



Architectures for Wafer-Scale Integration

Rob Aitken, Arm Research

UCLA CHIPS Workshop
Nov 6, 2019

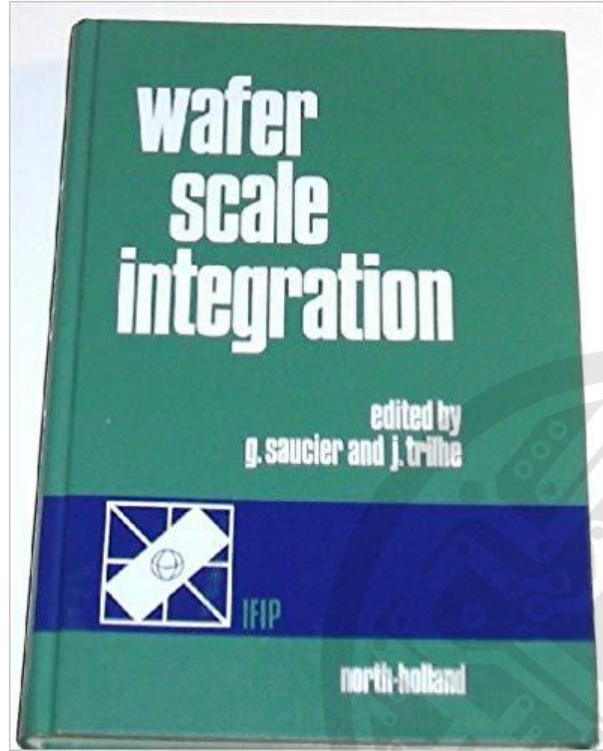
Outline

- Why wafer scale?
- Drivers, history
- Implications of wafer scale
- Architecture possibilities

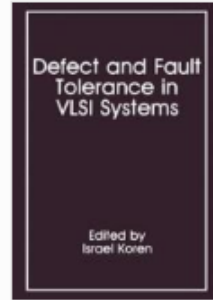


CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

History



© 1986



[Defect and Fault Tolerance in VLSI Systems](#) pp 327-338 | [Cite as](#)

Defect Tolerance in a Wafer Scale Array for Image Processing

Authors

Authors and affiliations

G. Saucier, J.-L. Patry, E.-F. Kouka, T. Midwinter, P. Ivey, M. Huch, M. Glesner

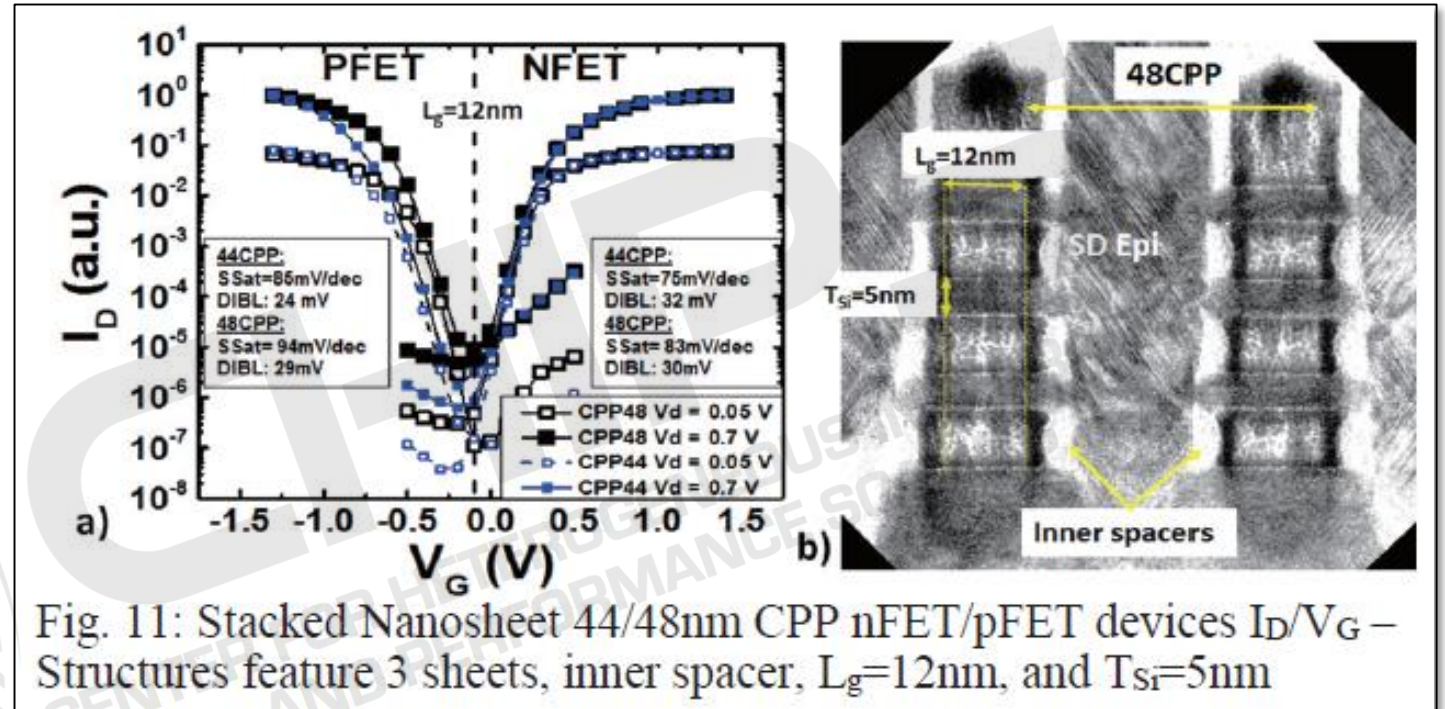
© 1989

Amdahl, who raised \$200 million to fail with his massive brute force WSI programme in his company Trilogy, was so revered for this failure, that he was then flown over to England by the TEE to tell us that a five processor machine ran no faster than a single processor machine, concluding that there was no point in having more than one processor.

- Ivor Catt, "Dinosaur Computers", Electronics World, June 2003

Back to the future

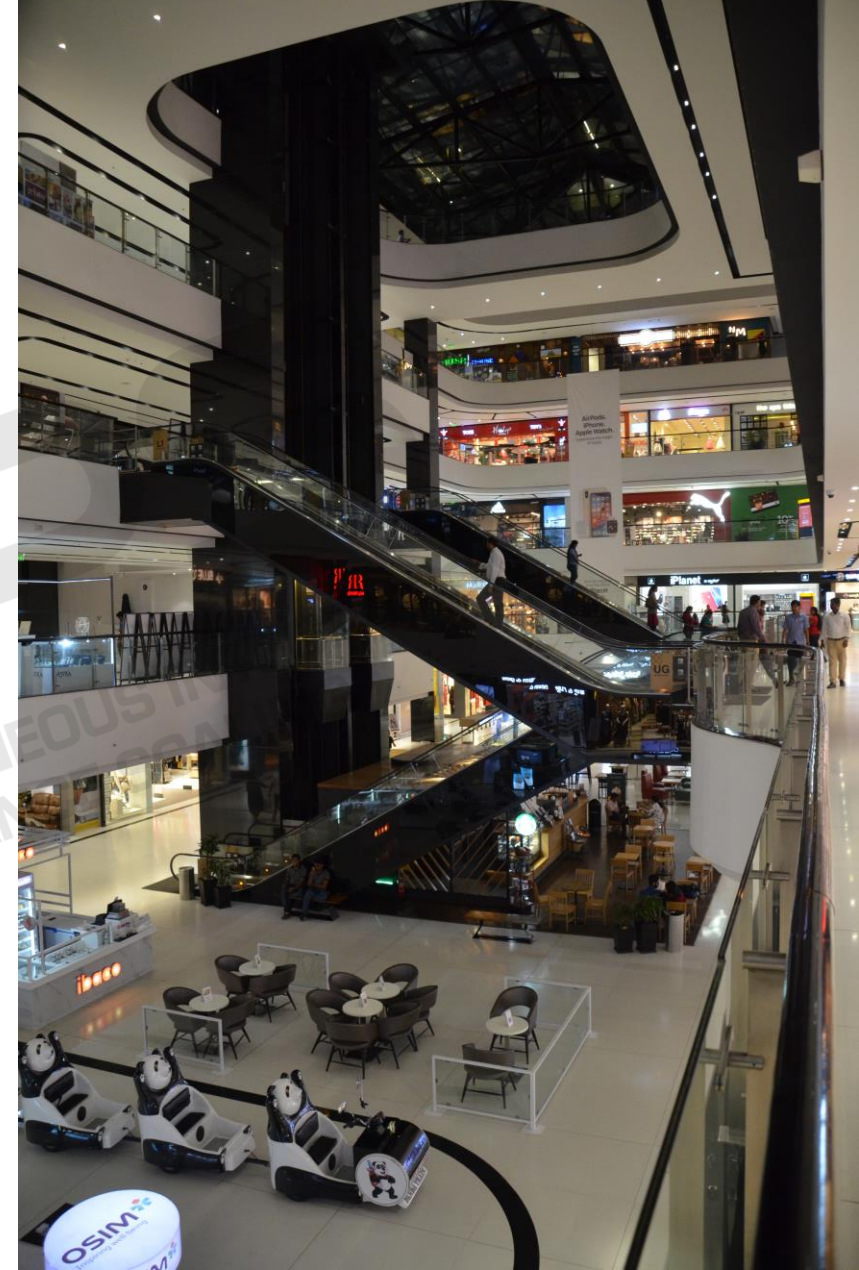
- You can't shrink atoms
 - 5nm = ~25 silicon atoms
- Existing oxides ~1nm
 - 1 layer of silicon oxide
 - 1 layer of high K material



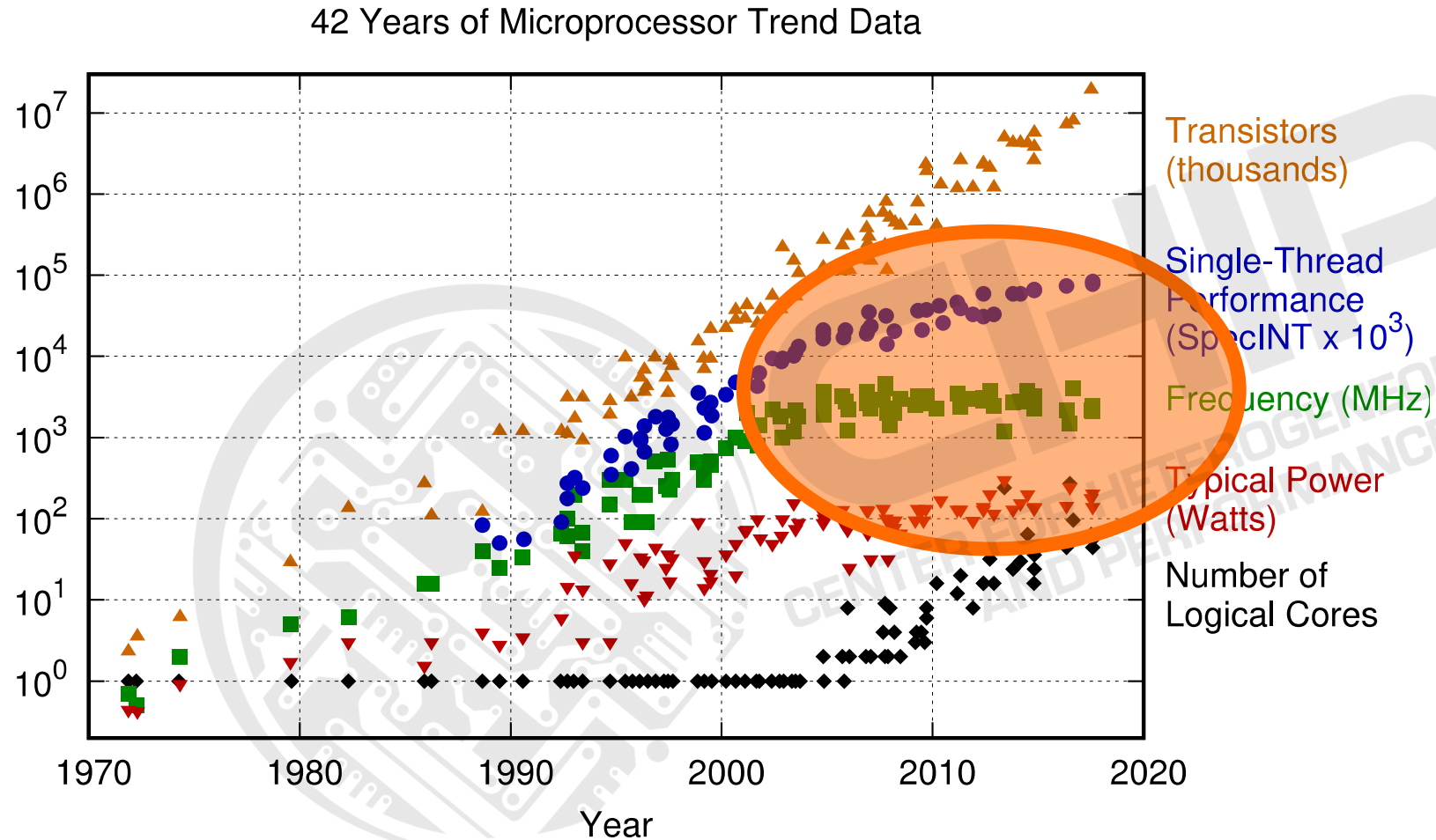
Loubet et al, IEDM 2017

Evolution of care-about

- Mainframe era: performance
 - Number crunching
- PC era: performance/cost
 - Graphics
- Mobile era: performance/(cost*energy)
 - Video
- Now: (performance***capability**)/(cost*energy)
 - Machine learning



Trends

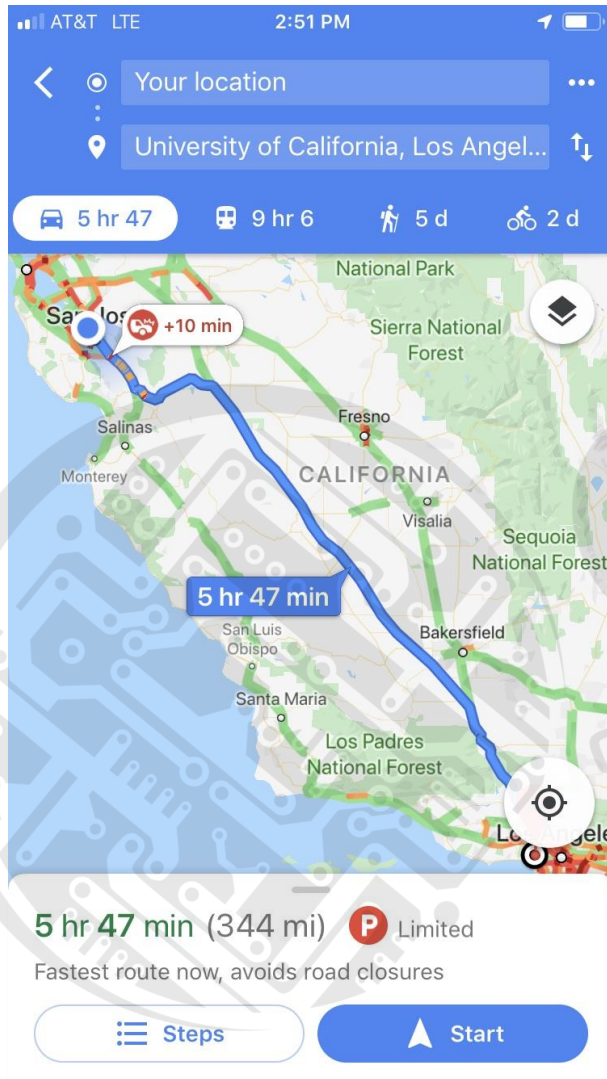


- Single thread performance growing slowly
- Gains coming from vector and other parallel operations

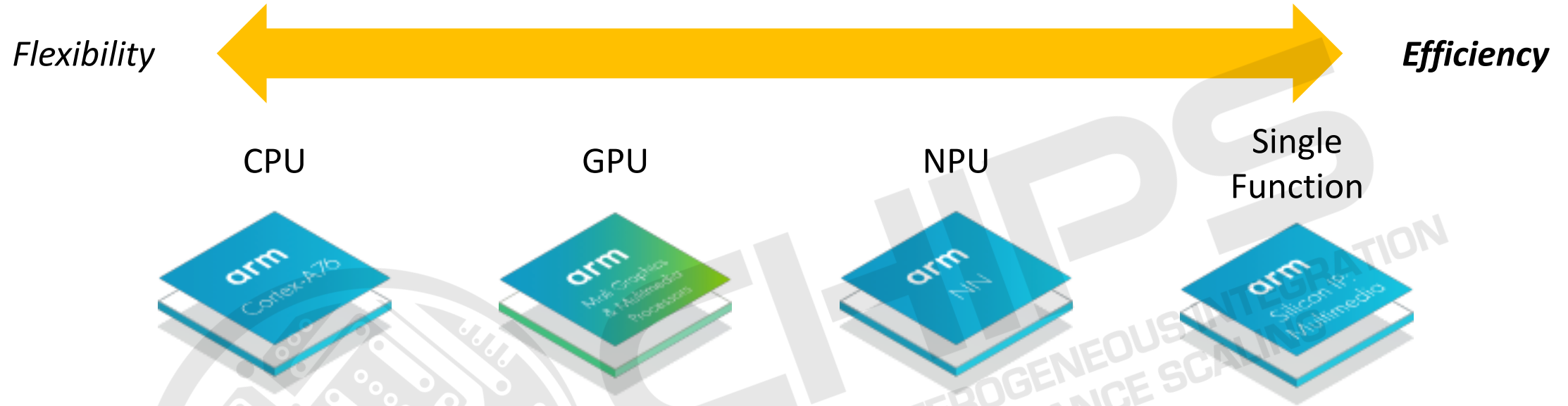
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Source: <https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/>

Metrics digression

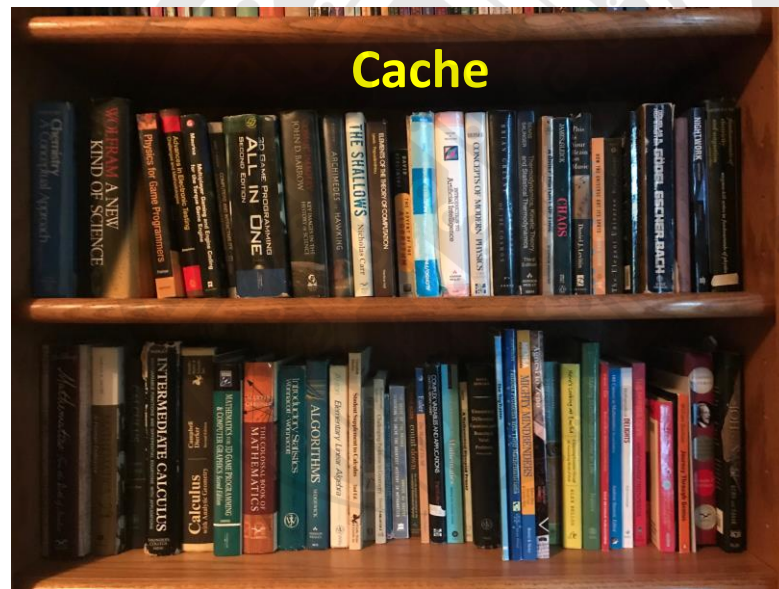
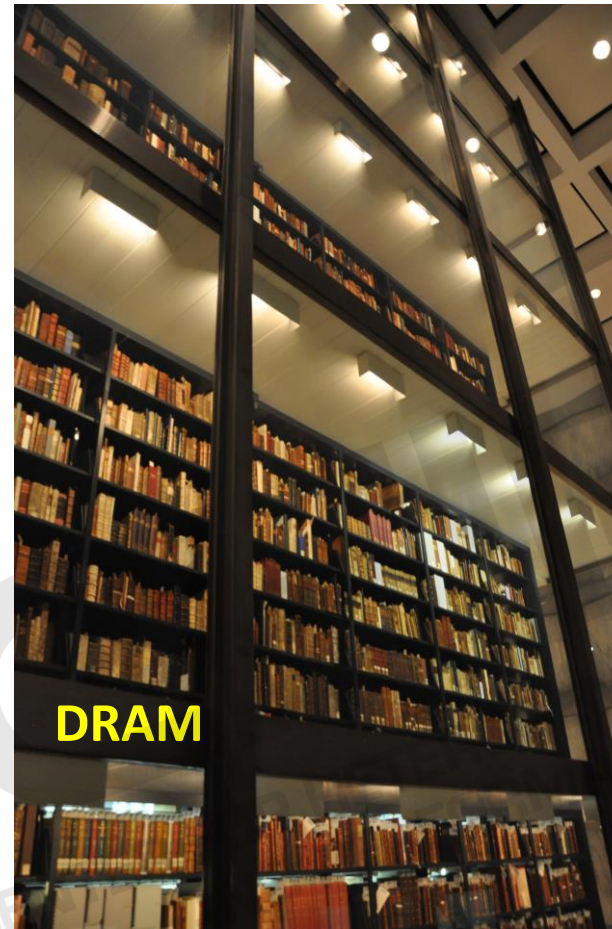
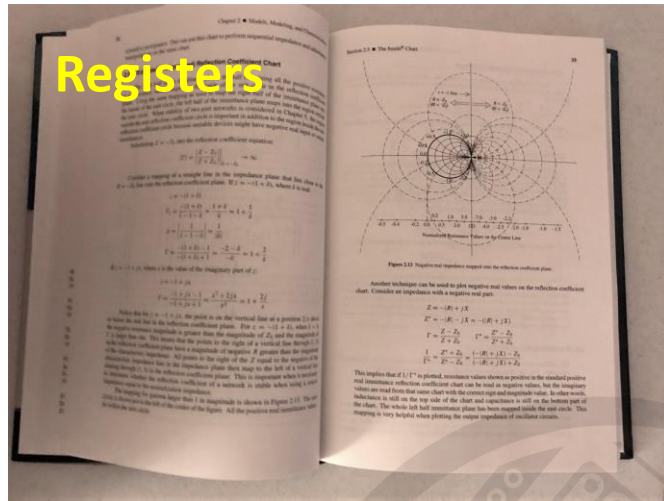


Hardware architecture specialization



- “Ease of programming” often follows “flexibility”
- How do we allow accelerators and processors to share data efficiently?
 - Coherency regimes
 - Standards

Memory at human scale



Problem scale

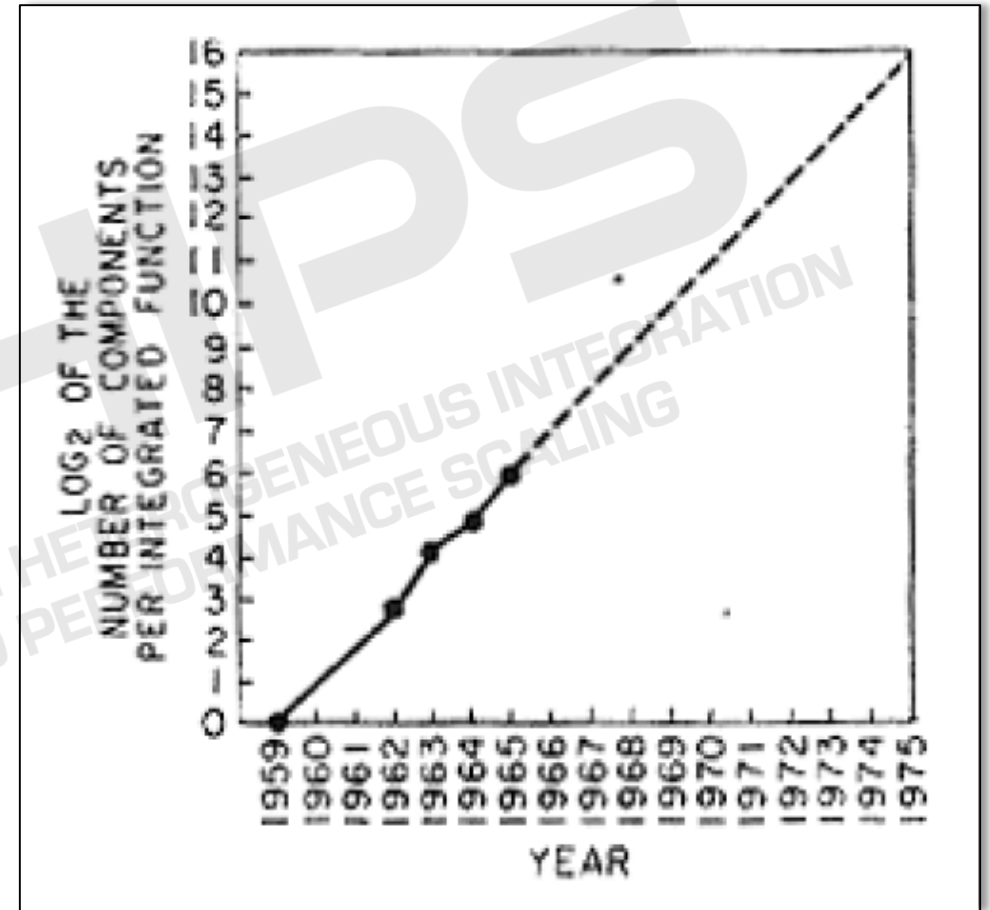
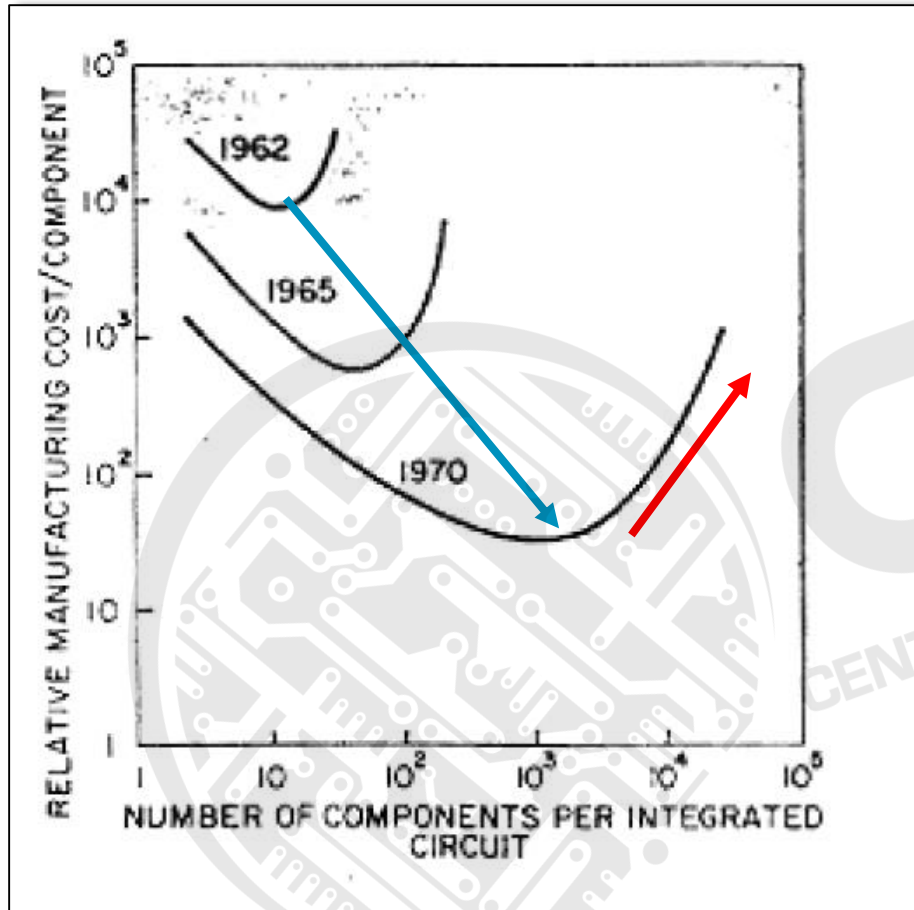


It's all about the memory system

- Specifically, it's all about data transport versus compute (or re-compute) cost



Figure 1 in Moore's paper...



Others may have thought of this before...

- From CHIPS web page
- Get rid of packages and use a silicon substrate for dense high-performance interconnect

Mega SolFs by re-integration on an Interconnect Fabric



The "right" interconnect fabric

- Mechanically robust (flat, stiff, tough...)
- Capable of fine wiring, fine pitch interconnects
- Thermally conductive
- Can have active and passive built-in components

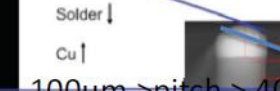
Silicon Fits the Bill in many cases

Challenges:

- Warpage
- Topography
- Assembly /Thru'put
- Thermal



>100 μm pitch
Mass reflow



100 μm > pitch > 40 μm
Mass re-flow+TCB



Full contact – TCB
Proximate
• Inductive
• Capacitive

UCLA ENGINEERING
Henry Samueli School of
Engineering and Applied Science
Birthplace of the Internet



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

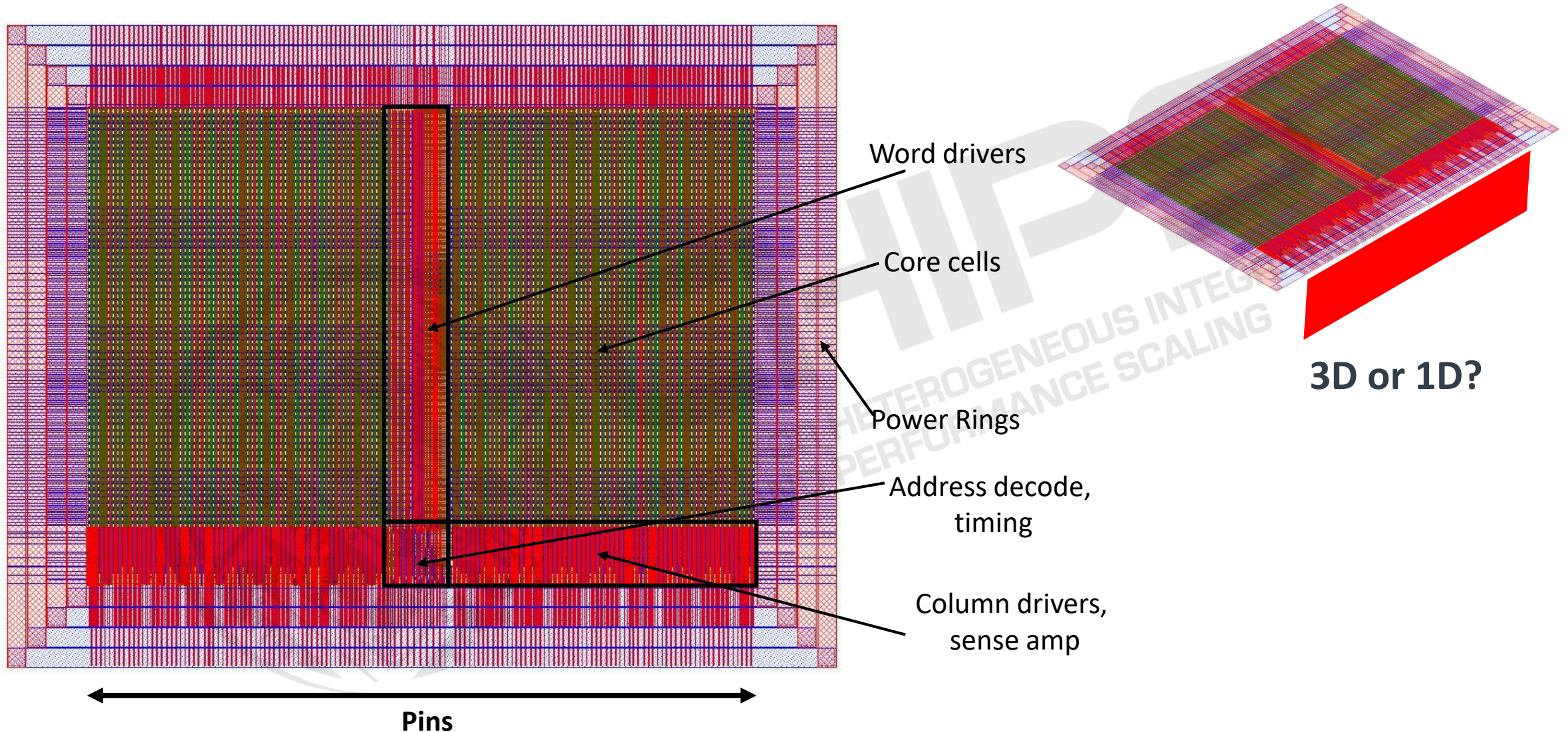
3D vs 2.5D

- Stacked 2D objects are not 3D
- Limited freedom to move vertically
- Wafer scale is localized pseudo-3D combined with massive global amounts of 2.5D

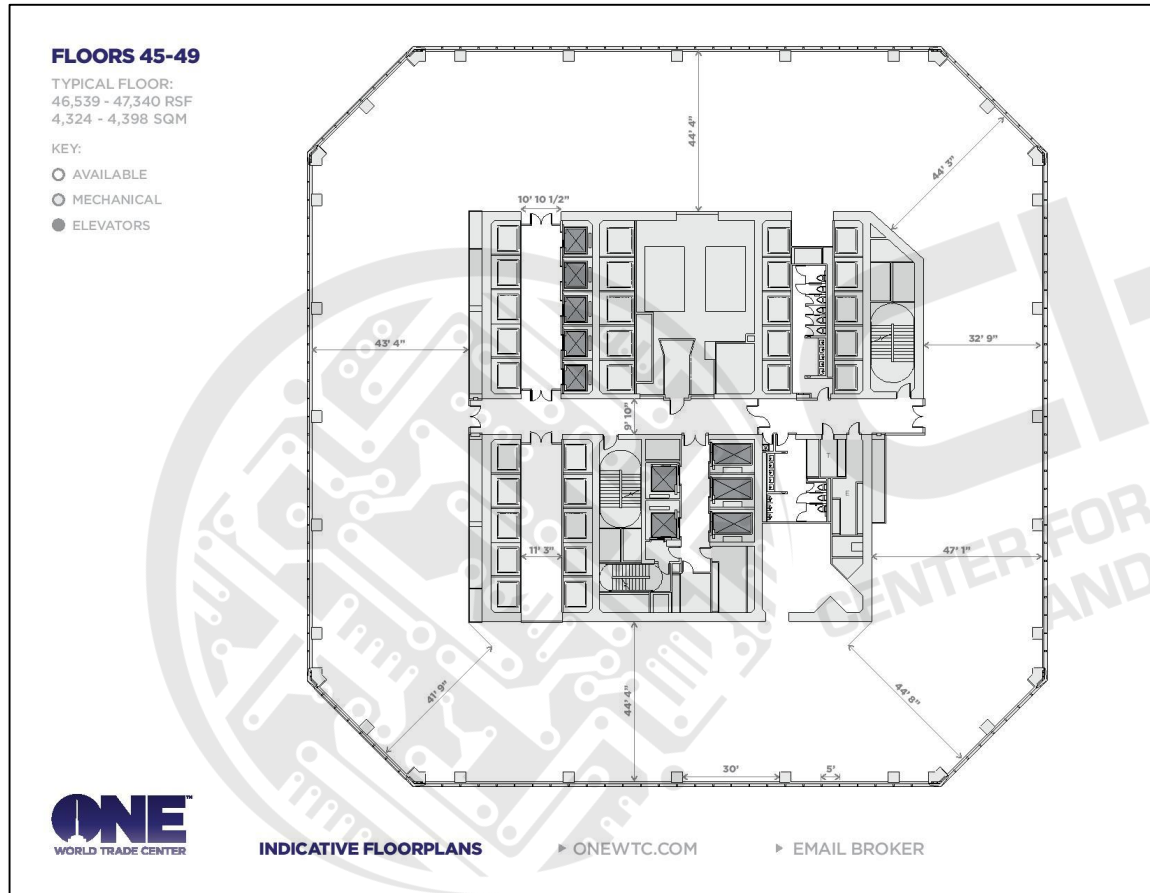


Source: Amazon.com

Pins on the edge are the Amdahl's law of 3D



3D is not about the layers, it's about the interconnect

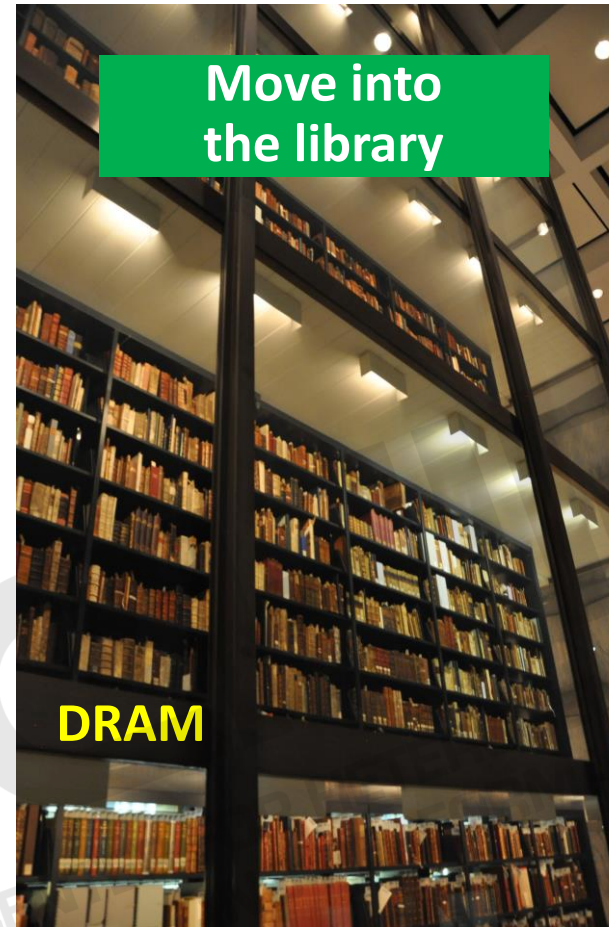
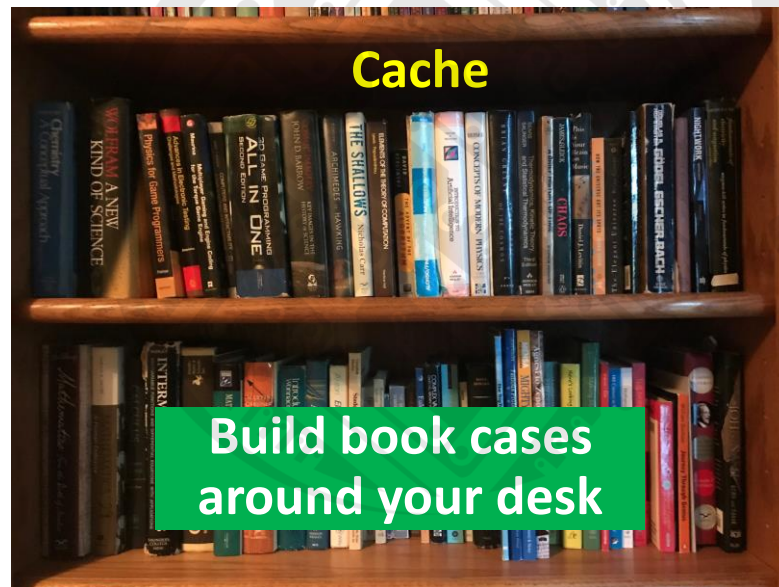
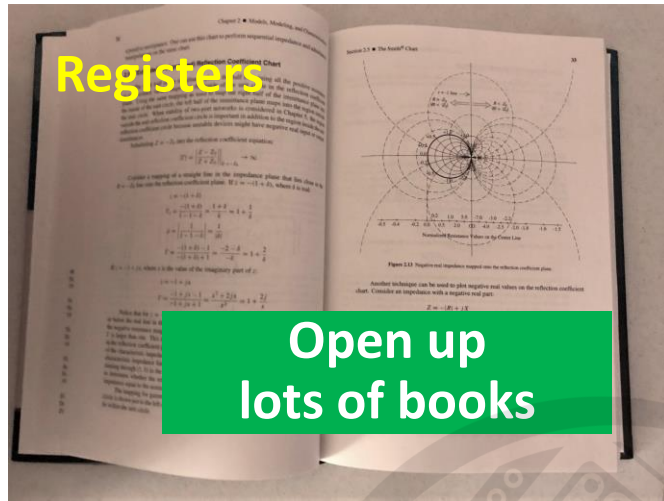


Source: worldfloorplans.com

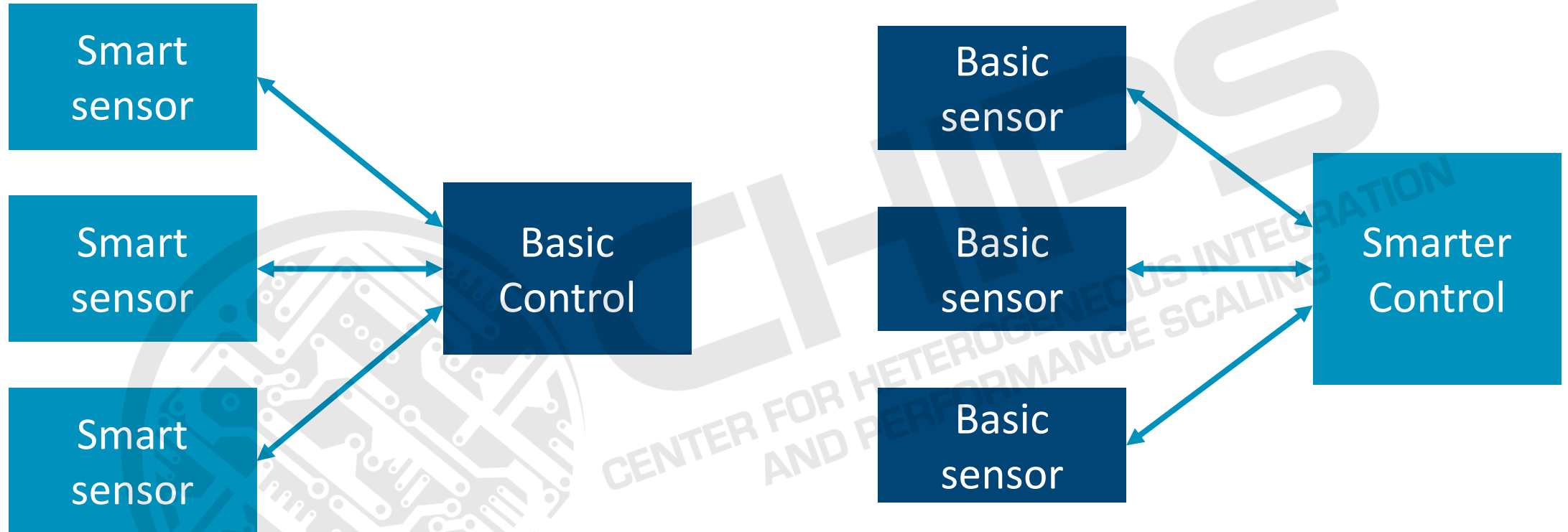


Memory at human scale

wafer



Distributed systems choices

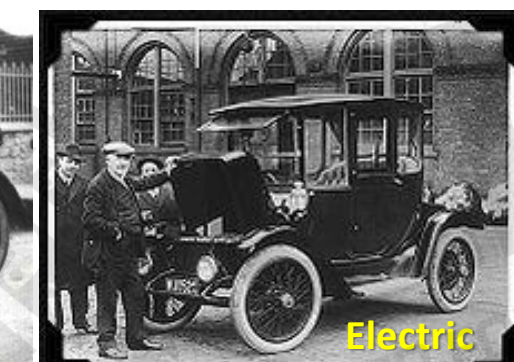
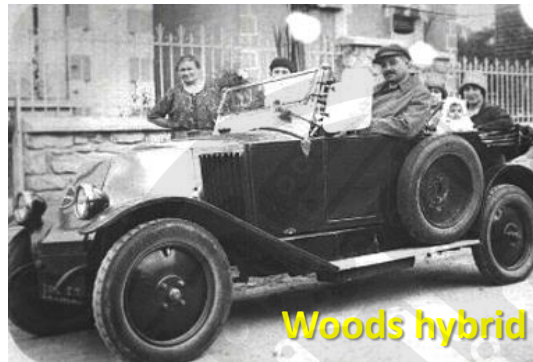
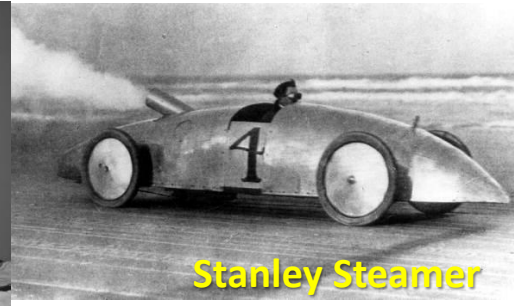


One size doesn't fit all

	3D-SIP			3D-SIC	3D-SOC		3D-IC		
3D Technology	“PoP”	“Chip last”	“Chip first”	Die stacking		Parallel W2W		Sequential FEOL	
3D-Wiring level	Package I/O	Chip I/O Interposer I/O	Chip I/O	Global	Semi-global	Intermediate	Local	FEOL	
				Chip BEOL Wiring Hierarchy					
Partitioning	Functional unit	subsystem	Embedded die	Die	Blocks of standard cells		Standard cells	Transistors	
Technology	Package-to Package reflow	Multi-die SIP 3D/2.5D stack	FO-WLP Embedded die	3D D2D, D2W 2.5D Si-interposer	Wafer-to-Wafer bonding		Active layer transfer or deposition		
					Hybrid bonding	Via-last			
2-tier stack Schematic									
Characteristic	Solder ball Stack	• C4, Cu-pillar Si-Organic • Through- Mold-vias	• Bumpless • Si-RDL • Through- Package-vias	• μ bump • Si-to-Si • Through- Silicon-Via	BEOL between 2 FEOL layers			FEOL stack	
					Overlay 2 nd tier defined by W2W alignment/bonding		Overlay 2 nd tier defined by litho scanner alignment		
Contact Pitch	400 $\Rightarrow$ 350 $\Rightarrow$ 300 μ m	120 $\Rightarrow$ 80 $\Rightarrow$ 60 μ m	60 $\Rightarrow$ 40 $\Rightarrow$ 20 μ m	40 $\Rightarrow$ 20 $\Rightarrow$ 10 $\Rightarrow$ 5 μ m	5 μ m $\Rightarrow$ 1 μ m	2 μ m $\Rightarrow$ 0.5 μ m	200nm $\Rightarrow$ 100nm	< 100 nm	
Relative density:	1/100 $\Rightarrow$ 1/77 $\Rightarrow$ 1/55	1/9 $\Rightarrow$ 1/4 $\Rightarrow$ 1/2.3	1/2.3 $\Rightarrow$ 1 $\Rightarrow$ 4	1 $\Rightarrow$ 4 $\Rightarrow$ 16 $\Rightarrow$ 64	64 $\Rightarrow$ 1600	400 $\Rightarrow$ 6400	4 10 ⁴ $\Rightarrow$ 1.6 10 ⁵	> 1.6 10 ⁵	

Source: IMEC via Electronics Weekly, Jan 18

Disruptive technologies: diversity leads to uniformity

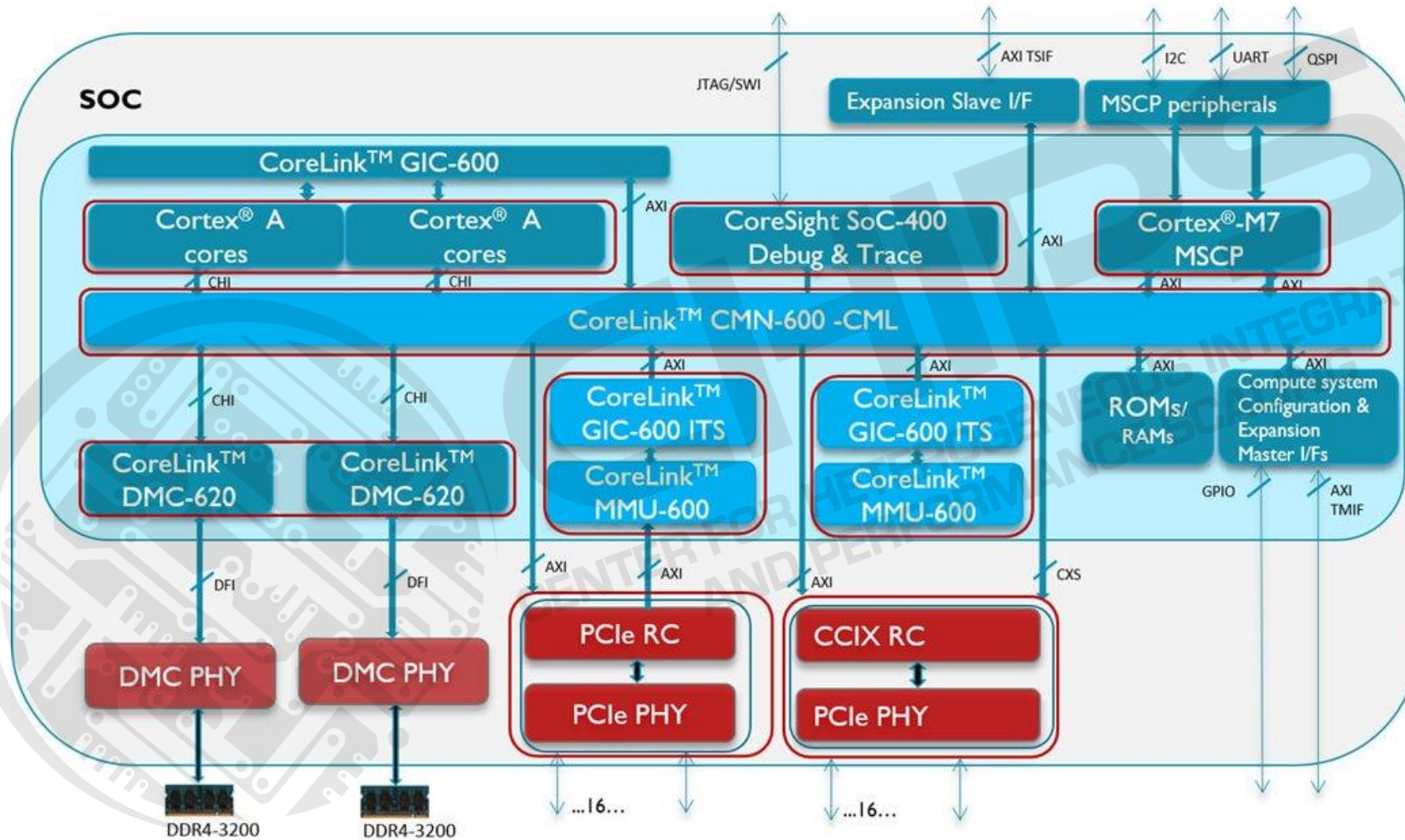


- Initially wide variety of creative solutions to a problem
- Some of these do better than others, eventually leading to uniformity
 - Still creativity, but focused on details
- Change in underlying problem can bring about new creative era

Heterogeneity: AMBA 5 Coherent Hub Interface (CHI)

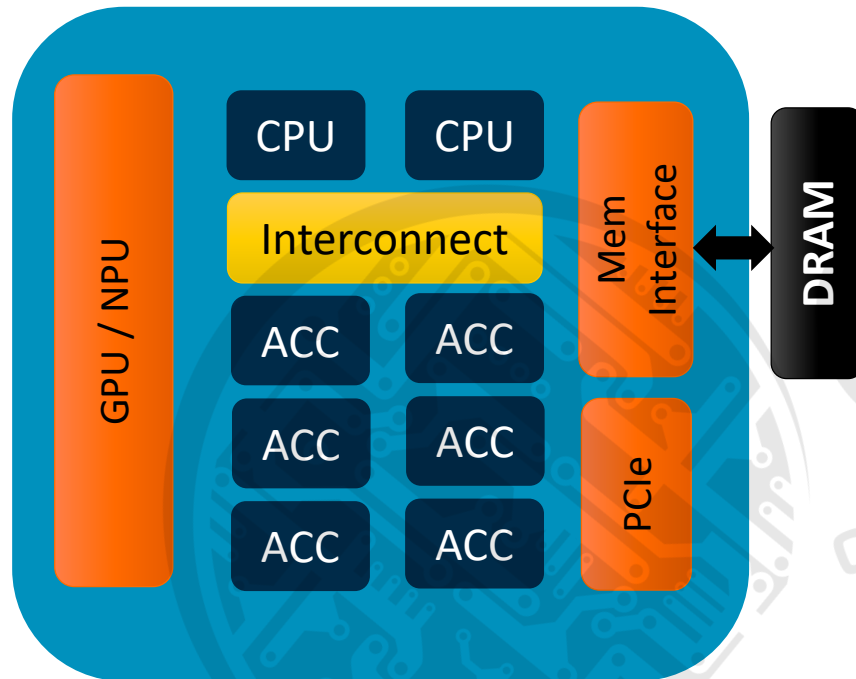
- Support for high frequency, non-blocking coherent data transfer between many processors.
- A layered model to allow separation of communication and transport protocols for flexible topologies, such as a cross-bar, ring, mesh or ad hoc.
- Cache stashing to allow accelerators or IO devices to stash critical data within a CPU cache for low latency access.
- Far atomic operations enable the interconnect to perform high-frequency updates to shared data.
- End-to-end data protection and poisoning signaling.

Cache coherent interconnect for accelerators (CCIX)

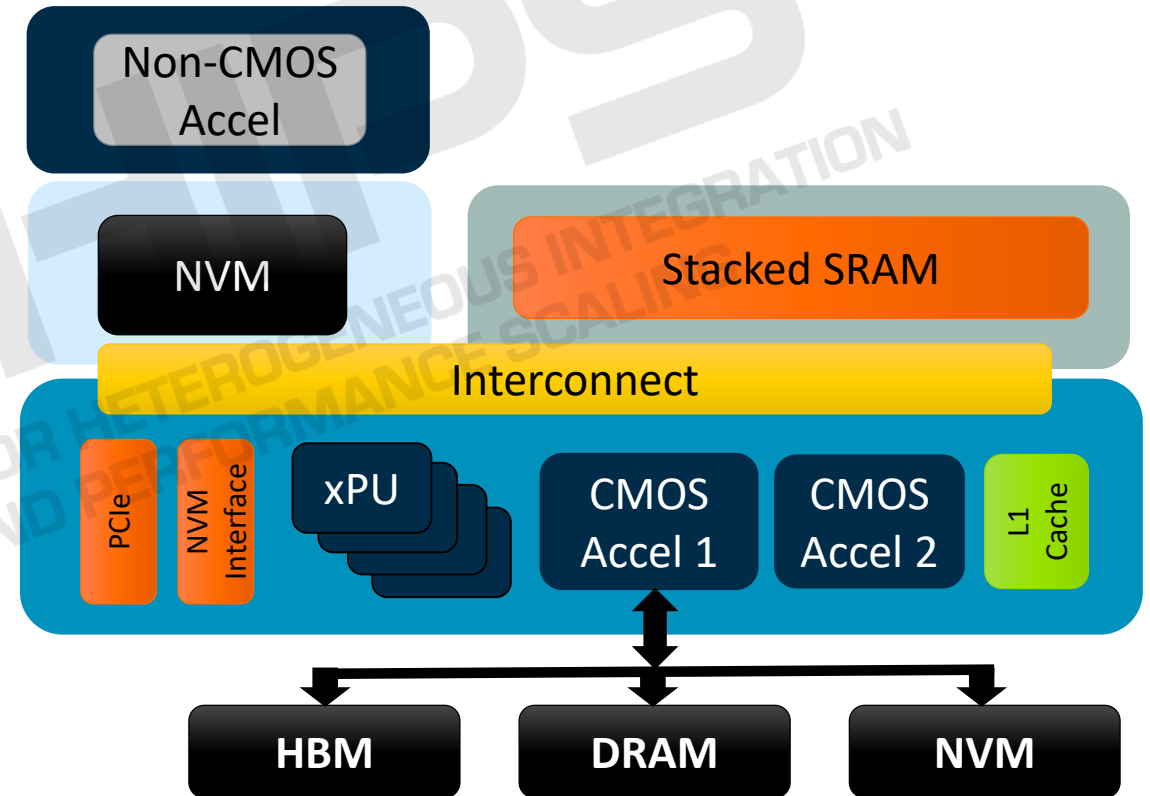


Heterogeneity at wafer scale

Current Heterogeneity



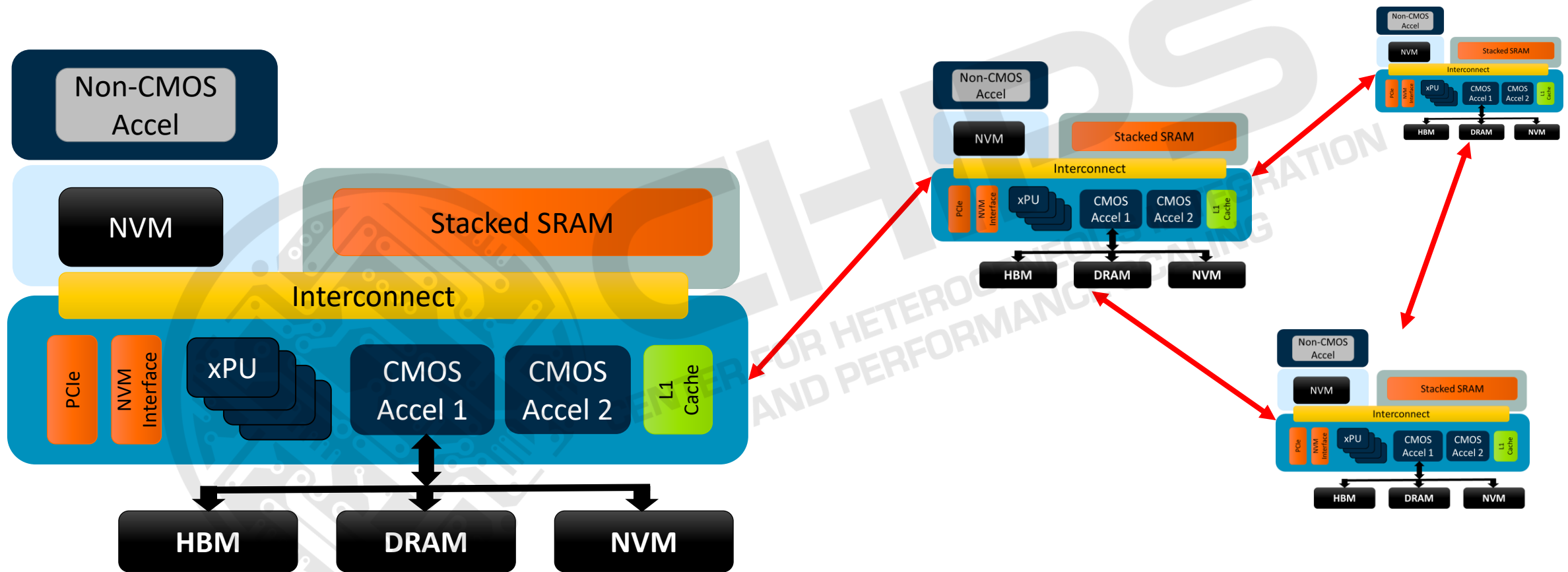
Localized 3D Heterogeneity Example



Could have different variants across wafer

Architecture decisions

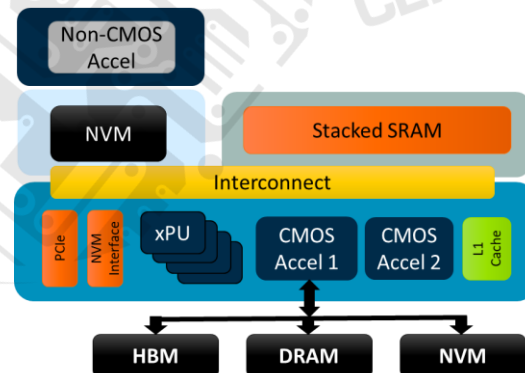
Massive, distributed 2.5D heterogeneity with localized 3D capability



Questions and constraints

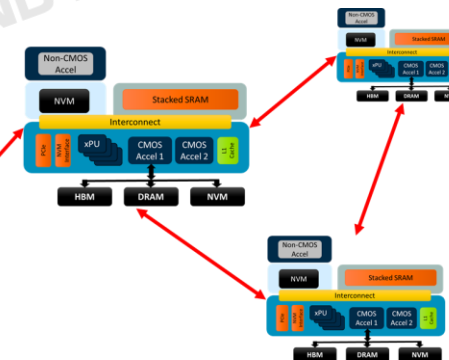
Questions

- What kinds of problems can it solve?
- How is it programmed?
- How many clusters does it have?
- How do they connect?
- What is in each cluster?
- How do those pieces connect?



Constraints

- What cost/time/energy/space/heat bounds do I have to meet?
- How many do I need?
- What does the yield have to be?
- How reliable does it have to be?
- How secure does it have to be?



Making use of 3D

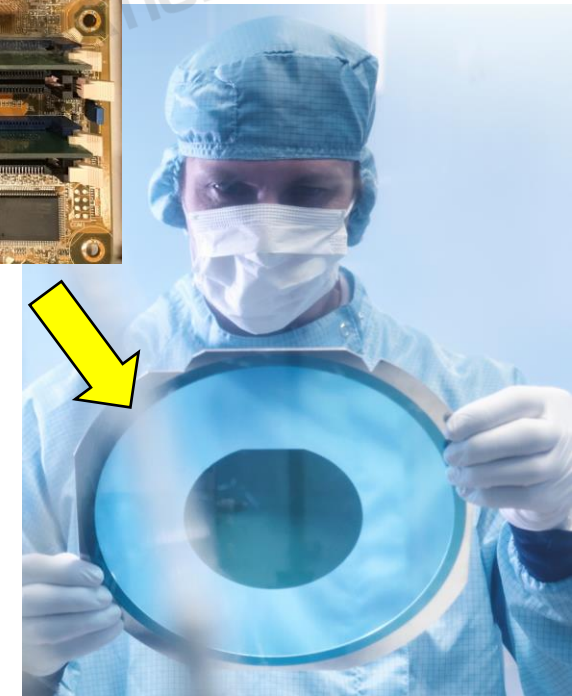
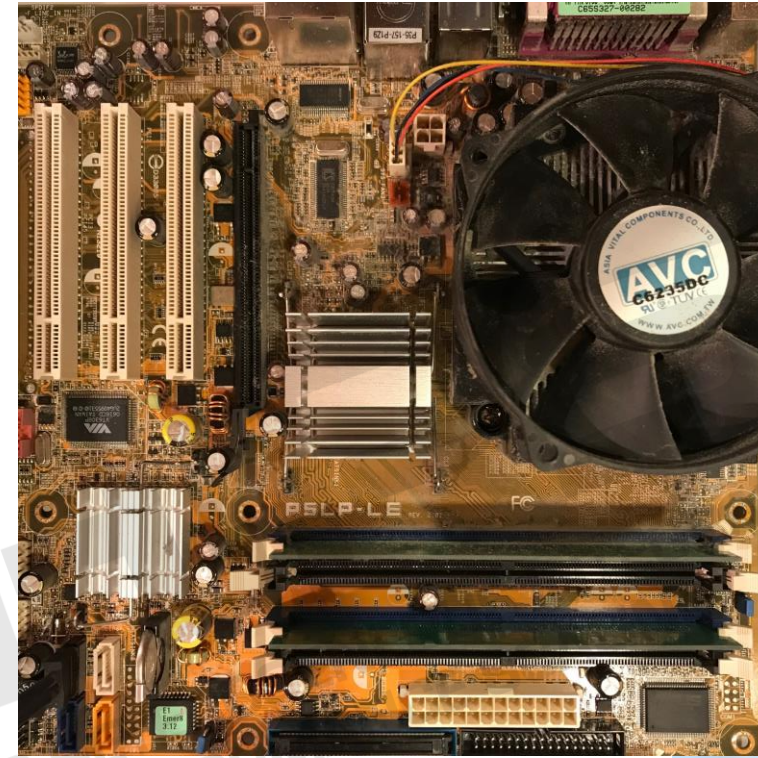
- Edge connections grow linearly with edge dimensions
 - Folding a 2D interface gains some benefits
 - Spreading a single object across layers gains additional benefits
- True 3D connections grow as the square of edge dimensions
- Using 3D effectively requires work
 - Physical connection
 - Wide interface
 - Bottleneck avoidance
 - Standards
 - DfX (short answer: “AND” yield bad, “OR” yield good, “known (mostly) good die”)

Something is going to win

In high performance, it will look something like wafer-scale

Some likely features

- Scalable, modular design
- Commercial EDA support
- Heterogenous compute
- Solution to memory bottleneck
- Solution to I/O bottleneck
- Reasonable answers on test, yield etc.



So where does that leave us?

- Several key workloads can use wafer scale compute power
- Effective wafer scale architectures need effective memory systems
- Opportunities for 3D solutions, but many chicken and egg problems
- Now is the time for experimentation



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

The End. For now.



Thanks to Arm colleagues

- Brian Cline
- Stephan Diestelhorst
- Wendy Elsasser
- Doug Joseph
- Alejandro Rico
- Saurabh Sinha
- Dam Sunwoo
- Greg Yeric



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

arm



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

Thank You

Danke

Merci

谢谢

ありがとう

Gracias

Kiitos

감사합니다

धन्यवाद

شكراً

תודה

arm

The Arm trademarks featured in this presentation are registered trademarks or trademarks of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All rights reserved. All other marks featured may be trademarks of their respective owners.

www.arm.com/company/policies/trademarks



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING