Talk by
PHILIPP OTT

DARK SILICON...

...AND THE END OF MULTICORE SCALING

Appears in the Proceedings of the 38th International Symposium on Computer Architecture (ISCA '11)

Dark Silicon and the End of Multicore Scaling

Hadi Esmaeilzadeh[†] Emily Blem[‡] Renée St. Amant[§] Karthikeyan Sankaralingam[‡] Doug Burger[◦]
[†]University of Washington [‡]University of Wisconsin-Madison
[§]The University of Texas at Austin [◦]Microsoft Research
hadianeh@cs.washington.edu blem@cs.wisc.edu stamant@cs.utexas.edu karu@cs.wisc.edu dburger@microsoft.com

EXECUTIVE SUMMARY

Problem:	Dennard- and multicore-scaling don't work anymore
	Transistor underutilisation is underrated
Goal:	Predict how technology-scaling will affect transistor-underutilisation
Method:	Model future chip-development
	Calculate future transistor underutilisation
Result:	Power- & multicore-scaling won't sustain Moore's Law due to transistor underutilisation

OUTLINE

I. THE PROBLEM/ NOVELTY

THE MODEL

THE RESULTS

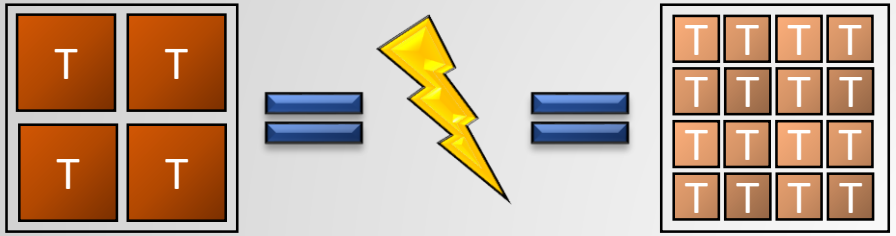
II. THE GOOD / THE IMPROVABLE

TAKE-AWAYS / THE FUTURE

III. QUESTIONS / DISCUSSION

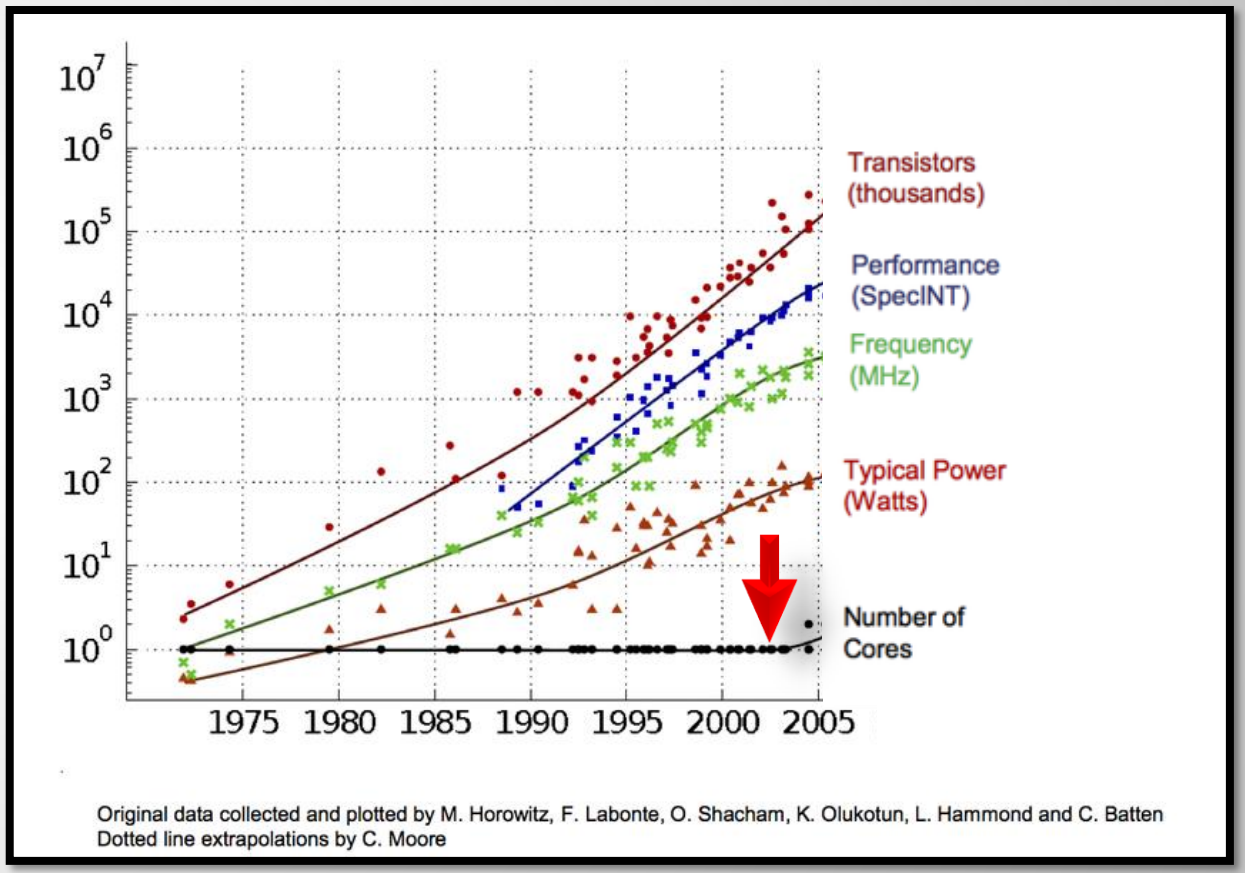
THE PROBLEM / THE NOVELTY

- Dennard scaling:



  - Broke down ~2006

- Solution: Multicore-scaling
  - Exploit parallelism by adding more cores
  - Keeping Moore's Law alive



Source: Benchmarking Adiabatic Quantum Optimization for Complex Network Analysis

We aren't getting the performance-gains, we'd expect – but why?

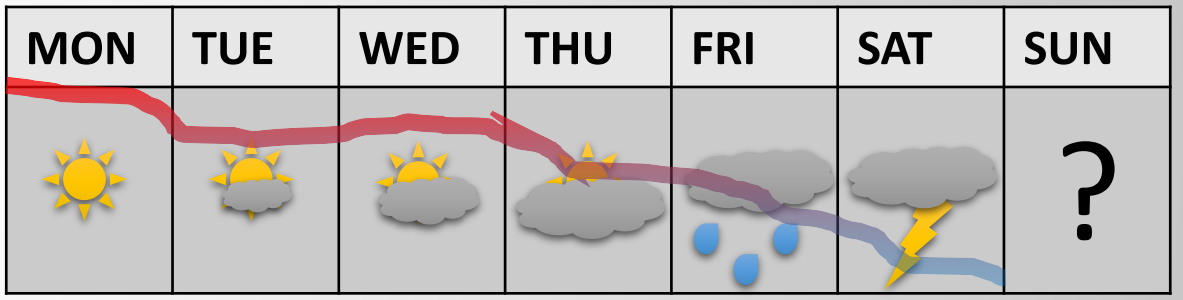
Dark Silicon – […]*the amount of circuitry of an integrated circuit that cannot be powered-on at the nominal operating voltage for a given* ***thermal design power*** *(TDP) constraint.*

- Wikipedia

Dark Silicon – *Transistors which suffer from* ***under-utilization***

AN ANALOGY

- Weather-forecast for performance-scaling
- Based on models
- Explain why the weather is getting rought



OUTLINE

I.
- THE PROBLEM/ NOVELTY
- THE MODEL
- THE RESULTS

II.
- THE GOOD / THE IMPROVABLE
- TAKE-AWAYS / THE FUTURE

III.
- QUESTIONS / DISCUSSION

THE MODEL

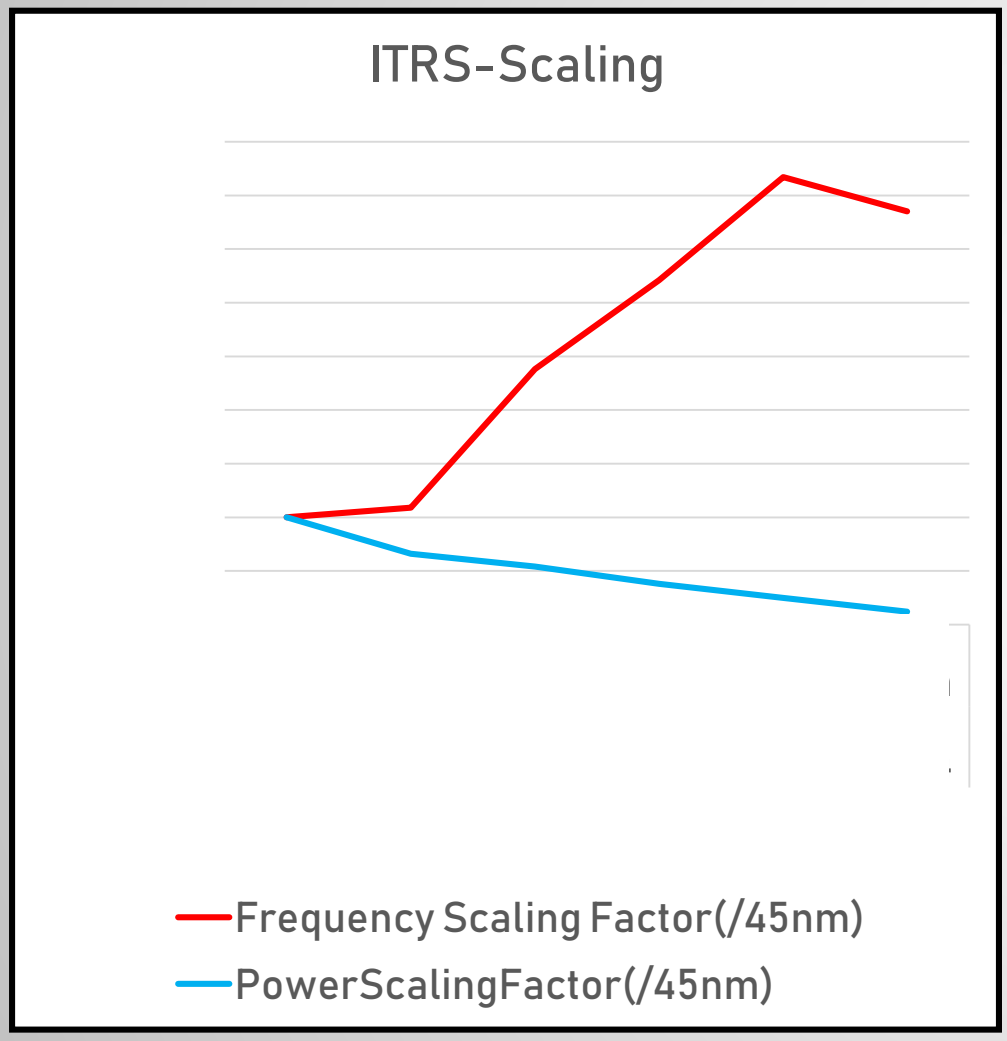
Scaling-Models

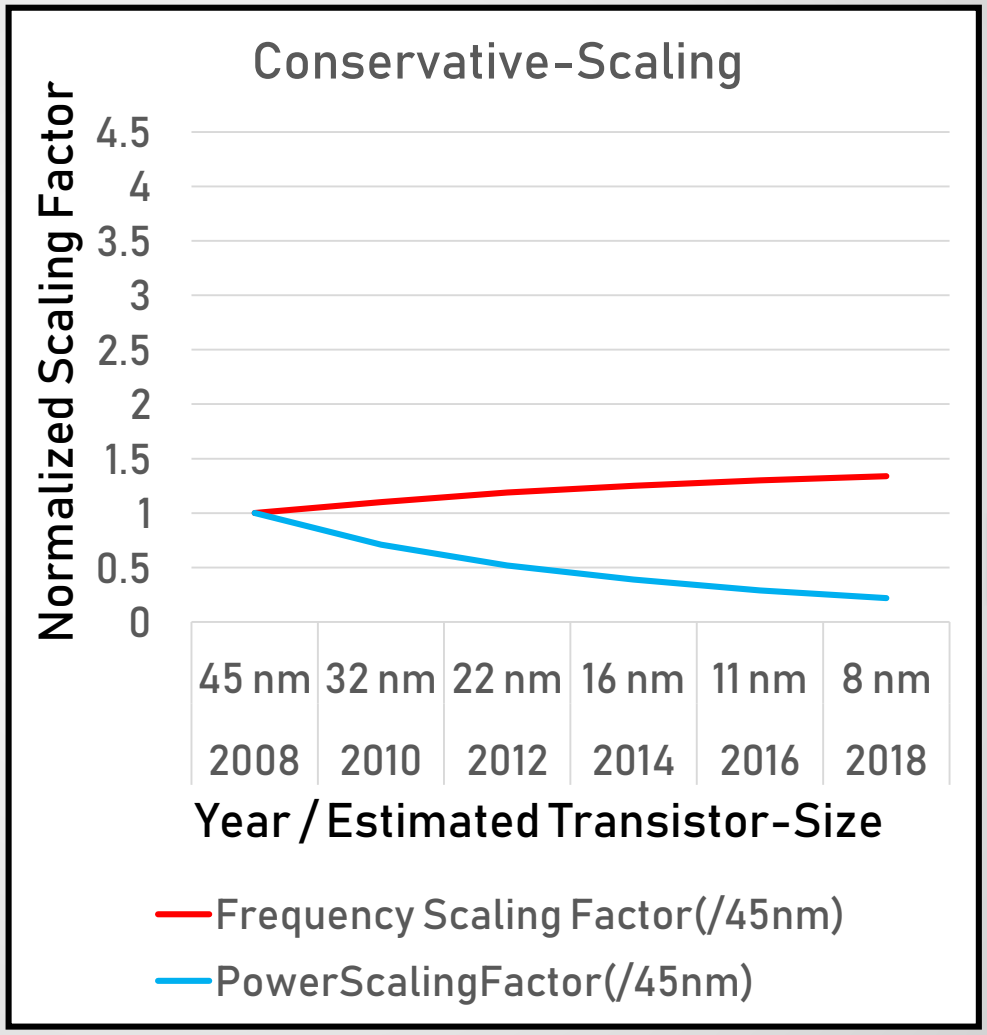
This model represents a best-case scenario!

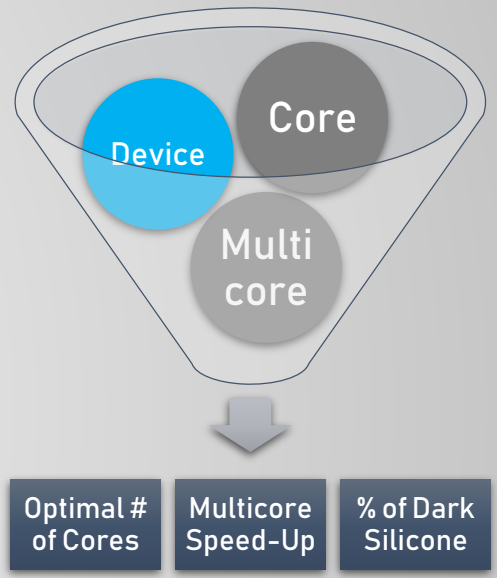
Predictions

- Optimal # of Cores
- Multicore Speed-Up
- % of Dark Silicon

DEVICE MODEL

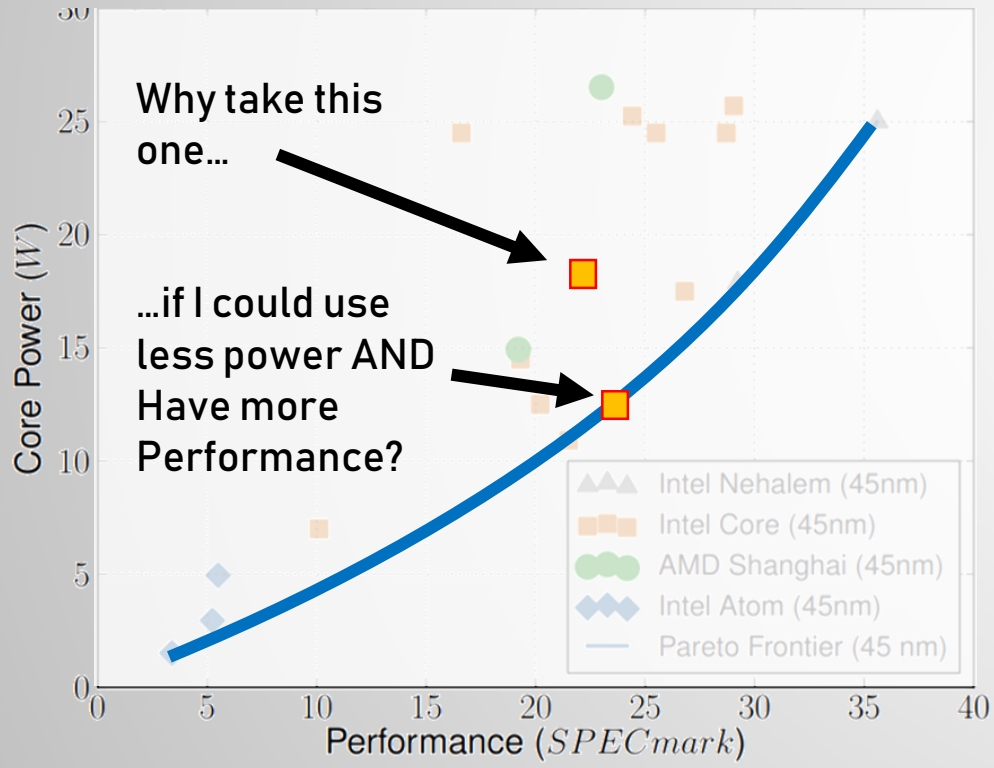




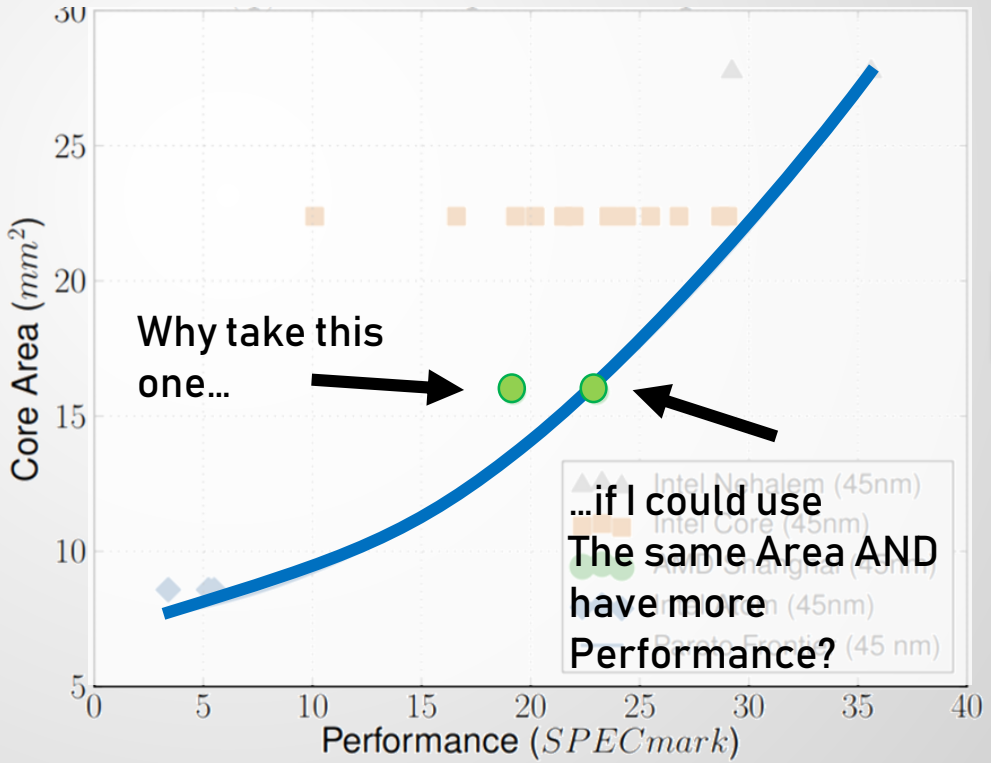


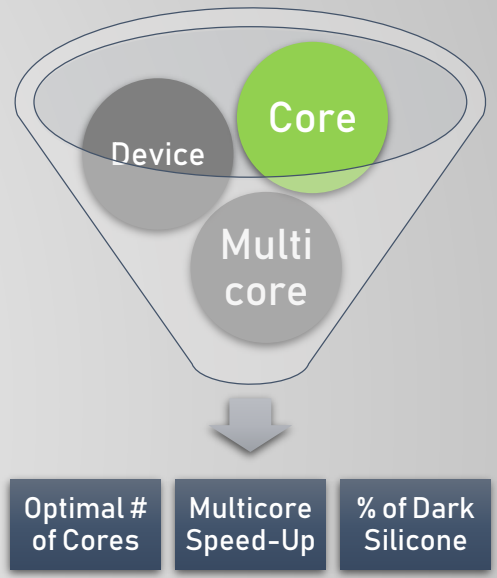
CORE MODEL

Power vs. Performance



Area vs. Performance

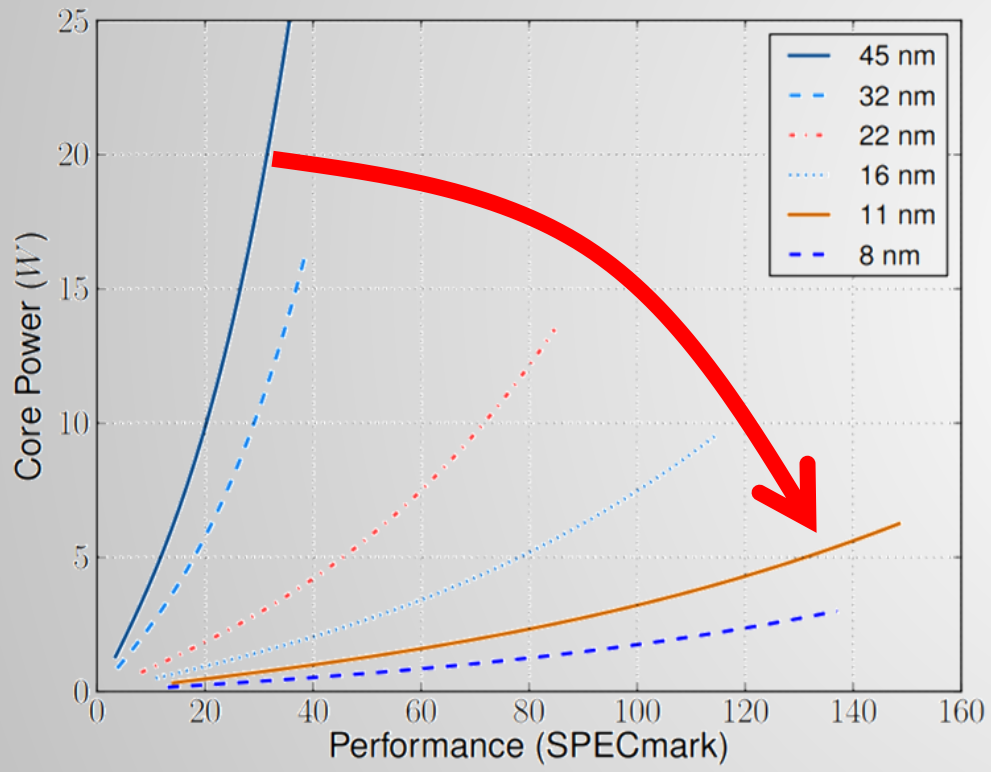




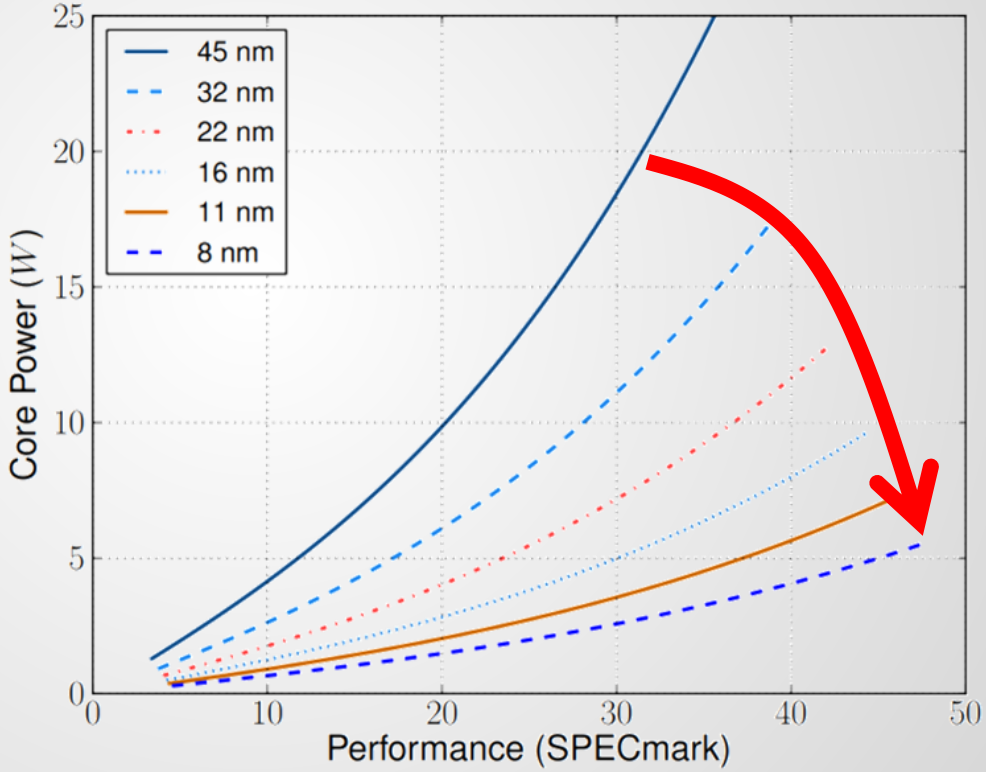
SPEC = Standard Pervormance Evaluation Corporation

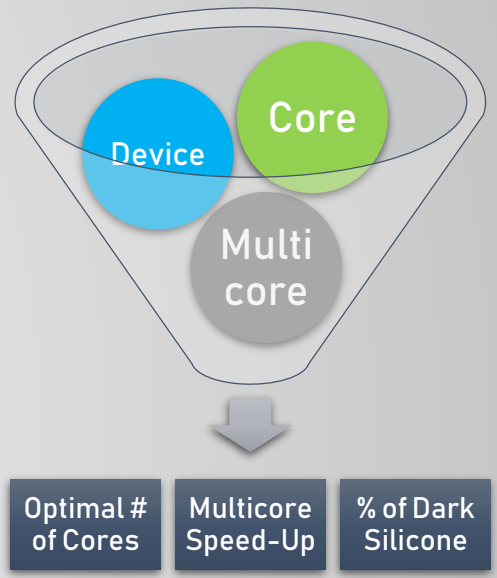
DEVICE × CORE

ITRS scaling



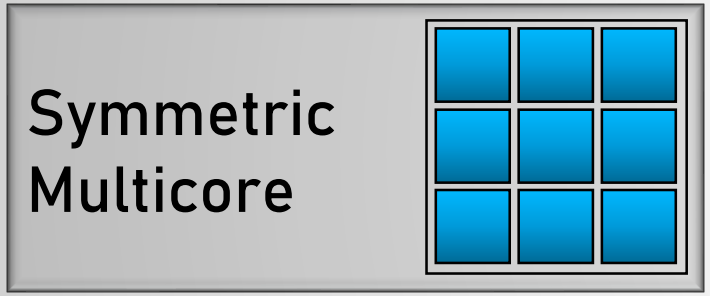
Conservative scaling

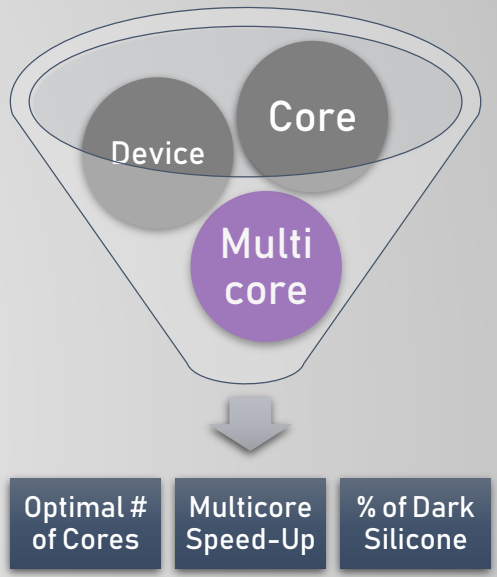




MULTICORE MODEL

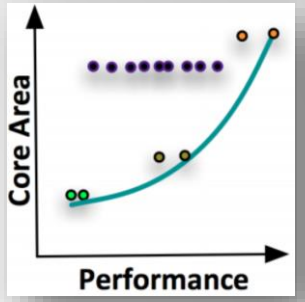
- Chip Organisations: CPU vs. GPU
- 4 microarchitectural features:

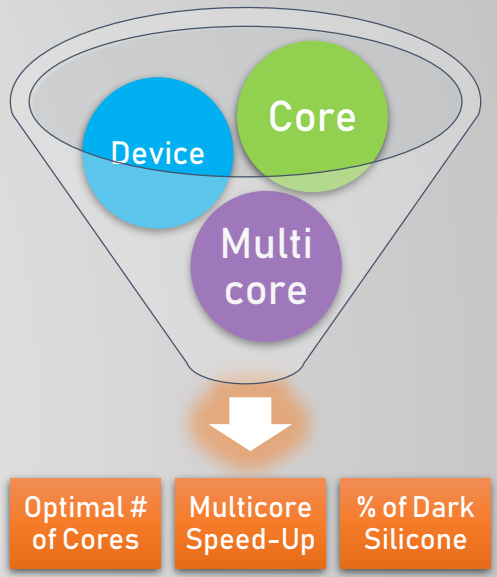




DEVICE × CORE × MULTICORE

1. All points along the area/performance Pareto-frontier are considered
2. Adding cores
3. Speed-up is computed
4. As the area or power-limit is hit, we obtain the optimal Number of cores and its speed-up
5. Dark Silicon = ChipArea – UsedArea





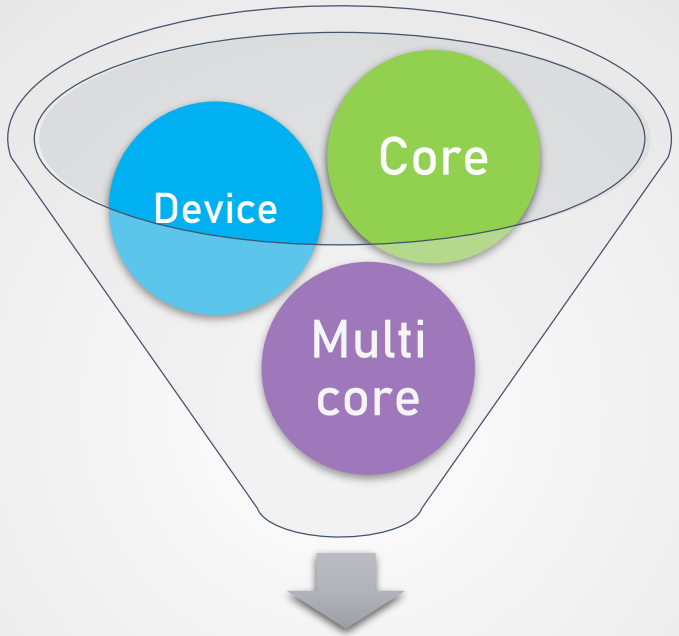
OUTLINE

I.
- THE PROBLEM/ NOVELTY
- THE MODEL
- THE RESULTS

II.
- THE GOOD / THE IMPROVABLE
- TAKE-AWAYS / THE FUTURE

III.
- QUESTIONS / DISCUSSION

THE MODEL



Scaling-Models

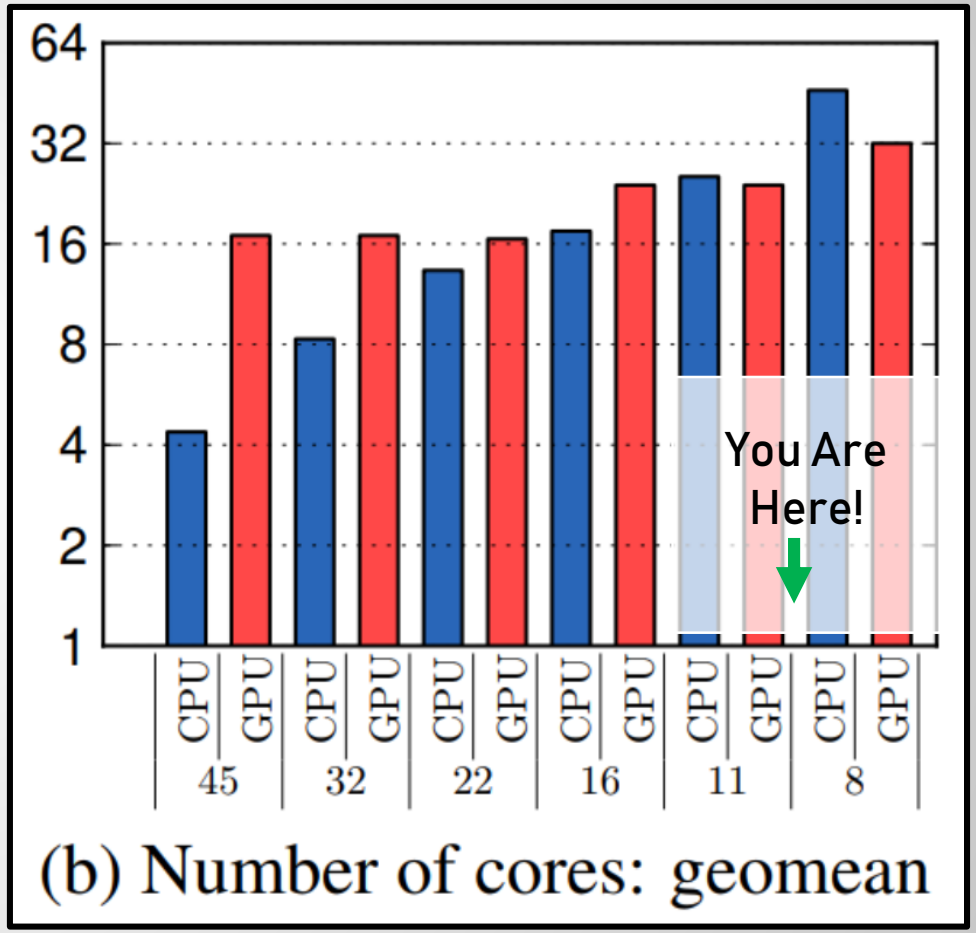
Predictions

- Optimal # of Cores
- Multicore Speed-Up
- % of Dark Silicon

RESULTS: # OF CORES

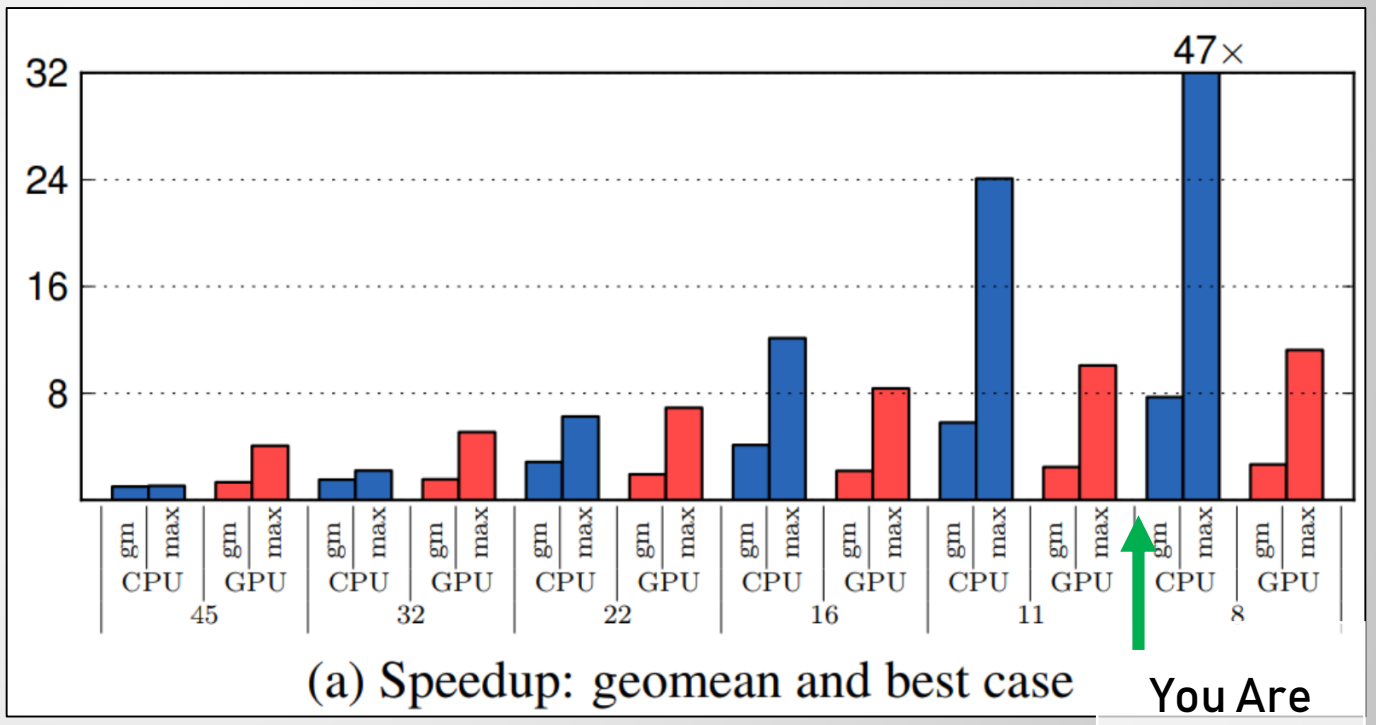
- Using Optimistic ITRS – scaling
- Tested with PARSEC
- No explosive growth in core-count

- The GPU-core-count low due to PARSEC
- For raytracing-application:
  - Transistor-size: 8 nm
  - Core count: 4864



RESULTS: SPEED-UP

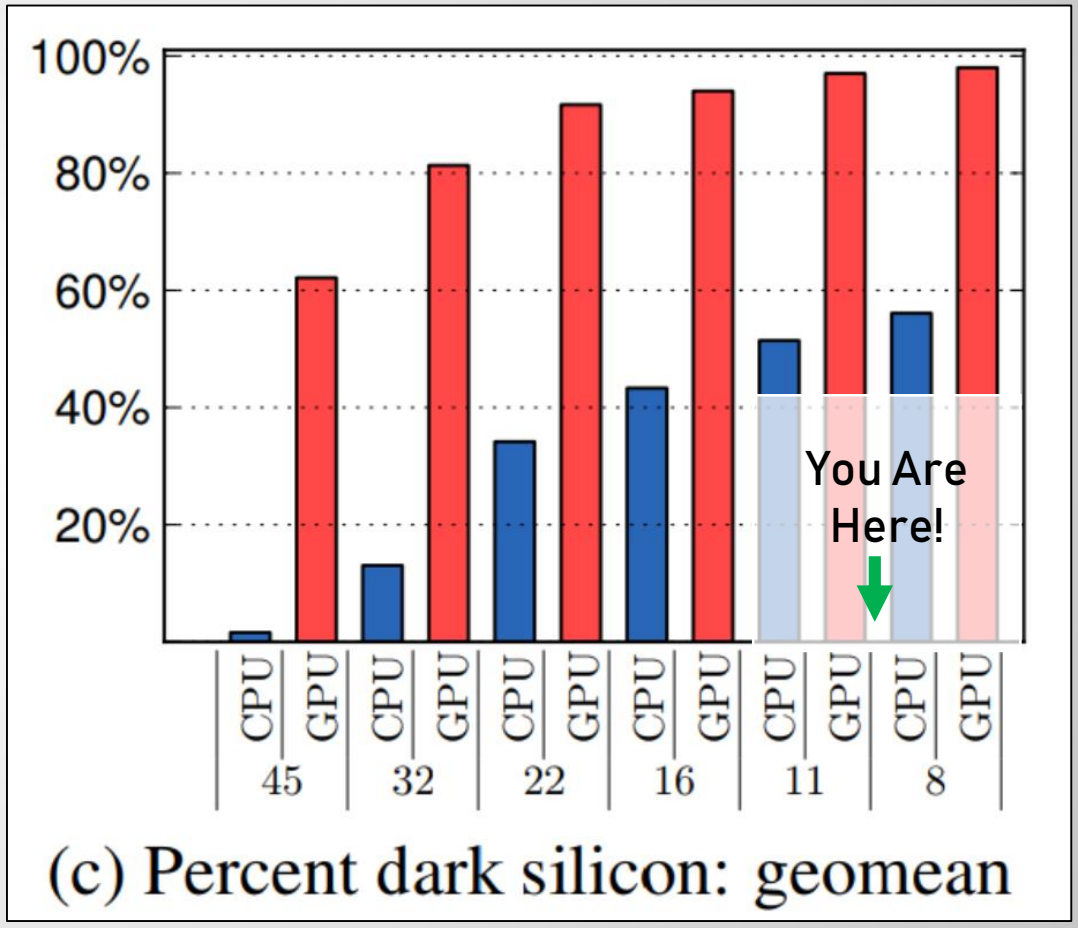
- Normalized speed-up, to
  - quadcore
  - Nehalem
  - 45 nm
- Speed-up in geomean saturates



(a) Speedup: geomean and best case

RESULTS: % OF DARK SILICON

- At 8nm:
  - CPU > 50%
  - GPU > 90% !!!
- GPU Raytracing
  - 8 nm
  - Dark silicon: 8 %

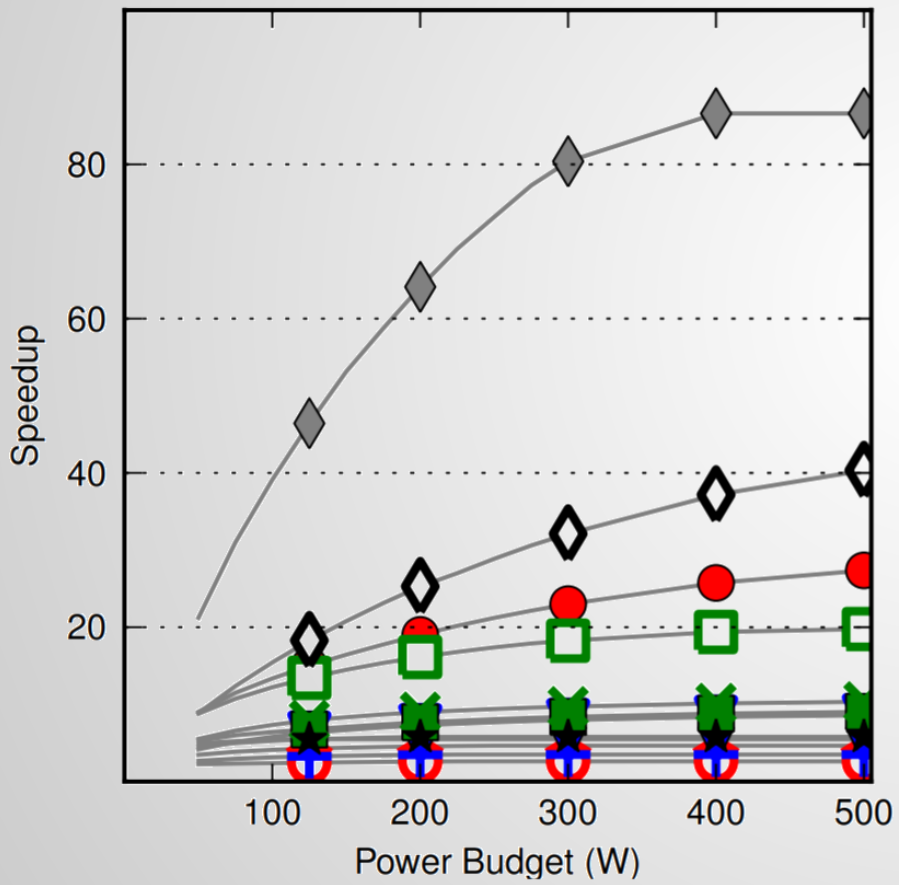


A Case for Core-Assisted Bottleneck Acceleration in GPUs: Enabling Flexible Data Compression with Assist Warps

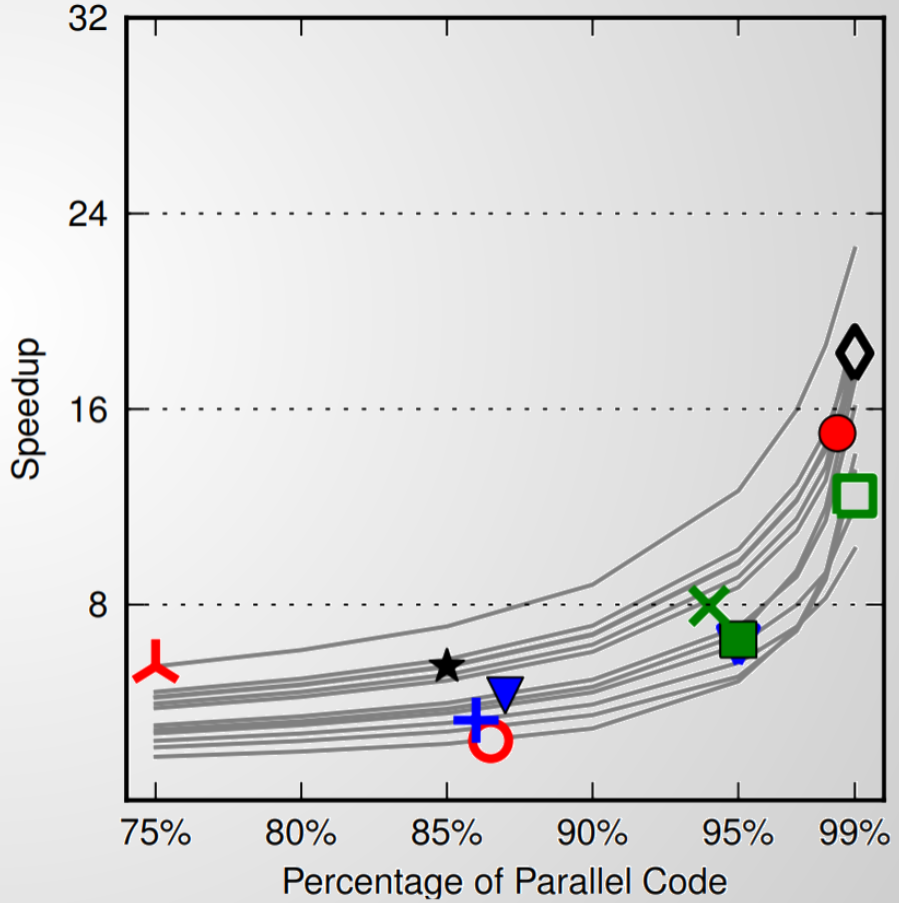
Nandita Vijaykumar Gennady Pekhimenko Adwait Jog† Abhishek Bhowmick
Rachata Ausavarungnirun Chita Das† Mahmut Kandemir† Todd C. Mowry Onur Mutlu
Carnegie Mellon University † **Pennsylvania State University**
{nandita,abhowmick,rachata,onur}@cmu.edu
{gpekhime,tcm}@cs.cmu.edu {adwait,das,kandemir}@cse.psu.edu

THE SOURCE OF DARK SILICON

SpeedUp vs. Power

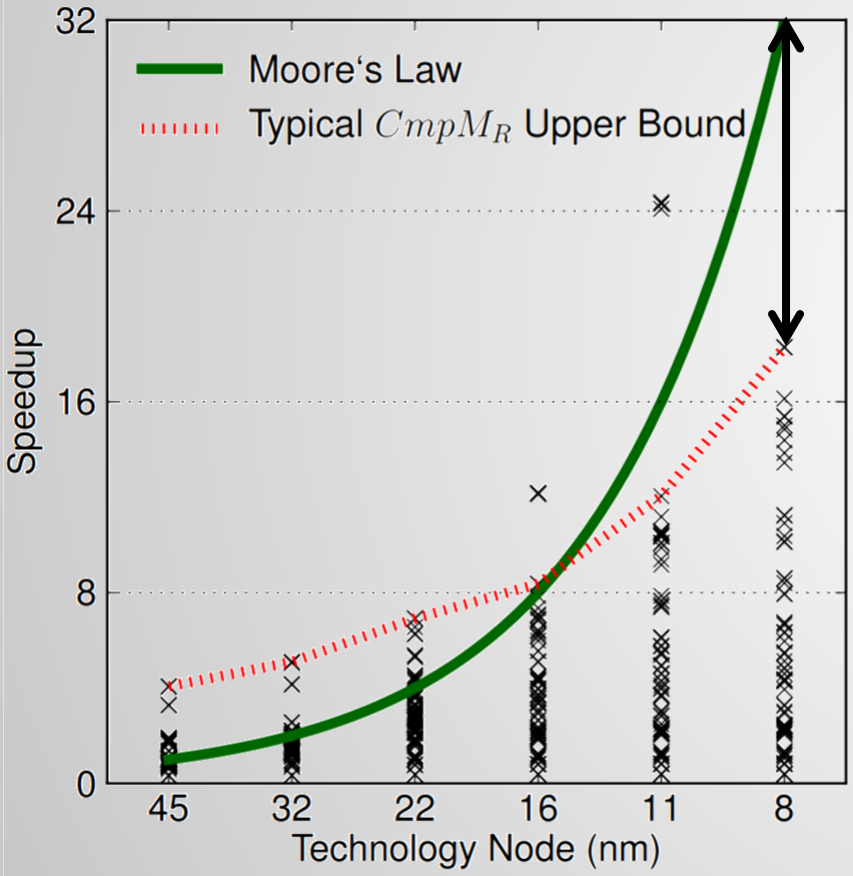


SpeedUp vs. Parallelism



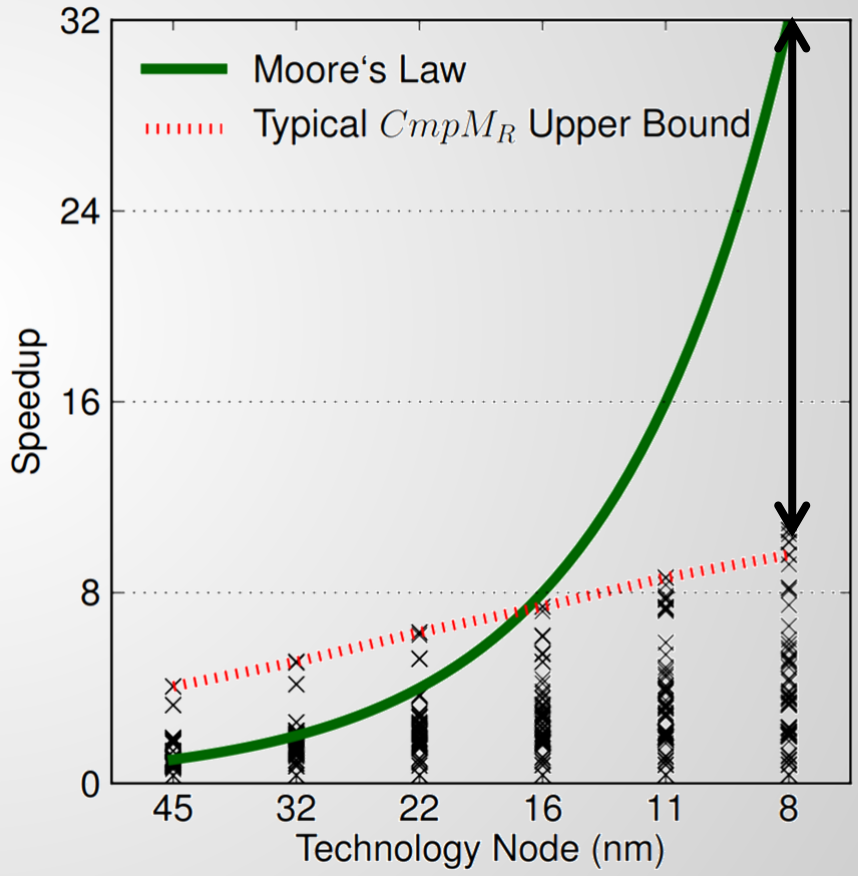
MOORE VS. REALITY

ITRS scaling



Dark Silicon

Conservative scaling



Dark Silicon

CONCLUSION

Problem:	Dennard- and multicore-scaling don't work anymore
	Dark Silicon is taking over
Goal:	Find the source of dark silicon
	Predict the development of core-numbers, performance and dark silicon
Method:	Model future chip-designs
	Calculate future dark silicon from model
Result:	Predictions toward's % of dark silicon, # of cores and speedUp
	Power and multicore-scaling is limited by dark silicon
	Limited parallelism is the primary source of dark silicon

OUTLINE

I.
- THE PROBLEM/ NOVELTY
- THE MODEL
- THE RESULTS

II.
- THE GOOD / THE IMPROVABLE
- TAKE-AWAYS / THE FUTURE

III.
- QUESTIONS / DISCUSSION

THE GOOD

- Big impact on the industry
- Shows where the problems are and how they will develop
- Covers various architectures and applications
- Well structured paper

THE IMPROVABLE

- Number-of-cores for GPU is a somehow sparse prediction
- ARM was not considered in this model
- Calculation of Dark Silicon percentage is explained in only one sentence
- Is Dark Silicon the sole enemy of Moore's Law?
- More information to replicate test-environment
- How would it behave in an multi-application-environment?

OUTLINE

I.
- THE PROBLEM/ NOVELTY
- THE MODEL
- THE RESULTS

II.
- THE GOOD / THE IMPROVABLE
- TAKE-AWAYS / THE FUTURE

III.
- QUESTIONS / DISCUSSION

TAKE-AWAY'S

- Moore's and Dennard's Law will no longer apply
- Multicore-scaling might not be the answer for everything
- Limited parallelism is at least as problematic as power-constraints

BEYOND THE PAPER: 4 HORSEMEN

THE SHRINKING	THE DIMMED	THE SPECIALICED	THE EX-MACHINA
«Simply remove» Dark Silicon physically	Underclock or only use in bursts	Build specialized cores, and only turn on the ones we need	Beyond Silicon

Graphics & content: http://darksilicon.org/

OUTLINE

I.
- THE PROBLEM/ NOVELTY
- THE MODEL
- THE RESULTS

II.
- THE GOOD / THE IMPROVABLE
- TAKE-AWAYS / THE FUTURE

III.
- QUESTIONS / DISCUSSION


X

DISCUSSION

X

- What new technologies come to mind, that might prevent the end of performance-scaling?
- Philosophical: What if we hit the end? What might be the consequences?
- What could Neuro-Science tell us?

Appears in the Proceedings of the 38th International Symposium on Computer Architecture (ISCA '11)

Dark Silicon and the End of Multicore Scaling

Hadi Esmaeilzadeh[†] Emily Blem[‡] Renée St. Amant[§] Karthikeyan Sankaralingam[‡] Doug Burger[◦]
[†]University of Washington [‡]University of Wisconsin-Madison
[§]The University of Texas at Austin [◦]Microsoft Research
hadianeh@cs.washington.edu blem@cs.wisc.edu stamant@cs.utexas.edu karu@cs.wisc.edu dburger@microsoft.com

Thank you and Happy Halloween!

...and Thank you Giray and Geraldo :D

Beyond MOSFET's

X

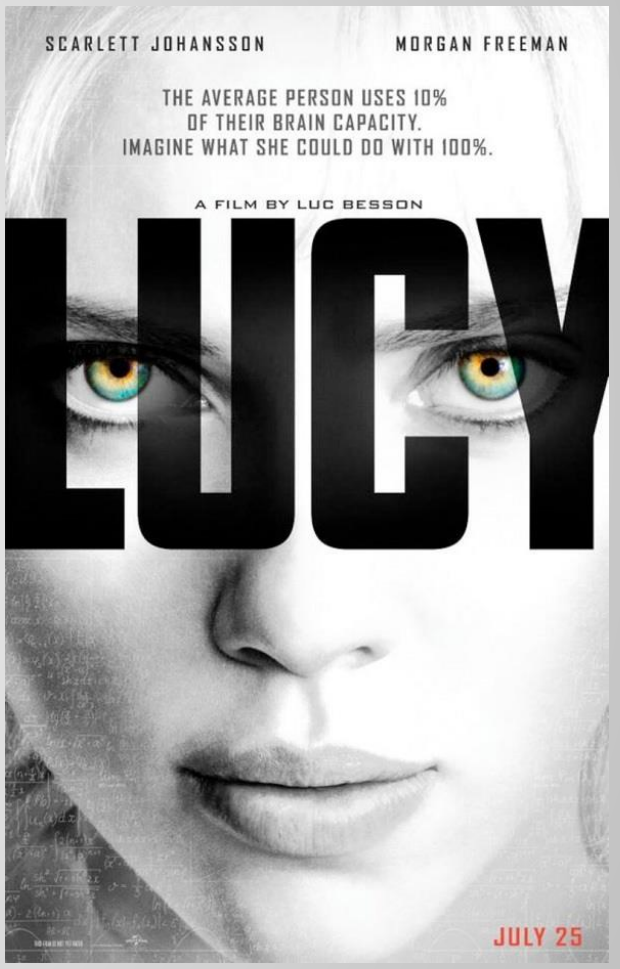
Silicon-in-insulator	Vertical replacement gate FET	GaN MOSFET	Neuro-Informatics
Silicon-on-"nothing"	Ballistic FET	Superconductive FET	And many more!
Double-gate FETs	Tunneling FET	Quantum	
FinFETs	CN-FET	Graphene	
Vertical FETs	MESFET	Nano-Tubes	

Source: http://darksilicon.org/horsemen/horsemen_slides.pdf

PARALLEL TO THE HUMAN BRAIN

- 100 trillion synapses
- Embody an existence proof of highly parallel, mostly dark operation

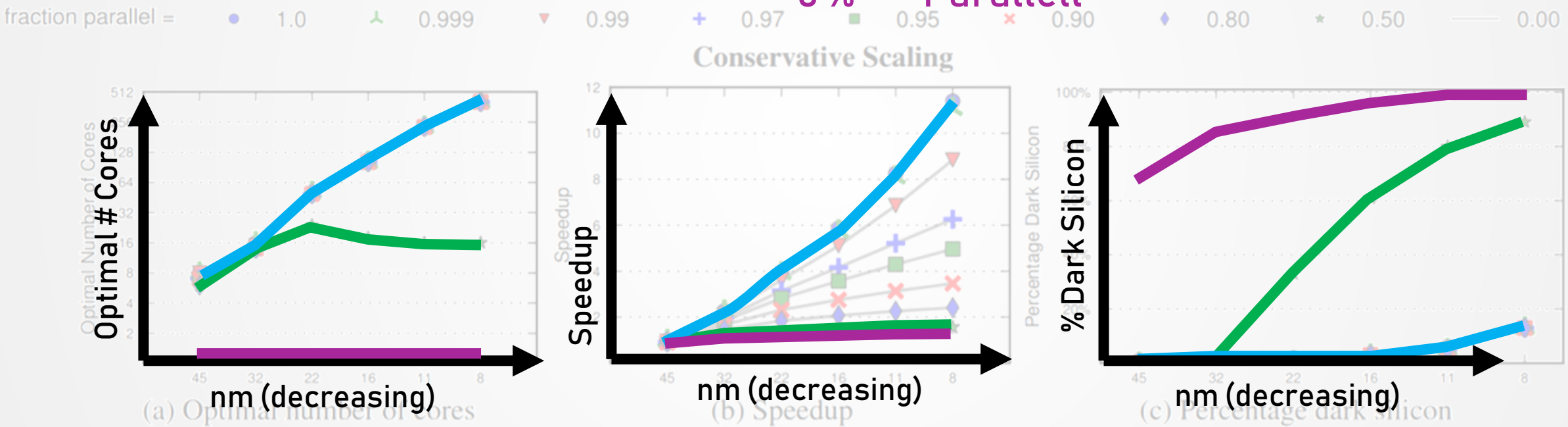
Source: http://darksilicon.org



AMDAHL'S LAW

X

100 % Parallell
50 % Parallel
0 % Parallell



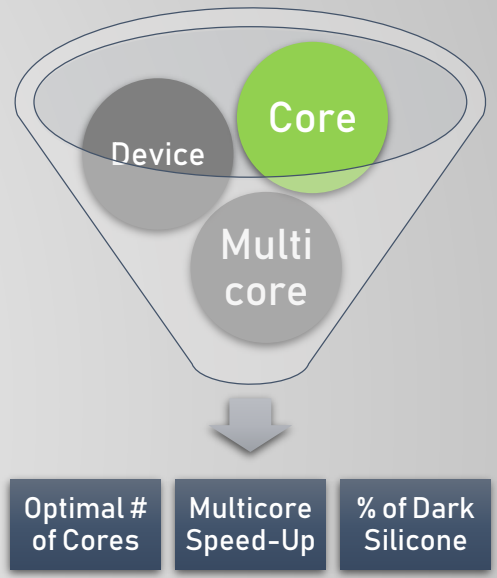
Source Graphics and Data: Presented Paper

CORE MODEL – (DITCHED)

- Pollacks Rule:

$$\Delta Performance \cong \sqrt{\Delta Area}$$

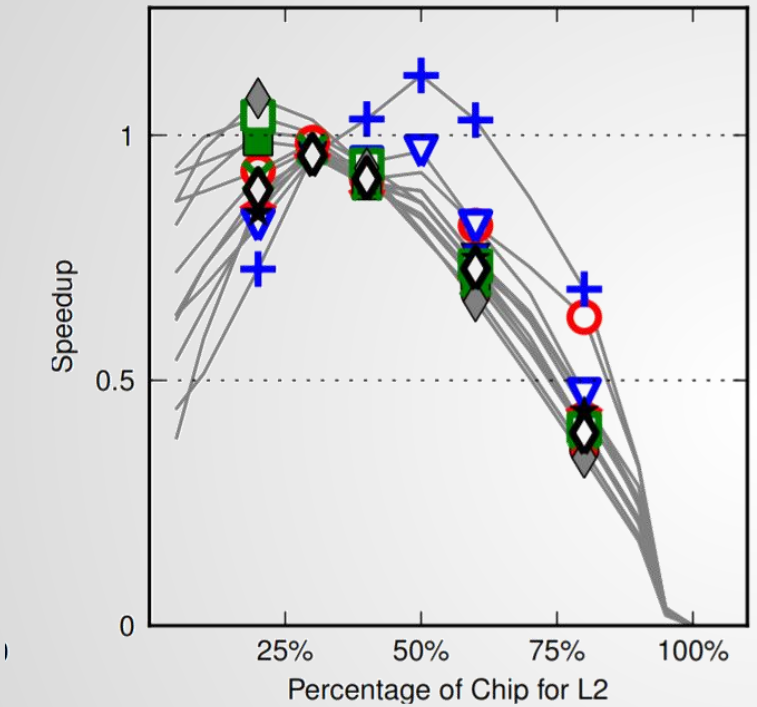
- Power is no longer only constrained by area



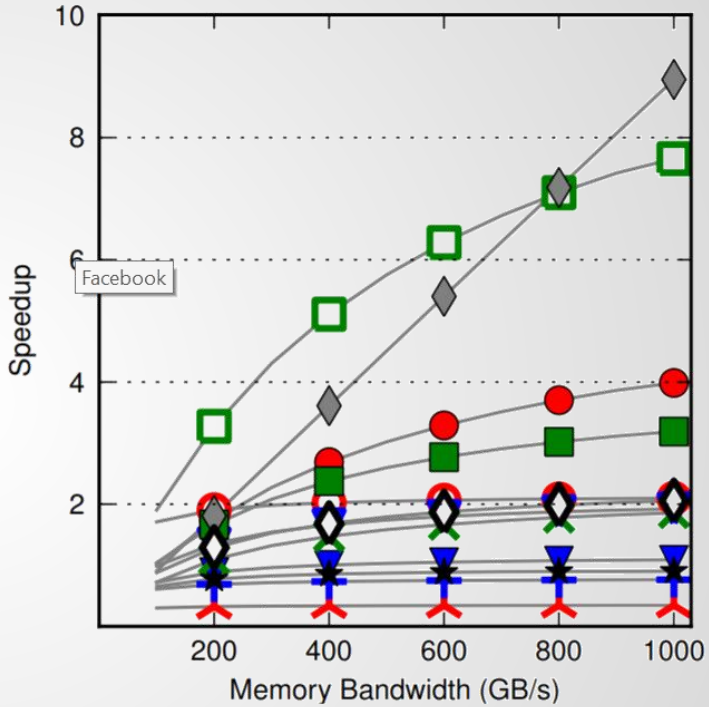
- Empirical Data from 152 Processors
- Deriving Pareto-Frontiers

WORK-AROUNDS

X



(a) L2 size (CPU)



(b) Memory bandwidth (GPU)

Figure 7: Sensitivity studies of L2 size and memory bandwidth using symmetric topology at 45 nm