Hierarchical Modeling and Exploration of Large-Scale Superconducting Supercomputer Architectures

Andreas Gerstlauer, George Biros, Lizy John, Jaydeep Kulkarni
The University of Texas at Austin

Lingda Li, Adolfy Hoise
Brookhaven National Lab

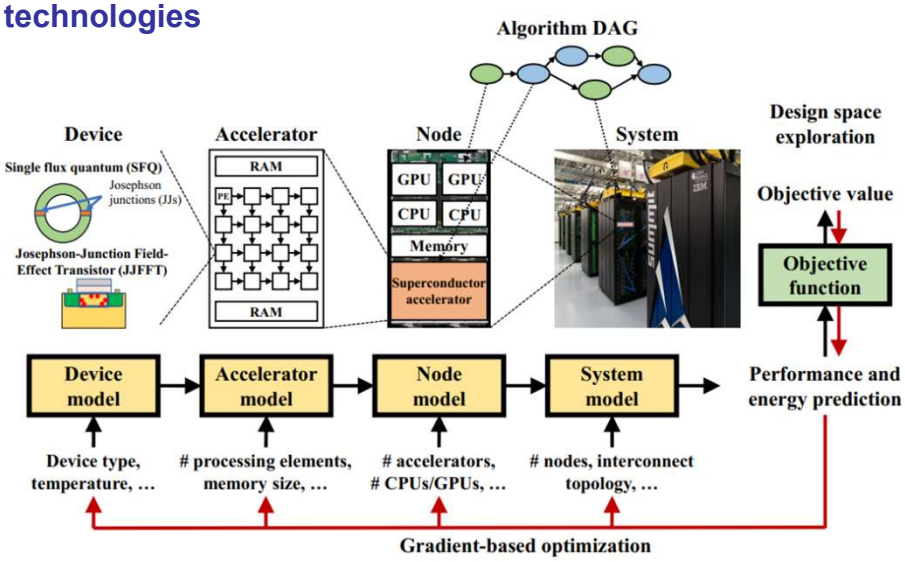
Sponsored by DOE EXPRESS

Background

- **Limits of traditional CMOS technology**
  - Increasingly challenging to address future high-performance computing (HPC) needs
  - Beyond-CMOS technology is promising
  - ➢ Superconducting digital (SCD) computing
- **Lack of modeling and simulation technology for beyond-CMOS HPC systems**
  - Advance understanding of technology landscape
  - ➢ How much system-wide benefit can latest technology deliver in practice?
  - ➢ How sensitive are these benefits to technology assumptions?
  - ➢ How do technology improvements impact architectural and application-level design decisions?

Objective

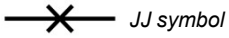
- **Modular & hierarchical modeling framework to explore and optimize system-level impacts of beyond-CMOS technologies**
  - Analytical modeling & simulation
  - Across abstraction levels
  - AI-enabled modeling & exploration
- **Focus on**
  - Superconducting technology
  - Compute-intensive applications (GEMM, LLM, …)



Superconducting Digital (SCD) Technology

- **Long carried a lot of promise, but scaling challenges**
  - Josephson Junction (JJ), zero-resistance
  - Ultra-low energy and ultra-high speed
  - But: cooling, low density, lack of memory

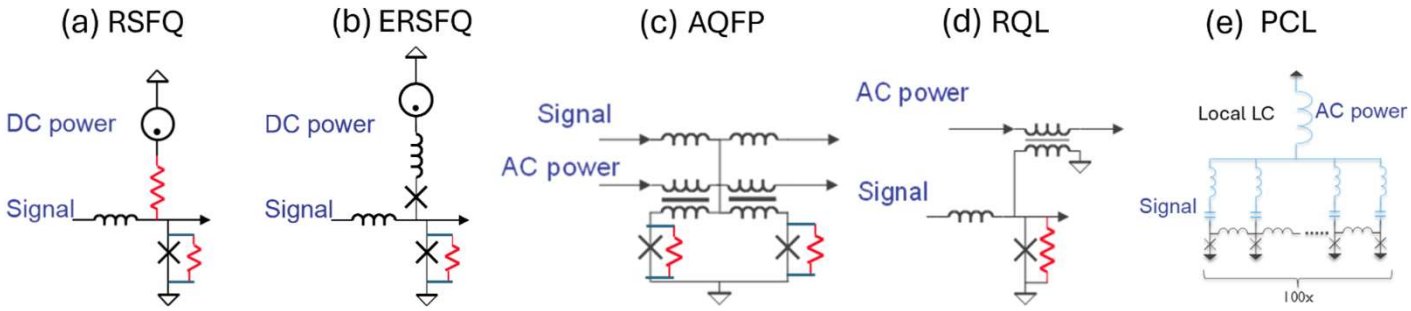
	CMOS (300K)	Cryo-CMOS (77K)	SCD (4K)
Density	Very high	Very high	Low
Energy	fJ	fJ	aJ
Speed	4GHz	6GHz	50GHz
Leakage	High	Low	Very low
Cooling	0 W/W	22 W/W	325 W/W



- **Recent technological SCD advances**
  - Novel materials and power delivery
  - High density and low overhead
  - True static random-access memory (JJ-SRAM)

➢ **Time for a fresh look at superconducting technology for future supercomputers?**
  - Esp. given ever-increasing demands for AI/ML datacenters

SCD Technology Evolution

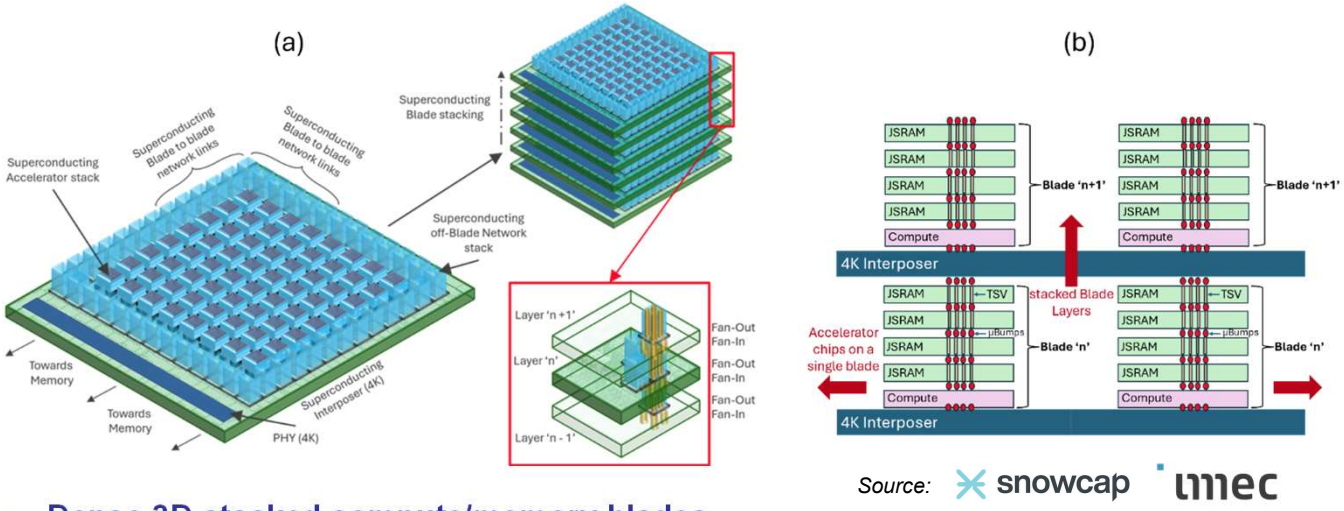


(a) RSFQ	(b) ERSFQ	(c) AQFP	(d) RQL	(e) PCL
High frequency High static power Low scalability	Low static power High dynamic energy Low density (inductors)	Zero static power Shallow (1 level) logic Low density (transformers)	Combinational logic Setup time overhead Low density (transf.)	NbTiN material Dual-rail logic High density (capacit.) Fast, low energy

(Shift-)Register based memory only, low bandwidth, high overhead

JJ SRAM
Fast interconnect

SCD-Based 3D-Stacked Architecture



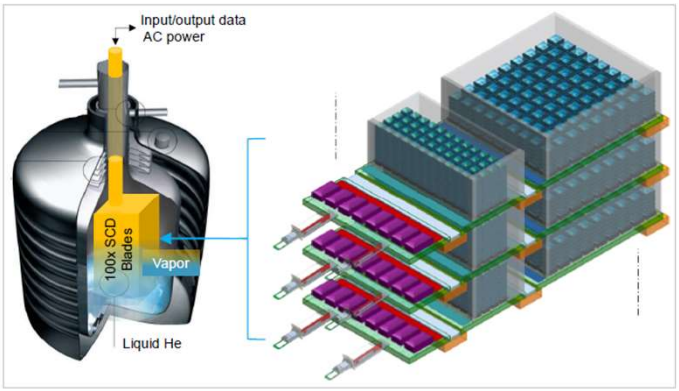
Source: snowcap imec

- **Dense 3D stacked compute/memory blades**
  - Array of SCD compute/SRAM stacks (4K) + cryo-HBM DRAM per blade (77K)
  - Local through silicon via (TSV) chip-to-chip and blade-to-blade interconnect

SCD Cooling

- **Modern cryocoolers**
  - Helium liquification
  - Up to 2.5KW cold (4K)
- **Cooling efficiency**
  - Improves with scale/size
  - 320 W/W at large scales
- **Supercomputer in a shoebox**
  - 100 stacked SCD+DRAM blades

Source: Linde



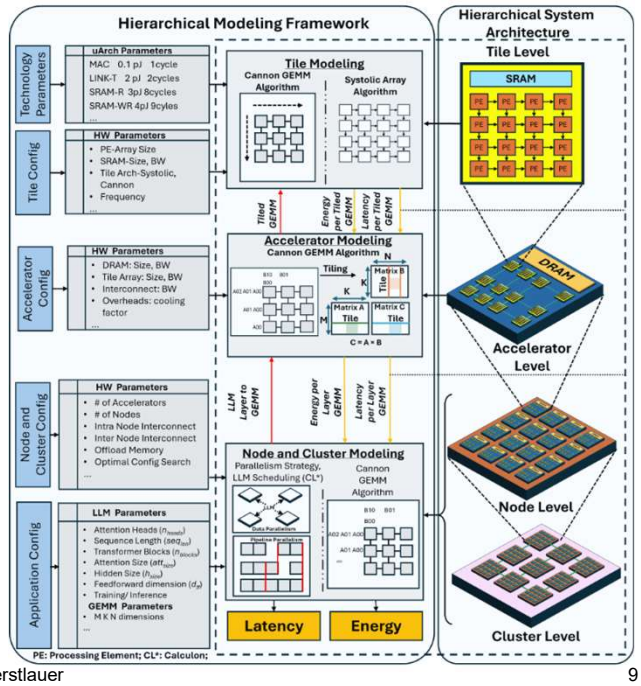
Source: snowcap imec

Outline

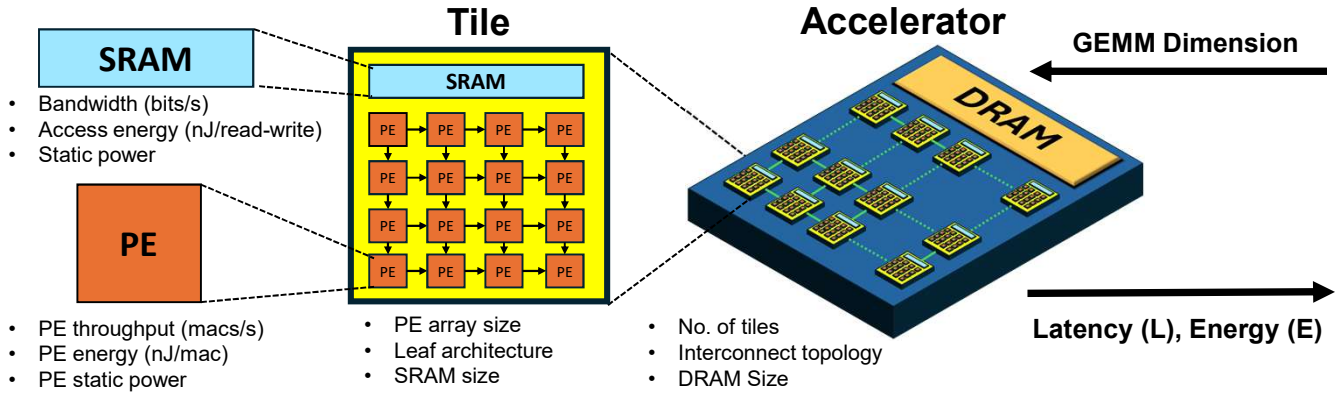
- ✓ **Introduction**
- ✓ **SCD technology background**
- **Modeling framework**
- **Experiments & results**
- **Summary & Conclusions**

Hierarchical Modeling Framework

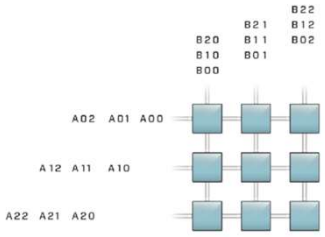
- **Semi-analytical models**
  - Latency and energy estimation
- **Processing element (PE)**
  - Technology characterization
  - Operation-level latency/energy
- **Tile**
  - Cannon GEMM or systolic [ScaleSim]
- **Accelerator**
  - Analytical Cannon GEMM model
- **Node/cluster**
  - Cannon or LLM [Calculon]



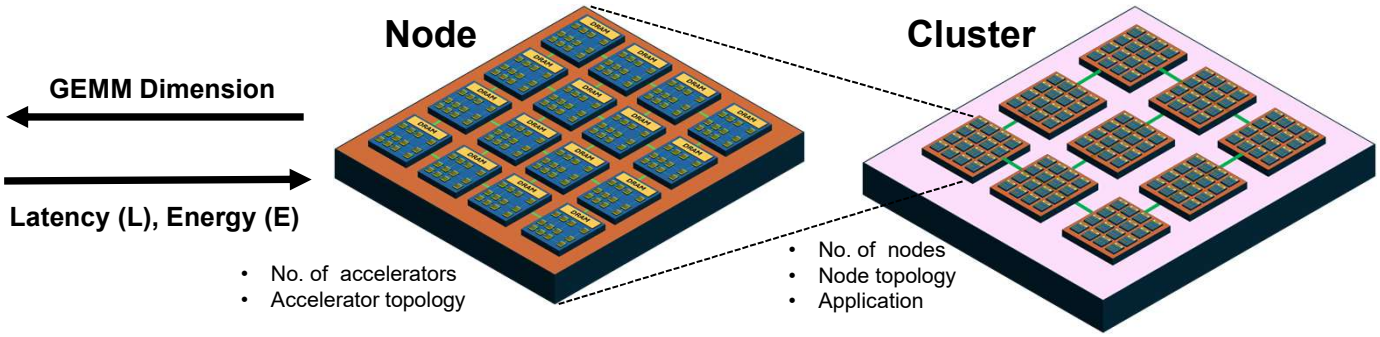
Tile & Accelerator Level Modeling



- **GEMM operations**
  - Recursive application of Cannon-GEMM / systolic mapping
  - Temporal and spatial tiling of a given GEMM loaded into DRAM
  - Maximally exploit local SRAM buffering



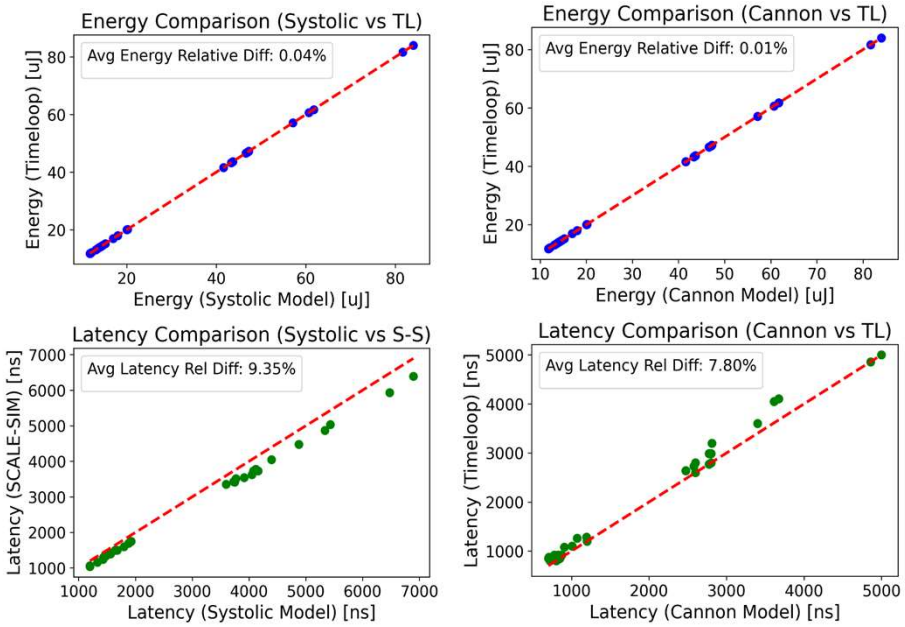
Node & Cluster Level Modeling



- **Cannon-GEMM applications**
  - Recursive application of tile/accelerator-level Cannon-GEMM
- **Transformer-based large language model (LLM) applications**
  - Optimized LLM mapping & parallelization into GEMM kernels using [Calculon]
  - Data parallelism (DP), tensor parallelism (TP), pipeline parallelism (PP)
  - Added custom analytical energy model to Calculon's access statistics

Model Validation

- **Cluster/node**
  - Calculon validated against GPU systems
  - We replace GPU model w/ our GEMM
- **Accelerator/tile**
  - Validation of our Canoon and systolic GEMM models
  - Against SoTA AI/ML models [Timeloop/TL] & [Scale-Sim/SS]



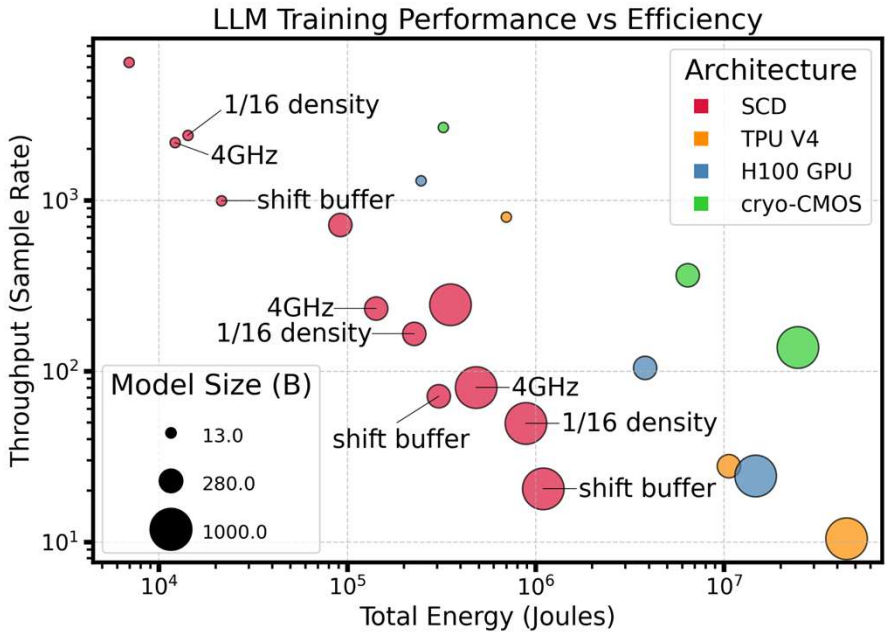
Experimental Setup

- **Systems (iso-area)**
  - SCD: superconducting system using imec technology at 4K
  - H100, TPUv4: 2 state-of-the-art CMOS systems at room temperature
  - Cryo-CMOS: 1 CMOS system at 77K with similar architecture as SCD

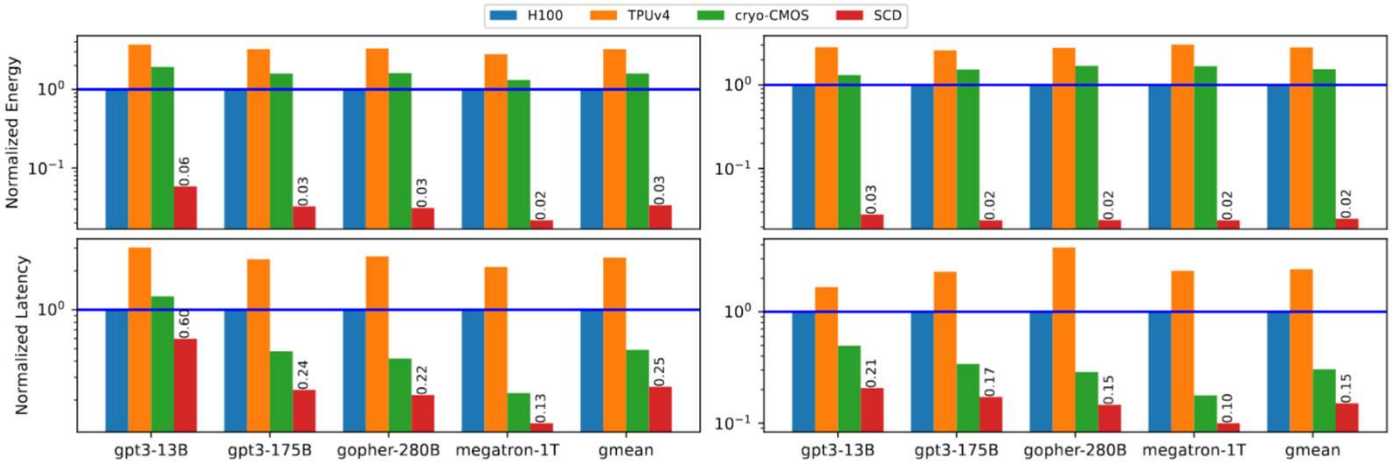
	Accelerator Count	Accelerator Configuration	Tech Node	Die Area	3D Stacking	Frequency	TOPs	DRAM Bandwidth	DRAM Capacity	Interconnect
h100	768	tensorCores with 50MB SRAM	4nm	$814mm^2$	-	1.83GHz	495	3.4TBps	80GB	Nvlink and NvSwitch
tpuv4	1024	2 tensorCores, each TC has 4 128x128PE arrays, 128MB SRAM	7nm	$600mm^2$	-	1.05GHz	137	1.2TBps	32GB	Optical switches
cryo-CMOS	64	64 chips, each chip has 1 500x500 PE arrays and 256MB SRAM	7nm	$9216mm^2$	2 layers	4GHz	67109	30TBps	1024GB	Cryo-CMOS switches
SCD	64	64 chips, each chip has 2 200x200 PE arrays and 16MB JSRAM	28nm	$9216mm^2$	6 layers	30GHz	76800	30TBps	1024GB	SCD switches

- **Workloads**
  - LLM: GPT3-13B, GPT3-175B, Gopher-280B, Megatron-1T

Overall Results



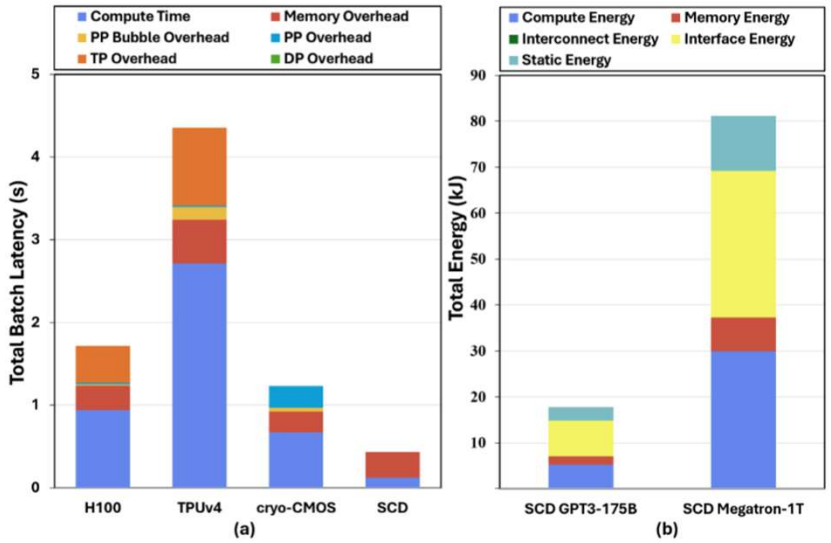
Q1: System-Wide Benefits of SCD?



LLM Inference LLM Training

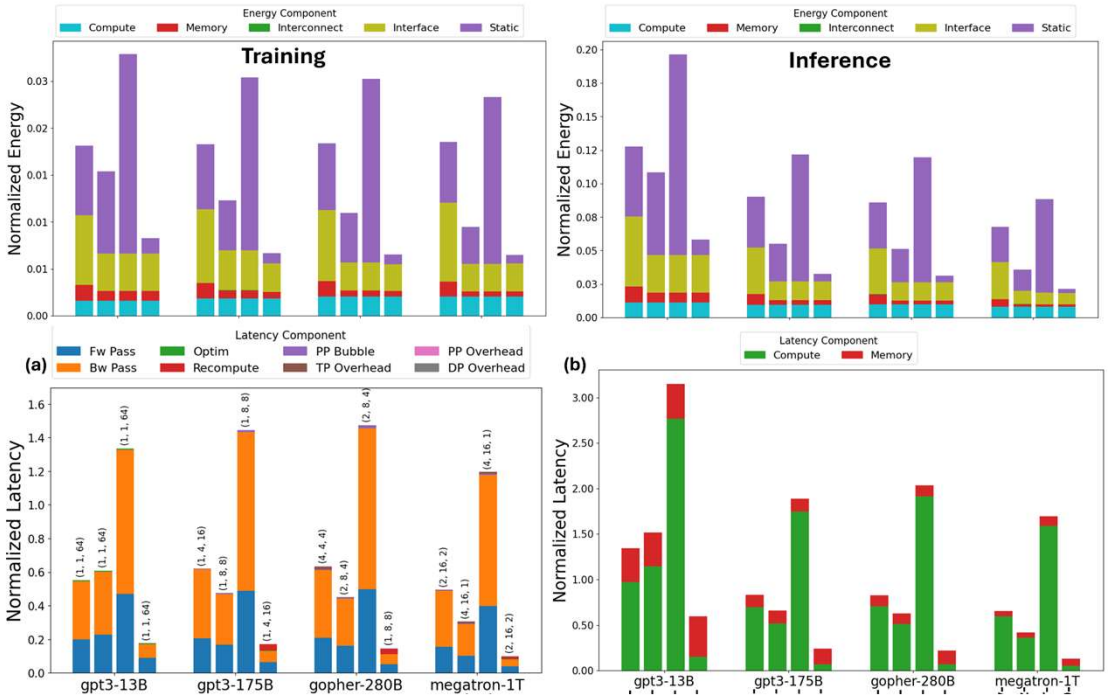
Q2: Where do Benefits Come From?

- **High compute capability**
- **High interconnect bandwidth**
- **Ultra-efficient SCD PE operations & free data movement**



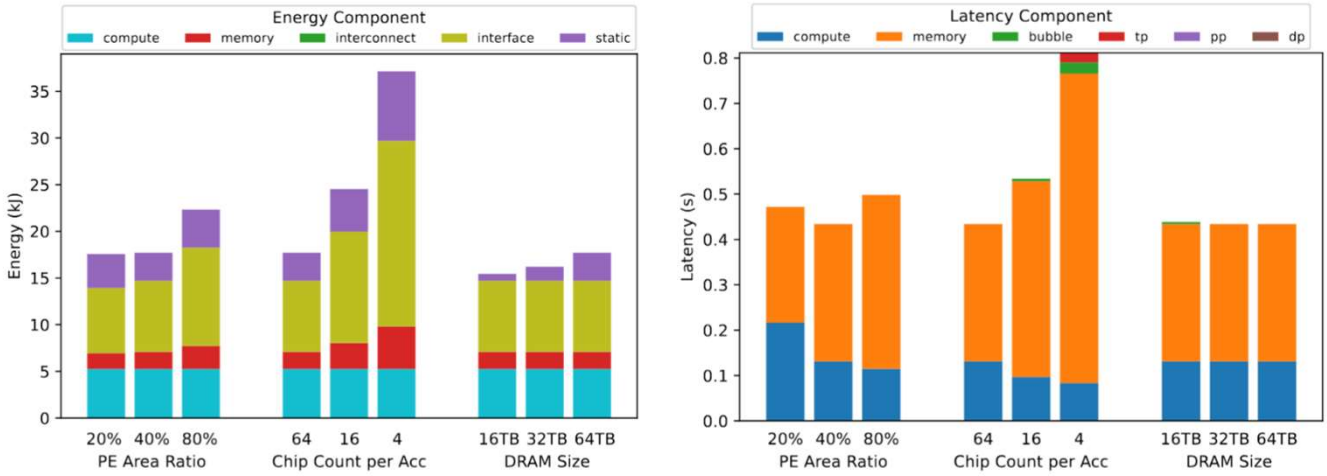
Q3: Sensitivity to SCD Assumptions?

- **SCD ablation**
  - 1/16 dens.
  - 4GHz freq.
  - Shift mem.
  - Full SCD
- **Contributors**
  - SRAM BW > dens. > freq.



Q4: Architecture Impact of SCD Technology?

- **Design space exploration**
  - PE vs. SRAM area tradeoff
  - SC interconnect allows efficient scaling, smaller DRAM size slightly reduce energy/lat.



Summary & Conclusions

- **Emerging technology for supercomputing**
  - Traditional CMOS scaling hitting physical limits
  - Recent superconducting digital (SCD) technology advancements
  - True system-level impact of technology assumptions, limitations & unknowns?
- **Hierarchical modeling framework**
  - Semi-analytical modeling across abstraction levels
  - From devices to clusters
  - On-going work on ML-enabled (gradient-based) design space exploration
- **SCD carries a lot promise, but open questions remain**
  - Significant potential energy (despite cooling) & performance benefits
  - Scalability, reliability, …? (Recent startups exploring the field)