IEEE 16th Workshop on Low Temperature Electronics WOLTE16

imec

Superconducting Array of Arrays for Acceleration of Transformers

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

IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024. Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.

OVERVIEW

You can add a subtitle if you want.

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

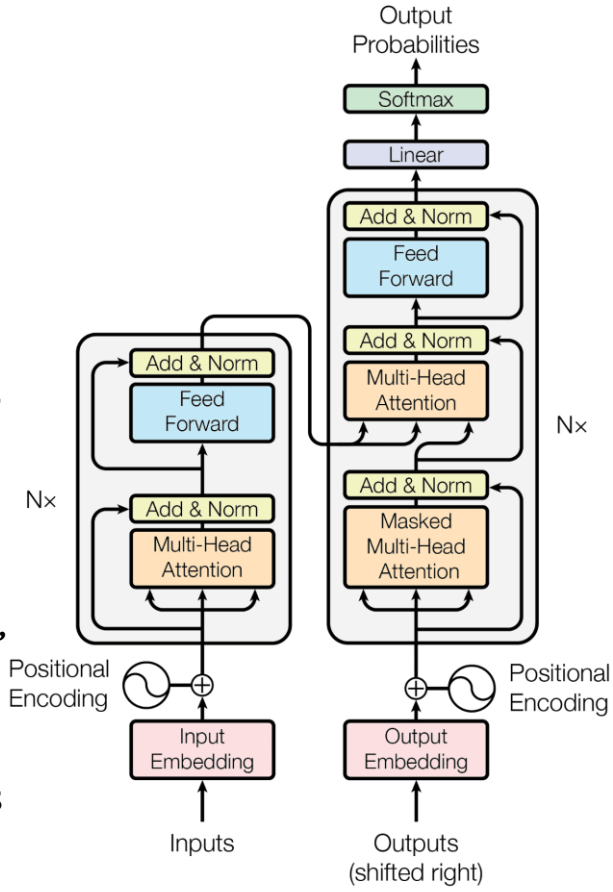
IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024. Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.

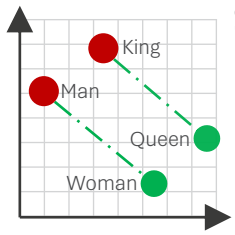
Introduction & Motivation

Introduction & Motivation

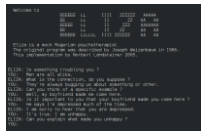
Why Transformers/LLMs

- 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





word2vec



1st chatbot

1966

2013

GPT and BERT

2018

GPT-3

2020

ChatGPT

2021

GPT-4

2022

Midjourney, CoPilot etc....

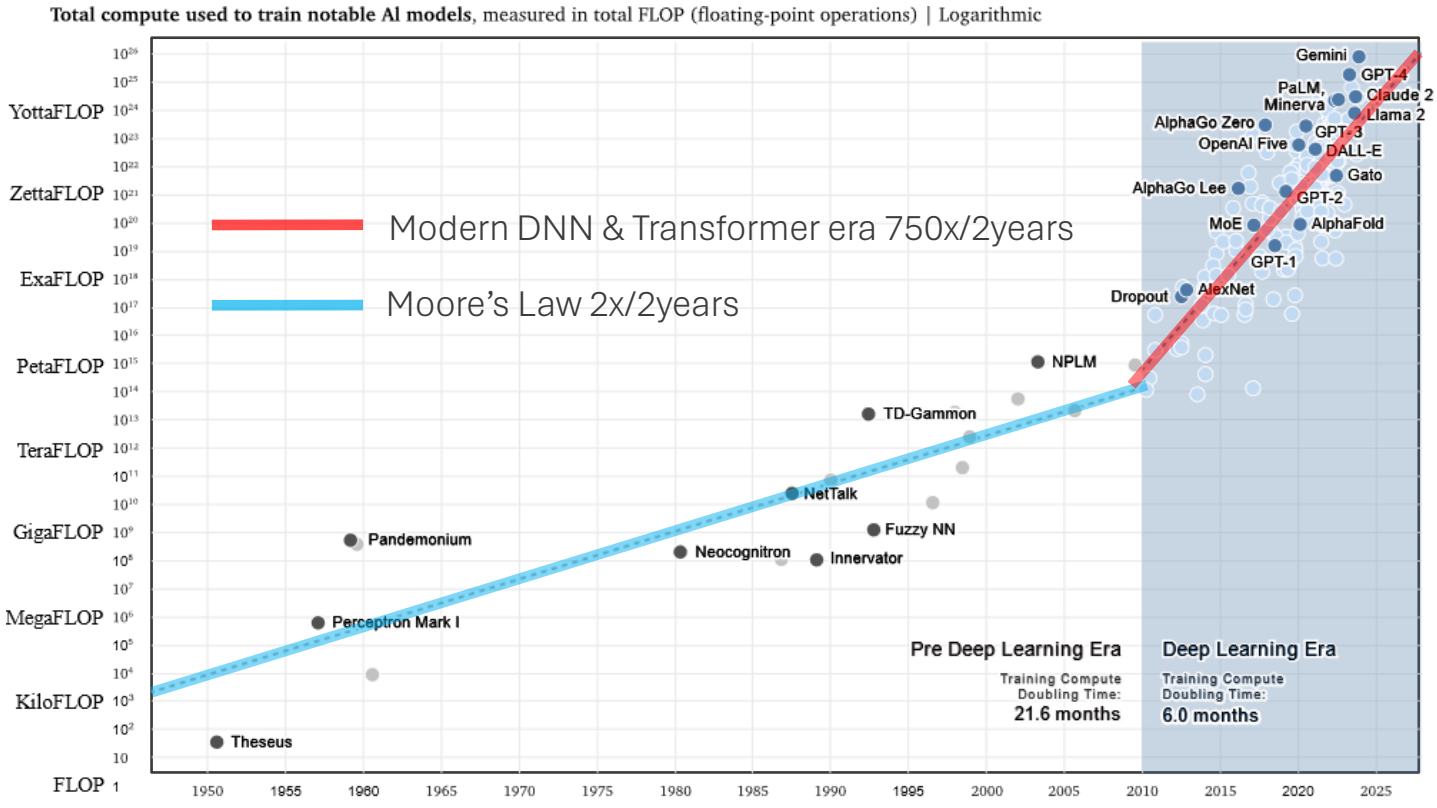
2023

imec

Introduction & Motivation

Cost of AI Training

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



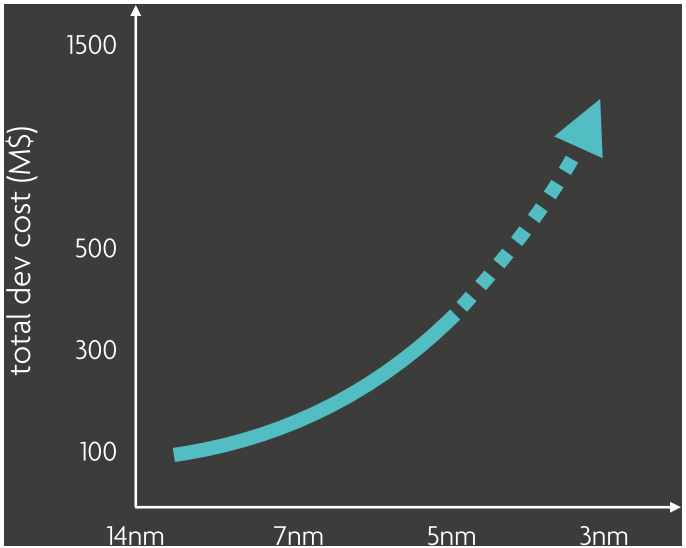
Computing Power and the Governance of Artificial Intelligence: https://doi.org/10.48550/arXiv.2402.08797

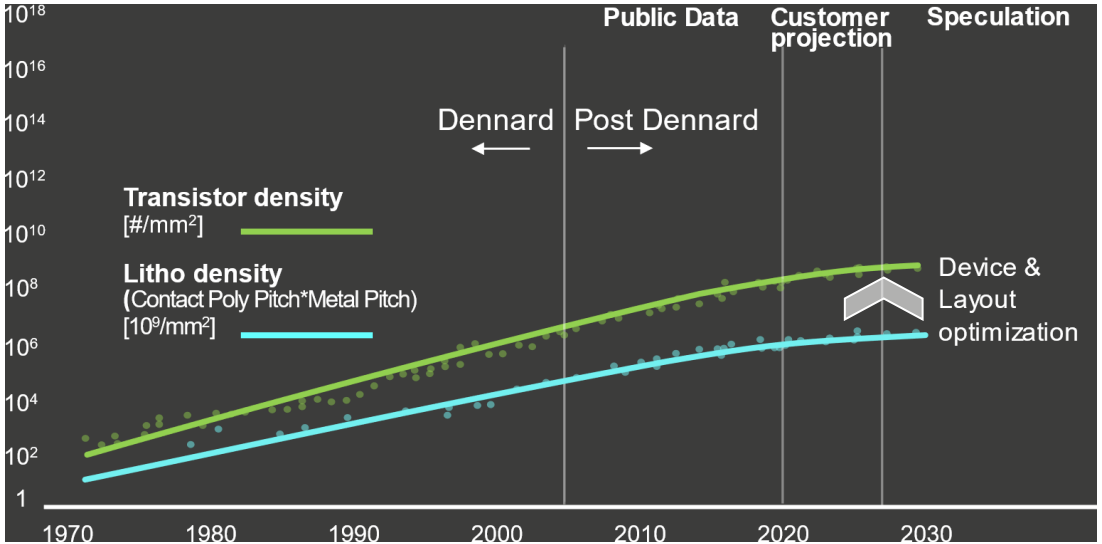
imec

Introduction & Motivation

Cost of AI Training

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





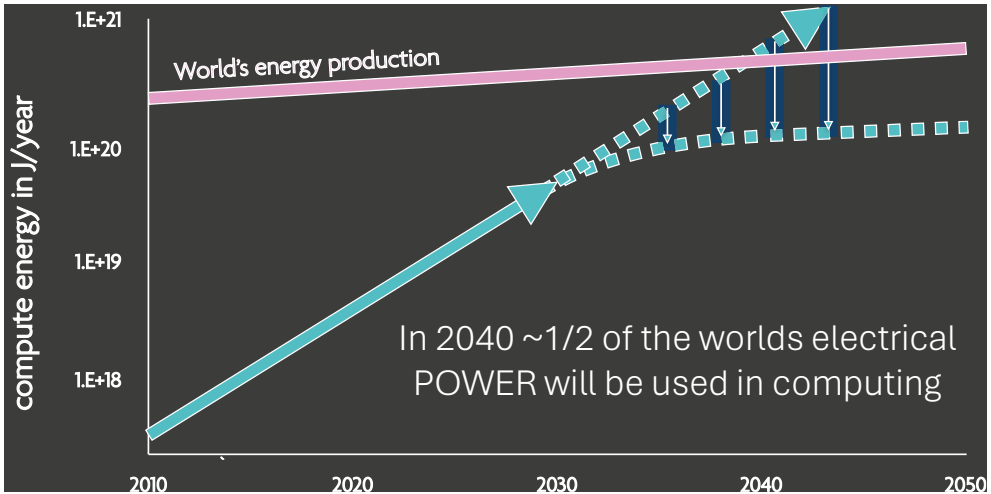
asml-investor-day-2021-technology-strategy---martin-van-den-brink.pdf

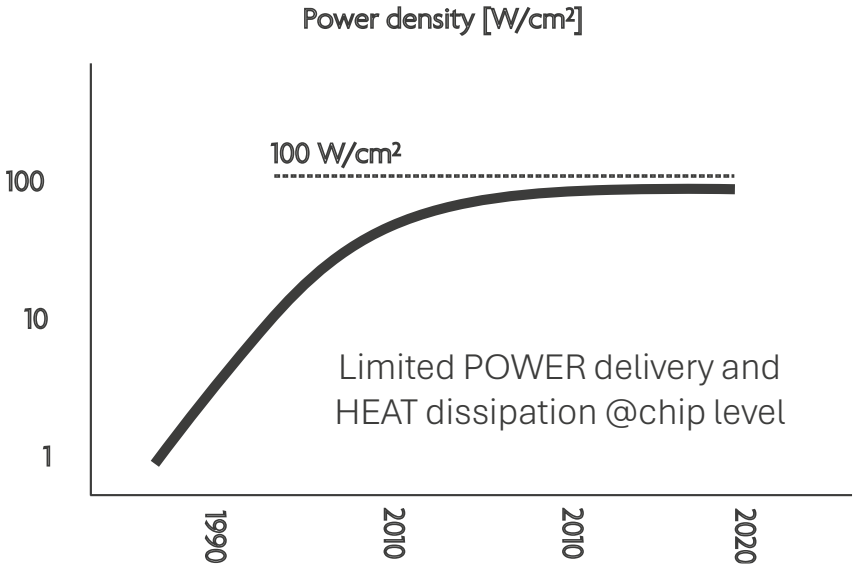
imec

Introduction & Motivation

Cost of AI Training

- Power distribution & thermal management cost (including Infra)

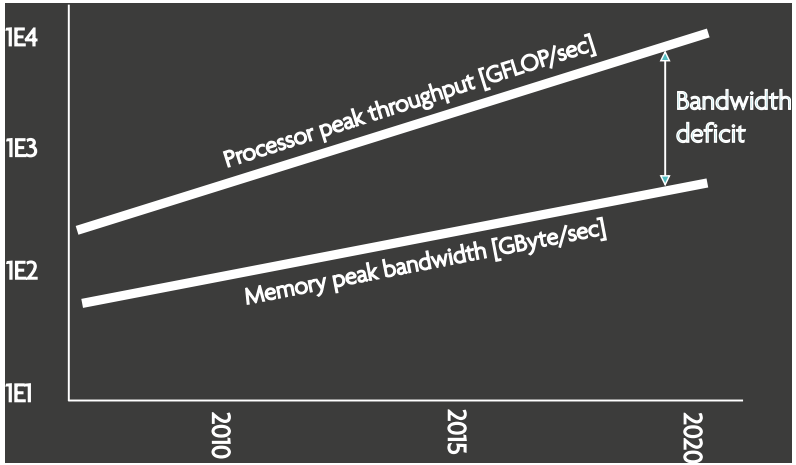


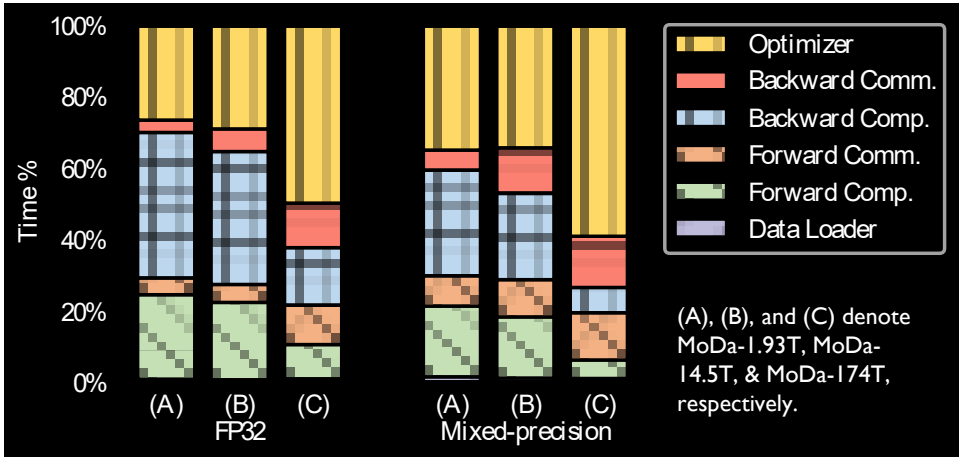


Introduction & Motivation

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

IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024. Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.

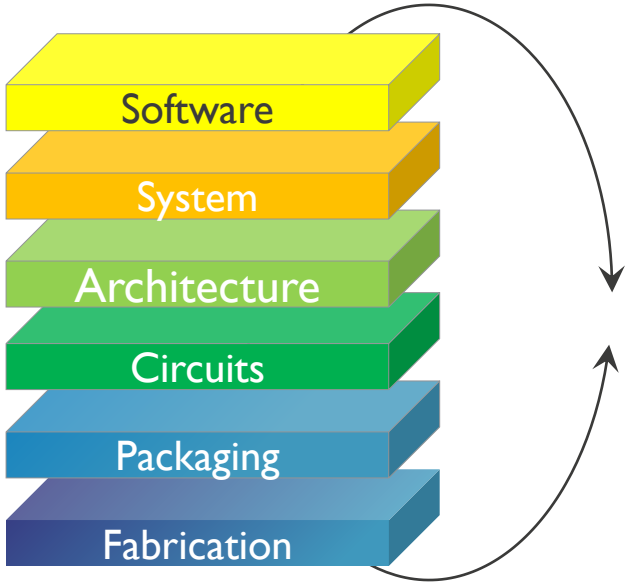
Accelerating LLMs via Superconducting Digital Systems

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

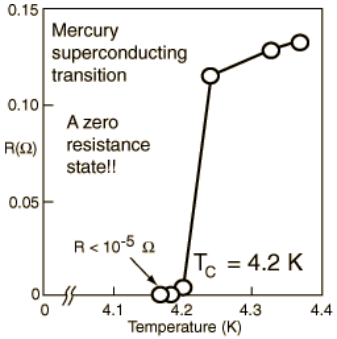


IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024. Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.

Accelerating LLMs via Superconducting Digital Systems

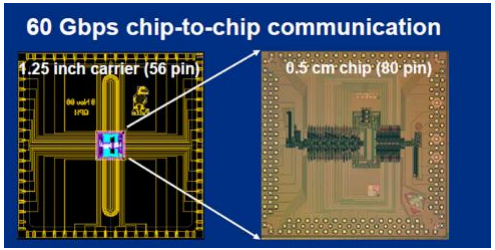
Superconducting technology

Zero resistance wires



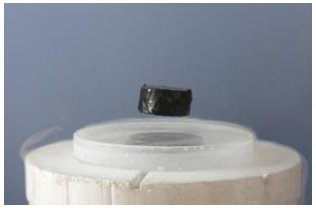
Enables THz bandwidth (optical like) interconnects that require no signal amplification

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

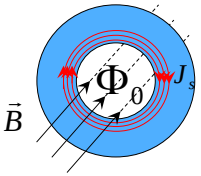


An 8-bit carry look-ahead adder with 150 ps latency and sub-microwatt power dissipation at 10 GHz; https://doi.org/10.1063/1.4776713

Meissner effect

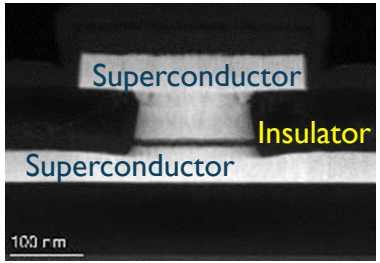


Allows for Quantum accurate digital bits (Fundamental limit for energy) with a large noise margin

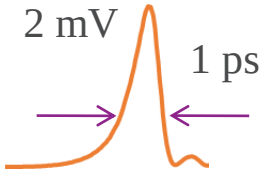


$$\Phi_0 = \frac{h}{2e} \approx 2\ \text{mApH}$$

Josephson effect



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



$E = 2\times10^{-20}$ J

imec

IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024.
Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.

Superconducting Pulse Conserving Logic and Josephson-SRAM; https://doi.org/10.1063/5.0148235

Scaling NbTiN-based ac-powered Josephson digital to 400M devices/cm²; https://doi.org/10.48550/arXiv.2303.16792

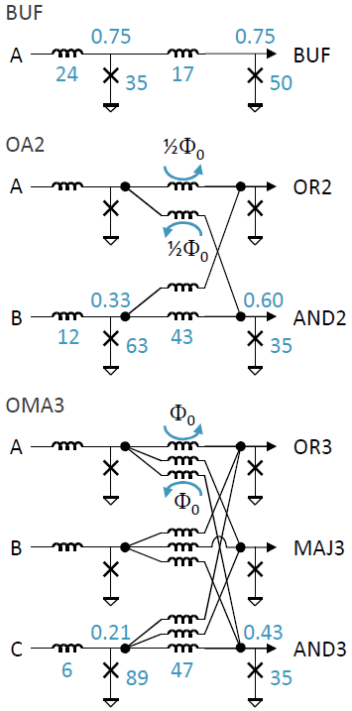
Accelerating LLMs

Superconducting digital logic and memory

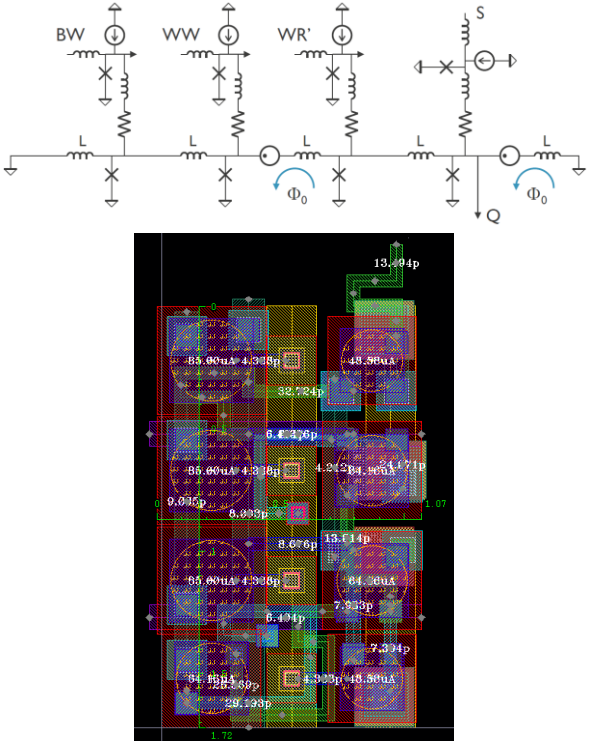
- Superconducting digital Pulse-Conserving Logic (PCL), Josephson SRAM (JSRAM) & AC power distribution via resonant networks allow for classical digital systems that are scalable

- 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



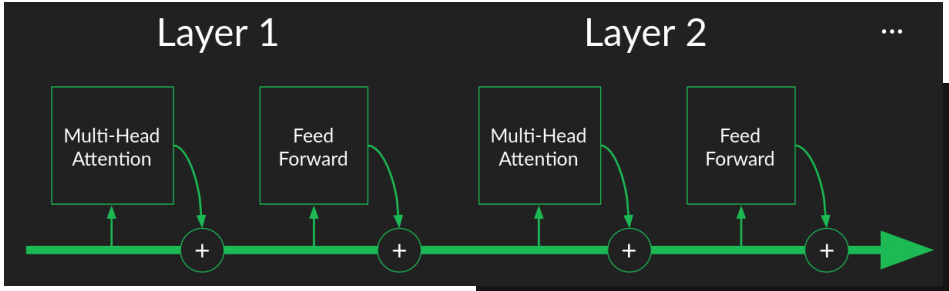
JSRAM unit cell schematics & Layout

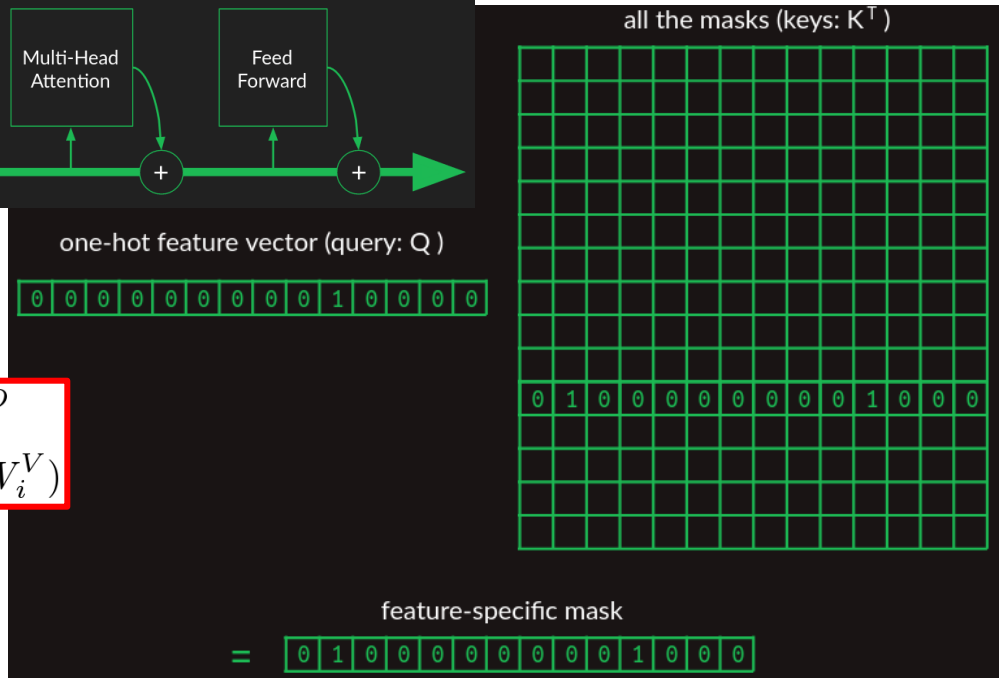


imec

Accelerating LLMs

Workload





$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

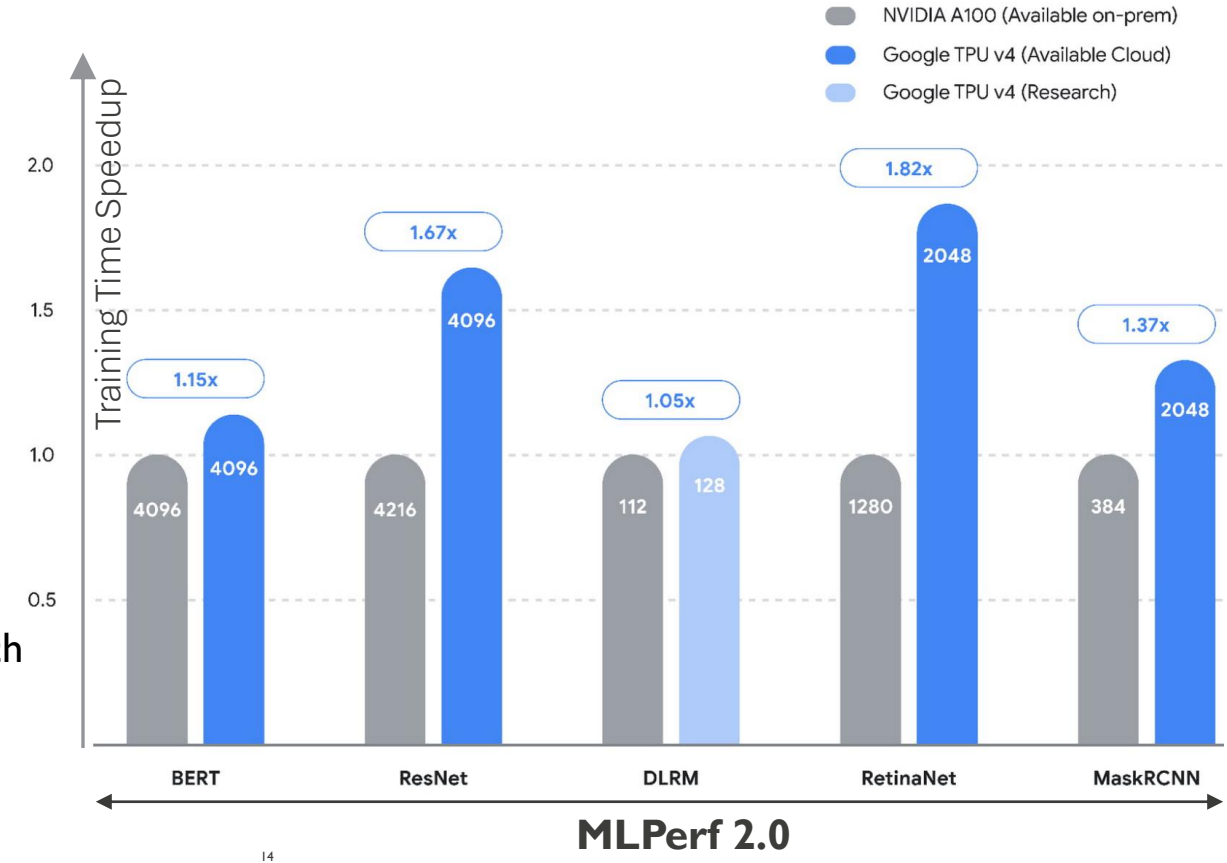
$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- LLM training (& inference) workloads are mainly composed of Matrix-Multiplication kernels

Accelerating LLMs

Architecture

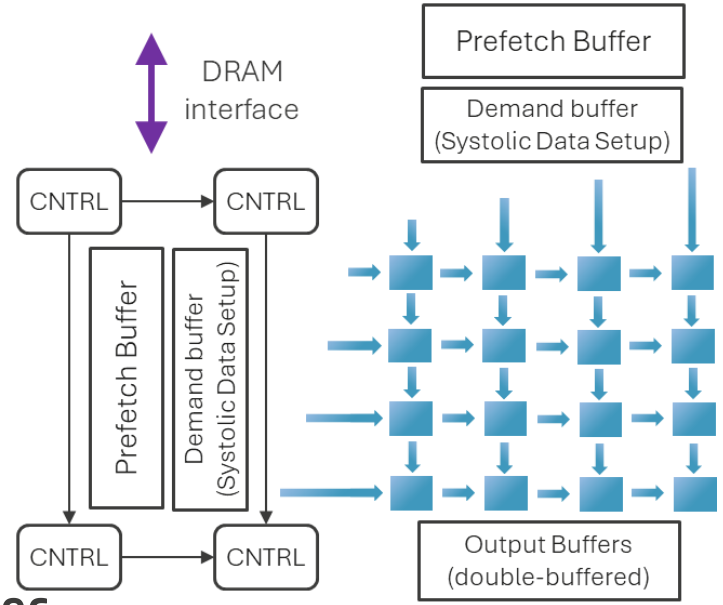
- Array architectures (*like systolic arrays*) are best suited to efficiently execute MMM kernels (*>>FLOPS/chip due to >>MACs*)
- They also have minimal control overhead → reduces design complexity, NRE & cost of adoption for superconducting tech



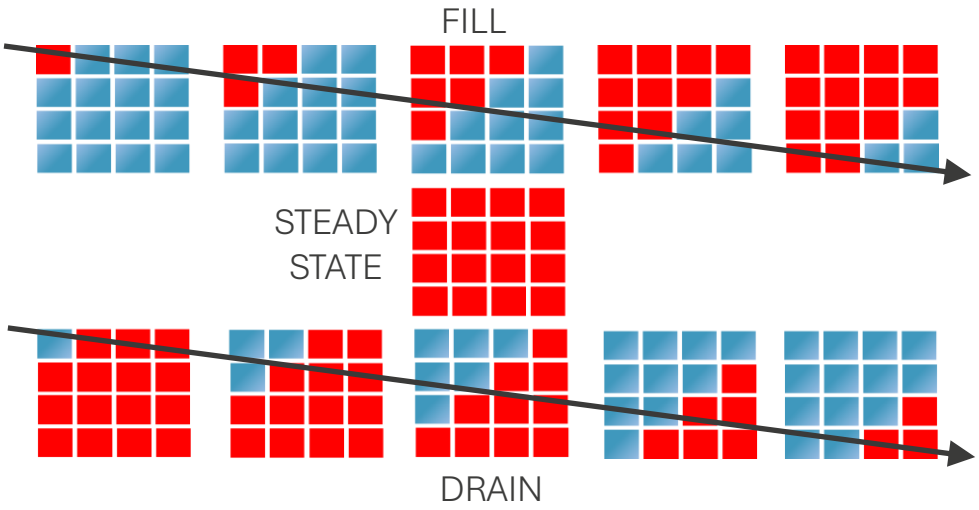
Accelerating LLMs

Architecture

- Control units to buffer dataflow from DRAM to MAC units

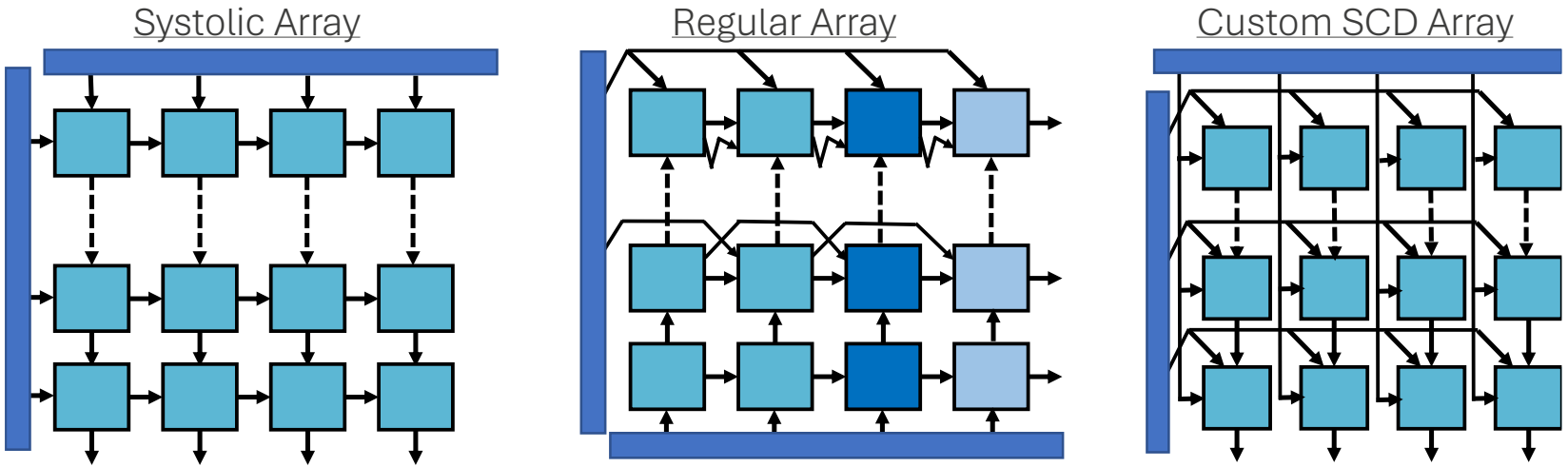


- However, simple systolic arrays incur pipeline bubbles during "FILL" & "DRAIN"



Accelerating LLMs

Array Architecture



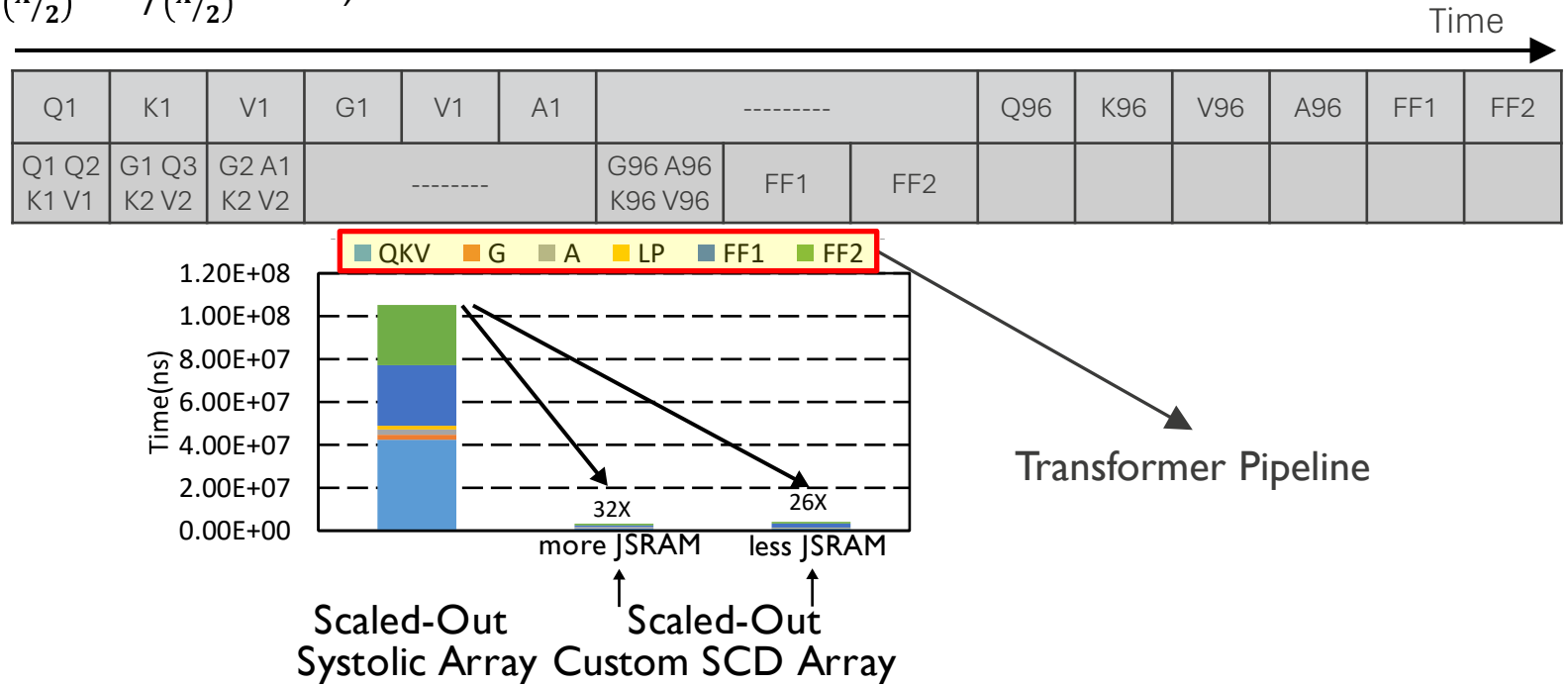
- Proposed Custom Superconducting Array is a homogenous version of regular arrays with superconducting wires.
- While this comes at the cost of more wires, the minimal wiring overhead (Power & Performance) incurred in superconducting digital technology makes this an ideal solution

Scaling the system

Scaling-Up

Array Architecture

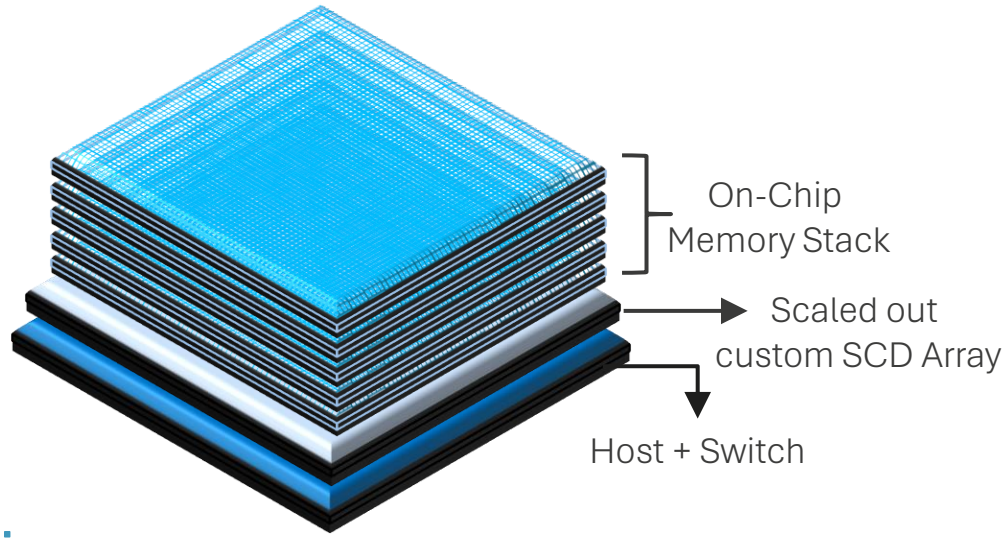
- Scale-up (*Single Large PE array consisting of '**N** × **N**' MACs*) vs Scale-out (*'**X**' PE arrays consisting of '$^{N}/_{(X/2)}$ × $^{N}/_{(X/2)}$' MACs*)



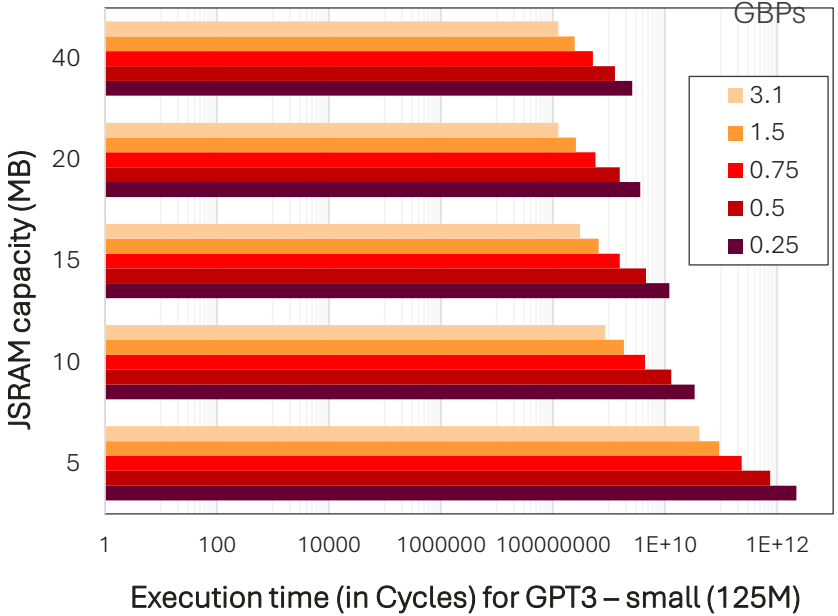
Scaling-Up

Superconducting Processing Stack

- Superconducting processing stack: multifunctional die stack 3D integrated via superconducting TSVs (*≤30µm pitch, 10µm diameter*)



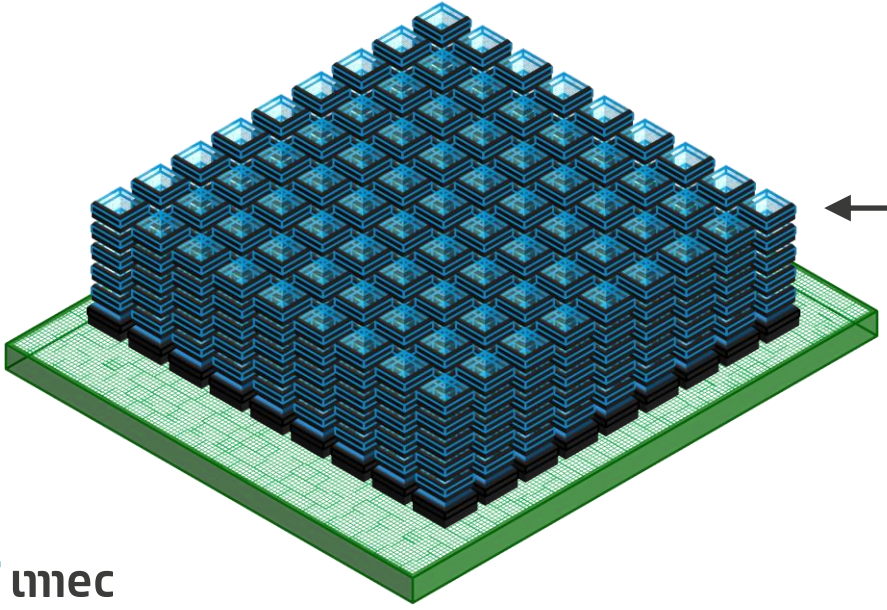
- Provisioning of JSRAM on-chip memory vs Main Memory BW

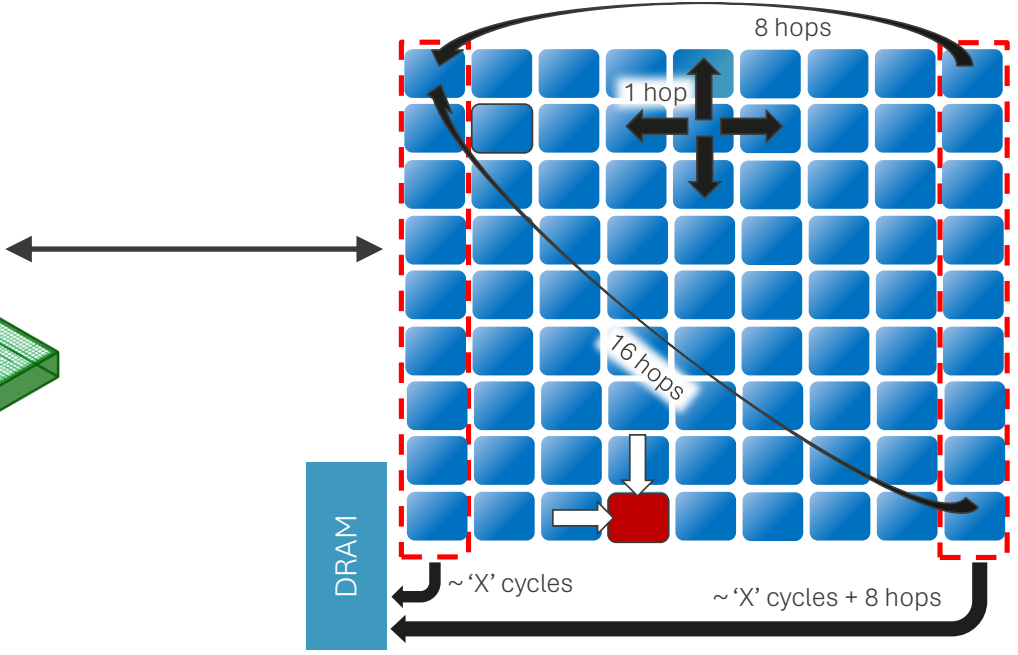


Scaling-Up

Superconducting Array of Arrays

- Shifting data to enable distributed matrix-multiply across the chip stack array is necessary: wrap around @edges → 2D torus topology
- Superconducting bumps (≤*30µm pitch, 10µm diameter*) to facilitate inter-array communication





Scaling-Up

High Level Metrics

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

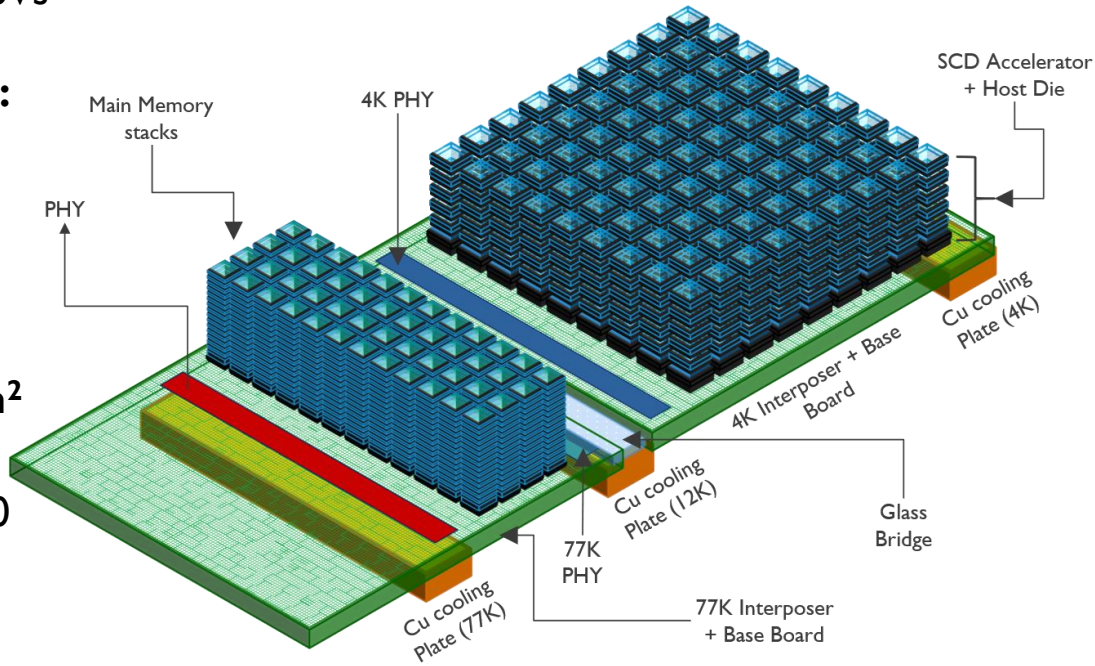
- **Intra-array & Inter-array bandwidth: ~0.5TBps/mm²**

@ 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)***



imec

IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024. Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.

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

IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 56, Sept 2024. Presentation given at WOLTE-16 2024, June 2024, Cagliari, Italy.

imec

embracing a better life