The Dark Silicon Implications for Microprocessors

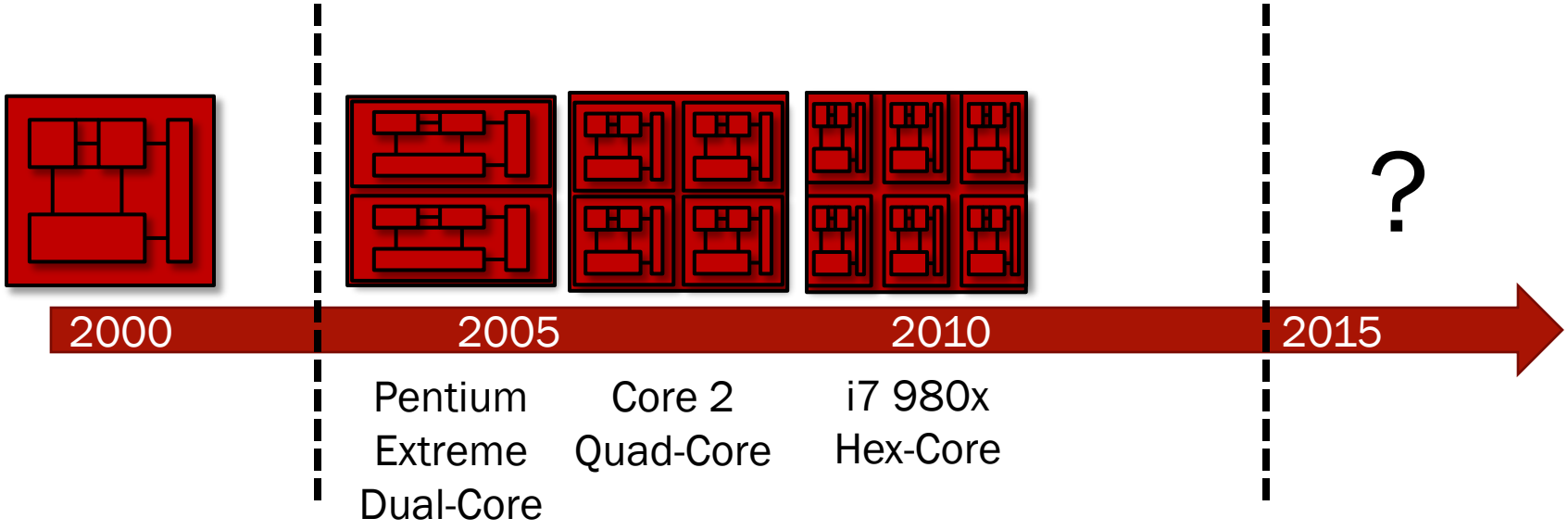
Karu Sankaralingam

University of Wisconsin-Madison

Collaborators: Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, and Doug Burger

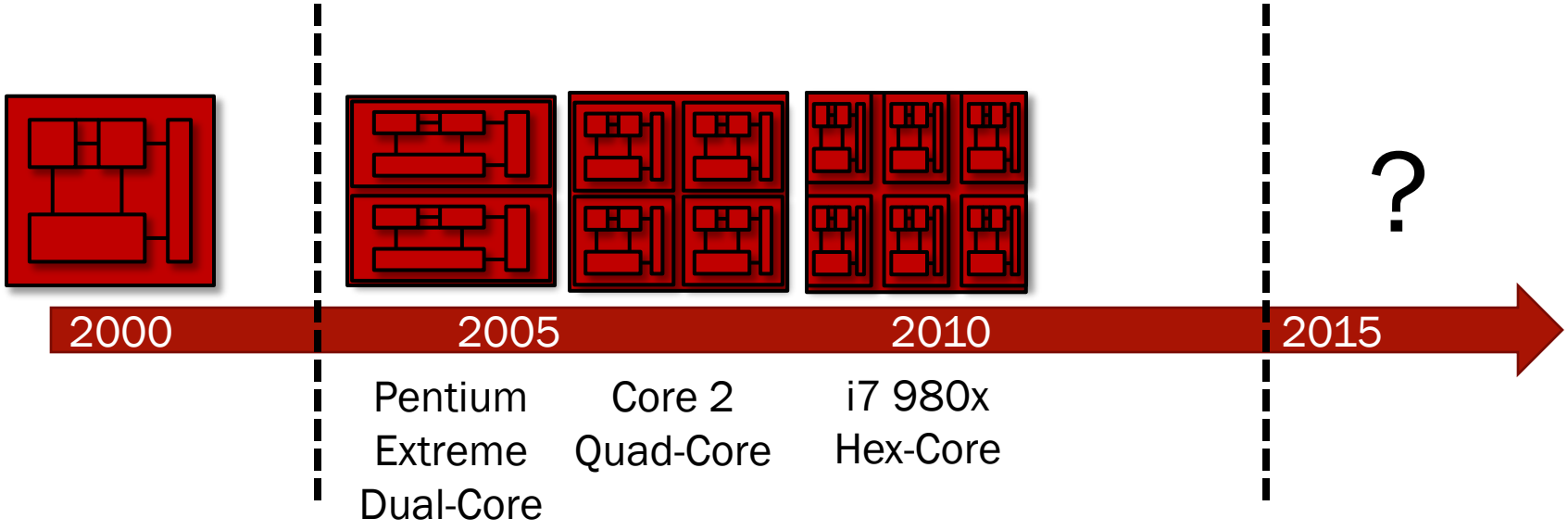
Multicore Decade?

We have relied on multicore scaling for over five years.



Multicore Decade?

We have relied on multicore scaling for over five years.



How much longer will it be our primary performance scaling technique?

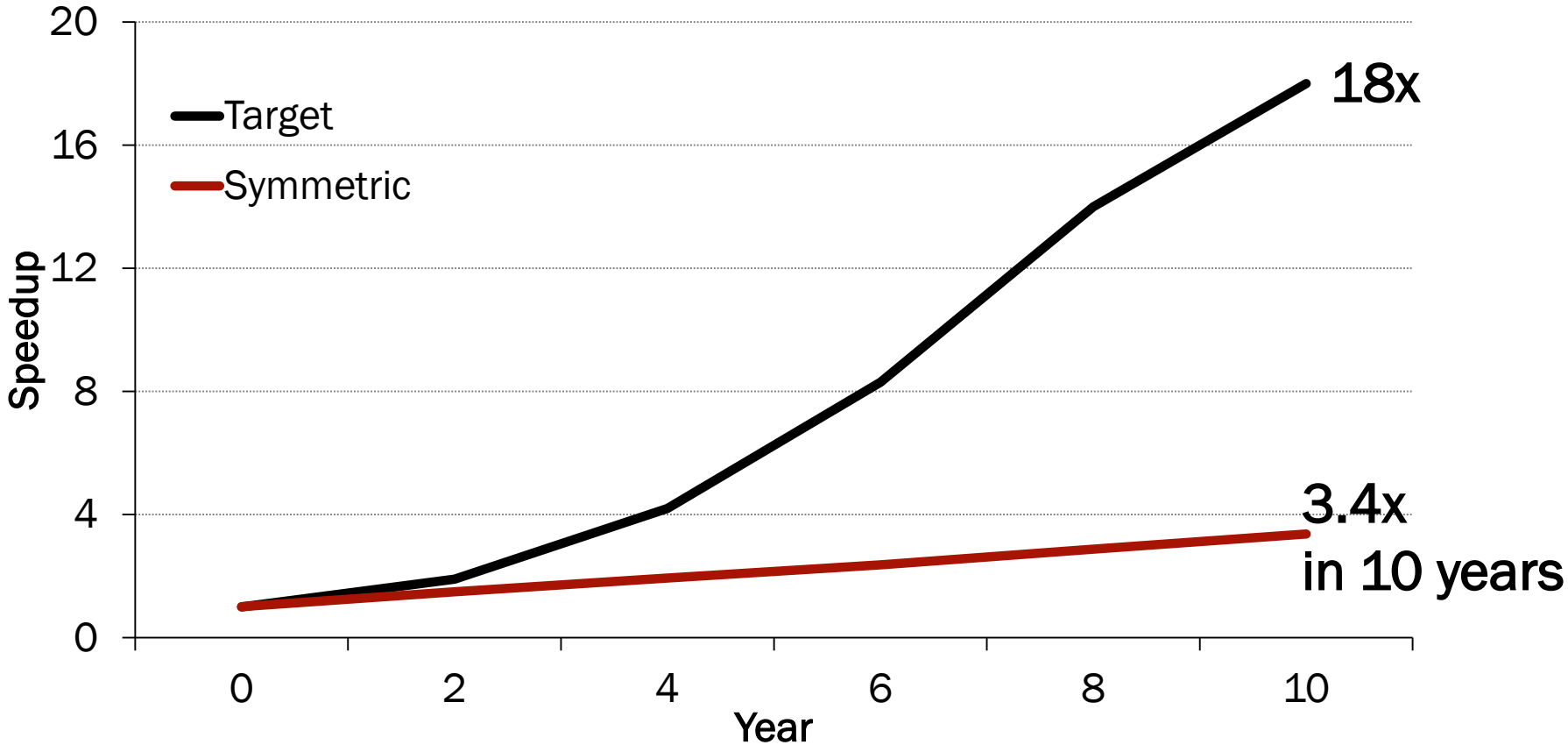
Finding Optimal Multicore Designs

Comprehensive design space:

- Fixed area budget
- Fixed power budget
- Two sets of CMOS scaling projections
- Optimal core and diverse multicore organizations
- Parallel benchmarks

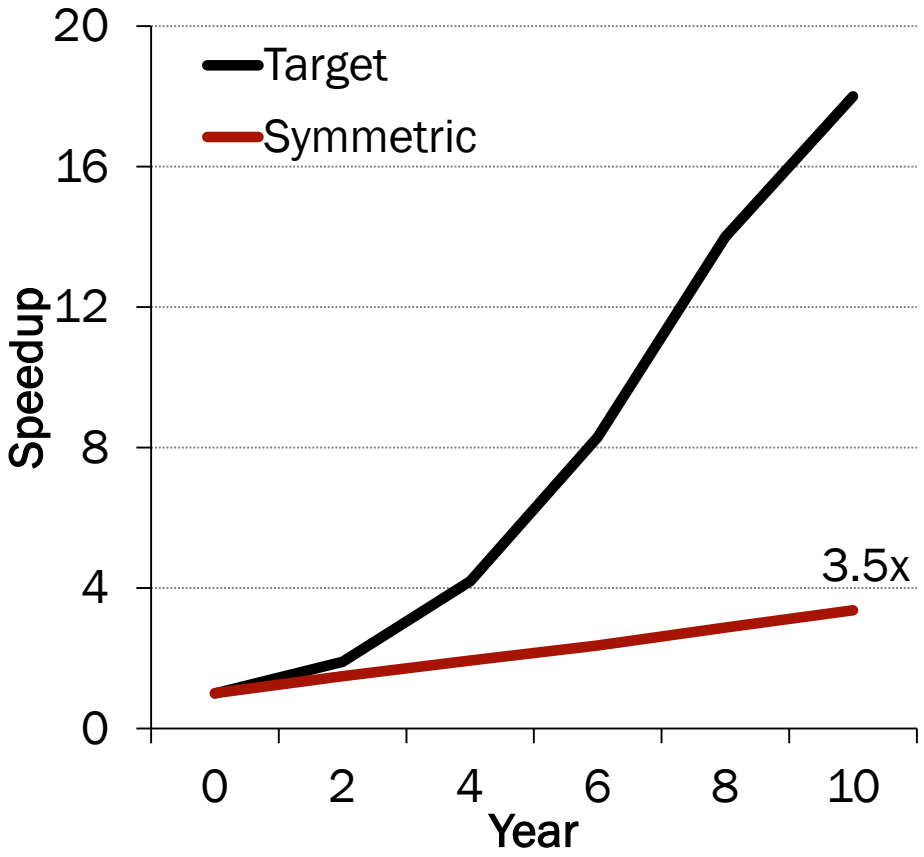
For next 5 technology generations, we find the best performing multicore from a comprehensive design space search for each of the PARSEC benchmarks

Symmetric Multicore Projections

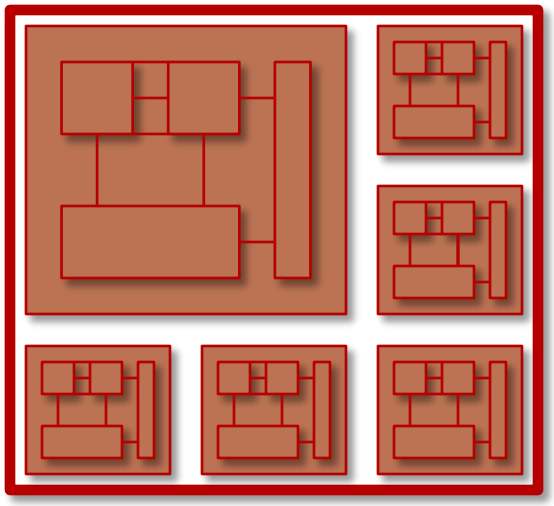


Symmetric multicores alone will not sustain the multicore era.

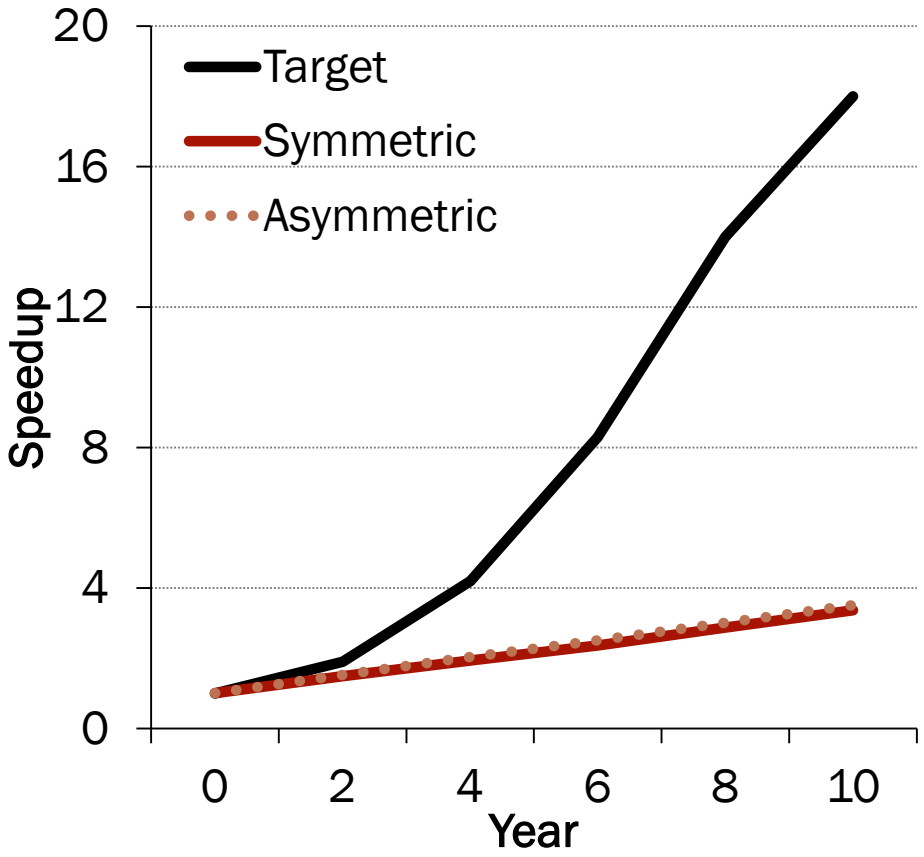
Multicore Solutions



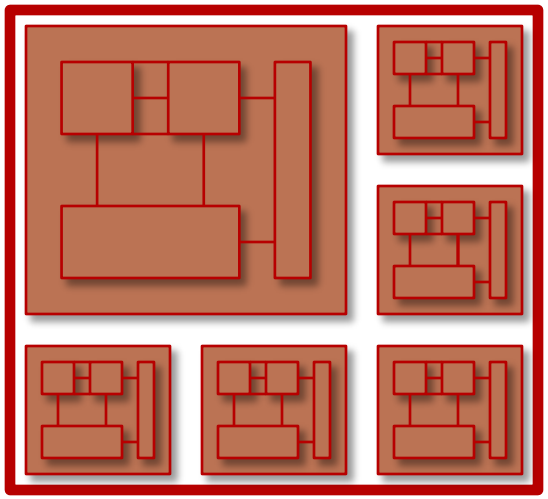
Asymmetric Topologies



Multicore Solutions

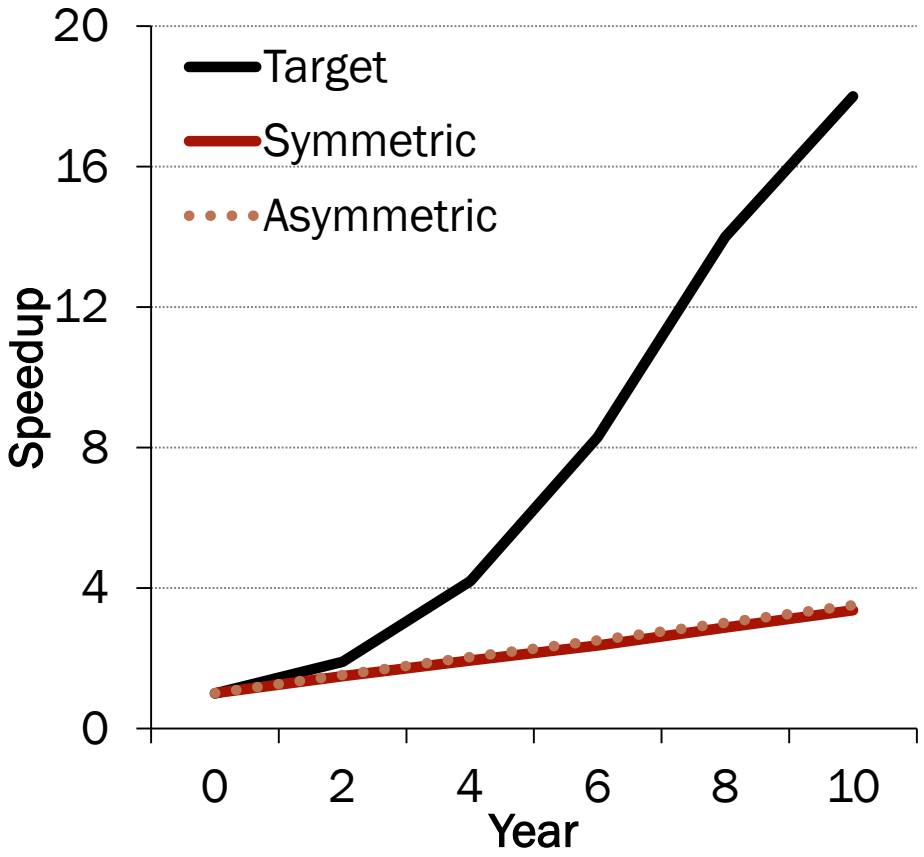


Dynamic Topologies

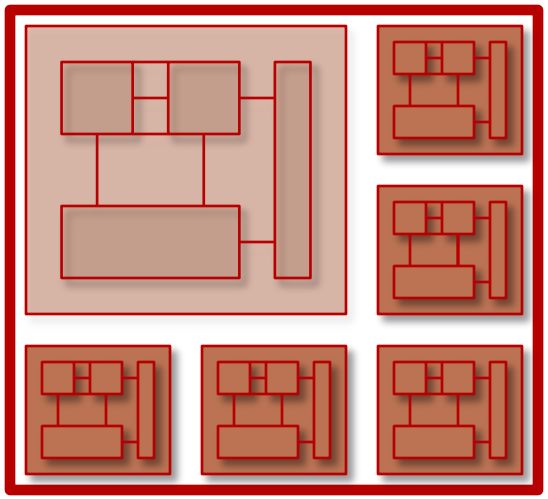


[Chakraborty (2008), Suleman et al (2009)]

Multicore Solutions

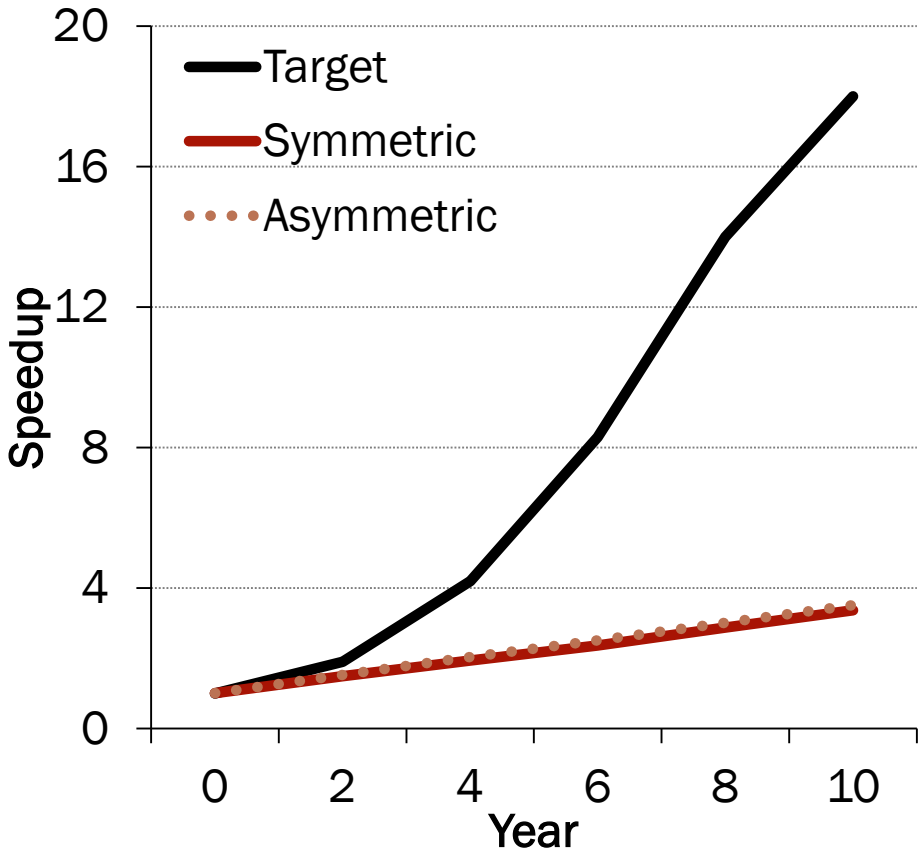


Dynamic Topologies

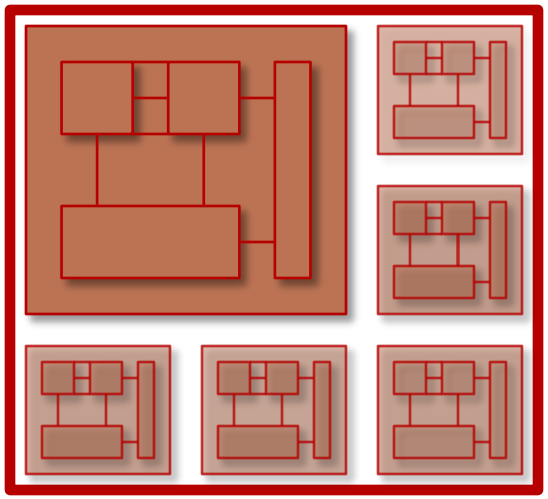


[Chakraborty (2008), Suleman et al (2009)]

Multicore Solutions

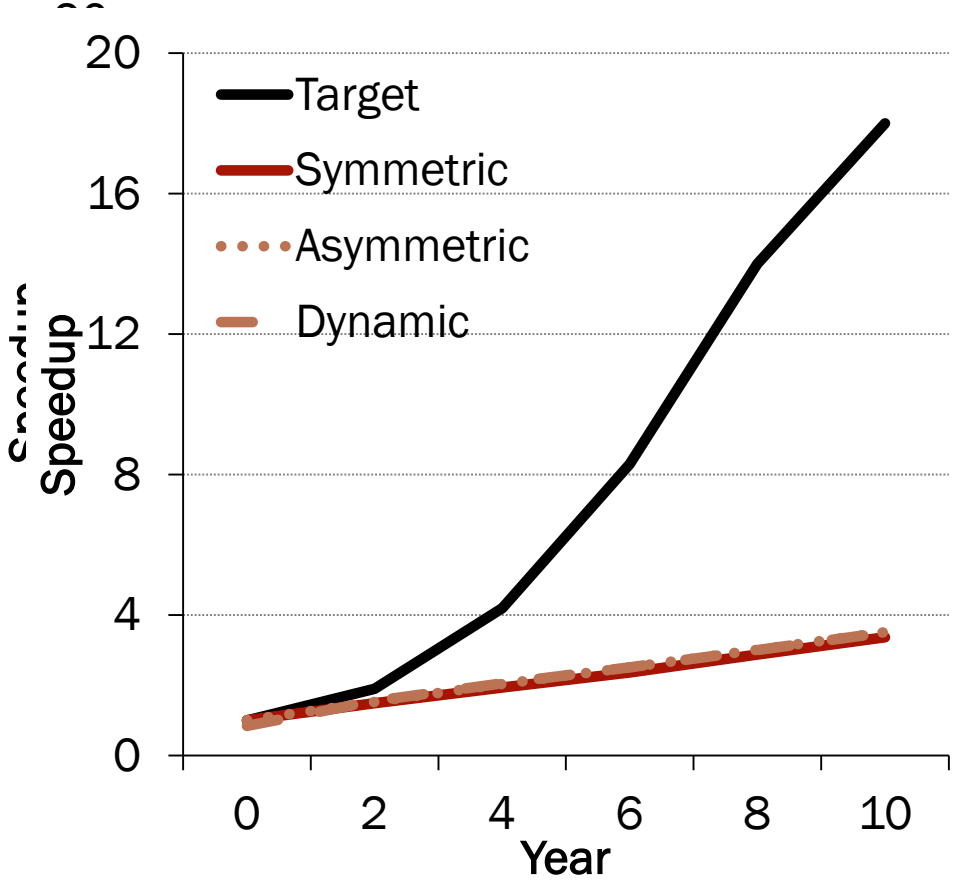


Dynamic Topologies

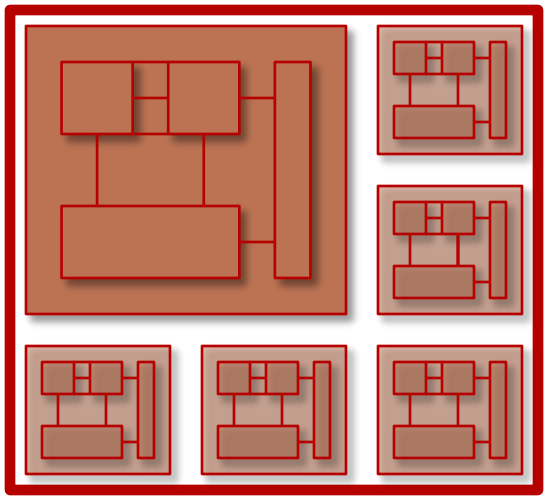


[Chakraborty (2008), Suleman et al (2009)]

Multicore Solutions

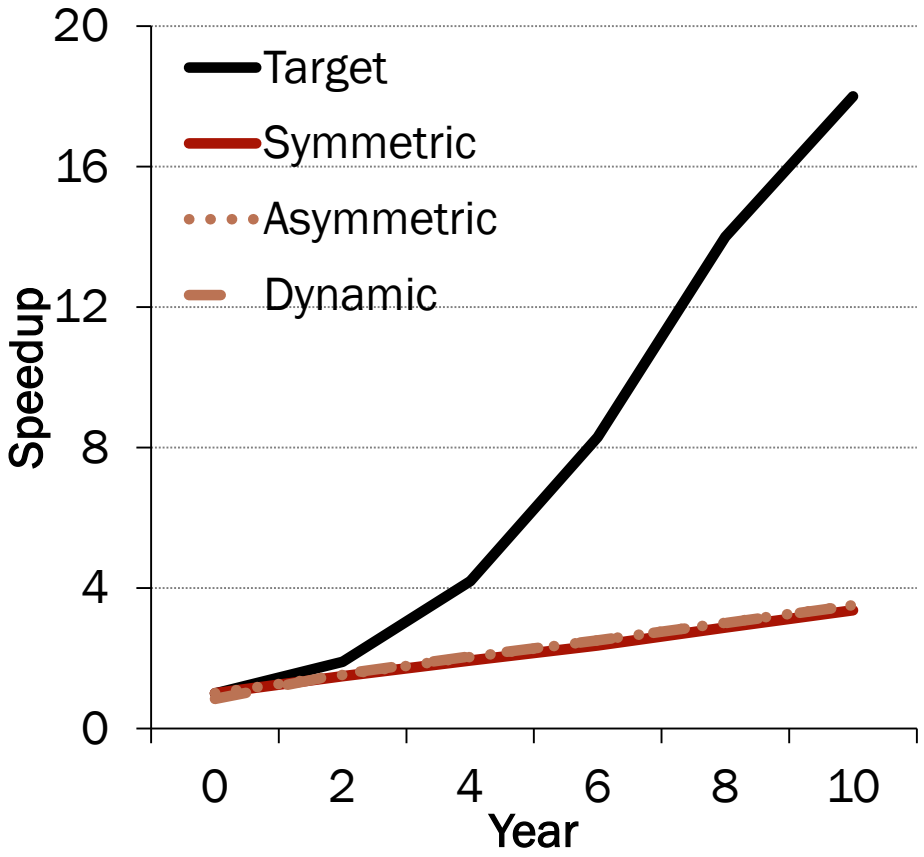


Dynamic Topologies

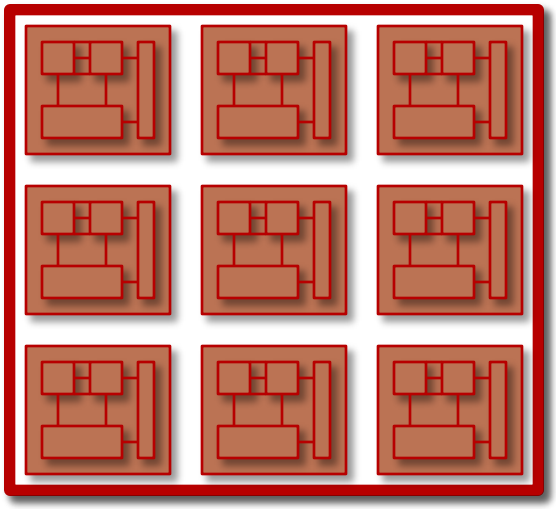


[Chakraborty (2008), Suleman et al (2009)]

Multicore Solutions

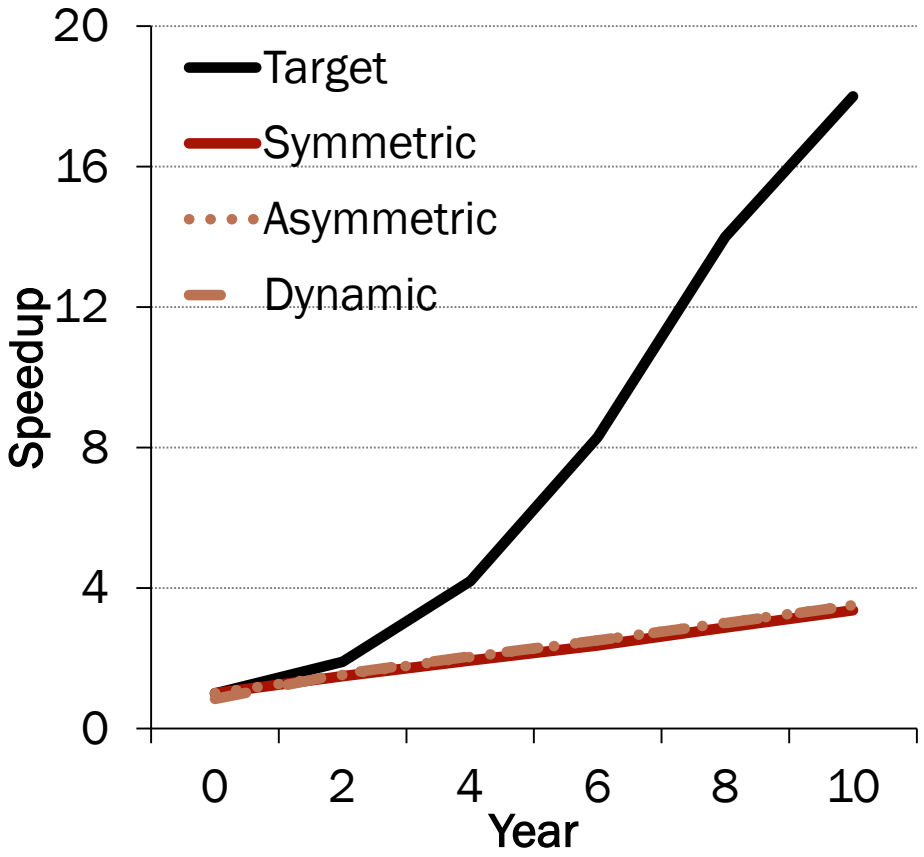


Composed/Fused Topologies

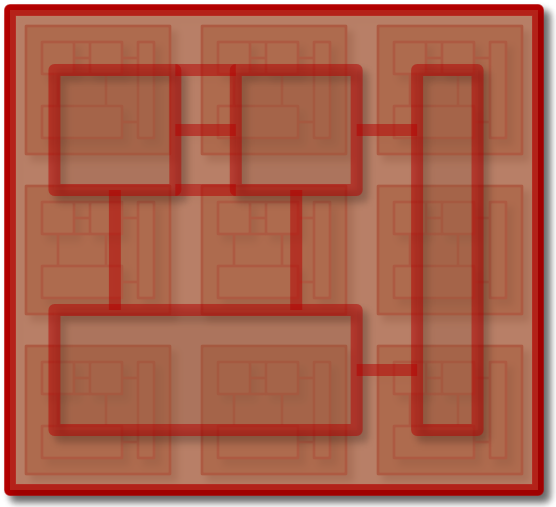


[Ipek et al (2007), Kim et al (2007)]

Multicore Solutions

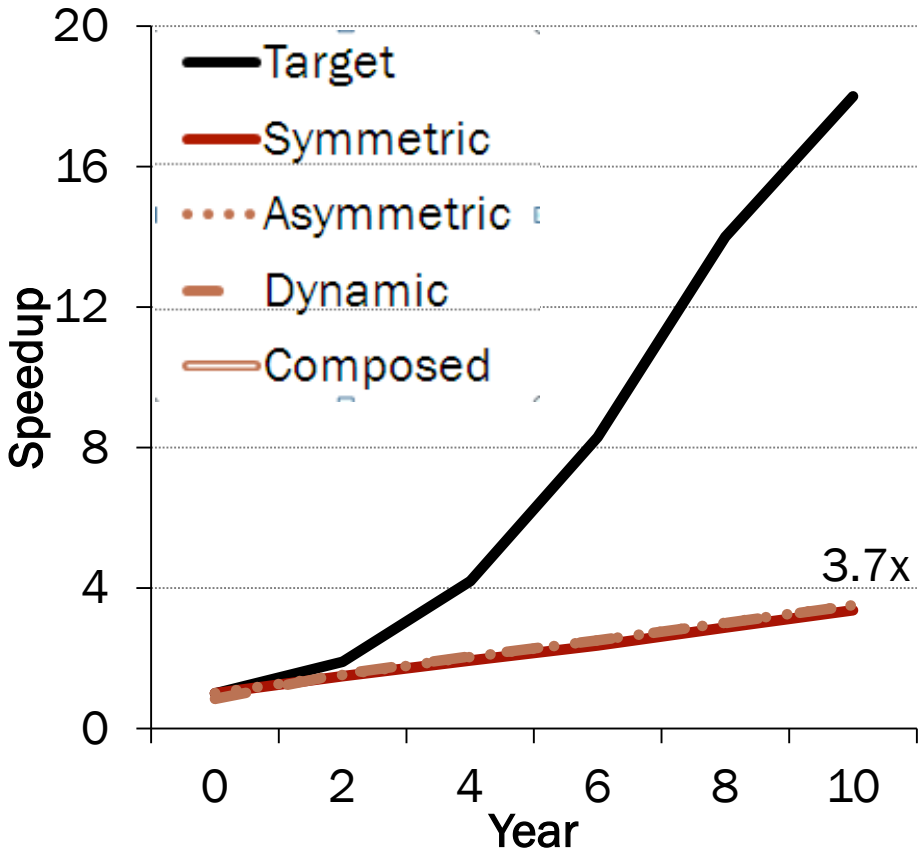


Composed/Fused Topologies

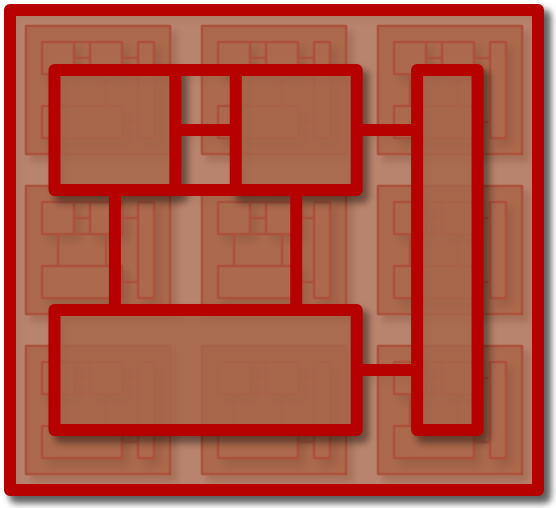


[Ipek et al (2007), Kim et al (2007)]

Multicore Solutions

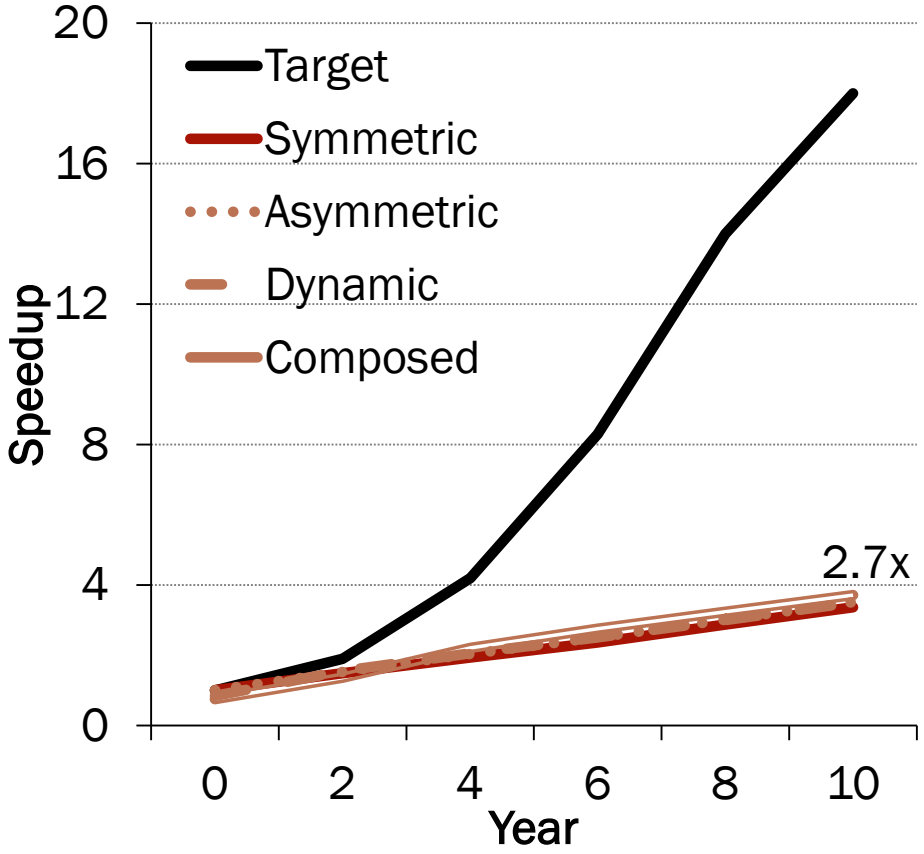


Composed/Fused Topologies

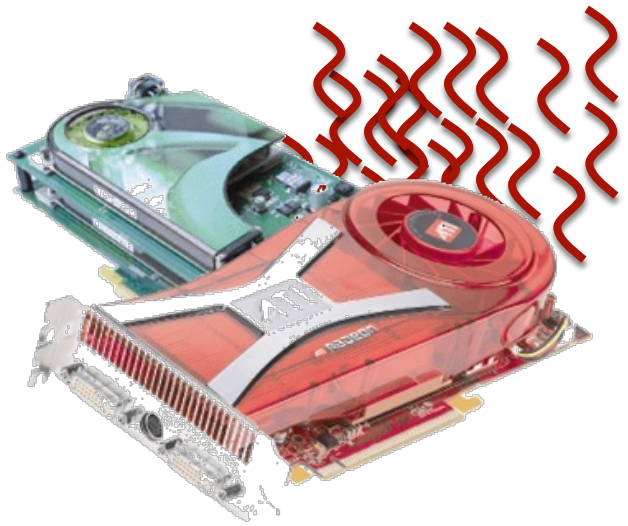


[Ipek et al (2007), Kim et al (2007)]

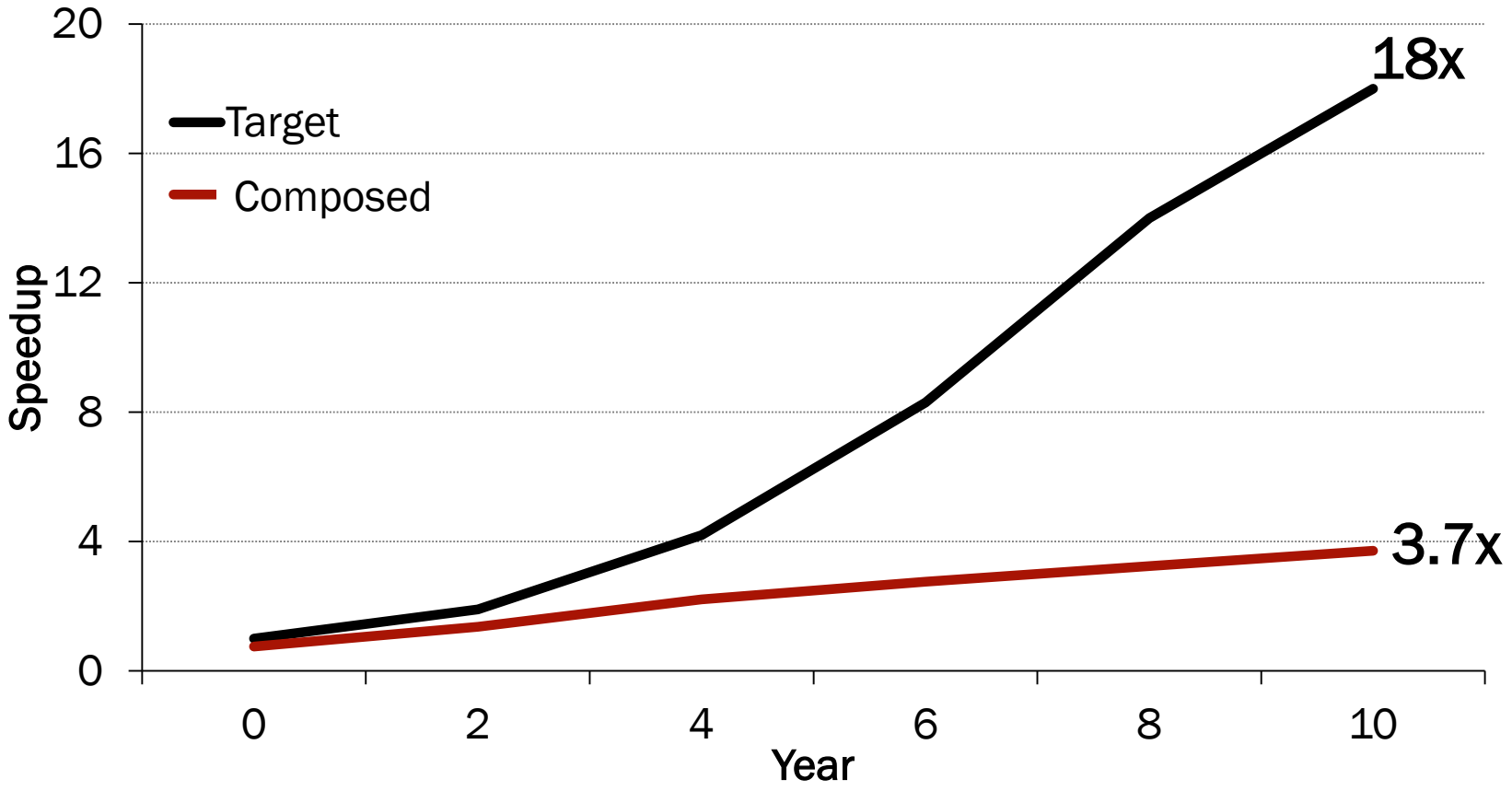
Multicore Solutions



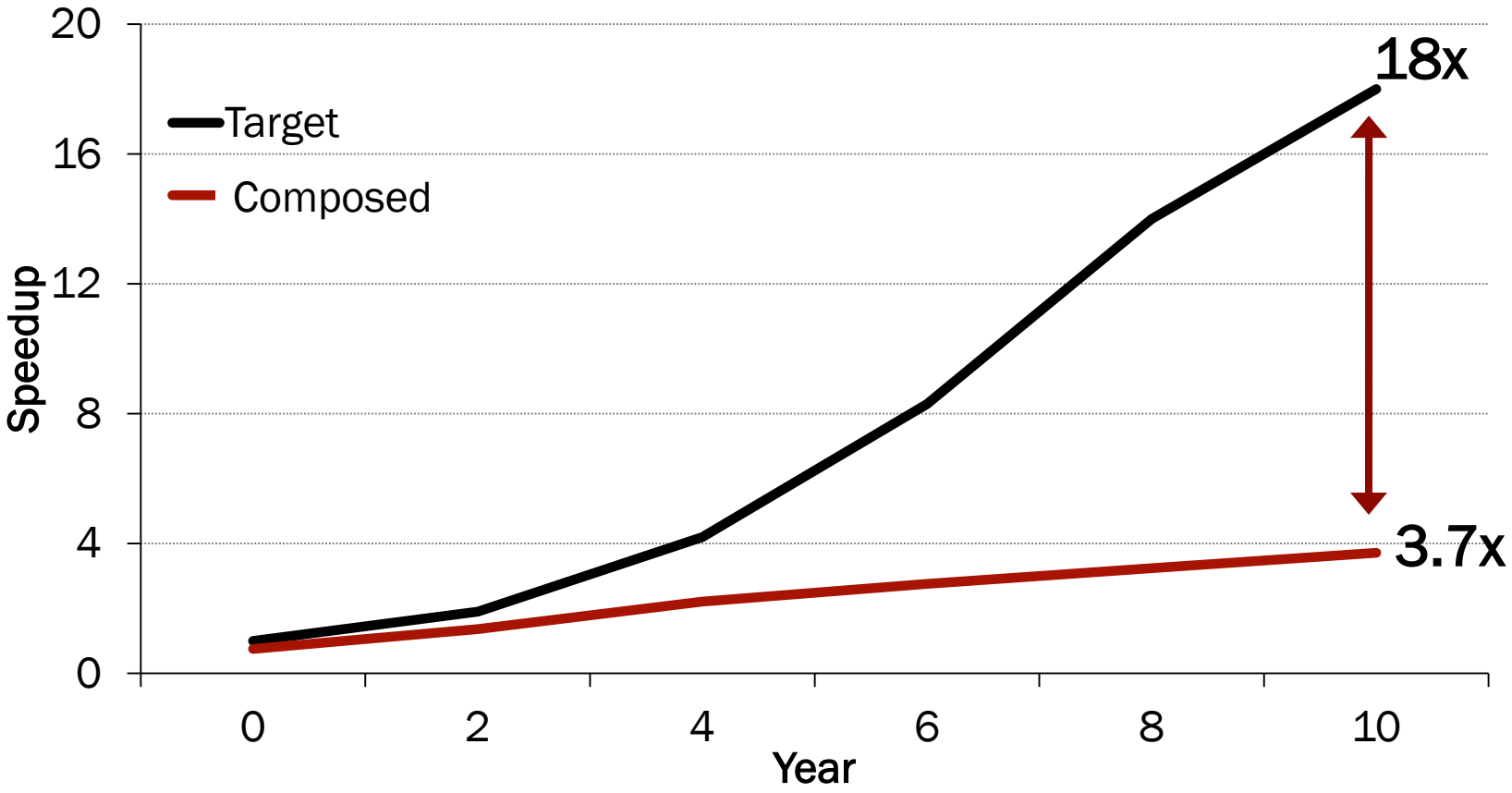
GPU-Style Cores



Multicore Era Projections

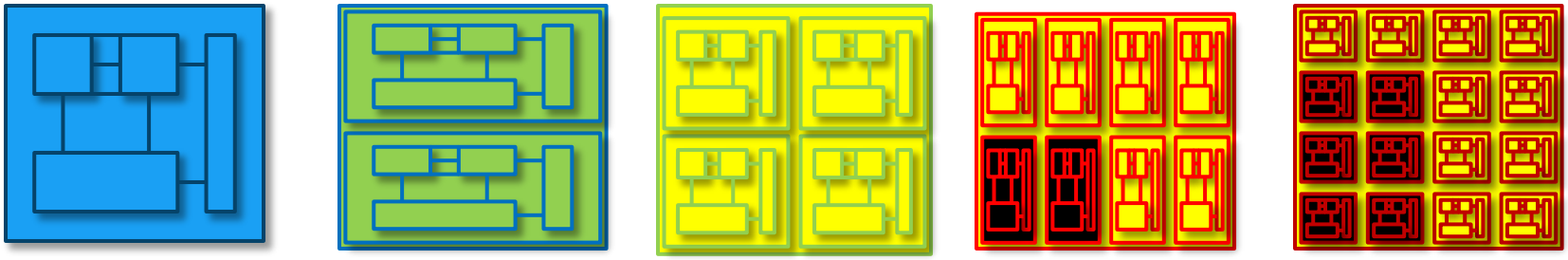


Multicore Era Projections



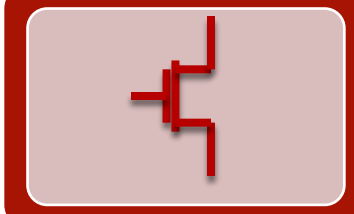
The best designs speed up 14% per year
rather than the recent trend of 34% per year

Why Diminishing Returns?



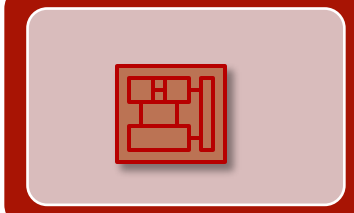
- Transistor area is still scaling
- Voltage and capacitance scaling have slowed
- Result: designs are power, not area, limited

Overview



Devices

- Find the best case technology scaling

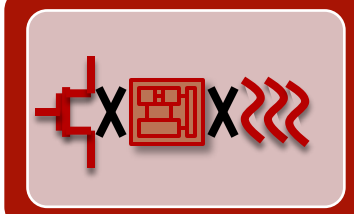


Cores

- Find the best cores

Multicores

- Find the best multicore organization



Projections

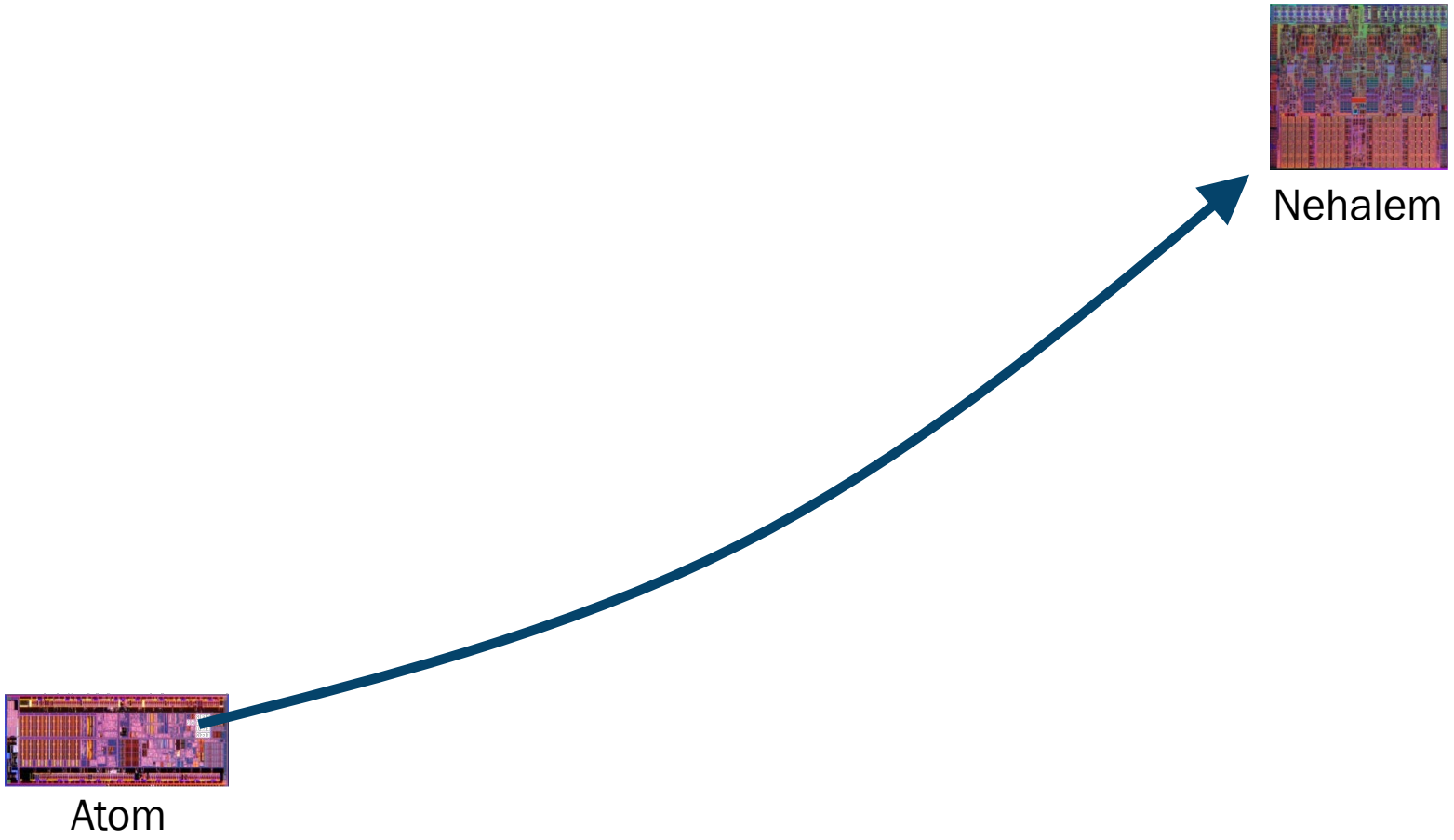
- Predict best case multicore performance for each technology generation

Device Scaling Projections

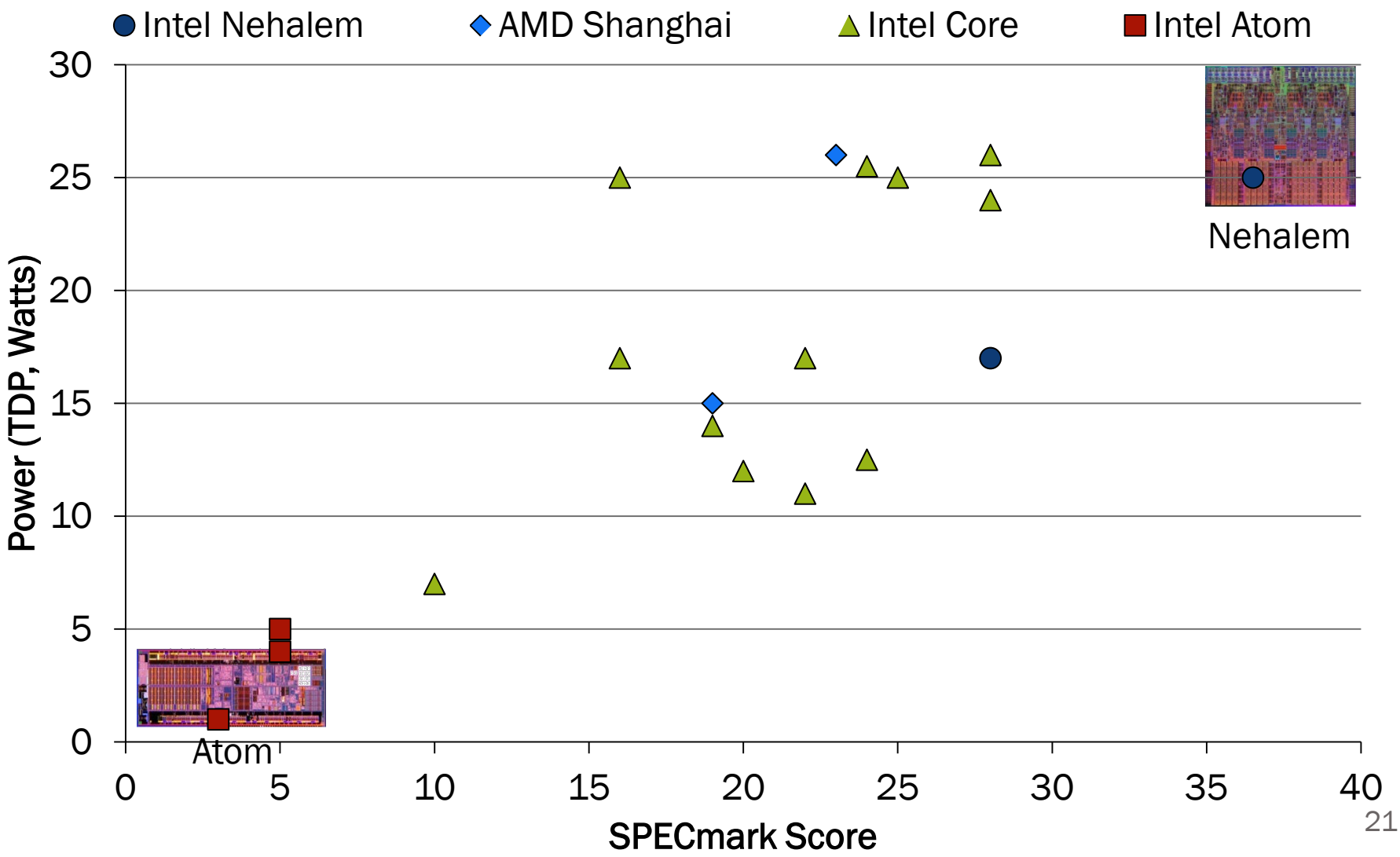
From 45 nm to 8 nm:

	Conservative	Optimistic
Area	32x ↓	32x ↓
Power	4.5x ↓	8.3x ↓
Frequency	1.3x ↑	3.9x ↑
	[Borkar 2007]	[ITRS 2010]

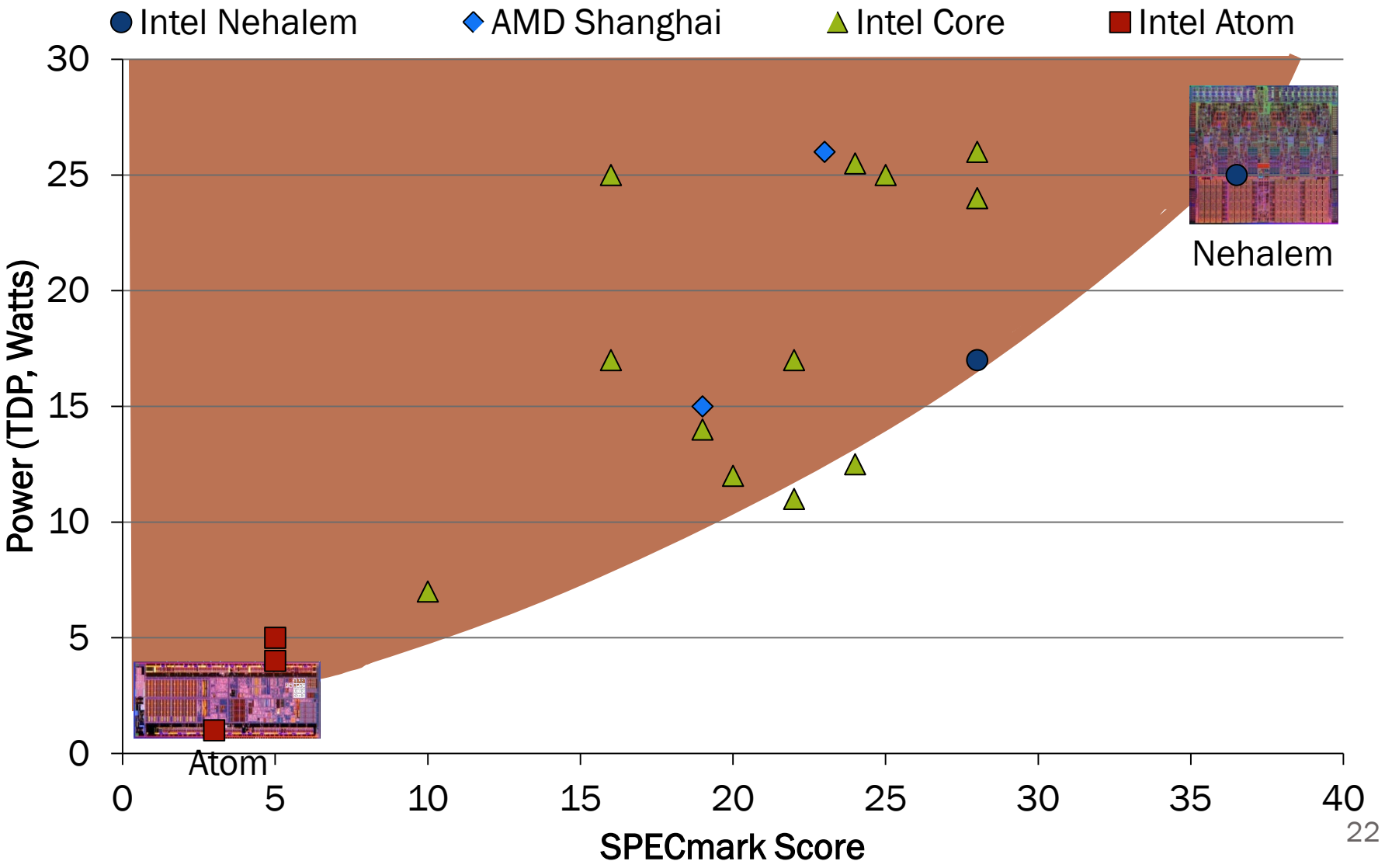
Modeling Ideal Core Power/Perf.



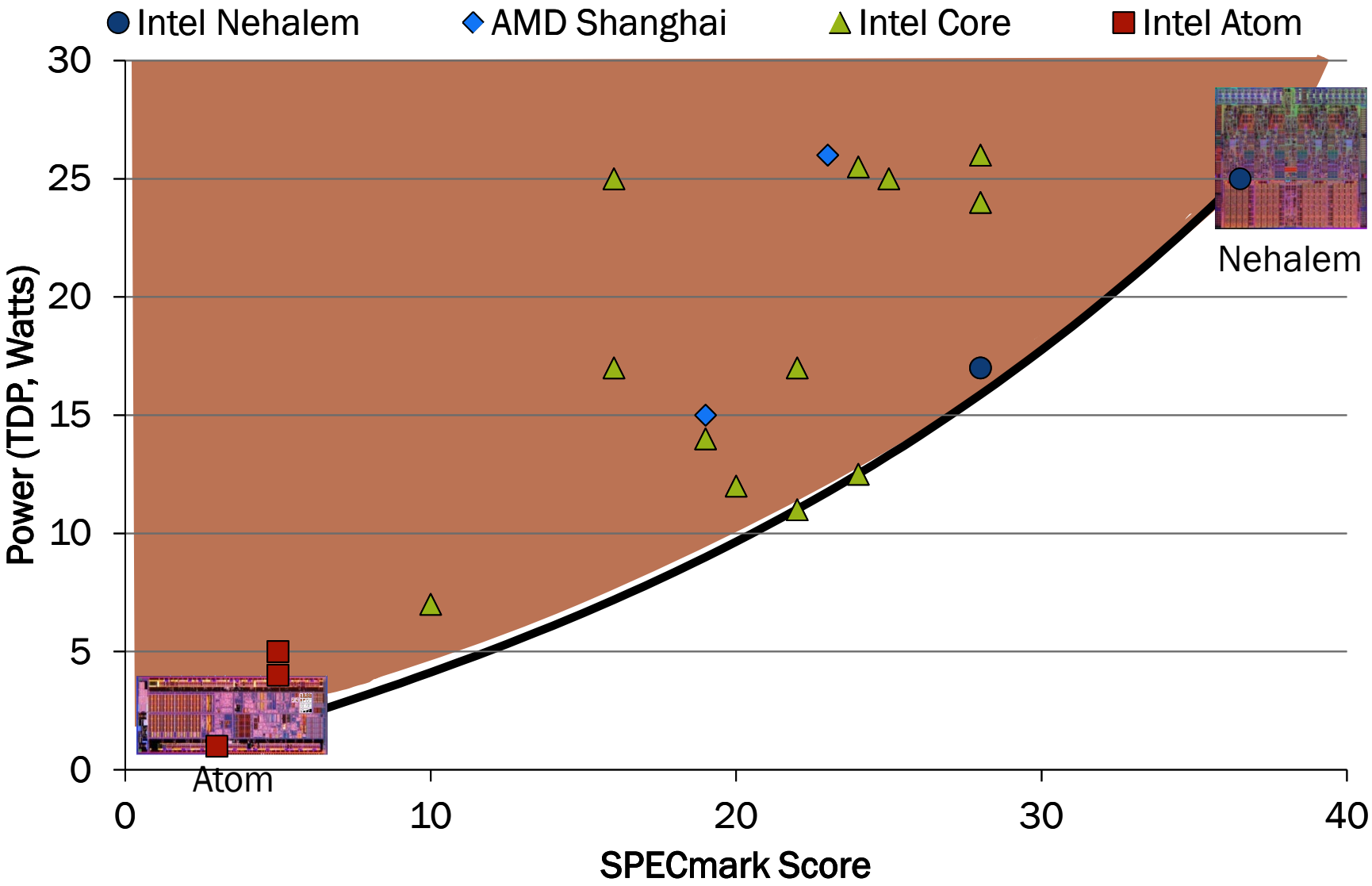
Modeling Ideal Core Power/Perf.



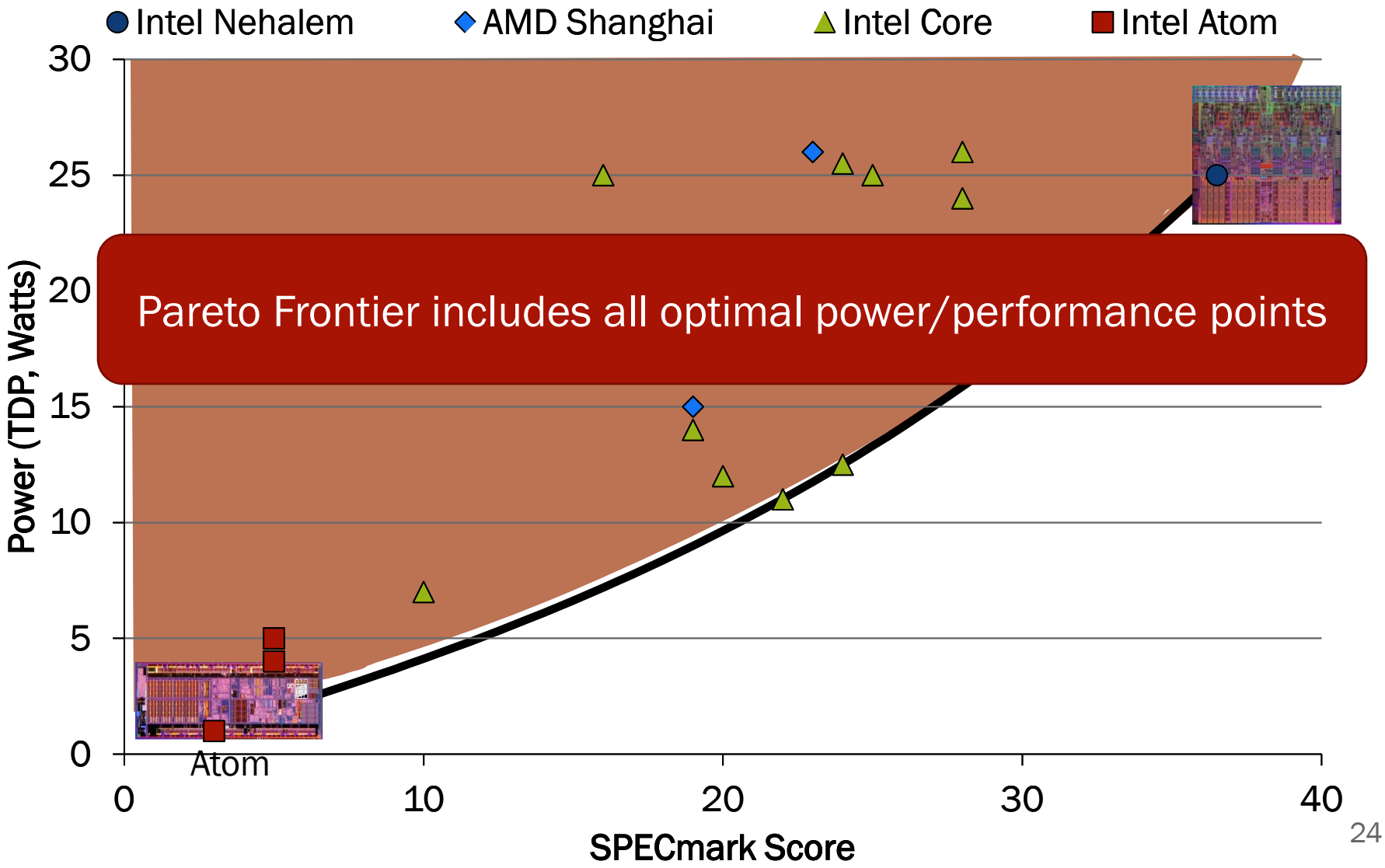
Modeling Ideal Core Power/Perf.



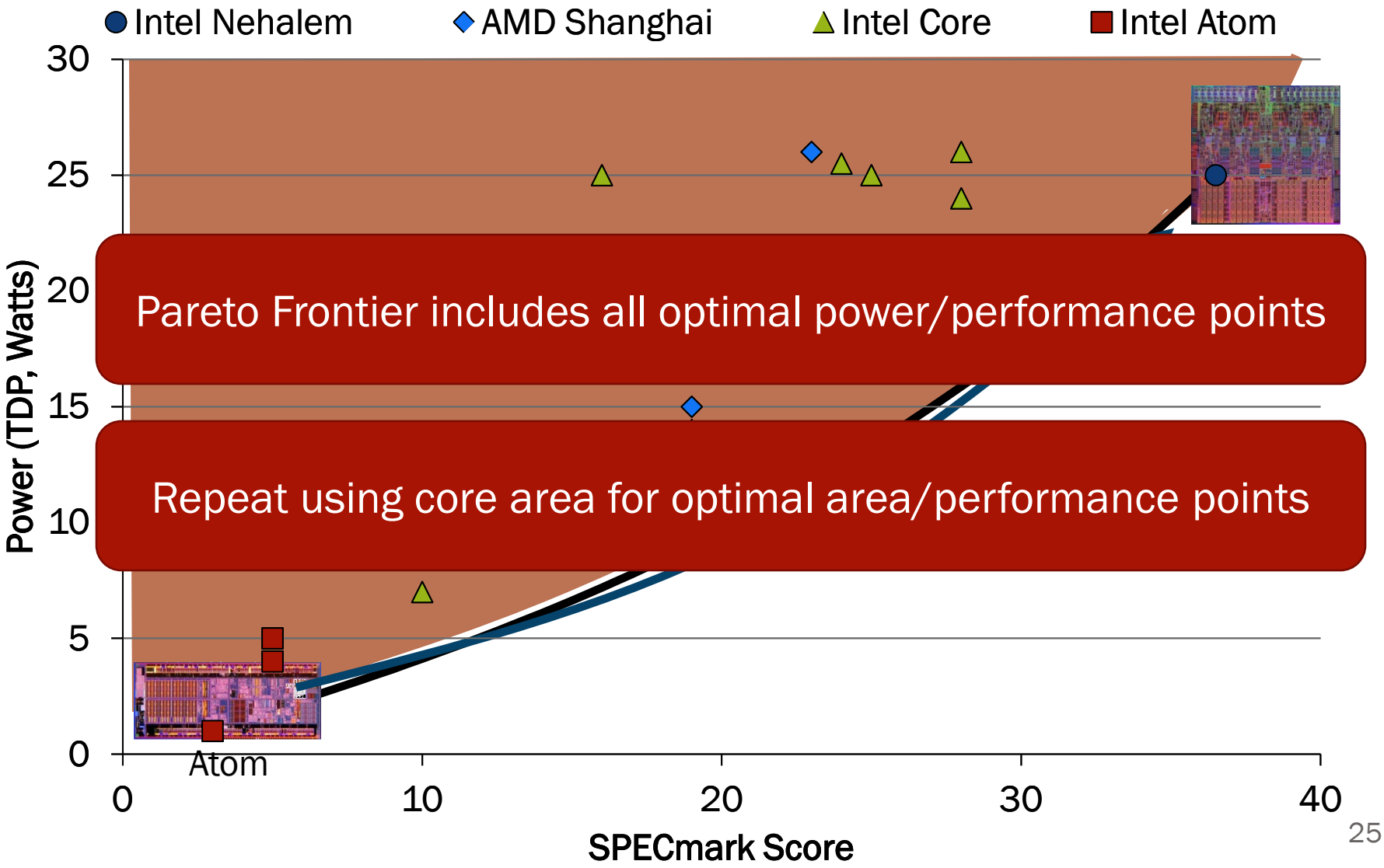
Modeling Ideal Core Power/Perf.



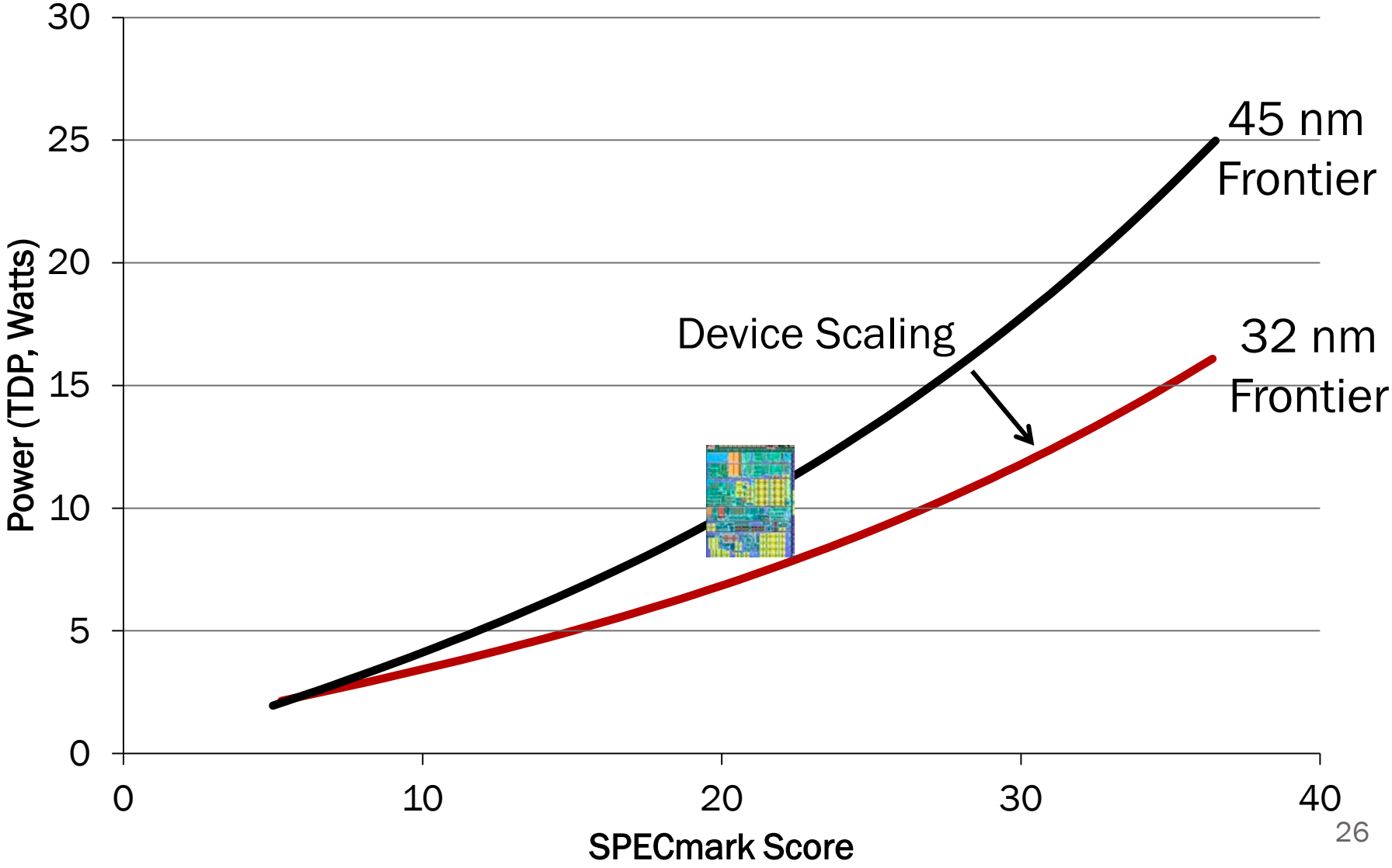
Modeling Ideal Core Power/Perf.



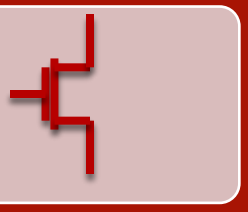
Modeling Ideal Core Power/Perf.



Combining Device and Core Models

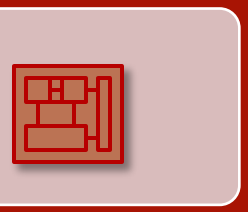


Overview



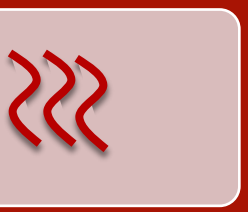
Devices

- Find the best case technology scaling



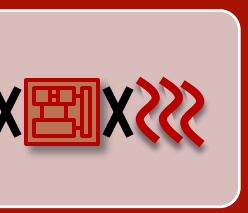
Cores

- Find the best cores



Multicores

- Find the best multicore organization

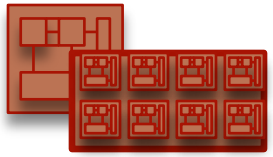


Projections

- Predict best case multicore performance for each technology generation

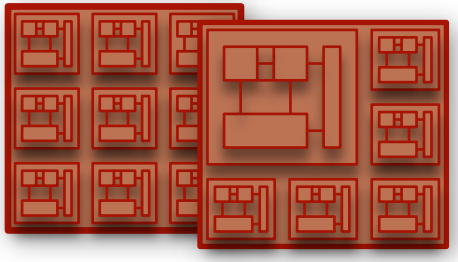
What belongs in multicore model?

Styles



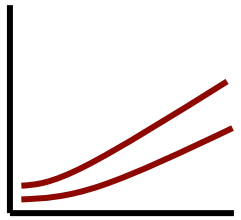
Number of Threads, Cache Sizes

Topologies



Area & Power Budget

Pareto Frontiers



Area & Power / Performance Tradeoffs

Architectures

Cache & memory latencies, memory bandwidth

Applications

PARSEC $f_{parallel}$, Data Use

Multicore Speedup Model

$$\text{Multicore Speedup} = \frac{1}{\frac{1-f_{parallel}}{\text{Serial Speedup}} + \frac{f_{parallel}}{\text{Parallel Speedup}}}$$

Multicore Performance Model

Performance is limited by:

Memory bandwidth

BW_{max} / (instructions per byte from memory)

and

Computation

$N_{cores} \times$ (core frequency/CPI_{exe}) $\times$ core utilization

[Guz et al, 2009]

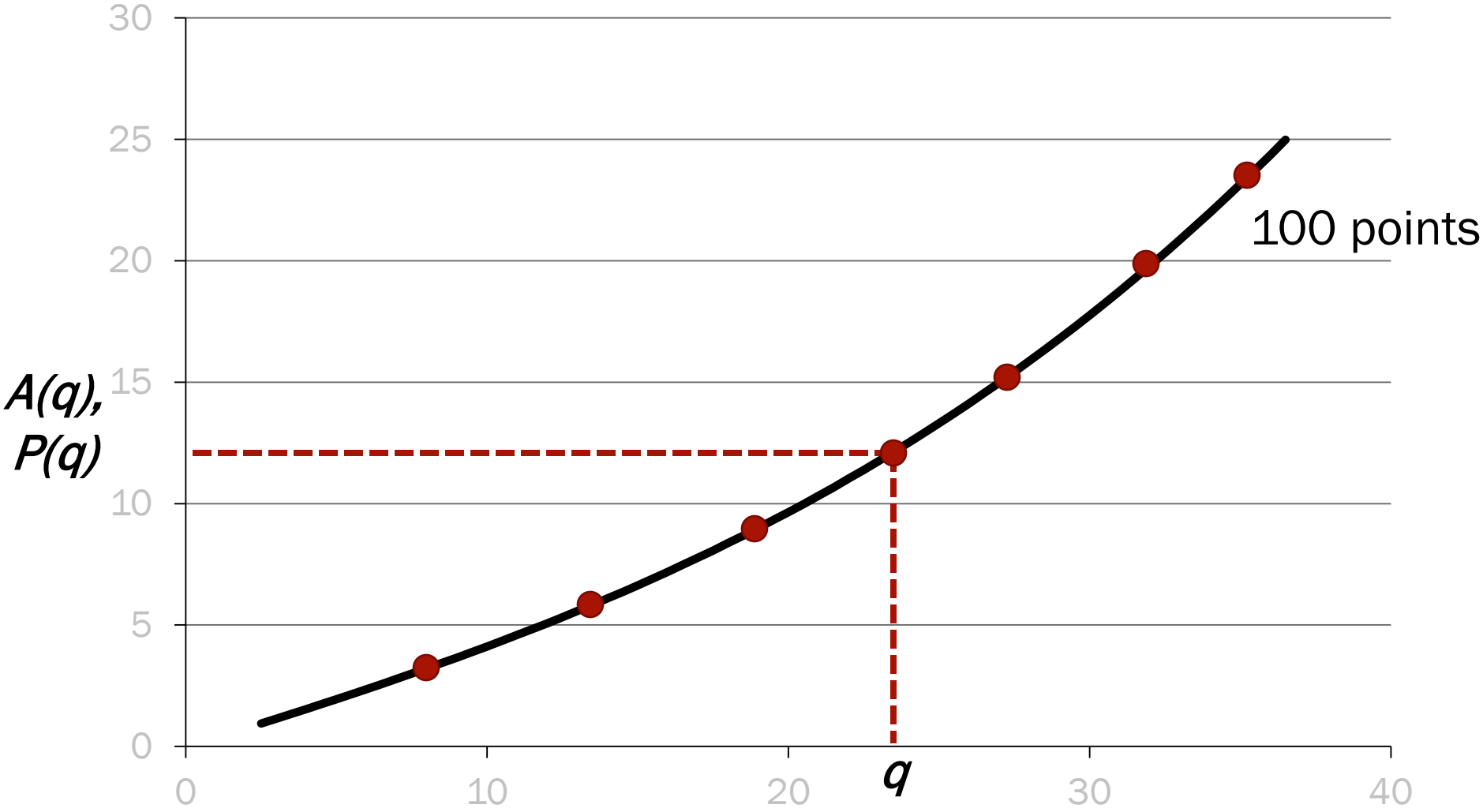
Core Utilization Model

Core utilization is limited by:

Fraction of Time Core is Ready to Issue
Number of Threads in Core / Number of Threads to Keep Busy

[Guz et al, 2009]

Multicore Model & Pareto Frontiers



Translating from SPECmark

1. From *q*, find core's SPECmark speedup

2. Frequency linearly distributed from Atom to Nehalem

3. Recall: model predicts benchmark performance as
   f(benchmark chars, frequency, CPI_{exe})

4. Compute CPI_{exe} such that
   Benchmark Speedup = SPECmark Speedup

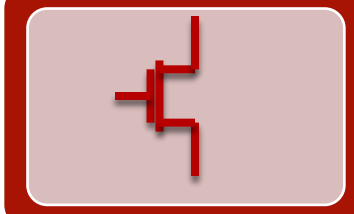
Area and Power Constraints

$$N_{cores} \; x \; A(q) \leq \text{Area Budget}$$

$$N_{cores} \; x \; P(q) \leq \text{ Power Budget}$$

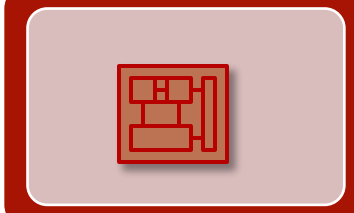
Dark silicon = N_{cores} / # of cores that fit in chip area

Overview



Devices

- Find the best case technology scaling

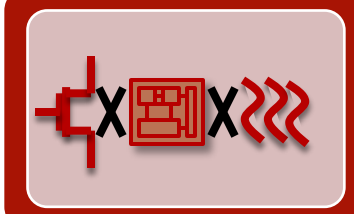


Cores

- Find the best cores

Multicores

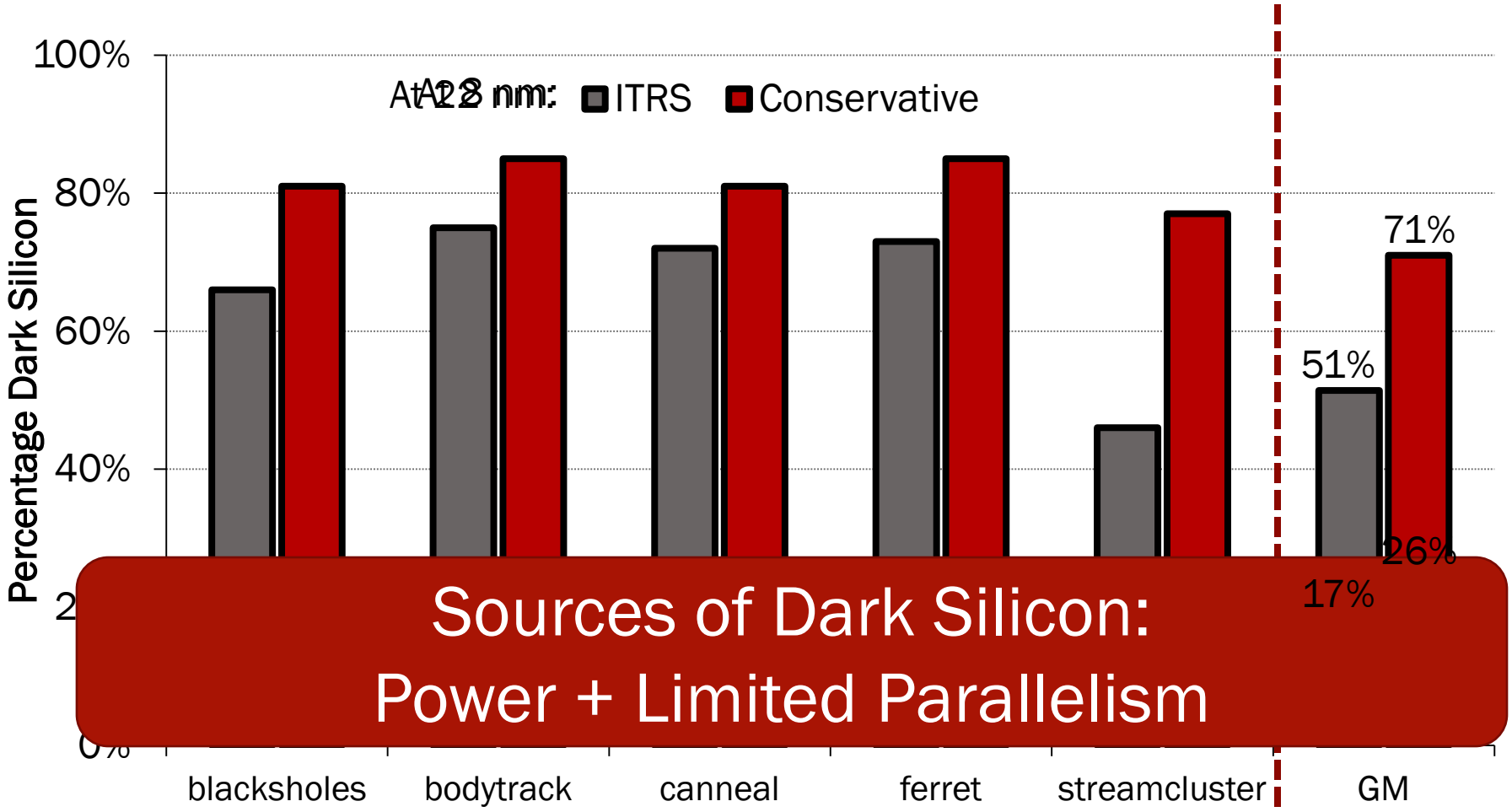
- Find the best multicore organization



Projections

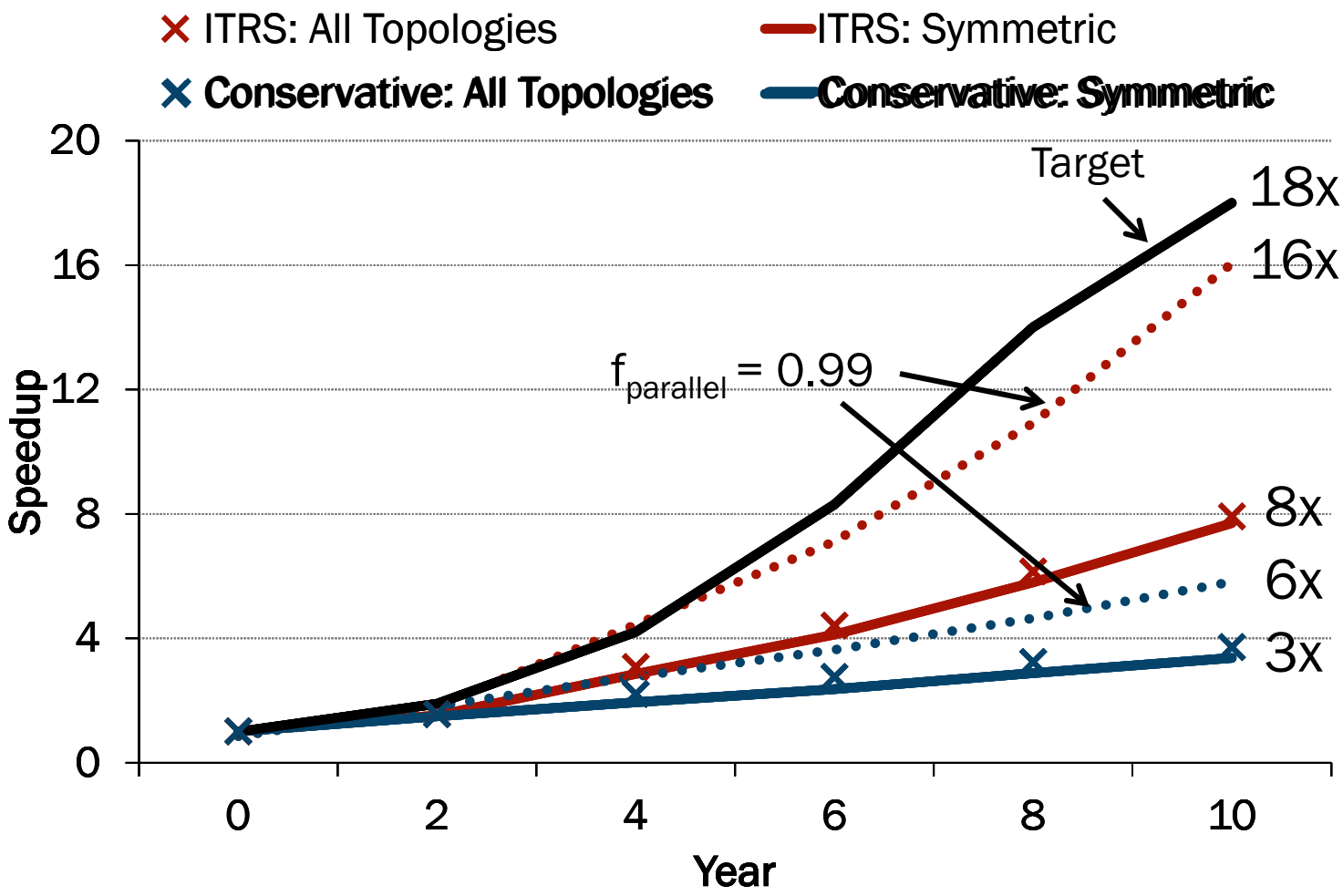
- Predict best case multicore performance for each technology generation

Dark Silicon



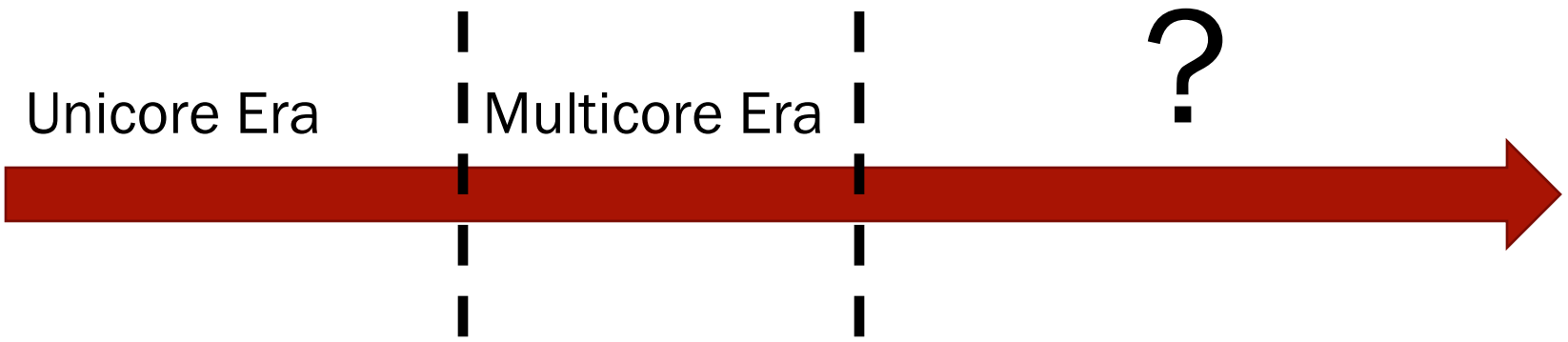
Sources of Dark Silicon:
Power + Limited Parallelism

Overall Performance

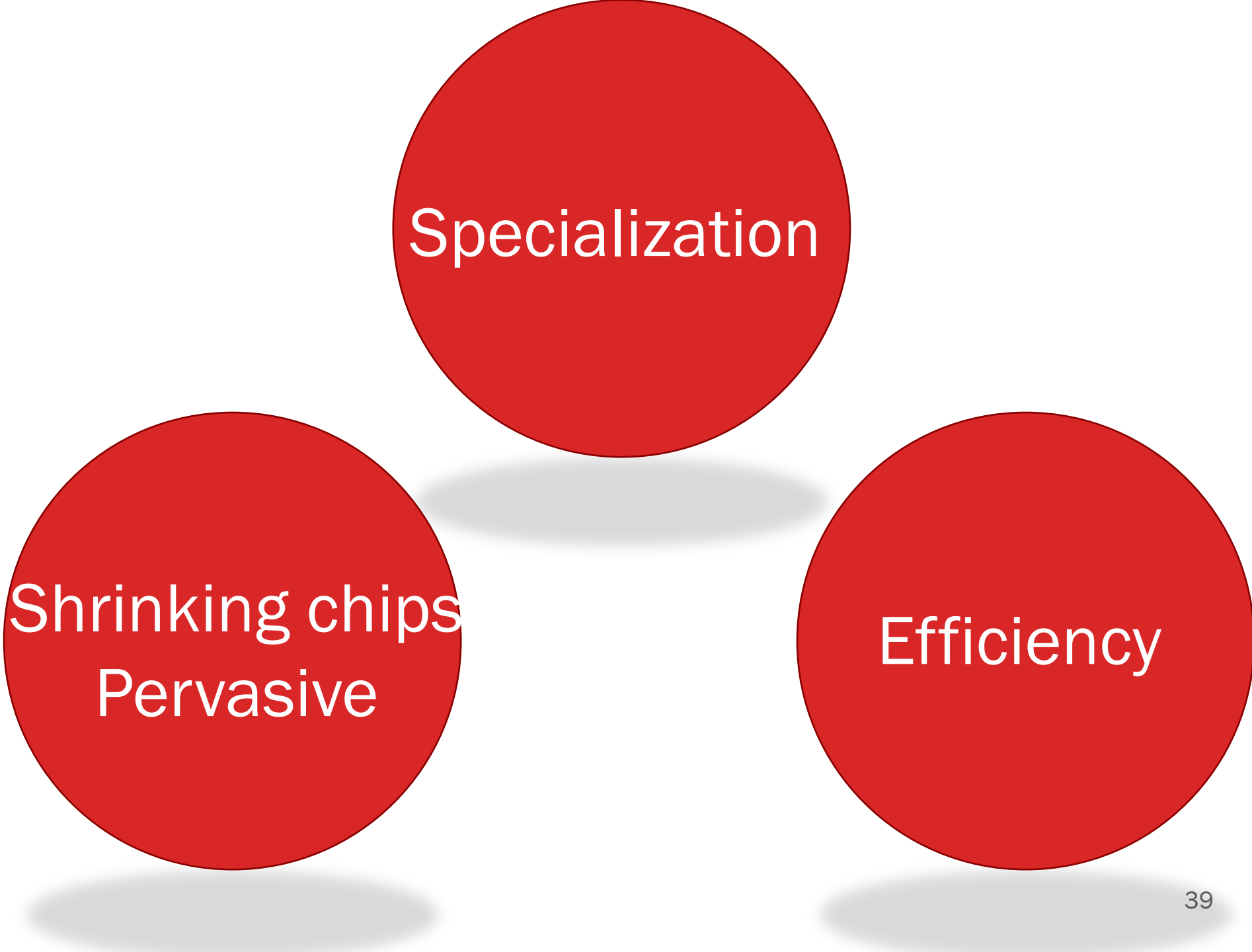


Conclusions

Multicore performance gains are limited



Need at least 18%-40% per generation from architecture alone without additional power


Specialization
Shrinking chips
Pervasive
Efficiency