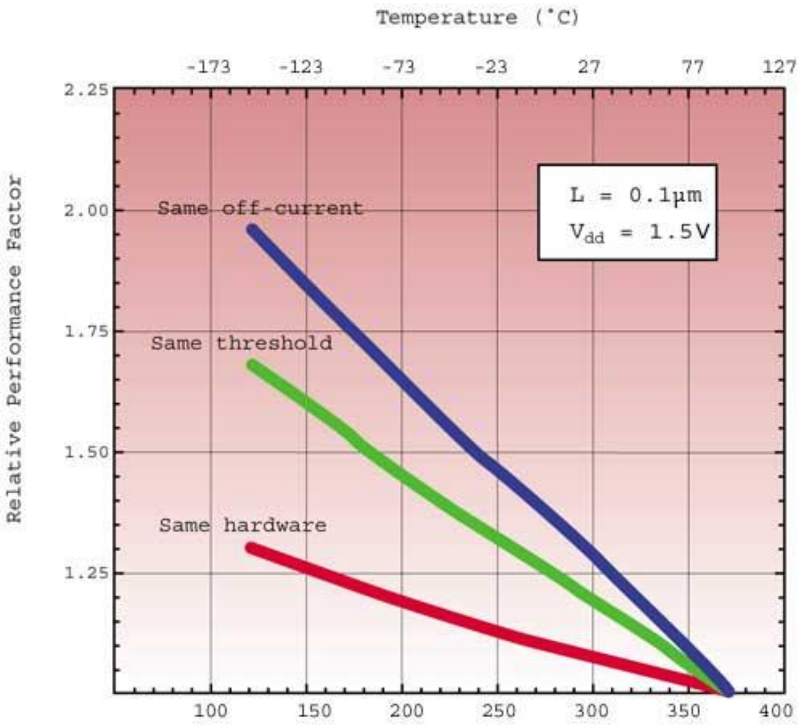
Implications of Full-System Modeling for Superconducting Architectures

Kunal Pai, Mahyar Samani, Anusheel Nand & Jason Lowe-Power

DArchR
DAVIS ARCHITECTURE RESEARCH

UCDAVIS
COMPUTER SCIENCE

Introduction

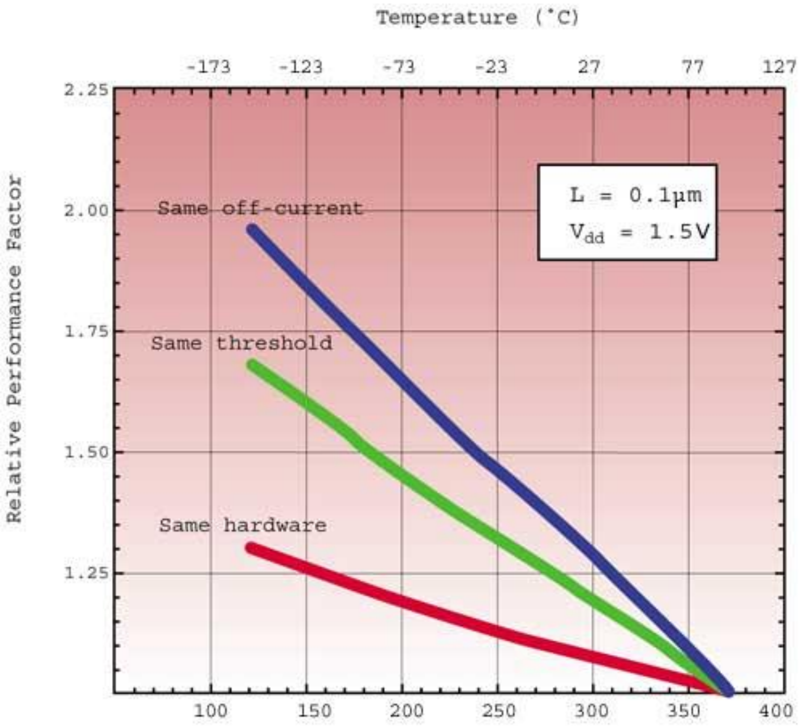


source: https://www.electronics-cooling.com/2001/08/the-challenge-of-operating-computers-at-ultra-low-temperatures/

CMOS -> high leakage currents, *reduced* perf. at high temp

CryoCMOS and superconductors -> *low temp., high perf., high energy efficiency*

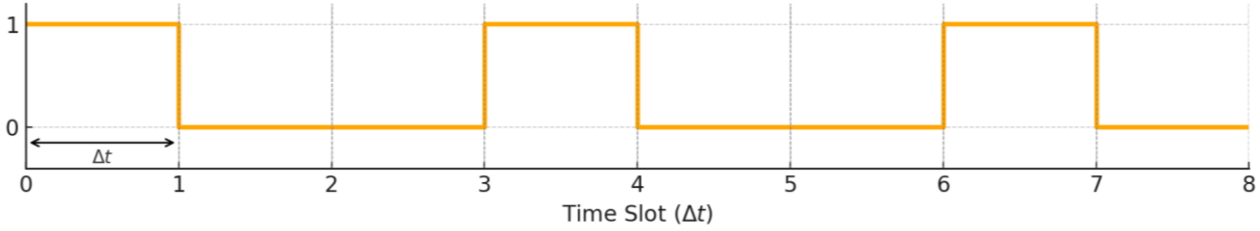
Introduction



source: https://www.electronics-cooling.com/2001/08/the-challenge-of-operating-computers-at-ultra-low-temperatures/

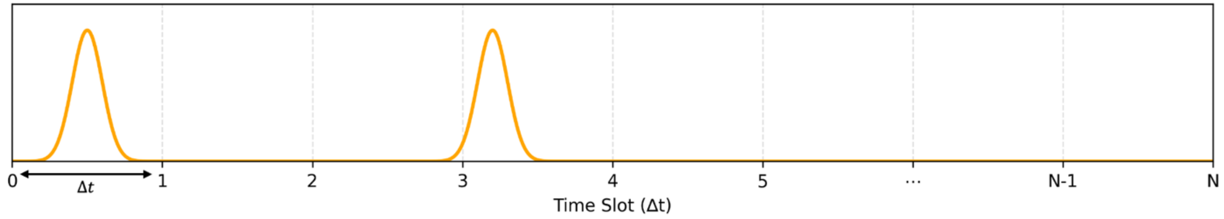
CMOS -> high leakage currents, *reduced* perf. at high temp

CryoCMOS and superconductors -> *low temp., high perf., high energy efficiency*



Cryogenic CMOS: 123 K, 4 GHz clk.

- Same logic as regular CMOS



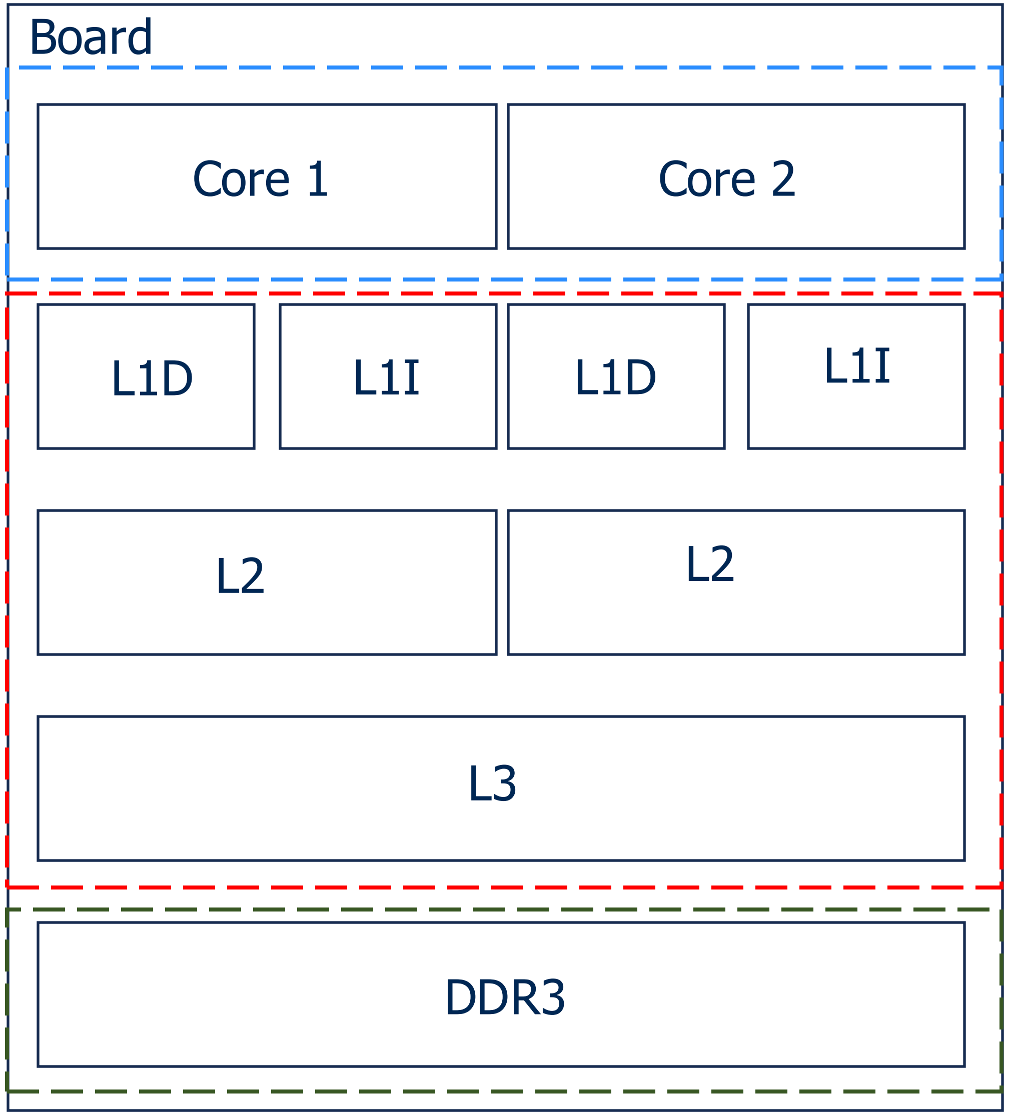
Superconducting electronics: 10 K, max. 100 GHz clk.

- Logic based on detection of pulse at time steps (race logic)

Contributions

- ***First*** full-system study on CryoCMOS & Superconductors
  - gem5: cycle-level simulator @ v23.1
  - Diverse workloads: SPEC CPU2006 (ref), BFS, PR, CC
  - Theoretical and realistic architectures

Theoretical Super- & Cryo- Architecture Modeling



Cryo (4 GHz) /
Super (100 GHz)

OOO (BOOM) /
In-order (HiFive Unmatched)

Cryo (4 GHz) /
Super (100 GHz)

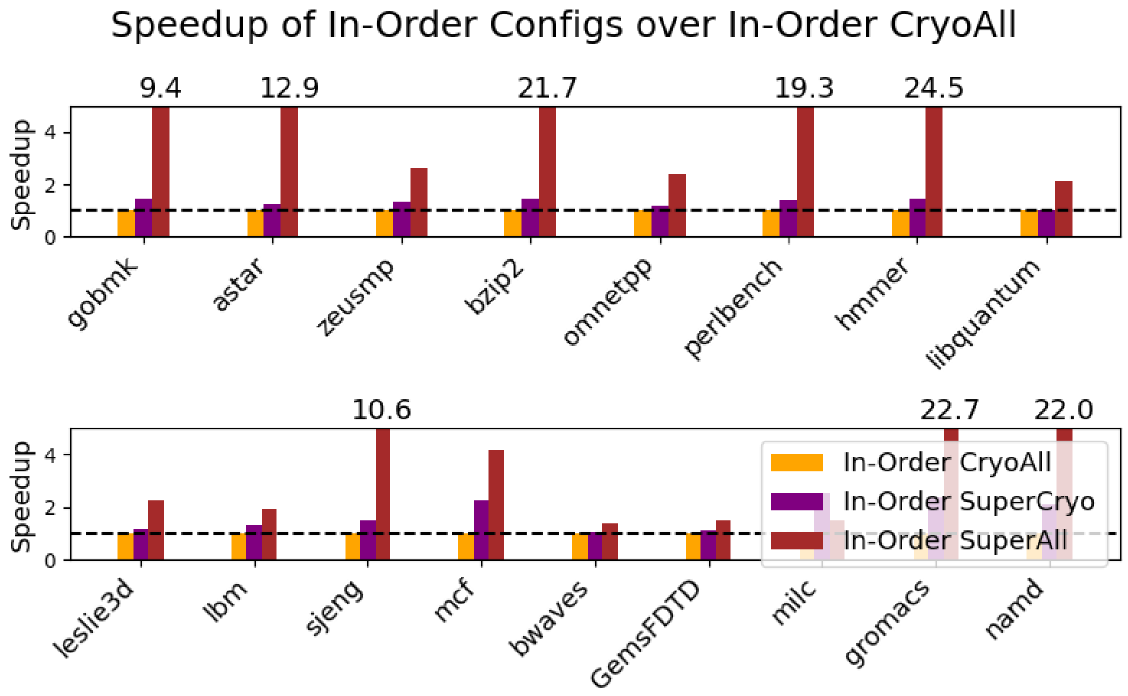
L1 32 kB, L2 512 kB, L3 16 MB

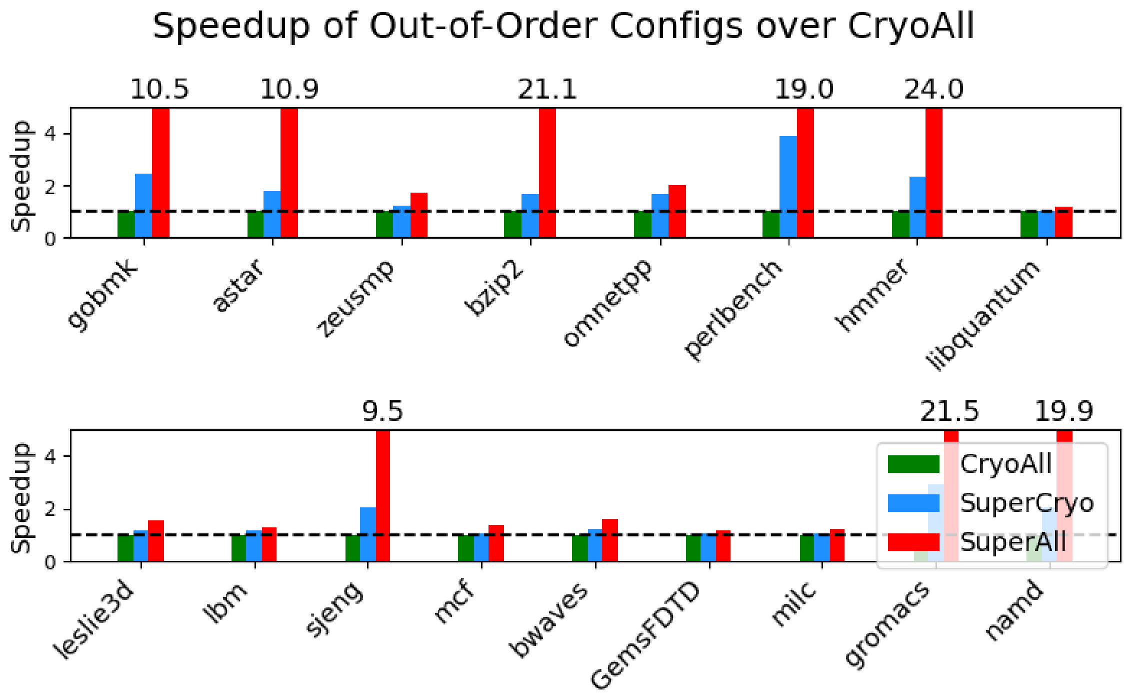
Room temp. (800 MHz)

Mem. hard to scale in cryo/super

Performance Improvement

- High potential bar:
  - but low freq. caches are bottleneck
- More abs. impact on in-order
  - Latency hiding less important
- Memory-intensive workloads:
  - minimal improvement
- Main bottleneck:
  - Room temp. DRAM



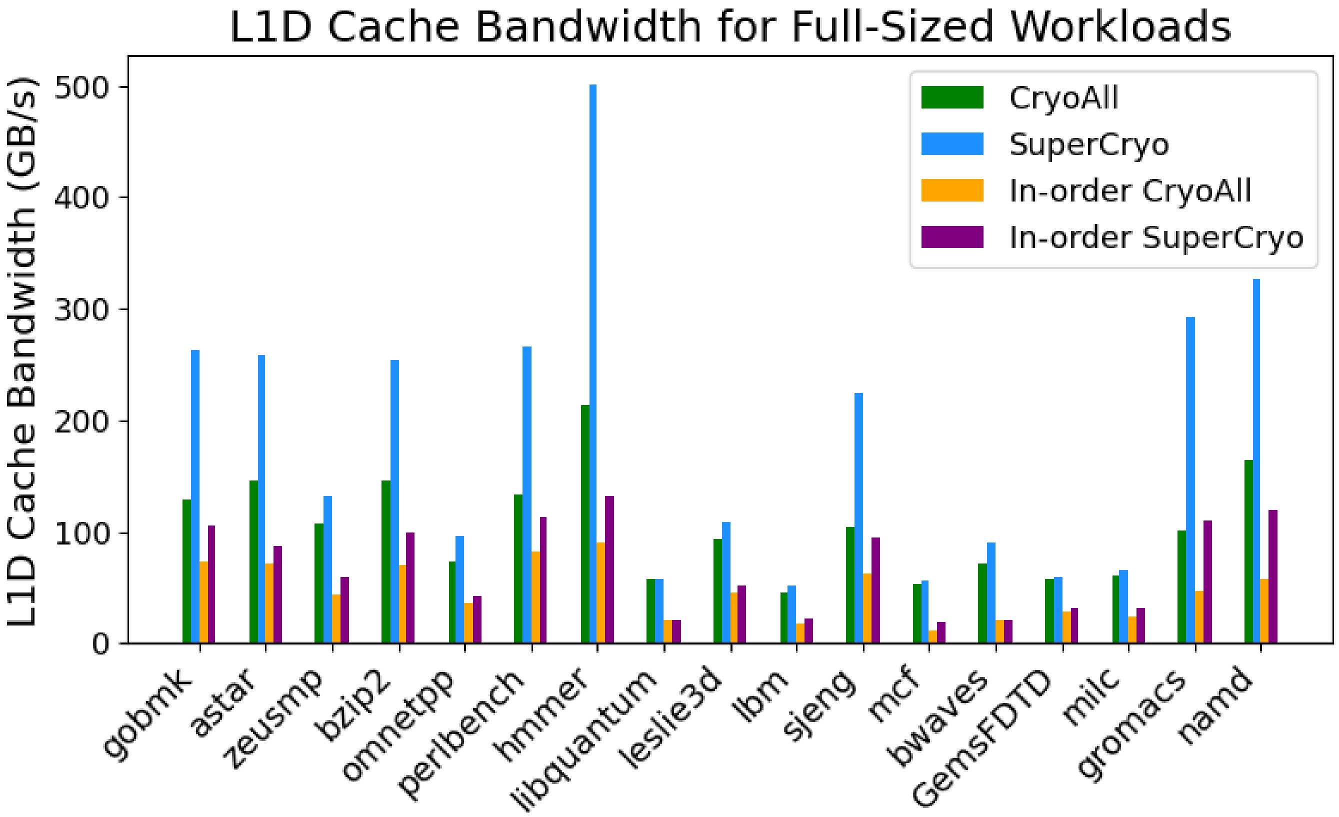


Performance Improvement

- Take away:
  - Big potential benefits, but only for some workloads: **Accelerator, Interconnect**

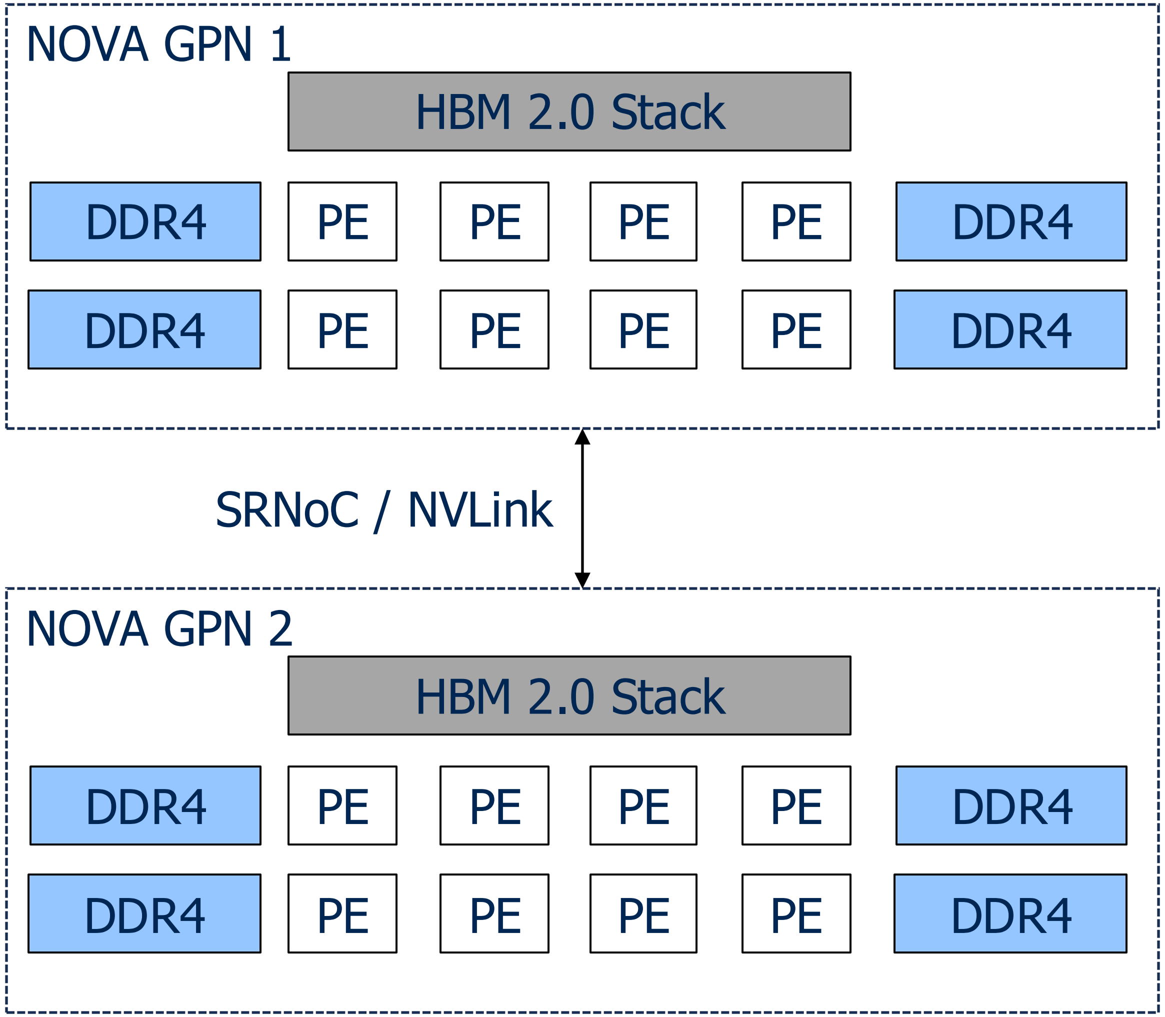
Data Movement

- CryoAll and SuperCryo (in-order and OOO) - realistic configs
- Max. 500 GB/s for L1D Cache in SuperCryo configuration.
  - Reasonable for optics!



Interconnect Model

- SRNoC: Circuit-switched, statically-scheduled
- Workloads: BFS, PR and CC
- Graph size: 12 k nodes, 60 k edges



SRNoC Results

Workload	Slowdown	NVLink Energy (J)	SRNoC Energy (J)	Efficiency Gain
BFS	**1.05×**	1.06×10^{-6}	2.95×10^{-8}	**35.98×**
CC	**1.31×**	1.31×10^{-4}	1.78×10^{-6}	**73.60×**
PR	**1246.28×**	2.32×10^{-6}	6.44×10^{-5}	**0.04×**

- All workloads: **slowdown**
- Narrow int data paths (8-bit): BFS and CC - **low** slowdown, **high** energy efficiency
- Float transmissions (32-bit): PR - **high** slowdown, **low** energy efficiency

Conclusion

- **Compute-intensive workloads**: gains as high as 24x
  - Limited by CMOS DRAM
- **General-purpose CPUs**: limited benefit
- **Best use case**: narrow-path, domain-specific accelerators (graph / ML)
- **Future:** Explore superconducting memory and SERDES conversion penalties