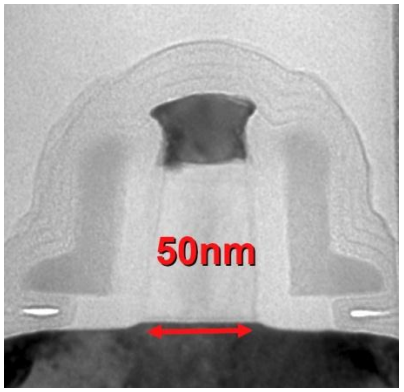
Exploiting Dark Silicon in Server Design

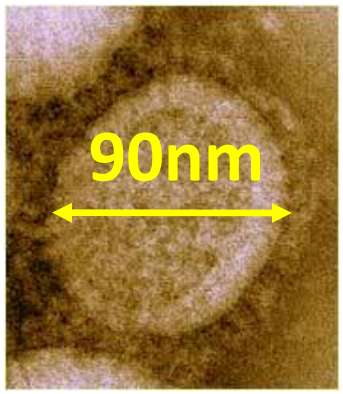
Nikos Hardavellas

Northwestern University, EECS

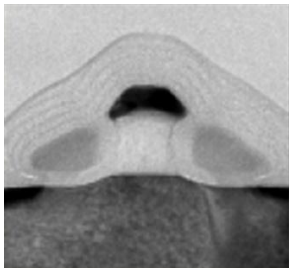
Moore's Law Is Alive And Well



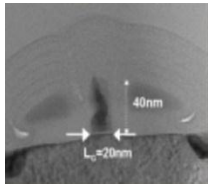
90nm transistor
(Intel, 2005)



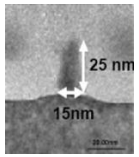
Swine Flu A/H1N1
(CDC)



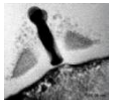
65nm
2007



45nm
2010

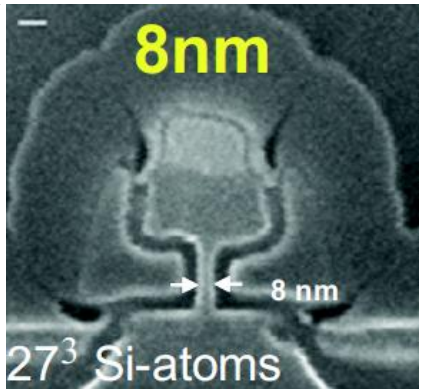


32nm
2013



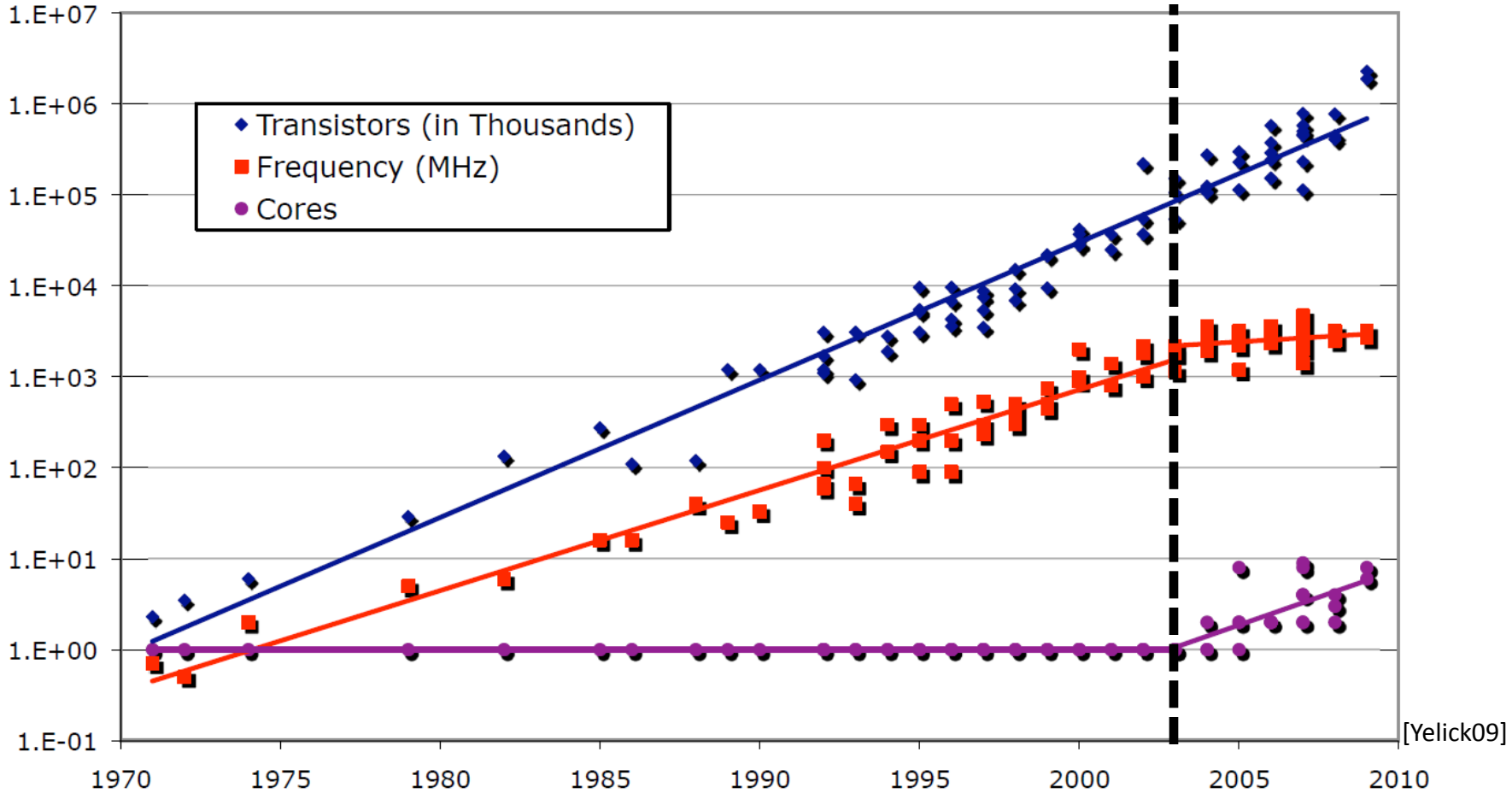
22nm
2016

16nm
2019



- Device scaling continues for at least another 10 years

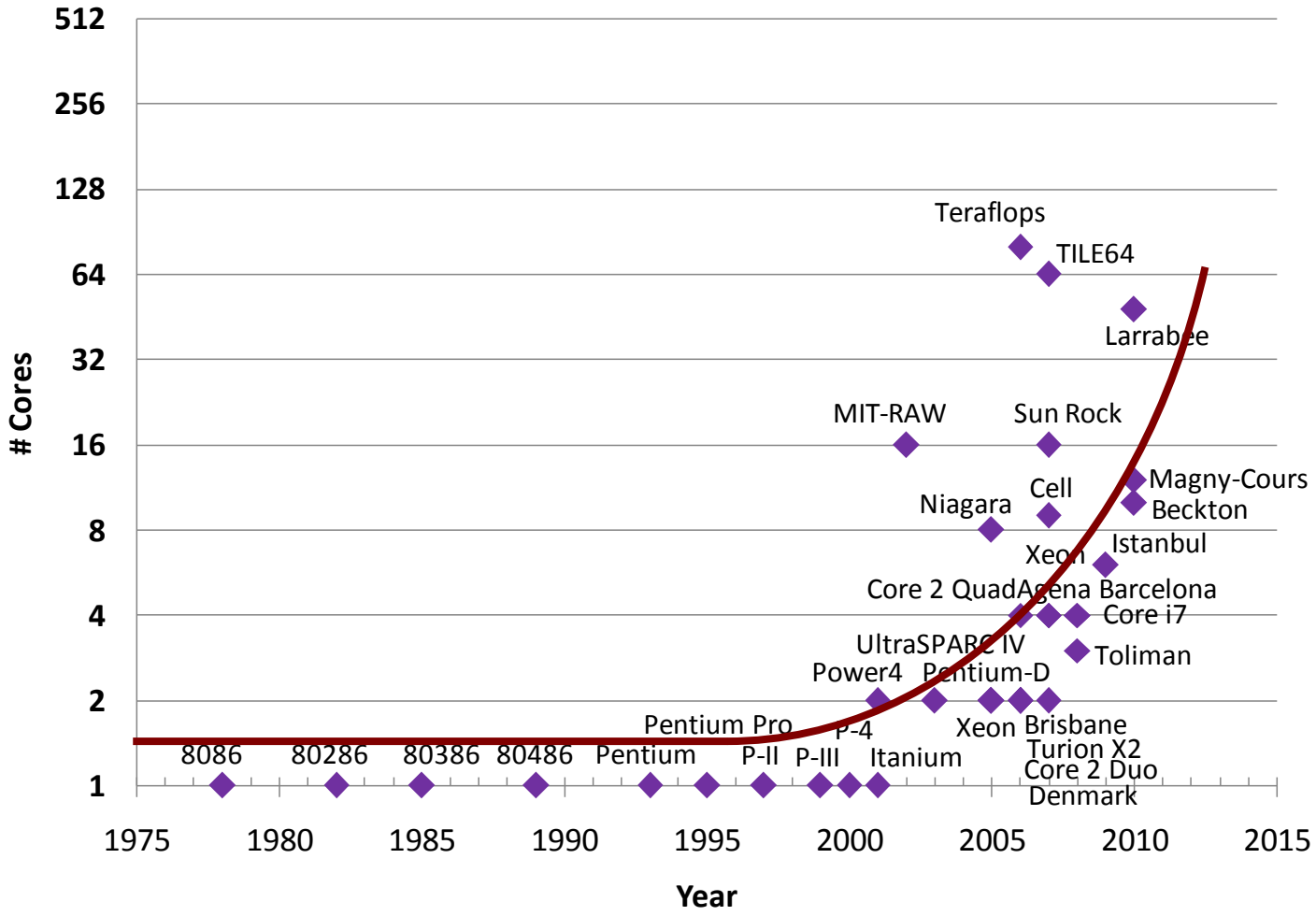
Moore's Days Are Numbered: The Multi-core Era Is Already Here



[Yelick09]

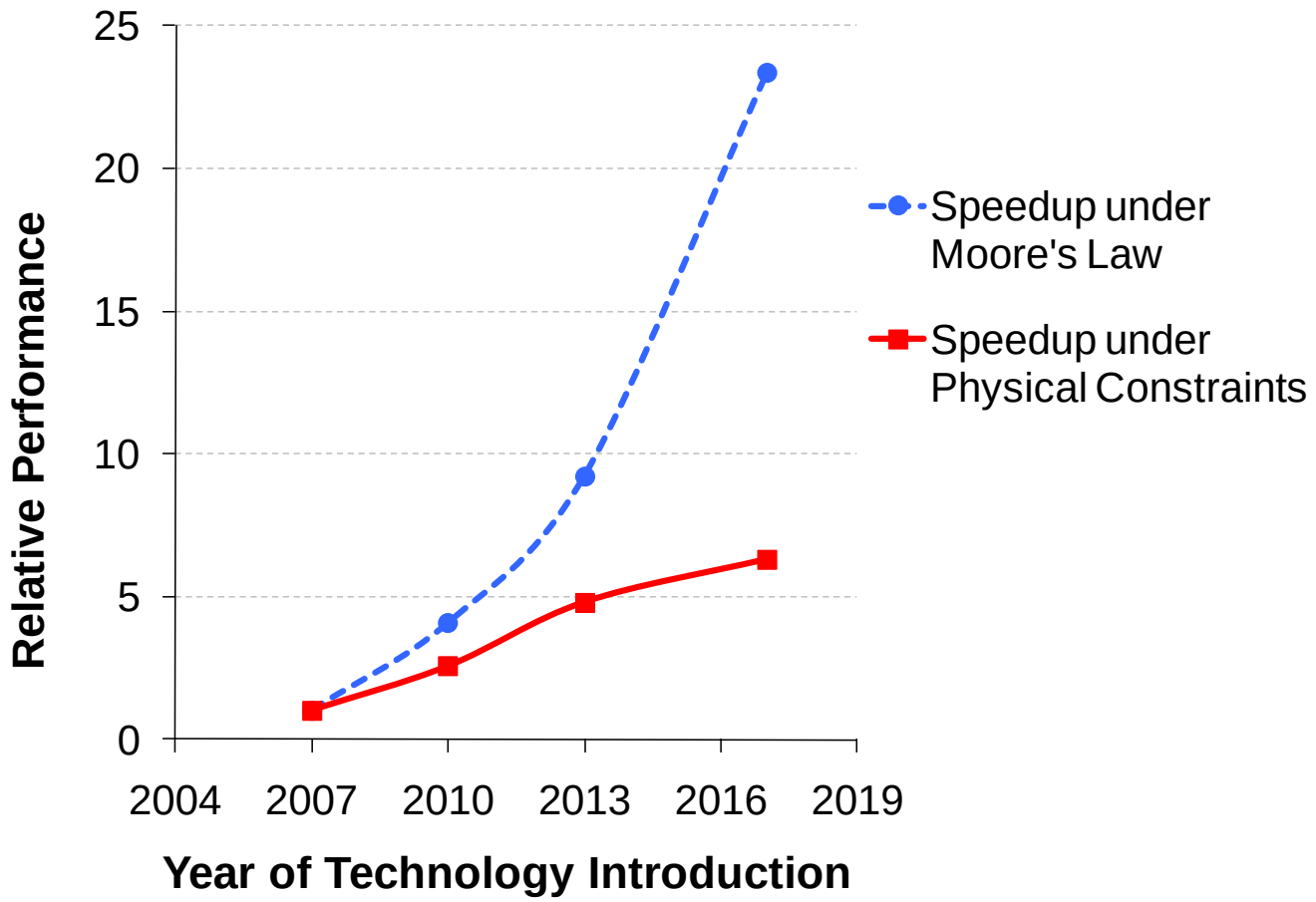
- "New" Moore's Law: 2x cores with every generation

Exponential Growth of Core Counts



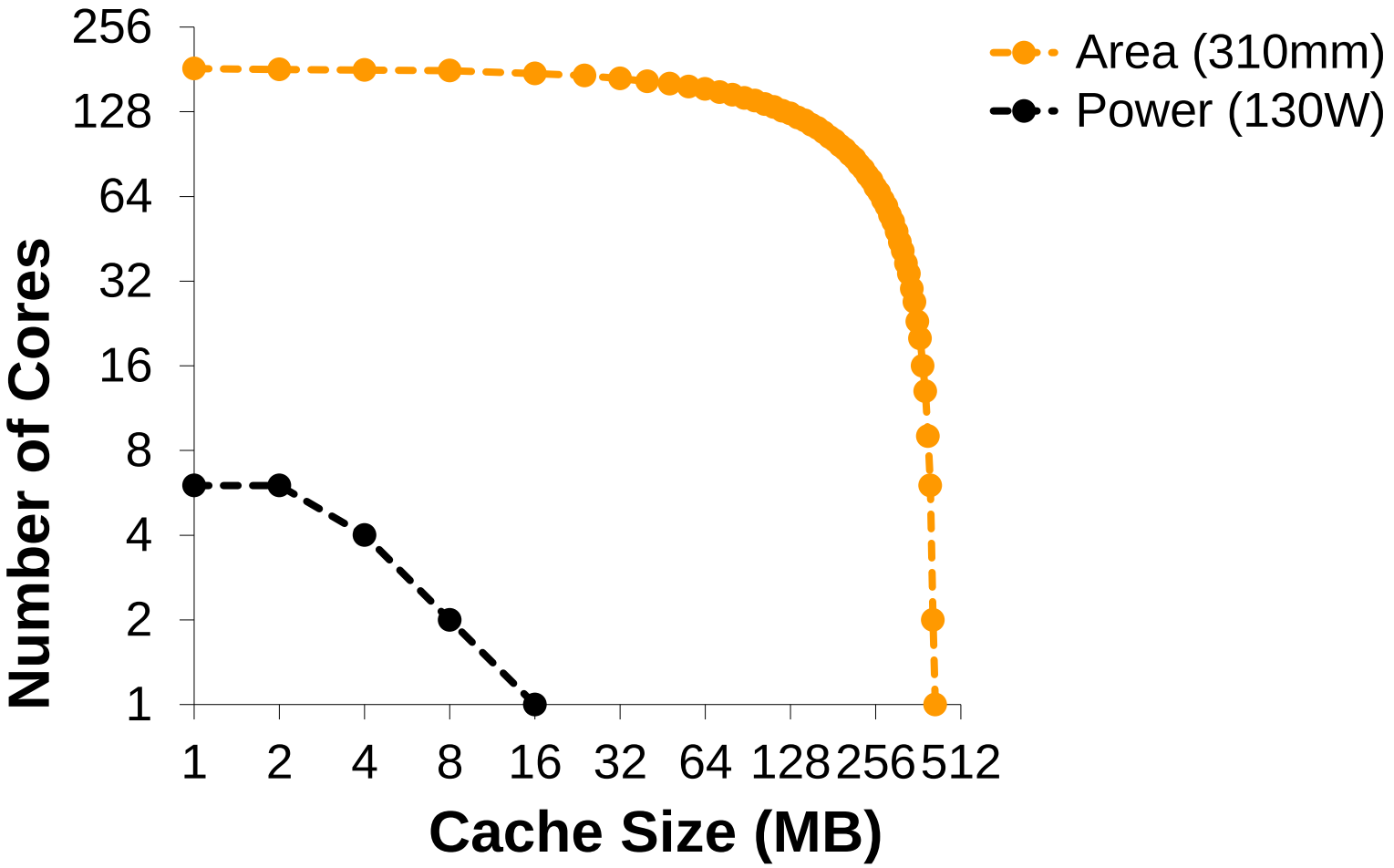
So, are 1000-core chips a viable architecture?

Performance Expectations vs. Reality



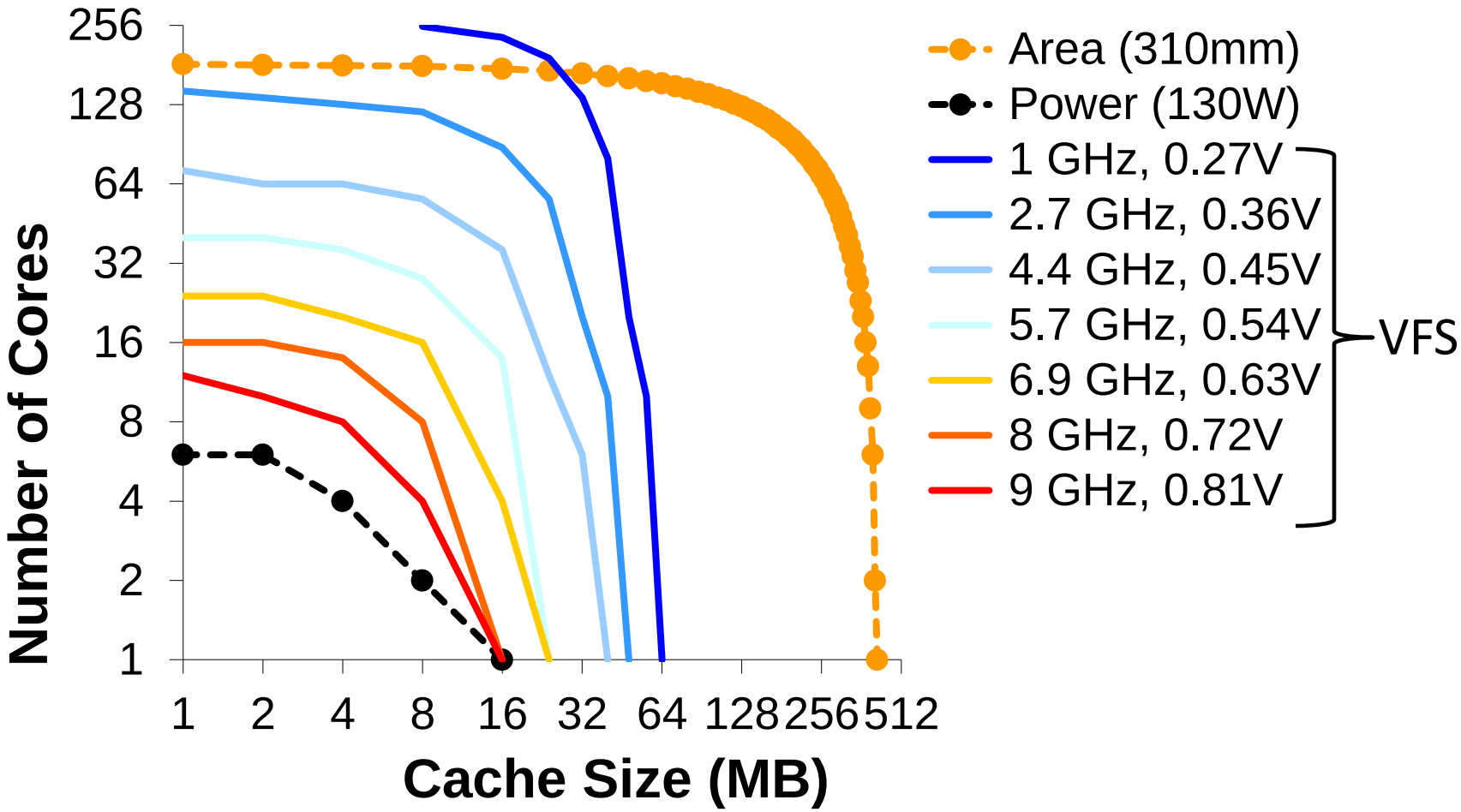
- Physical constraints limit speedup

Area vs. Power Envelope



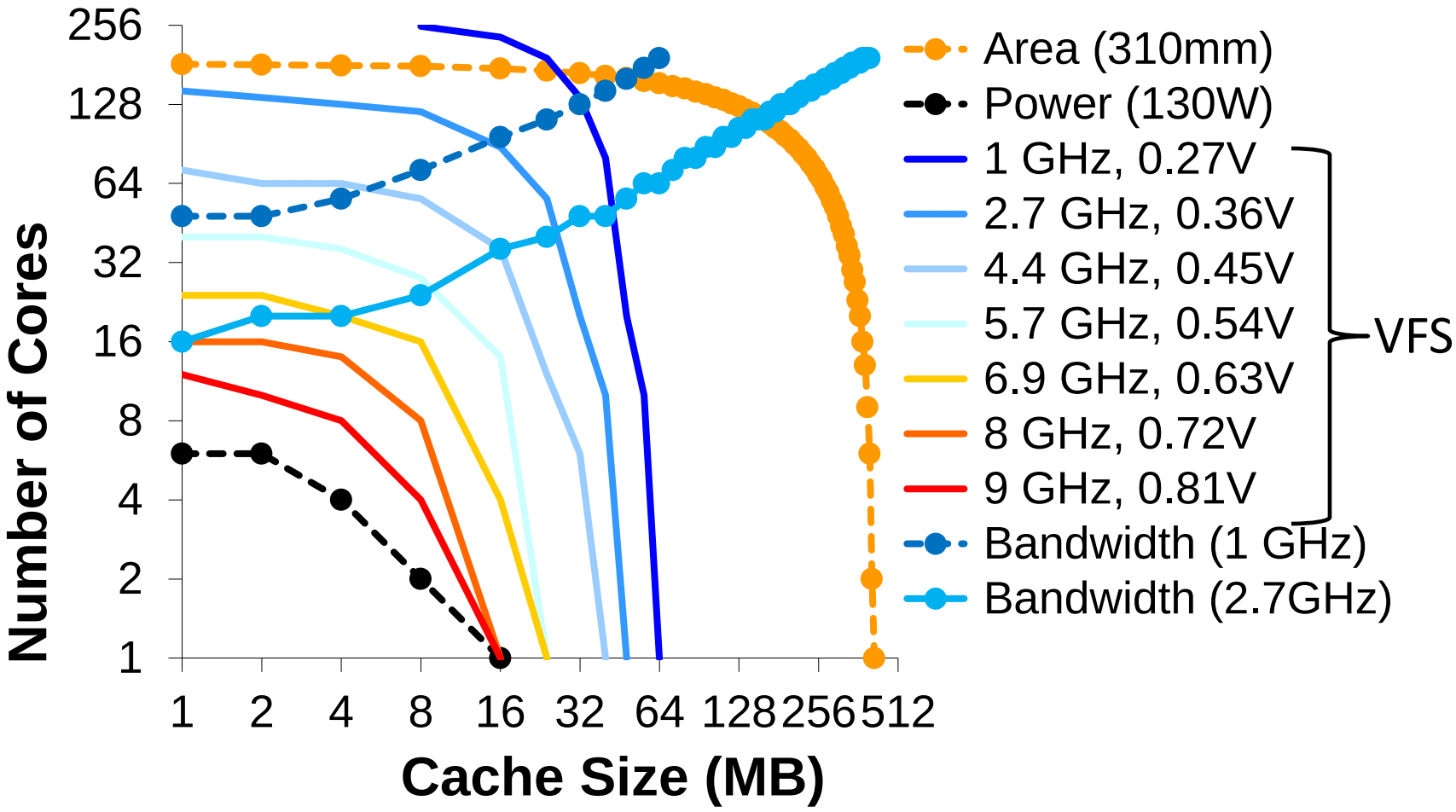
Good news: can fit 100's cores. Bad news: cannot power them all

Pack More Slower Cores, Cheaper Cache



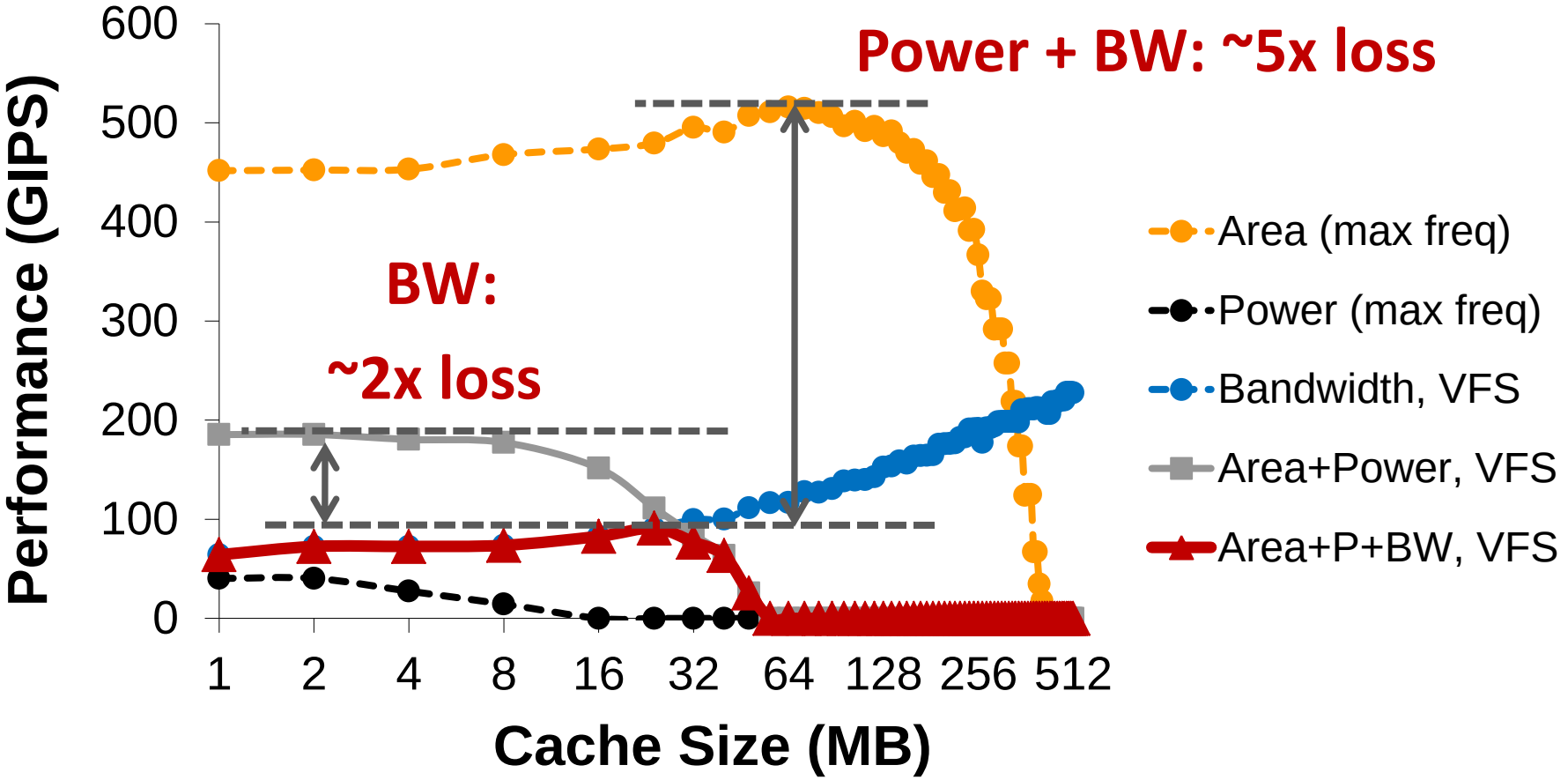
- The reality of The Power Wall: a power-performance trade-off

Pin Bandwidth Constraint



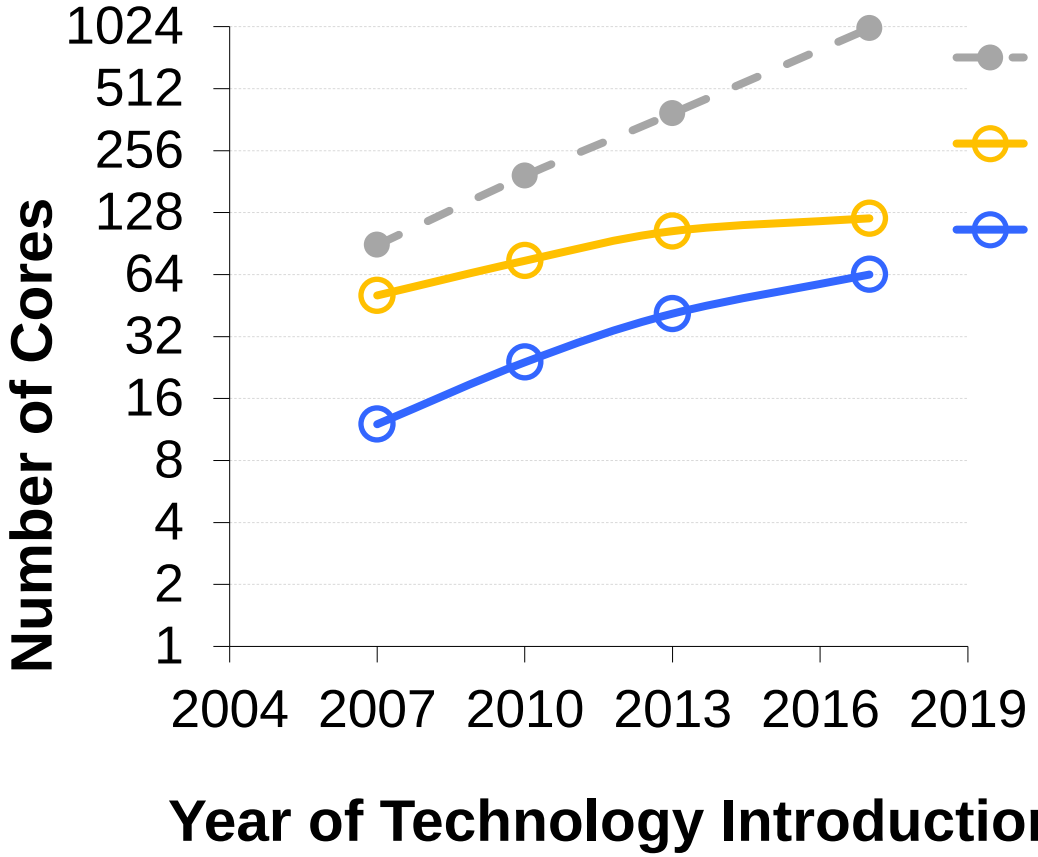
- Bandwidth constraint favors fewer + slower cores, more cache

Example of Optimization Results



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

Core Counts for Peak-Performance Designs

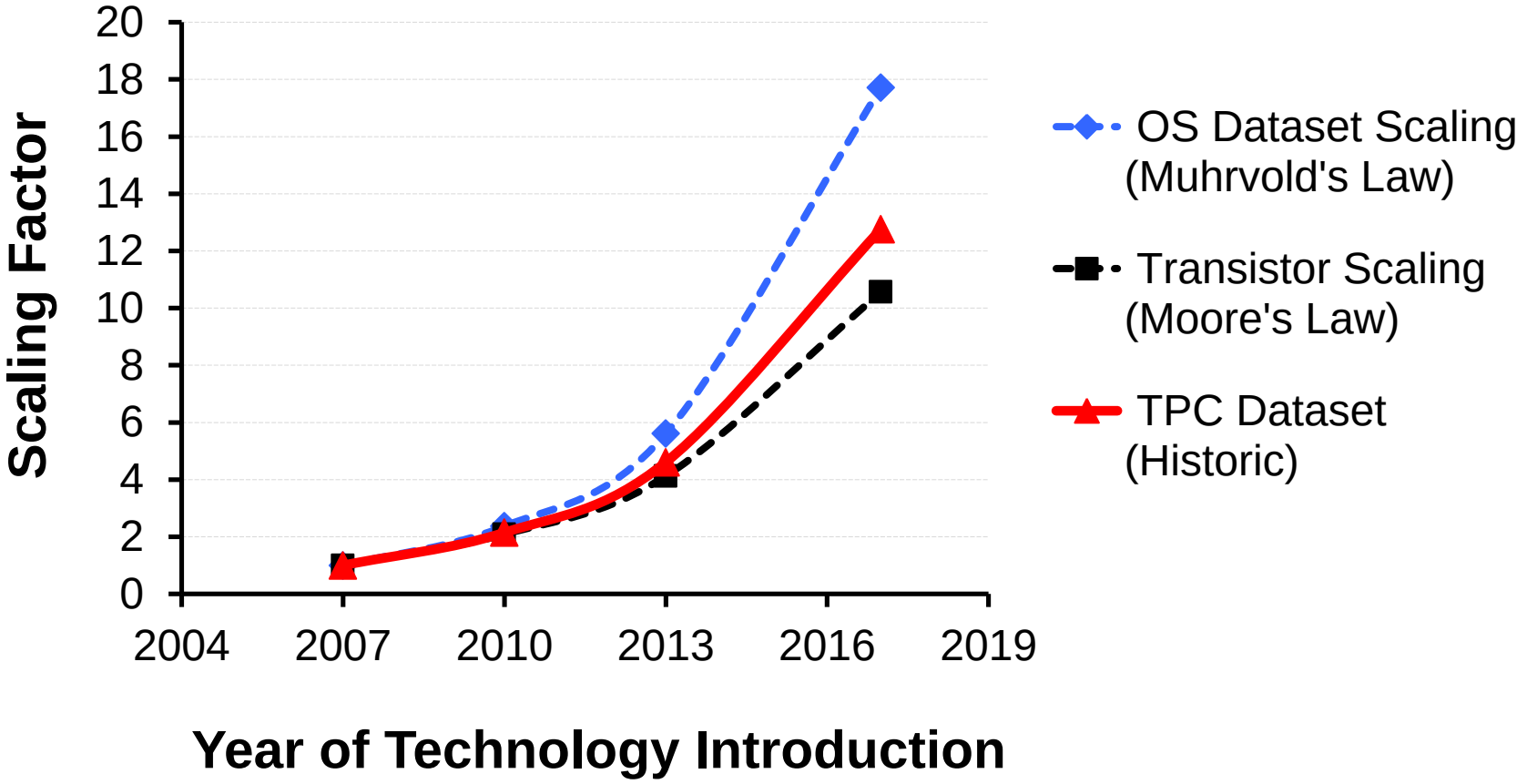


Physical characteristics modeled after

- UltraSPARC T2 (GPP)
- ARM11 (EMB)

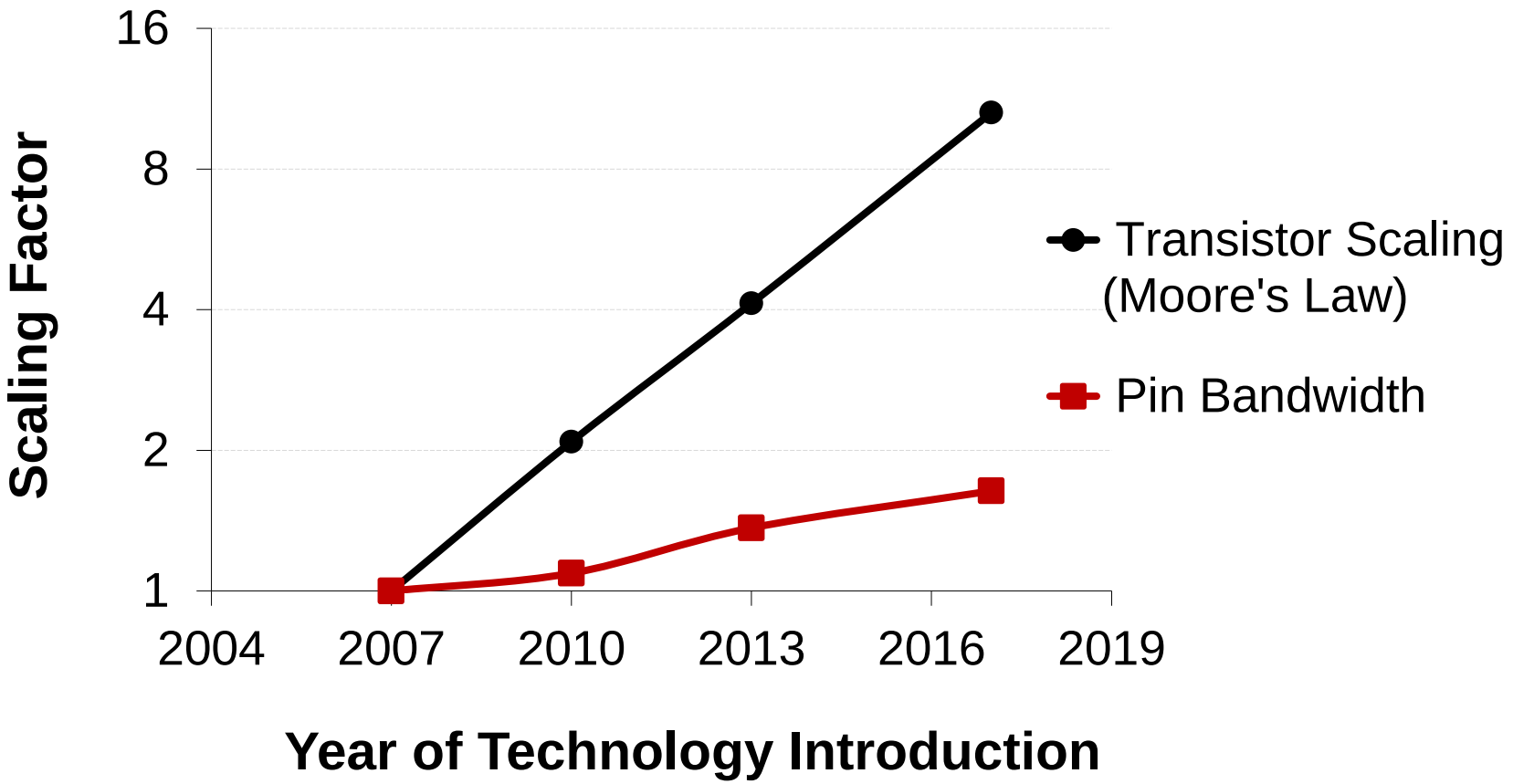
- Designs > 120 cores impractical for general-purpose server apps
- B/W and power envelopes + dataset scaling limit core counts

Application Dataset Scaling



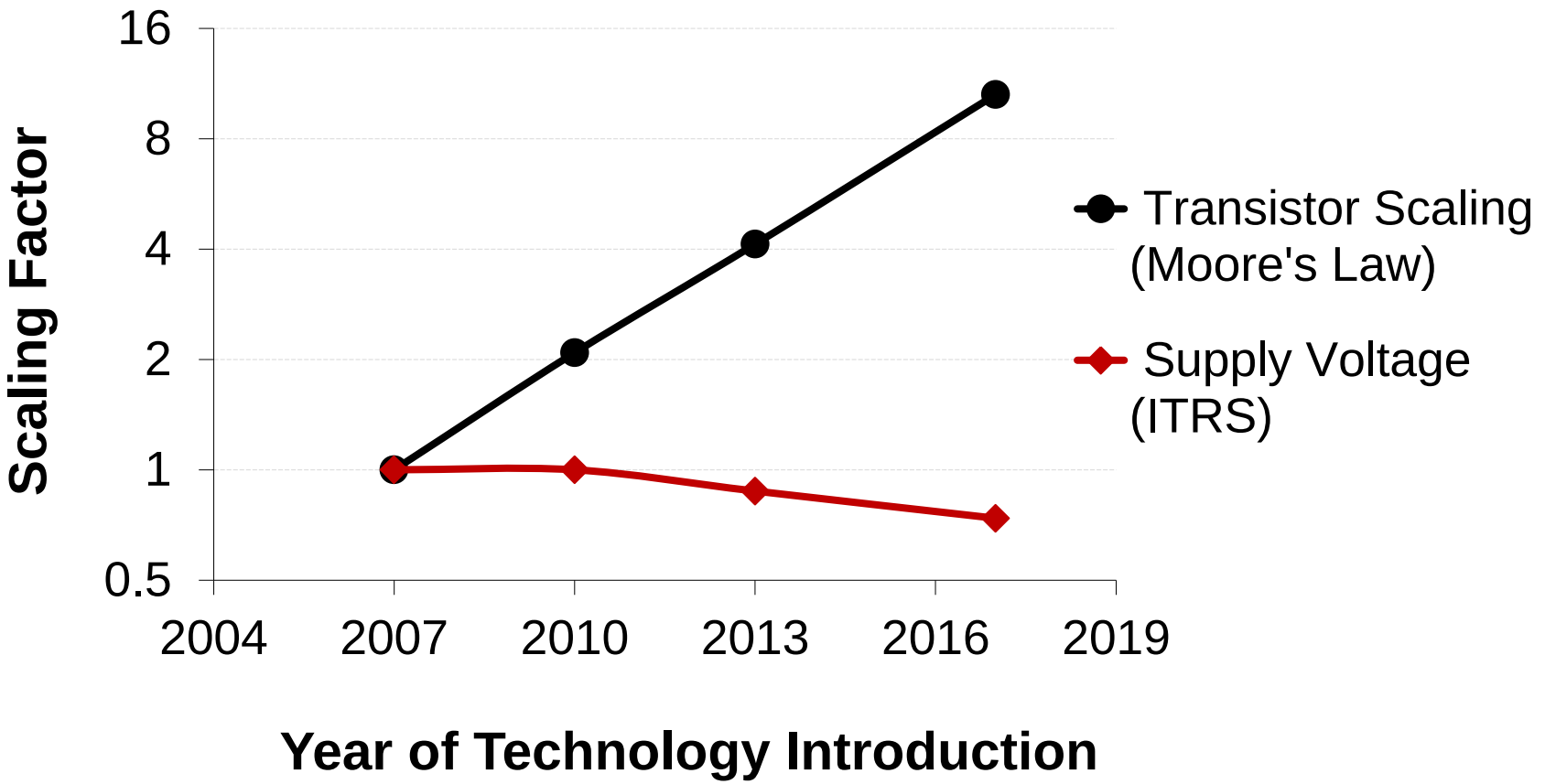
- Application datasets scale faster than Moore's Law! → Big Caches

Pin Bandwidth Scaling



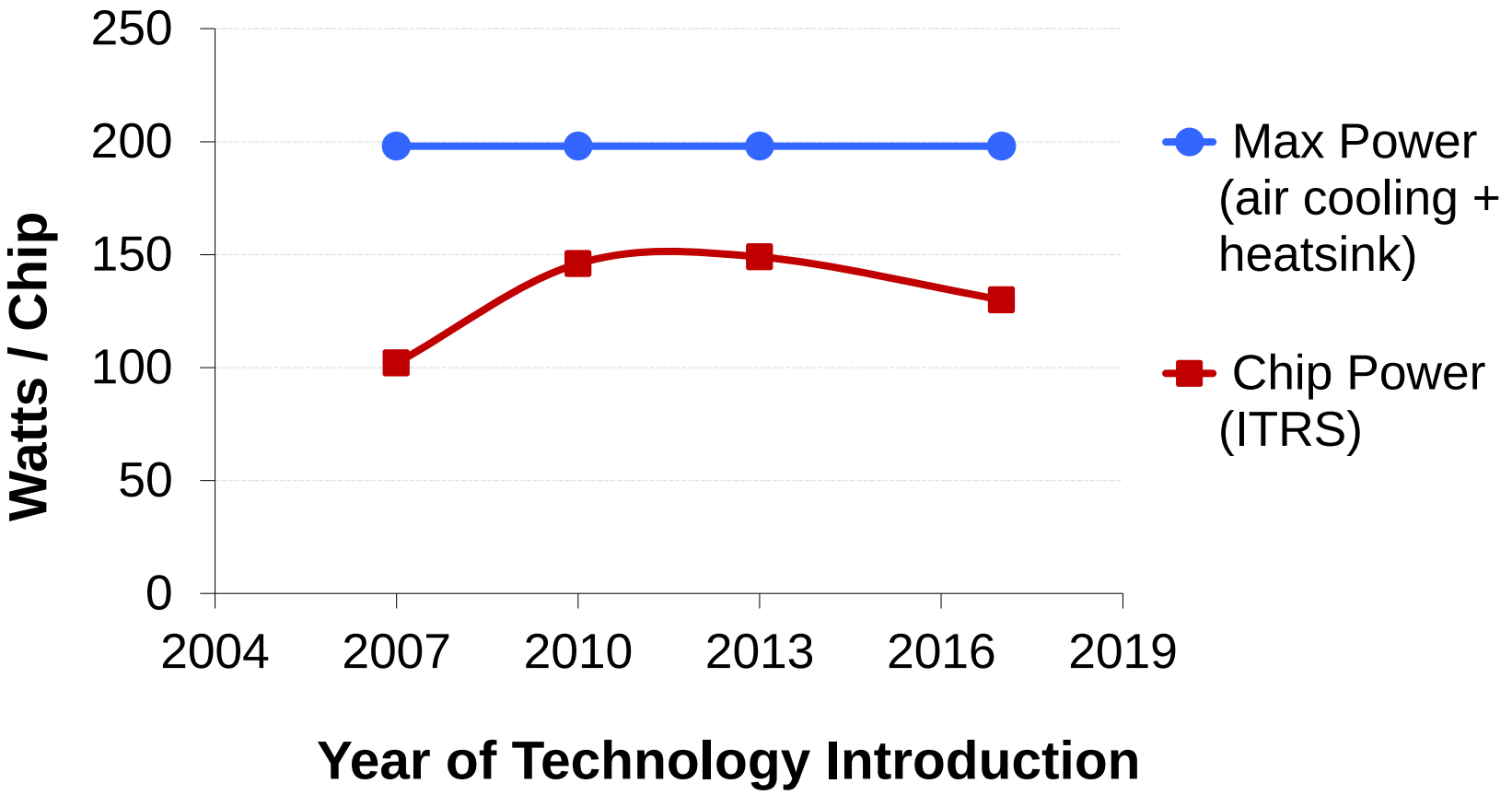
- Off-chip bandwidth scales slowly (#pins, off-chip clock) → Big Caches

Supply Voltage Scaling



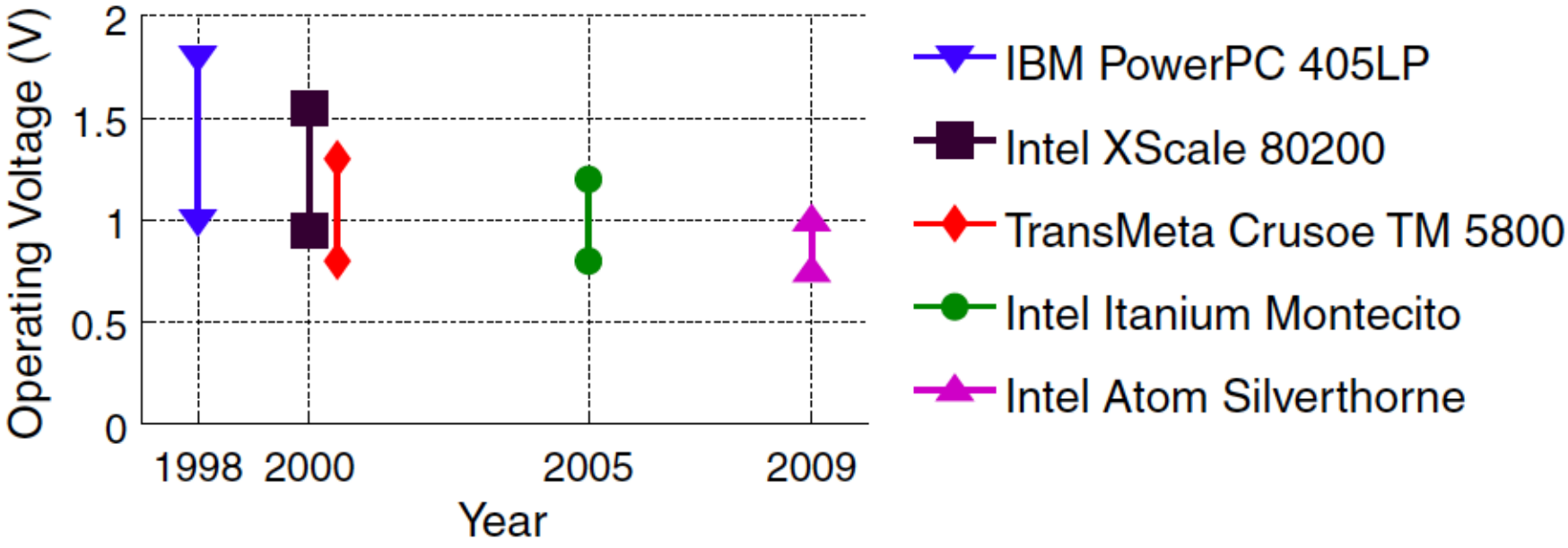
Supply voltage scaling is SLOW! → Dark Silicon

Chip Power Scaling



➧ Chip power does not scale!

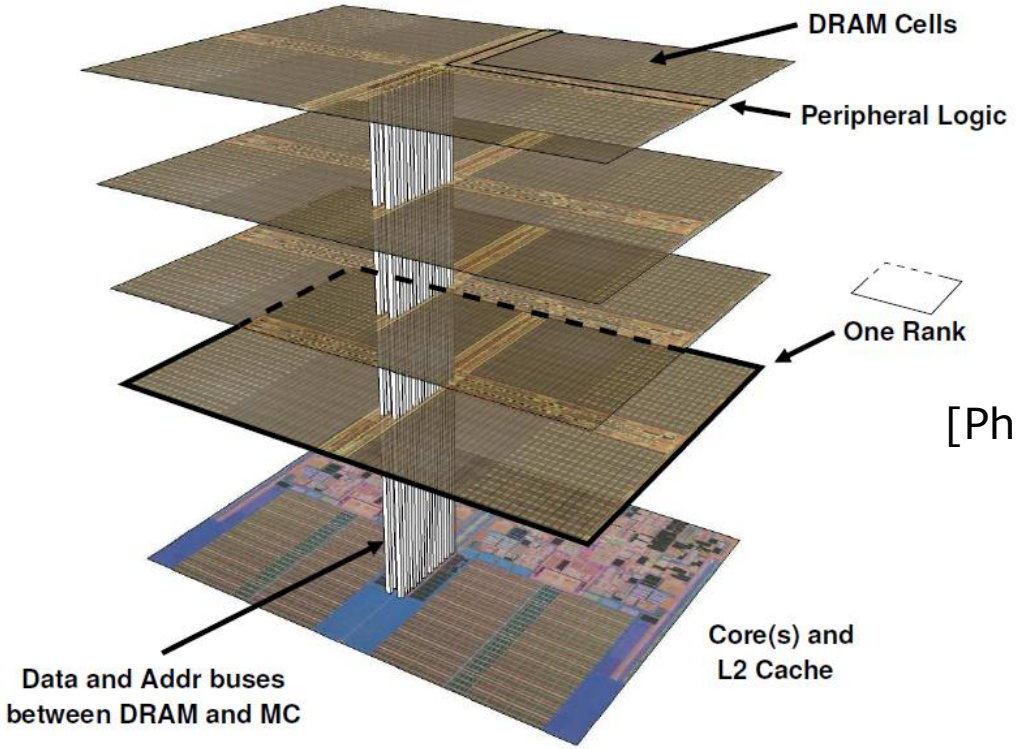
Range of Operational Voltage



[Watanabe et al., ISCA'10]

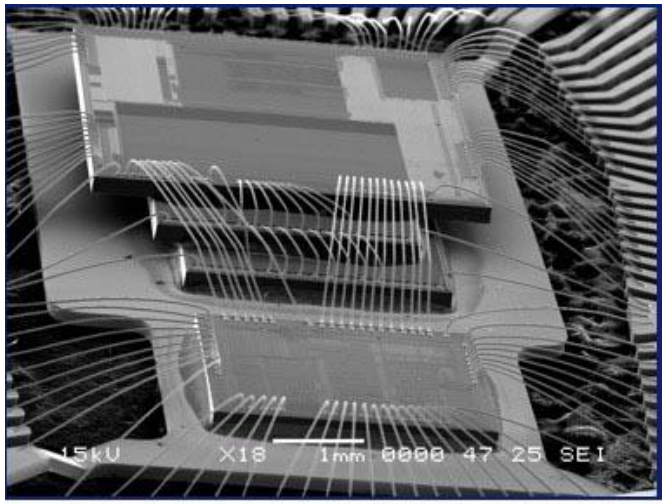
- Shrinking range of operational voltage hampers voltage-freq. scaling

Mitigating Bandwidth Limitations: 3D-stacking

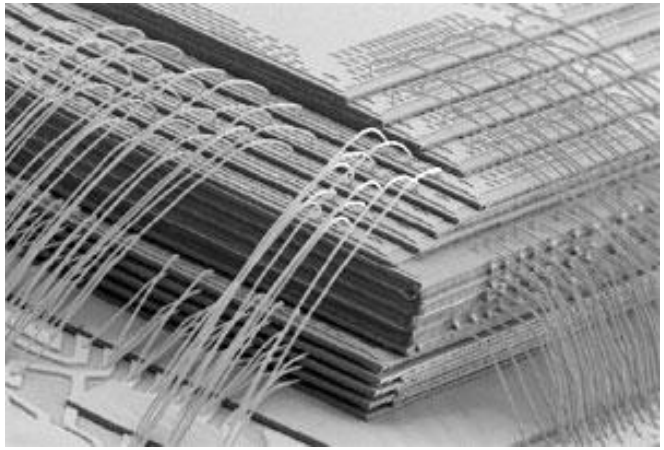


[Loh et al., ISCA'08]

[Philips]

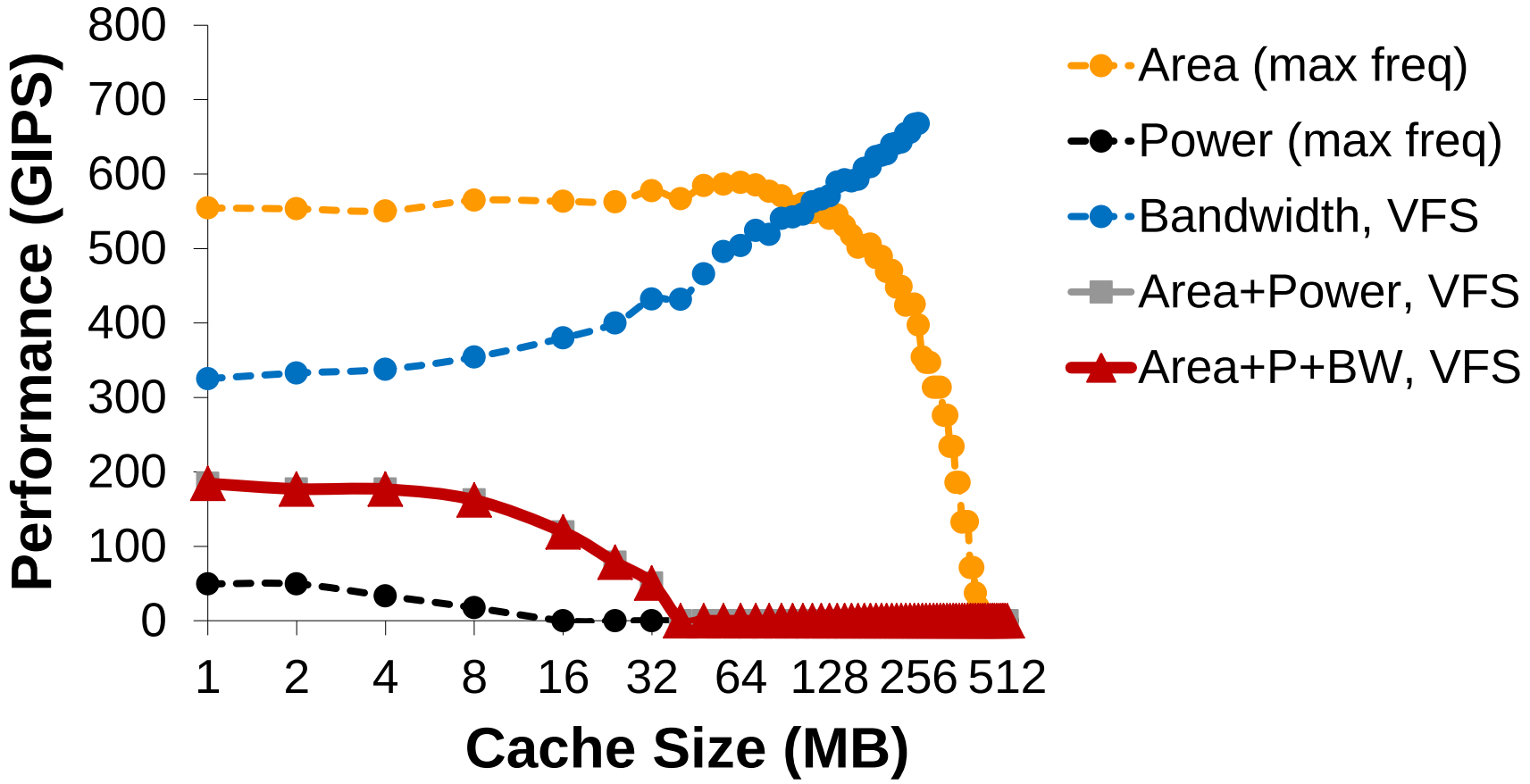


[Amcor Tech]



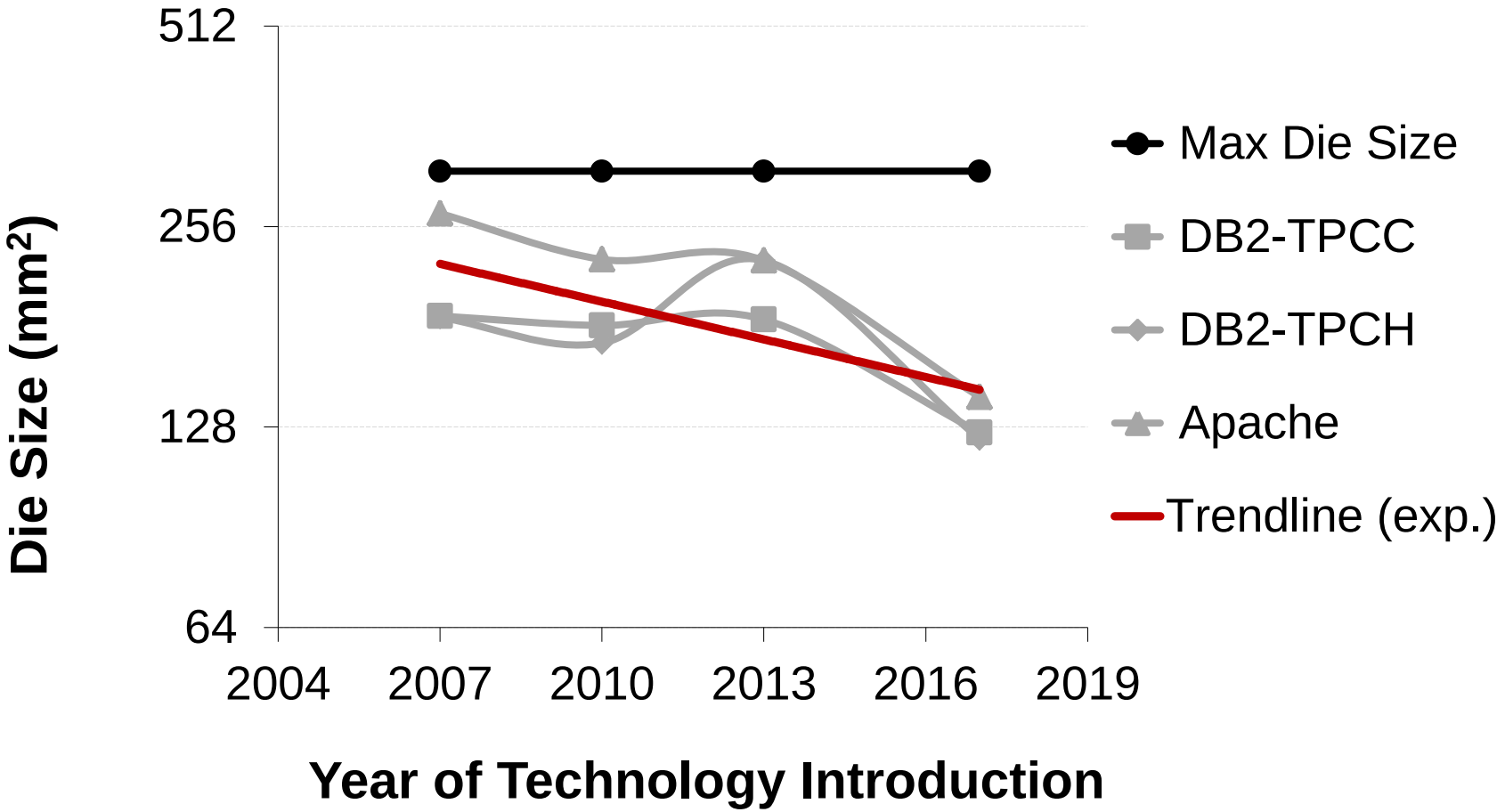
- Delivers TB/sec of bandwidth; use as large "in-package" cache

Performance Analysis of 3D-Stacked Multicores



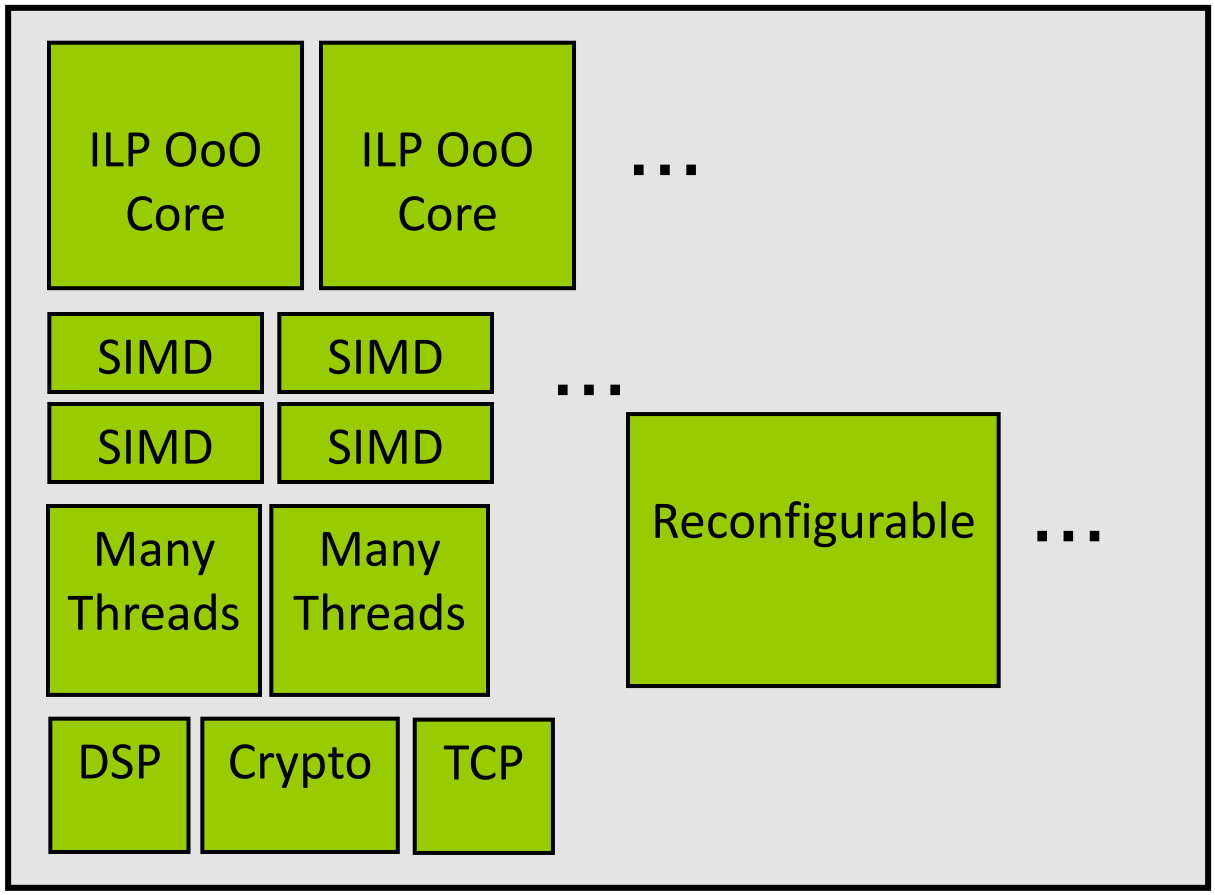
➧ Chip becomes power-constrained

Exponentially Large Die Area Left Unutilized



➧ Dark Silicon!!! Should we waste it?

Example of a Specialized Multicore Chip



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

Core Specialization

- Existing general designs
  - OoO for ILP, in-order MT for memory-latency-bound, SIMD for data-parallel, systolic arrays
- Customizable cores
  - Tensilica Xtensa (custom ISA and datapath, operation fusion)
- 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

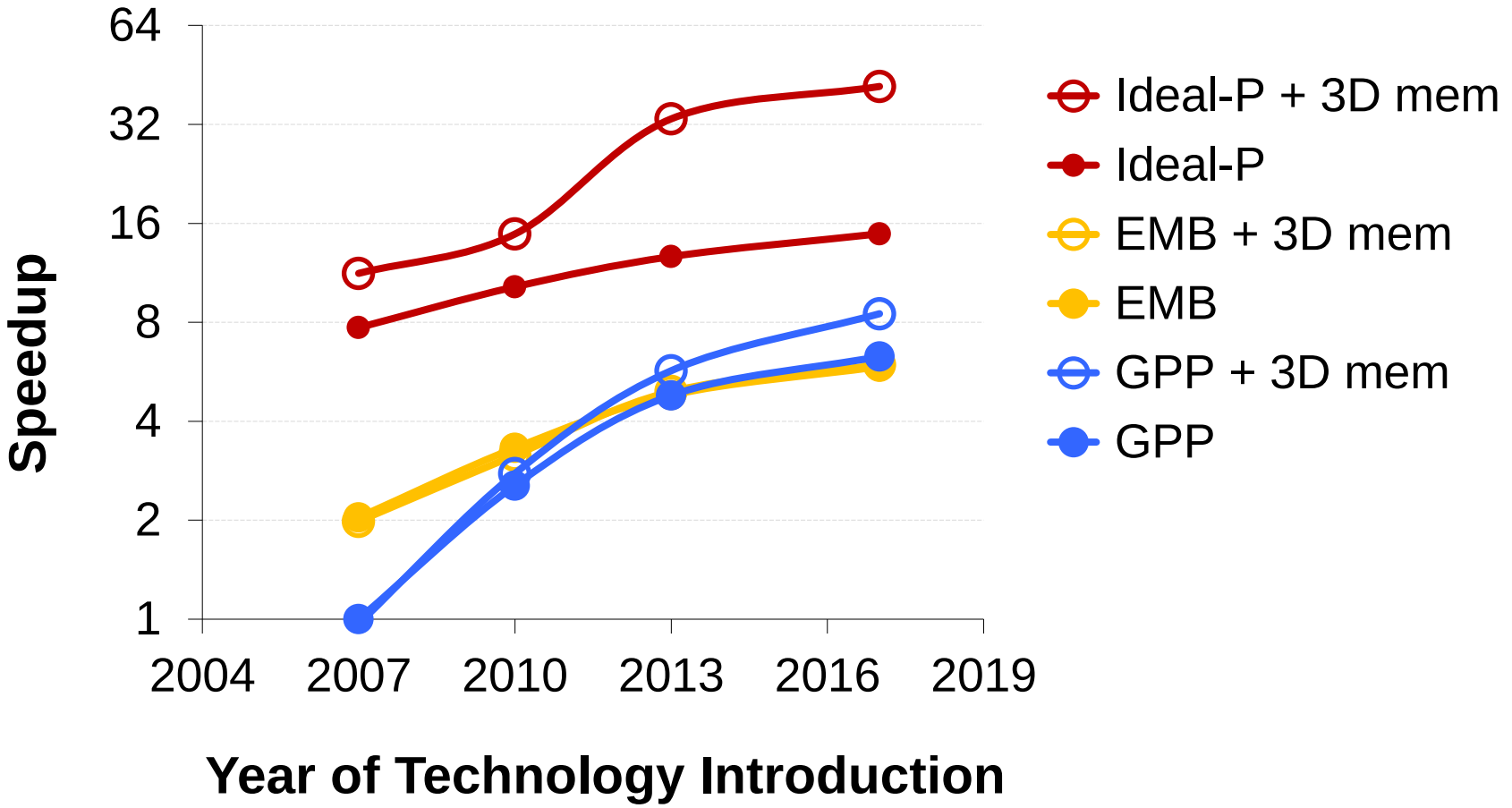
First-Order Core Specialization Model

- 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

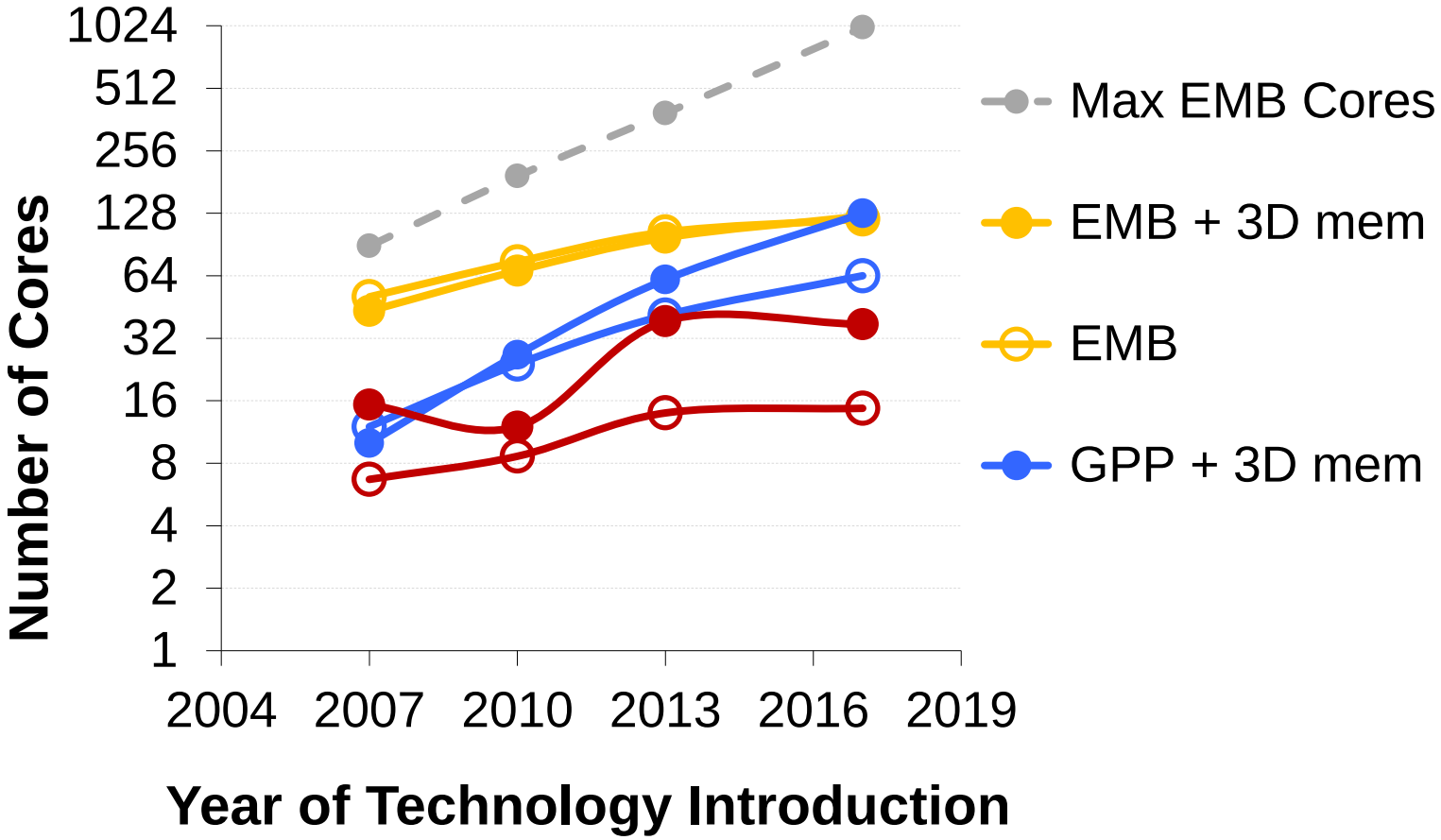
[Hameed et al., ASPLOS'10]

Performance of Specialized Multicores



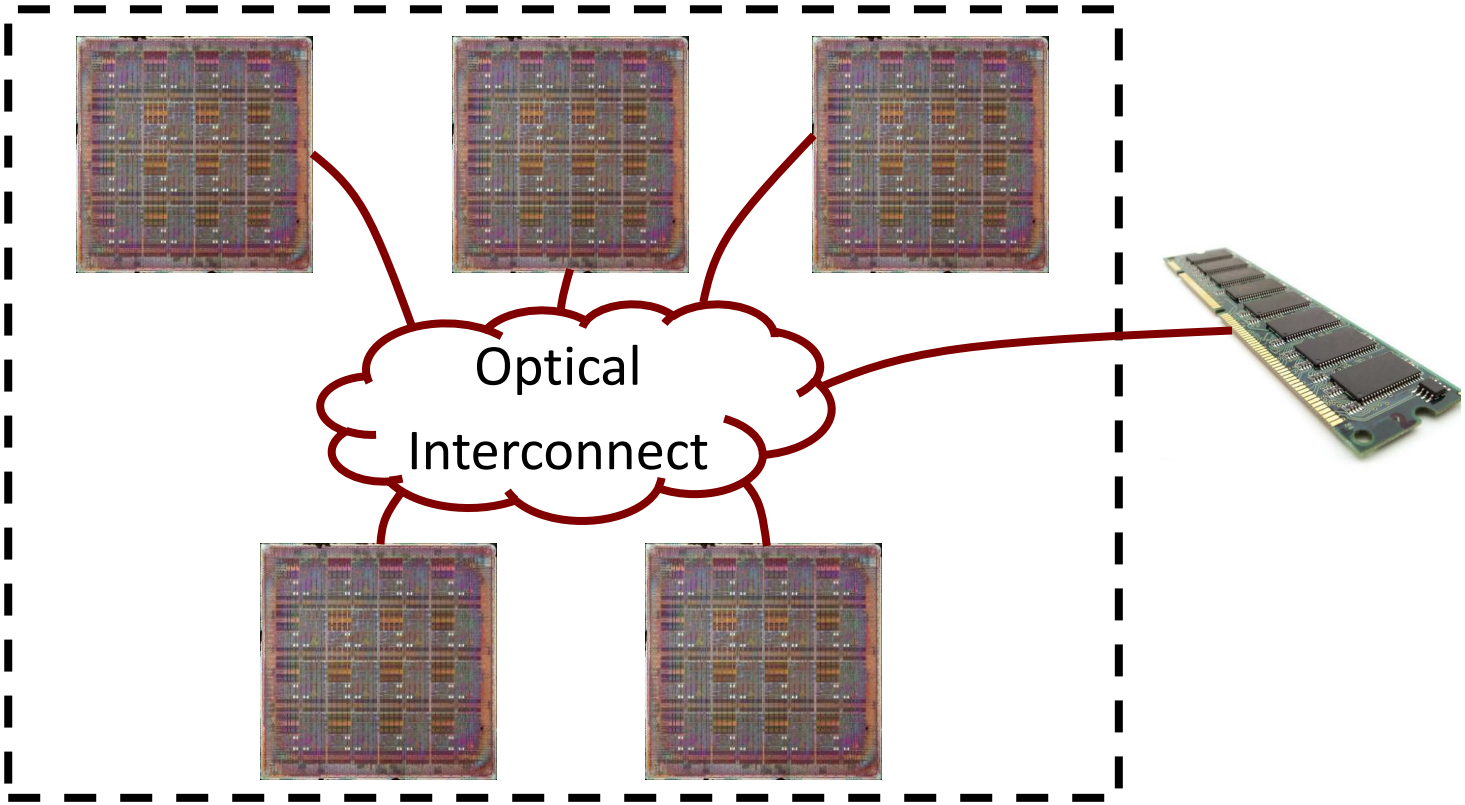
➧ Specialized multicores deliver 2x-12x higher performance

Core Counts for Specialized Multicores



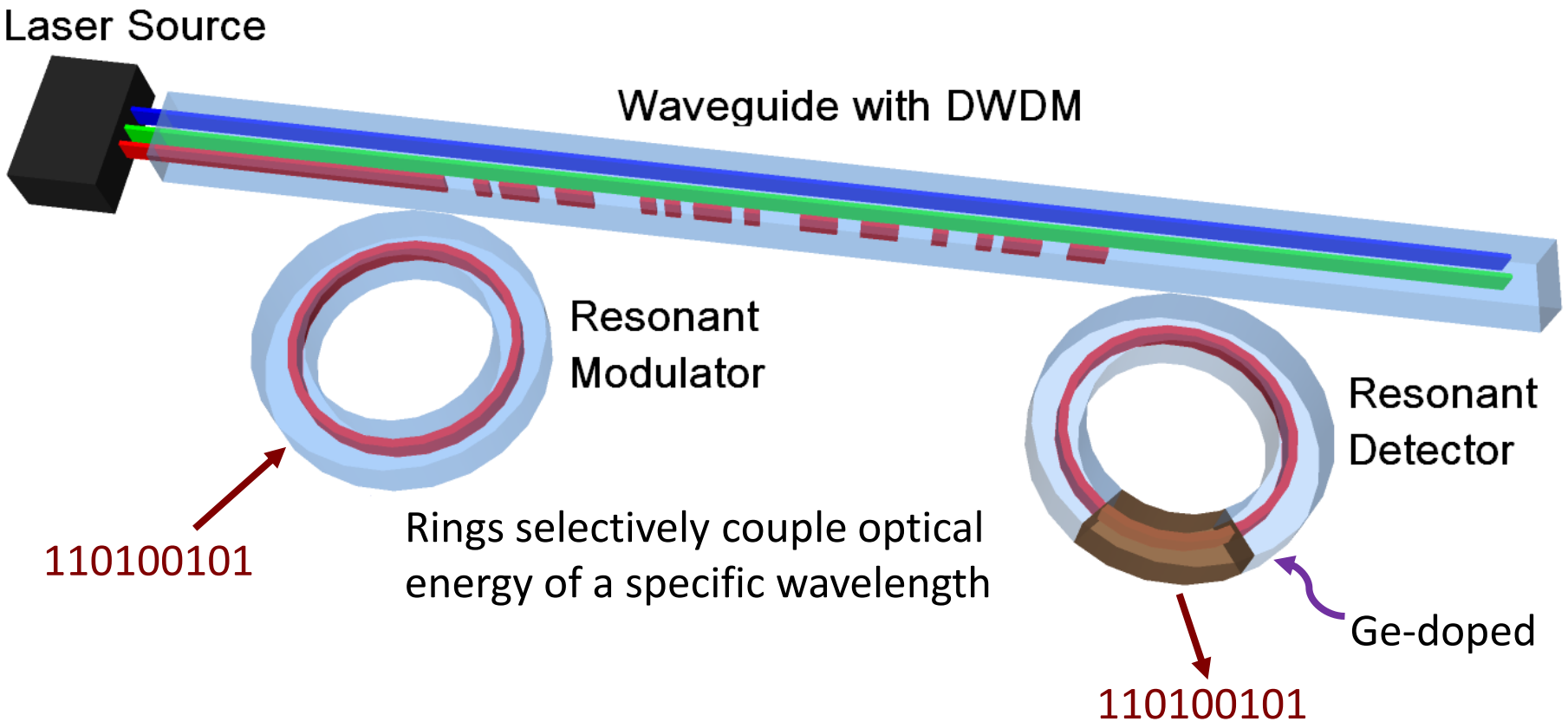
- Only few cores need to run at a time; large die area allow many cores
  - Power constraints? Yield?

Taming Power and Bandwidth : Nanophotonics



- Split chip into chiplets, spread in space
- Ease cooling and power delivery, high yield; photonics for bandwidth

Nanophotonic Components



- 64 wavelengths DWDM, 3 ~ 5µm waveguide pitch, 10Gbps per link
  - ~100 Gbps/µm bandwidth density !!! [Batten et al., HOTI'08]

Technology: Off-chip Channel Material

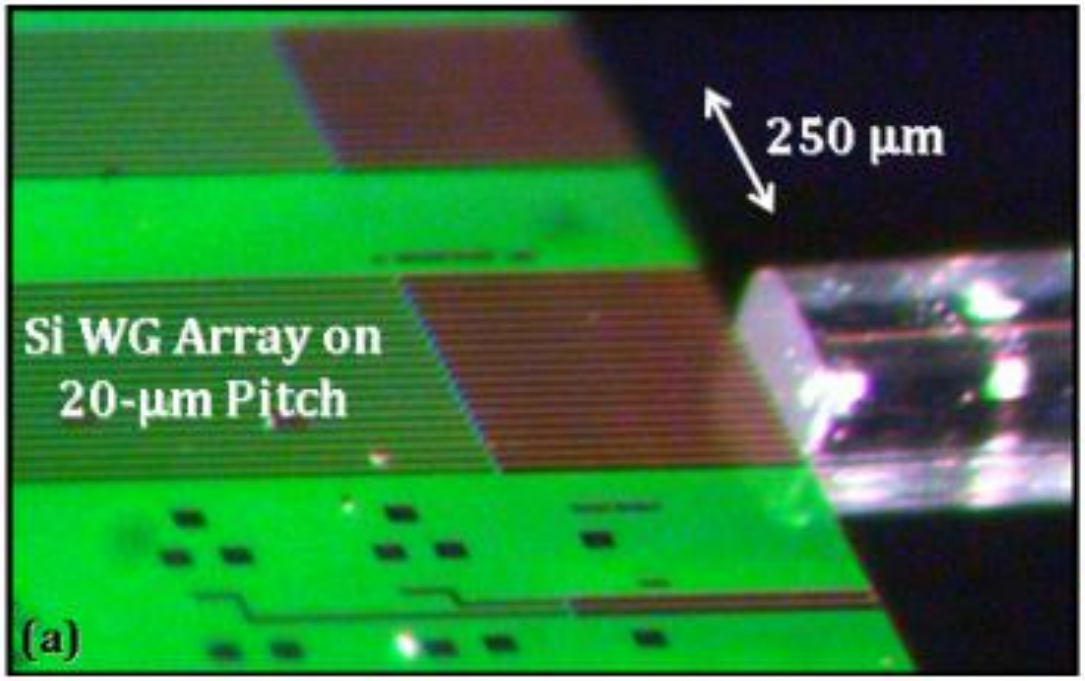
Material	Optical Loss	Propagation Speed	Pitch (density)
Silicon Waveguide	0.3 dB/cm*	0.286c	20um
Optic Fiber	0.2 dB/km	0.676c	250um

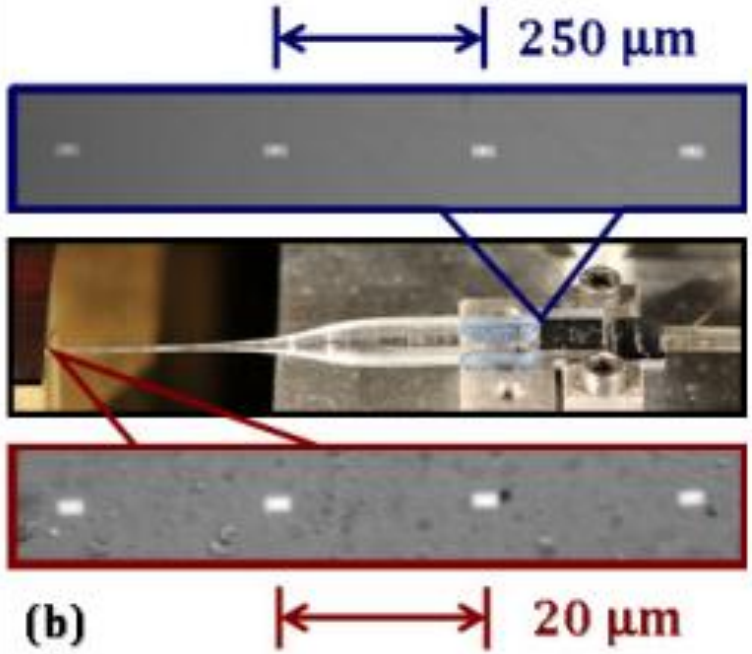
- Optical fiber is low-loss, high speed
  - Enables further spreading out chiplets.
  - *BW density was a challenge (fiber pitch size is large)*

* J. Cardenas et al., Optics Express 2009

➡ Fiber: low optical loss, high speed, flexibility eases assembly

Technology: Dense Off-Chip Coupling

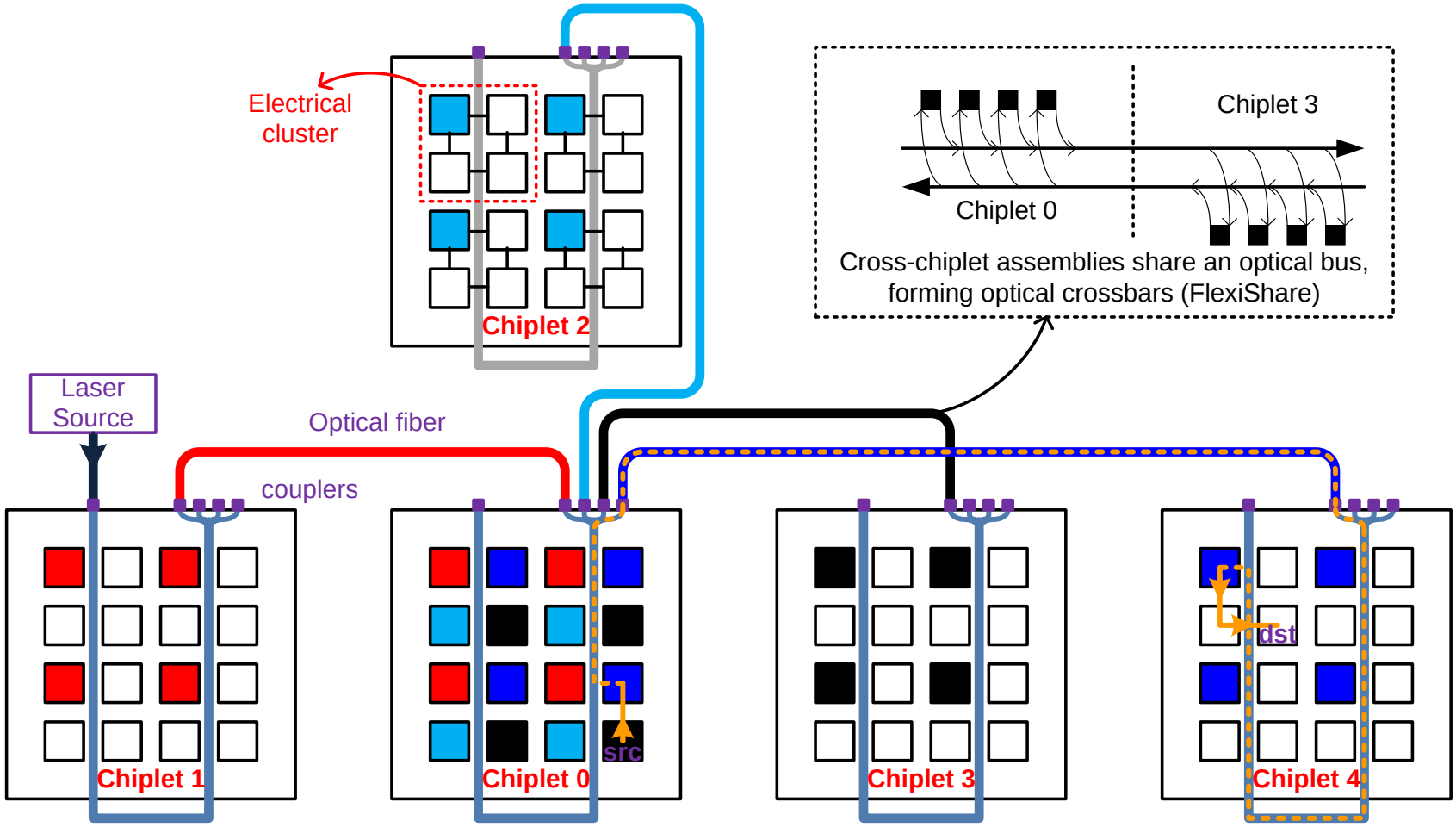




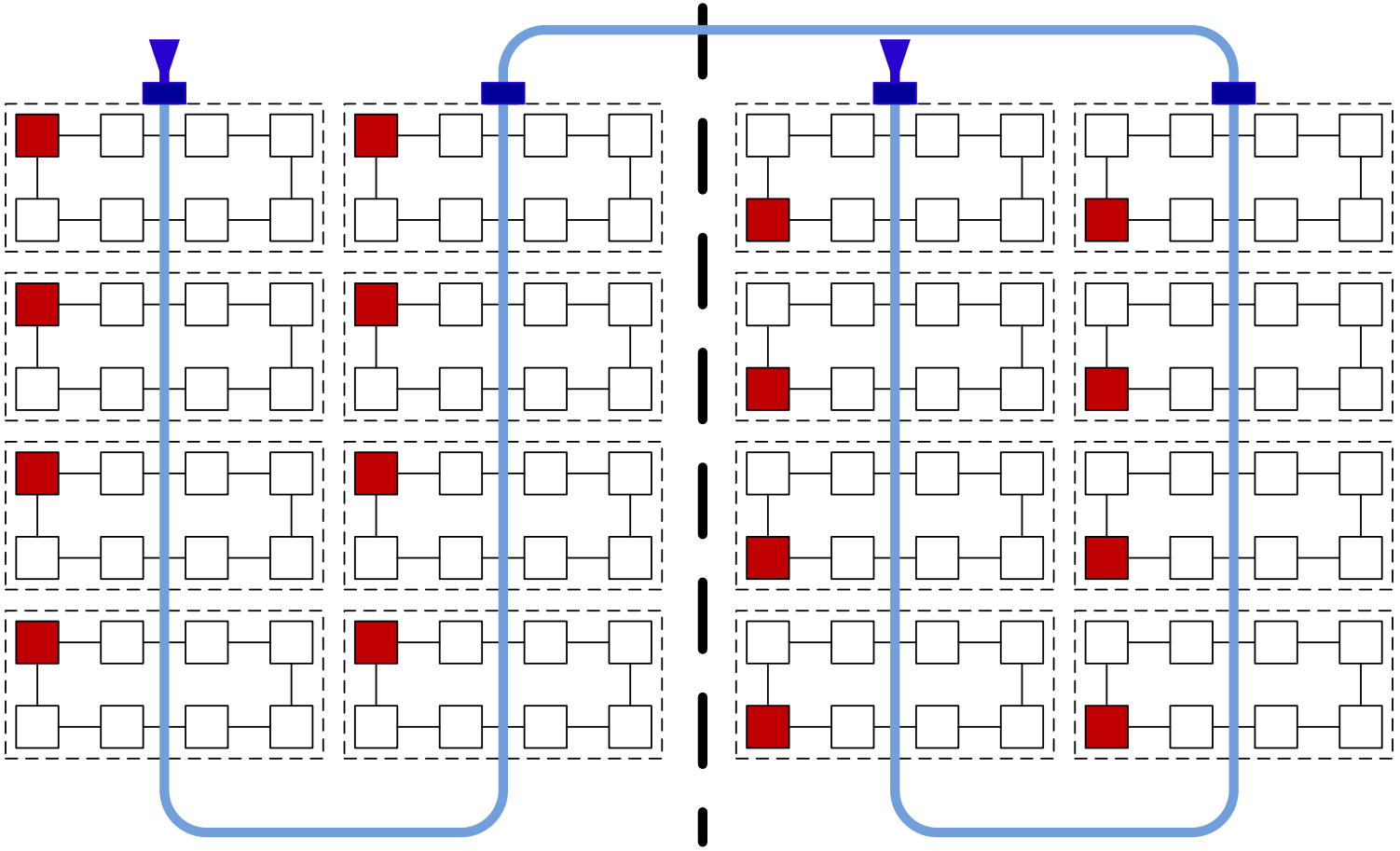
- Dense optical fiber array. [Lee et al., OSA / OFC/NFOEC 2010]
- <1dB loss, 8 Tbps/mm demonstrated.

➡ Tapered couplers solved bandwidth problem, demonstrated Tbps/mm

Galaxy Overall Architecture



Large-Scale Interconnects



- 200mm^2 die, 64 routers per chiplet, 9 chiplets, 16cm fiber
- Supports > 1K cores!

Conclusions

- Physical constraints and software pragmatics limit core counts
  - …and performance
- Emerging/exotic technologies may solve some problems
  - 3D-memory for bandwidth
  - Nanophotonics for bandwidth, power, yield
- Need to reduce wasted energy per unit of work
  - Heterogeneity, only power the few cores needed
- Need to innovate across software/hardware stack
  - Programmability, tools are a great challenge
- Scaling forces caches to grow exponentially
  - Address data management both at cache and software

Thank You!

Acknowledgements:

Y. Pan, J. Kim, G. Memik, M. Ferdman, B. Falsafi