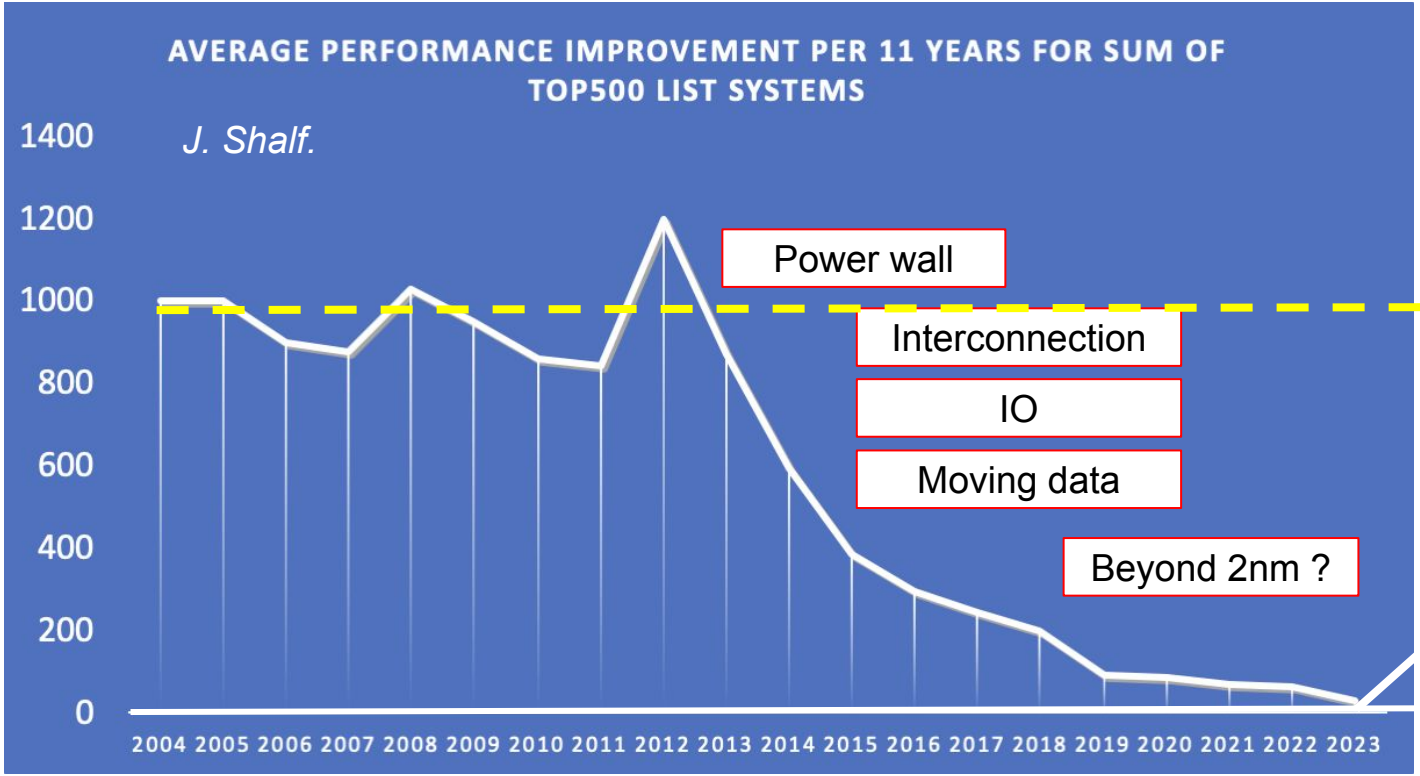
BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

Chiplets: A key ingredient for specialization in HPC

Patricia Gonzalez-Guerrero
Research Scientist
lg4er@lbl.gov

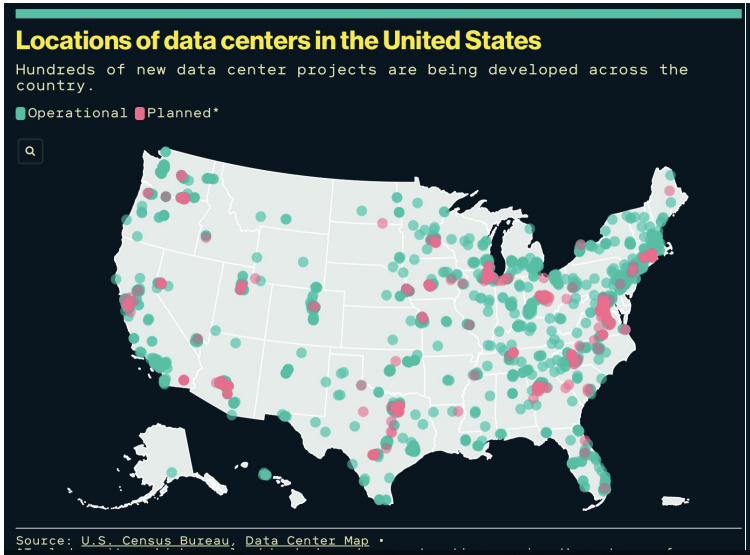
HPC Performance does not increase any more

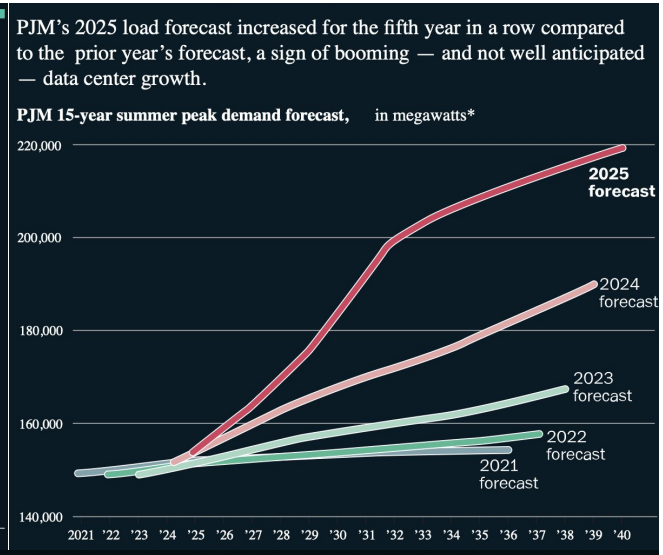


How do we get HPC performance back on track?

The problem? An energy crisis

GW AI factory!



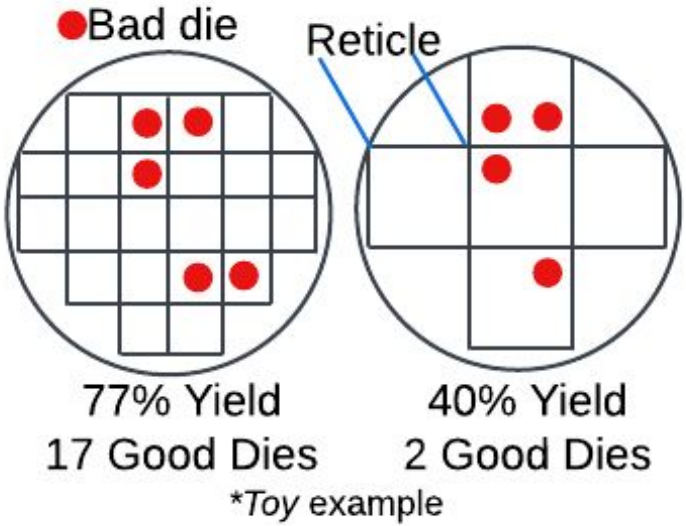


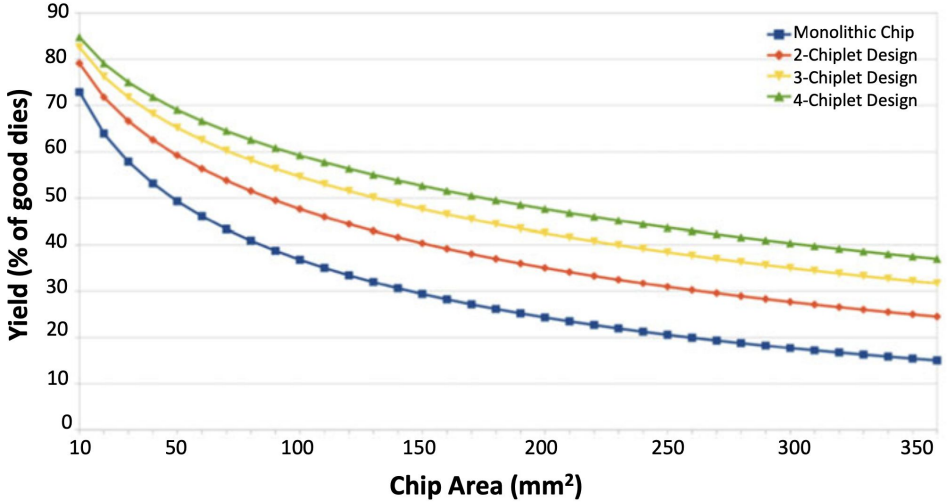
Series: The New Energy Crisis

https://thenewenergycrisis.com/

An investigative series uncovering how Virginia's data center boom—fueled by AI—threatens to overwhelm the power grid, drive up household energy costs, and put entire communities at risk.

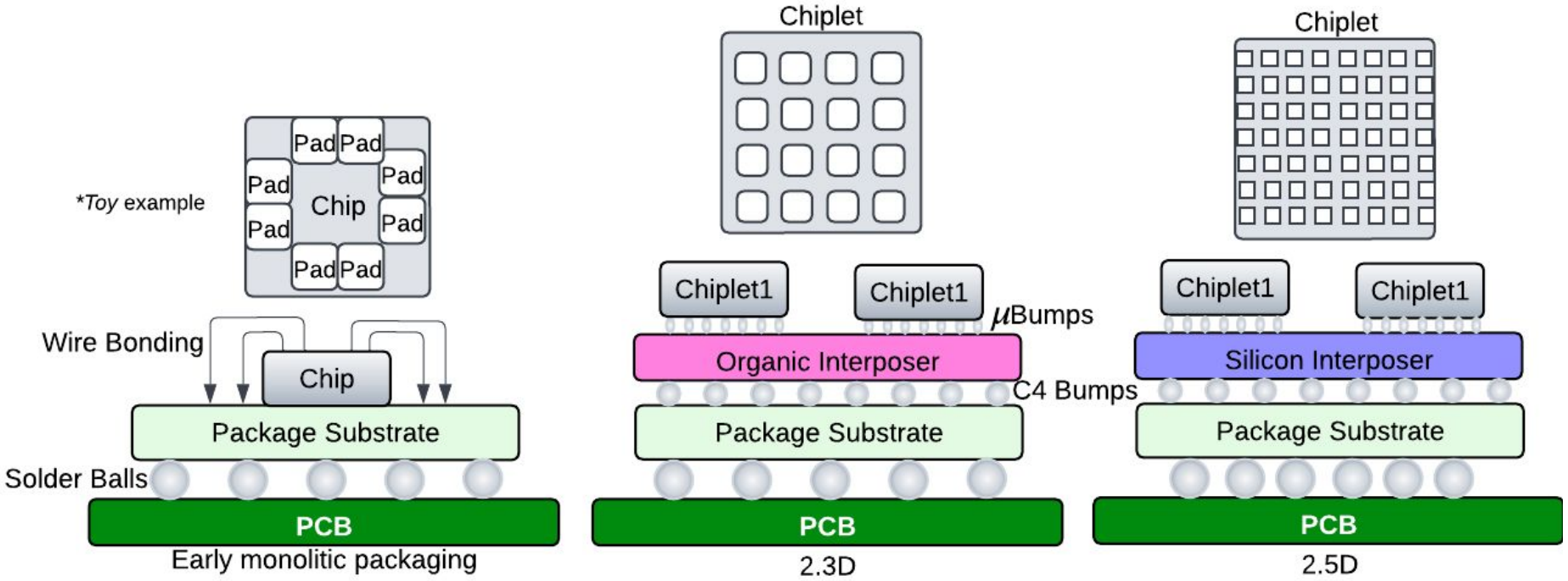
Solution 1: Make larger dies





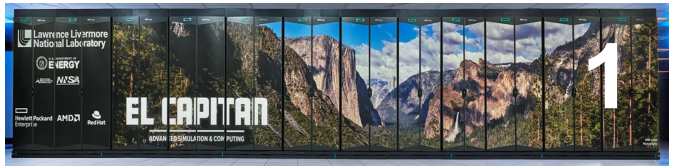
Lau. et al. Chiplet design and heterogeneous integration packaging

Solution 2: Advanced Packaging

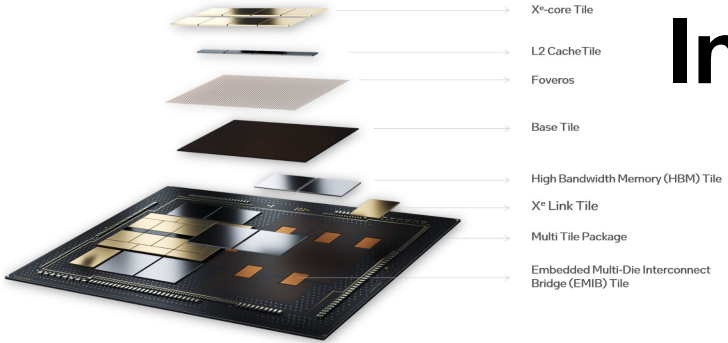


Chiplets made Exascale possible*

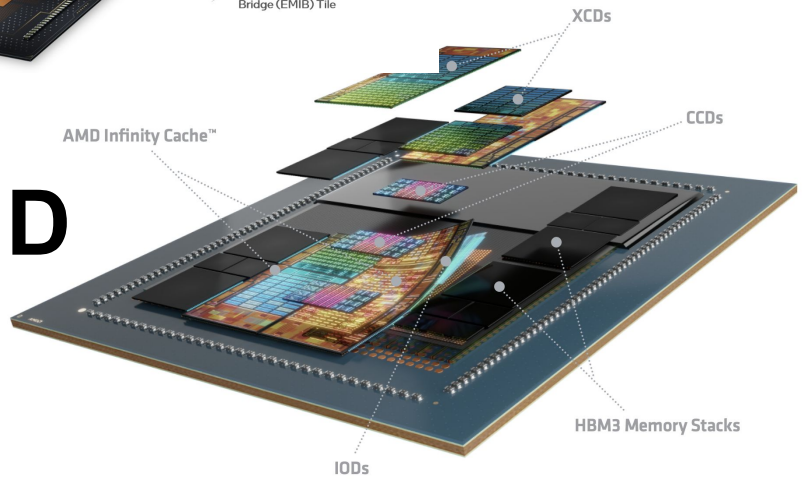
***And other key technologies: Novel interconnects, advanced cooling, power delivery… :-)*



Intel

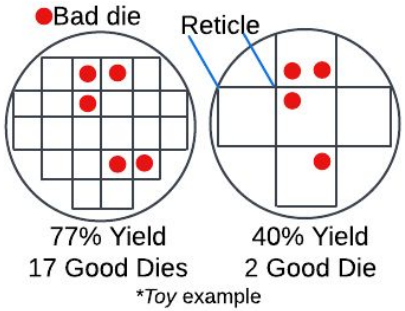


AMD



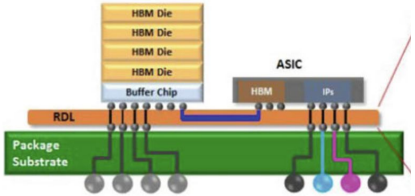
Chiplets have many well known advantages

Yield and cost efficiency



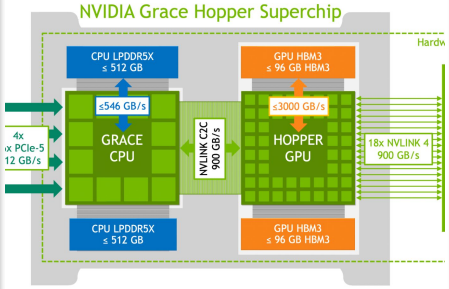
Overcome the memory wall

HBM is 3D integration



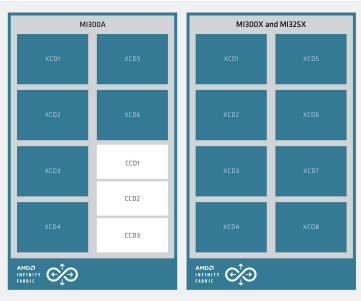
Lau. et al. Chiplet design and heterogeneous integration packaging

Heterogeneous integration



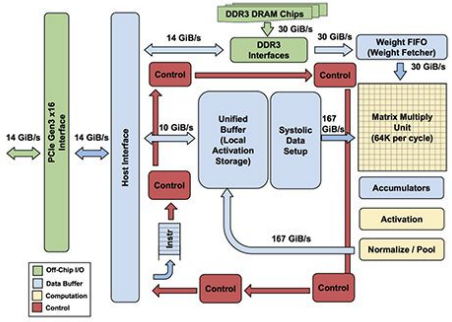
CPU + GPU

Scalability & Modularity

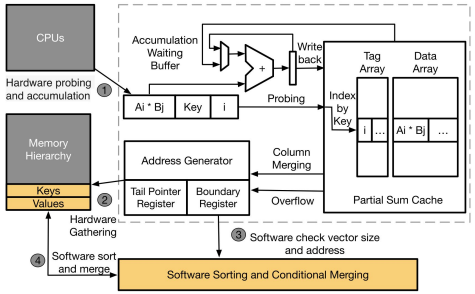


AMD. Two different products. Same chiplets

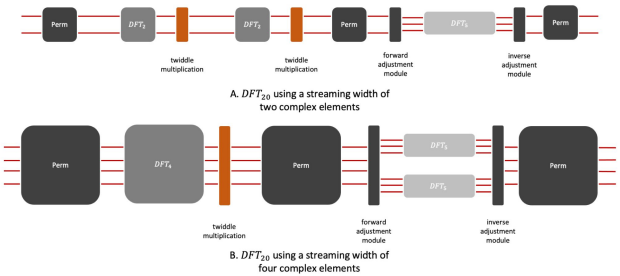
What is next: Specialization!



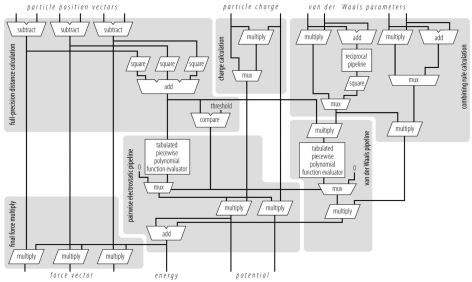
Tensor Processing Unit for Matrix-Matrix Multiplication. Google



Sparse Accumulation for SpGEMM. Lehigh University and LBNL



Hardware Implementation for Mixed-Radix Fourier Transform. LBNL



Pairwise Point Interaction Pipeline for Molecular Dynamics. Anton.

Potential

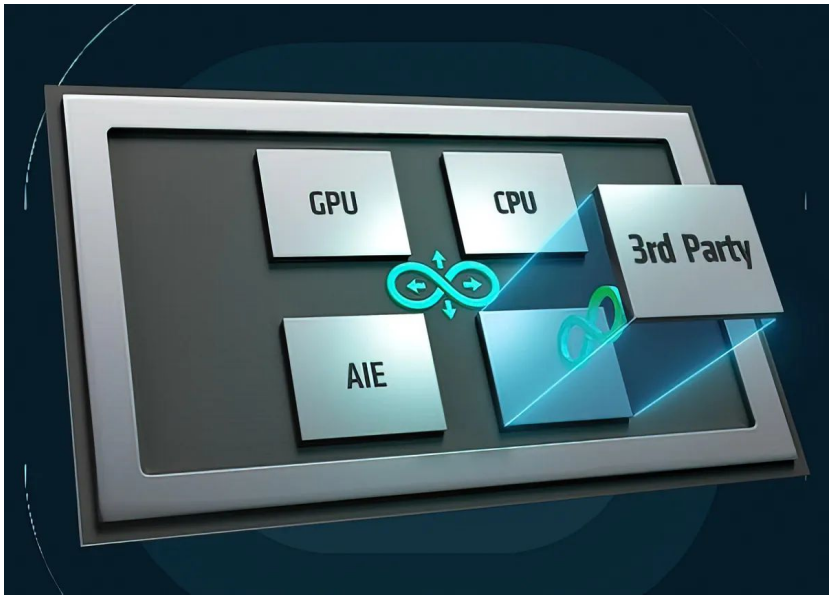
- 100x - 1000x more performance
- 100x - 1000x less energy consumption

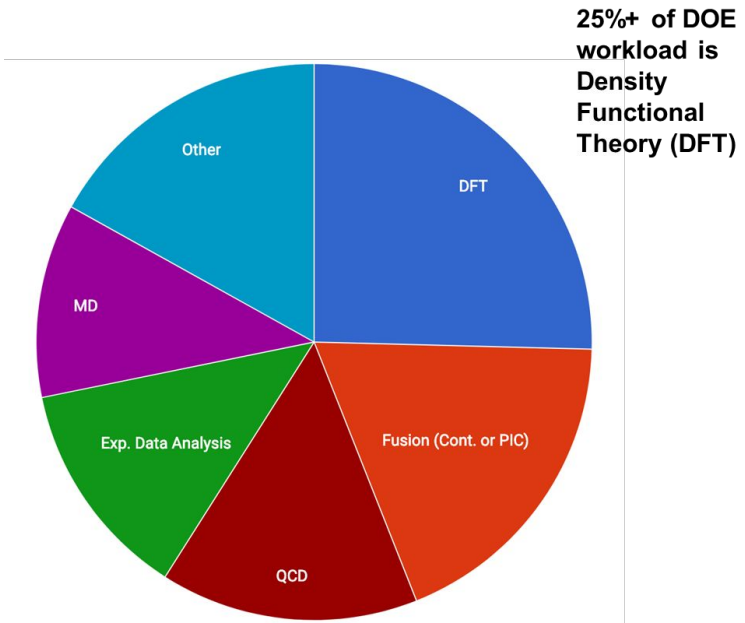
Challenges

- Overspecialization
- Algorithms evolve faster than hardware
  - RNN

How do we leverage chiplets modularity so hardware specialization makes it into HPC clusters?

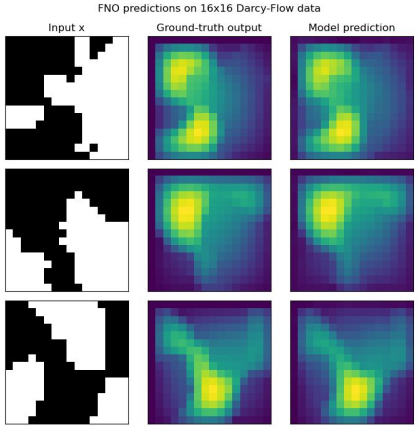
What would it take to add an specialized chiplet to an HPC SiP?

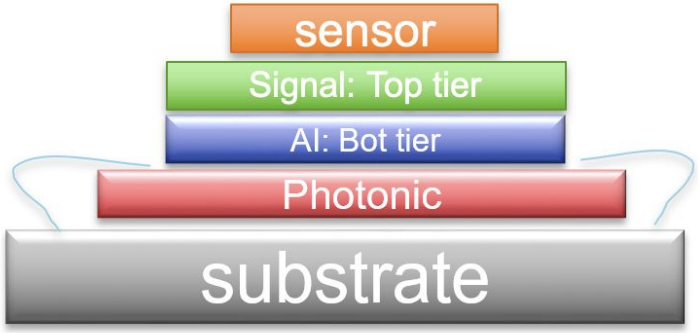




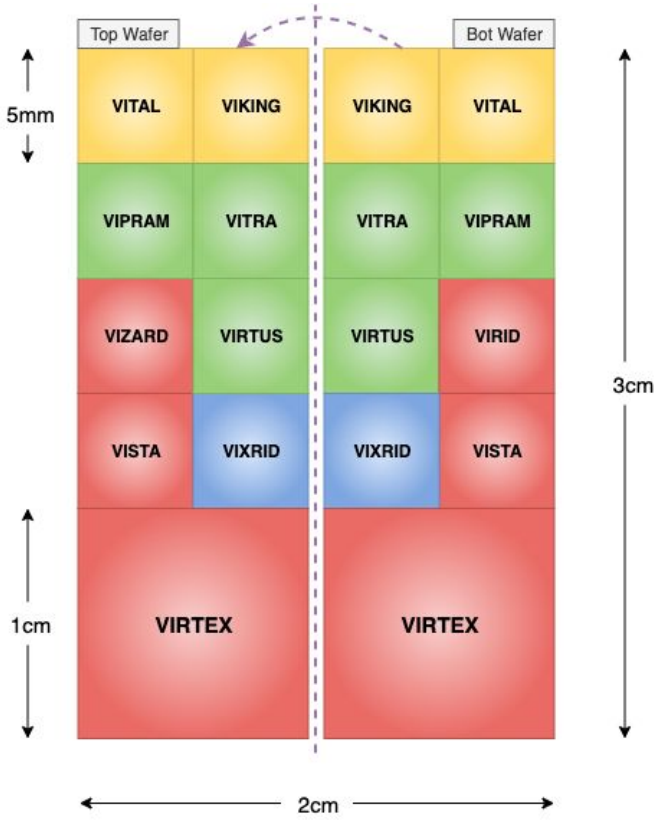
Project 38 (P38) - PI: John Shalf

VIAS: Vertically Integrated Artificial Intelligence for Sensing and High Performance Computing





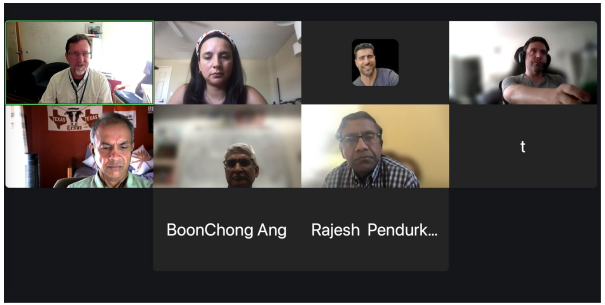
PI: Farah Fahim



OCP-Server-Workstream: Chiplets Modularity for HPC & AI

Standards for plug & play chiplets.

Bi-Weekly
Tuesdays 9 am PT



OPEN Compute Project®

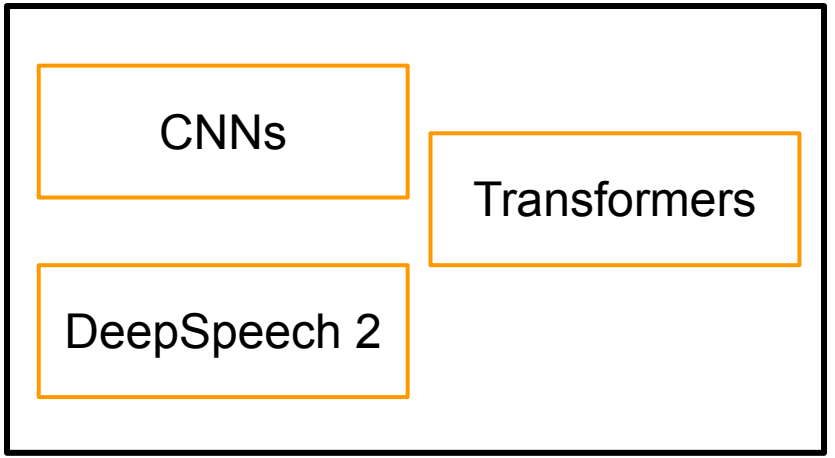
Recipe for Hardware Specialization enabled by chiplets modularity in 6 steps

Ingredients:

- *4 Chiplets*
- *1 Interposer*

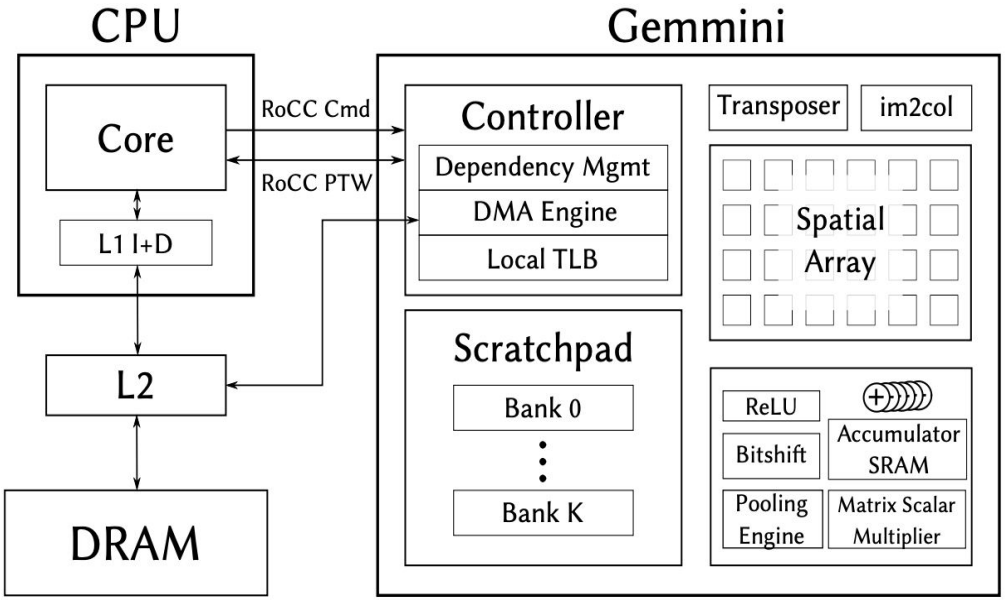
Chiplets Marketplace ;)

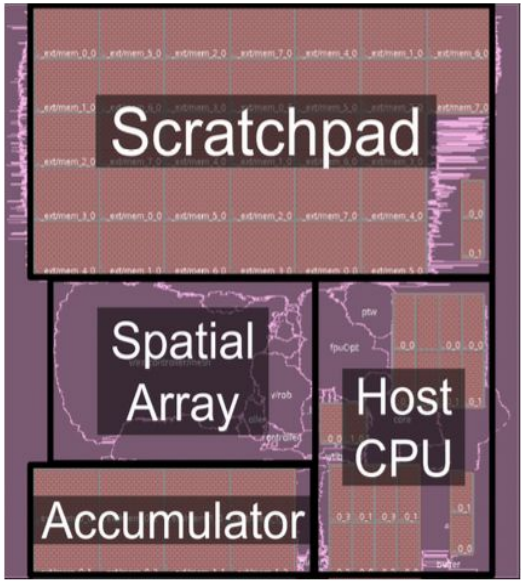
Example: AI



Three optimal Systems in Package (SiP) for AI

Step 0: Identify specialized hardware accelerator functional blocks

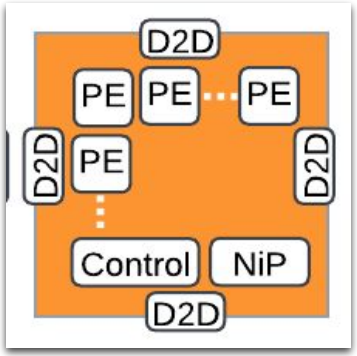




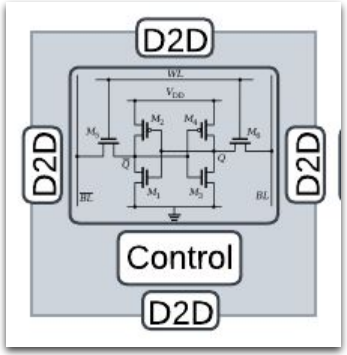
https://github.com/ucb-bar/gemmini

Shao, et al,

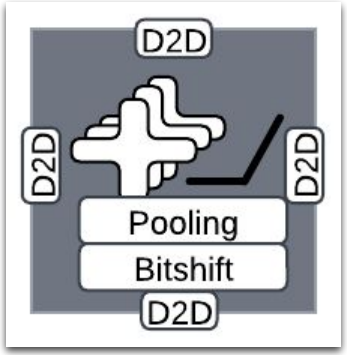
Step 1: Chipletization of hardware accelerator



Systolic Array Chiplet
Grid of PxP PE



Scratchpad Chiplet
SRAM based



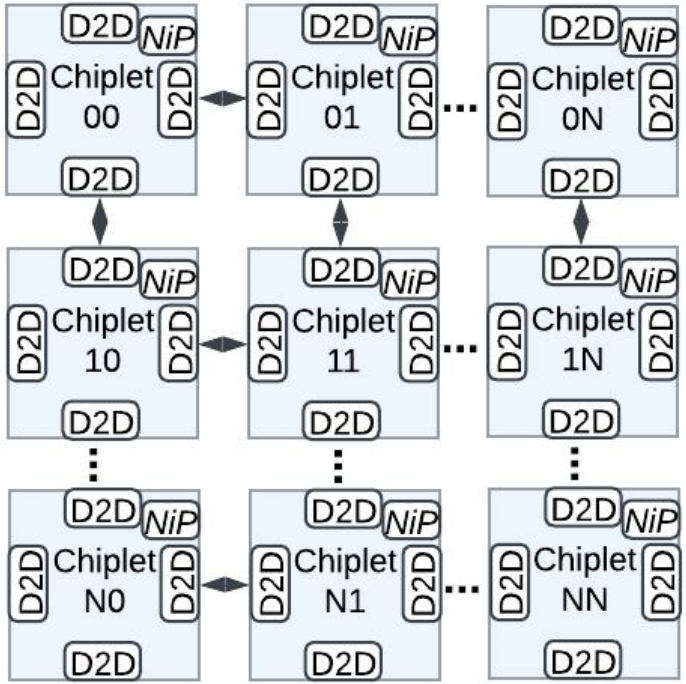
Accumulator Chiplet

Chiplets Marketplace ;)

Step 2: Scalable Interposer 2D grid with NxN Chiplets

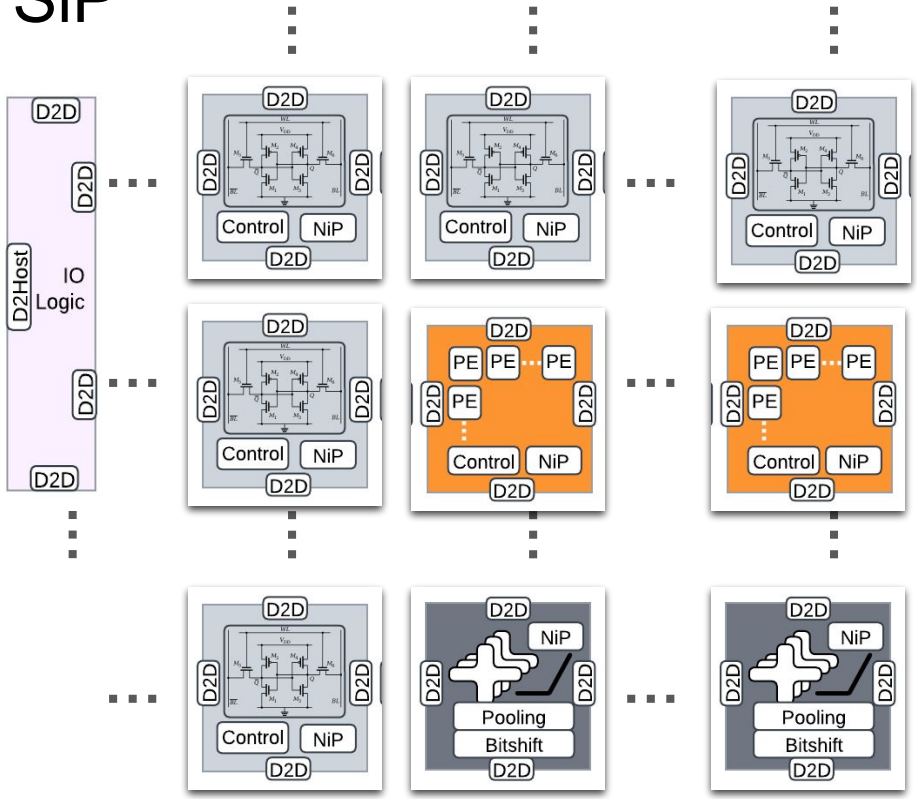
- **Network-In-Package (NiP)**
  - Packet switched
  - Light weight
  - Minimum transmission latency (1CC)
- **No Cache Coherency**
  - Message passing for parallel computing
    - Hardware Message Queues

Dataflow is regular and predictable



Step 3: Architecture for an AI SiP

- Scratchpad chiplets
  - Left: Input/Output Buffer
  - Top: Trained parameters
- KxK systolic arrays chiplets
- Accumulator chiplets : last row
- IO Chiplet: From and to host



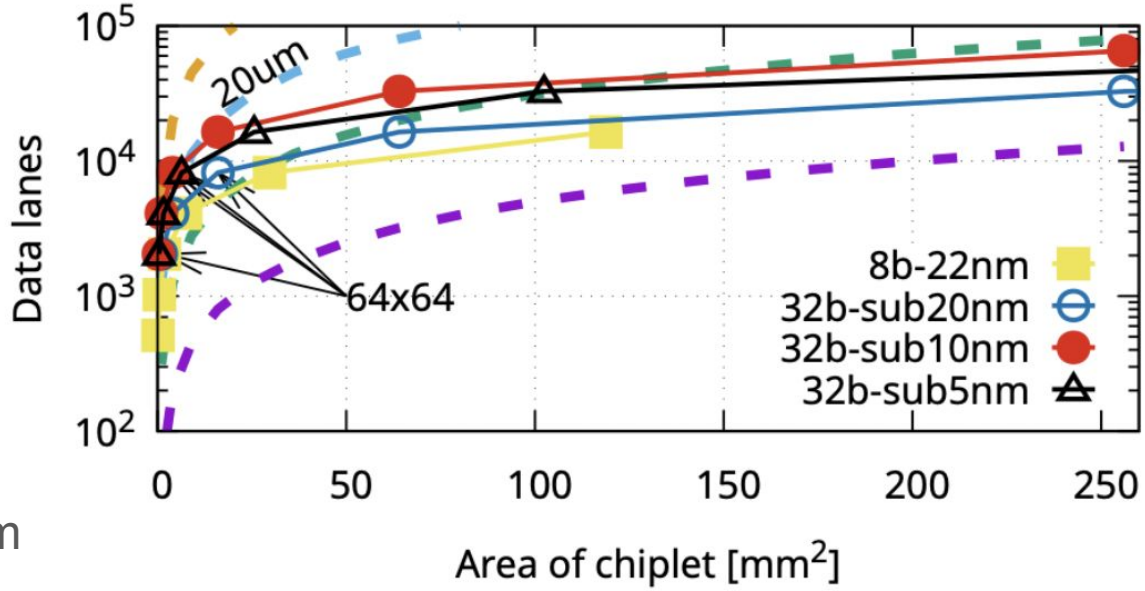
Step 4: Let's do some math and define our chiplets parameters

Bump Pitch > 40um

- Standard Interfaces: UCIe, BoW
- High frequency + serialization = more Power and latency

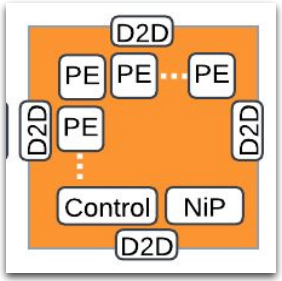
Bump Pitch = 20um

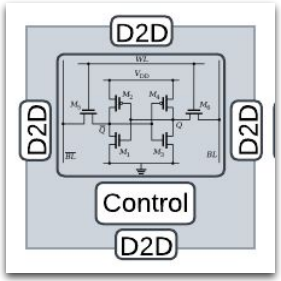
- Chiplet area 4mm x 4mm
- Oversized Buffers

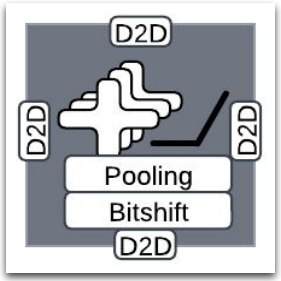


Area of chiplet: Logic (Systolic array) + NiP + D2D Interface

Step 4: Let's do some math and define our chiplets parameters



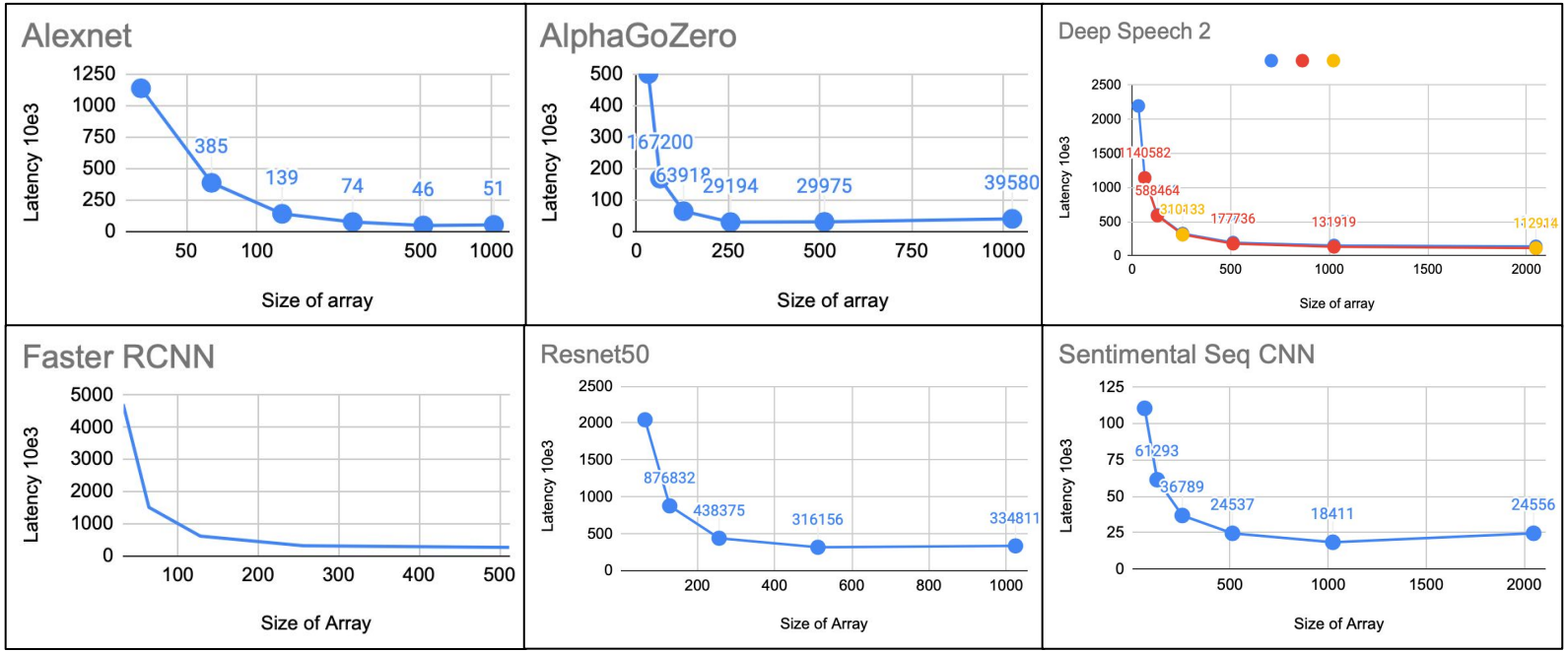




	Chiplet 4 mm x 4mm	Sub 20nm @ 1GHz	Sub 10nm @ 1GHz Sub 5nm @ 2GHz
1	**Systolic Array**	64 x 64 PEs	128 x 128 PEs
2	**Scratchpad**	1MB	9MB
3	**Accumulator**	64 acc	128 acc

Step 5: Profiling AI algorithms

- **Optimal Size of systolic array**
- **Optimal Scratchpad capacity (MB)**

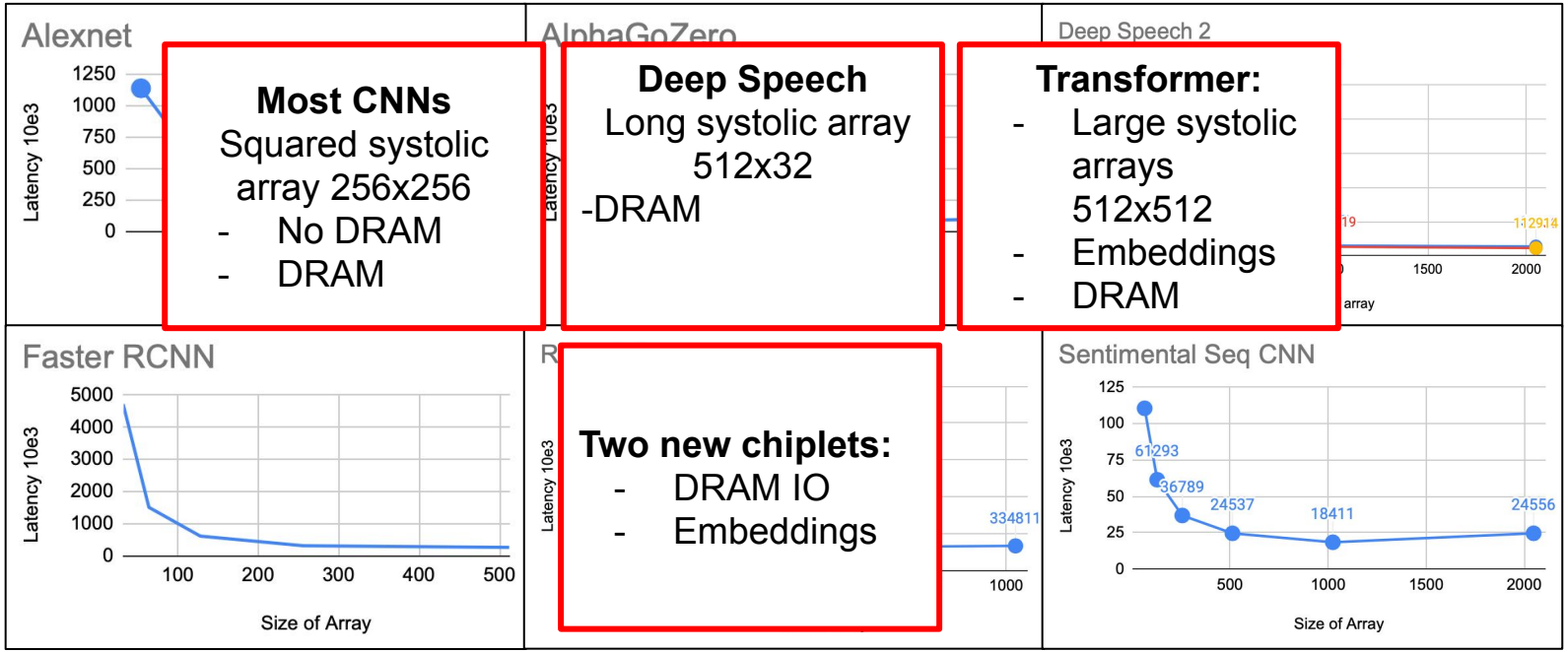


...

...

Step 5: Profiling AI algorithms

- **Optimal Size of systolic Array**
- **Optimal Scratchpad capacity (MB)**



Step 6: Three different SiPs for three dataflows

	Sub 20nm SiP		
	CNN	Deep Speech	Trans.
Syst. Array PEs	256x256	512x64	512x512
Syst. Array Chiplets Num.	16	8	64
Scratchpad Chiplets Num. (3D)	16	27	20
Accumulator Chiplets Num.	4	1	8
Embeddings Chiplets Num.	0	0	8
BW [TB/s] per Chiplet	1	1	1
Data lanes per Chiplet	8192	8192	8192
Chiplet Grid	6x6	6x6	10x10
Interposer Size* [mm x mm]	38x38	38x38	64x64

Step 6: Three different SiPs for three dataflows

	Sub 20nm SiP		
	CNN	Deep Speech	Trans.
Syst. Array PEs	256x256	512x64	512x512
Syst. Array Chiplets Num.	16	8	64
Scratchpad Chiplets Num. (3D)	16	27	20
Accumulator Chiplets Num.	4	1	8
Embeddings Chiplets Num.	0	0	8
BW [TB/s] per Chiplet	1	1	1
Data lanes per Chiplet	8192	8192	8192
Chiplet Grid	6x6	6x6	10x10
Interposer Size* [mm x mm]	38x38	38x38	64x64

Step 6: Three different SiPs for three dataflows

	Sub 20nm SiP		
	CNN	Deep Speech	Trans.
Syst. Array PEs	256x256	512x64	512x512
Syst. Array Chiplets Num.	16	8	64
Scratchpad Chiplets Num. (3D)	16	27	20
Accumulator Chiplets Num.	4	1	8
Embeddings Chiplets Num.	0	0	8
BW [TB/s] per Chiplet	1	1	1
Data lanes per Chiplet	8192	8192	8192
Chiplet Grid	6x6	6x6	10x10
Interposer Size* [mm x mm]	38x38	38x38	64x64

Step 6: Three different SiPs for three dataflows

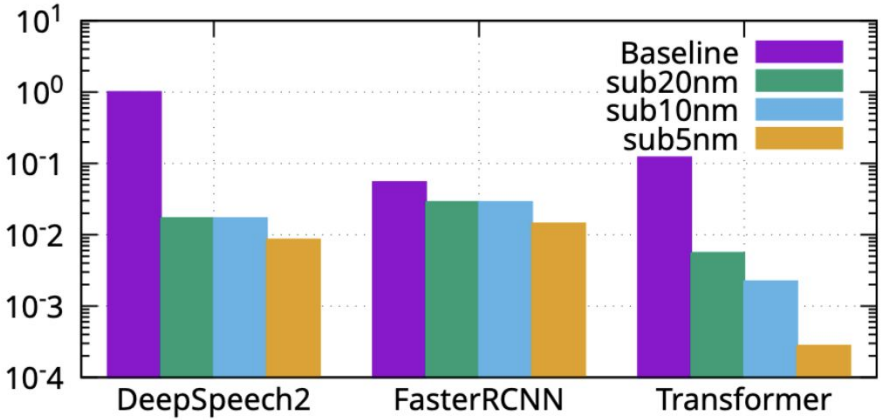
	Sub 20nm SiP		
	CNN	Deep Speech	Trans.
Syst. Array PEs	256x256	512x64	512x512
Syst. Array Chiplets Num.	16	8	64
Scratchpad Chiplets Num. (3D)	16	27	20
Accumulator Chiplets Num.	4	1	8
Embeddings Chiplets Num.	0	0	8
BW [TB/s] per Chiplet	1	1	1
Data lanes per Chiplet	8192	8192	8
Chiplet Grid	6x6	6x6	10
Interposer Size* [mm x mm]	38x38	38x38	64x64

Step 6: Three different SiPs for three dataflows

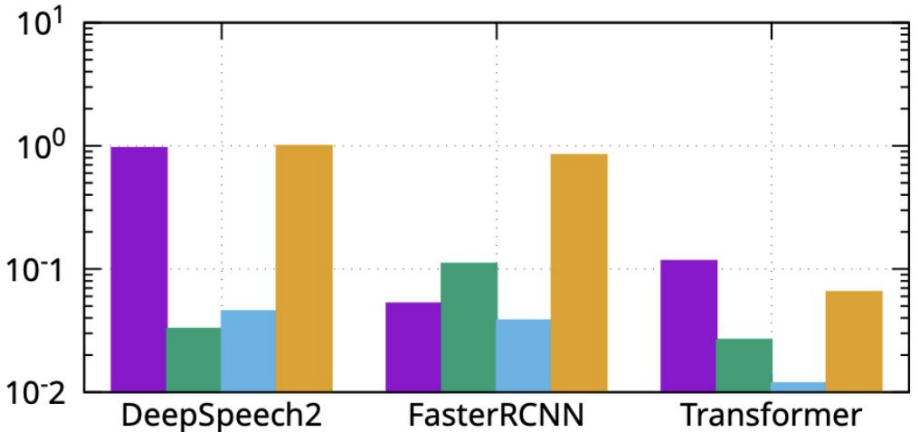
	Sub 20nm SiP			Sub 10nm/5nm SiP		
	CNN	Deep Speech	Trans.	CNN	Deep Speech	Trans.
Syst. Array PEs	256x256	512x64	512x512	256x256	512x128	512x512
Syst. Array Chiplets Num.	16	8	64	4	8	16
Scratchpad Chiplets Num. (3D)	16	27	20	10	27	12
Accumulator Chiplets Num.	4	1	8	2	1	4
Embeddings Chiplets Num.	0	0	8	0	0	4
BW [TB/s] per Chiplet	1	1	1	2/4	2/4	2/4
Data lanes per Chiplet	8192	8192	8	16384	8192	16384
Chiplet Grid	6x6	6x6	10	4x4	6x6	6x6
Interposer Size* [mm x mm]	38x38	38x38	64x64	26x26	38x38	38x38

Evaluation: Better Latency and Better Energy consumption

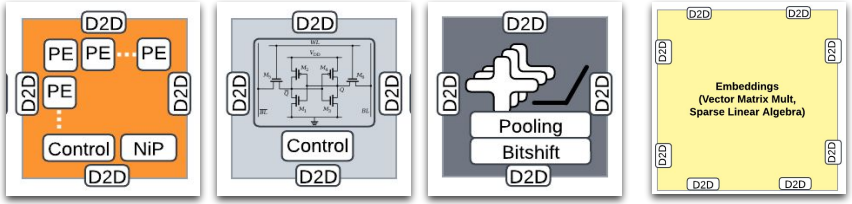
Normalized Latency



Normalized Energy Consumption



Overspecialization: Fear Not!



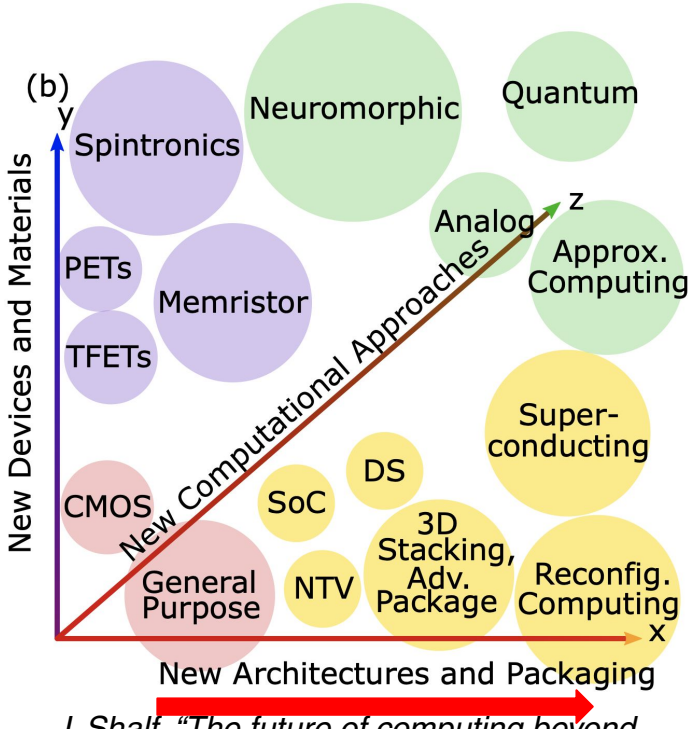
4 Chiplets + 1 interposer = 4 highly specialized SiPs for AI

+100x less latency than baseline
+10x better energy consumption than baseline

Final thoughts and next steps

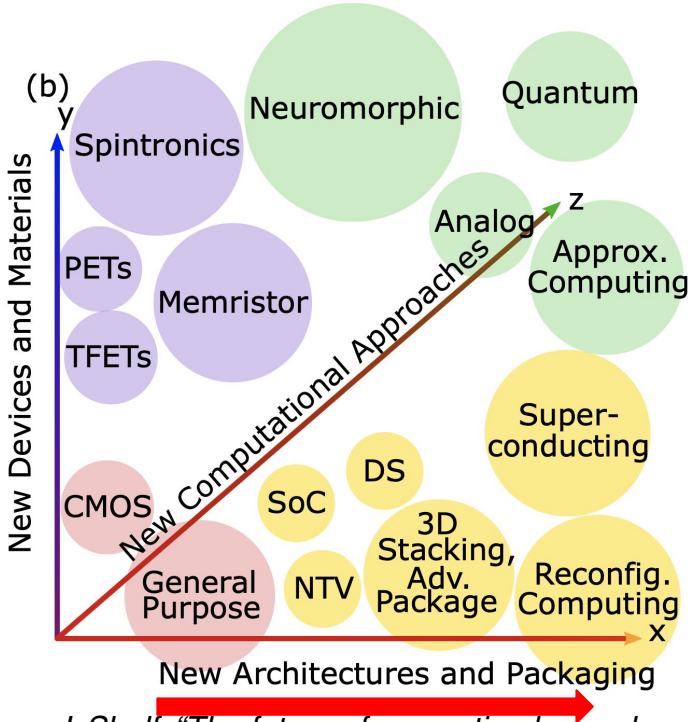
- Chiplet ≠ Application
- Chiplet = Kernel, *i.e. (GEMM)*
- **HPC applications:**
  - Density Functional Theory
  - Molecular Dynamics
  - Climate modeling
  - Processing In Cell
  - NEGFs
  - SpVM

Beyond Moore avenues to improve performance

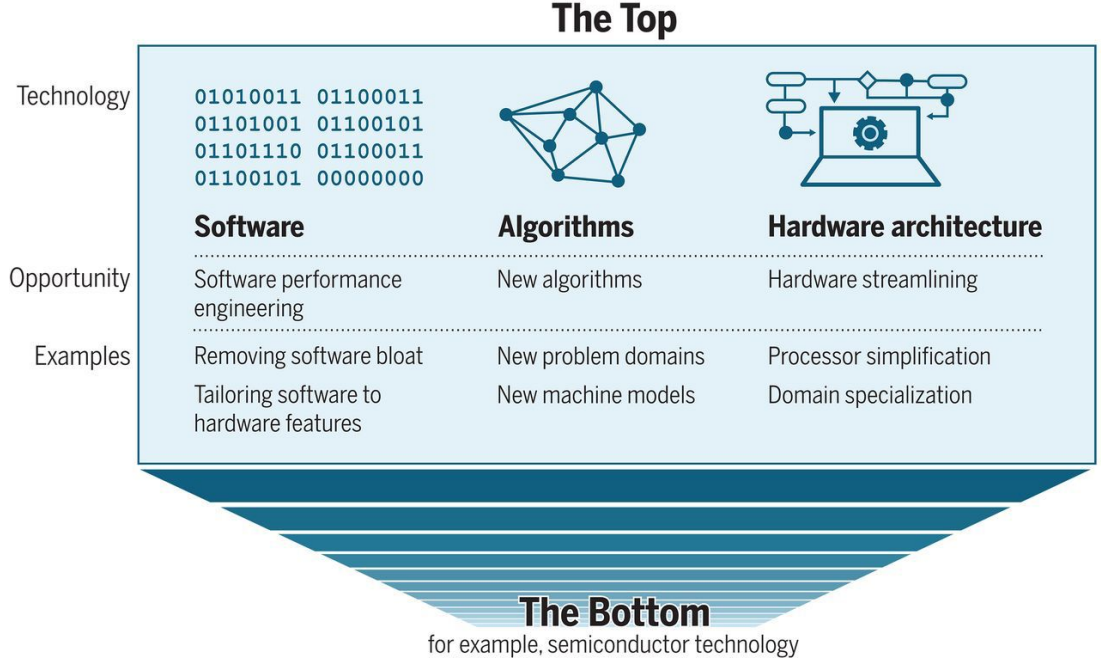


J. Shalf, "The future of computing beyond Moore ' s Law," 2020.

Beyond Moore avenues to improve performance (x axis)

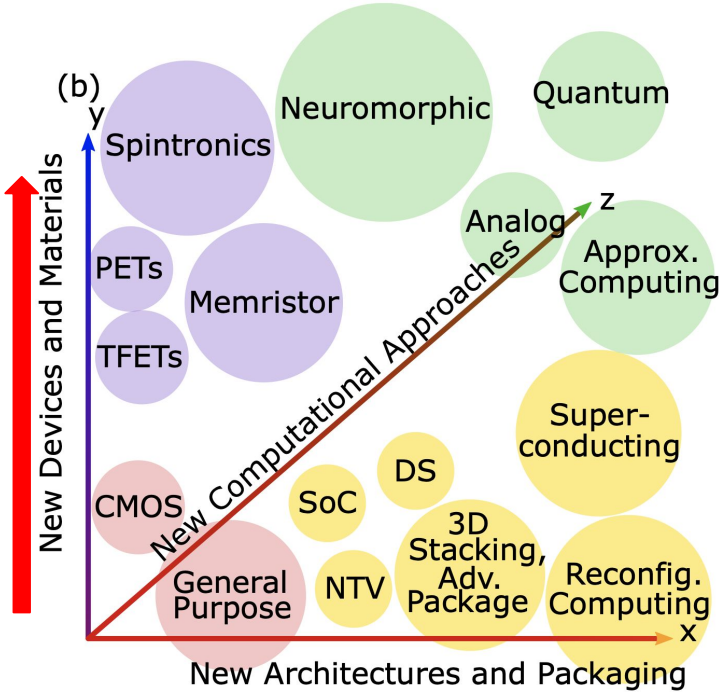


J. Shalf, "The future of computing beyond Moore ' s Law," 2020.

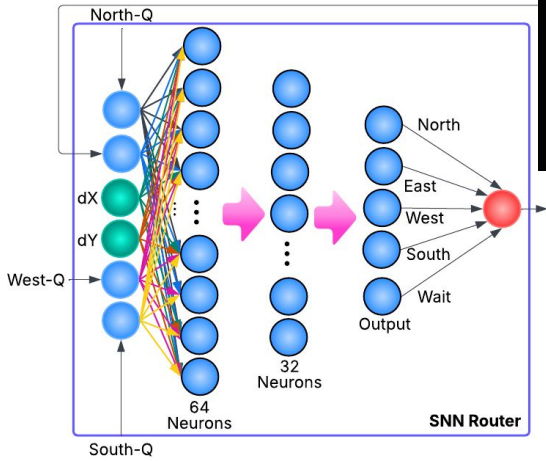


Charles E. Leiserson, *et al.* ,There's plenty of room at the Top: What will drive computer performance after Moore's law?. *Science* **368**, eaam 9744 (2020). DOI:10.1126/science.aam9744

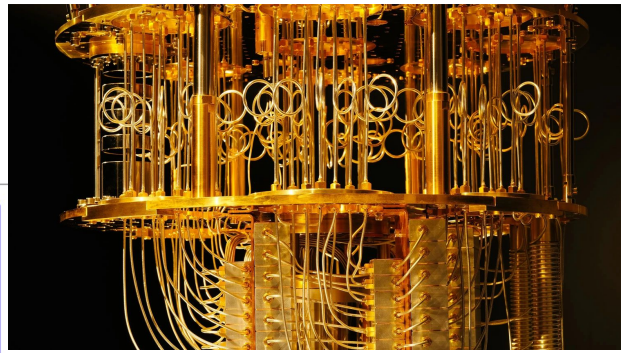
Beyond Moore avenues to improve performance (y axis)



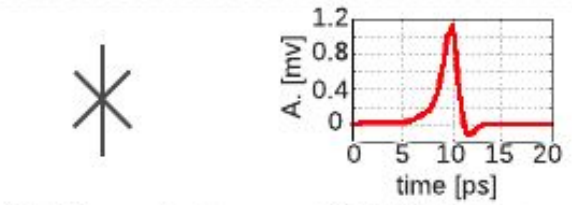
J. Shalf, "The future of computing beyond Moore ' s Law," 2020.



Neuromorphic

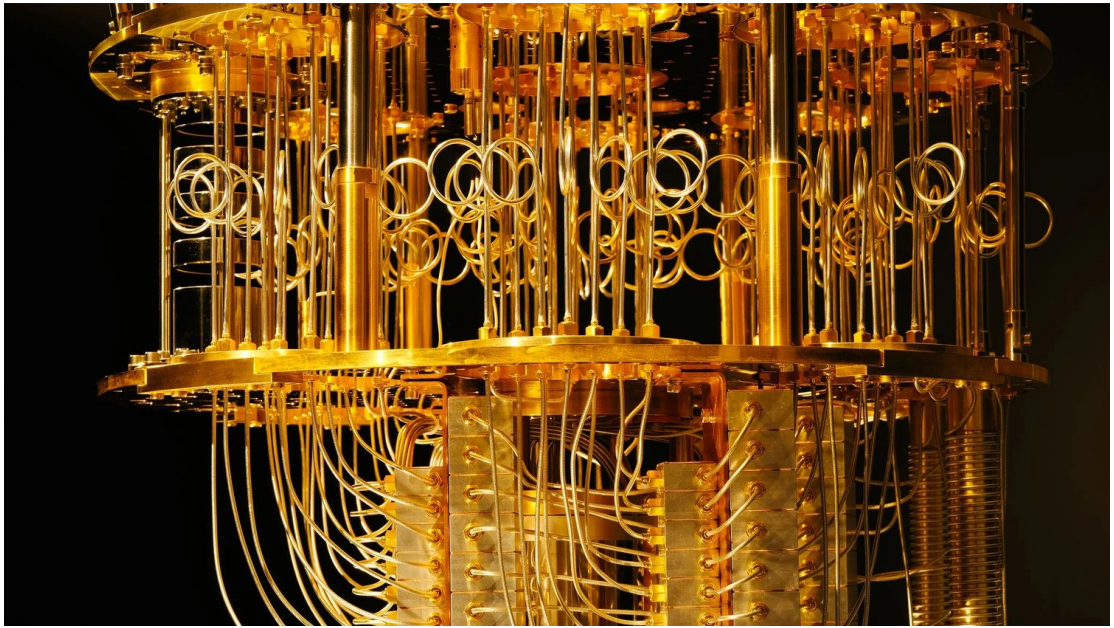


Quantum Computing



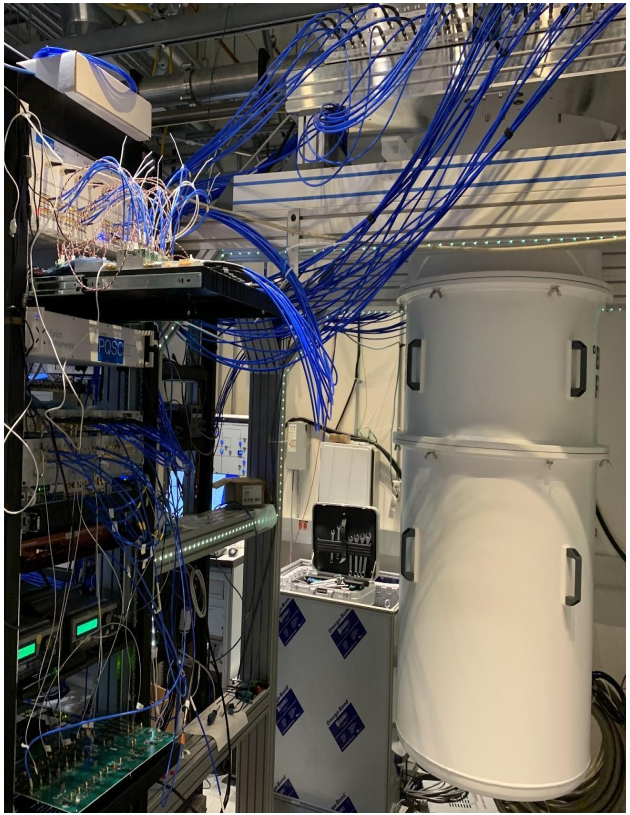
Superconducting computing

What I think about quantum computing



The real quantum computing

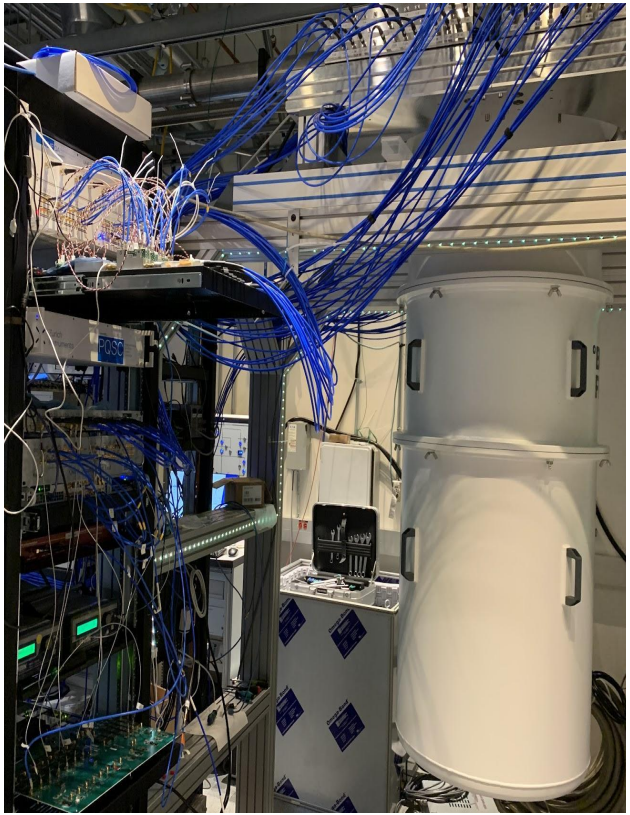
- Dilution refrigerator
  - Qubits, filters, amplifiers,
  - Superconducting coaxial line
  - Cryogenic isolators
- Room temperature control
  - Waveform generators
  - Data Converters (Digital-to-Analog & Analog-to-Digital)
  - Control and data flow processors
- Limitations
  - Not scalable
  - Expensive equipment
  - High latencies
  - IO bottleneck



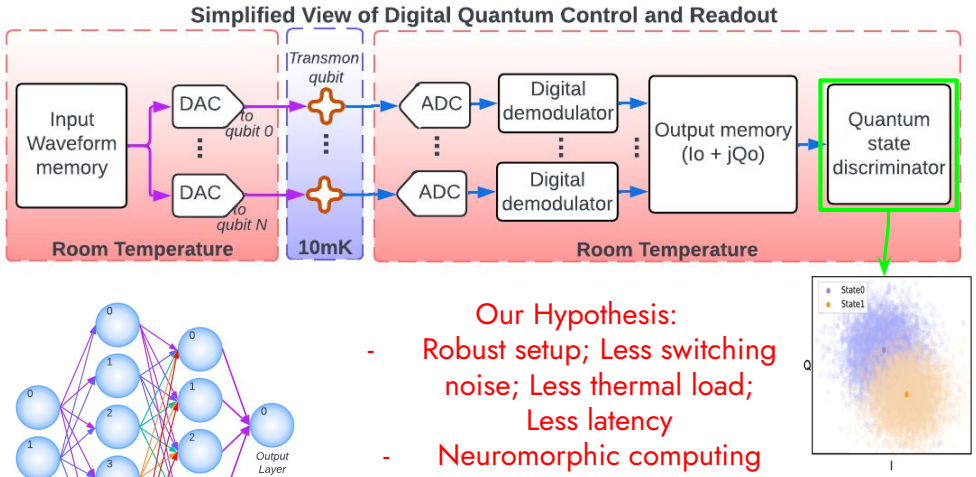
The real quantum computing

- Dilution refrigerator
  - Qubits, filters, amplifiers,
  - Superconducting coaxial line
  - Cryogenic isolators
- Room temperature control
  - Waveform generators
  - Data Converters (Digital-to-Analog & Analog-to-Digital)
  - Control and data flow processors
- Limitations
  - Not scalable
  - Expensive equipment
  - High latencies
  - IO bottleneck

Simply not scalable



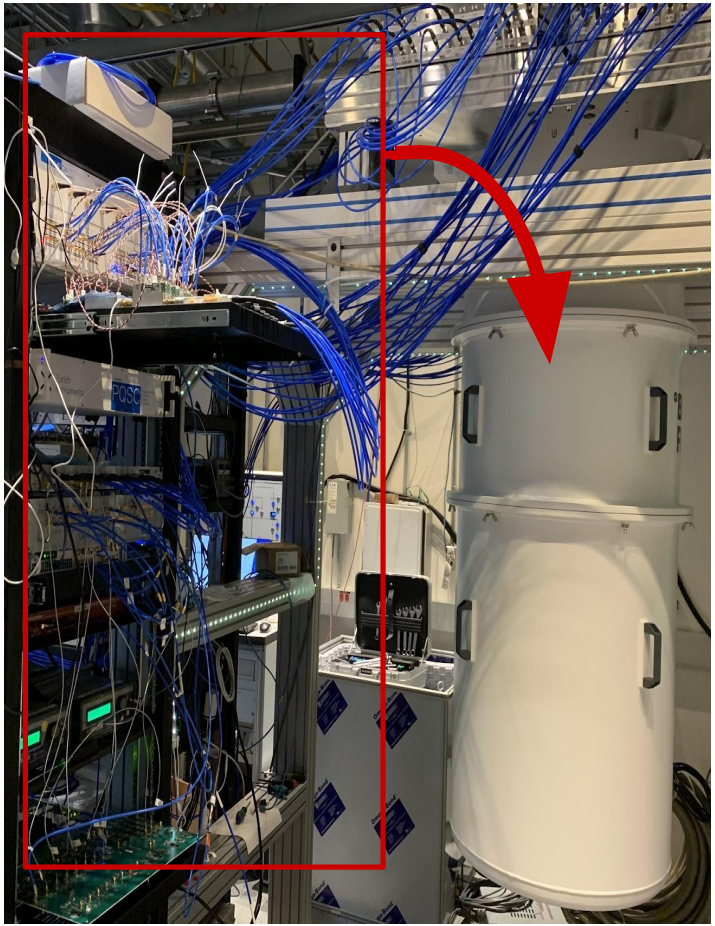
It is possible through superconducting computing



Our Hypothesis:

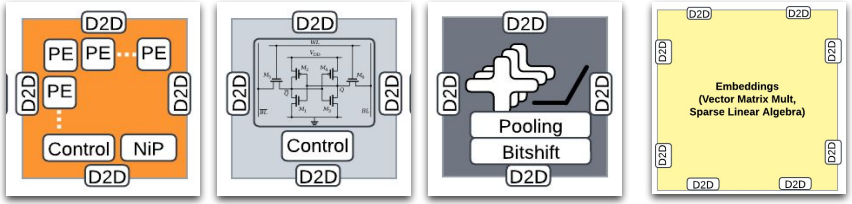
- Robust setup; Less switching noise; Less thermal load; Less latency
- Neuromorphic computing
- Native cryogenic operation

Just another example of specialization



Thanks!

lg4er@lbl.gov



4 Chiplets + 1 interposer = 4 highly specialized SiPs for AI

+100x less latency than baseline
+10x better energy consumption than baseline

Final thoughts and next steps

- Chiplet ≠ Application
- Chiplet = Kernel, *i.e. (GEMM)*
- **HPC applications:**
  - Density Functional Theory
  - Molecular Dynamics
  - Climate modeling
  - Processing In Cell
  - NEGFs
  - SpVM