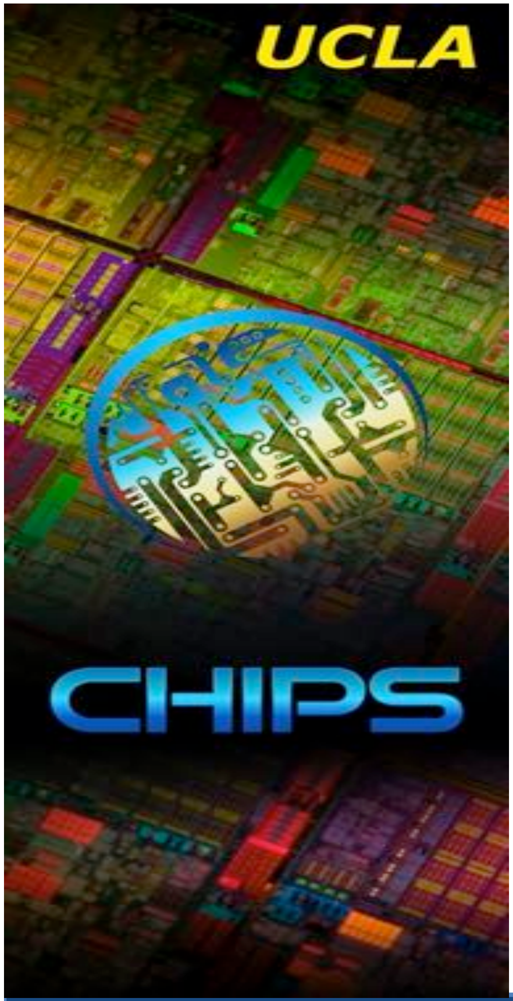


Signaling for wafer-scale systems

Subramanian S. Iyer
(s.s.iyer@ucla.edu)

Center for Heterogeneous Integration and Performance Scaling
chips.ucla.edu

Discussion with ODSA group on November 20,2020

UCLA CHIPS

A UCLA Led partnership to develop Applications, Enablement and Core technologies and the eco-system required for continuing Moore's Law at the Package and System Integration levels and develop our students & scholars to lead this effort

Simplify hardware development through novel architectures, integration methods, technologies, and devices.

What we do @UCLA CHIPS

Large Scale Energy Efficient Systems

Medical Engineering applications

Advanced Packaging Technologies

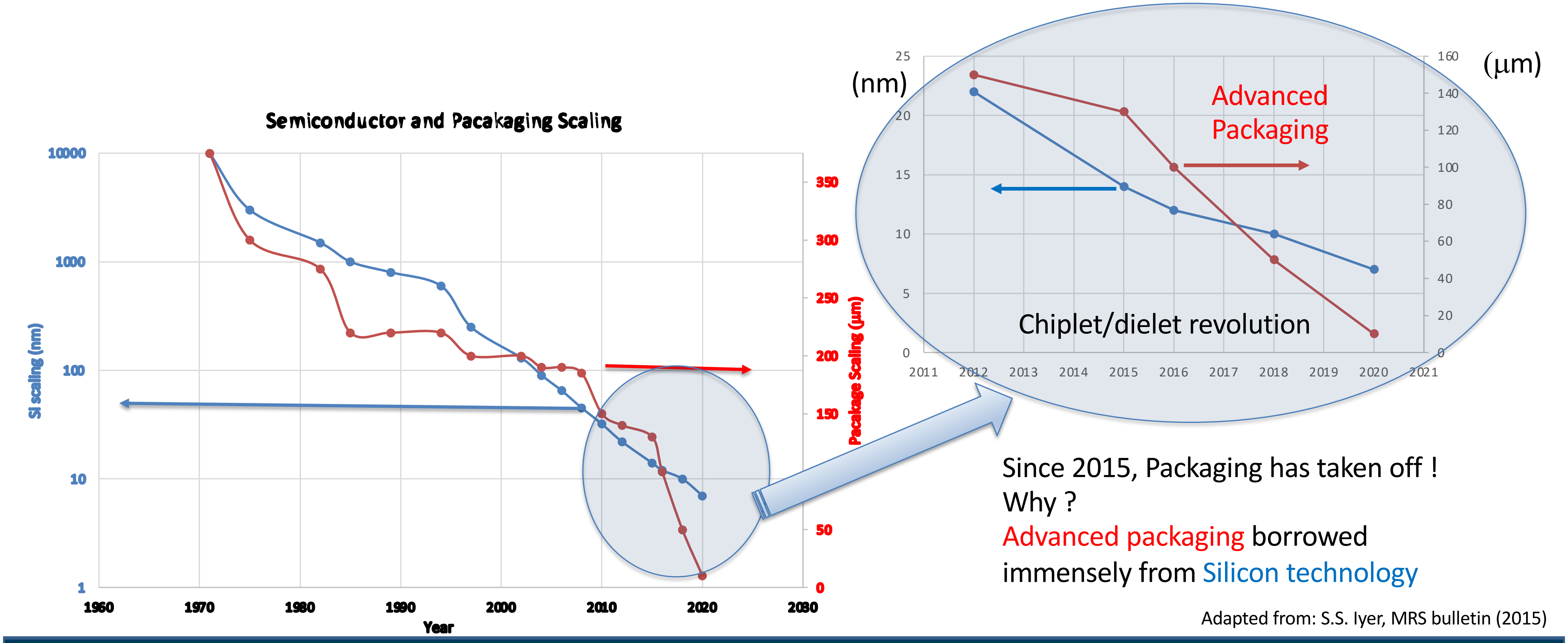
Novel Compute architectures

Silicon as a heterogeneous fine pitch packaging Platform, Si IF

FlexTrate as a flexible Biocompatible Heterogeneous Integration Platform

The CTT as an in-memory compute device

Silicon and Package scaling



Advanced Packaging

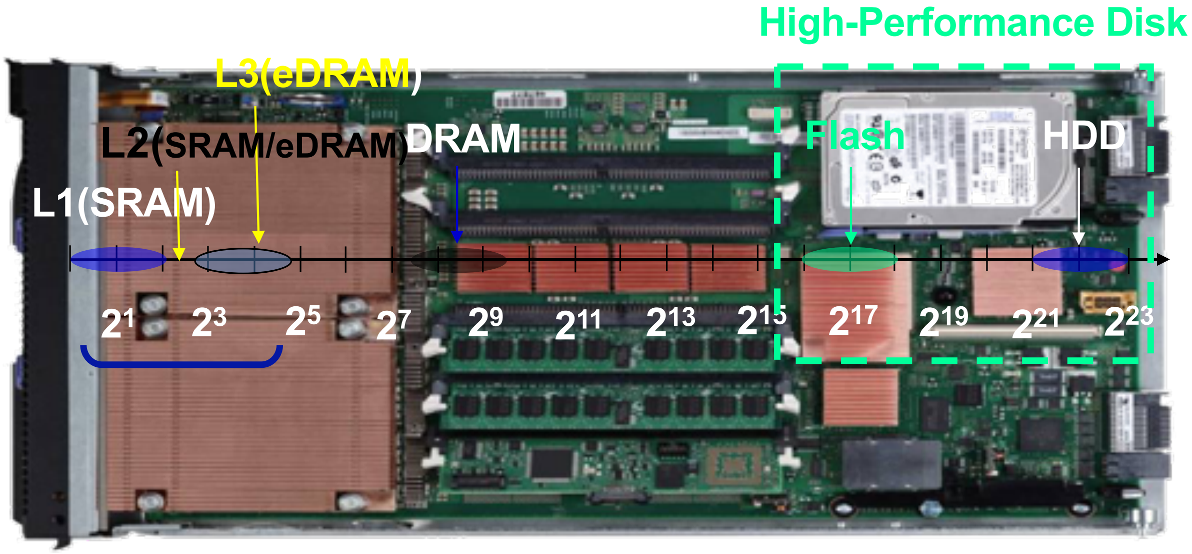
Chiplet/dielet revolution

Since 2015, Packaging has taken off !
Why ?
Advanced packaging borrowed immensely from Silicon technology

Adapted from: S.S. Iyer, MRS bulletin (2015)

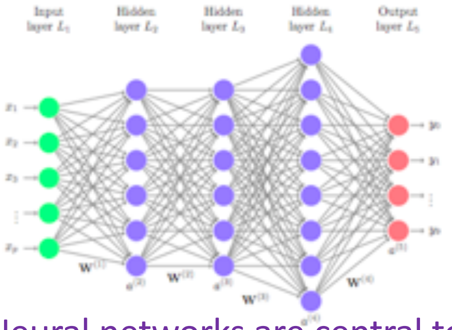
Why is heterogeneity assuming sudden importance ?

- Packaging has always been about assembling heterogeneous dies/chips onto a Printed Circuit Board

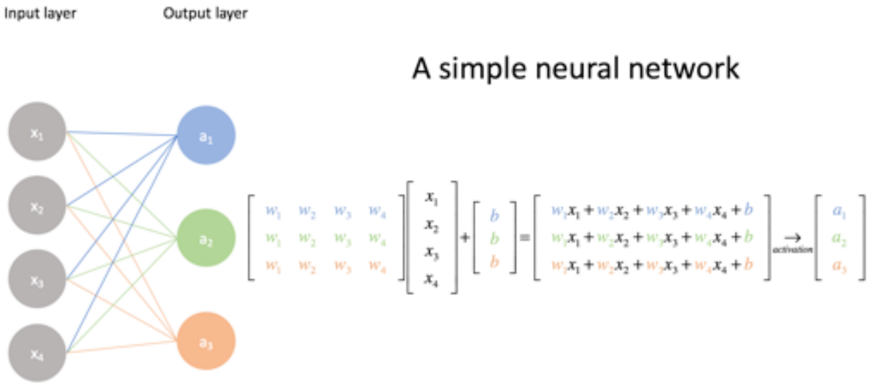


- The problem with PCBs has to do with Latency and Bandwidth between the chips as well as energy per bit transferred

Packaging and AI - A one-page illustrative primer



Neural networks are central to AI
Accuracy requires these networks
to be extremely deep (many hidden layers)
Eg. Residual Net (ResNet) has ~1000+ layers
Also the width of these hidden layers can
also be quite large



A simple neural network

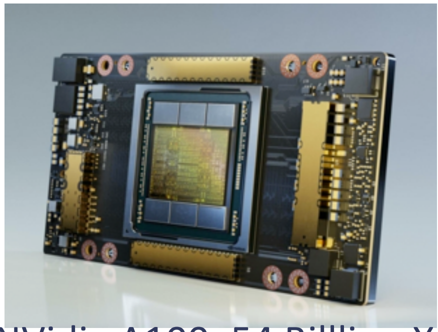
$$\begin{bmatrix} w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix} = \begin{bmatrix} w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \end{bmatrix} \xrightarrow[\text{activation}]{} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

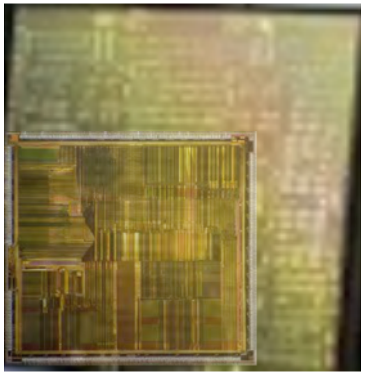
Vector multiplications are a key operation in neural networks
And the vector multiply and accumulate (MAC) function is central
The bit precision of the inputs, weights and outputs can exceed 16 bit,
leading to unprecedented computational complexity .

Even with today's very powerful processors, processors need
to time multiplex, ***constantly*** moving inputs, weights and
outputs of each layer between the processor and memories
So the memory bottleneck is quite severe.
This is where packaging comes in ! - BW, energy-per-bit Xferred
(and latency) define system performance (and processer speed)

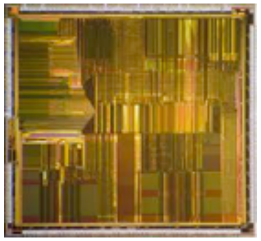
Some observations

- If Moore's law has enabled miniaturization, why have chips gotten larger ?
  - More complex problems
  - More cores @ higher clock speeds
  - More cache memory
- Main memory capacity and access limits performance
- Power density challenges - more "dark" silicon
- I/Os take up more space and power as system size increases >30%





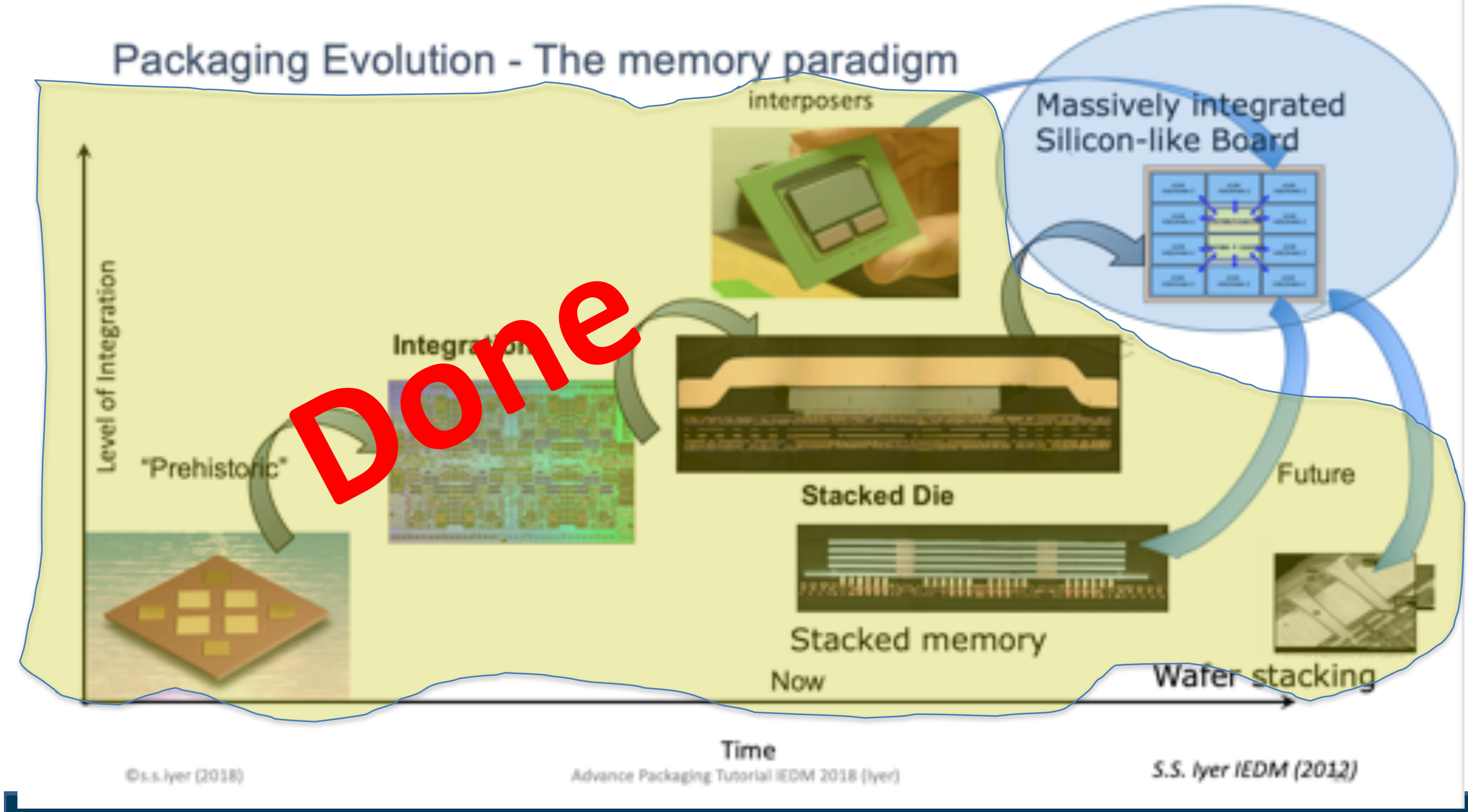
NVidia A100: 54 Billlion Xtors - 826 mm^2 (2020)
In TSMC 7 node



17,000 more transistors

Intel Pentium cpu ~300mm^2
-3.1 Million Xtors (1993)
0.8 µm technology

Can this be Done practically ?



©s.s.iyer (2018)

Advance Packaging Tutorial IEDM 2018 (Iyer)

S.S. Iyer IEDM (2012)

Some more observations:

- Interposers are getting bigger
- 3D stacks are getting taller

- Interposers are an additional level in the packaging hierarchy

Going to a silicon-like board With fine pitch interconnect and short die to die spacings will allow us to build massive systems

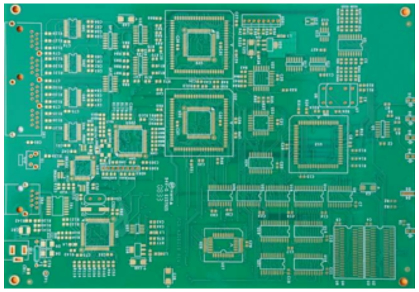
But many issues need to be addressed

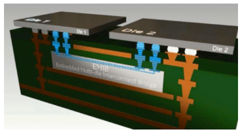
The “Right” Rigid Interconnect Fabric

Requirements:

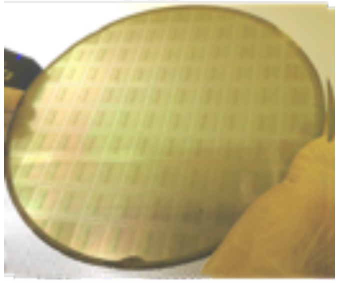
- **Mechanically robust (flat, stiff, tough…)**
- **Processability: fine pitch wiring, & interconnects**
- **Thermally conductive**
- **Can have passive (and active) built-in components**
- **Economical**

Organic (e.g. FR-4)



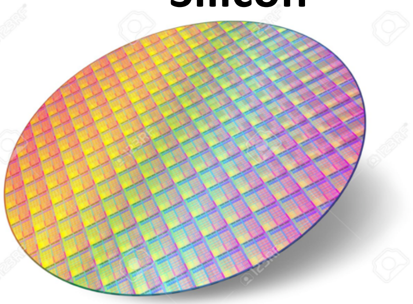


Glass

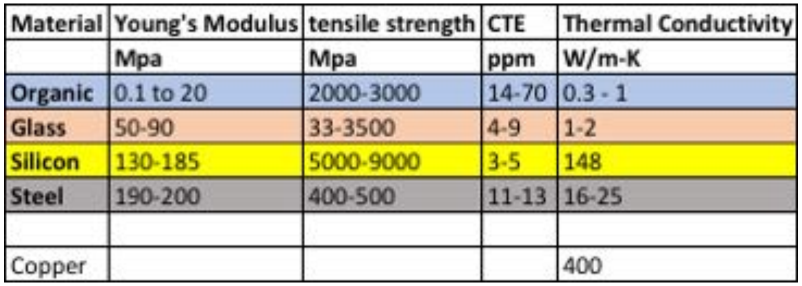


Hedrick et al ECTC 2016

Silicon



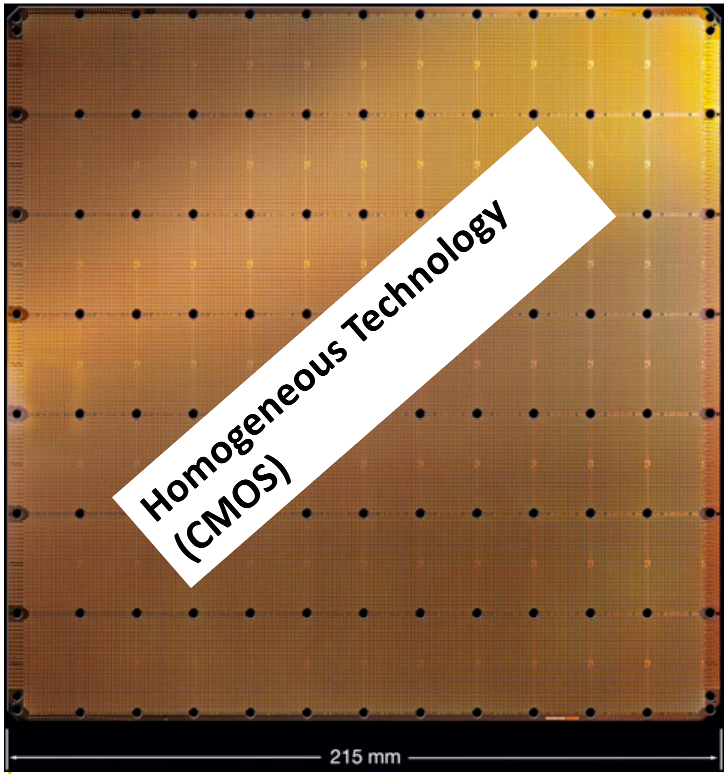
Hybrid approaches(EMIB by Intel



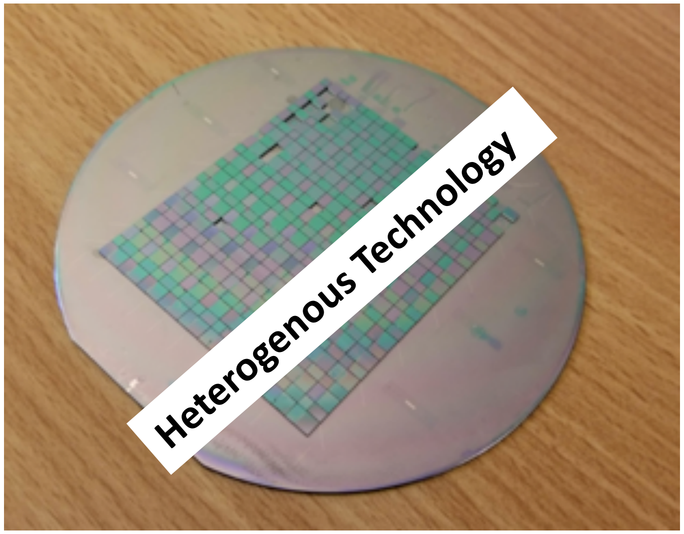
Material	Young's Modulus	tensile strength	CTE	Thermal Conductivity
	Mpa	Mpa	ppm	W/m-K
Organic	0.1 to 20	2000-3000	14-70	0.3 - 1
Glass	50-90	33-3500	4-9	1-2
Silicon	130-185	5000-9000	3-5	148
Steel	190-200	400-500	11-13	16-25
Copper				400

Going to a silicon wafer scale is not new - there is a

Bumped 100 mm wafer (Ca 1982)
Trilogy Systems



Cerebras (2019) - wafer scale AI processor
Homogeneous



UCLA CHIPS 2019
Wafer scale Heterogeneous assembly

Our approach:
Integrate lots of dielets on a silicon substrate at fine pitches

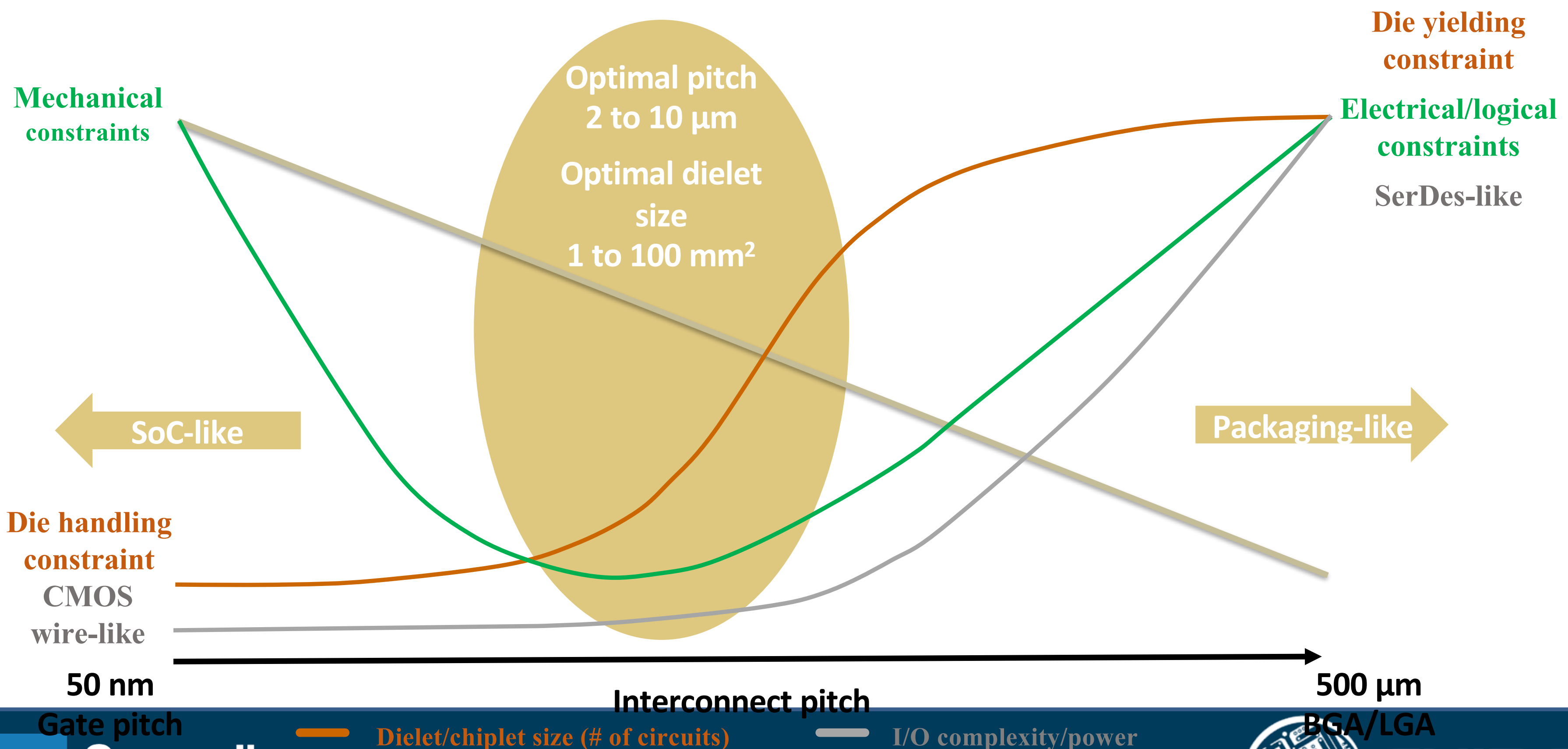
* Wafer-Scale Integration

Important Questions

- What is the optimal pitch at which dies should be interconnected ?
- What is the optimal dielet size
- How close should we assemble dies
- What level of heterogeneity should we aim for

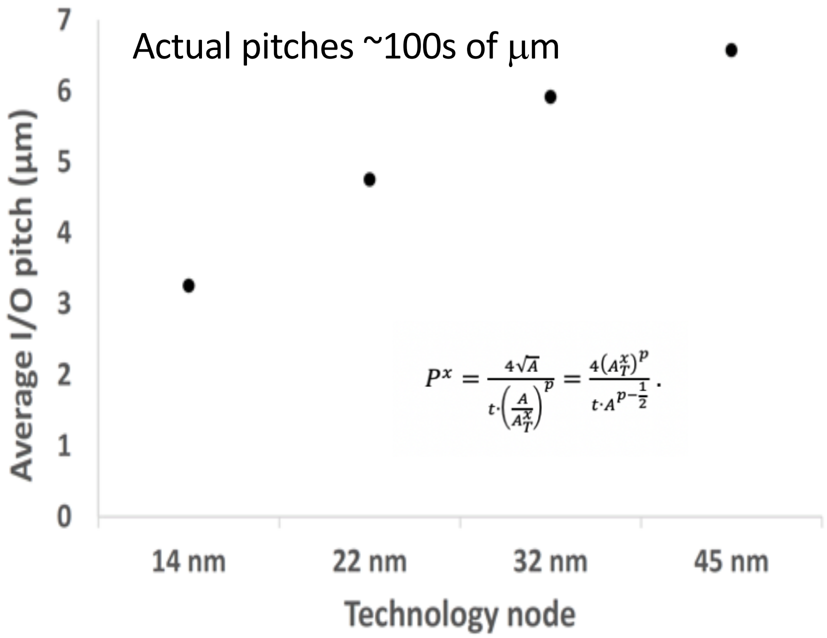
Hint: how do we make a SOW look like an ginonormous SOC

The CHIPLET Golden Regime



What is the optimal I/O pitch ?

Chip	Area (mm^2)	Transistor count (x10^9)	Technology node (nm)
IBM POWER9 [26]	695	8	14
AMD Zen [27]	44	1.4	
IBM POWER8 [28]	649	4.2	22
Intel Xeon Haswell E5 [29]	663	5.56	
IBM POWER7 + 80 MB [30]	567	2.1	32
Intel Itanium Poulson [31]	544	3.1	
IBM POWER7 + 32 MB [32]	567	1.9	45
Intel Xeon 7400 [33]	503	1.9	



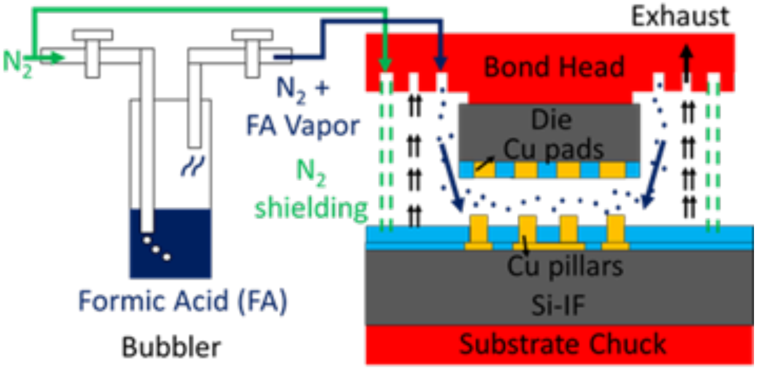
Actual pitches ~100s of µm

$$P^{x} = \frac{4\sqrt{A}}{t\cdot\left(\frac{A}{A_T^x}\right)^p} = \frac{4\left(A_T^x\right)^p}{t\cdot A^{p-\frac{1}{2}}}.$$

Practical limits in heterogeneous integration

- Fine pitch ?
  - like ”fat wires” on a Silicon wafer - 2-10 μm - this is the bump pitch (BGA pitch is >500μm)
  - Trace pitch < 1 μm (compared to ~30 μm on PCB)
- Precision alignment ?
  - similar to fat wire alignment <0.2 μm (bumps alignment accuracy is several μm)
- Close Spacing
  - As close as possible <20 μm (dies on a PCB are spaced at least a few 10’s of mm away)
- Typical block sizes on an SoC are typically a few ~100 μm on a side
  - So dielets should be small (1 to 100 mm^2 in area)
- Heterogeneity:
  - multiple nodes - use the node that is optimal from a performance, area and cost perspective
  - multiple technologies - logic, DRAM, sensors etc.
  - multiple materials Si , III-Vs………

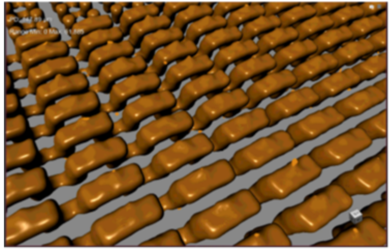
A versatile Fine pitch wafer-scale assembly (Si IF)

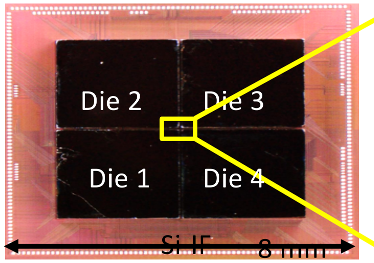


Process parameters
Pressure
Temperature
Surface prep

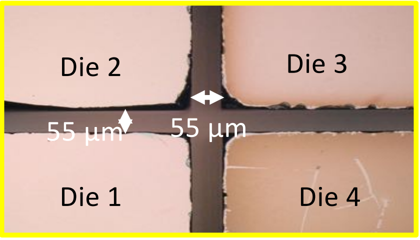
Direct Cu-Cu Thermal Compression Bonding using formic acid vapor

X-Ray Tomograph of
10µm Cu-Cu pitch die to
wafer connects

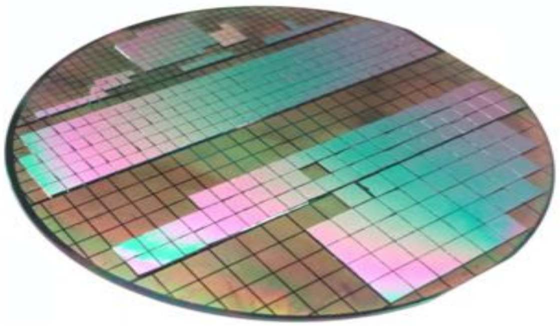




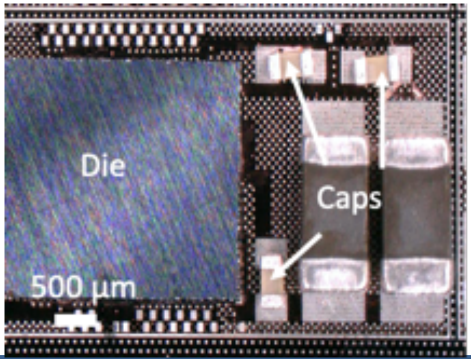
2x2 array of TSMC 16FF
dies on the Si-IF.



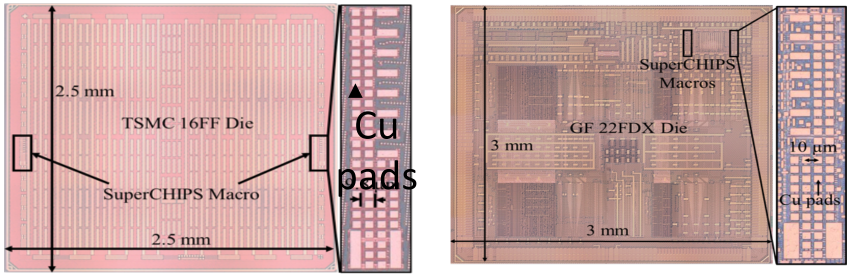
55 µm inter-die spacing



Wafer scale assembly at fine pitch
Both Si and III-Vs



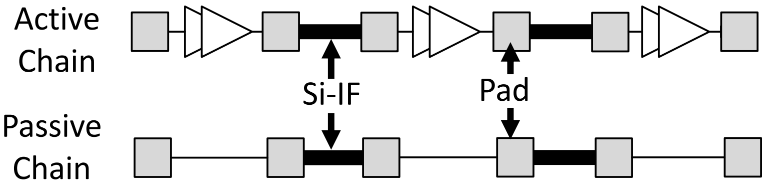
Legacy dies & passives on Si-IF

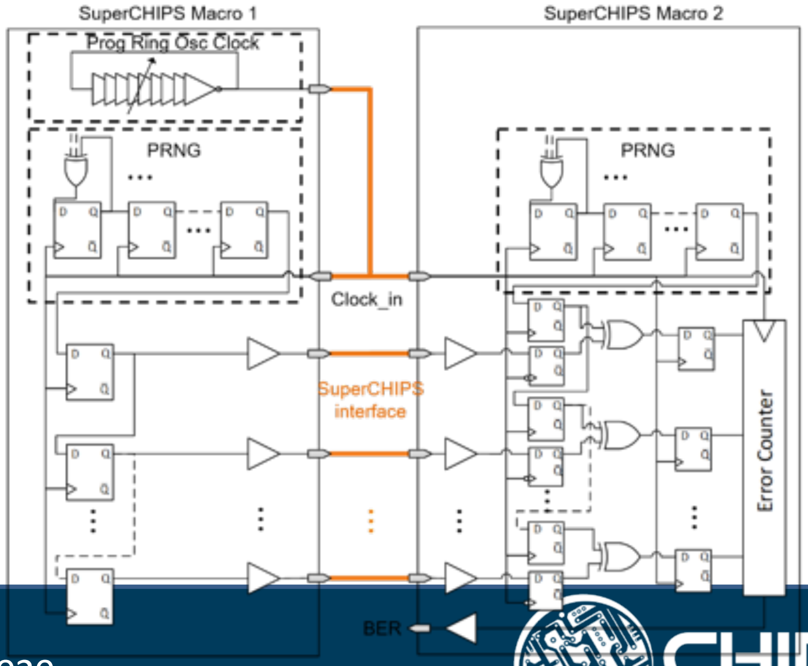


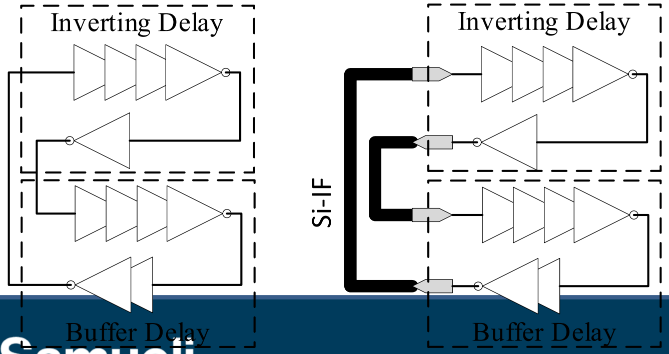
Developed termination protocols
with most major foundries Cu for Si, Au for III-Vs

Established CHIPS metrics using SuperCHIPS macros

- Continuity check
- Latency characterization
  - Reference & Si-IF ring oscillator: 3-4 GHz
  - On-chip frequency divider (2^{12}) & cycle counter
- High-speed data transfer & Bit error rate (BER)
  - Programmable ring oscillator clock: 0.5-3 GHz
  - Pseudo Random Number Generator (PRNG)
  - On-chip comparator and error counter



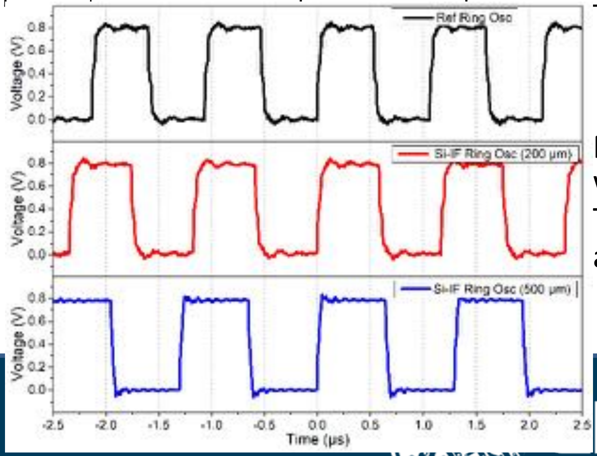




Results: SuperCHIPS macros (GF 22FDX, TSMC 16FF)

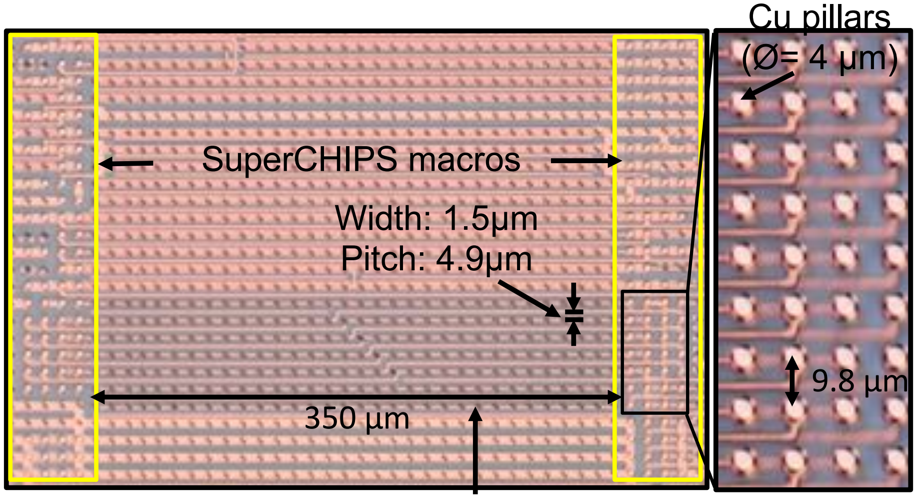
- Successfully passed continuity tests of both passive and active daisy chains
- Measured latency verified with on-chip counter
  - –Latency comparable to on-chip buffer delays
  - –Overall latency is <30 ps
- Demonstrated data transfer up to 3 Gbps
  - –Bandwidth: 1200 Gbps/mm for 2-layer Si-IF
  - –No errors were observed even after 43 hrs
  - –BER: <10^{-14} with 99% confidence (Estimate: <10^{-25})
- Measured energy/bit: 0.028 pJ/b
- No electrostatic discharge protection (ESD) used
  - –For ESD protection of 50 fF : Latency & Energy increase by 2X

Oscillator	*Measured frequency [kHz]*	*Actual frequency [GHz]*	*Latency of Si-IF links [ps]*
TSMC 16FF Die			
On-chip reference	921.1	3.77	NA
200 µm Si-IF links	836.8	3.43	6.67
500 µm Si-IF links	762.3	3.12	13.80
GF 22FDX Die			
On-chip reference	1033.9	4.23	NA
200 µm Si-IF links	877.6	3.59	10.51
500 µm Si-IF links	760.3	3.11	21.26

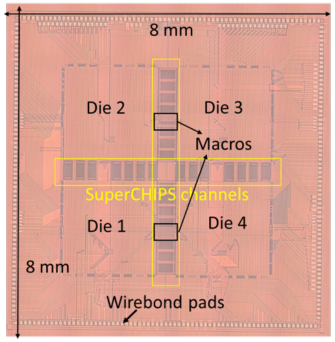


Measured waveforms for TSMC 16FF die assembly

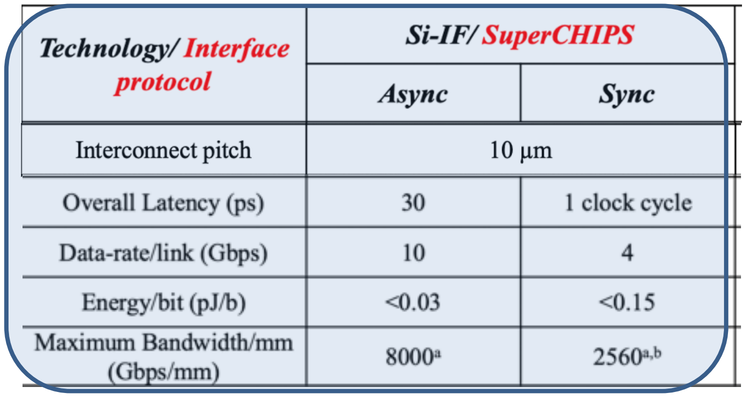
SuperChips - a versatile communication protocol



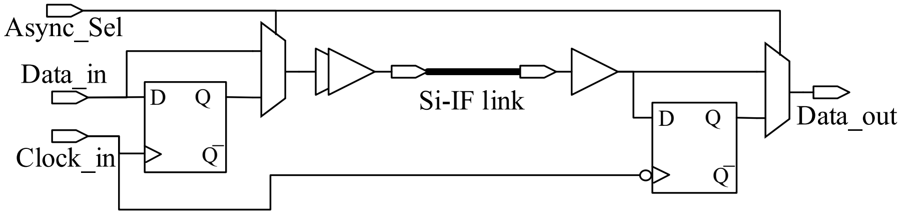
Micrograph of the fabricated SuperCHIPS interface



8111 I/O interdie Connections
22291 power Connections



Technology/ Interface protocol	Si-IF/ SuperCHIPS	
	Async	Sync
Interconnect pitch	10 µm	
Overall Latency (ps)	30	1 clock cycle
Data-rate/link (Gbps)	10	4
Energy/bit (pJ/b)	<0.03	<0.15
Maximum Bandwidth/mm (Gbps/mm)	8000[a]	2560[a,b]



Schematic of the SuperCHIPS I/O

Longer Range connections can be done daisy chaining through Intervening dies using porosity rules and multiple buffer stages - for a few die over
or
using pico-SerDes for longer (~ cms) lengths.

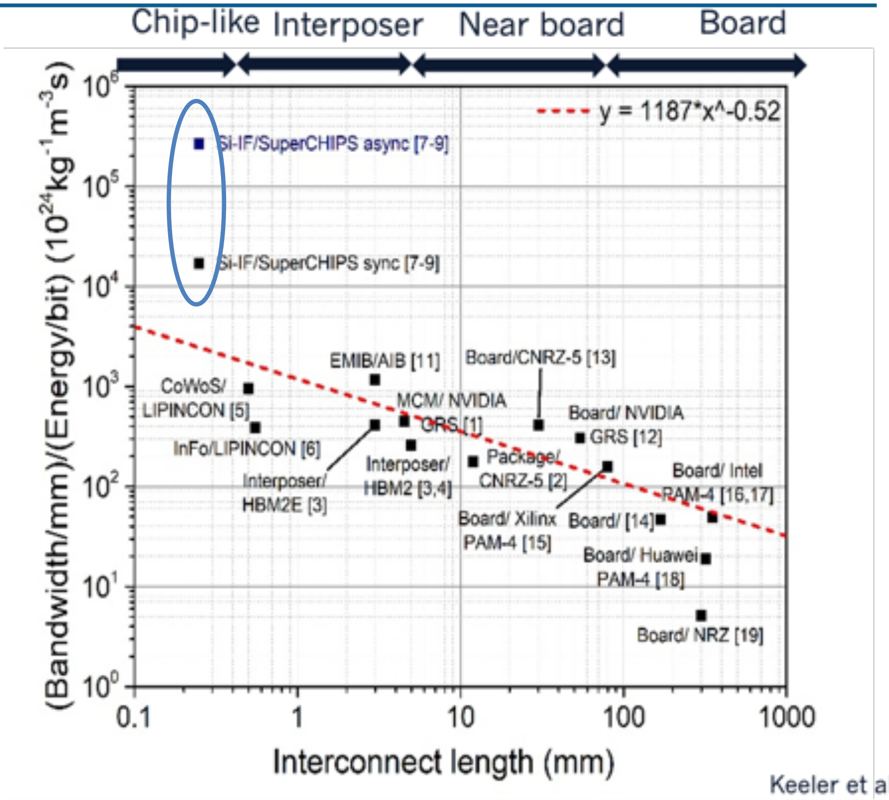
Using "utility dies" which may also provide redundant routing options to manage assembly defects

Technology Comparison using s-FOM$_k$

Tech/ Interface protocol	*Si-IF/ SuperCHIPS*		*Interposer/ AIB*	*PCB/ SerDes*		*Improvement*
	Async	*Sync*				
Reach	Neighbor		Neighbor	Neighbor	Long Reach	
Overall Latency (ps)	30	500	1500[1]	~2000	~6000	3-65X
Energy/bit (pJ/b)	<0.03	<0.15	0.8-0.85[3,4]	1.17[7]	6.9[13]	5-40X
Bandwidth/ mm (Gbps/mm)	8000[a]	2560[a,b]	707.7[b]	354	149-298[c]	4-23X

[a] 4 wiring levels, [b] Assuming 20% overhead, [c] Estimated from data in [10-13]

$$s - FoM_k = \frac{\left(\frac{Bandwidth}{mm}\right)}{\left(\frac{Energy}{bit}\right)}$$



Keeler et al

Jangam & Iyer T-CPMT (2020)

[1] AIB interface [2] HBM JEDEC Standard JESD235C, 2020. [3] M. O'Connor et al, MICRO, 2017. [4] M. Lin, JSSC, 2020. [5] M. Lin, et al, HCS, 2016. [6] J. W. Poulton, et al, JSSC, 2013. [7] J. W. Poulton et al., JSSC, 2019. [8] A. Shokrollahi, ISSCC, 2016. [9] A. Tajalli, et al, JSSC, 2020. [10] Y. Krupnik et al, JSSC, 2020. [11] J. Kim et al, JSSC 2019. [13] M. LaCroix et al, ISSC 2019. [14] E. Depaoli, JSSC, 2019.

Technology Comparison s-FoM$_{ucla}$ - shows the benefit of technology

- Does not account for area used by I/Os
  - SerDes occupy significant chip area
  - Especially when we have deep I/Os that go several layers in
  - This can be >30% of die area !
  - *Note: this is influenced by Technology node*

- Does not account for latency
  - ToF is not always the main contributor
  - Serialization, Deserialization, equalization, clock recovery etc. are the major contributors
  - *Note: this is influenced by Circuit design*

- No credit for load that is driven
  - *This is influenced by Packaging Technology*

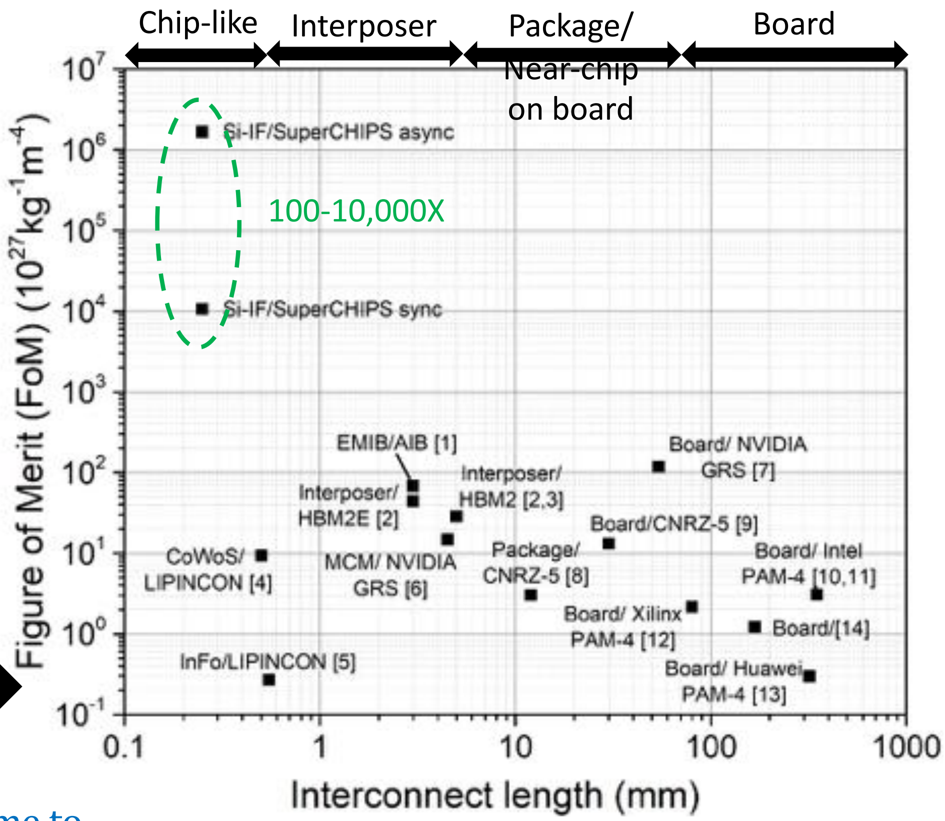
BW/mm

Surrogate for Load

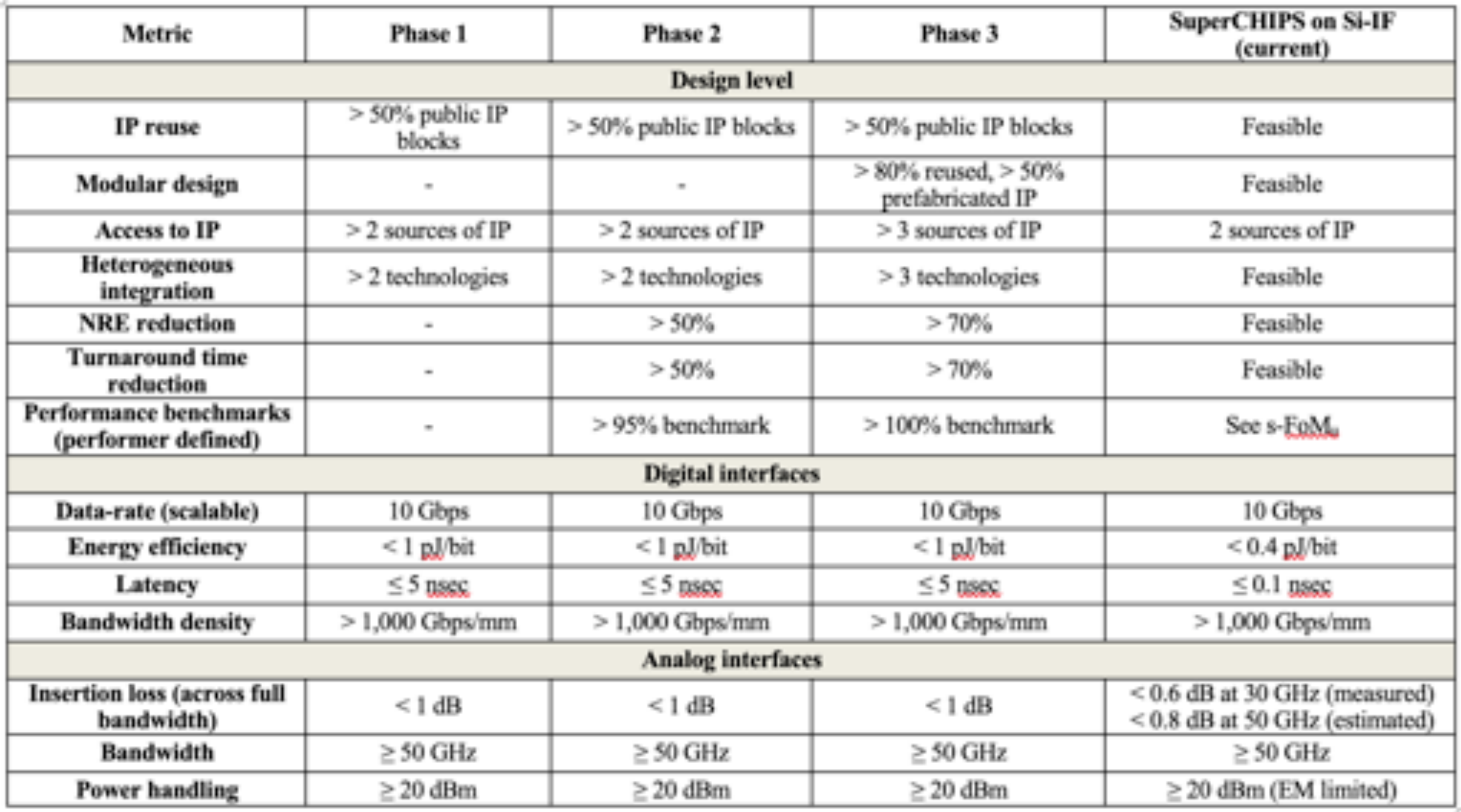
$$s-FoM_u = \frac{\left(\frac{Bandwidth}{shoreline * IOcols}\right) * Length_{link}}{\left(\frac{Energy}{bit}\right) * \left(\frac{TransceiverArea}{Link}\right) * Latency}$$

This is the die area "wasted" by IOs and cant be used for compute

Overhead time to serialize/deserialize data, ECC, + ToF



CHIPS Project Goals and Milestones



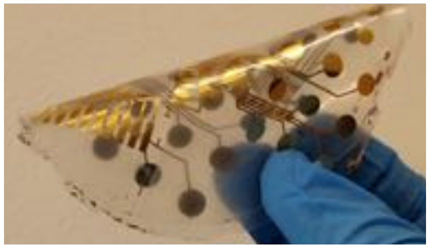
Metric	Phase 1	Phase 2	Phase 3	SuperCHIPS on Si-IF (current)
Design level				
IP reuse	> 50% public IP blocks	> 50% public IP blocks	> 50% public IP blocks	Feasible
Modular design	-	-	> 80% reused, > 50% prefabricated IP	Feasible
Access to IP	> 2 sources of IP	> 2 sources of IP	> 3 sources of IP	2 sources of IP
Heterogeneous integration	> 2 technologies	> 2 technologies	> 3 technologies	Feasible
NRE reduction	-	> 50%	> 70%	Feasible
Turnaround time reduction	-	> 50%	> 70%	Feasible
Performance benchmarks (performer defined)	-	> 95% benchmark	> 100% benchmark	See s-FoM.
Digital interfaces				
Data-rate (scalable)	10 Gbps	10 Gbps	10 Gbps	10 Gbps
Energy efficiency	< 1 pJ/bit	< 1 pJ/bit	< 1 pJ/bit	< 0.4 pJ/bit
Latency	≤ 5 nsec	≤ 5 nsec	≤ 5 nsec	≤ 0.1 nsec
Bandwidth density	> 1,000 Gbps/mm	> 1,000 Gbps/mm	> 1,000 Gbps/mm	> 1,000 Gbps/mm
Analog interfaces				
Insertion loss (across full bandwidth)	< 1 dB	< 1 dB	< 1 dB	< 0.6 dB at 30 GHz (measured) < 0.8 dB at 50 GHz (estimated)
Bandwidth	≥ 50 GHz	≥ 50 GHz	≥ 50 GHz	≥ 50 GHz
Power handling	≥ 20 dBm	≥ 20 dBm	≥ 20 dBm	≥ 20 dBm (EM limited)

So What are the issues ?

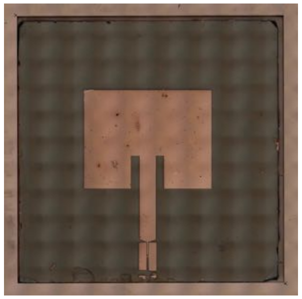
- Developing the assembly technology: fine pitch, close spacing tight alignment etc…
- Establishing a communication protocol for both near and far dielets
- Communicating with the outside world
- Delivering power - huge amounts of power !
- Extracting heat - huge amounts of heat !
- Making such system reliable
- Ensuring the costs are economical

Communicating with the outside world

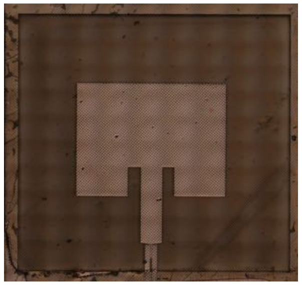
- Flexible high speed wired connectors (FlexTrate^{tm})



- RF links using embedded fused quartz or PDMS and III-V drivers

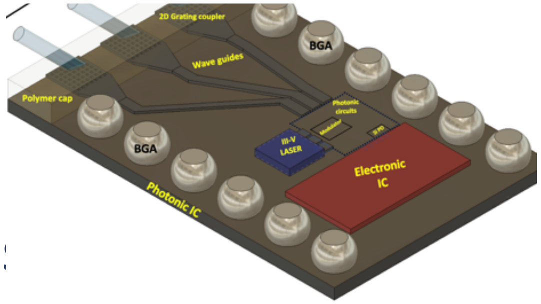


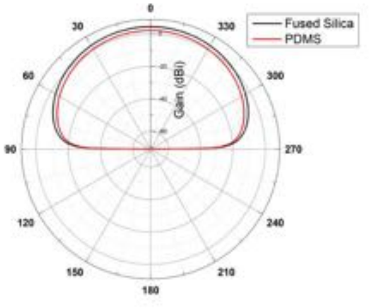
Antenna on Fused Silica substrate



Antenna on PDMS substrate

- Photonic Interconnects





Simulated radiation pattern (20GHz)

Summary

- Packaging has scaled significantly in the last few years
  - Driven by need, more investment, More silicon-like processing
  - Silicon as a base packaging material has significant potential
- The challenges are
  - Assembly - especially at high throughput
  - Connections to the outside world
  - Power delivery and heat extraction
  - Reliability and yield
  - Supply chain for bare dies
- We can extend this concept to flexible hybrid electronics (did not talk about it much today)

Selected Bibliography (more here)

- S. Jangam and S. S. Iyer, "A Signaling Figure of Merit (s-FoM) for Advanced Packaging," in IEEE Transactions on Components, Packaging and Manufacturing Technology, doi: 10.1109/TCPMT.2020.3022760
- S. Jangam, U. Rathore, S. Nagi, D. Markovic and S. S. Iyer, "Demonstration of a Low Latency (<20 ps) Fine-pitch (≤10 µm) Assembly on the Silicon Interconnect Fabric," 2020 IEEE 70th Electronic Components and Technology Conference (ECTC), Orlando, FL, USA, 2020, pp. 1801-1805, doi: 10.1109/ECTC32862.2020.00281.
- S. S. Iyer, S. Jangam, and B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for AI systems," in IBM Journal of Research and Development, vol. 63, no. 6, pp. 5:1-5:16, 1 Nov.-Dec. 2019.
- Boris Vaisband and S. S. Iyer, "Global and Semi-Global Communication on Silicon Interconnect Fabric", Proceedings of the IEEE/ACM International Symposium on Networks-on-Chip, pp. 15:1-15:5, October 2019.
- P. Gupta and S. S. Iyer, "Goodbye, motherboard. Bare chiplets bonded to silicon will make computers smaller and more powerful: Hello, silicon-interconnect fabric," in IEEE Spectrum, vol. 56, no. 10, pp. 28-33, Oct. 2019, doi: 10.1109/MSPEC.2019.8847587.
- Boris Vaisband and S. S. Iyer, "Communication Considerations for Silicon Interconnect Fabric," Proceedings of the ACM/IEEE International Workshop on System Level Interconnect Prediction, June 2019.
- Kannan K. Thankappan, B. Vaisband, S. S. Iyer, "On-Chip ESD Monitor", IEEE 69th Electronic Components and Technology Conference (ECTC), May 28-31, 2019, Las Vegas, NV.
- Saptadeep Pal, D. Petrisko, M. Tomei, P. Gupta, S. S. Iyer, and R. Kumar, "Architecting Waferscale Processors: A GPU Case Study", in 25th IEEE International Symposium on High-Performance Computer Architecture (HPCA), February 16-20, 2019, Washington D.C., USA.
- SivaChandra Jangam, A. Bajwa, K. K. Thankappan, P. Kittur, and S. S. Iyer, "Electrical Characterization of High Performance Fine Pitch Interconnects in Silicon-Interconnect Fabric," IEEE 68th IEEE Electronic Components and Technology Conference (ECTC), May 29-June 1, 2018, San Diego, CA.
- Saptadeep Pal, D. Petrisko, A. Bajwa, P. Gupta, S. S. Iyer, and R. Kumar "A Case for Packageless Processors", 24th IEEE International Symposium on High-Performance Computer Architecture (HPCA), February 24-28, 2018, Vienna, Austria.
- Saptadeep Pal, S. S. Iyer, and P. Gupta, "Advanced packaging and heterogeneous integration to reboot computing," in IEEE International Conference on Rebooting Computing (ICRC), November 8-9, 2017, Washington, DC, USA. (Invited)
- Arvind Kumar, Z. Wan, W. Wilcke, and S. S. Iyer, "Towards Human-Scale Brain Computing Using 3D Wafer Scale Integration," ACM Journal of Emerging Technologies in Computing, vol. 13, no. 3, article no. 45, Apr. 2017.
- Subramanian S. Iyer, "Heterogeneous Integration for Performance and Scaling," in IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 6, no.7, pp. 973-982, Jul. 2016. doi: 10.1109/TCPMT.2015.2511626

UCLA CHIPS Thanks you for your Support !

GLOBALFOUNDRIES

Kulicke & Soffa®

IBM®

MivaTek Global

TEL

tsmc

APPLIED MATERIALS®

Veeco

KEYSIGHT TECHNOLOGIES

ZUKEN®

semi™
Semiconductor Equipment and Materials International

ON
ON Semiconductor®

GINTI

i3 electronics

THE UNIVERSITY OF CALIFORNIA
LET THERE BE LIGHT
1868

DISCO
Kiru·Kezuru·Migaku Technologies

LOCKHEED MARTIN

MDLSoft Inc
LINKING MATERIALS TO DEVICE AND SYSTEMS

Nissan Chemical
America Corporation

ELECTRON DEVICES SOCIETY®

nbmc
Nano-Bio Manufacturing Consortium

Atoms to devices

COVENTOR®
A Lam Research Company

TECHNIC

nanotronics

MaXentric
Technologies LLC

IEEE ELECTRONICS PACKAGING SOCIETY

DARPA

DEFENSE THREAT REDUCTION AGENCY
UNITED STATES OF AMERICA

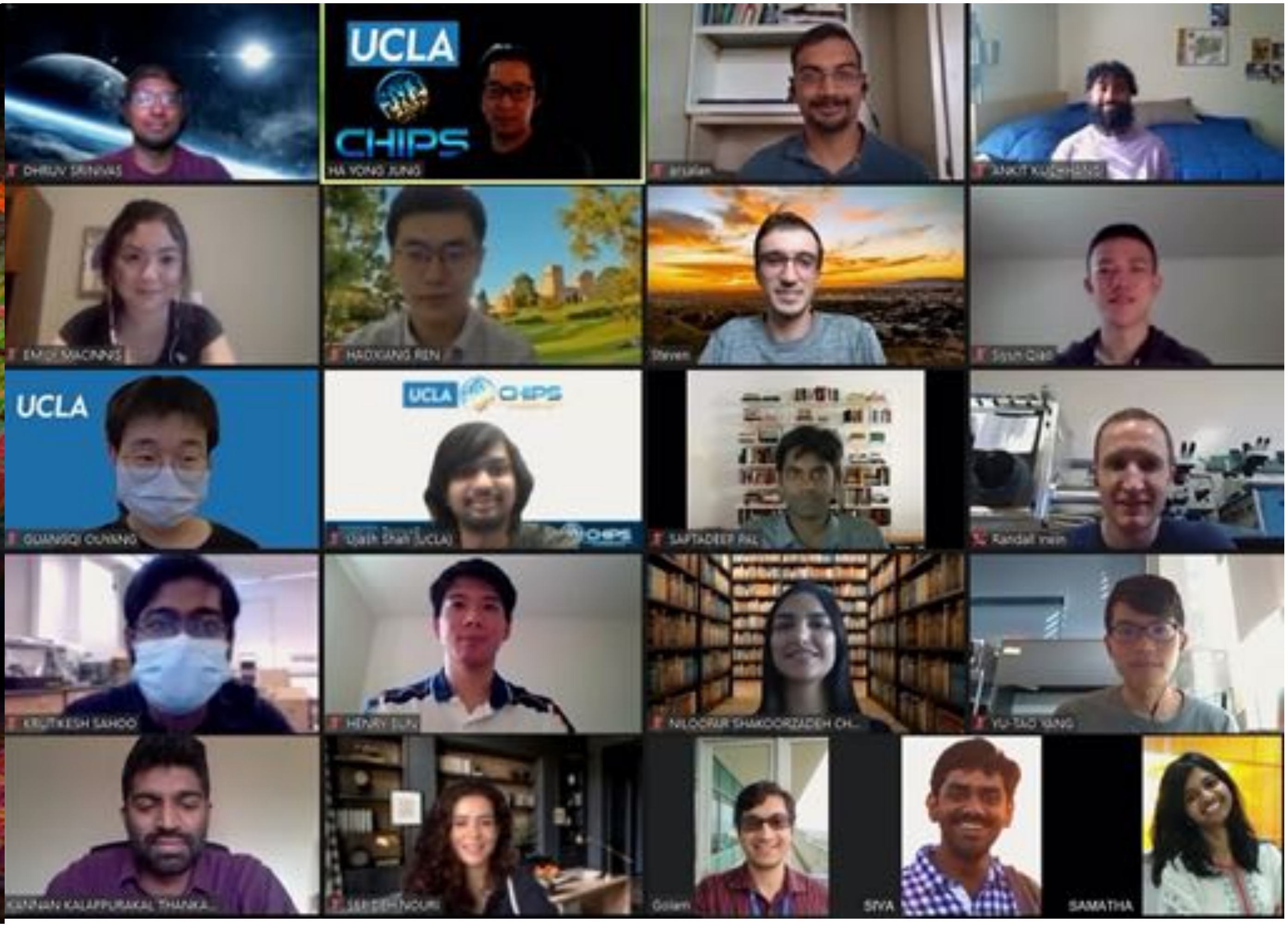
Office of Naval Research
USN
ONR
Science & Technology

SRC®
Semiconductor Research Corporation

FlexTech
SEMI Strategic Association Partner

UCLA Samueli School of Engineering

CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION AND PERFORMANCE SCALING



Our Students

In a recent zoom Group meeting