



Superconducting Array of Arrays for Acceleration of Transformers

Manu Perumkunnil, Kartik Lakshminarasimhan, Udara De Silva, Debjoyti Bhattacharjee, Trent Josephson, Quentin Herr, Anna Herr

OVERVIEW

You can add a subtitle if you want.

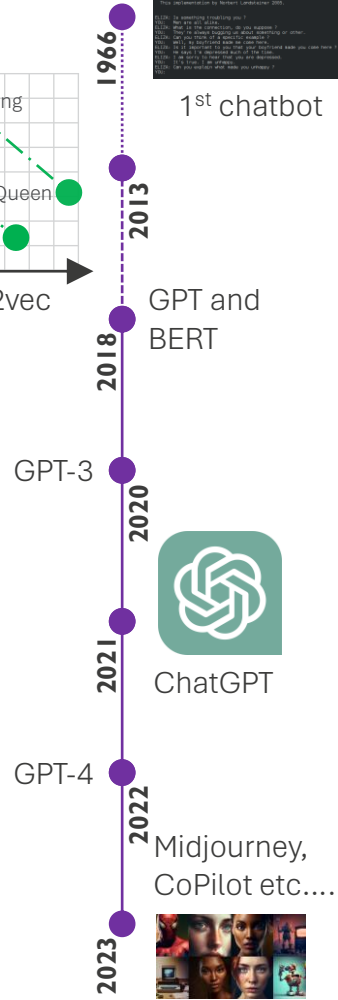
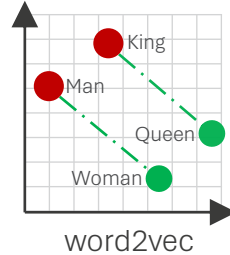
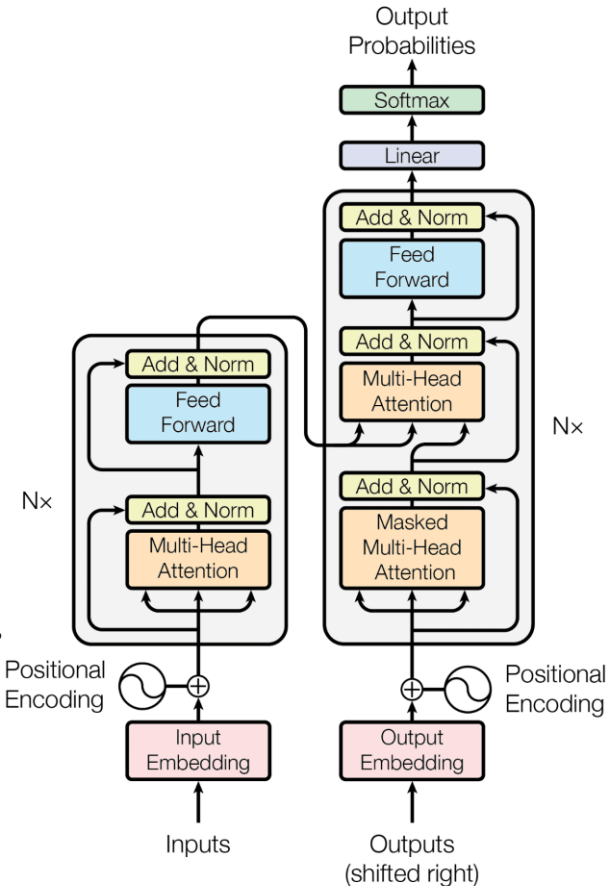
- Introduction & Motivation
- Accelerating LLMs via Superconducting Digital Systems
- Scaling up
- Conclusions & Future

Introduction & Motivation

Introduction & Motivation

Why Transformers/LLMs

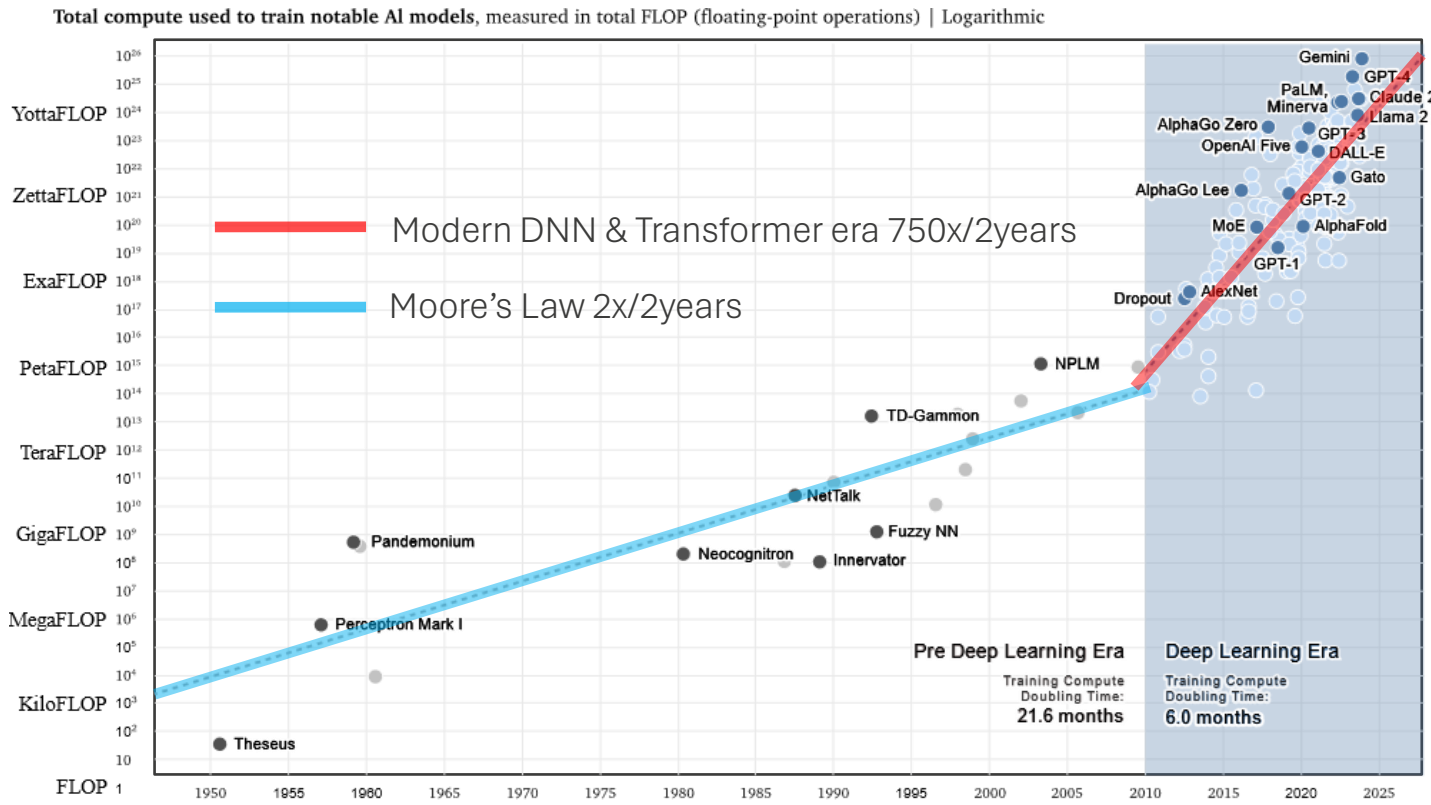
- Increasingly common for a variety of tasks – *text generation, vision, translation, code generation, classification, etc*
- Superior performance in sequence-to-sequence family of ML algorithms compared to RNNs & LSTMs
- Exploits all kinds of parallelisms – *data, pipeline, tensor/model & expert*
- Consists of encoder & decoder blocks



Introduction & Motivation

Cost of AI Training

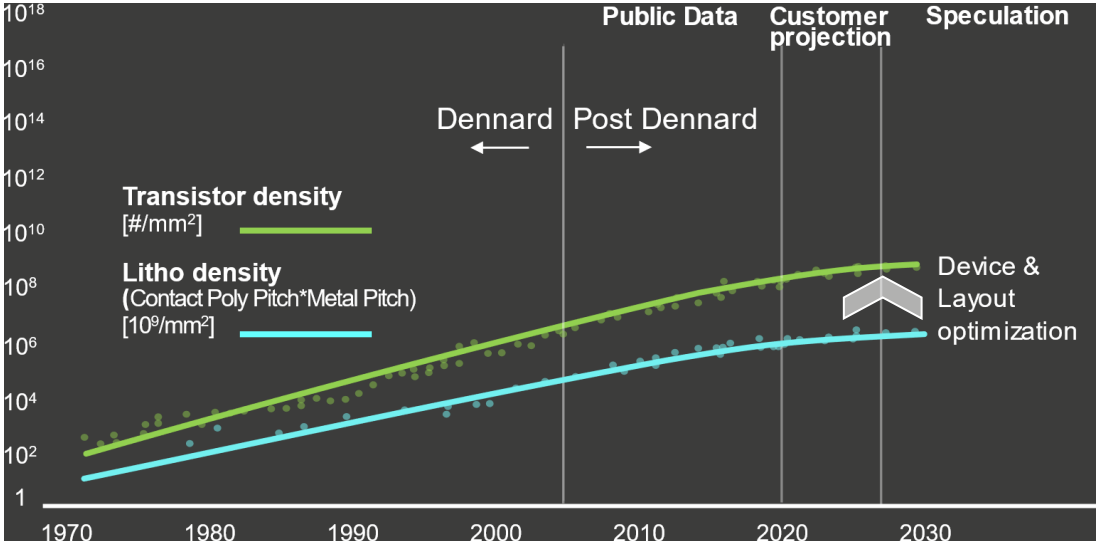
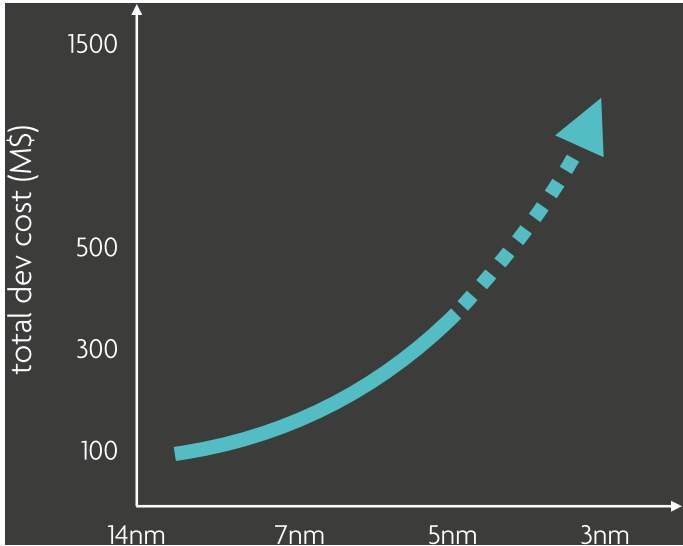
- Exponential rise in overall cost of training large AI models → Silicon, NRE, Infrastructure & Power/Thermals



Introduction & Motivation

Cost of AI Training

- Soaring cost of Silicon (& design) as we hit process technology scaling limits

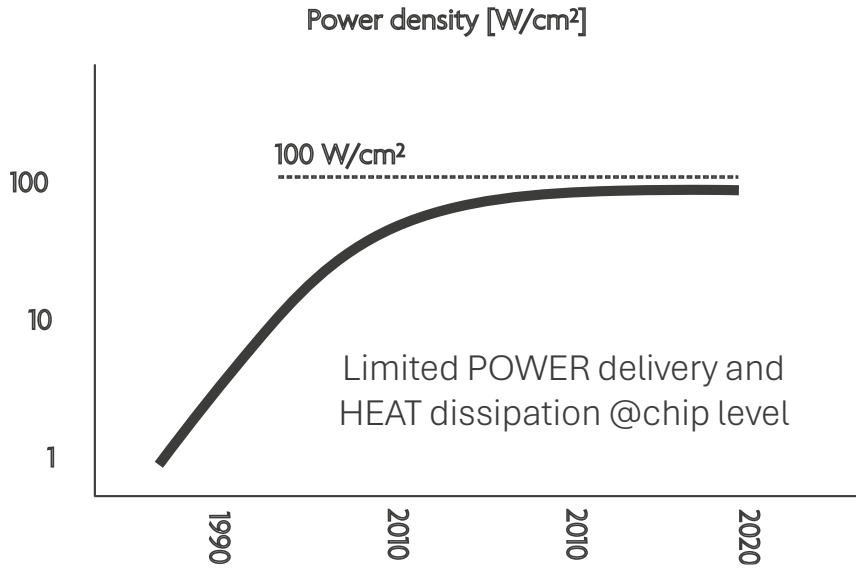
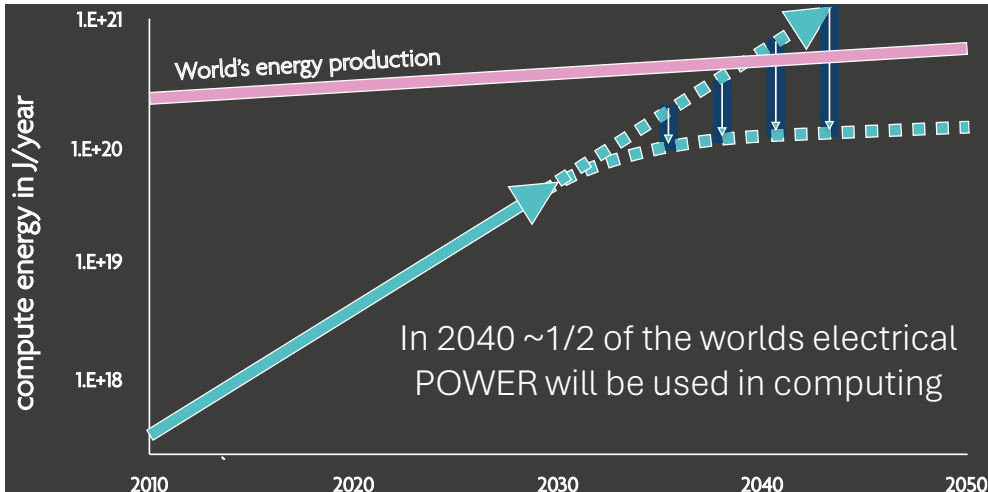


[asml-investor-day-2021-technology-strategy---martin-van-den-brink.pdf](#)

Introduction & Motivation

Cost of AI Training

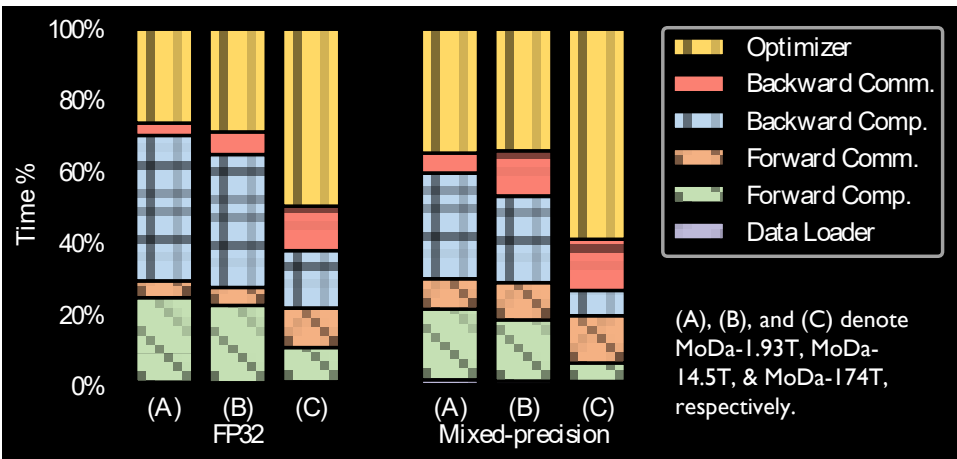
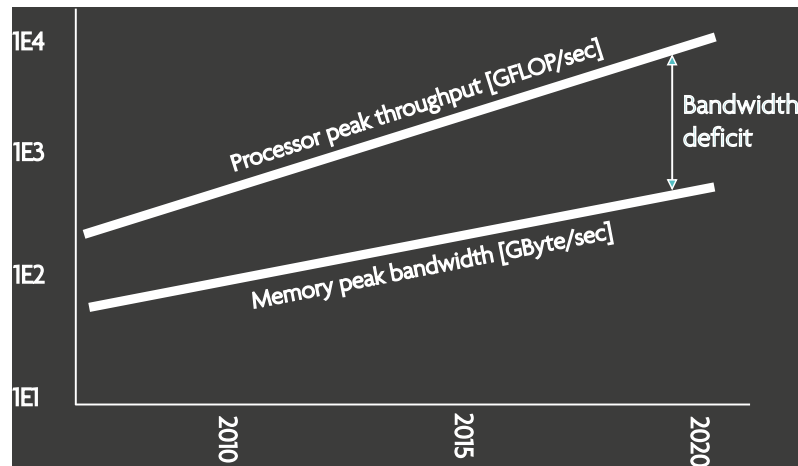
- Power distribution & thermal management cost (including Infra)



Introduction & Motivation

Cost of AI Training

- Cost to overcome Memory & Interconnect (*Infra and otherwise*) bottlenecks that can severely underutilize compute in scaling AI/HPC clusters



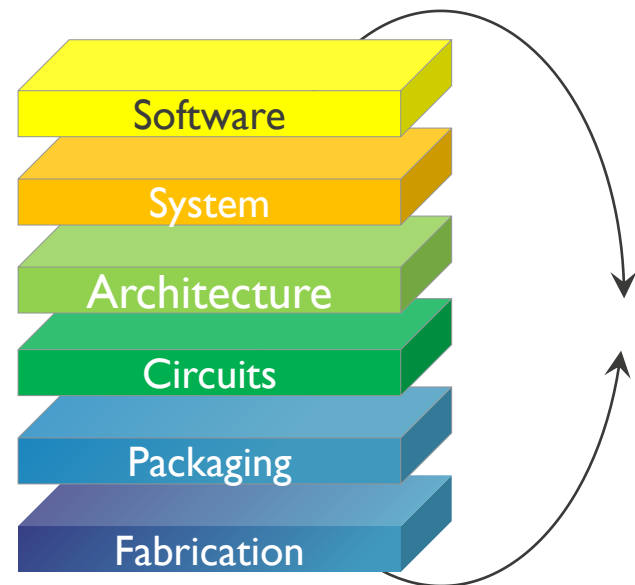
BaGuaLu: <https://doi.org/10.1145/3503221.3508417>

Accelerating LLMs via Superconducting Digital Systems

Accelerating LLMs via Superconducting Digital Systems

Full stack solution

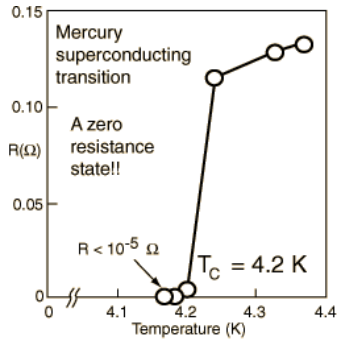
- Co-optimization across the stack is essential to exploit novel technology solutions to the maximum keeping in mind cost (*including NRE, infra & cost of adoption*)
- Superconducting technology offers →
 - THz bandwidth interconnects
 - Quantum accurate digital bits
 - Fast, low energy logic & memory



Accelerating LLMs via Superconducting Digital Systems

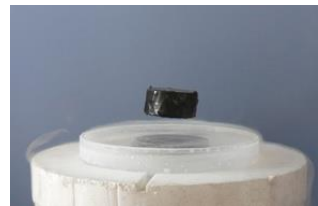
Superconducting technology

Zero resistance wires



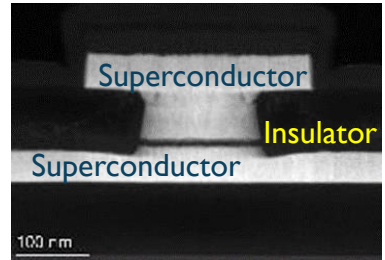
Enables THz bandwidth (optical like) interconnects that require no signal amplification

Meissner effect



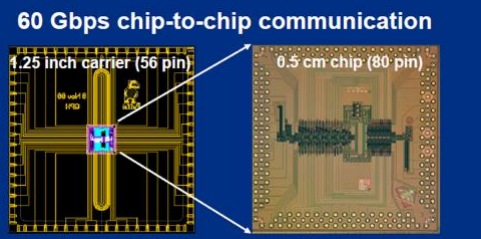
Allows for Quantum accurate digital bits (Fundamental limit for energy) with a large noise margin

Josephson effect

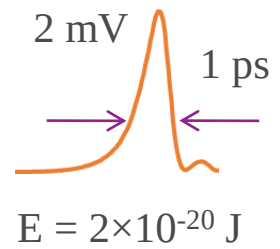
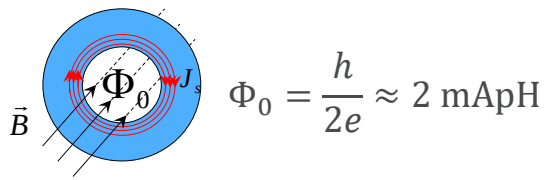


Enables fast, low energy logic & memory devices that can be scaled and used like classical CMOS technology at ultra-high frequencies (~30GHz)

High speed data link between digital superconductor chips:
<https://doi.org/10.1063/1.1473687>



An 8-bit carry look-ahead adder with 150 ps latency and sub-microwatt power dissipation at 10 GHz; <https://doi.org/10.1063/1.4776713>

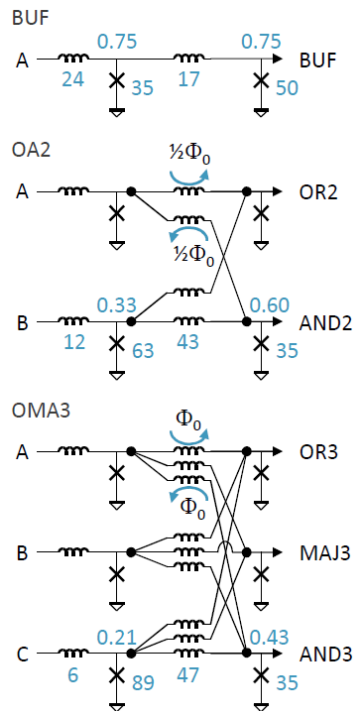


Accelerating LLMs

Superconducting digital logic and memory

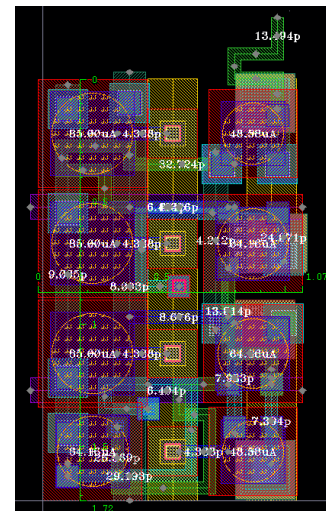
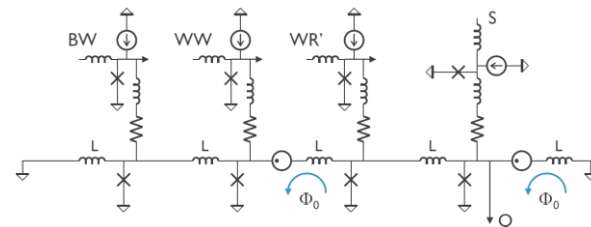
- Superconducting digital Pulse-Conserving Logic (PCL), Josephson SRAM (JSRAM) & AC power distribution via resonant networks allow for classical digital systems that are scalable
- Fabrication stack allows for digital logic with 16 layers, achieving up-to 400 MJJ/cm² device density. The stack allows scaling beyond 28 nm lithography & is compatible with standard high-temperature CMOS processes.

Gate Schematics for JJ based PCL



12

JSRAM unit cell schematics & Layout



The diagram illustrates the internal structure of a Transformer block, showing two layers (Layer 1 and Layer 2) and an ellipsis indicating further layers. Each layer consists of two sub-layers: Multi-Head Attention and Feed Forward. The input sequence (represented by a blue line) passes through these sub-layers, and the output of each sub-layer is added to the original input sequence via residual connections (indicated by '+' signs in circles). The output of the first layer then passes through the second layer, and so on.

one-hot feature vector (query: Q)

0	0	0	0	0	0	0	0	0	1	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---

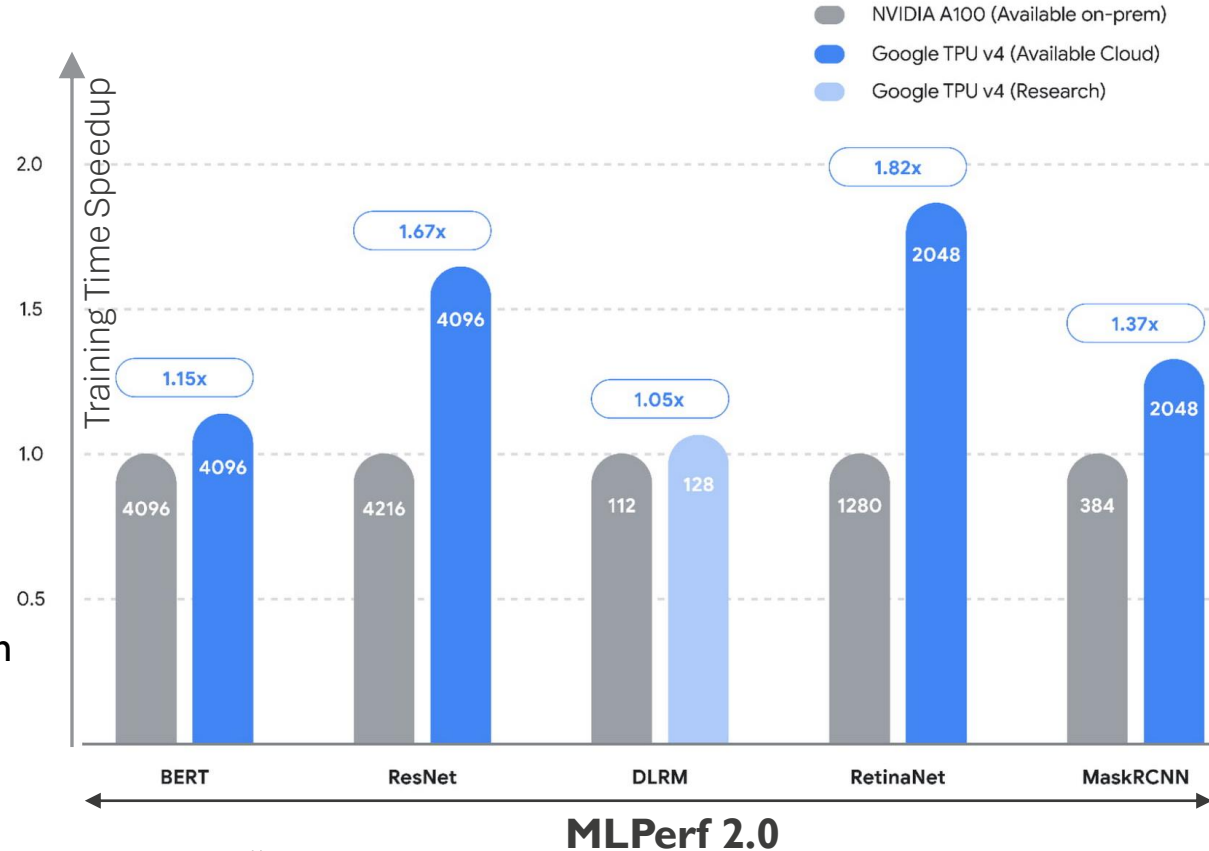
feature-specific mask

[illegible]

- 

Accelerating LLMs Architecture

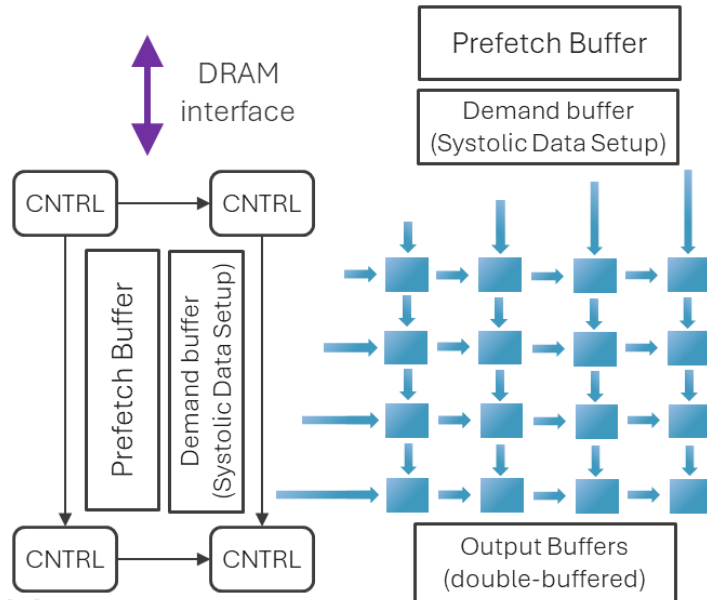
- Array architectures (like systolic arrays) are best suited to efficiently execute MMM kernels ($\gg$ FLOPS/chip due to $\gg$ MACs)
- They also have minimal control overhead $\rightarrow$ reduces design complexity, NRE & cost of adoption for superconducting tech



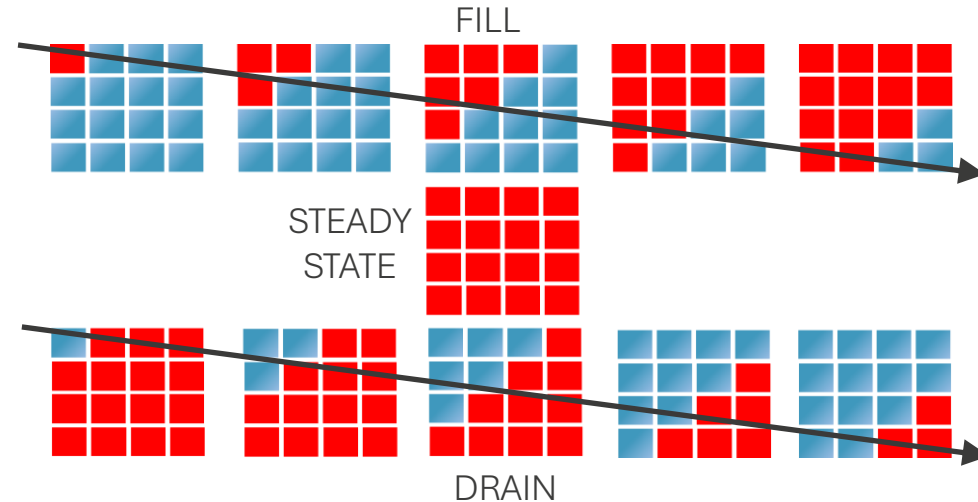
Accelerating LLMs

Architecture

- Control units to buffer dataflow from DRAM to MAC units

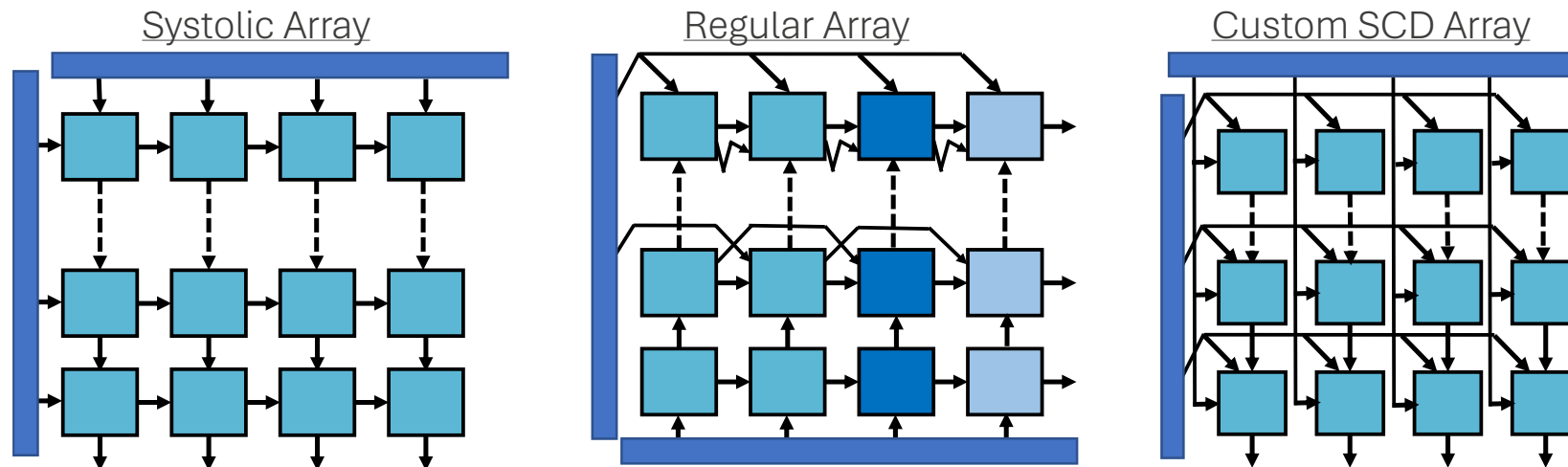


- However, simple systolic arrays incur pipeline bubbles during “FILL” & “DRAIN”



Accelerating LLMs

Array Architecture

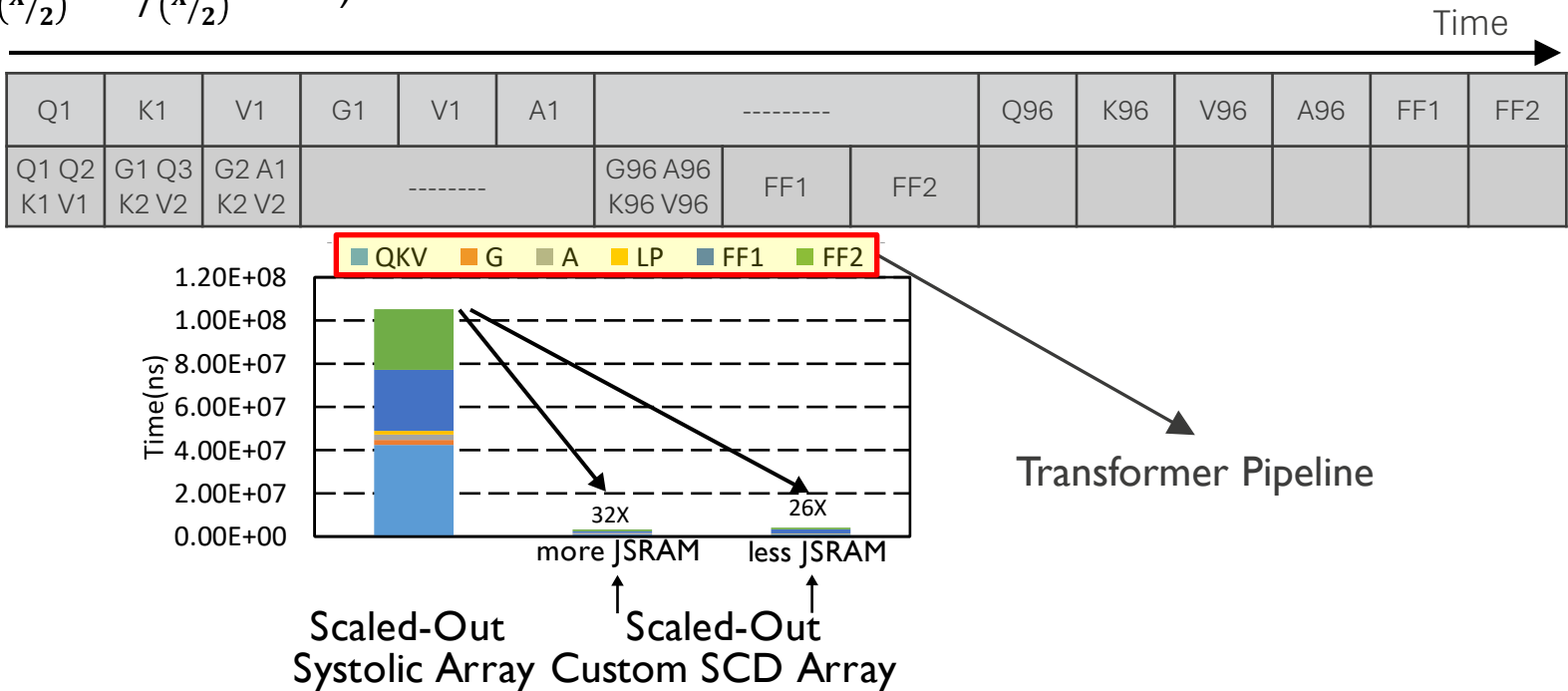


- Proposed Custom Superconducting Array is a homogenous version of regular arrays with superconducting wires.
- While this comes at the cost of more wires, the minimal wiring overhead (Power & Performance) incurred in superconducting digital technology makes this an ideal solution

Scaling the system

Scaling-Up Array Architecture

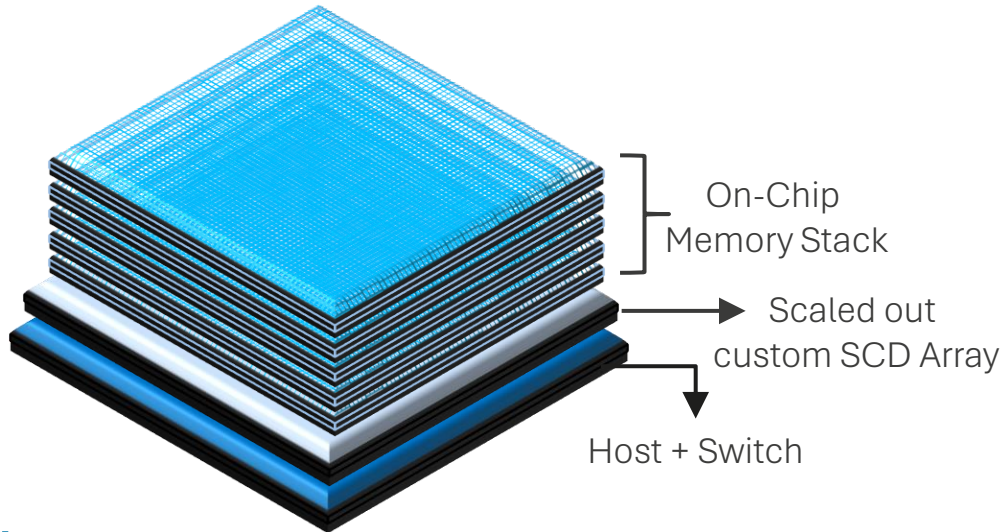
- Scale-up (Single Large PE array consisting of ' $N \times N$ ' MACs) vs Scale-out (' X ' PE arrays consisting of ' $N/(x/2) \times N/(x/2)$ ' MACs)



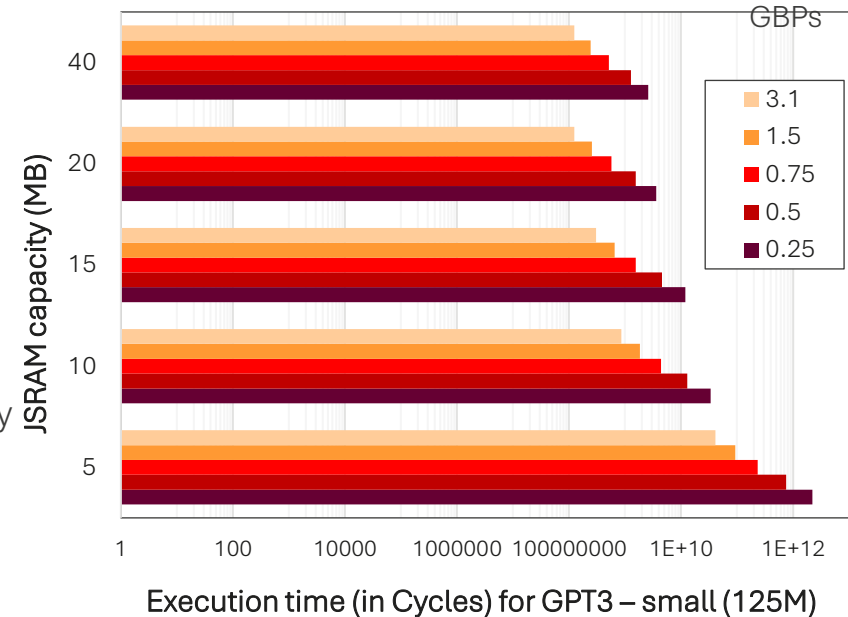
Scaling-Up

Superconducting Processing Stack

- Superconducting processing stack: multifunctional die stack 3D integrated via superconducting TSVs ($\leq 30\mu\text{m}$ pitch, $10\mu\text{m}$ diameter)



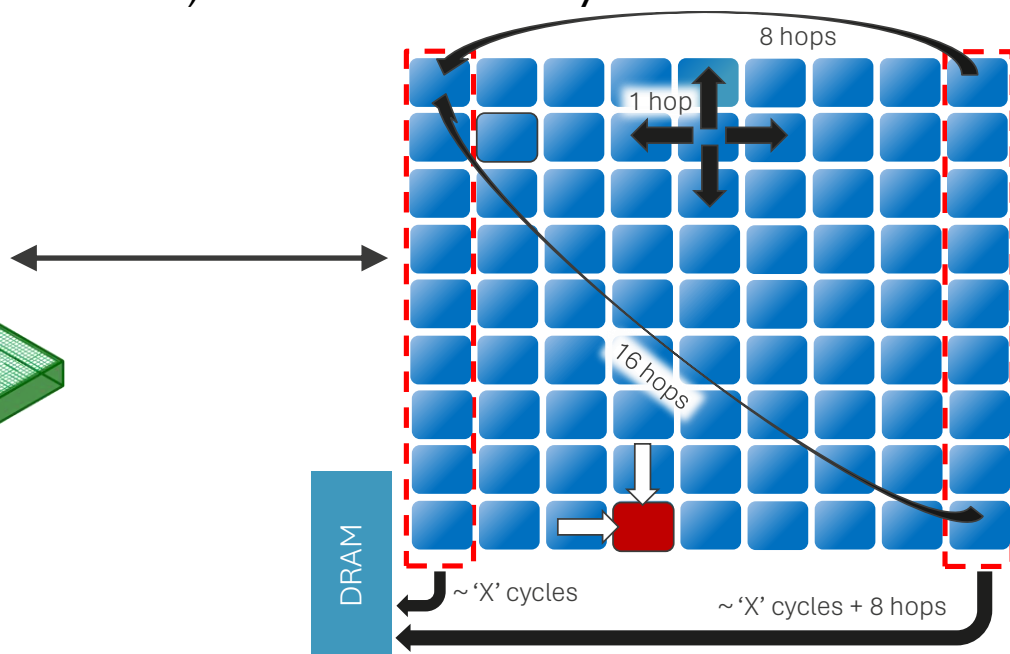
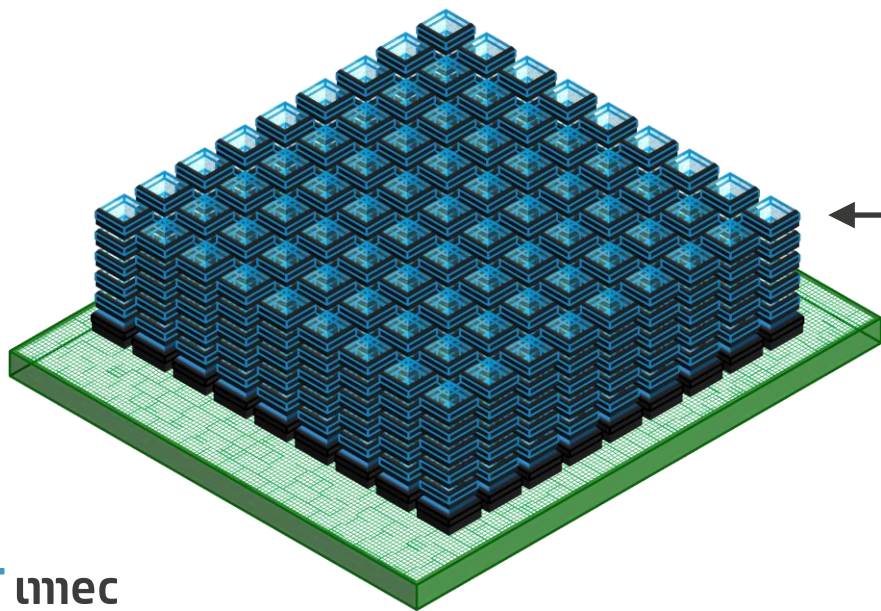
- Provisioning of JSRAM on-chip memory vs Main Memory BW



Scaling-Up

Superconducting Array of Arrays

- Shifting data to enable distributed matrix-multiply across the chip stack array is necessary: wrap around @edges $\rightarrow$ 2D torus topology
- Superconducting bumps ($\leq 30\mu\text{m}$ pitch, $10\mu\text{m}$ diameter) to facilitate inter-array communication



Scaling-Up

High Level Metrics

@ 30Gbps per for superconducting wires, TSVs & bumps →

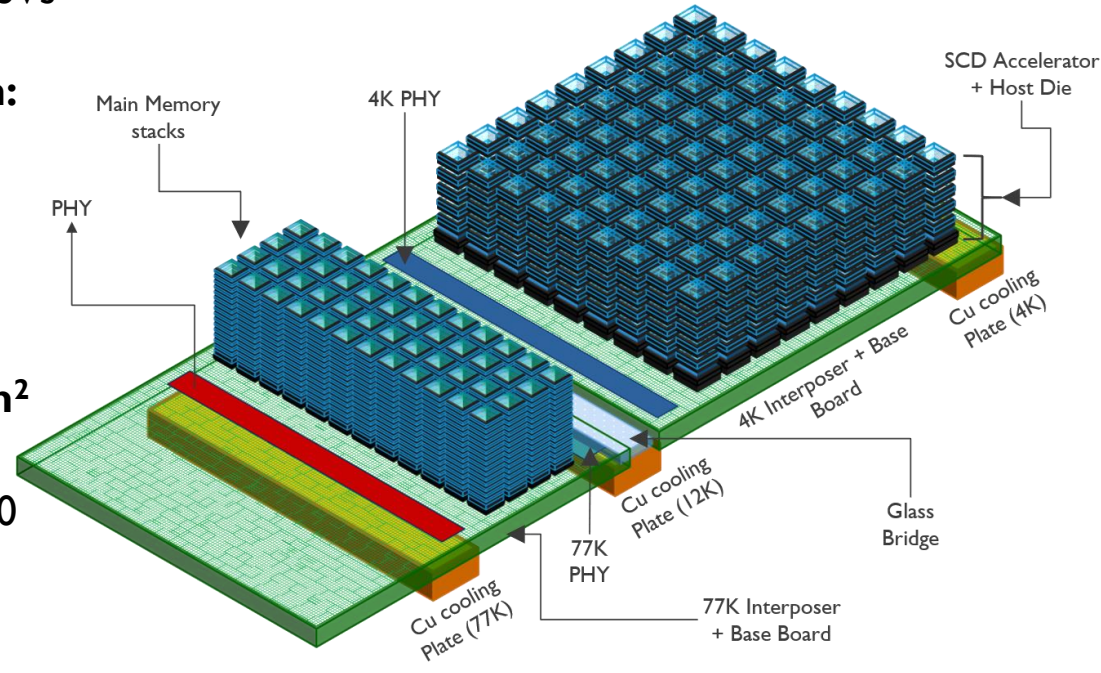
- **Intra-array & Inter-array bandwidth:**
~0.5TBps/mm²

@ JJ device dimension of 200nm, minimum metal pitch of 90nm & JSRAM density of ~4MB/cm² →

- **AI compute density: ~17TFlops/mm²**

@ Cooling efficiency of 325 (4K regime), 290 attojoules per JJ transition →

- **Energy efficiency: >50TOPs/Watt,**
(~100x vs conventional systems)



Conclusions & Future Outlook

Conclusions & Future Outlook

- Path for realizing highly efficient Superconducting Array of Arrays for acceleration of next generation AI/ML algorithms like Transformers
 - Sneak peek on scaling towards supercomputing/post-exascale clusters: “A Datacenter in a shoebox, IEEE Spectrum, June 2024”
- Towards high fidelity, calibrated and accurate simulations based on experimental data
- Connectivity to Main Memory (77K-regime), and Room Temperature links need to be investigated for a feasible appliance
- Utilizing open standards for ISA, and software stack to help in wider community adoption



mec

embracing a better life