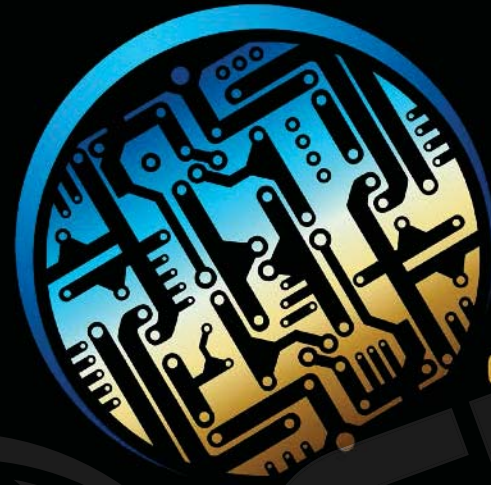


UCLA



CHIPS

**CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING**

2019 UCLA CHIPS Report

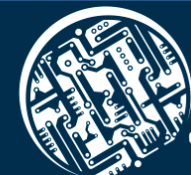
Subramanian Iyer

s.s.iyer@ucla.edu

CHIPS

UCLA

Samueli
School of Engineering



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

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

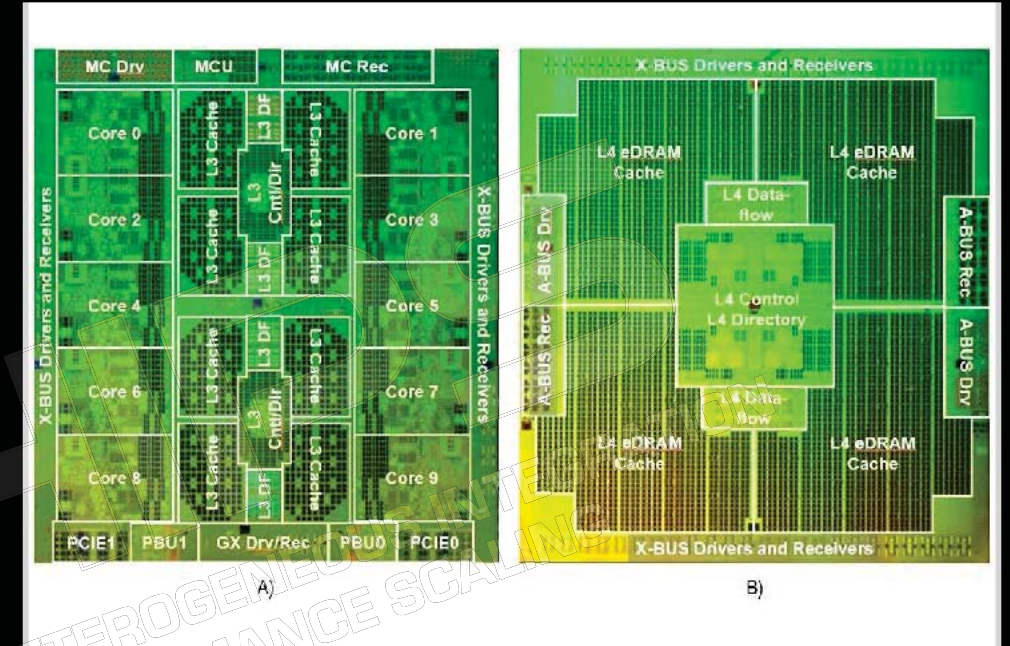
Silicon as a
heterogeneous fine pitch
packaging Platform, Si IF

FlexTrate as a flexible
Biocompatible Heterogeneous
Integration Platform

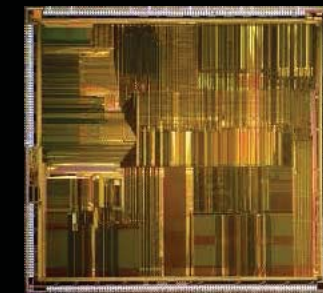
The CTT as an in-
memory compute
device

Some observations

- If Moore's law has enabled miniaturization, why have chips gotten larger ?
 - More complex problems
 - More cores
 - More cache memory
- Main memory capacity and access limits performance
- I/Os take up more space and power as system size increases
- Power density and thermal challenges limit performance



Eg. IBM Z14 cpu and cache are each $\sim 700 \text{ mm}^2$ with 6.8B Xtors (2018) and a separate cache die

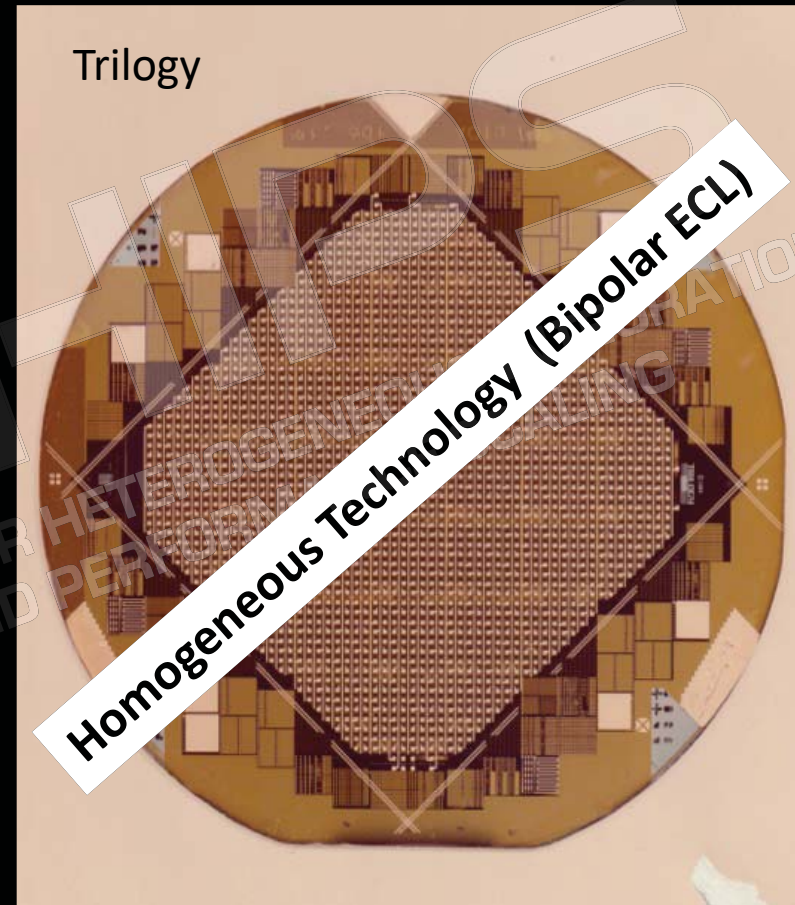


Intel Pentium cpu $\sim 300 \text{ mm}^2$ -3.1 Million Xtors (1993)

The “Early” Origins of Wafer Scale Integration



Gene Amdahl with Michael Current (2008)



Bumped 100 mm wafer (Ca 1982)

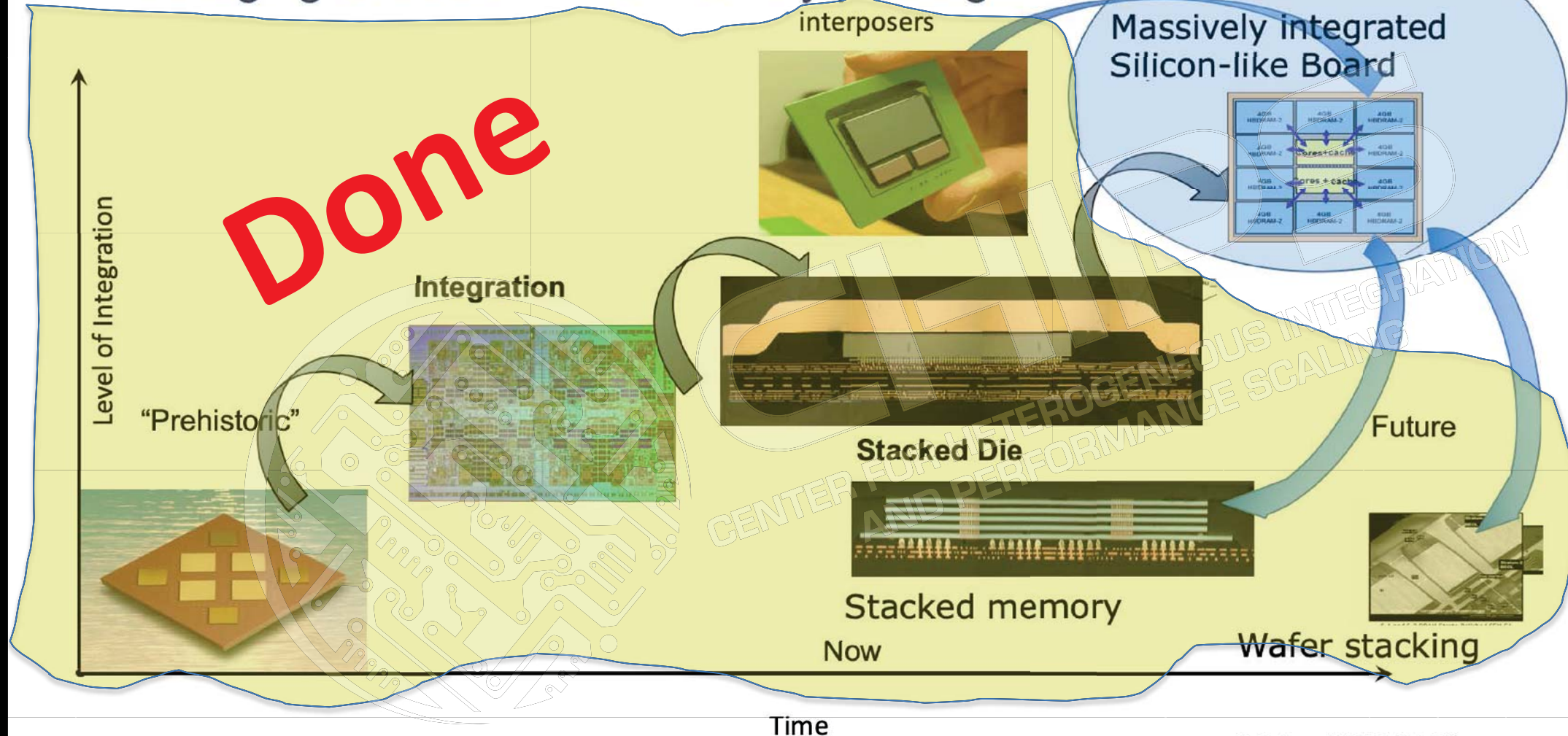
The public talk at the time (and echoed in the (generally ill-informed) modern blogosphere) was that the motivation was processing speed. Going on about reduced interconnect line length, etc.

But the reality was that computer system speed was dominated by architecture and memory access (not much change in that in ≈ 40 years). The real advantages were in "board" physical size and the mechanical ruggedness of the assembled wafer package.

One could put the logic of an entire PDP-11 computer on a single wafer and DEC (one of the supporting consortium) was pleading with Amdahl to build some modules for sale to the US Air Force for airborne radar processing (bombers have lots of electrical power on board). But Gene kept after the goal of building an IBM-scale mainframe (a la his earlier Amdahl Corp that Fujitsu took over). And we lacked the scale to compete with IBM out of the box. So it goes.

(- Michael Current via email)

Packaging Evolution - The memory paradigm



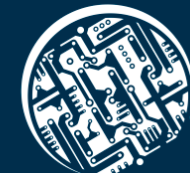
©s.s.ilyer (2018)

Advance Packaging Tutorial IEDM 2018 (Iyer)

S.S. Iyer IEDM (2012)

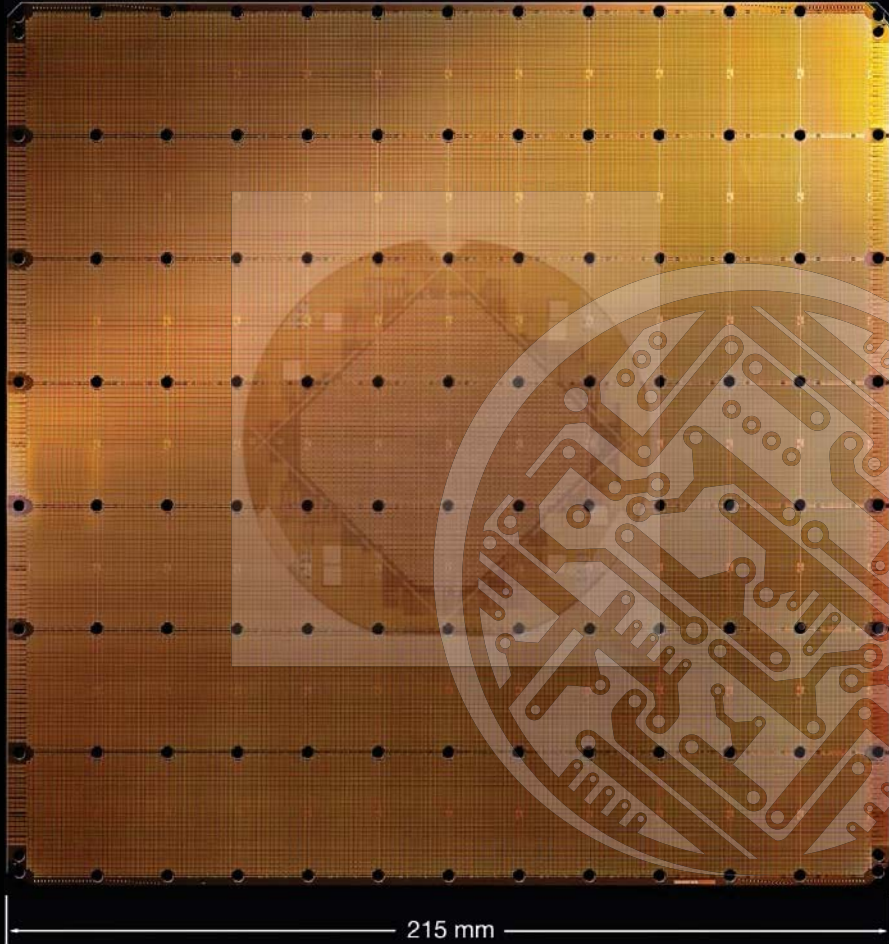
UCLA

Samueli
School of Engineering



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

Fast Forward to today



Cerebras has taken WSI to a new scale

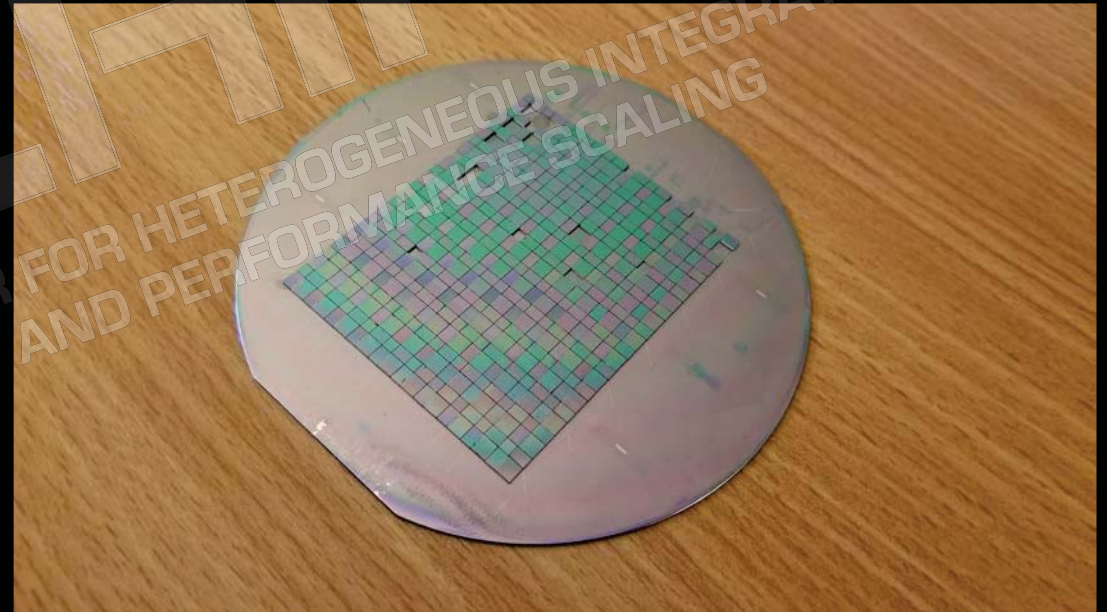
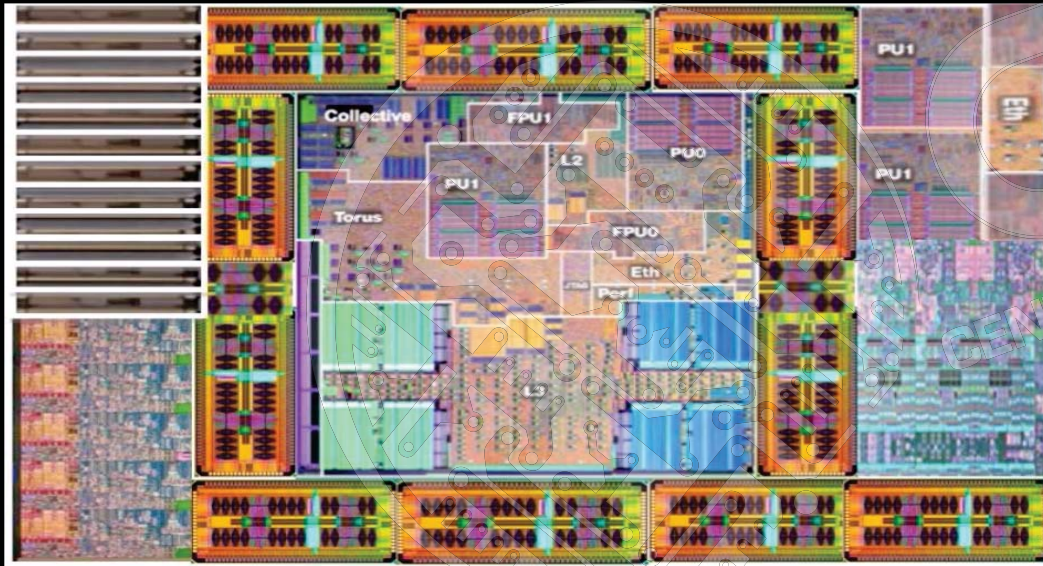
And addressed a significant number of problems

This is absolutely the right first Step

But at its core the system is still homogeneous and probably memory starved

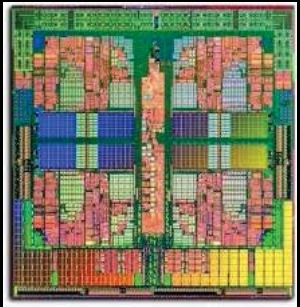
Wafer Scale Integration @UCLA CHIPS

Lets revisit this concept but
with a different approach



Chiplets or Dielets ?

A **die**, in the context of integrated circuits, is a small block of semiconducting material on which a given functional circuit is fabricated. We usually think of dies as bare. Dies are diced out of a wafer



A chip is more or less the same thing but connotes the design aspects rather than the physical attributes. We usually think of chips as packaged.



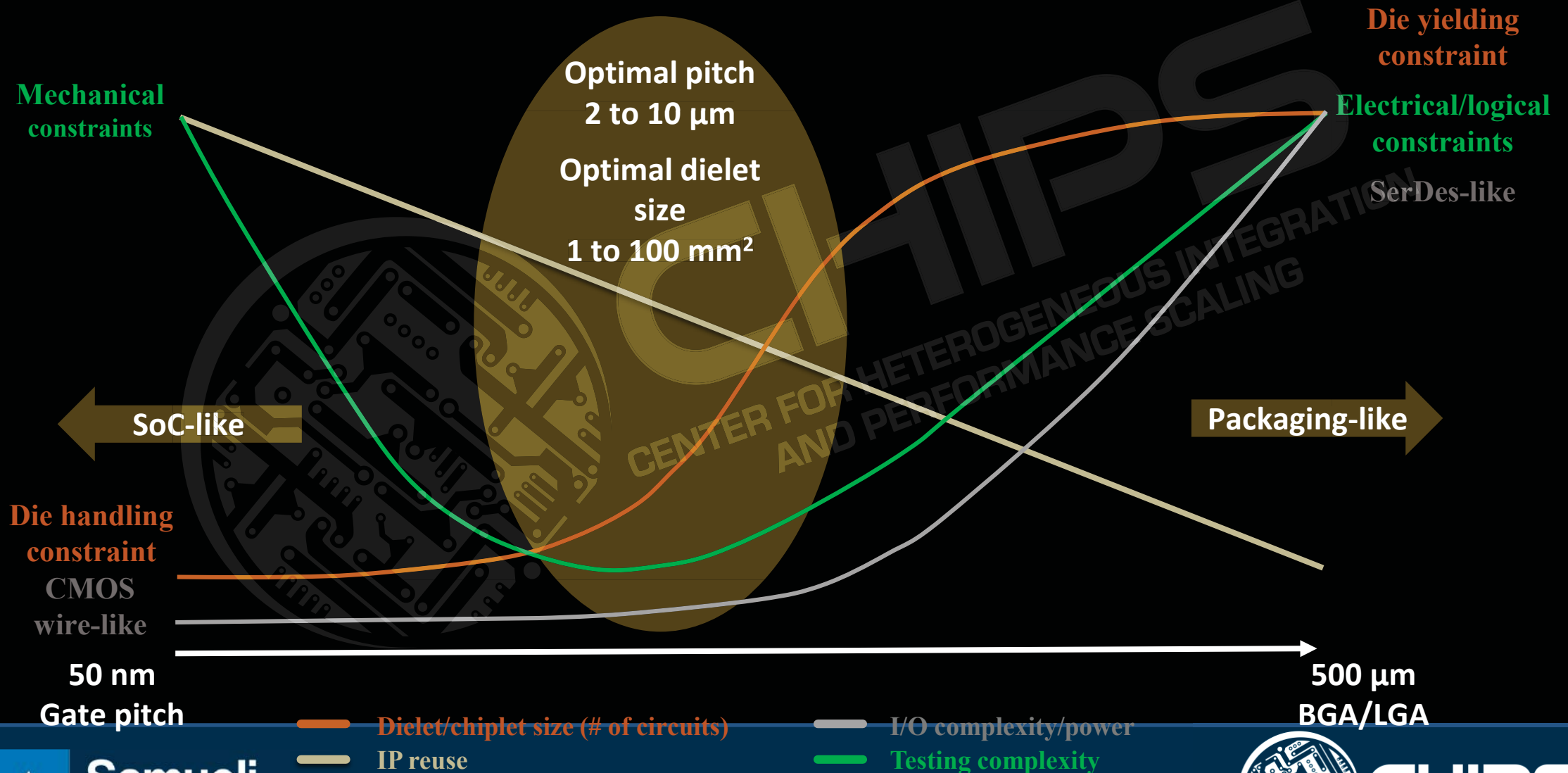
We will use these terms interchangeably though dielet is more correct

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

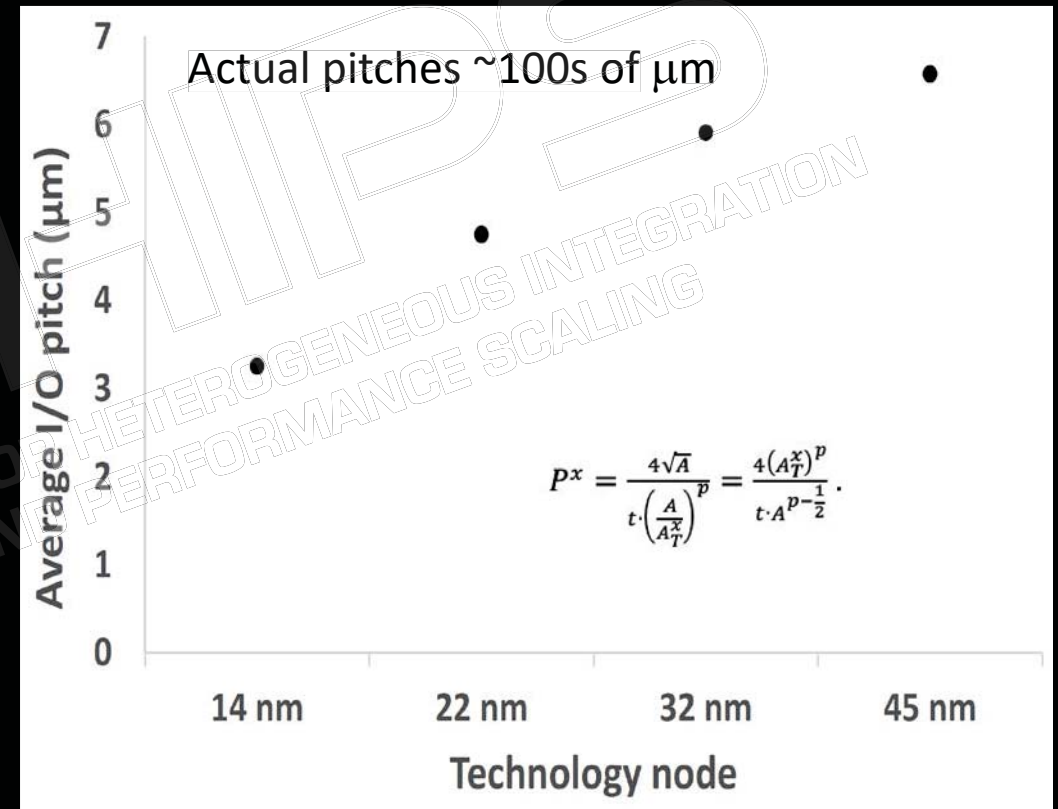
Hint: how do we make a SOW look like an ginonormous SOC

The Dielet Golden Regime



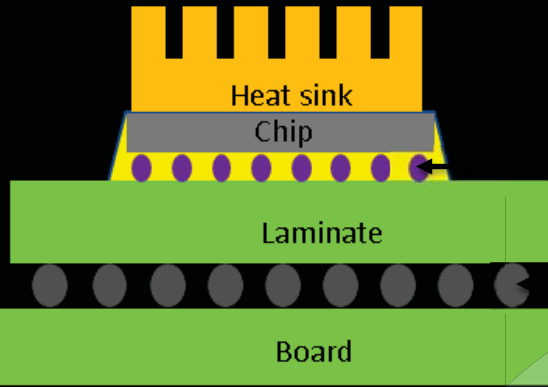
What is the optimal I/O pitch ?

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	

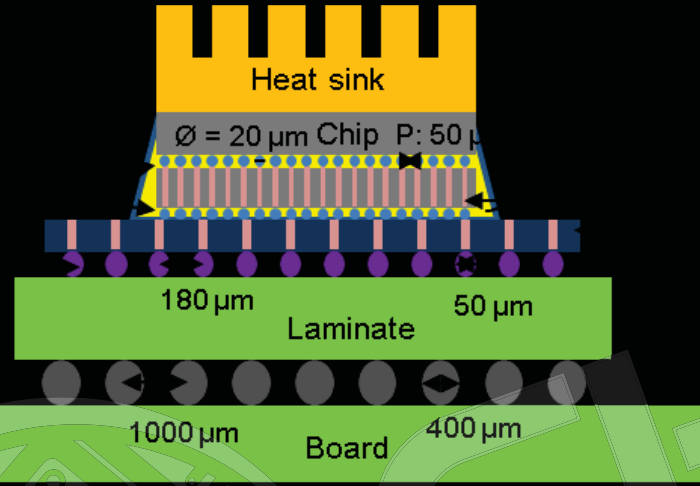


Iyer, Jangam & Vaisband, IBM J R&D 2019

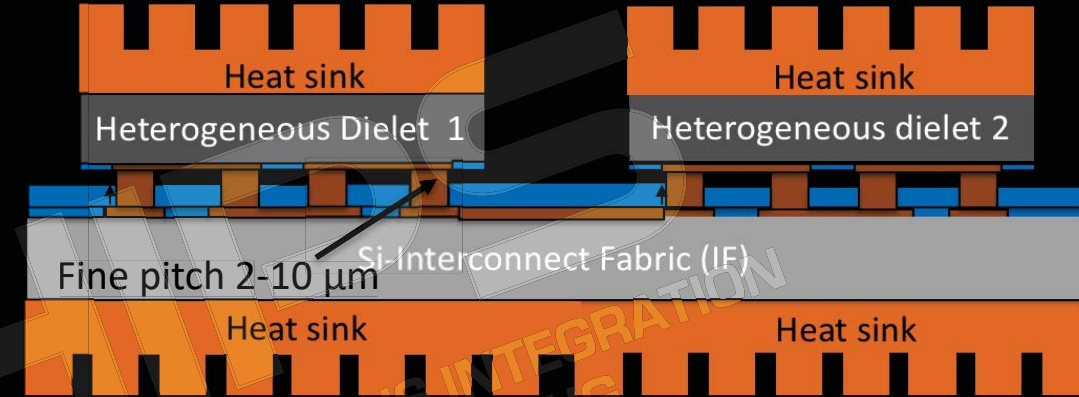
PCBs



Interposer



Si-IF

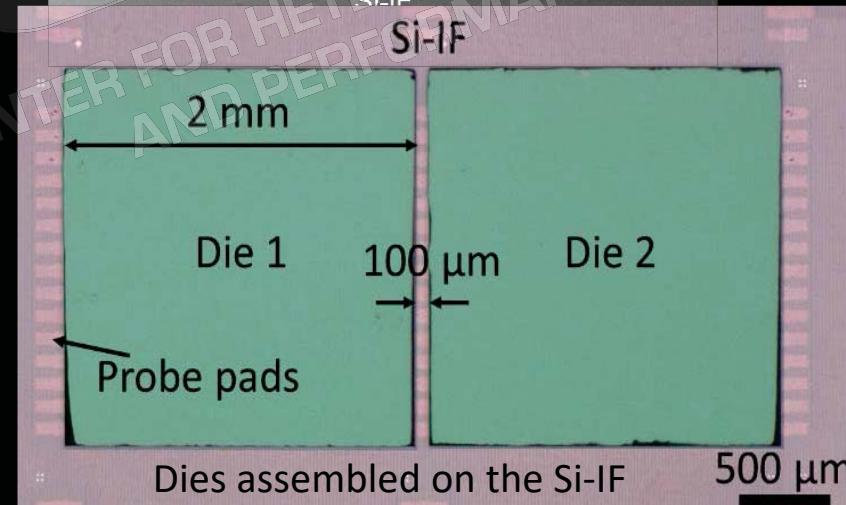
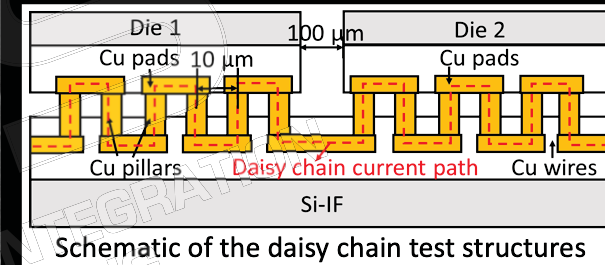
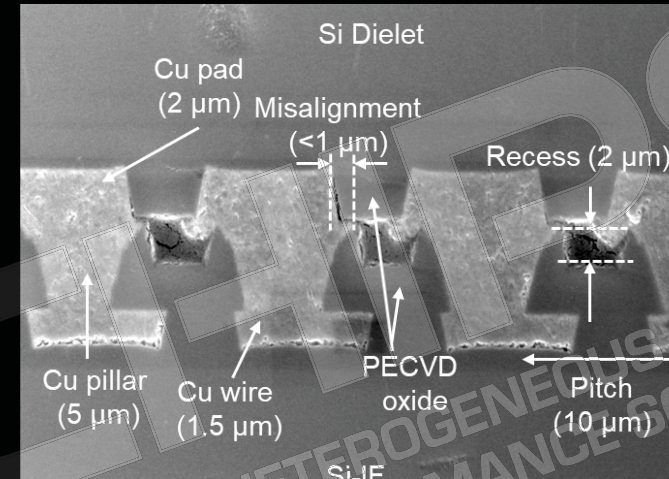
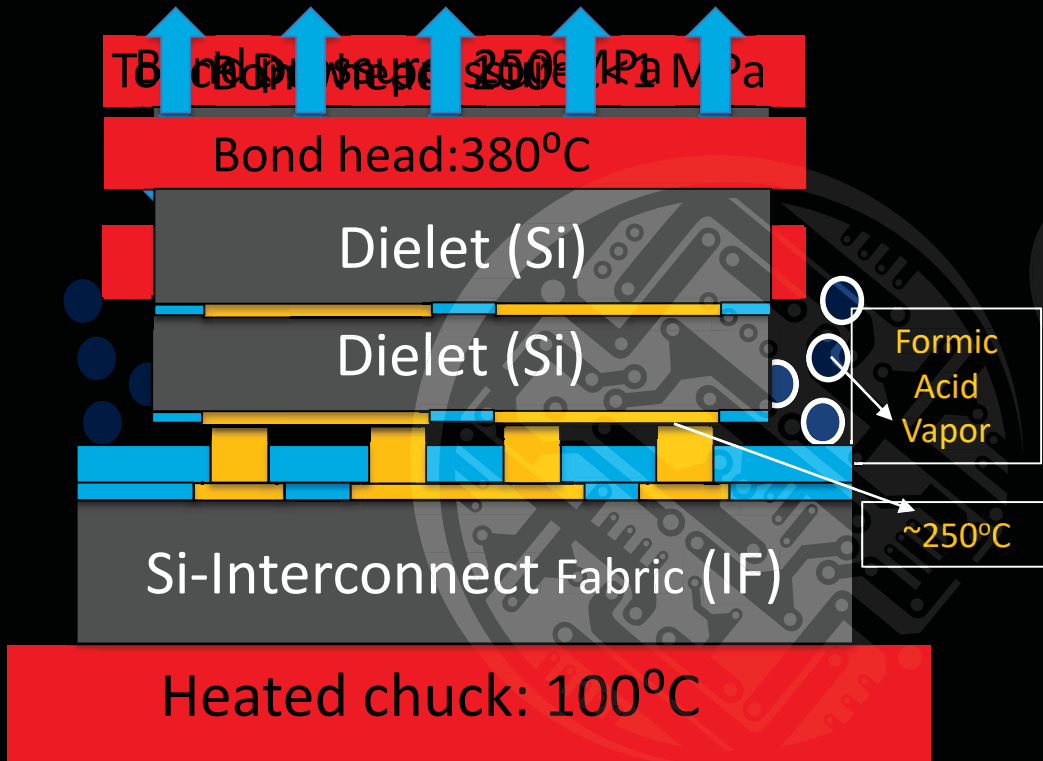


- Multiple packaging levels
- Disparate materials (Si, Cu, FR4, Molding compound)
- Limited heat sinking
- Solder based interconnects
- Underfill needed
- Die-to-die connections limited by BGA pitch

- Additional level in packaging hierarchy
- **All PCB limitations still exist** especially thermo-mechanicals
- Limited in size: Sub-system integration
- Requires fine TSVs that add cost
- Interconnect pitch: $50 \mu\text{m}$

- Single package level
- Mainly three materials (Si, Cu, Oxide)
- Excellent heat sinking
- Metal-metal interconnects
- No underfill
- Allows for Wafer Scale Integration
- Interconnect pitch: $2\text{-}10 \mu\text{m}$

We do this by Thermal Compression Bonding



Pillar resistance: 35 mΩ
Shear strength 160 MPa

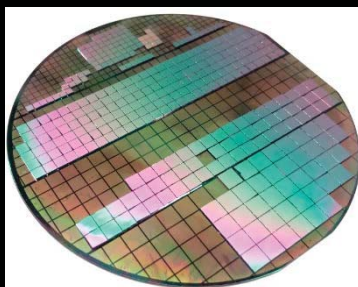
The eco system

- The Silicon IF is a versatile platform for heterogeneous integration, but several issues remain
 - Testing, Repair and Reliability
 - Communication on and off the IF
 - Clock distribution
 - Power delivery and heat extraction
 - Dielet supply-chain and ecosystem

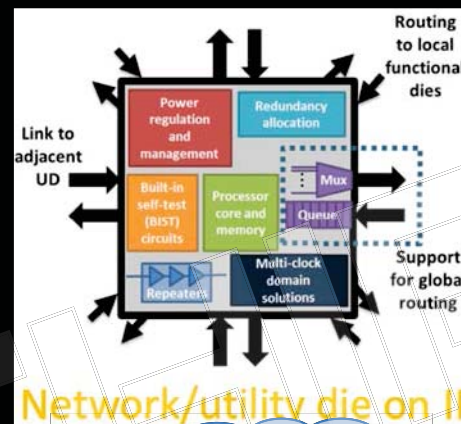
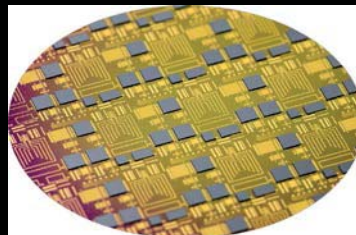
A 50KW High Performance Wafer Scale System on Si IF



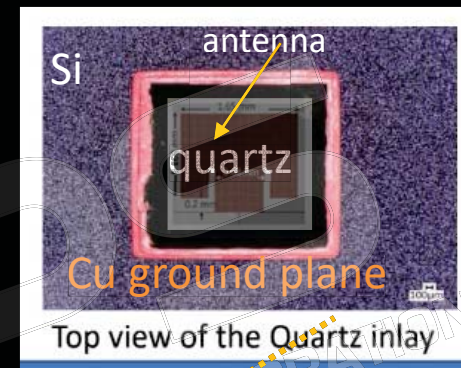
Connector collar



Populated Si-IF (Si and III-V dies)

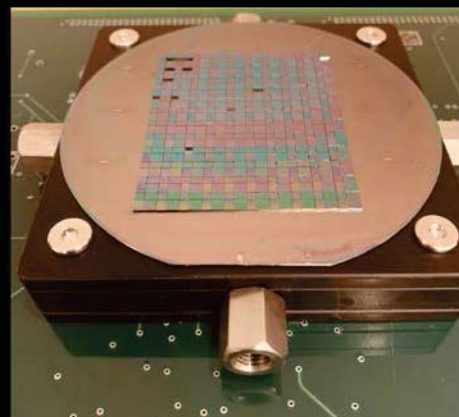


Network/utility die on IF

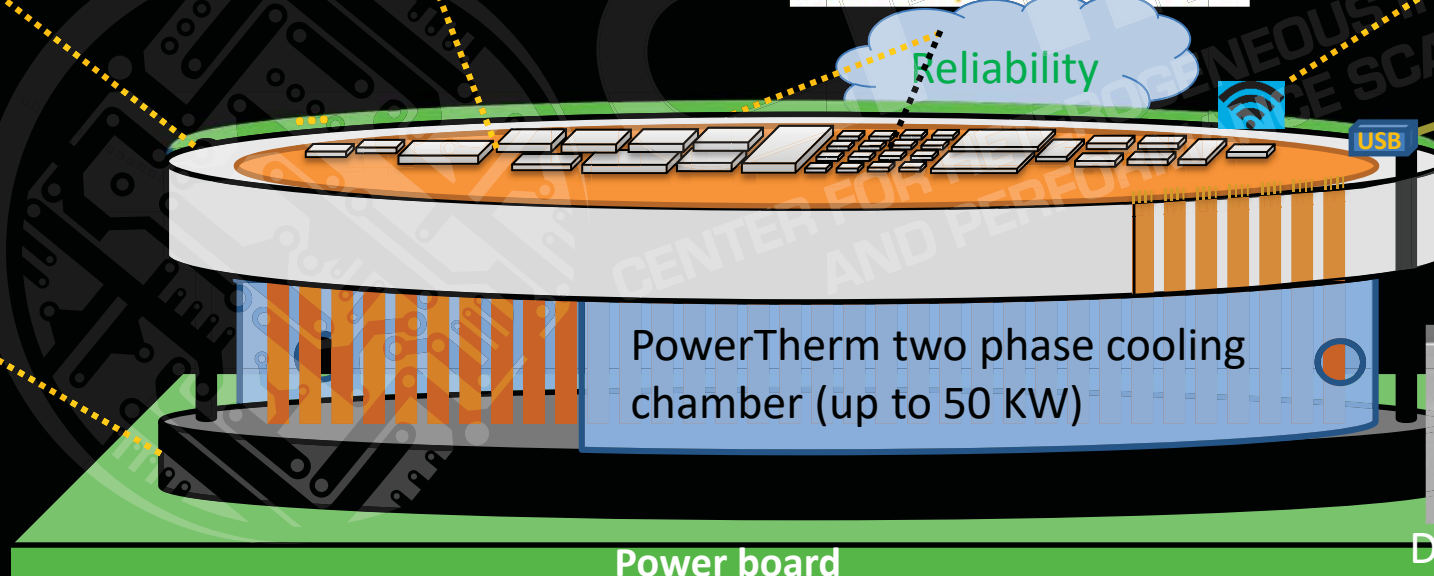


RF interconnect

Photonic interconnect

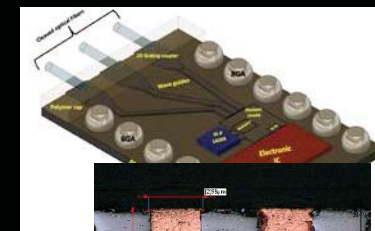


PowerTherm Assembly

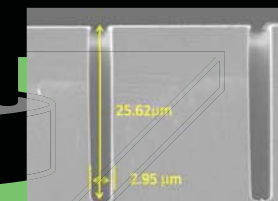


Power board

PowerTherm two phase cooling chamber (up to 50 KW)

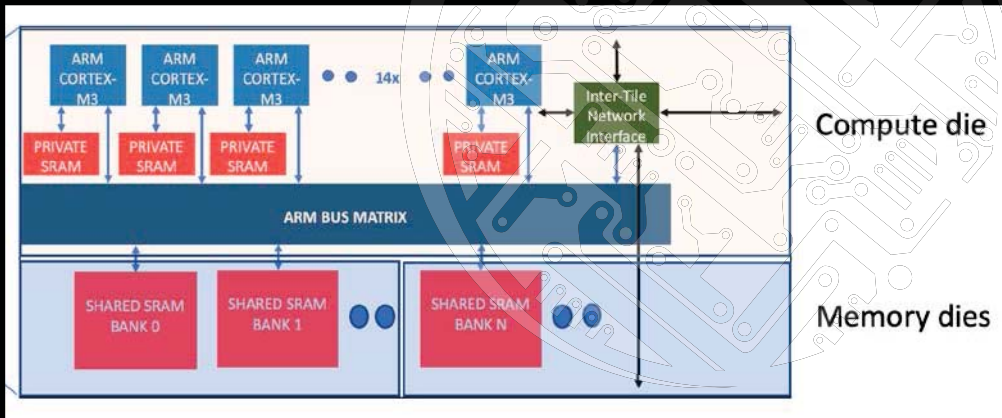
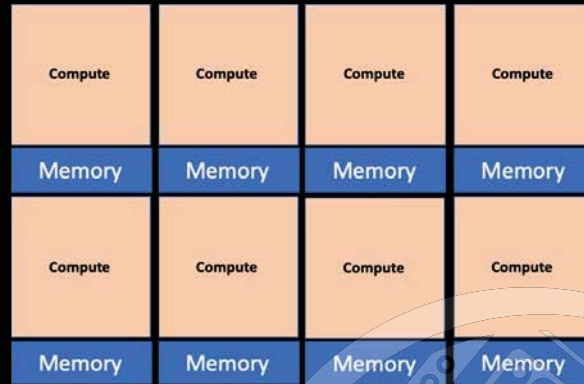


TWVs to deliver power to front



Deep Trench decoupling

A Modest Graph Processor implementation



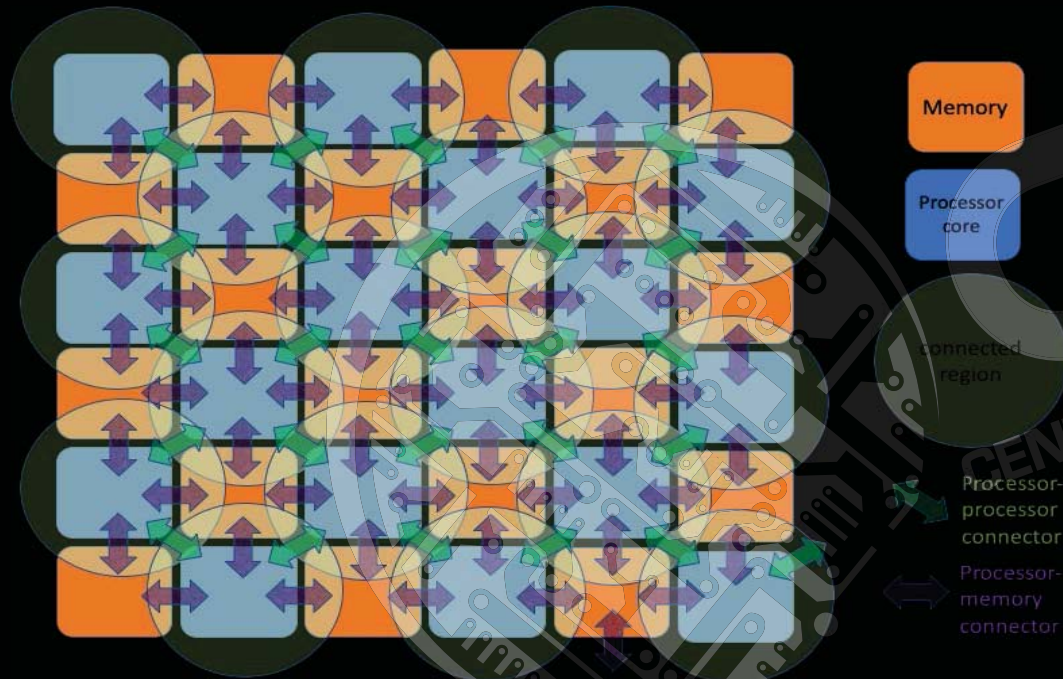
Design Details

Compute dielet area	6 mm ² (mostly memory)
Memory Dielet area	2 mm ²
Power per tile	0.5W
tiles per 100mm wafer	625
# of ARM cores	8750
Memory BW Network BW	9 TB/s 8TB/s

To be taped out in TSMC 40nm
Technology in January 2020

To be Assembled at UCLA using the SI IF
at 10 μ m die to wafer pitch Cu pillars

The “Ultimate” System on Si IF



3D stacked combination of memories (DRAM, emerging etc.)

With additional High performance compute in the bottom layer

Memory capacity >12 TB

Intranode BW >1TB/s

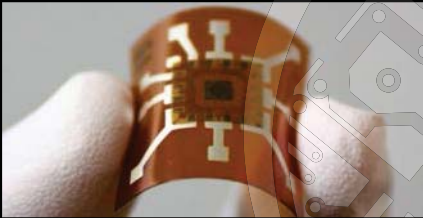
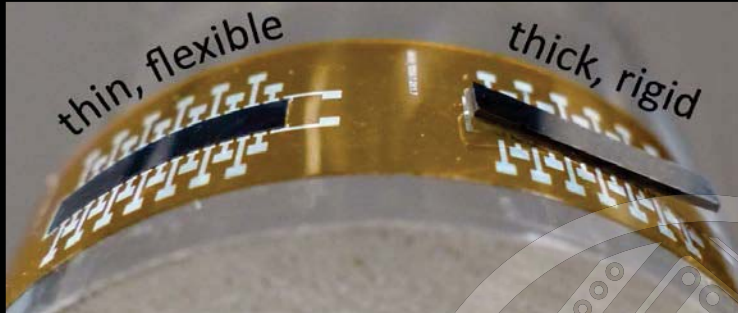
Internode BW > 7.5 PB/s

PowerTherm cooled

Scalable to stacked system

Gen 2 FHE* Packaging for Medical Electronics

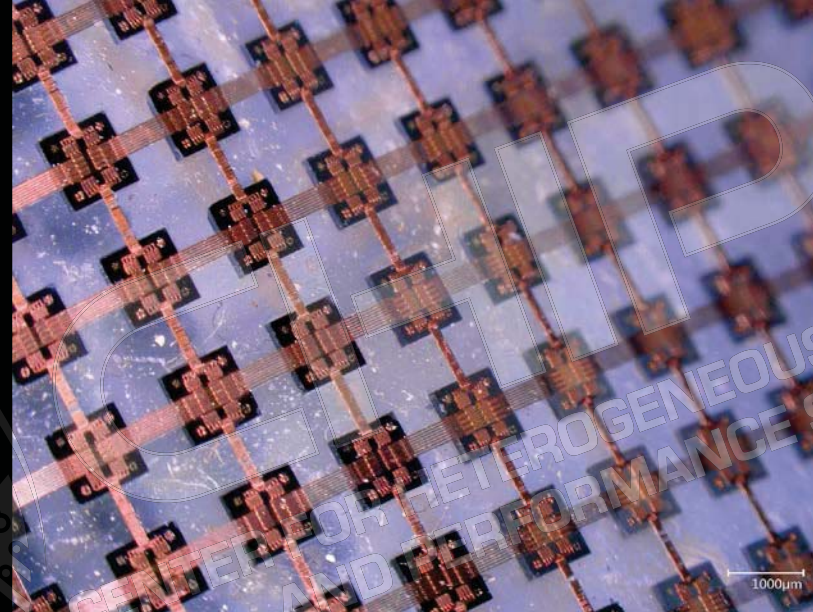
Today



Printed electronics & wiring, ultra-thin die, organic thin film semiconductors - at best bendable

* Flexible Hybrid Electronics

FlexTrate™ @ UCLA



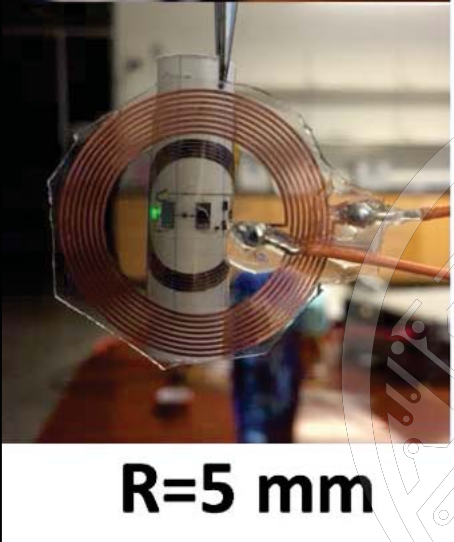
- Fan-Out Wafer Level Packaging with ultralow die shift
- Dielet concept on flexible molding compound (PDMS) – bicycle chain principle
- Corrugated lithographically defined high performance interconnects
- Thermally conductive PDMS using metallic nanowires
- Biocompatible
- Transparent

UCLA

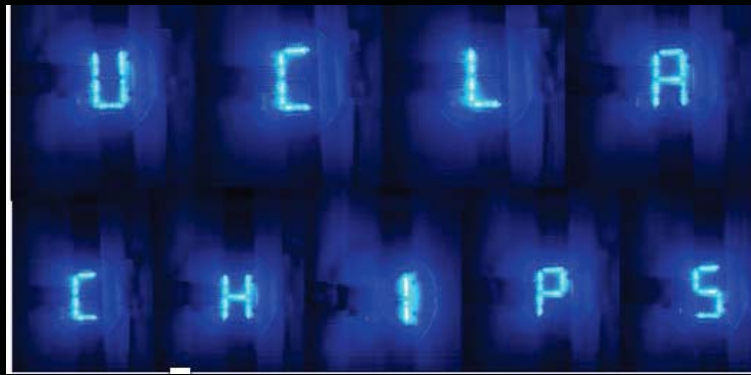
Samueli
School of Engineering



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

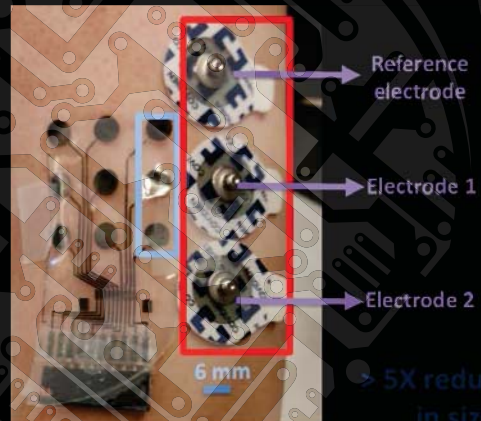


Wireless Power Transfer



Segmented flexible displays

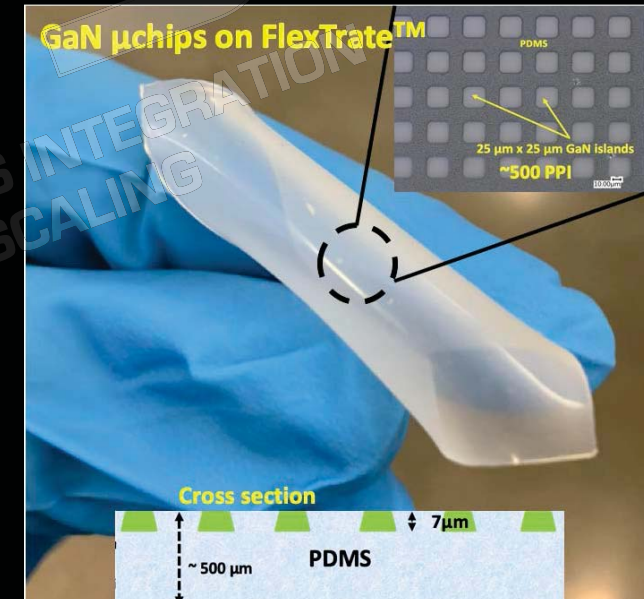
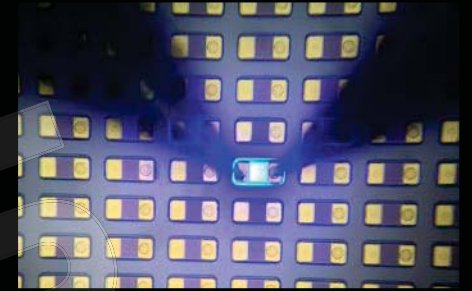
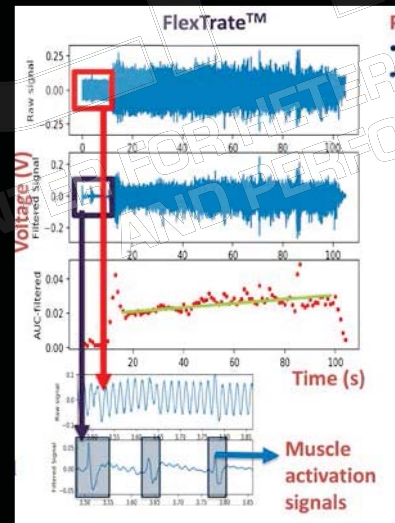
Applications



Size comparison for 1 channel

5X reduction in size

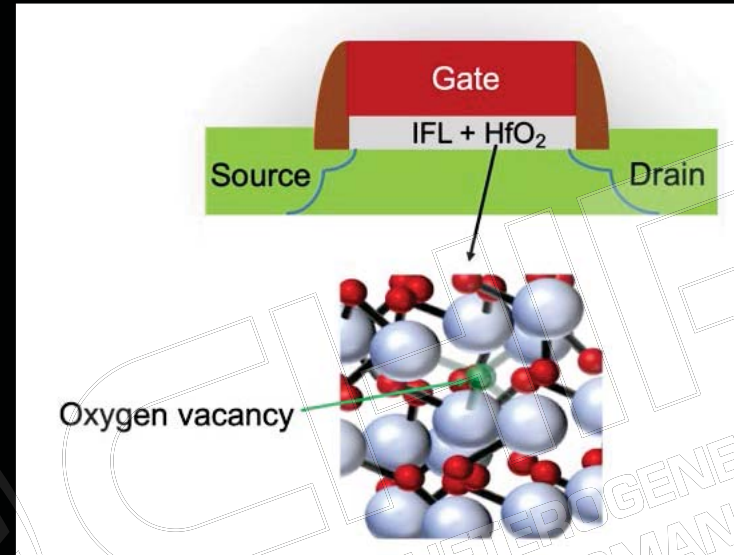
Wireless Surface electromyography



Flexible μ LED displays fabricated using Laser Lift Off mass Transfer

Background of CTT

- Novel multi-time-programmable non-volatile memory element for HKMG CMOS technologies
- CTTs are as-fabricated CMOS logic devices operated under enhanced charge trapping mode
- Intrinsic device self-heating enhanced charge trapping in HKMG
- Use as an analog memory element for unsupervised learning and in memory compute

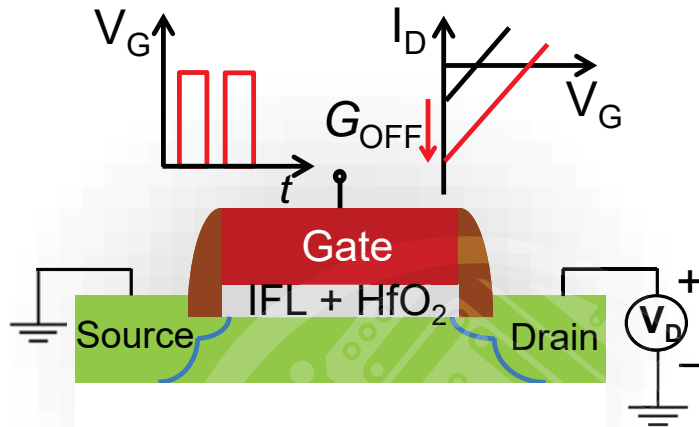


	WL	BL	SL
Write	~2V (VPP)	0V	~1.5V (VSL1)
Read	1V (VDD)	Floating	1V (VDD)
Erase	~-1V (VWL)	Floating	2V (VSL2)
Standby	0V	Floating	0V

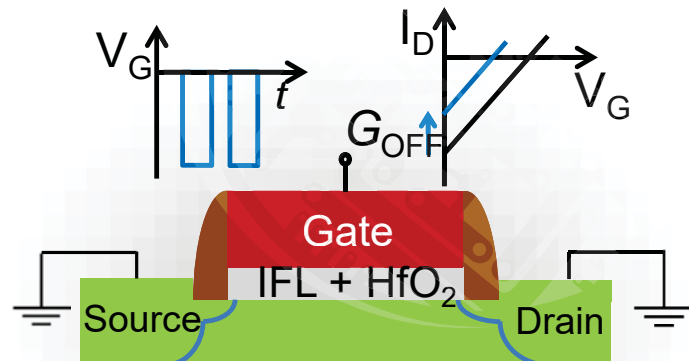
CTTs have been demonstrated as digital MTPM in multiple nodes - And is being productized

Use of CTT as Analog Memory Device

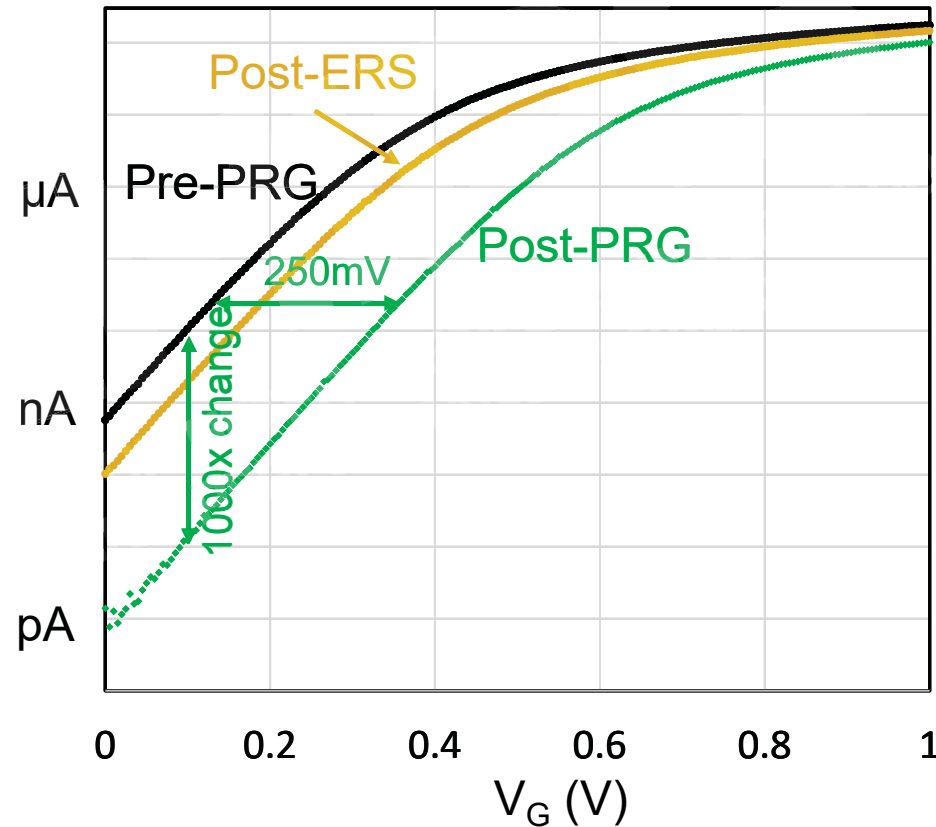
Increasing threshold voltage:



Reducing threshold voltage:



The CTT can be programmed and erased



NeuroCTT V1 (in GF 22FDX)

- An inference engine
- with 1024 x 20 CTTs
 - Fabricated in GF 22FDX
- Die size: 2.5x2 mm²
- Packaging: Wire-bond
- Lots of test macros

Test macros:
Devices, arrays, WL driver, neuron, etc.

Neuron Macro

Array macro
(170nm)

INPUT
BUFFERS

DIGITAL
BLOCK

WL DRIVER
& ARRAY

OUTPUT
BUFFERS

DECAPS

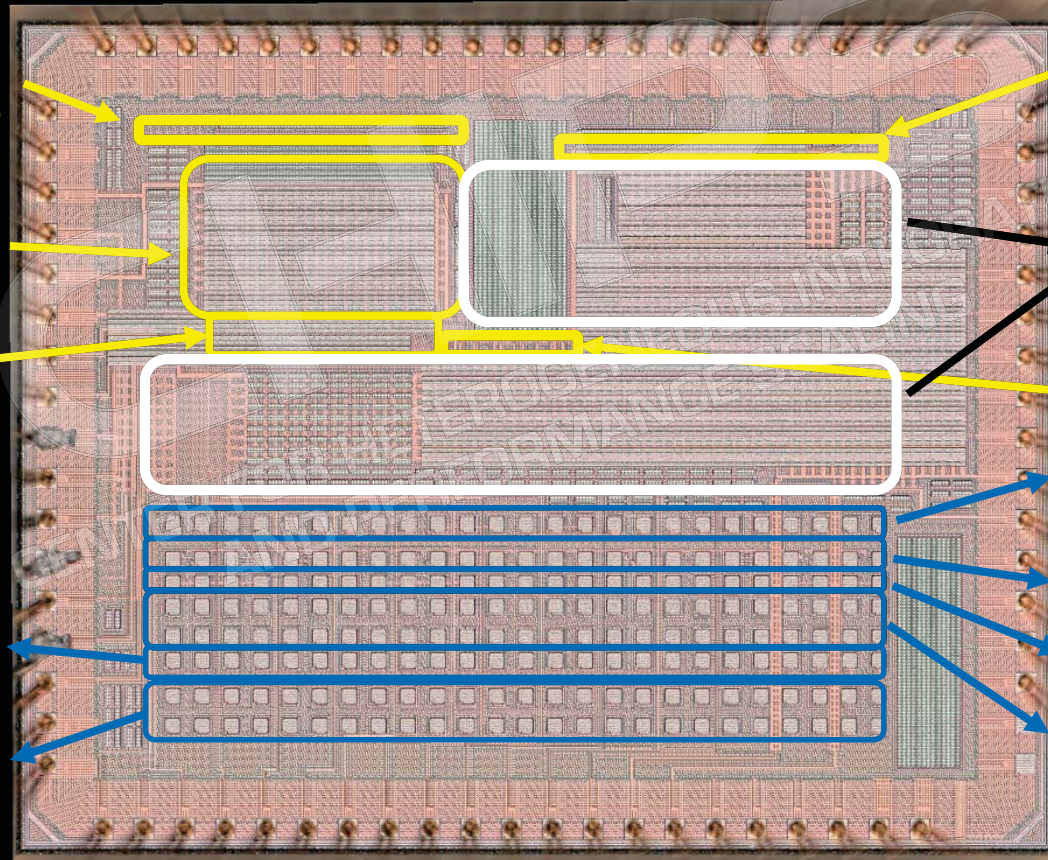
NEURON

WL Driver Macro

ESD Macro

Device macro

Array macro
(428nm)



Schematic of inference array

The schematic illustrates the inference array architecture. Input data is encoded as binary pulses of variable width, represented by a graph of $T \propto \sum_{i=1}^N I_i t_i$. These pulses are fed into an integrator block. The output of the integrator is connected to a network of neurons. Each neuron consists of a synaptic array (SL) and a feedback loop (BL). The synaptic array is connected to the integrator output via a synaptic weight $\Delta I = I_i - I_j$, which represents the weight of the (i,j) -th synapse. The neurons are connected to each other via a feedback loop (BL) and a synaptic array (SL). The diagram also shows the input data as a series of pulses (WL1, WL2, WL3) and the output of the neurons as a series of pulses (BLt1, BLt2, BLt3).

CTT array Vector-Matrix Multiplication (VMM)

This scatter plot shows the measured VMM value (nA) versus the target VMM value (nA) for the CTT array. The data points are categorized by time after programming: After 2hrs (blue dots), After 20hrs (red dots), and After 200hrs (green dots). A solid black line represents the ideal case $y = x$. The plot demonstrates that the measured VMM value closely follows the target VMM value, indicating high accuracy in vector-matrix multiplication.

Room Temperature stability (< 6% over 10 years)

This scatter plot shows the standard error (Std_err) / range (%) versus time after programming (hr) for the CTT array. The data points are categorized by time after programming: After 2hrs (blue dots), After 20hrs (red dots), and After 200hrs (green dots). A dashed red line represents the model. The plot demonstrates that the standard error / range (%) remains below 6% over a period of 10 years, indicating high room temperature stability.

Figure 1 illustrates the schematic diagram of the proposed CTT synapse. The diagram shows the internal structure of the synapse and its connection to two neurons.

The synapse consists of an integrator block and a "Twin-Cell" CTT Synapse block. The integrator block receives input data encoded as binary pulses of variable width (WL1, WL2, WL3) and outputs BLt1 and BLc1. The "Twin-Cell" CTT Synapse block receives BLt1 and BLc1 and outputs BLt2 and BLc2.

The BLt2 and BLc2 signals are then fed into two neurons, Neuron 1 and Neuron 2. Neuron 1 has inputs BLt2, SL2, and BLc2. Neuron 2 has inputs BLt3, SL3, and BLc3. The output of Neuron 1 is I_+ and the output of Neuron 2 is I_- .

The difference between I_+ and I_- represents the weight of the (i,j)-th synapse, $\Delta I = I_+ - I_-$.

Scatter plot showing Measured VMM value (nA) on the Y-axis versus Target VMM value (nA) on the X-axis. The plot includes data points for three time points: After 2hrs (blue), After 20hrs (red), and After 200hrs (pink). A solid black line represents the identity line $y = x$. The data points are clustered around the identity line, indicating a strong positive correlation between the measured and target VMM values. The 'After 200hrs' data points are generally higher than the 'After 2hrs' data points, suggesting an increase in VMM value over time.

Figure 6 Data Summary:

Time after programming (hr)	Std_err (%)	Type
~2	~4.05	Measurement
~20	~4.28	Measurement
~200	~4.75	Measurement
-	-	Model Fit

Benchmarking NeuroCTTs

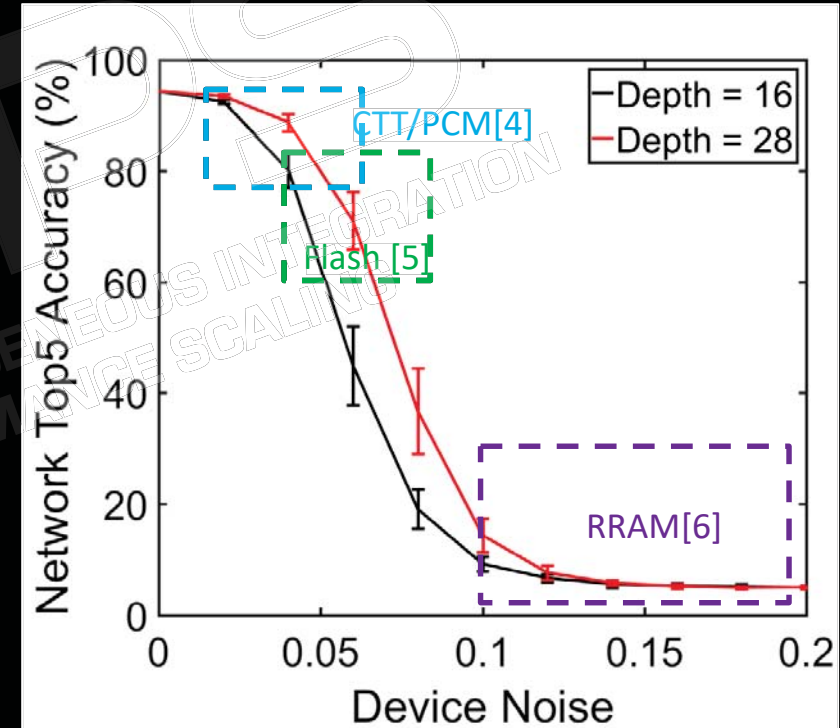
CTT is very competitive for neural network apps.

- CMOS-compatible and ready at advanced node
- Third terminal (access switch) is free
- Acceptable variation (no dead cells like in RRAM [6])
- Energy and area efficient both for the device and the periphery
- Accurate programming (including retention)

Using NeuroSim [1] (with Prof. Shimeng Yu @GaTech)

	PCM [2]	STT-MRAM [3]	RRAM	Twin-CTT cell	CTT vs. PCM
Tech.Node	32nm	32nm	32nm	22nm	N/A
Cell bit	4bit	1bit	4bit	4bit	1x
R _{ON}	40K	14.8K	6K	41.67K	~1x
Chip area (mm ²)	16.26	76.13	28.11	8.11	0.499x
Latency (ms)	2.75	8.42	16.77	1.63	0.593x
Energy efficiency (TOPS/W)	9.92	1.347	2.79	17.96	1.81x
Throughput (FPS)	363.02	118.65	59.61	612.87	1.69x

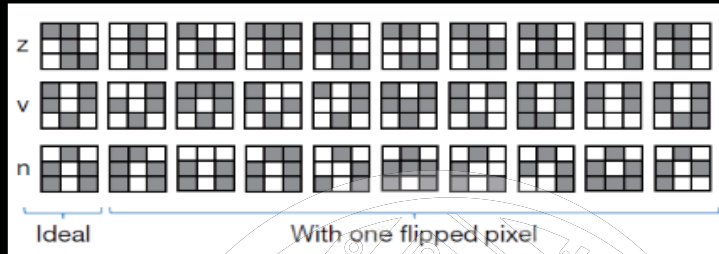
Accuracy of ResNet Using Analog Devices



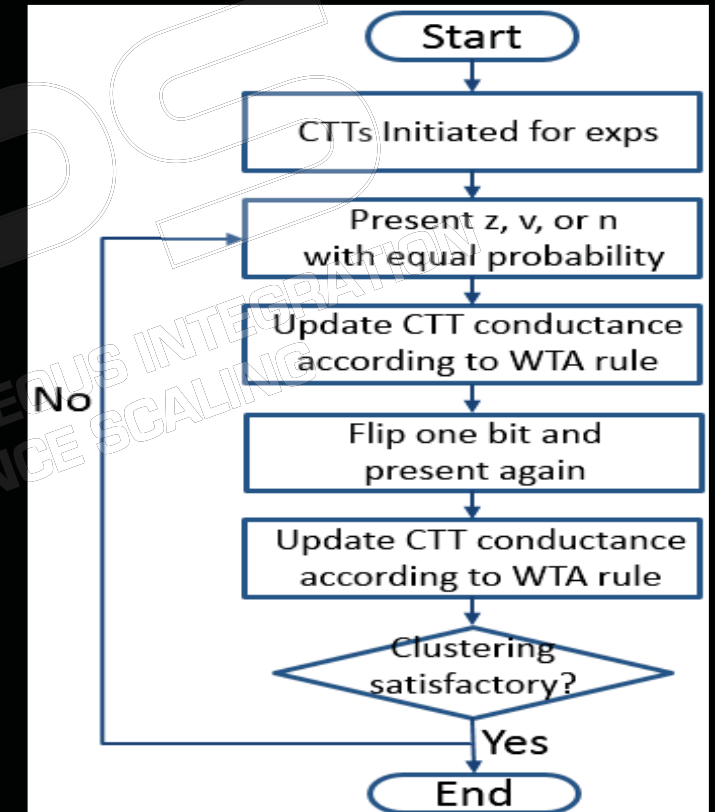
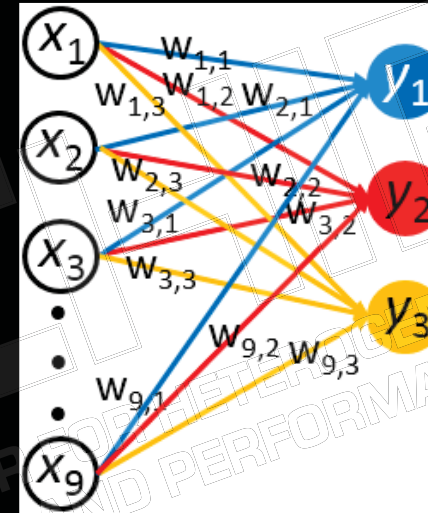
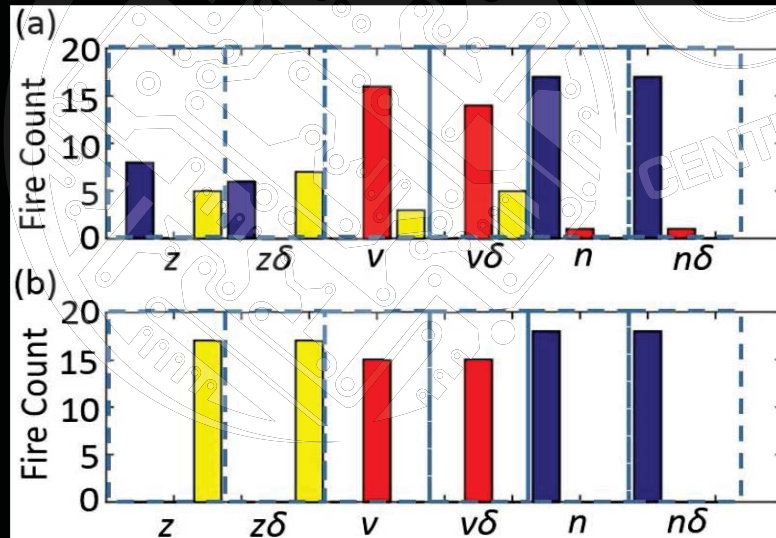
[1] Chen, Pai-Yu, Xiaochen Peng, and Shimeng Yu. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 37.12 (2018): 3067-3080. [2] G. W. Burr *et al.*, 2014 IEDM, pp. 29.5.1-29.5.4. [3] Y. Kim *et al.*, "Integration of 28nm MJT for 8~16Gb level MRAM with full investigation of thermal stability," 2011 Symposium on VLSI Technology - Digest of Technical Papers, Honolulu, HI, 2011, pp. 210-211. [4] G. Burr, *et al.*, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 2016 [5] X. Guo, *et al.*, IEDM 2017 [6] X. Zheng, *et al.*, IEDM 2018

Unsupervised Learning Using CTT

- Proof-of-concept network simulated using CTT characteristics to cluster stylized letters z, v, n, and their noisy versions

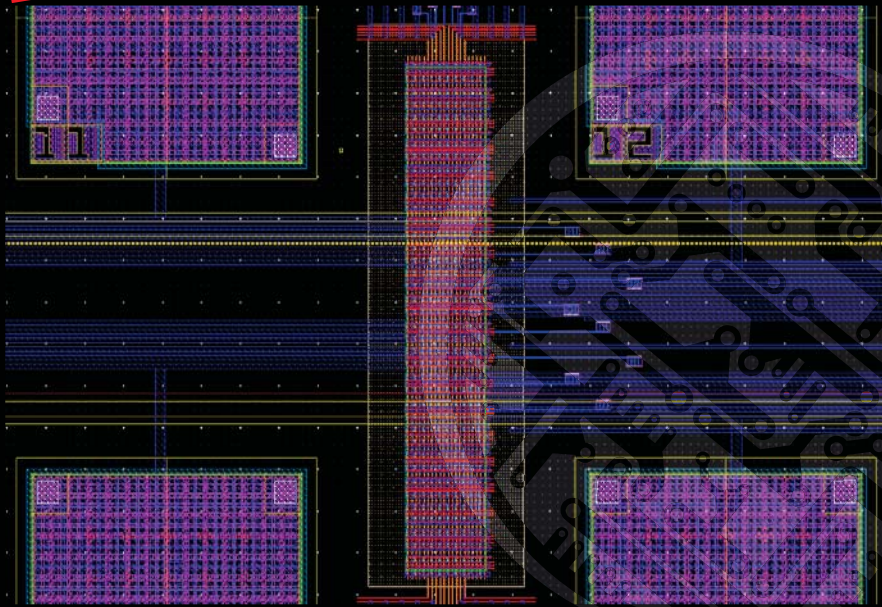
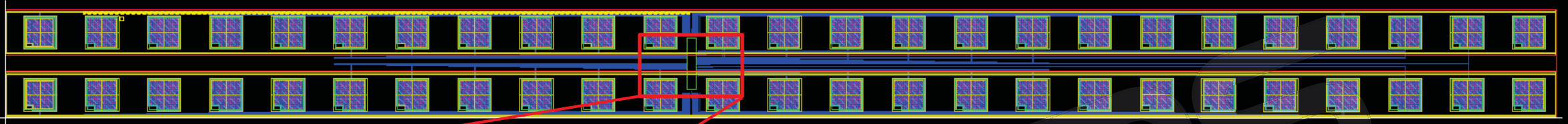


- Perfect clustering achieved after on average 24 presentations

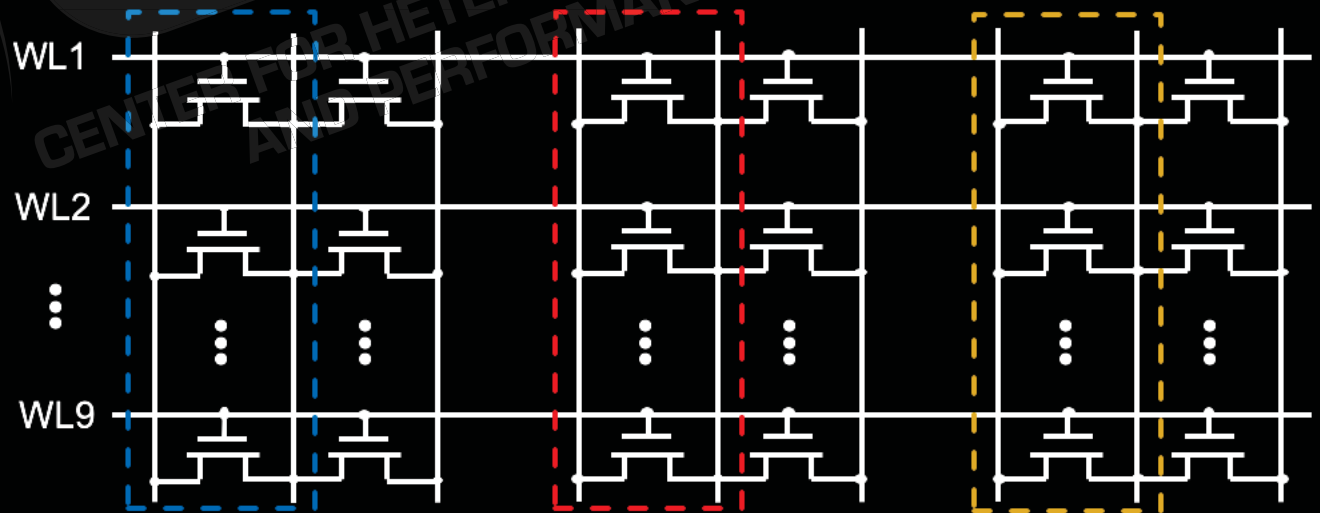


X. Gu and S. S. Iyer, IEEE EDL, 2017

Experimental Platform for Demonstration



- ❑ Standalone CTT array (10WL X 8BL)
- ❑ Probe card to access half of the array at once
- ❑ Largest hardware demonstration of WTA to our knowledge



Four years and going strong.....

- We have developed of three core technologies (Si IF FlexTratetm and CTT) to acceptable usability levels
- They promise to significantly impact the packaging, medical electronics and in-memory analog computing areas
- Our focus going forward is to
 - move these technologies into manufacturing entities with our consortium partners and helping them leverage these technologies
 - Focus on further technology innovations such as PowerTherm, RF and Photonic interconnects, adaptive patterning, high resolution displays
 - Leverage our technologies in novel architectural applications
- Turn out the best and most prepared students to continue in Industry and academia

Acknowledgements

<https://chips.ucla.edu/students>

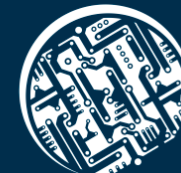
<https://www.chips.ucla.edu/faculty>

<https://www.chips.ucla.edu/alumni>



UCLA

Samueli
School of Engineering



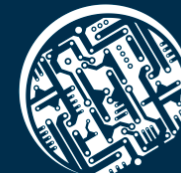
CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

Our Government and Industrial Sponsors



UCLA

Samueli
School of Engineering



CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING