NHANCED SEMICONDUCTORS

3D Integration and Highly Interconnected Systems

April 23, 2019

Robert Patti
rpatti@NHanced-semi.com
630-561-6813

Salient Points

- Interconnect and communication can (will) drive the future

- Reduced Latency
- Reduced Power
- Reduced Wiring

- Gives Simplicity

Our Belief: The Future of Compute is an Orchestra

- Many
  - Architectures
  - Technologies
  - Data Sources

- Agility
- Adaptability

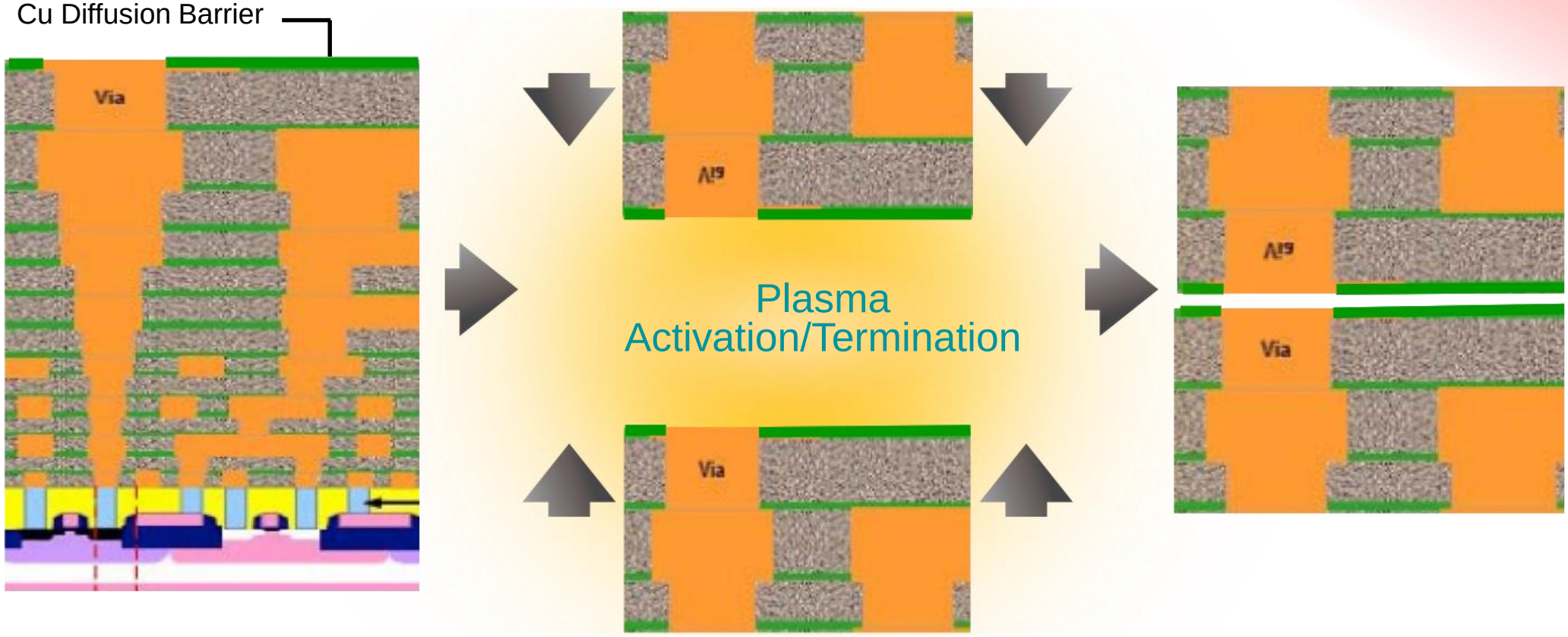
- Clever Design Can Overcome Physics

2.5/3D INTEGRATION

Hybrid Bonding Alignment and Placement

Leverages Industry Standard Processes and Equipment

Start with Industry Standard CMOS



Align/Place Wafers (or Die)

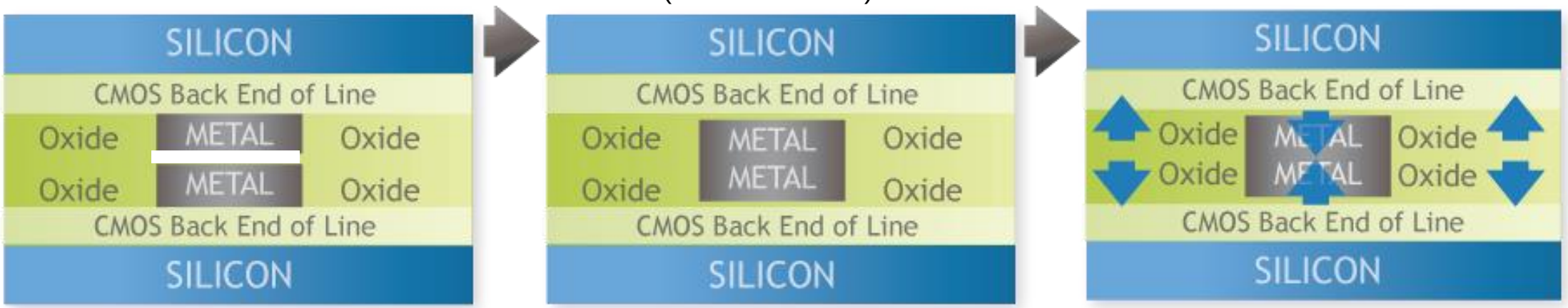
Hybrid Bonding Internal Thermo-Compression

Electrical Interconnections without External Pressure
Minimizes Stress and Cost of Ownership

Cross-Section after Pick/Place (example)

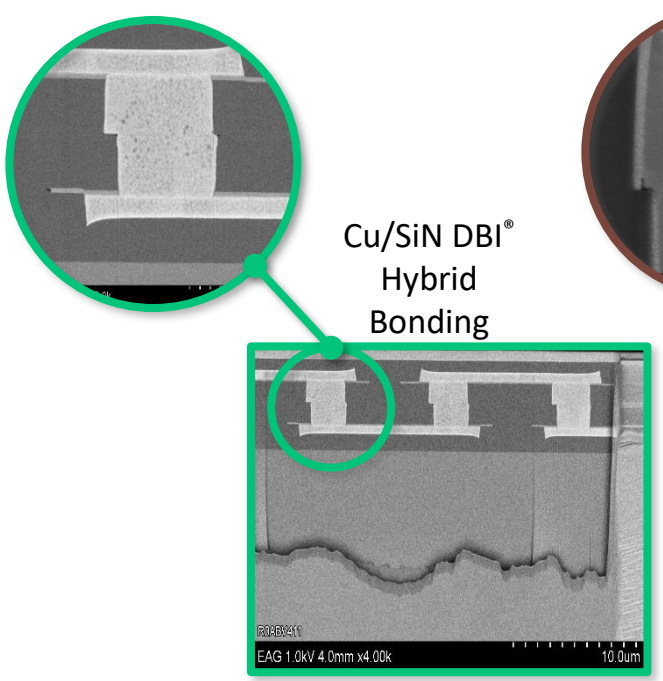
Heating Closes Dishing Gap (~ 1 nm/um/50ºC)

Further Heating Compresses Metal w/out External Pressure

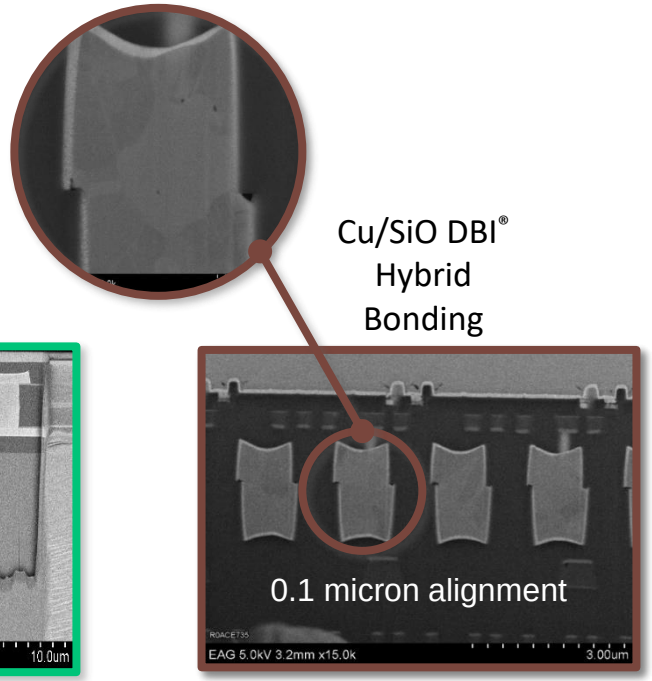


Spontaneous Chemical Reaction with Byproducts Diffusing Away from Bond Interface

Hybrid Bonding Interconnect Pitch Scaling

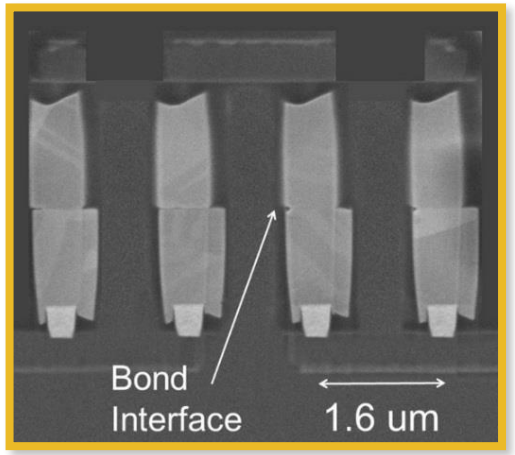


10 µm DBI® pitch, 300℃



1.9 µm DBI® pitch, 300℃

Scalable To < 1um Pitch



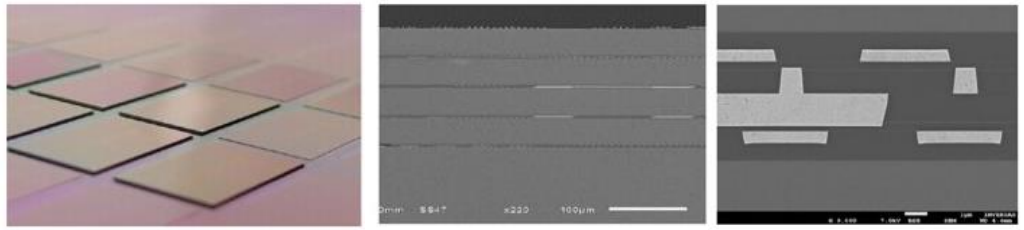
1.6 µm DBI® pitch, 300℃

Wafer-to-Wafer vs. Die-to-Wafer

• Wafer-to-Wafer

- Process implementable in foundry back end of line (BEOL) with a low cost-of-ownership
  - Particle control requirement easily met
  - Proven in many applications
    - CMOS BSI Image Sensors
    - RF switches
- Requires wafer and die sizes to be matched

50um die stacked 4-high, optical and SEM cross-sections



4-high 50um die stacks | 4-high cross section | Die bond interface

• Die-to Wafer

- Accommodates die tiling, stacking and mismatched die/wafer sizes
- Additional process steps of die singulation and handling required
  - Additional particulate/handling challenges

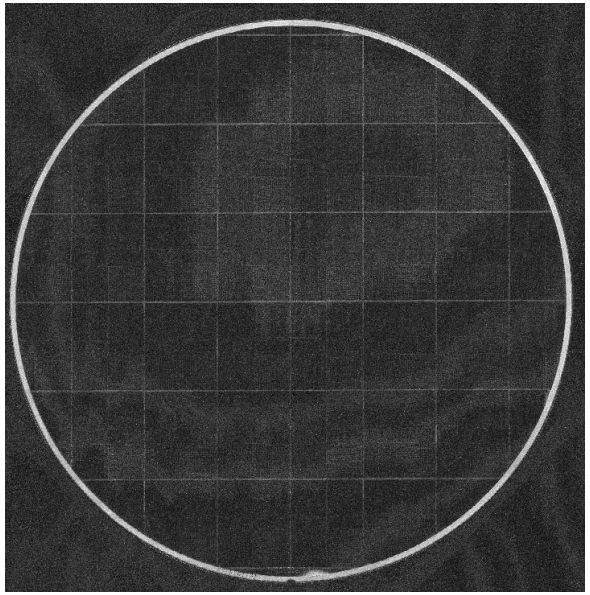
Die Stack with DBI® Hybrid Bonding

- Improved performance, cost, and yield/reliability potential
  - Throughput – no reflow/alloy, throughput improved x2
  - Thermals – no underfill, ΔT improved x5/10 for 4/8 high stack
  - Electrical parasitics – DBI® replaces bumps, RC improved ~ x20
  - Reduced stress – eliminate reflow/alloy and underfill
  - Reduced pitch – pick/place tool limited, throughput dependent

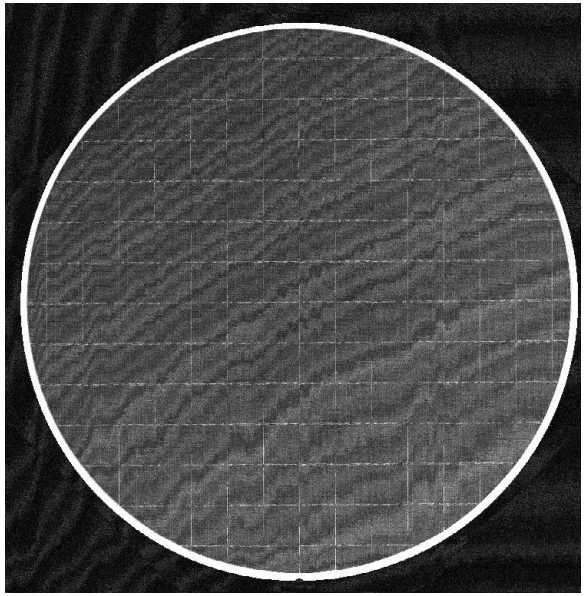
Hybrid Bonding:

Yield = 80% +/- 20% electrical yield depends on project

- 2 step anneal for SiO2 bonding first and followed by metal-to-metal bonding
- Cu and Ni are used for vertical interconnect bonding metal
- Front-to-front & back-to-front bonding depends on design
- Application for the high density fine pitch vertical interconnect



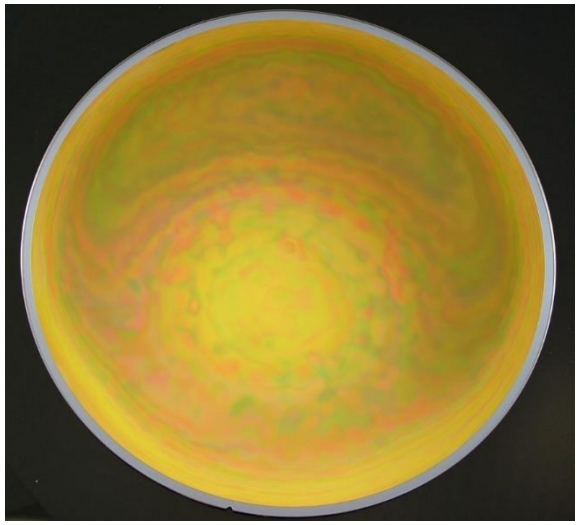
(a) C-SAM after anneal @350C of wafer bonded with large die size



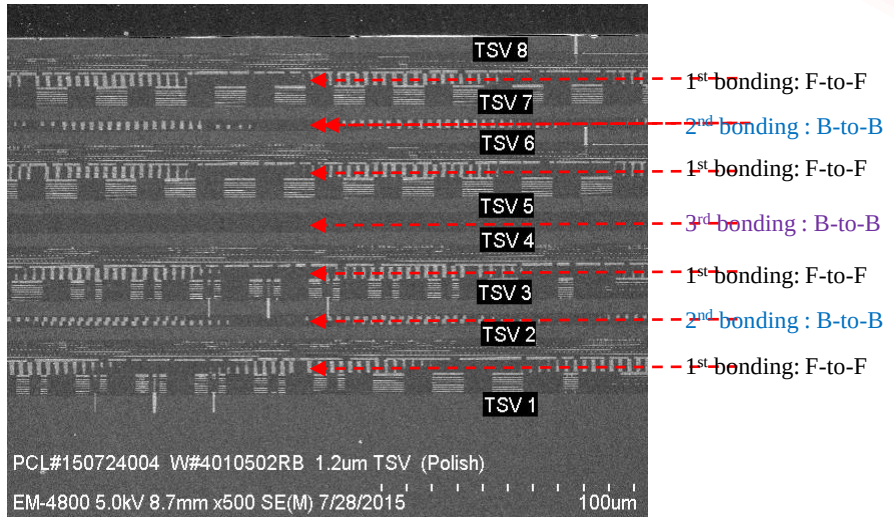
(b) C-SAM after anneal @350C of wafer bonded with small die size

SiO2 Bonding and Hybrid Bonding for Multi-Wafer Stacking

- 4 wafer stack : SiO2 bonding
- 8 wafer stack : Hybrid bonding
- 16 wafer stack : SiO2 bonding
- First article 20 wafer stacks : Hybrid bonding



(a) Picture of 4 wafer stack bonded using SiO2 bond Top Si has been removed

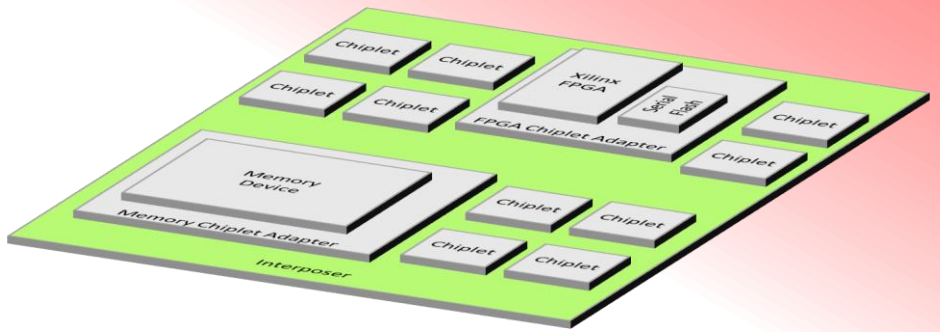


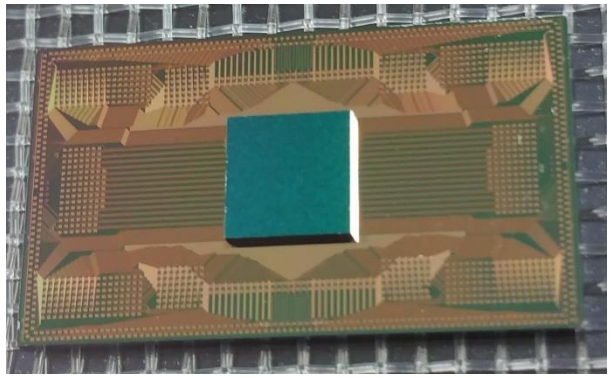
(b) SEM cross sectional micrograph for 8 device wafer stack

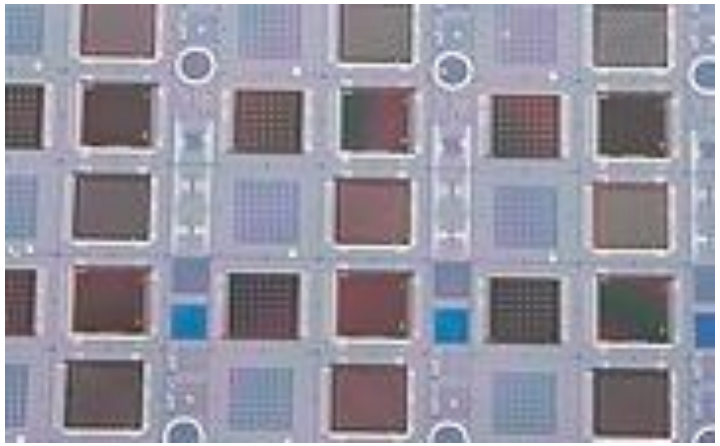
2.5D Systems

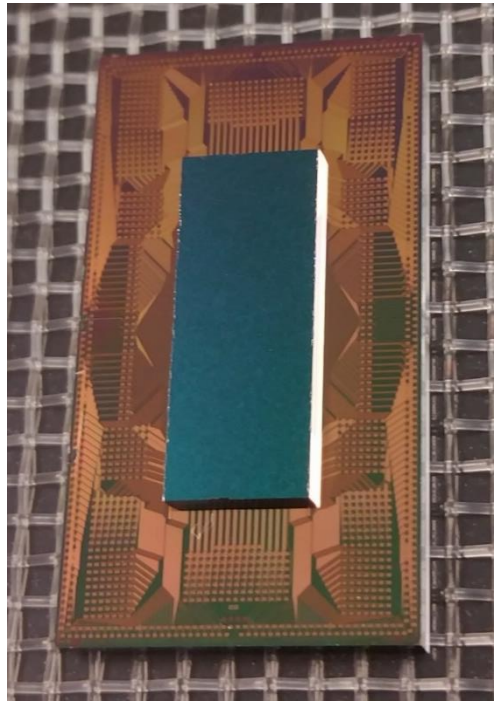
Mixed technology assemblies

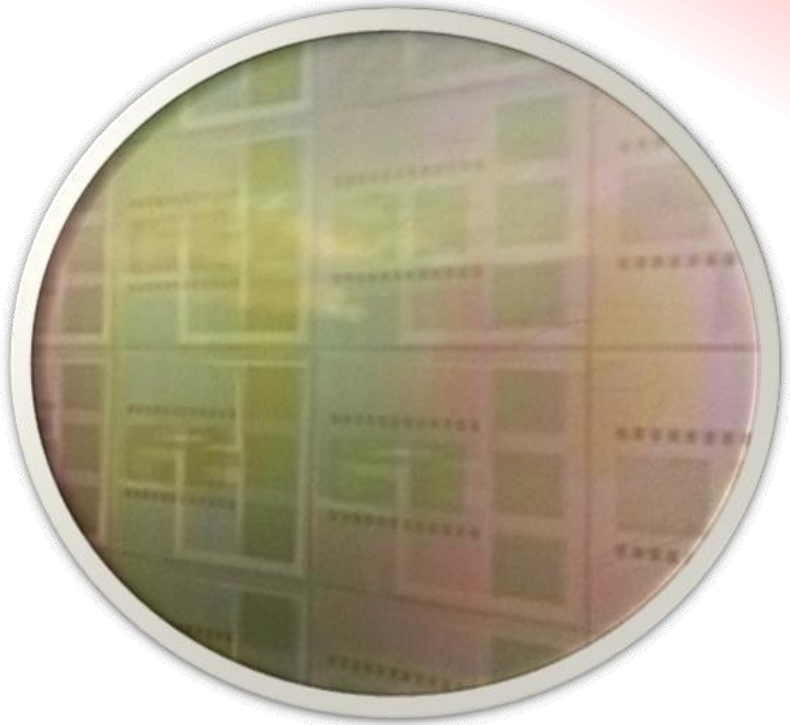
- Flip-chip
- Copper pillar
- DBI die to wafer
- Organics and silicon circuit boards





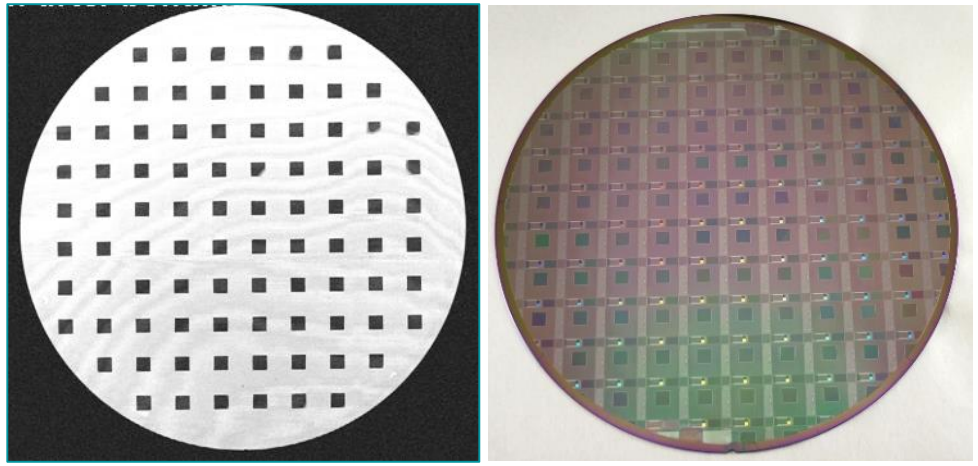




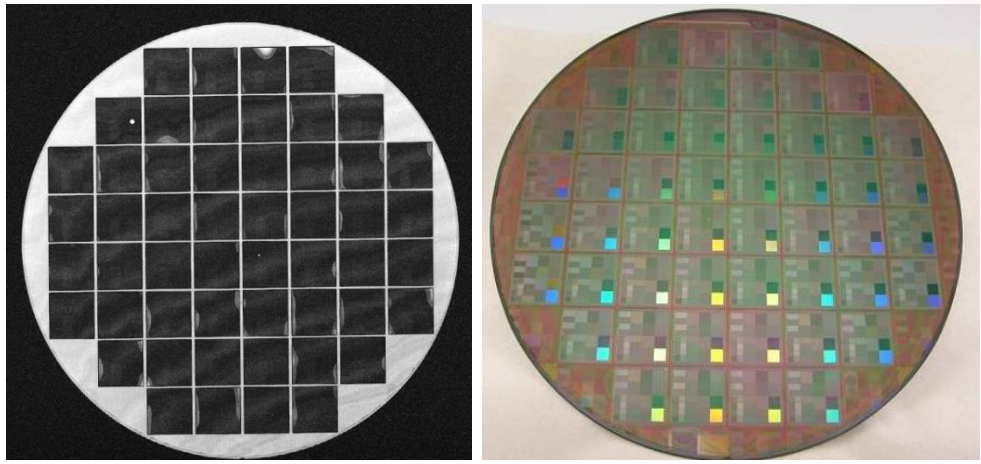


Hybrid Bonding of Die-to-Wafer:
Yield = 80% +/- 20% electrical yield depends on project

- 2 step anneal for SiO2 bonding first and followed by metal-to-metal bonding
- Cu and Ni are used for vertical interconnect bonding metal
- Pad can be opened on die back or host wafer front.



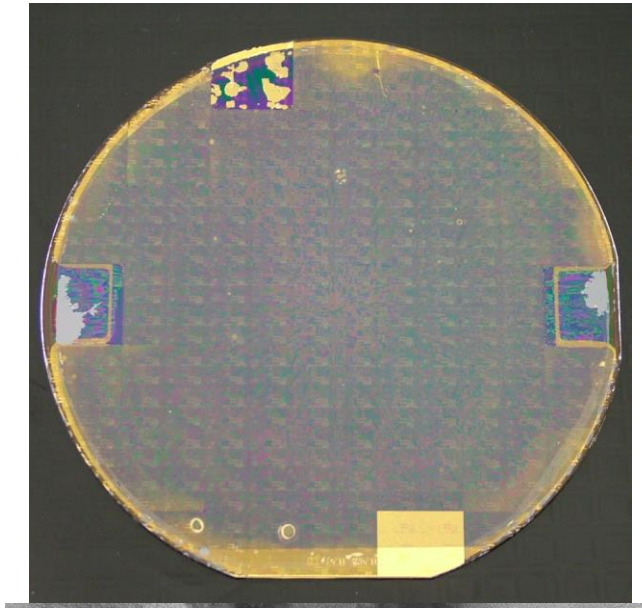
(a) Die-to-wafer bonding for smaller die: C-SAM and wafer picture

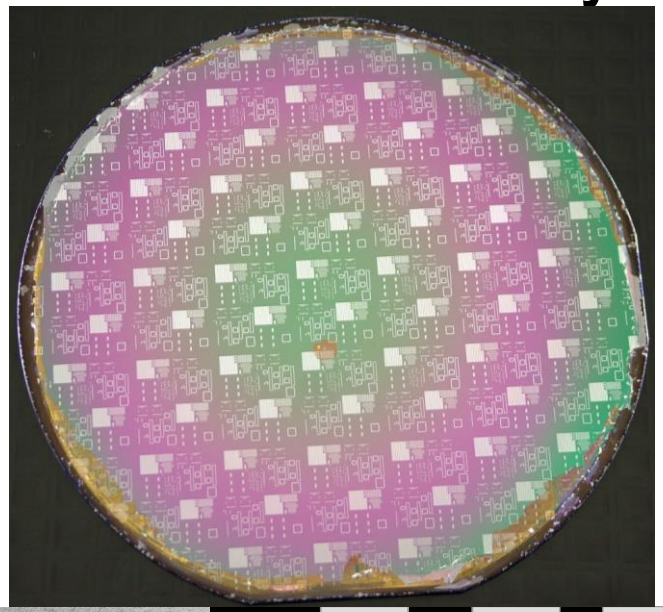


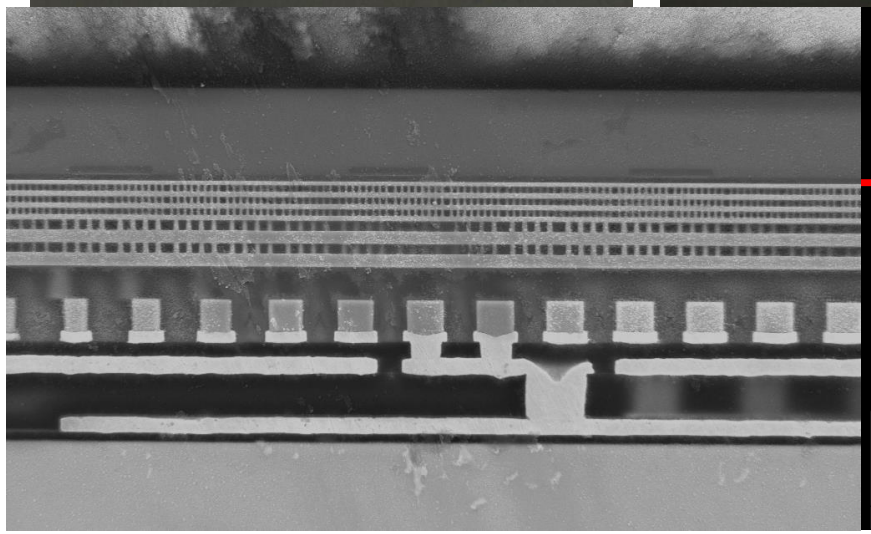
(b) Die-to-wafer bonding for bigger die: C-SAM and wafer picture

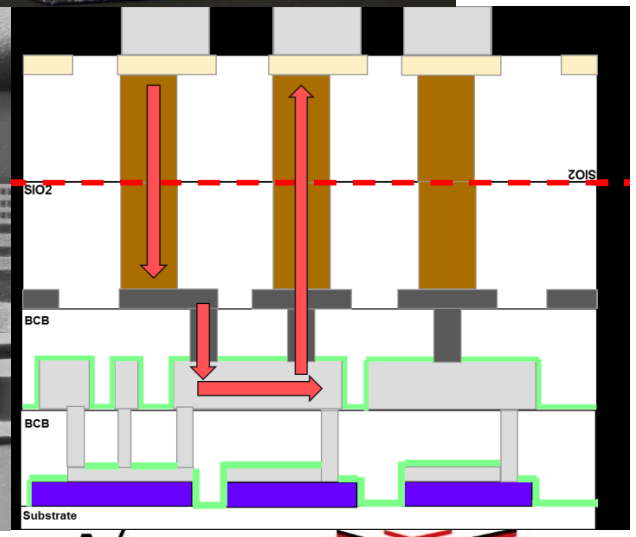
Hybrid Bonding of Heterogeneous Substrate:

Yield = 70% +/- 20% electrical yield depends on project

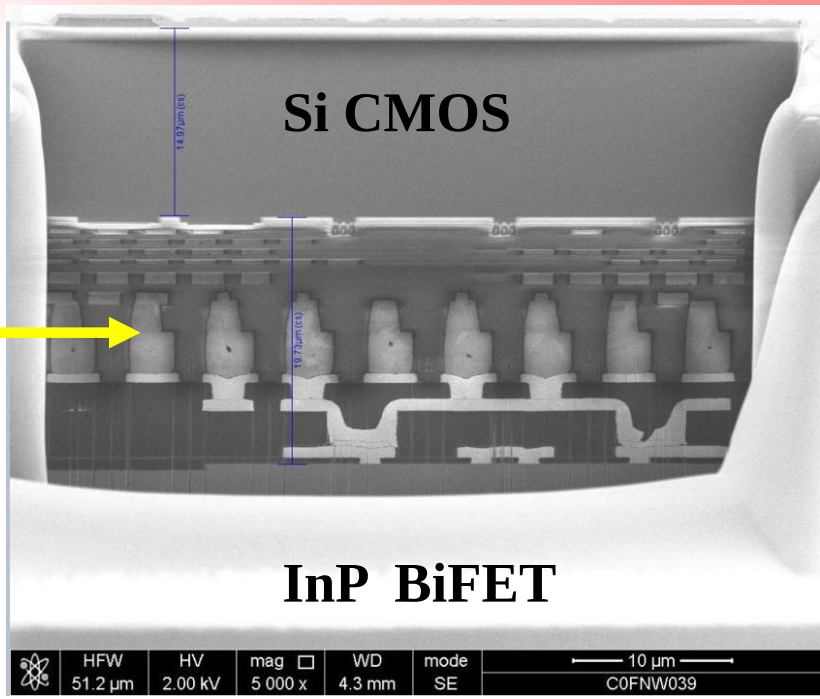








Bonding Interface



2.5/3D Integration Summary

- 2.5/3D integration overcomes the limitations of scaling
  - Reduces wire
    - Improves speed
    - Improves power
- Enables heterogenous materials, technologies…
  → enables heterogenous computing
- Runs in HV

- **More performance per watt, sqft, or $.**

NHANCED SEMICONDUCTORS

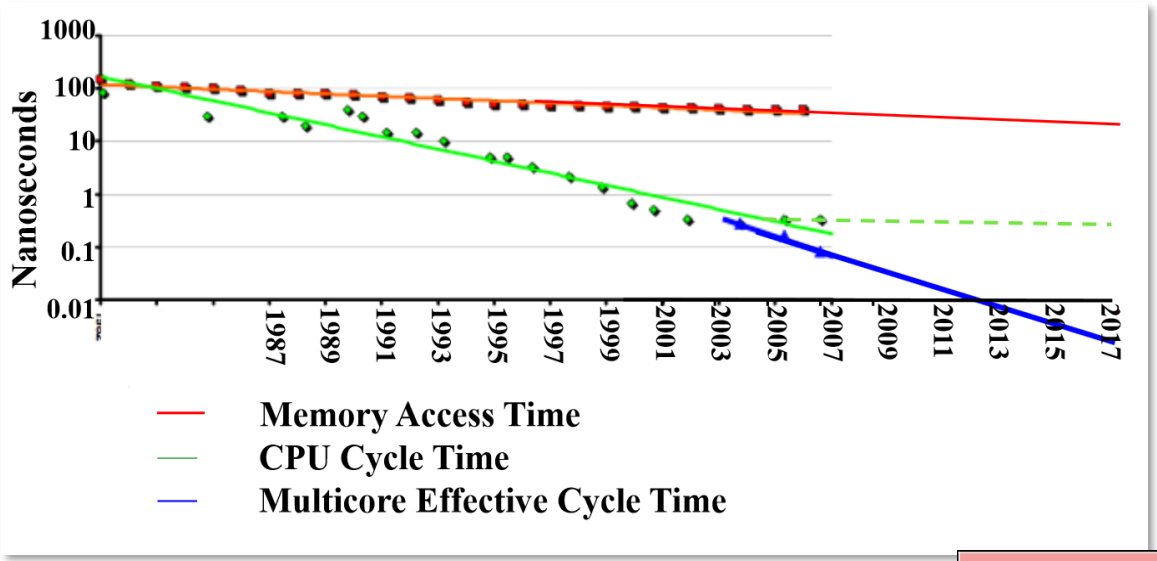
Qubulate

LambdaFabric®: A Scale-Invariant Computing Interconnect Scheme

NHANCED SEMICONDUCTORS

Current Generation Chips Have A Problem…

Today's processors: Time & Energy Dominated by Fetch



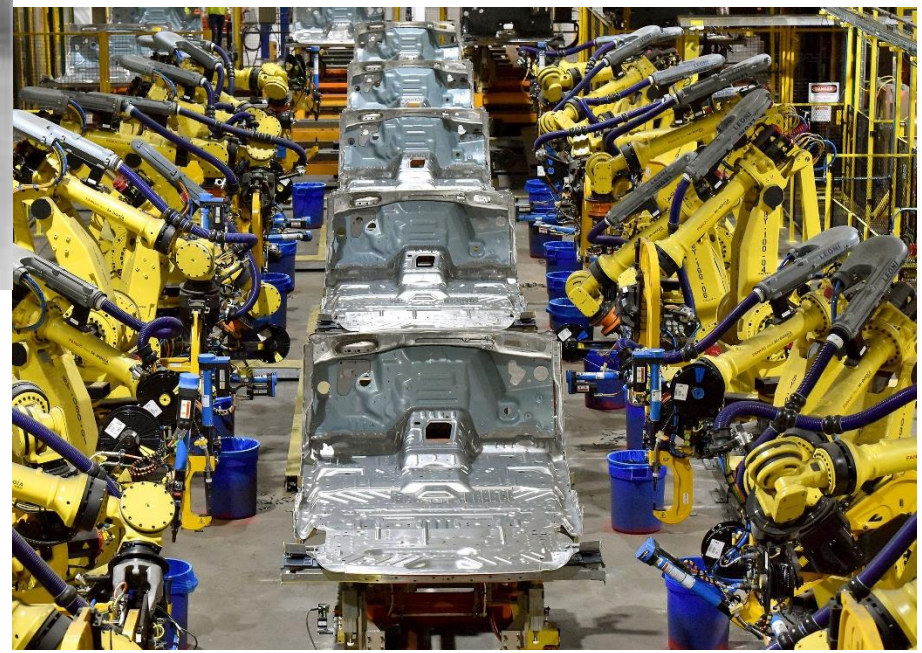
- **Memory gets larger but not faster**
- **Logic gets faster but spends more time waiting for memory**
- **Logic gets more energy efficient but memory transport does not**

Operation	*Energy consumed*
64-bit multiply-add	64 pJ
Read/store register data	6 pJ
Read 64 bits from DRAM	**4200 pJ**
Read 32 bits from DRAM	**2100 pJ**

The Babel Problem

- A modern “system on a chip” can have as many as several hundred different data busses and formats. To move information any distance many changes in format take place thereby wasting energy and time.
- When moving bits across die, between chips, between boards, and between racks, today’s systems change the format of the data communications scheme many times.

The Data Assembly Line



Solution: Wetware To Silicon

Nature's Design Principles

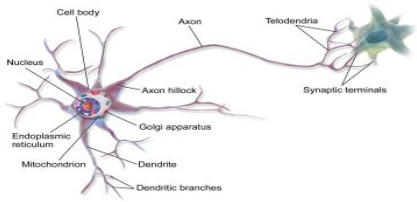
- Neuron body contains "context"
- Communication via synapses
- Axonal connections define geometry

LambdaFabric® Design Principles

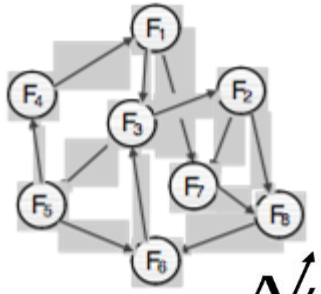
- Scalable heterogeneous parallelism
- Proximity of memory and processing
- Support for complex connections
- Communications driven architecture
- Developed for hardware based implementation
  - Ultra low latency
- All information is "pushed"
- Single modality
- Packets are instructions
- Must operate with noise and failures

ASIC

Neurobiology

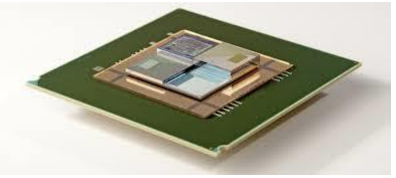


Directed Graph



Random Access Operations

$$\begin{pmatrix} 0 & 0 & W_{0,2} & W_{0,3} \\ W_{1,0} & 0 & 0 & 0 \\ 0 & W_{2,1} & 0 & W_{2,3} \\ 0 & W_{3,1} & 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix} = \begin{pmatrix} A_0' \\ A_1' \\ A_2' \\ A_3' \end{pmatrix}$$

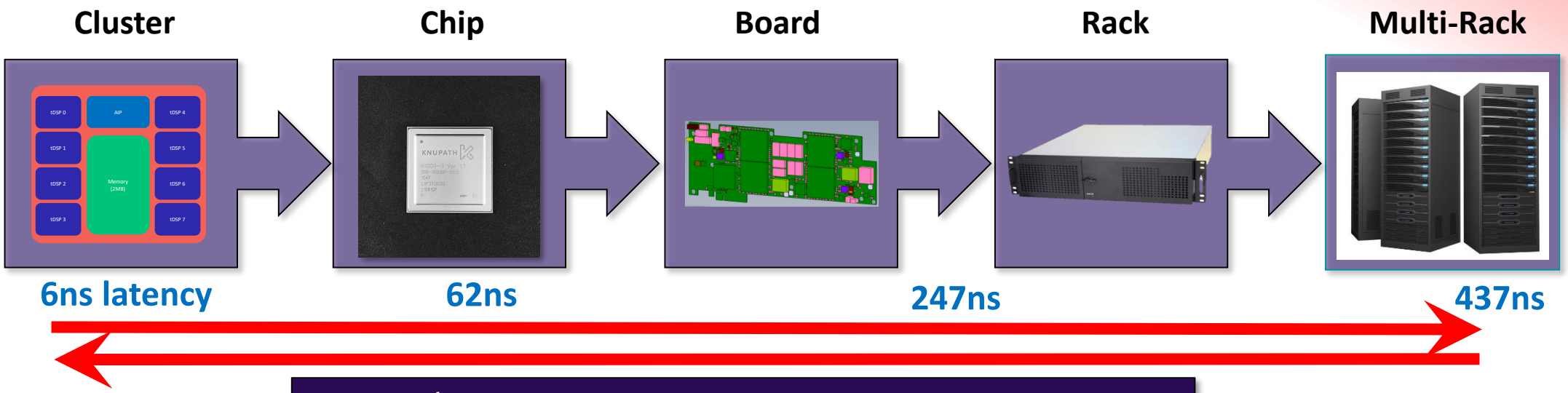


NHANCED SEMICONDUCTORS

Take away: Be flexible.

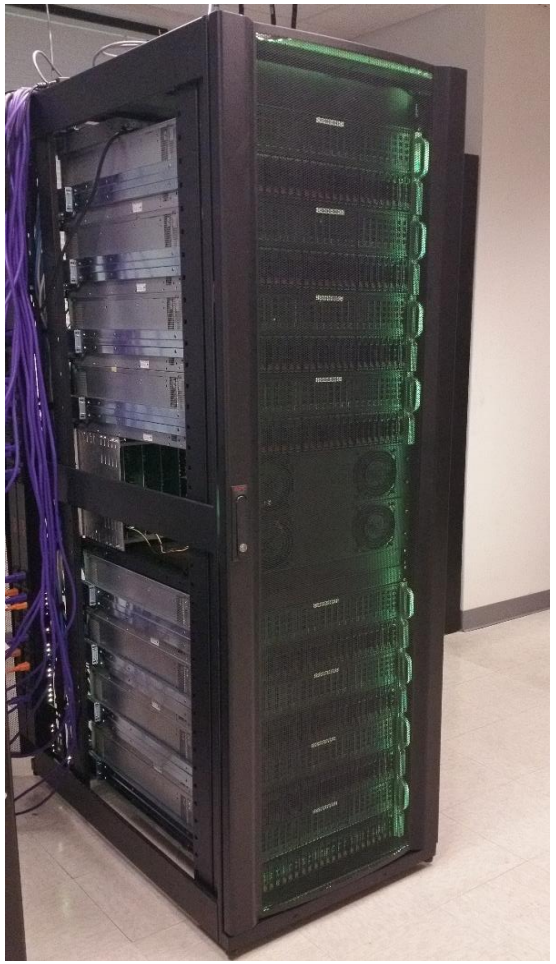
LambdaFabric® Scalable Computing

LOW-LATENCY, HIGH-THROUGHPUT, LOW-POWER COMPUTING FABRIC



- ✓ Scale invariant network architecture
- ✓ Low latency to everywhere
- ✓ High bandwidth
- ✓ Multi-dimensional connectivity
- ✓ Scalable up to 524,288 chips

Synthetic Quantum Computing
- Quantum Inspired Computing

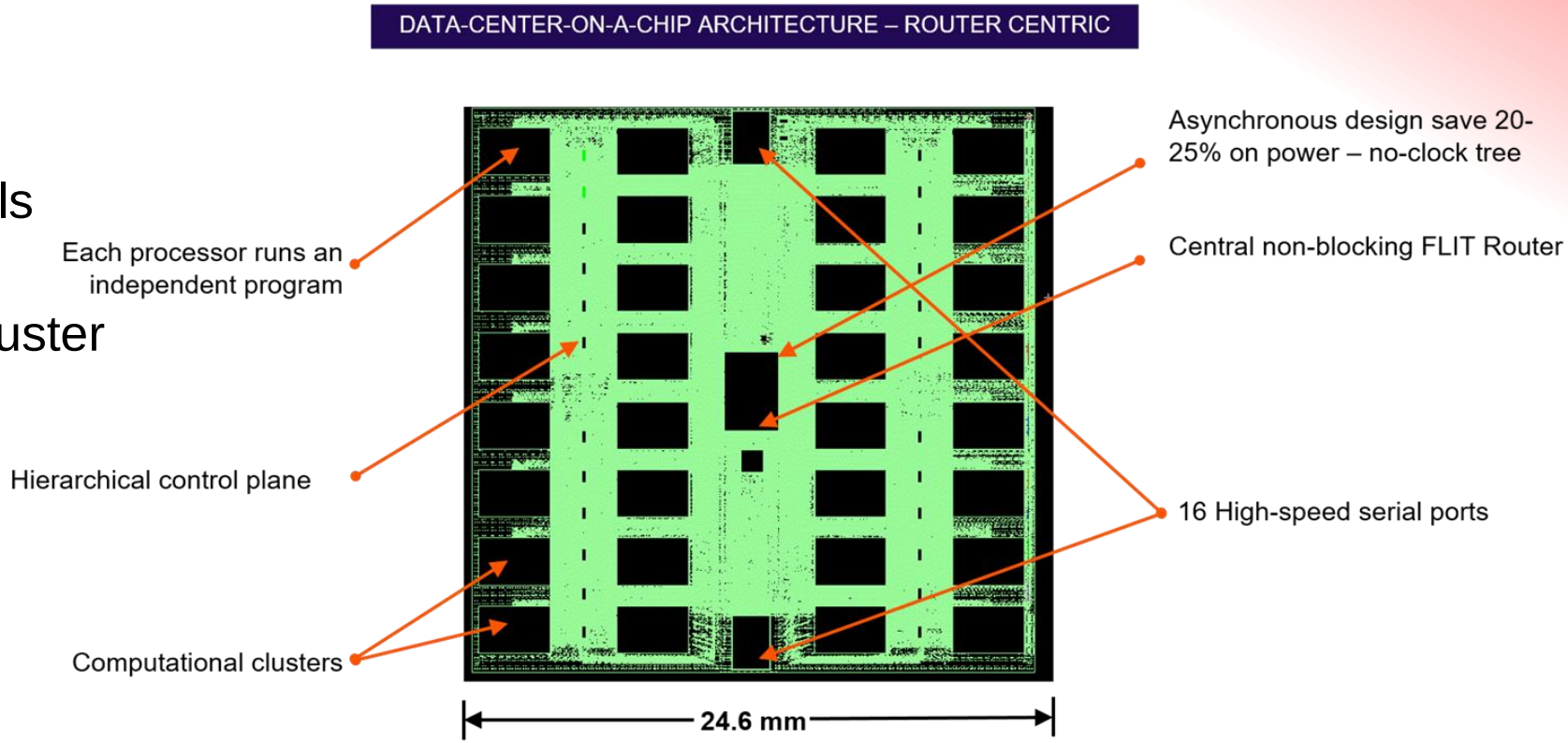


The Qubulate Miwok™ platform is the culmination of a more than $125M development effort targeting next generation computing.

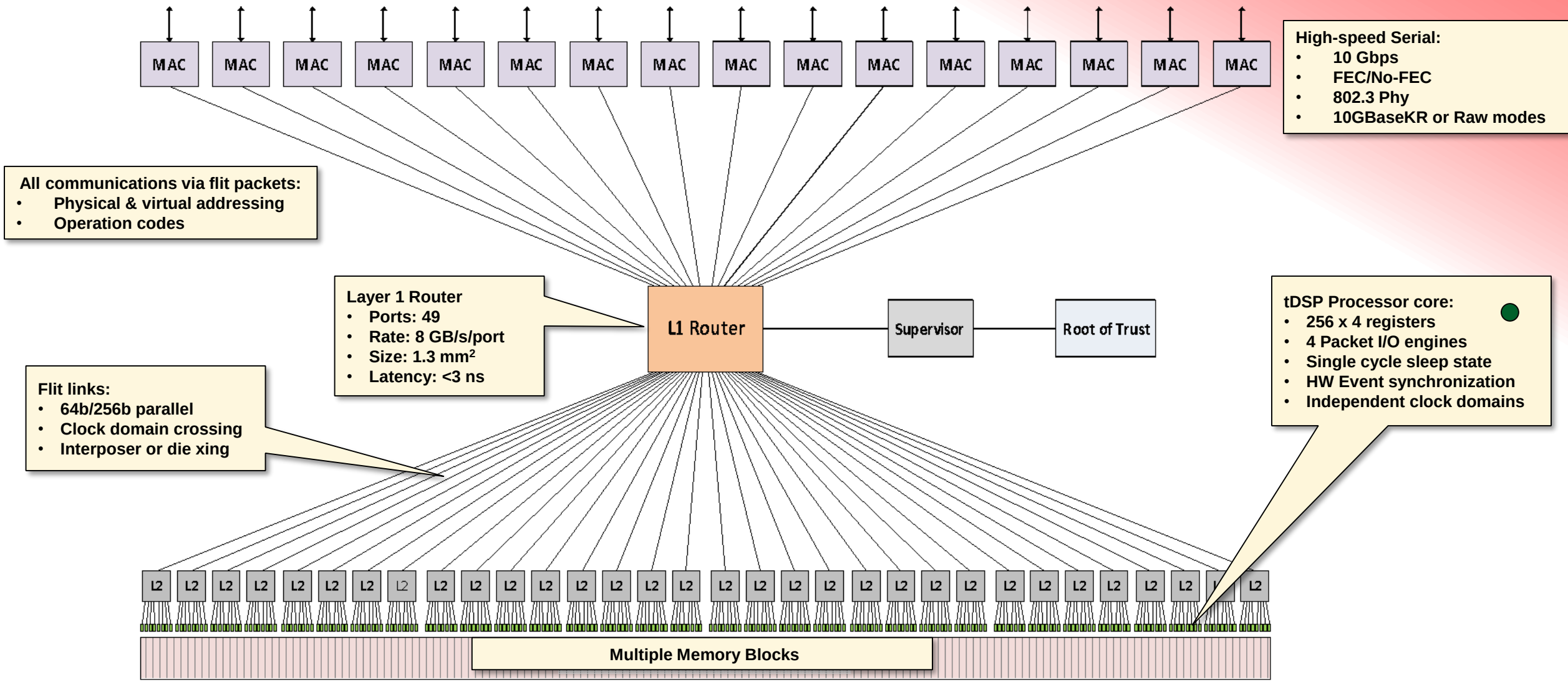
- >80k Heterogenous Cores
  - One Processor
- 5.6Tb/s Cross Sectional Backplane Bandwidth
- >50TB/s Fabric Bandwidth
- <250ns End-to-End System Communication Latency
  - Register to Register

Knureon CPU (KPU, K1000)
Many Core Processor Building Block

- 256 cores each with 288kBytes of near memory (L1)
- 256 registers for in-register computing
- Each core has 256 x 4 bytes multi-ported register file
- 1GHz clock rate
- 16 10Gb/s SERDES channels
- One 32b FPU/core
- One memory coprocessor/cluster
- One 32b transcendental math engine/cluster (8cores)

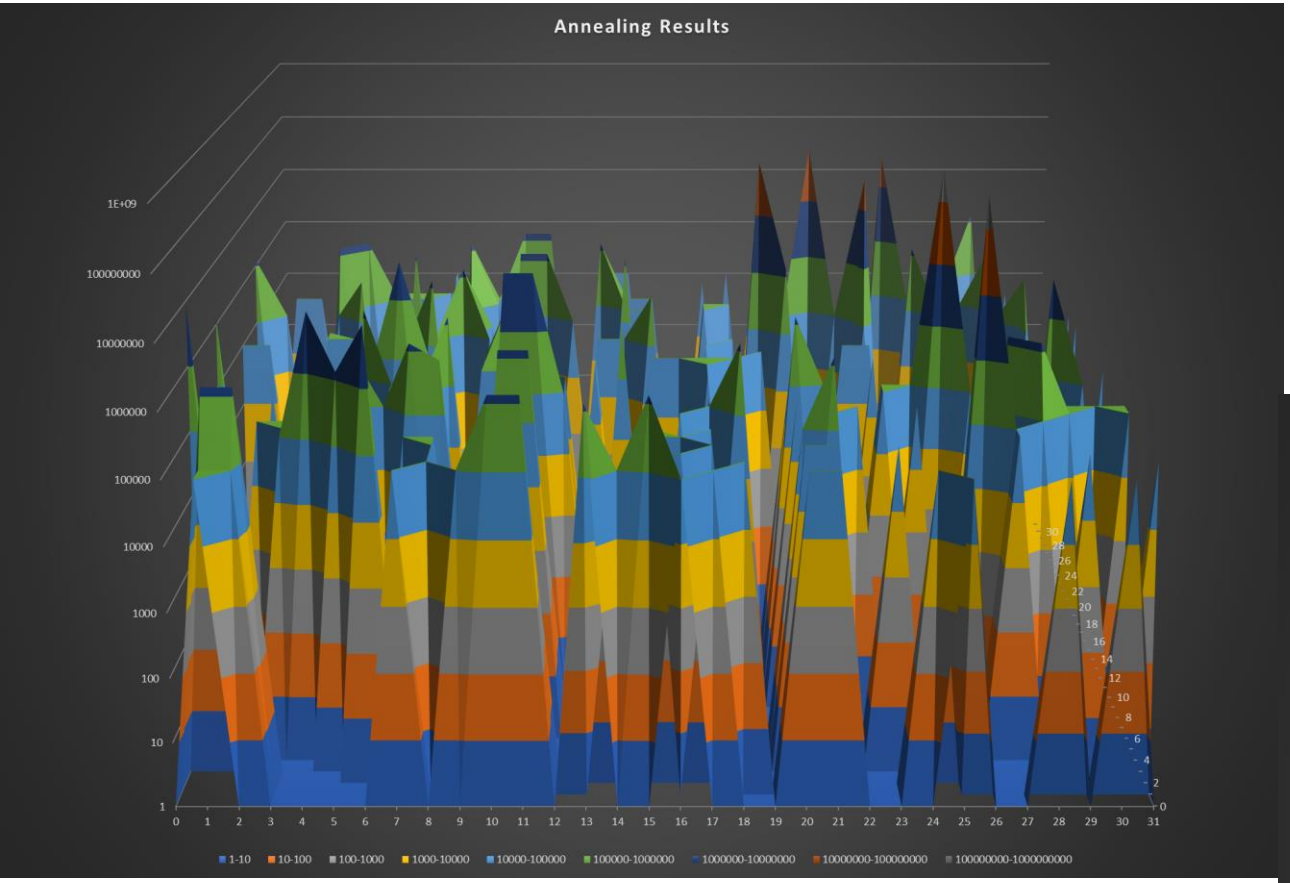


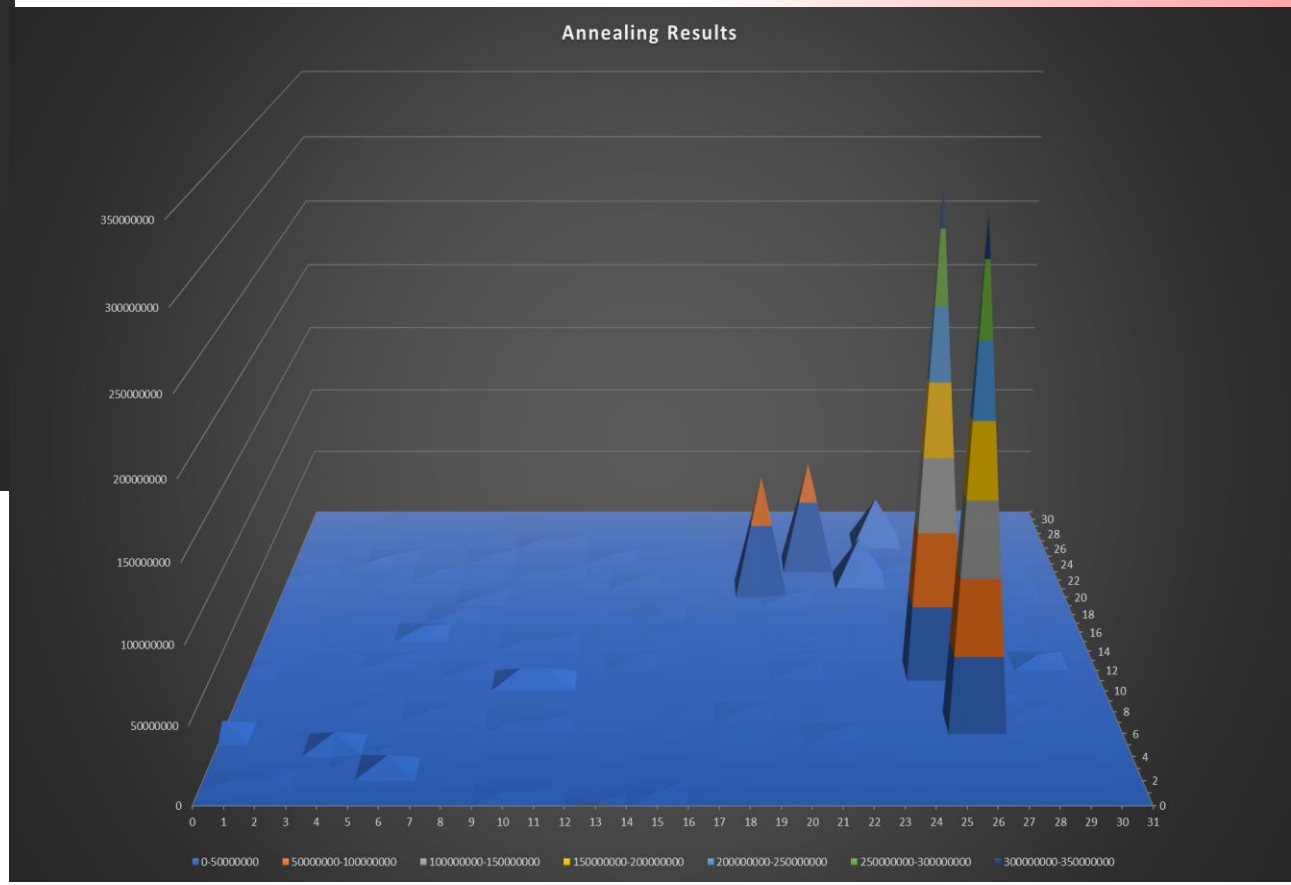
Datacenter-On-a-Chip



Classical System *Supremacy* by Architecture

- Current system capabilities for “quantum annealing”:
  - Up to 65,536 nodal variables (64k “qubits”)
  - 32 bit FP all to all nodal weights (4B values)
  - With 64k variables we perform ~350 data set solutions per second
    - Up to 1.5 trillion packetized variables exchanged between nodes every second
  - With 1024 state variables we perform ~250M data set evaluations providing ~4M solutions per second – 99.99% confidence <5 secs
  - 32b integer or floating point arithmetic
- Exceeds current Icing and Quantum machine performance
  - Will continue to exceed Quantum machine capabilities at least for several more years


Annealing Results
1E+09
100000000
10000000
1000000
100000
10000
1000
100
10
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
1-10
10-100
100-1000
1000-10000
10000-100000
100000-1000000
1000000-10000000
10000000-100000000
100000000-1000000000


Annealing Results
350000000
300000000
250000000
200000000
150000000
100000000
50000000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
0-50000000
50000000-100000000
100000000-150000000
150000000-200000000
200000000-250000000
250000000-300000000
300000000-350000000

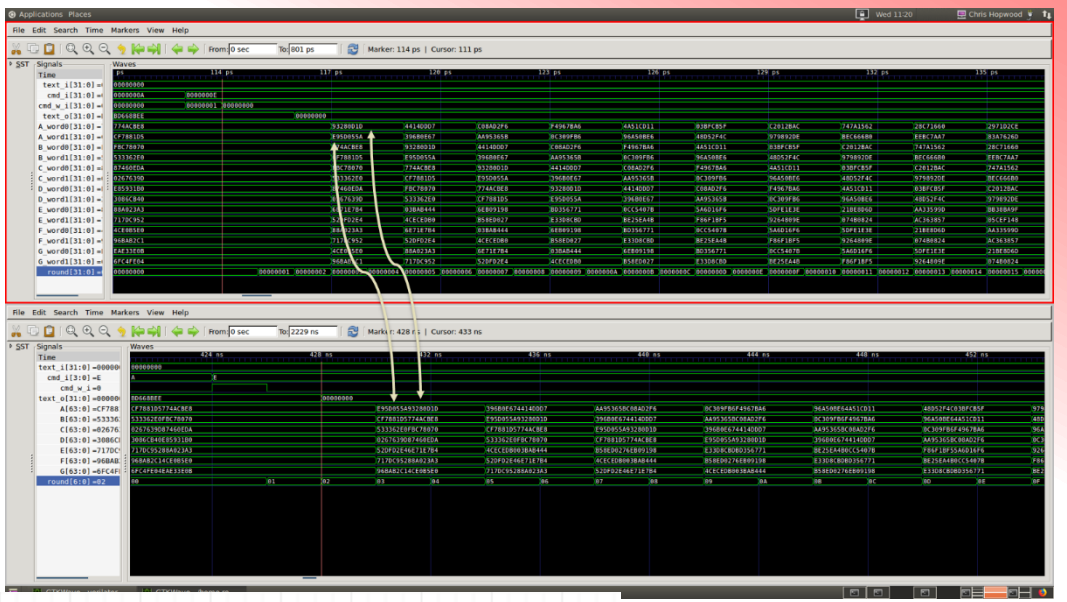
Scalability

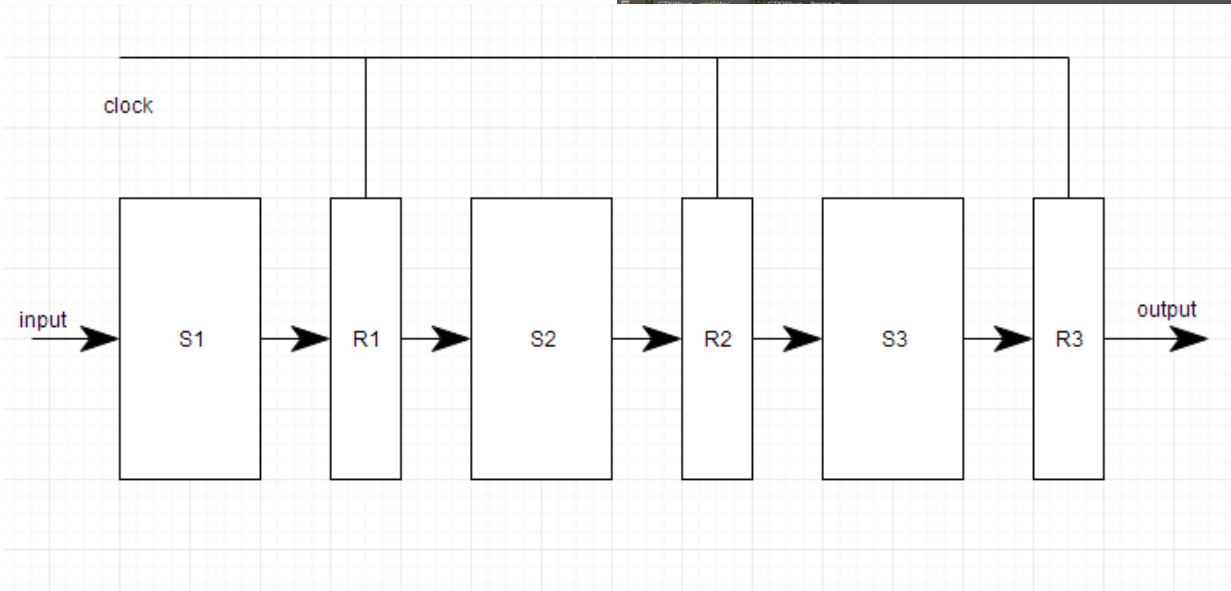
Site Set Size	Number of Coupling Weights	Solutions Per Second	Simultaneous Runsets Available On The Miwok System
1024	1M	1290	4,096
2048	4M	1260	1024
4096	16M	1190	256
8192	64M	1310	64
16,384	256M	947	16
32,768	1G	578	4
65,536	4G	326	1

Other Performance Validations

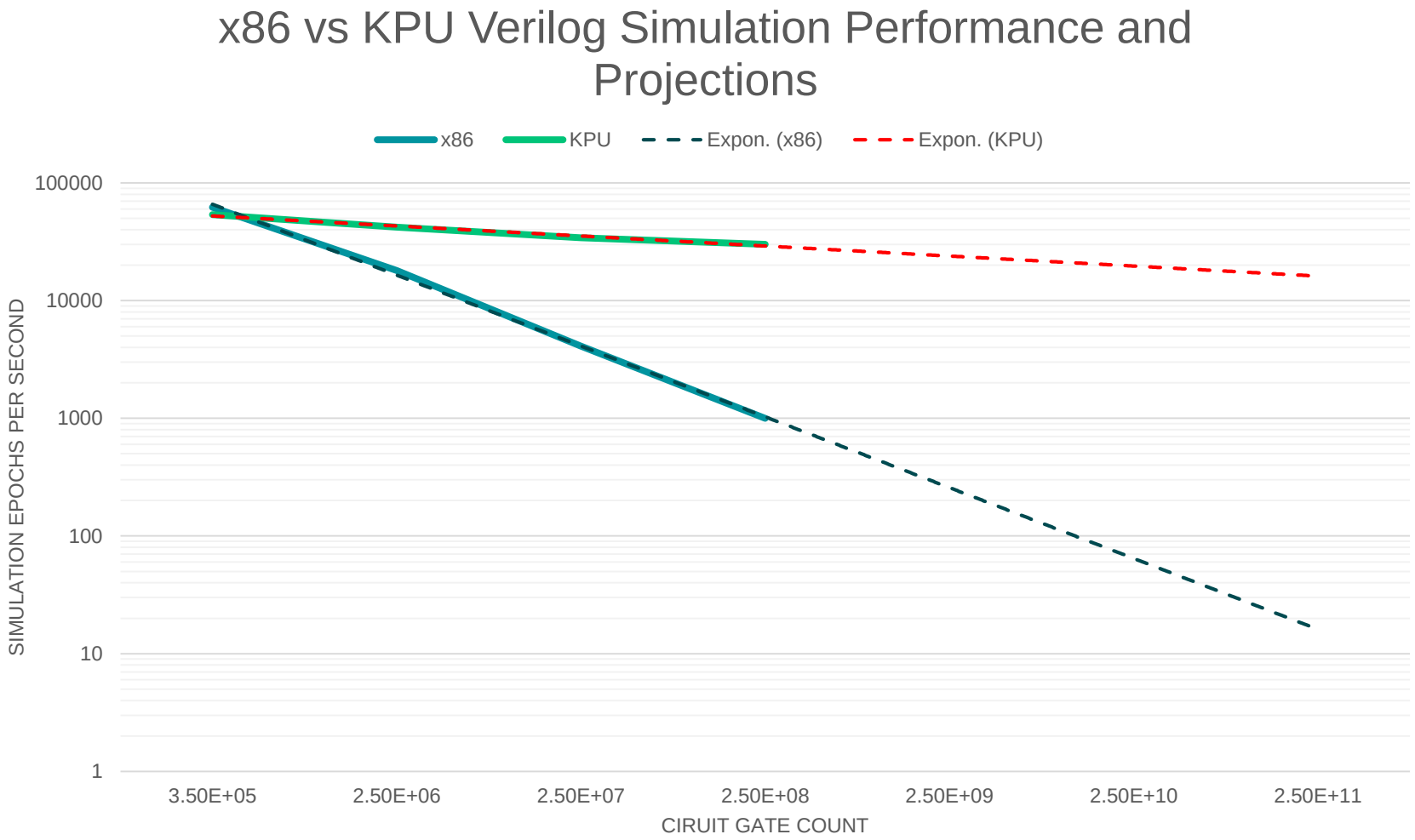
- We implemented a Verilog (timing mode) simulator on both KPUs and x86 processors
- We verified the results using a handful of circuits including AES512
- We then did benchmarking on arrays of 32bit wide register pipelines n wide X 10 deep.
- Tests were run on array widths from 10 to 500,000 registers –

– ~160M F-F, ~1.25B gate equivalents





KPU vs x86 for MPMD Verilog Simulator



LambdaFabric is expected to accelerate:

CFD, finite element analysis, SPICE, airfoil modeling, drug discovery, combinatorial chemistry, combustion engine design, etc.

Additional areas of technology that use the same basic software design include in-loop hardware simulation which is used for dynamic power grid control and beamforming for radar and 5G communications.

Epochs are nominally time steps or evaluation cycles in simulations.

The performance gains are due to the very low latency of the LambdaFabric and custom synchronization hardware built into the Qubulate Knureons.

LambdaFabric Summary

- Hyper connected in-memory computing provides effectively unlimited number of cores in a “processor”
- System extensibility to >100M cores
- Near linear scaling

- Can be combined with in-fabric FPGA compute and GPU acceleration

- **<u>More performance per watt, sqft, or $.</u>**