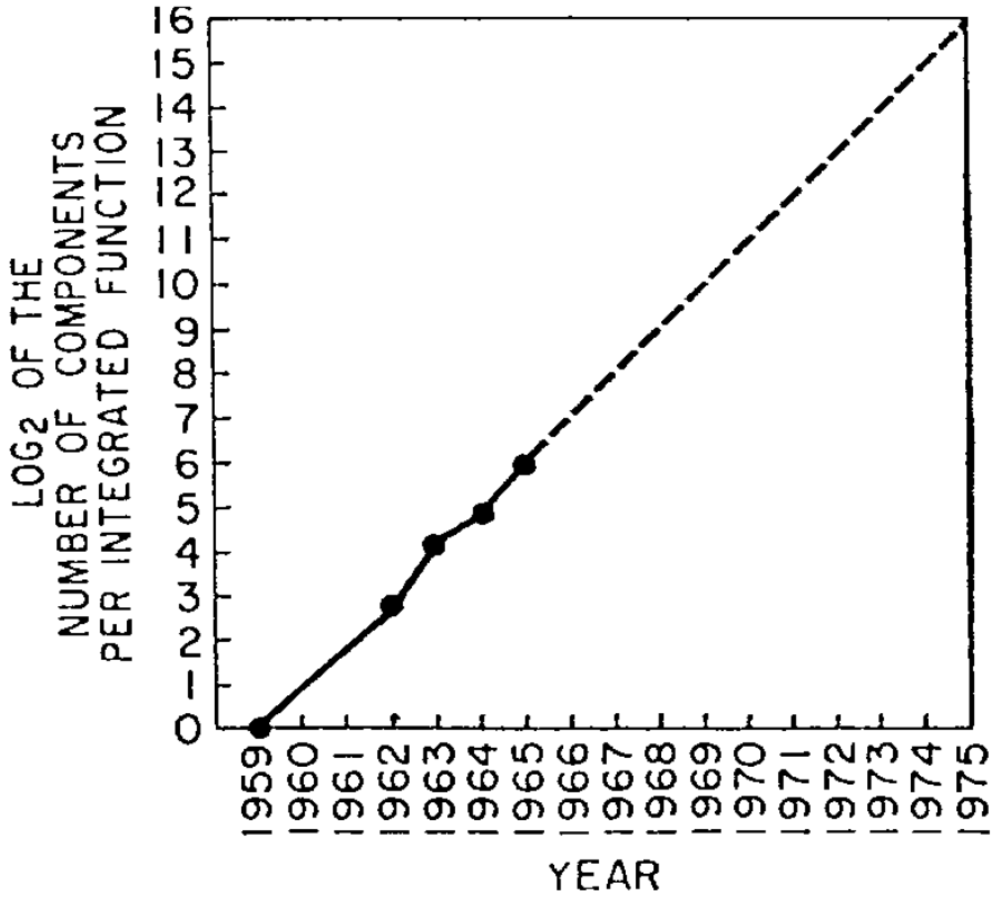
18-742:
Computer Architecture & Systems

Dark Silicon
and the End of Multicore Scaling

Prof. Phillip Gibbons

Spring 2025, Lecture 2

Recall: Moore's Law



Number of components [transistors] per integrated function [integrated circuit] will double every year (for at least ten years).

Moore revised in 1975 to doubling every two years.

Dennard Scaling (1974)

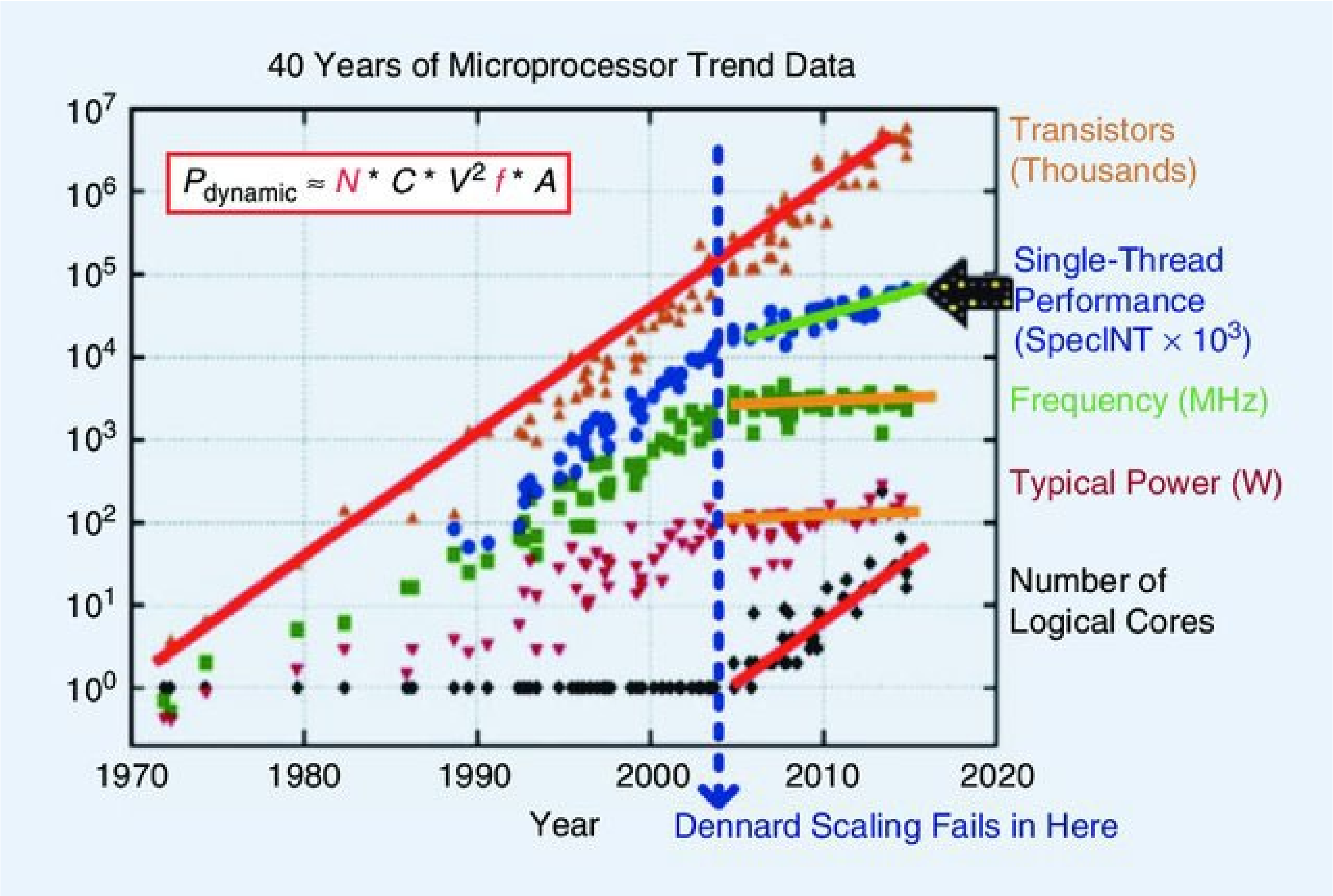
As transistors get smaller, their power consumption per unit area remains the same for every technology generation.

Alternatively, with every generation the number of transistors in a chip can be doubled with **no change in power consumption**.

Breakdown of Dennard Scaling around 2006
due to current leakage at small sizes.

Moore's Law w/o Dennard Scaling

- 2X transistors every 2 years, but frequency & power capped



- What to spend transistors on? Multiple cores

N = transistors per unit area

"Scaling the Bandwidth Wall: Challenges in and Avenues for CMP Scaling"

Brian Rogers, Anil Krishna, Gordon Bell, Ken Vu, Xiaowei Jiang, Yan Solihin 2009

- **Memory Bandwidth Wall:**
  - Compute improvements >> Memory BW improvements
  - On multicores: cores share off-chip BW
  - Each core's BW share declines with each generation
- **This limits effective CMP scaling**
  - E.g., from 8 to 24 in four generations, not 128
  - Most of chip is needed for caches

Chip Area for Cores vs. Caches

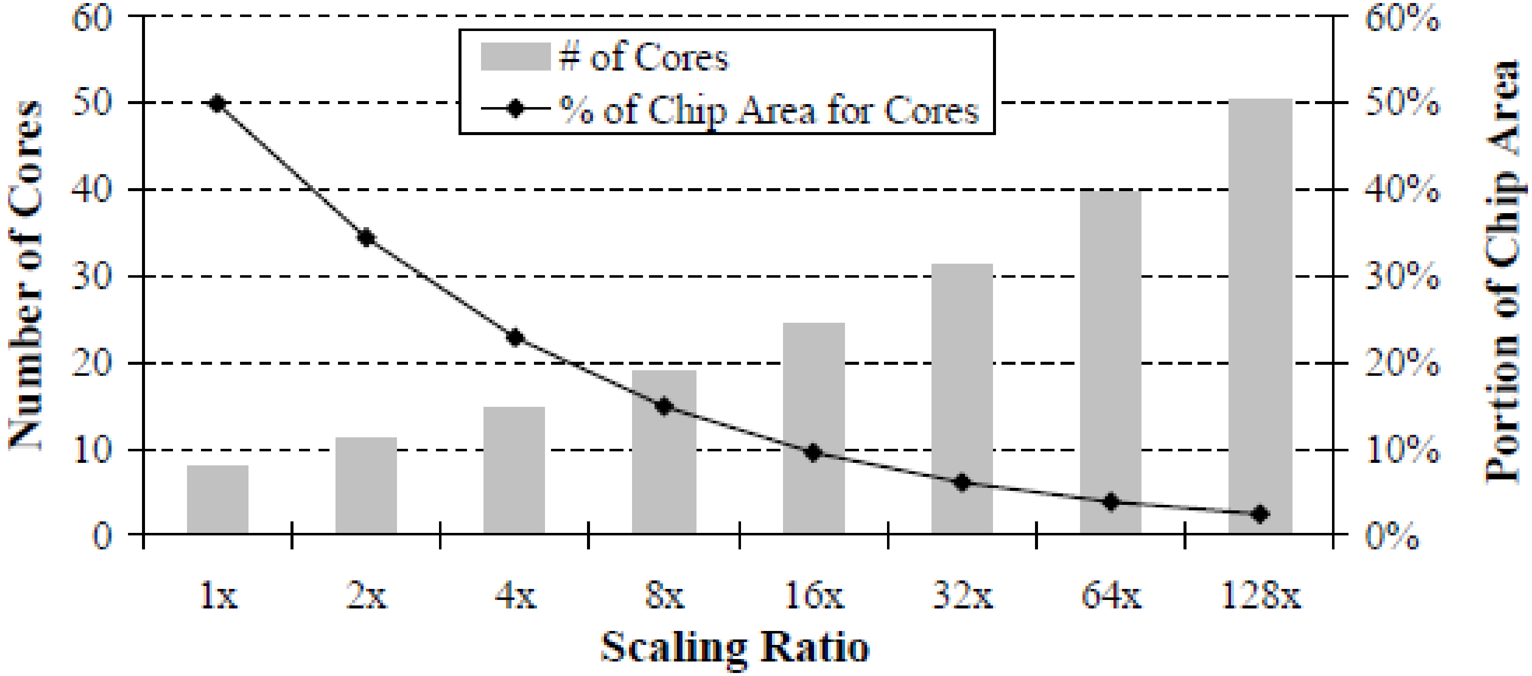
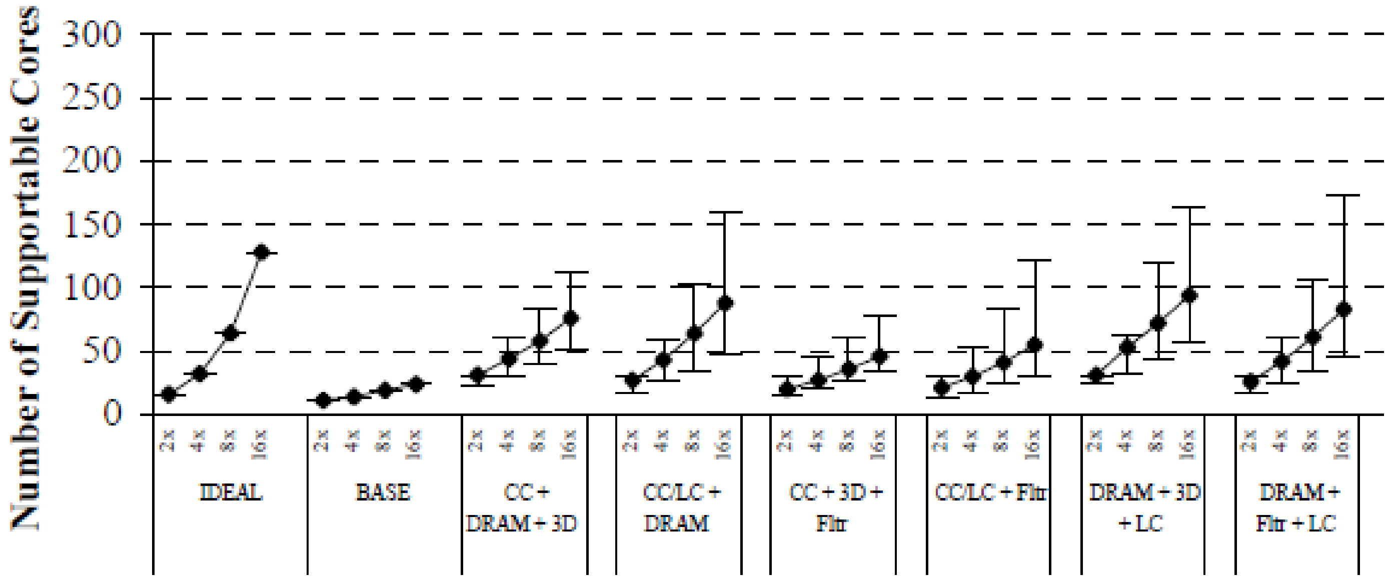


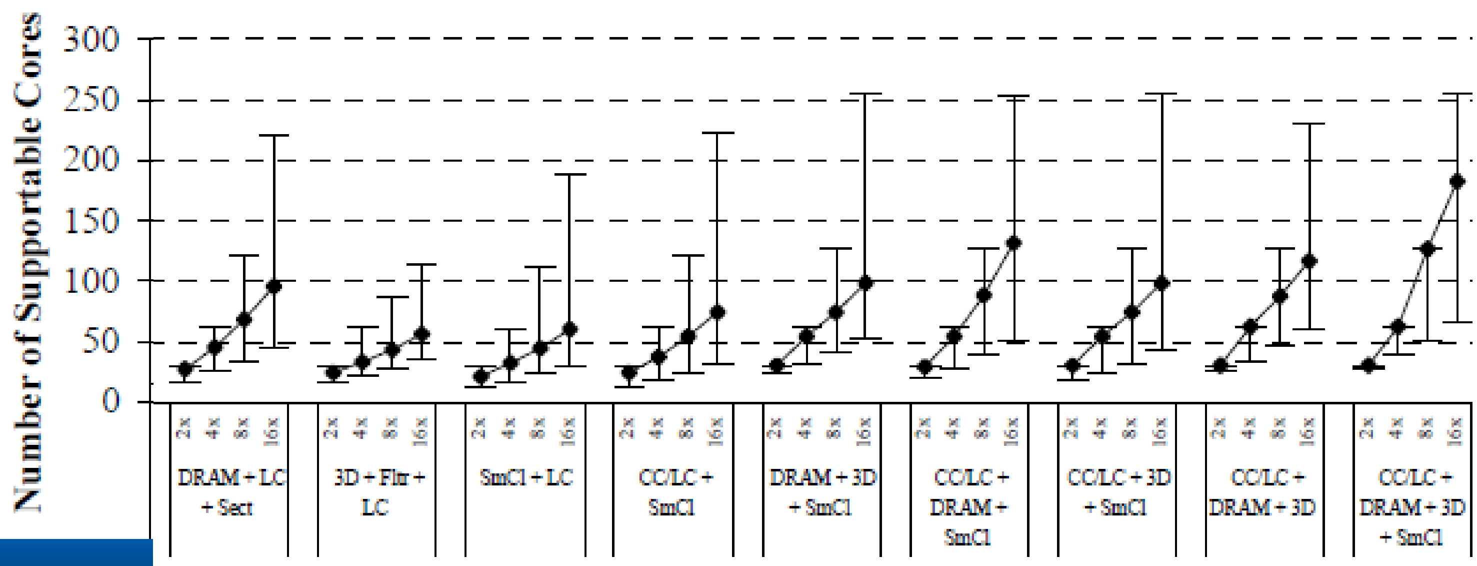
Figure 3: Die area allocation for cores and the number of supportable cores assuming constant memory traffic requirements.

BW Conserving Techniques

- **DRAM Caches [DRAM]**
  - DRAM is 8x-16x denser than SRAM
- **Smaller cores freeing up space for cache [SmCo]**
- **Link compression [LC] >> Cache compression [CC]**
- **3D-stacked caches [3D]**
- **Ideally-sized cache lines [SmCl]**

All techniques combined enables 183 cores





"Dark Silicon and the End of Multicore Scaling"

Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, Doug Berger 2011

- **Hadi:** Washington PhD, now UCSD prof
  - Young Architect Award 2018
- **Emily:** Wisconsin PhD, now Google
- **Renee:** UT Austin PhD, ex-Arm
  - Fellow World Economic Forum
- **Karu:** Wisconsin prof & NVIDIA
  - Young Architect Award 2012
- **Doug:** Microsoft Fellow
  - IEEE & ACM Fellow

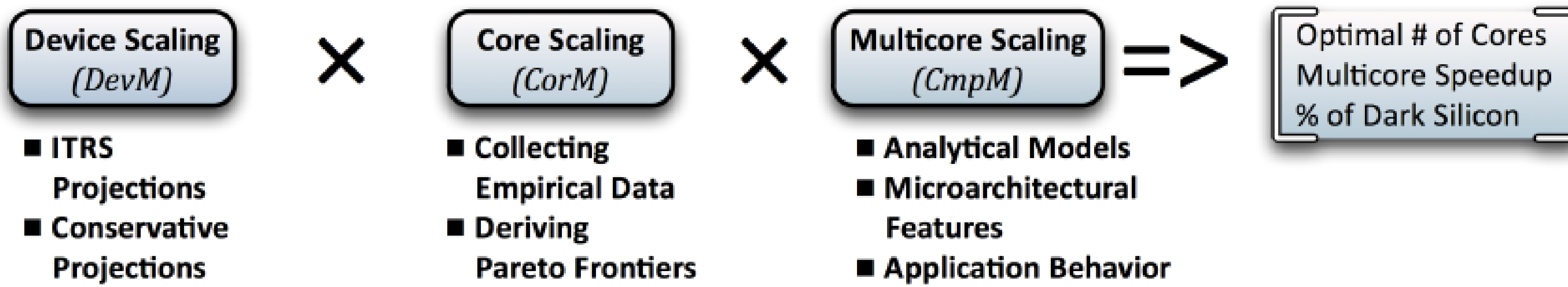
Paper’s Main Question

- **How effective would multicore designs be in a power-limited era, for differing degrees of parallelism?**

- **Gives upper bound on achievable speedup, accounting for**
  - Transistor scaling trends
  - Processor core design options
  - Chip multiprocessor organizations
  - Benchmark characteristics

All under area & power constraints

Analytical Model



Analytical Model

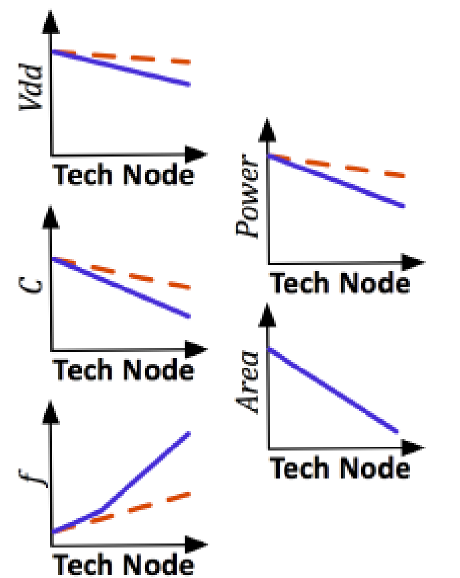
Device Scaling *(DevM)* × **Core Scaling** *(CorM)* × **Multicore Scaling** *(CmpM)* => Optimal # of Cores / Multicore Speedup / % of Dark Silicon

Device Scaling (DevM)
- ITRS Projections
- Conservative Projections

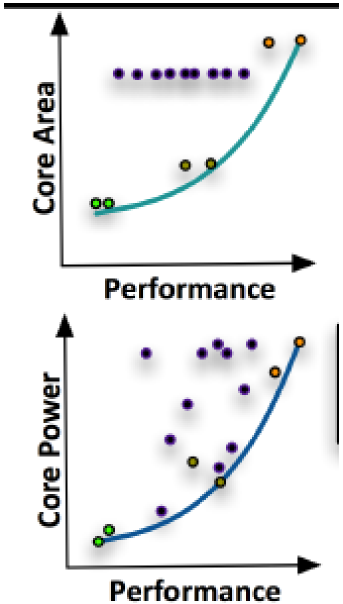
Core Scaling (CorM)
- Collecting Empirical Data
- Deriving Pareto Frontiers

Multicore Scaling (CmpM)
- Analytical Models
- Microarchitectural Features
- Application Behavior

Optimal # of Cores
Multicore Speedup
% of Dark Silicon

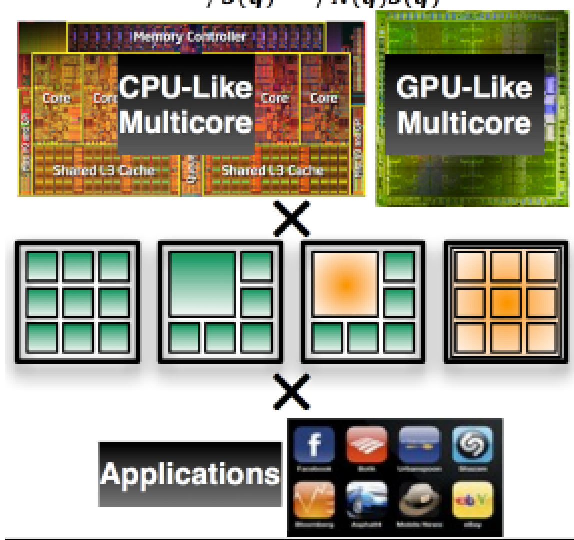


- 2 Projection Schemes

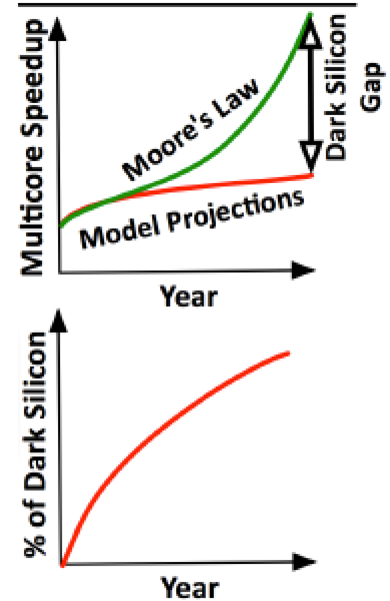


- Data for 152 Processors

$$\frac{1}{\frac{1-f}{S(q)} + \frac{f}{N(q)S(q)}}$$



- 2 Chip Organizations × 4 Topologies
- 12 Benchmarks



- Search 800 Configs for 12 Benchmarks

Discussion: Summary Question #1

What Did the Paper Get Right?

State the 3 most important things the paper says.

These could be some combination of the motivations, observations, interesting parts of the design, or clever parts of the implementation.

Analytical Model

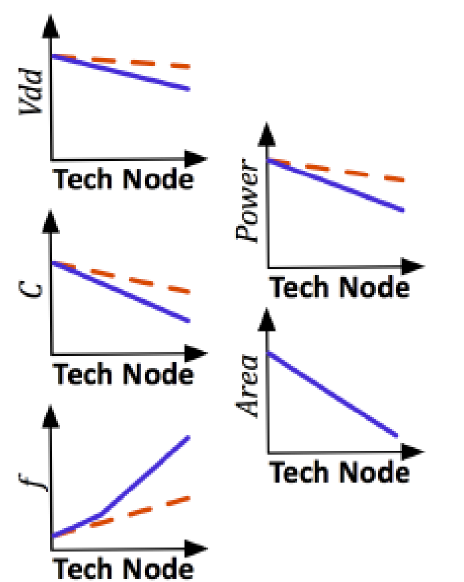
Device Scaling *(DevM)* × **Core Scaling** *(CorM)* × **Multicore Scaling** *(CmpM)* => **Optimal # of Cores, Multicore Speedup, % of Dark Silicon**

Device Scaling (DevM)
- ITRS Projections
- Conservative Projections

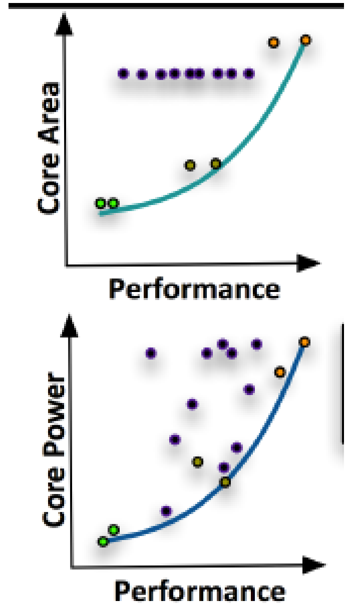
Core Scaling (CorM)
- Collecting Empirical Data
- Deriving Pareto Frontiers

Multicore Scaling (CmpM)
- Analytical Models
- Microarchitectural Features
- Application Behavior

Optimal # of Cores
Multicore Speedup
% of Dark Silicon

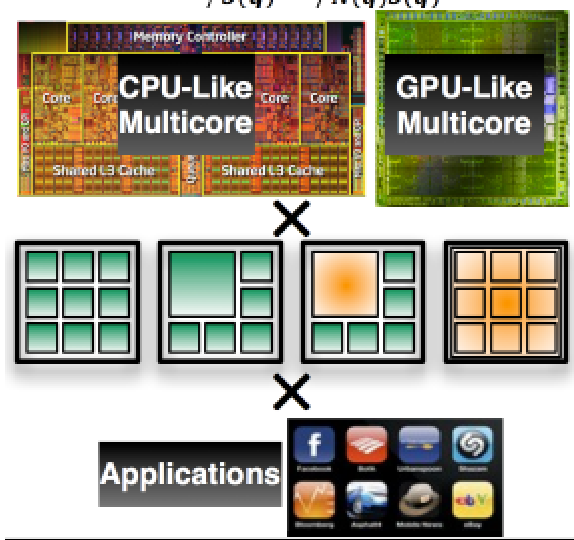


- 2 Projection Schemes

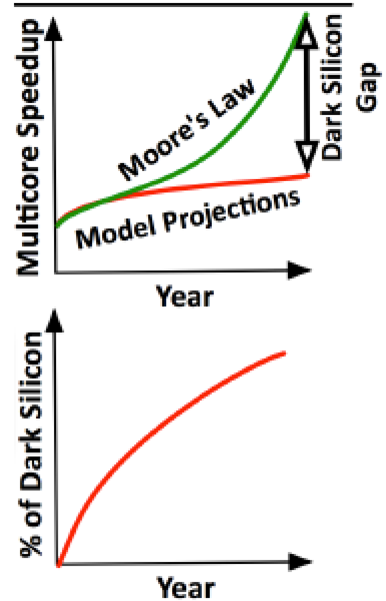


- Data for 152 Processors

$$\frac{1}{\frac{1-f}{S(q)} + \frac{f}{N(q)S(q)}}$$



- 2 Chip Organizations × 4 Topologies
- 12 Benchmarks



- Search 800 Configs for 12 Benchmarks

Take-Aways

- **2008 to 2024 would see only a 7.9x benchmark speedup**
  - 24-fold gap from 2x every generation

- **Multicore scaling would NOT be the principal vector of performance scaling**

- **Dark Silicon: Substantial fraction of the chip will be powered off**

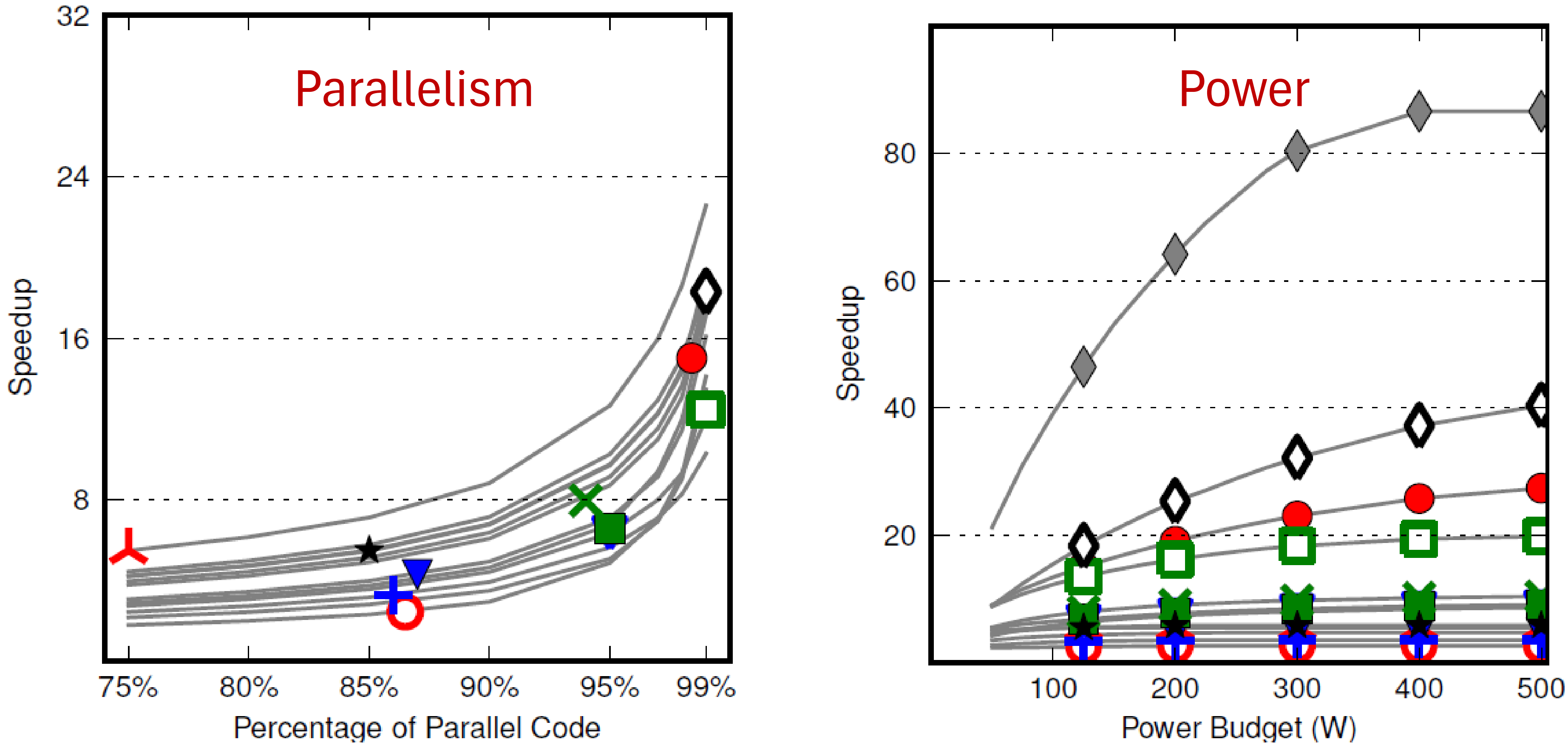
Results for PARSEC

Characteristic	Conservative CPU	Conservative GPU	ITRS CPU	ITRS GPU
Symmetric GM Speedup	3.4×	2.4×	7.7×	2.7×
Dynamic GM Speedup	3.5×	2.4×	7.9×	2.7×
Maximum Speedup	10.9×	10.1×	46.6×	11.2×
Typical # of Cores	< 64	< 256	< 64	< 256
Dark Silicon Dominates	2016	2012	2021	2015

Different Workload: Ray tracing
(on asymmetric topology, at 8 nm)

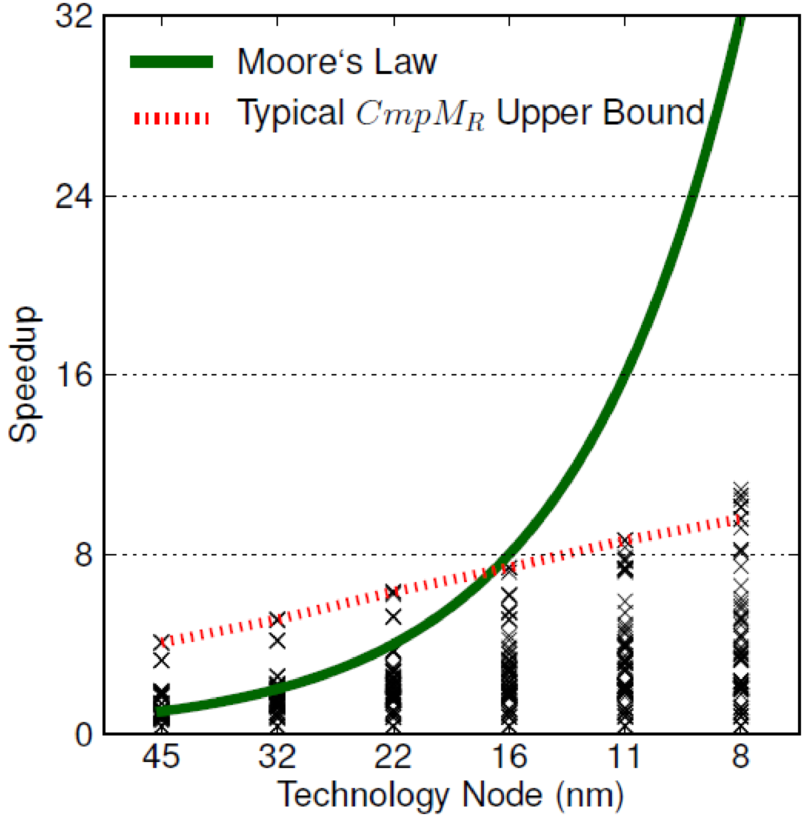
Optimal core count = 4864
8% Dark silicon

Two Sources of Dark Silicon

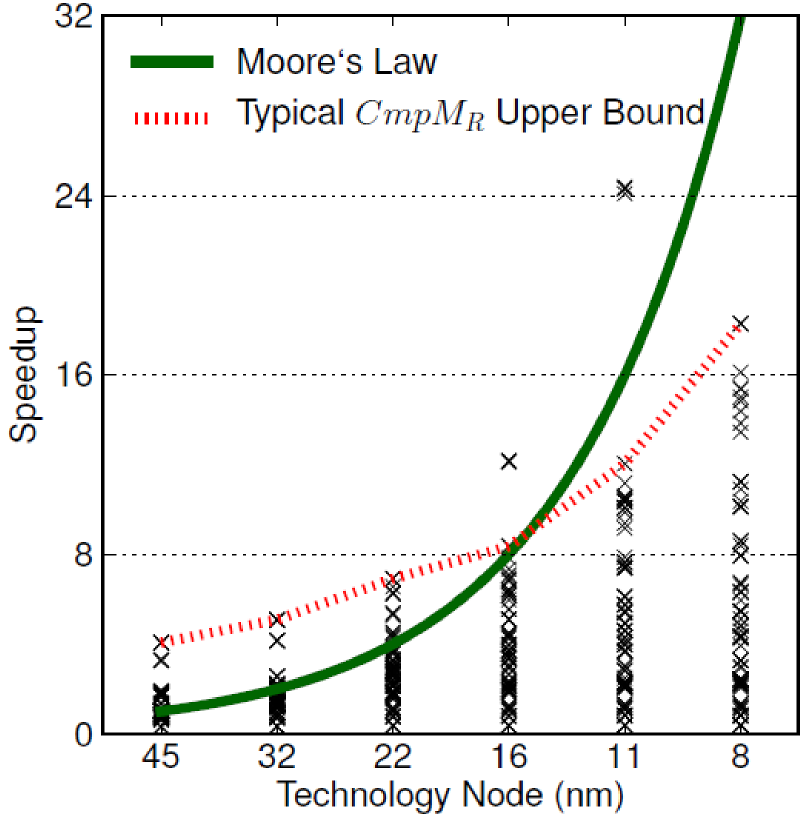


CPU, dynamic topology, 8nm, ITRS

Summary: All Organizations/Topologies



(a) Conservative Scaling



(b) ITRS Scaling

Predictions

- Multicore era will likely end in 2014

- Moore's Law may end by 2016,
  creating massive disruptions in our industry!

- Silver lining:

"The onus will be on computer architects—and computer architects only—to deliver performance and efficiency gains that can work across a wide range of problems"

Discussion: Summary Question #2

What Did the Paper Get Wrong?

Describe the paper's single most glaring deficiency.

Every paper has some fault. Perhaps an experiment was poorly designed or the main idea had a narrow scope or applicability.

Authors' Retrospective (2023)

- Do we have multicores with hundreds of cores today?
- Do we have large swaths of dark silicon?
- Proliferation of specialized on-chip accelerators
- Deep Learning changed everything
- Approximate computing becomes viable (e.g., reduced precision)
- Future: Deep Learning & generative inference as a second class of general-purpose computation

“Clearing the Clouds: A Study of Emerging Scale-out Workloads on Modern Hardware”

Michael Ferdman, Almutaz Adileh, Onur Kocberber, Stavros Volos, Mohammad Alisafaee, Djordje Jevdjic, Cansu Kaynak, Adrian Daniel Popescu, Anastasia Ailamaki, Babak Falsafi 2012

Scale-out workloads are ill-suited to Multicores (2012)

- Suffer from high instruction-cache miss rates
- Aggressive out-of-order cores are overkill
- On-chip caches are way too small
- On-chip & off-chip BW requirements are lower

To Read for Friday

“The Case for a Single-Chip Multiprocessor”

Kunle Olukotun, Basem Nayfeh, Lance Hammond,
Ken Wilson, Kunyung Chang 1996

Optional Further Reading:

“Simultaneous Multithreading: Maximizing On-chip Parallelism”

Dean Tullsen, Susan Eggers, Henry Levy 1995

“Single-chip Heterogeneous Computing: Does the Future Include Custom Logic, FPGAs, and GPGPUs?”

Eric Chung, Peter Milder, James Hoe, Ken Mai 2010

BACKUP SLIDES

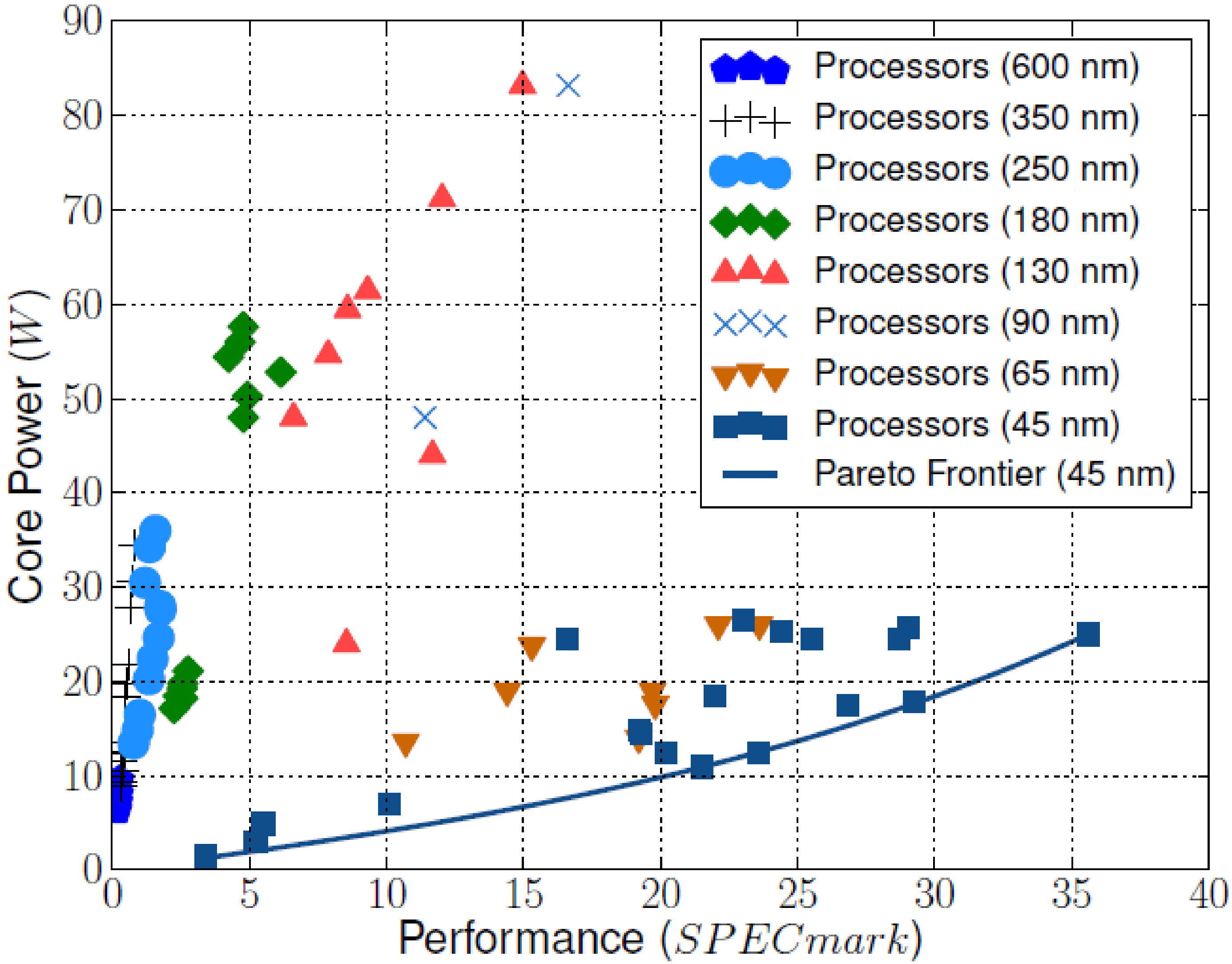
Pollack’s Rule

Pollack's Rule states that microprocessor performance increase due to microarchitecture advances is roughly proportional to [the] square root of [the] increase in complexity.

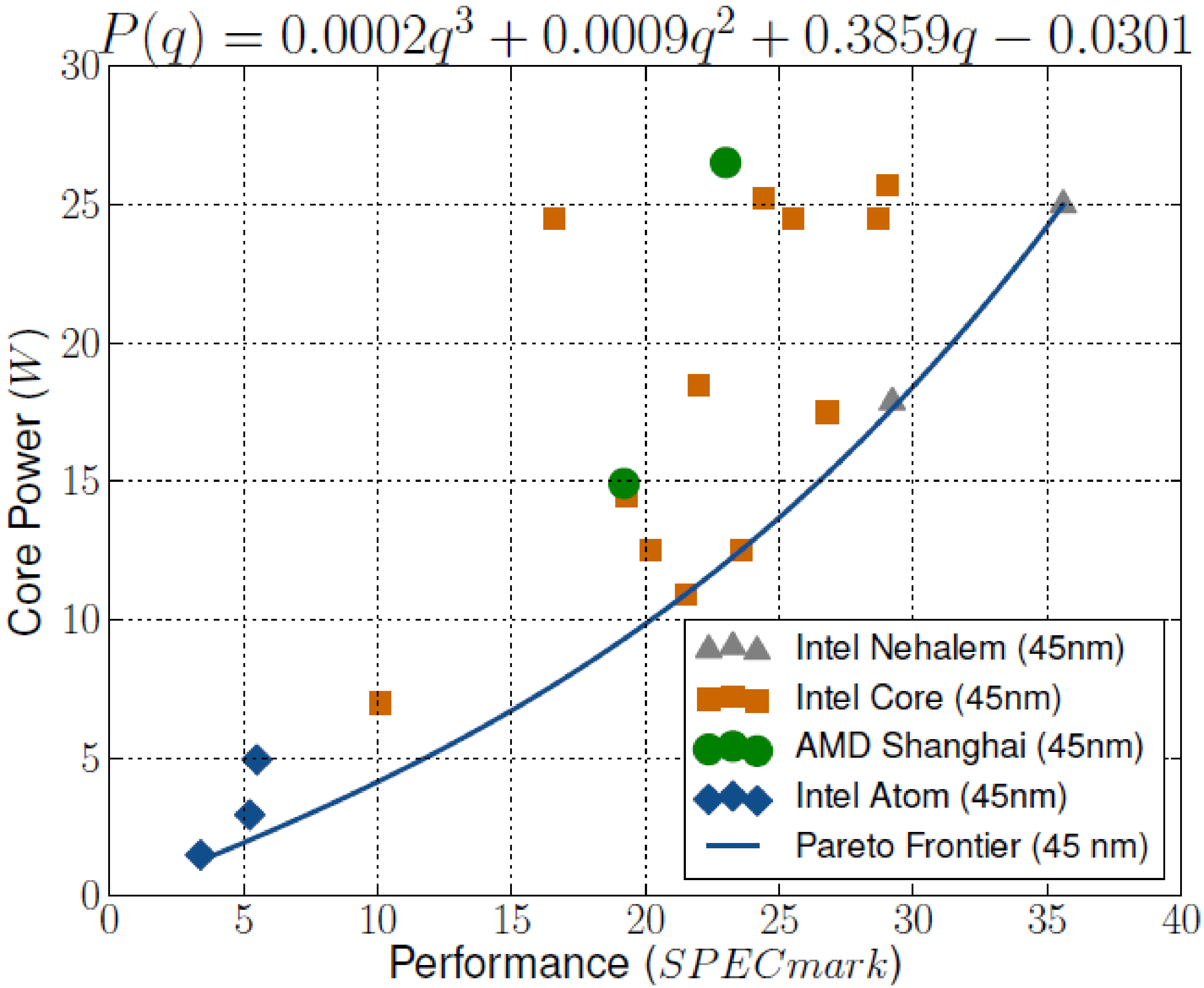
This contrasts with power consumption increase, which is roughly linearly proportional to the increase in complexity.

The performance of a core is proportional to the square root of its area.

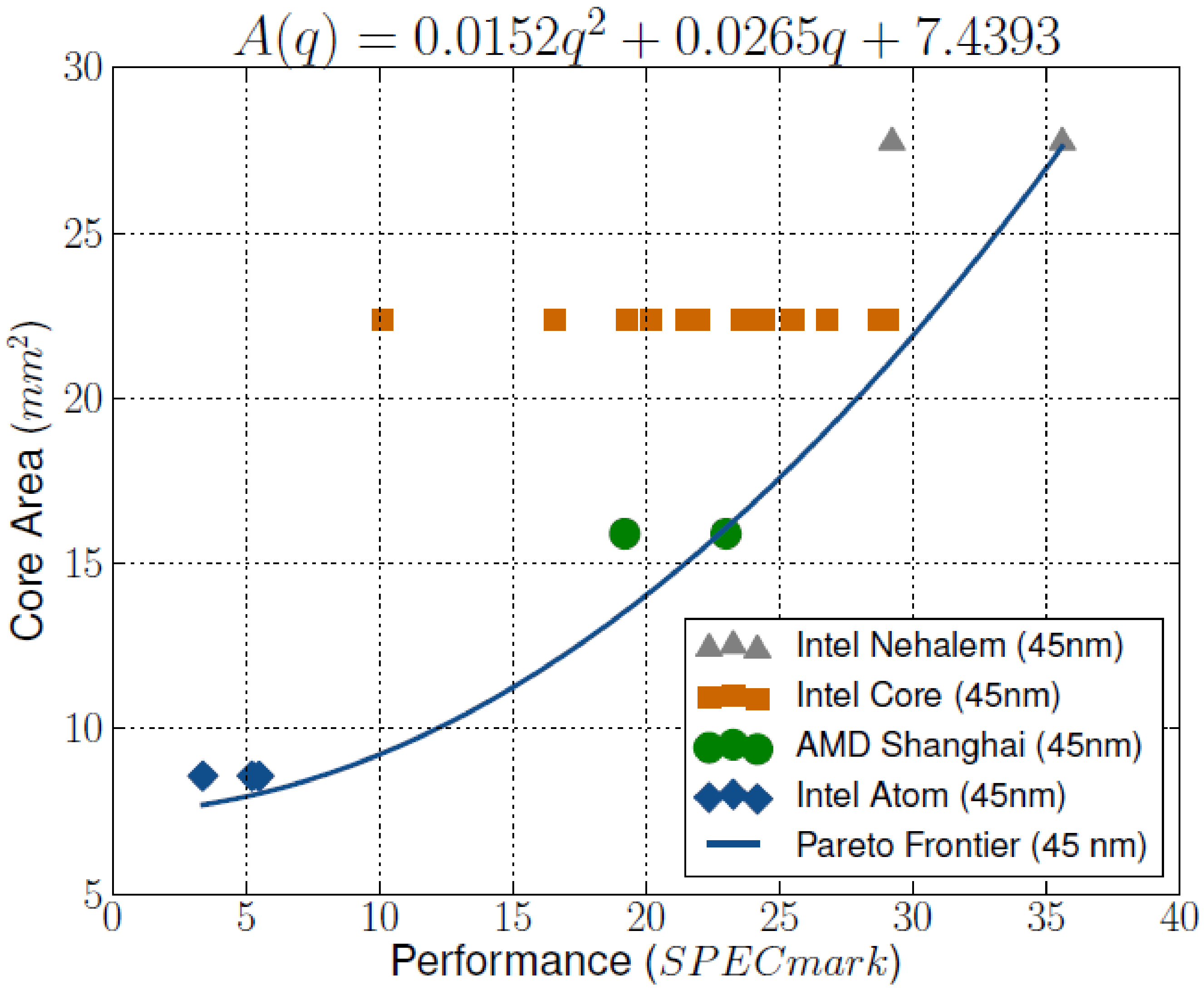
Power/performance



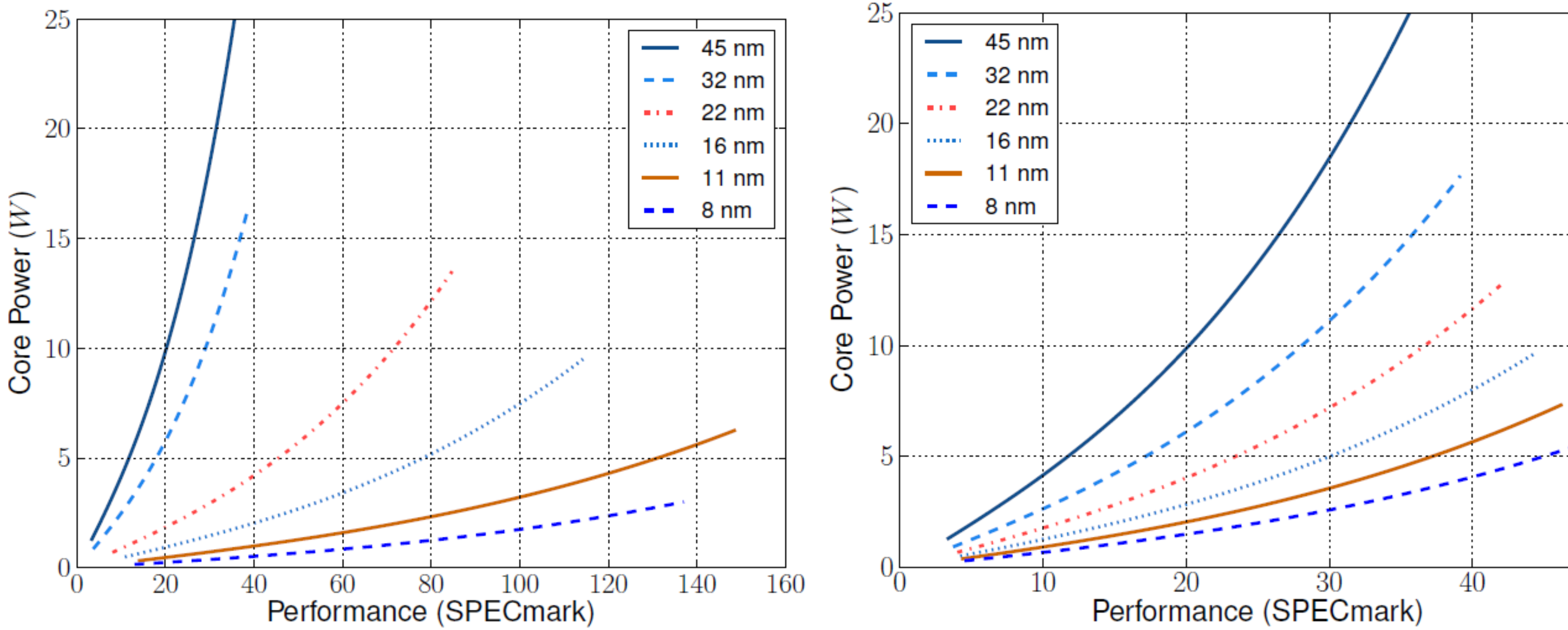
Power/Performance Frontier, 45nm



Area/Performance Frontier, 45nm



Frontier Scaling (Single Core)



(e) ITRS frontier scaling (f) Conservative frontier scaling