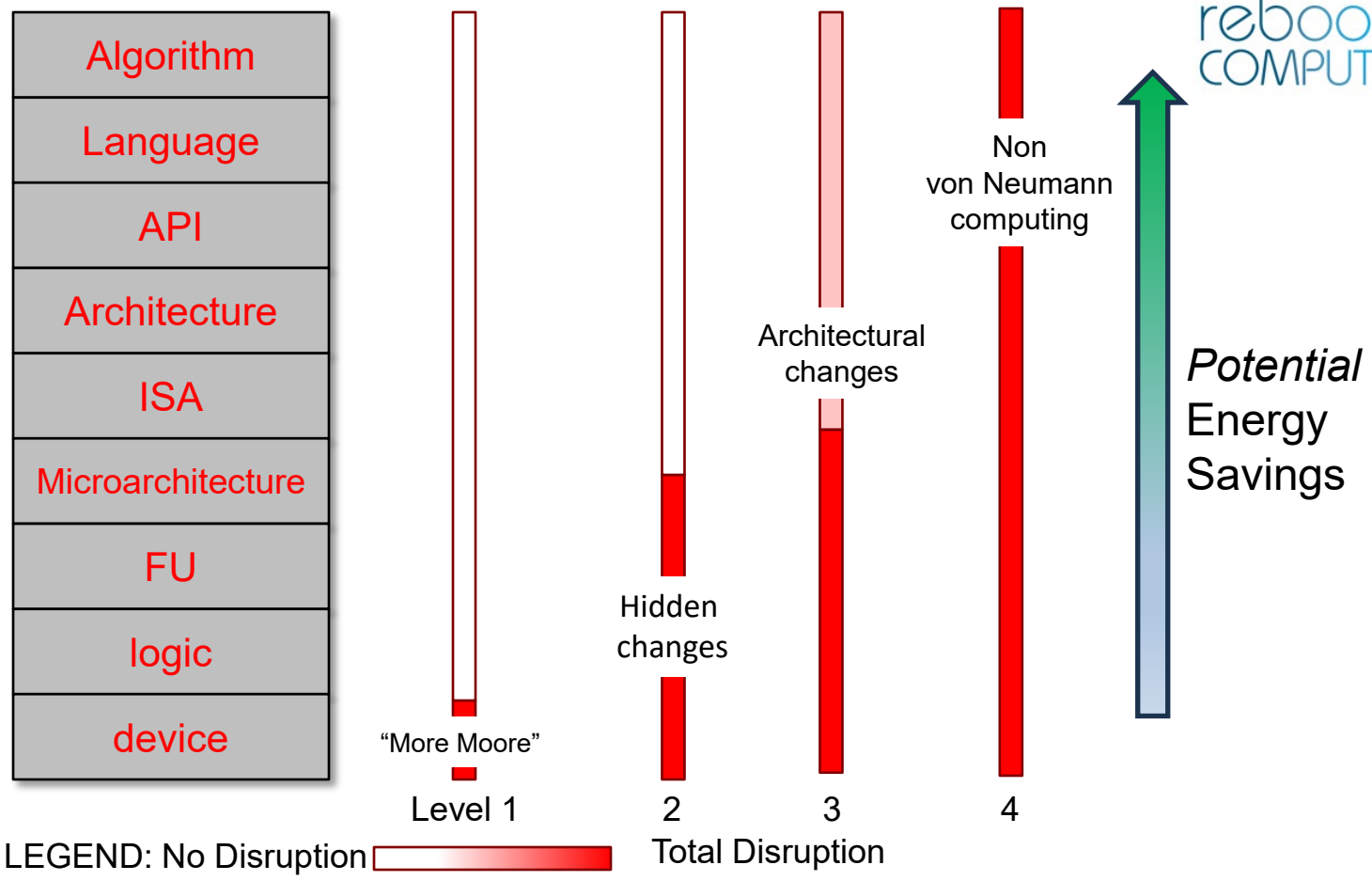
A *Rebooting Computing* Tour of Energy Efficiency Approaches

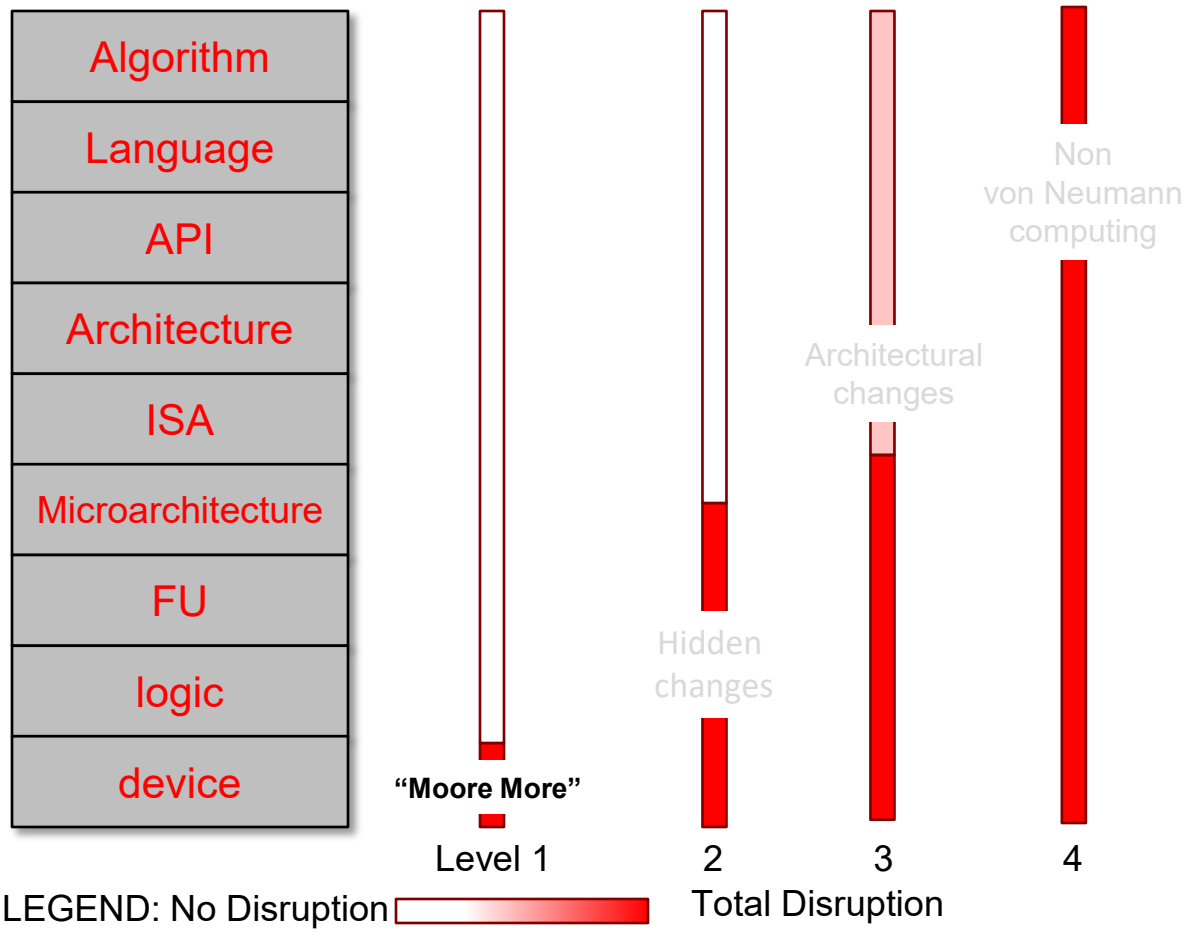
Tom Conte
Georgia Tech and vice chair, IRDS
conte@gatech.edu

Georgia Tech

Potential Approaches vs. Disruption in Computing Stack

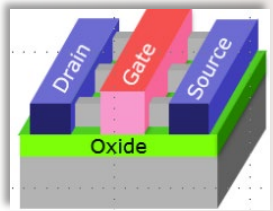


Potential Approaches vs. Disruption in Computing Stack



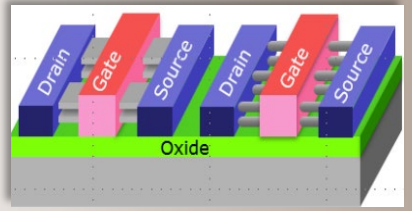
Evolution of Device Architectures

FinFET
2011-2025



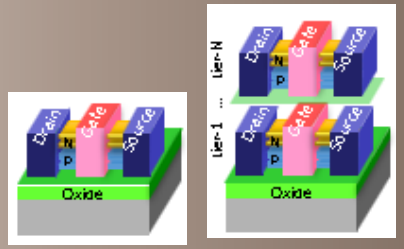
- Increasing drive by taller fin
- Better channel control for better perf-power

Lateral GAA
2025-2037



- Increasing drive by stacked devices
- Better channel control for better perf-power
- Reduced footprint stdcell

3D VLSI
2031 - 2037

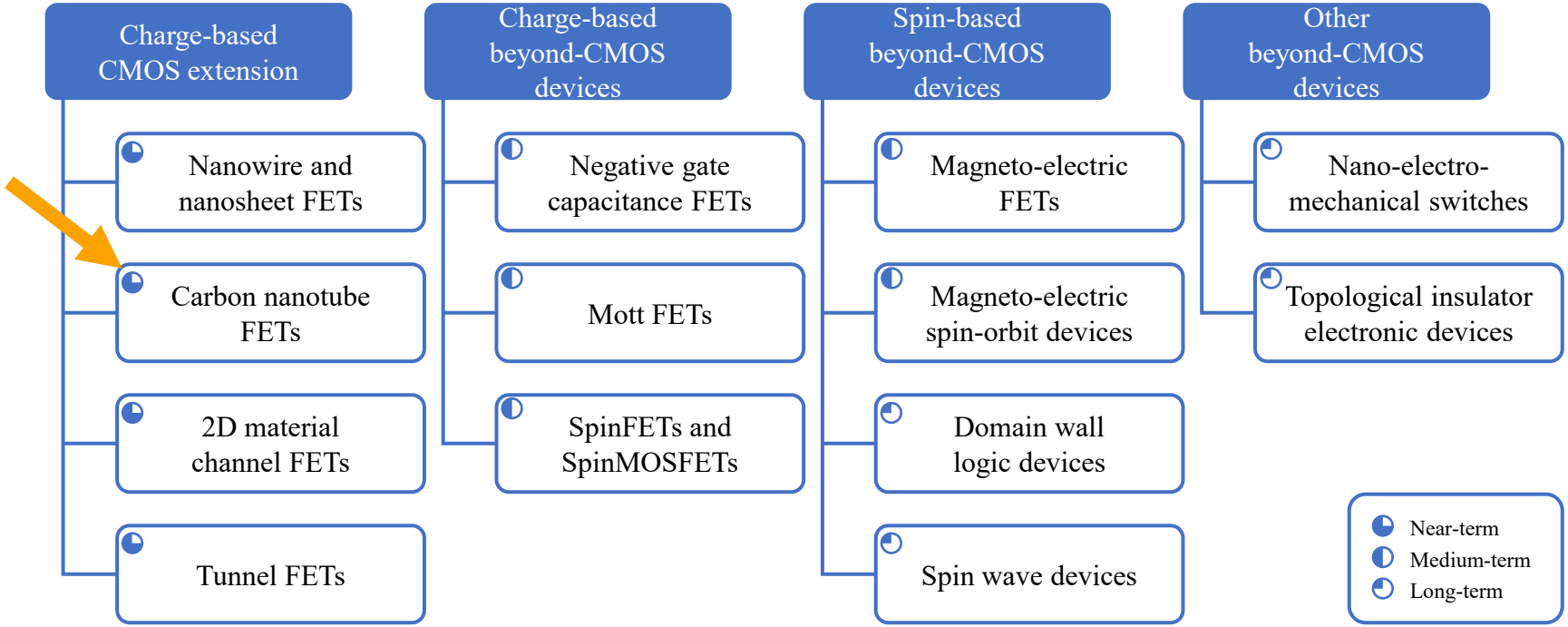


- Monolithic NMOS-over-PMOS
- Sequential integration of multi-device tiers

Power scaling in CMOS (IRDS roadmap)

- CMOS moves to **Gate All Around (GAA) by 2028, true 3D by 2031**
  - Makes power density *worse*
- Some mitigations: **System/Technology Co-optimization**
  - Backside power (wafer PDN)
  - Move away from Copper routing, especially system level
  - Integrate memory on-die and chiplet-based systems
- **New switches** …

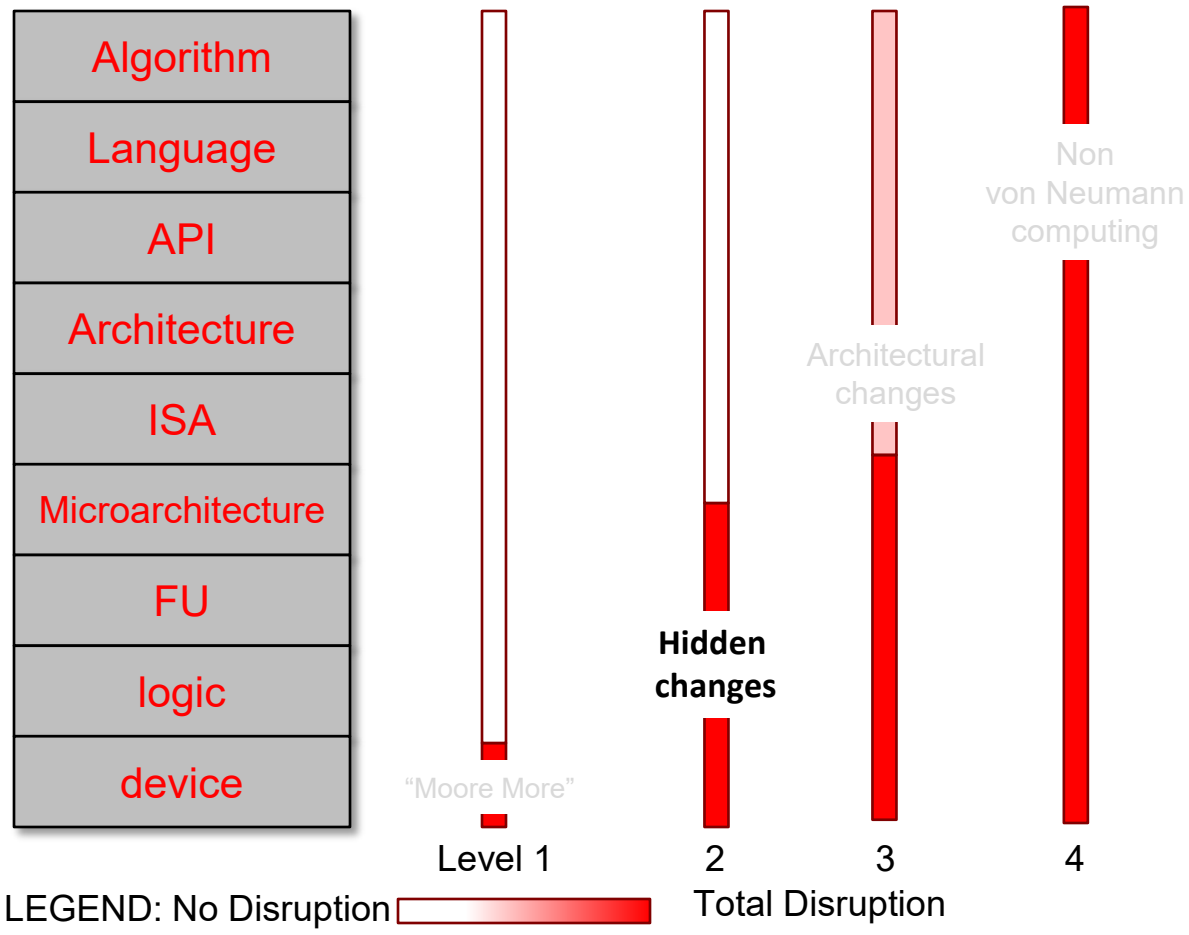
Emerging Logic and Information Processing Devices



Level 1: More Moore

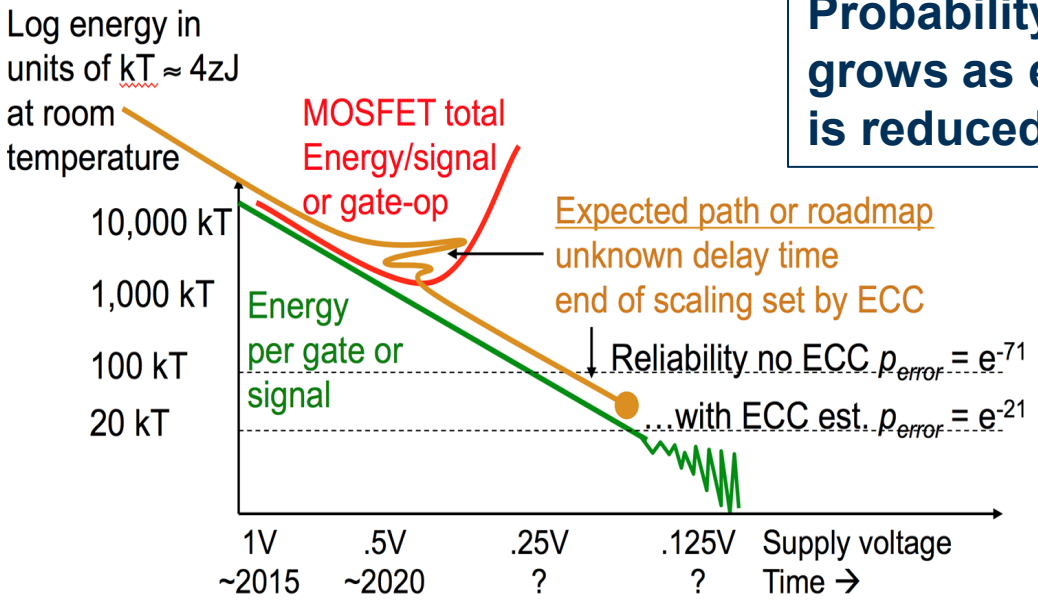
- **No silver bullet from the device world**
- **CMOS goes to monolithic 3D ~ 2028**
  - ***Power improvements then stall***
  - **But, CMOS will remain relevant**
- ***Beyond CMOS:*** *my money is on CNFETs*

Potential Approaches vs. Disruption in Computing Stack



Low Power But *Unreliable* Switches

Idea: keep lowering V_{DD}



Probability of signal error grows as energy of signal is reduced below 20kT

Georgia Tech

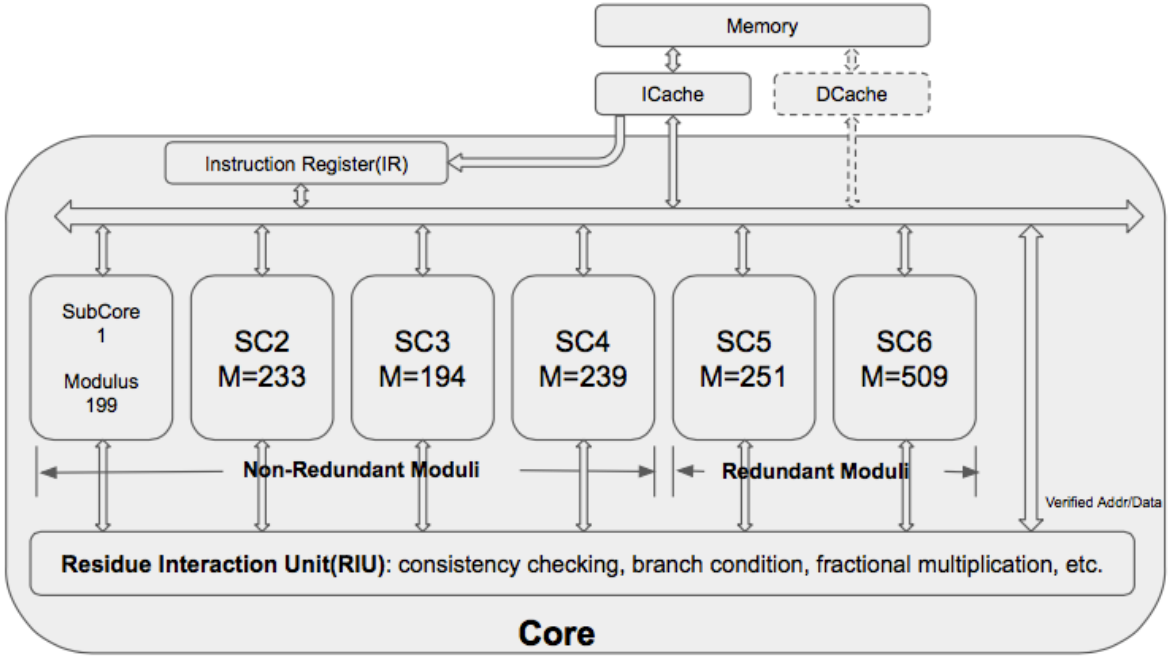
Redundant Residue Numbers (RRNS): Correct Errors *After Computation*

		Range = 3*5*2*7 = 210				Redundant	
	decimal	**mod 3**	**mod 5**	**mod 2**	**mod 7**	**mod 11**	**mod 13**
	13	1	3	1	6	2	0
	14	2	4	0	0	3	1
add case 1	27	0	2	1	6	**7->5**	1
add case 2	27	0	**1->2**	1	6	5	1
add case 3	27	0	**1->2**	1	6	**7->5**	1

	Chinese Remainder Theorem	**|X'|11 & |X'|13**	**|X'|mc == |X|mc ?**	**How to correct?**
Case 1	(0,2,1,6) ⇔27 ; X' = 27;	|27|11 = 5; |27|13 = 1	|X'|m5 = 5 |X|m5 = 7 |X'|m6 = 1 |X|m6 = 1	replace |X|m5 with |X'|m5
Case 2	(0,1,1,6) ⇔111 ; X' = 111;	|111|11 = 1; |111|13 = 7	|X'|m5 = 1 |X|m5 = 5 |X'|m6 = 7 |X|m6 = 1	check error correction table
Case 3	Two errors; Double Errors Detection algorithm could be used to detect errors. But unable to correct.			

Georgia Tech

Demonstrators of RRNS Cores



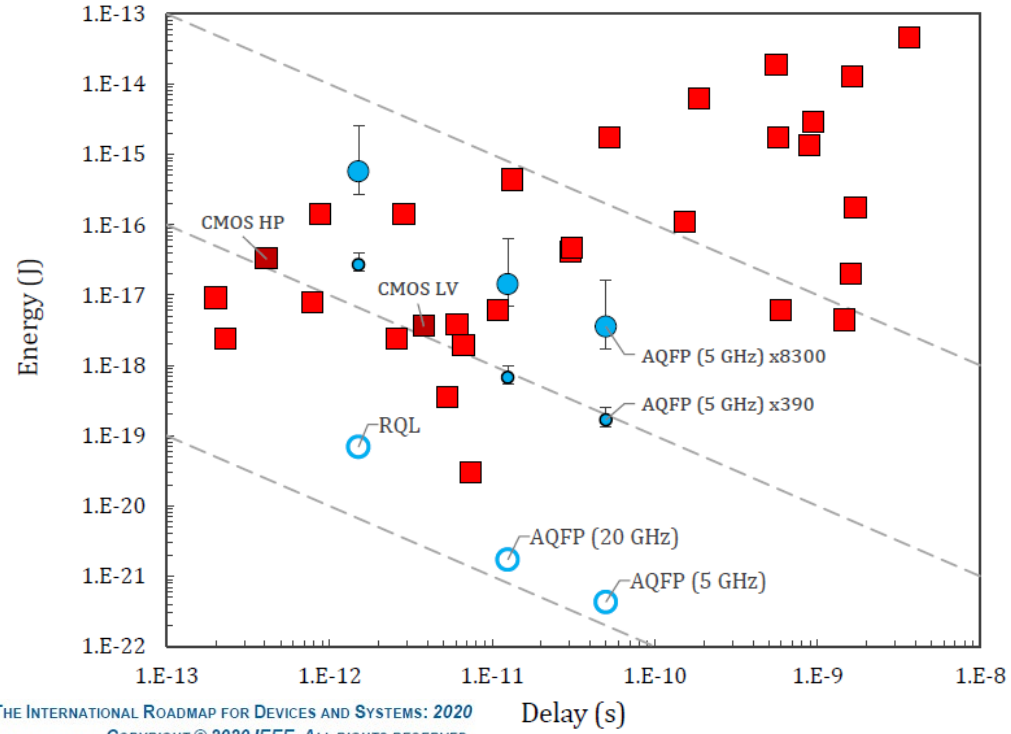
B. Deng, et al., "Computationally-redundant energy-efficient processing for y'all (CREEPY)," Proceedings of the 2016 IEEE International Conference on Rebooting Computing (ICRC), (San Diego, CA), Oct. 17-19, 2016.

Problems:

- Not enough demonstrators exist
- Need new floating point beyond 754

Superconducting Logic

Superconducting logic vs. CMOS

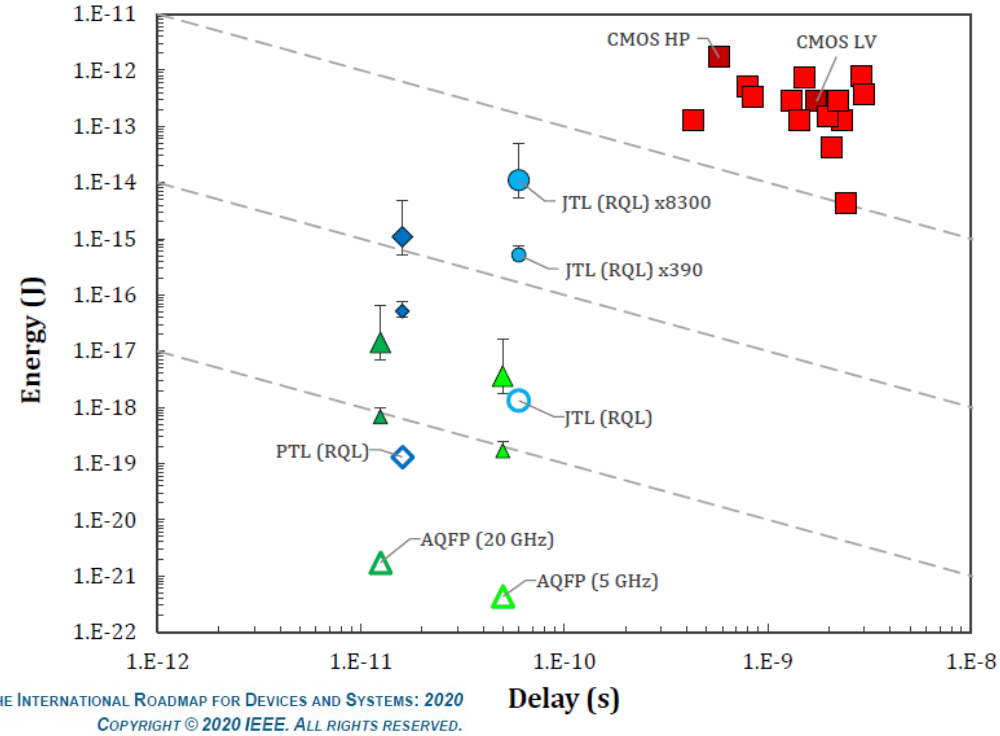


THE INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS: *2020*

- *w/o refrigeration*
- *w/ refrigeration*

Georgia Tech

Superconducting WIRES vs. CMOS



THE INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS: 2020

w/o refrigeration

- JTL (RQL)
- PTL (RQL)
- AQFP (5GHz)
- AQFP (20Ghz)

w/ refrigeration

- JTL (RQL)
- PTL (RQL)
- AQFP (5GHz)
- AQFP (20Ghz)

Georgia Tech

Level-2 ex 2: Superconducting

- Superconducting computing wins on the ***wires***

But:

- **Need better cryostats**
- **Feature size must come down** (~200nm today)
- **Memory scarce**: can get ~10s of kilobytes at 4 K
- **Logic families are immature, EDA still lags**
- **Need demonstrators *at scale***
- **Luckily** *riding along on massive investment in QC*

Reversible Logic

Reversible challenges

- Need **experimental demonstrators *at scale***
- **EDA tools lag** (*even when compared to superconducting's sad tool situation*)
- Effects of tuning "Degree of Reversibility" needs study
  - One extreme: only portions of datapath are reversible, reducing benefit
  - The other extreme: entirely reversible algorithms needed

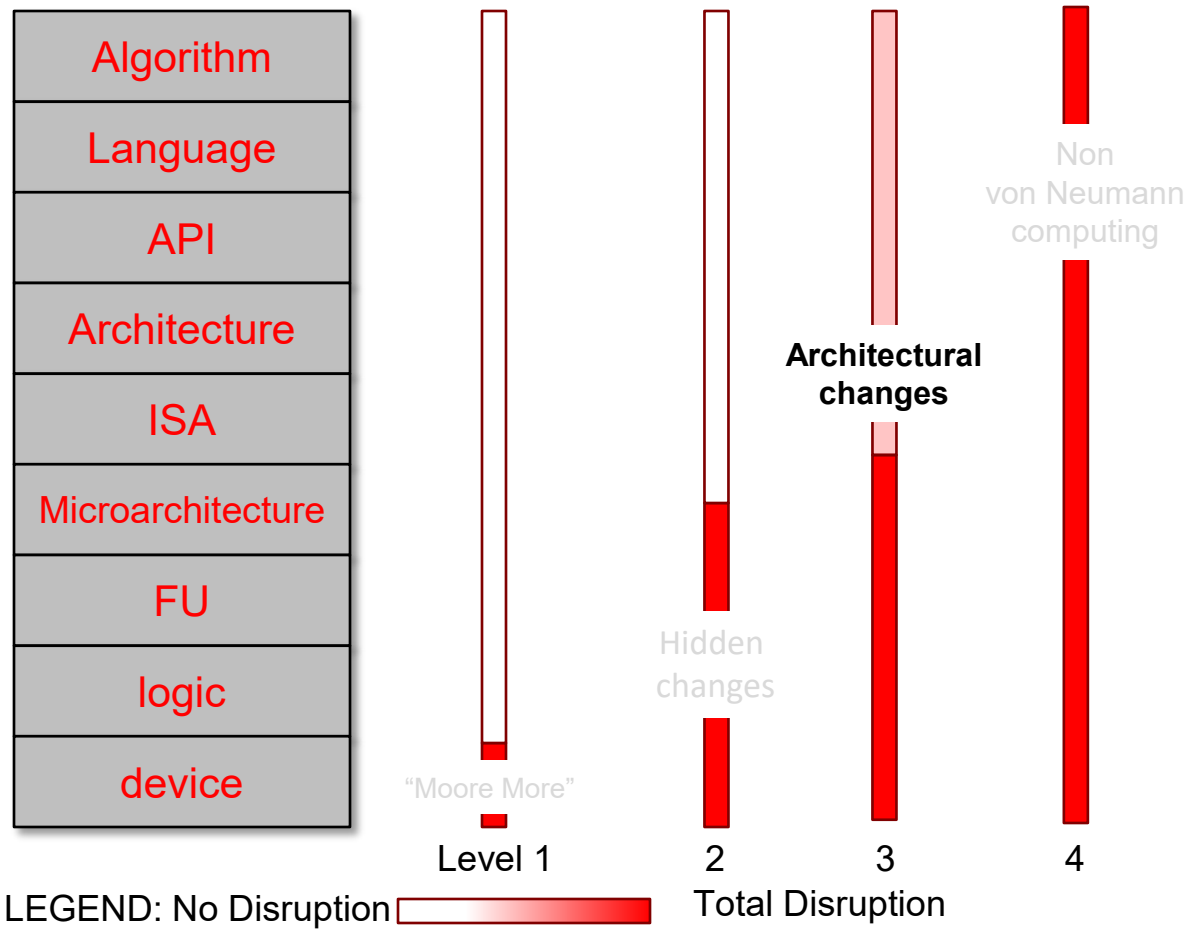
Level 2: Not CMOS, but “hidden” from the programmer

Potential for significantly lower-energy computation

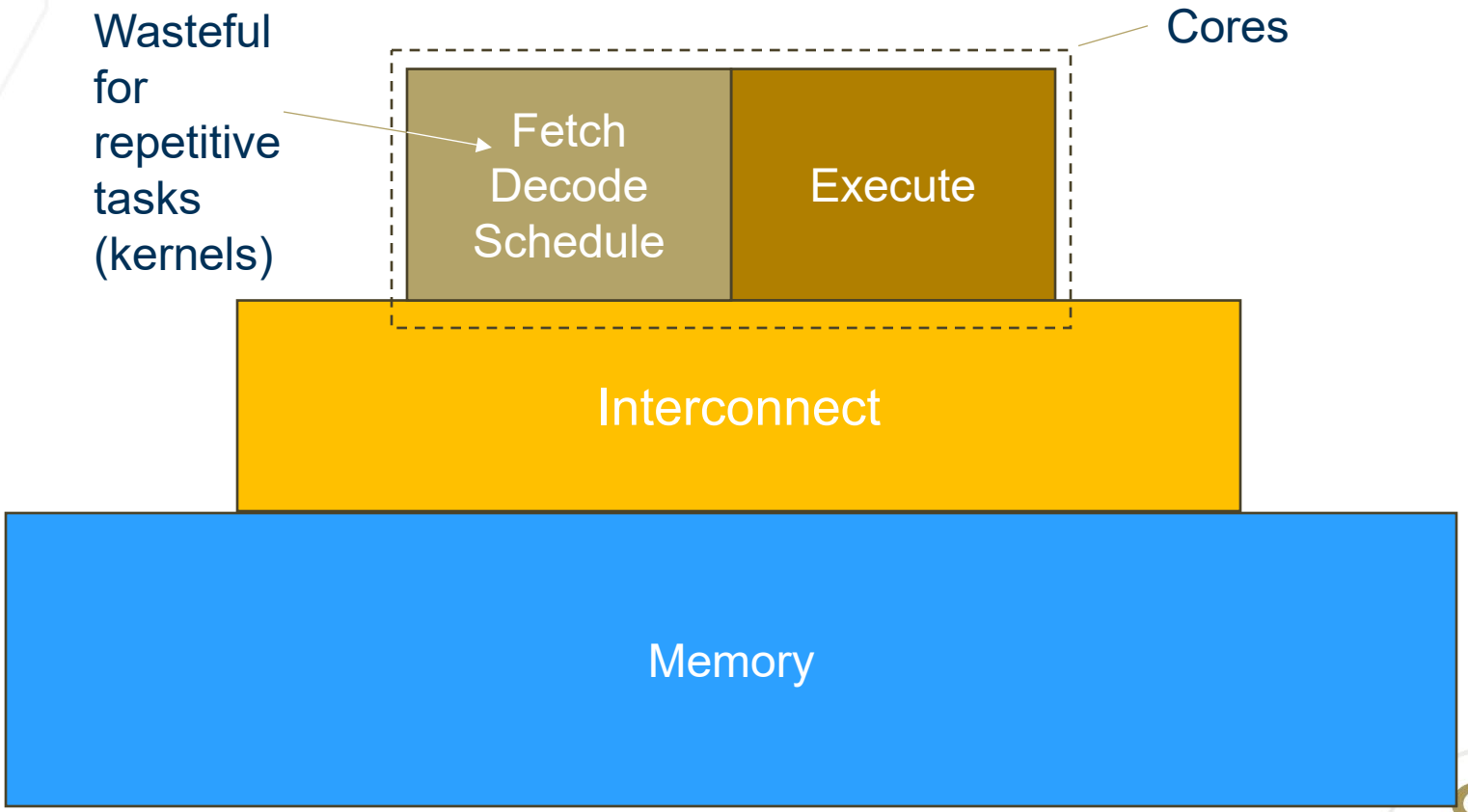
Programmer (mostly) **won’t know / care**

Massive **amount of R&D** *and* **demonstrators needed**

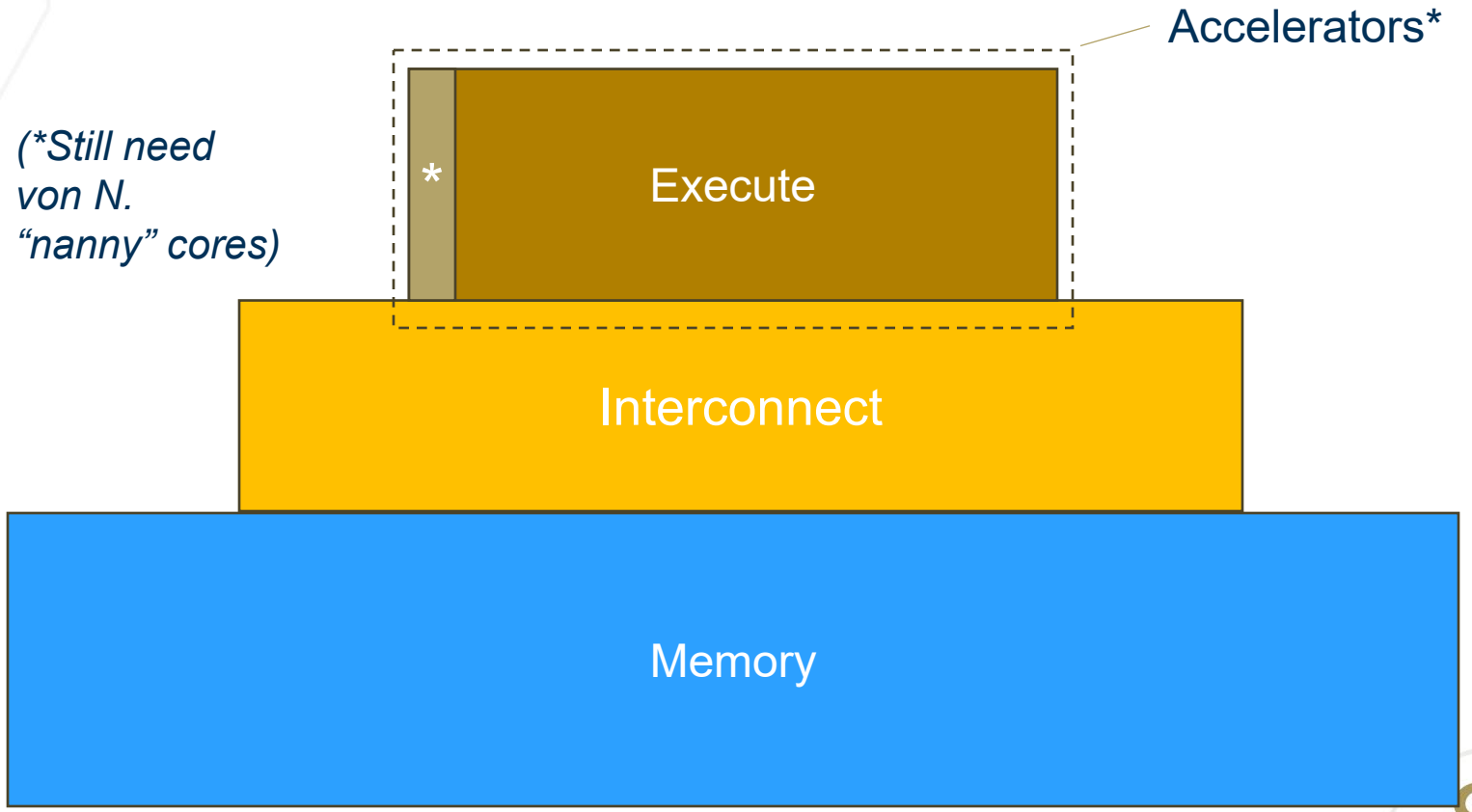
Potential Approaches vs. Disruption in Computing Stack



Hardware Accelerators: Remove Overhead



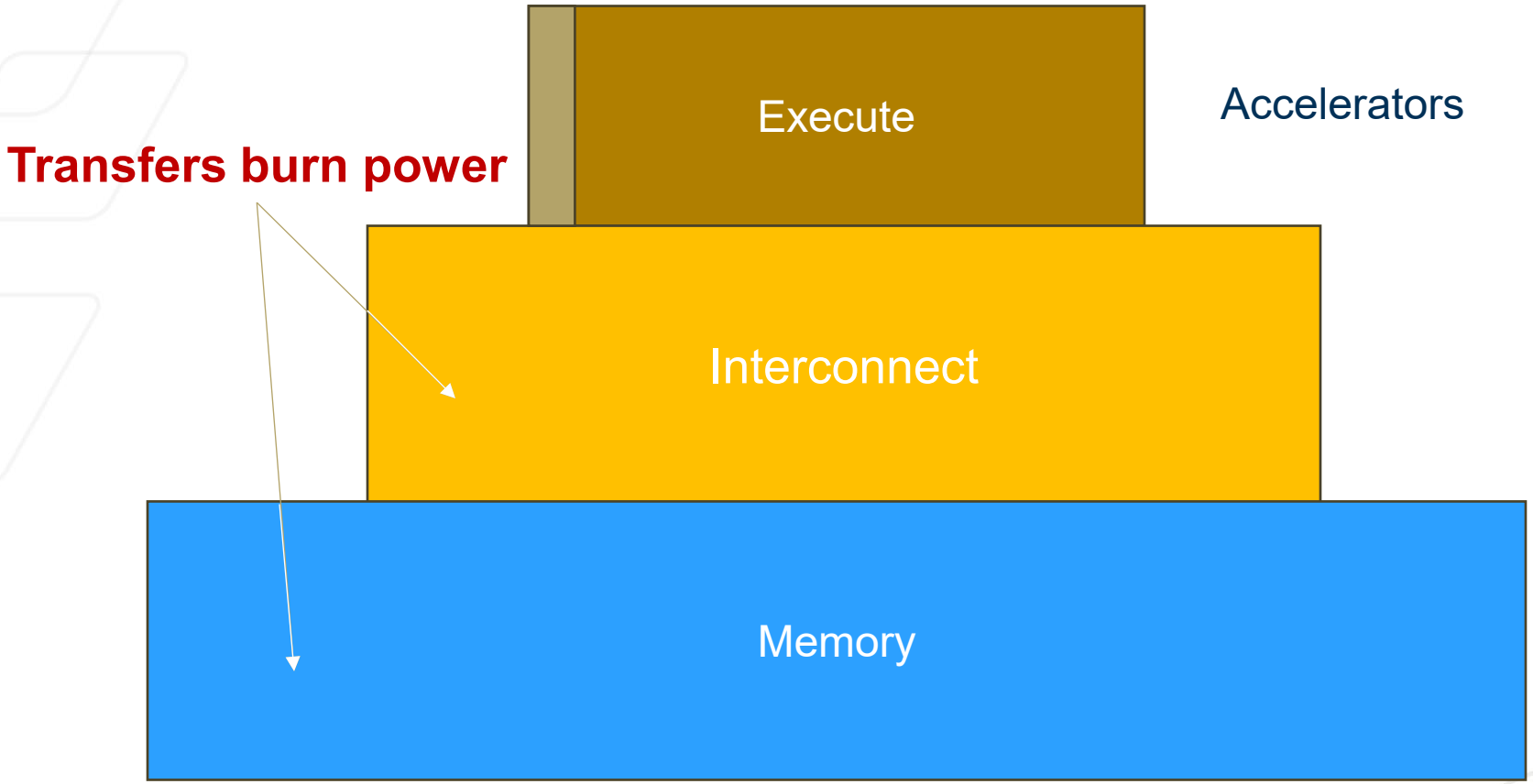
Hardware Accelerators: Remove overhead



Accelerators: Problem: Amdahl's Law

- Hardware acceleration will give us **huge gains**, *and then we're done*
- Or in other words: ***after we accelerate all the kernels, what do we do next?***

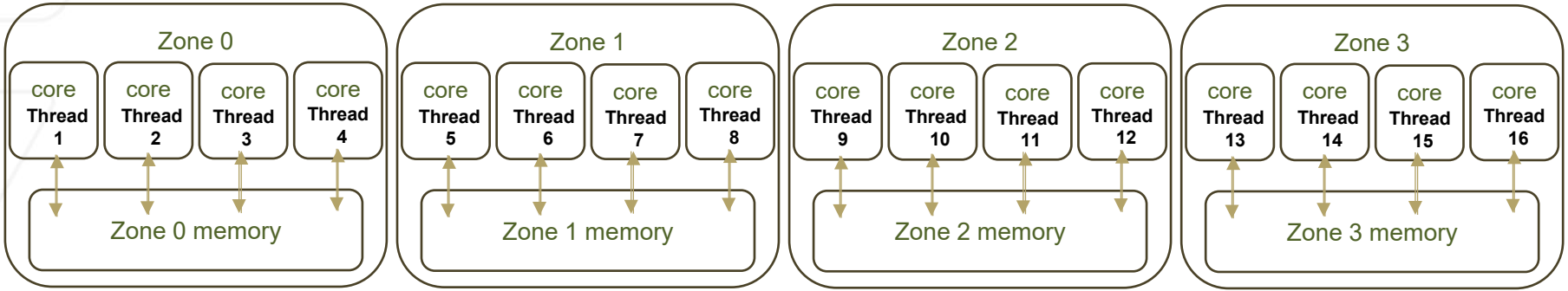
  Can't let other approaches atrophy in the interim



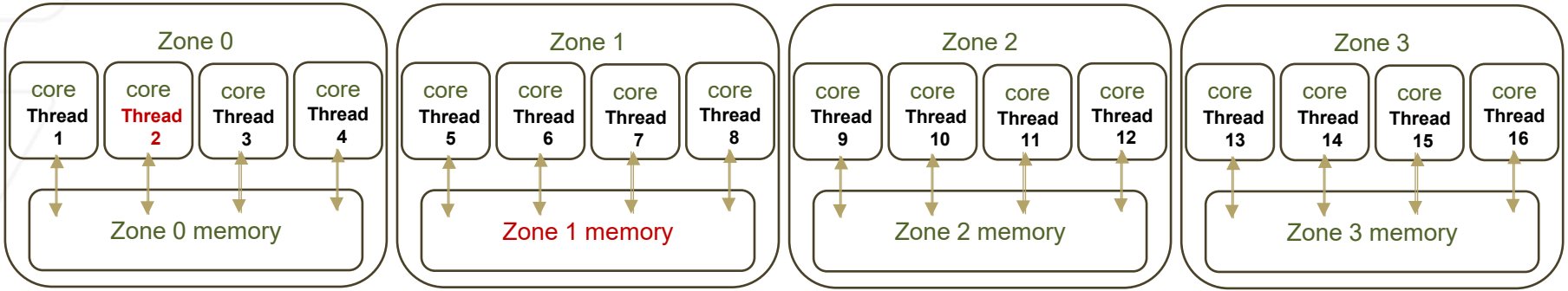
Level 3, ex#2: *Avoid Moving Data*

- **Moving data is an energy-consuming "fact of live"**
  - *But not all data needs to move*
- **What about moving the computation instead**?
  - Examples: *Emu Chick, Lucata Pathfinder, FORZA*

Example -
PGAS and threads can only access their local zone memory

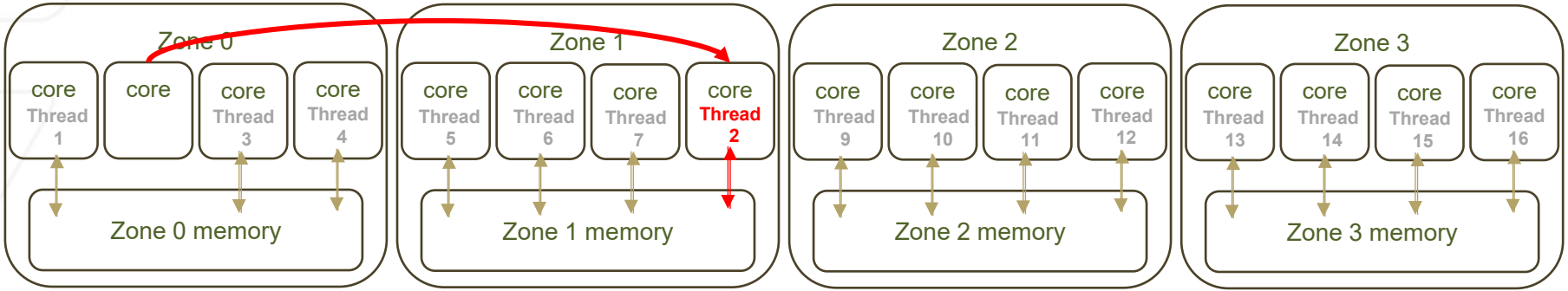


Example -
PGAS and threads can only access their local zone memory



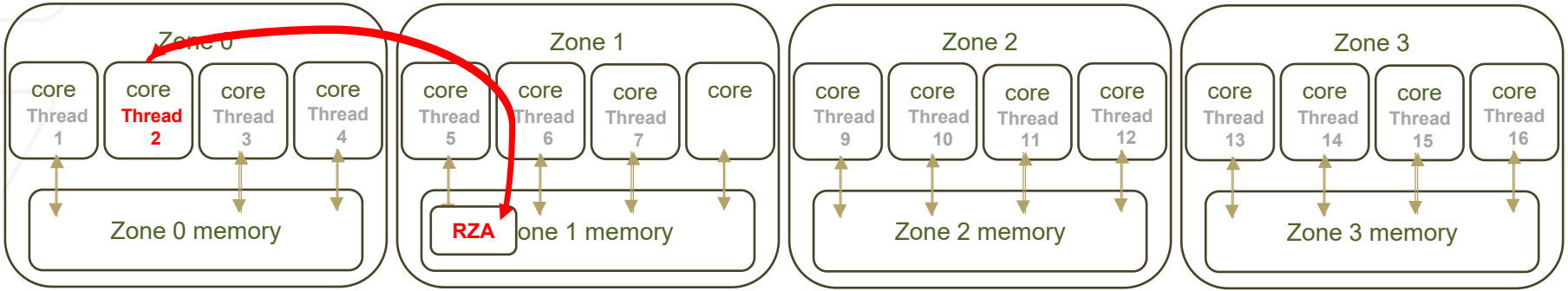
Ex: Thread 2 wants to work on data in Zone 1

Thread Migration: Move Computation to the Data

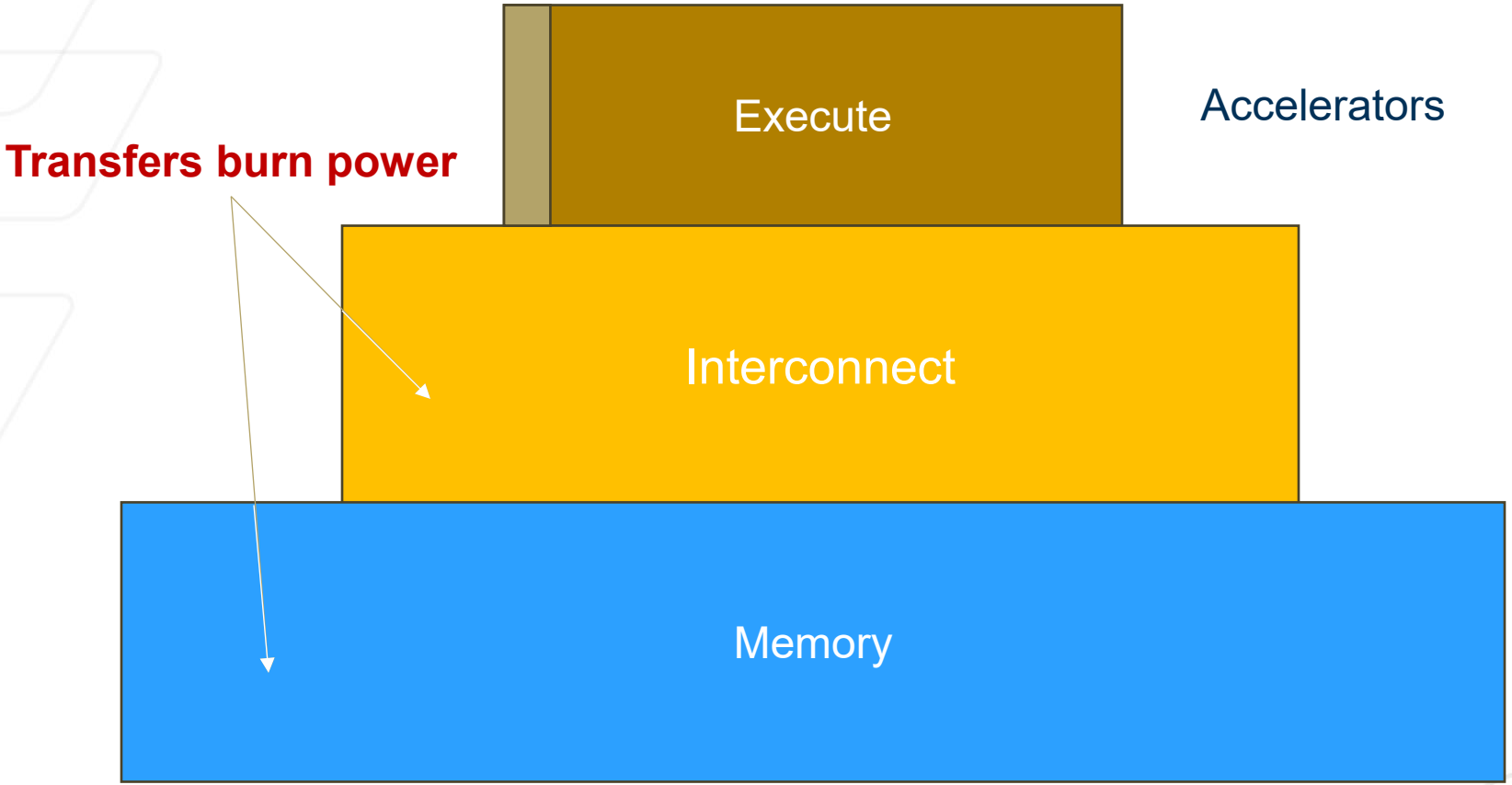


Hazard: Ping-ponging across zones consumes more energy than conventional approaches

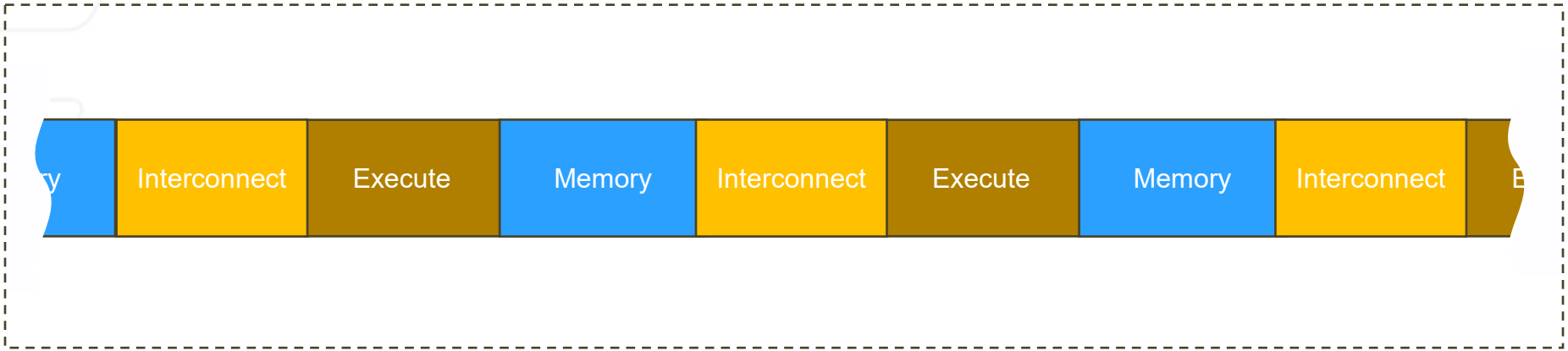
Or…Ask for a Favor: Remote Zone ALU Operations



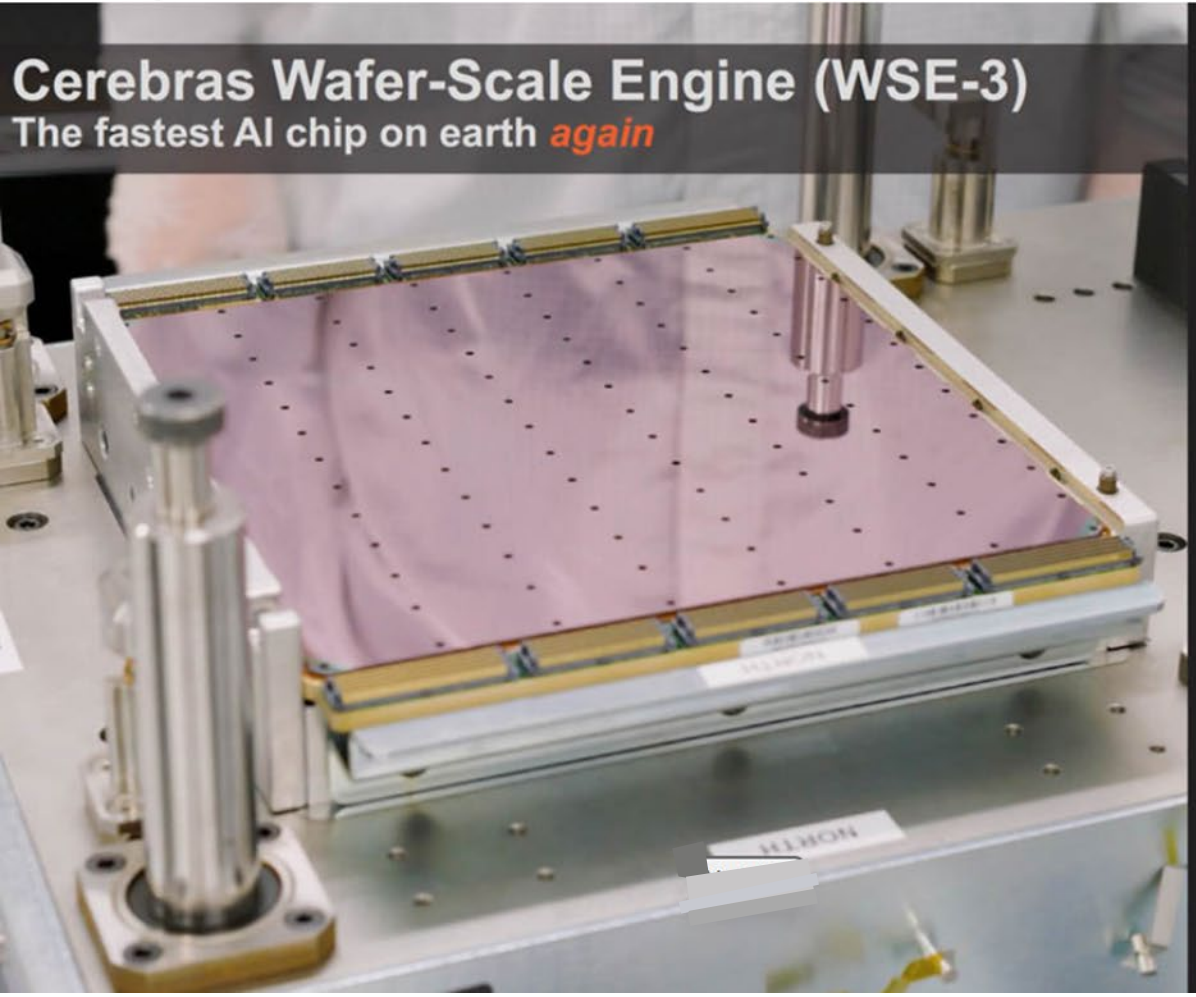
For light-weight tasks, don't migrate, instead operate remotely at-a-distance
More potential if RZA operations are programmable


Execute
Accelerators
Transfers burn power
Interconnect
Memory

Level 3: ex 3: Monolithic integration



Cerebras WSE-3



4 trillion transistors

46,225 mm² silicon

900,000 cores optimized for sparse linear algebra

5nm TSMC process

125 petaflops of AI compute

44 gigabytes of on-chip memory

21 PByte/s memory bandwidth

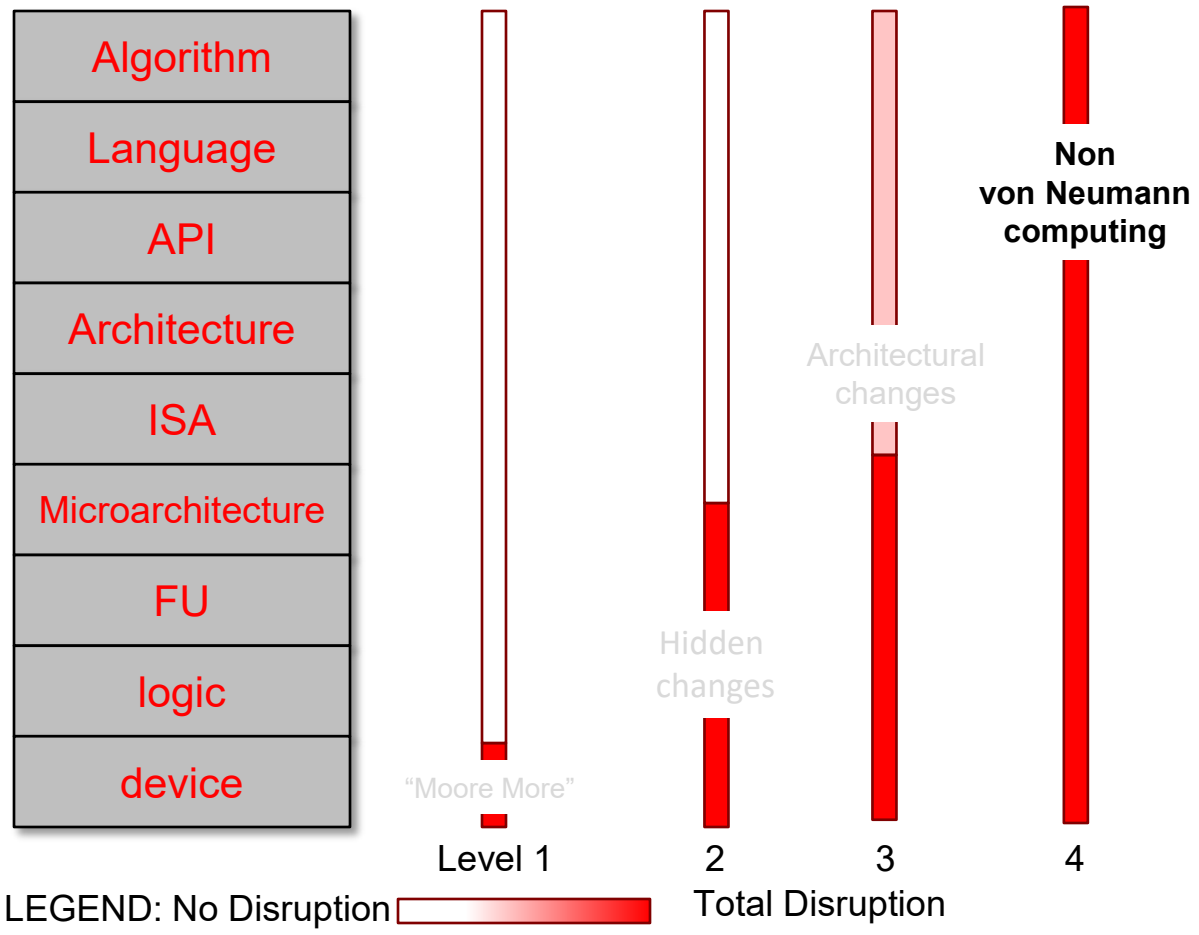
214 Pbit/s fabric bandwidth

Level 3: Architectural changes

- Accelerators shouldn’t consume all the oxygen
- Save energy by removing overheads & not moving bits
- Or, go cold! Superconducting deserves more attention

- For all level 3, programmer pain but for *significant* gain

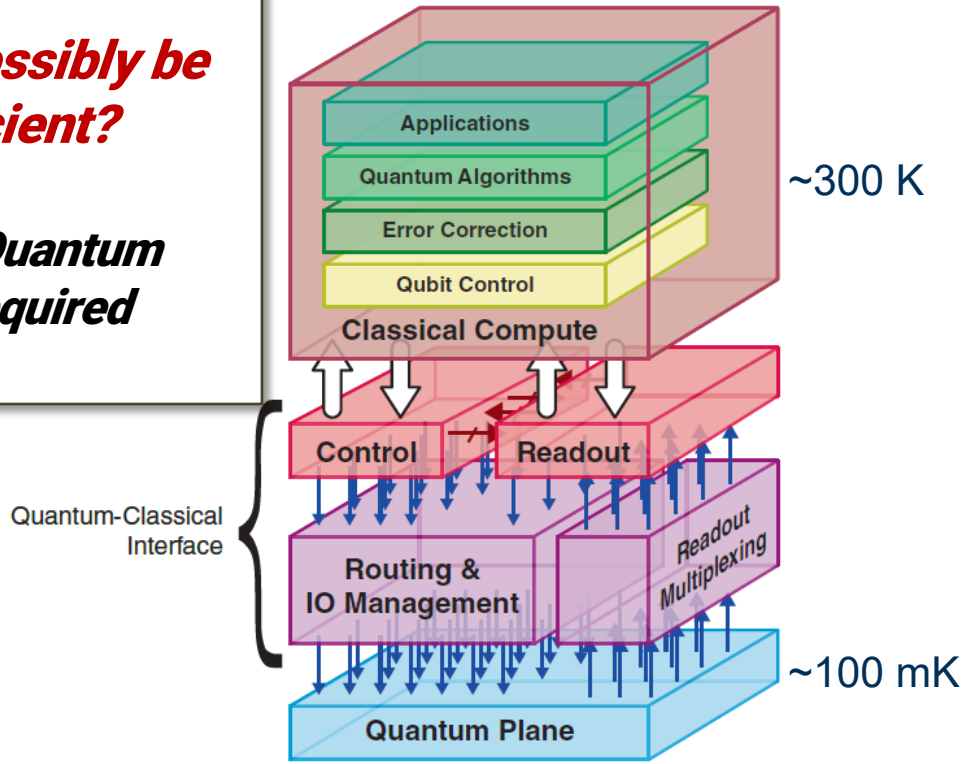
Potential Approaches vs. Disruption in Computing Stack



Quantum Computing: Energy Advantage?

How can this possibly be energy efficient?

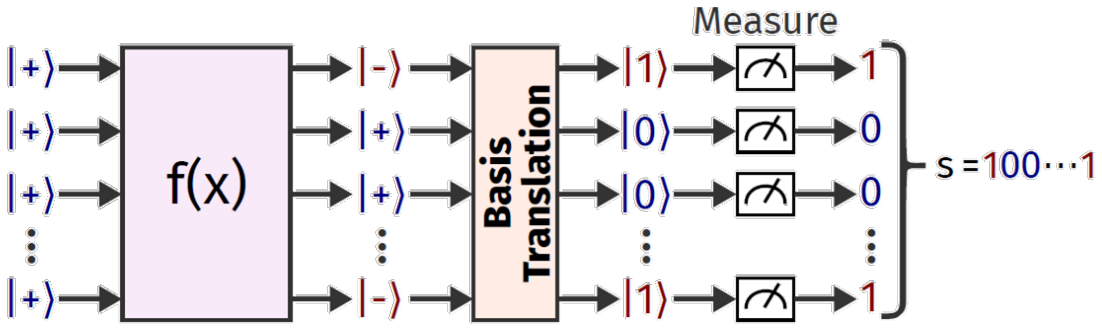
SIGNIFICANT Quantum Advantage Required



S. J. Pauka, et al., "A cryogenic interface for controlling many qubits," arXiv:1912.01299v1

Georgia Tech

Bernstein-Vazirani (QC's "hello world!")



Q#

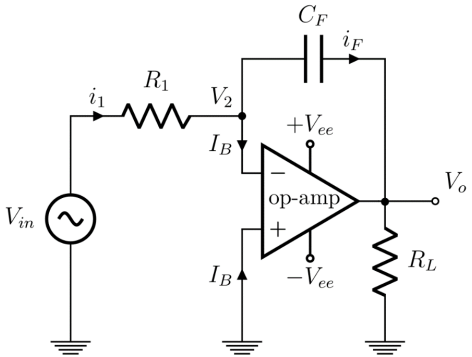
```
use queryRegister = Qubit[n];
use target = Qubit();
X(target);
within {
    ApplyToEachA(H, queryRegister);
} apply {
    H(target);
    Uf(queryRegister, target);
}
let results = ForEach(MResetZ,
                      queryRegister);
Reset(target);
return results;
```

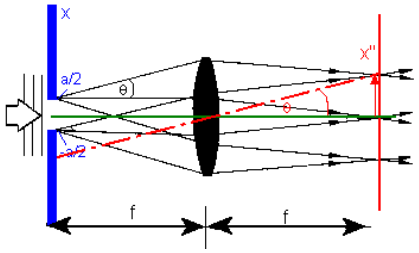
- **Can you teach an undergrad CS major to think up *new algorithms* in this language?**
  - *(No, not easily)*
- **Need new layers of abstraction**

Georgia Tech

Non-von ex 2: Analog (physical) Computing

Physical process
y = F(x)

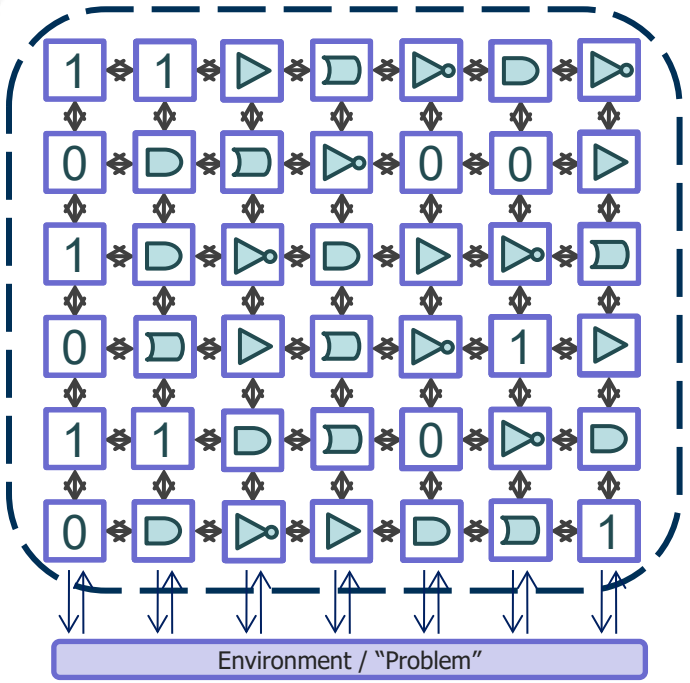




Problems: **Conversion** (DAC/ADC) and **Precision**

Georgia Tech

Non-von ex 2: Thermodynamic Computing



- TDCs work at the *theoretical limit* of energy efficiency
- Problems: **Programming** and **Demonstrators**

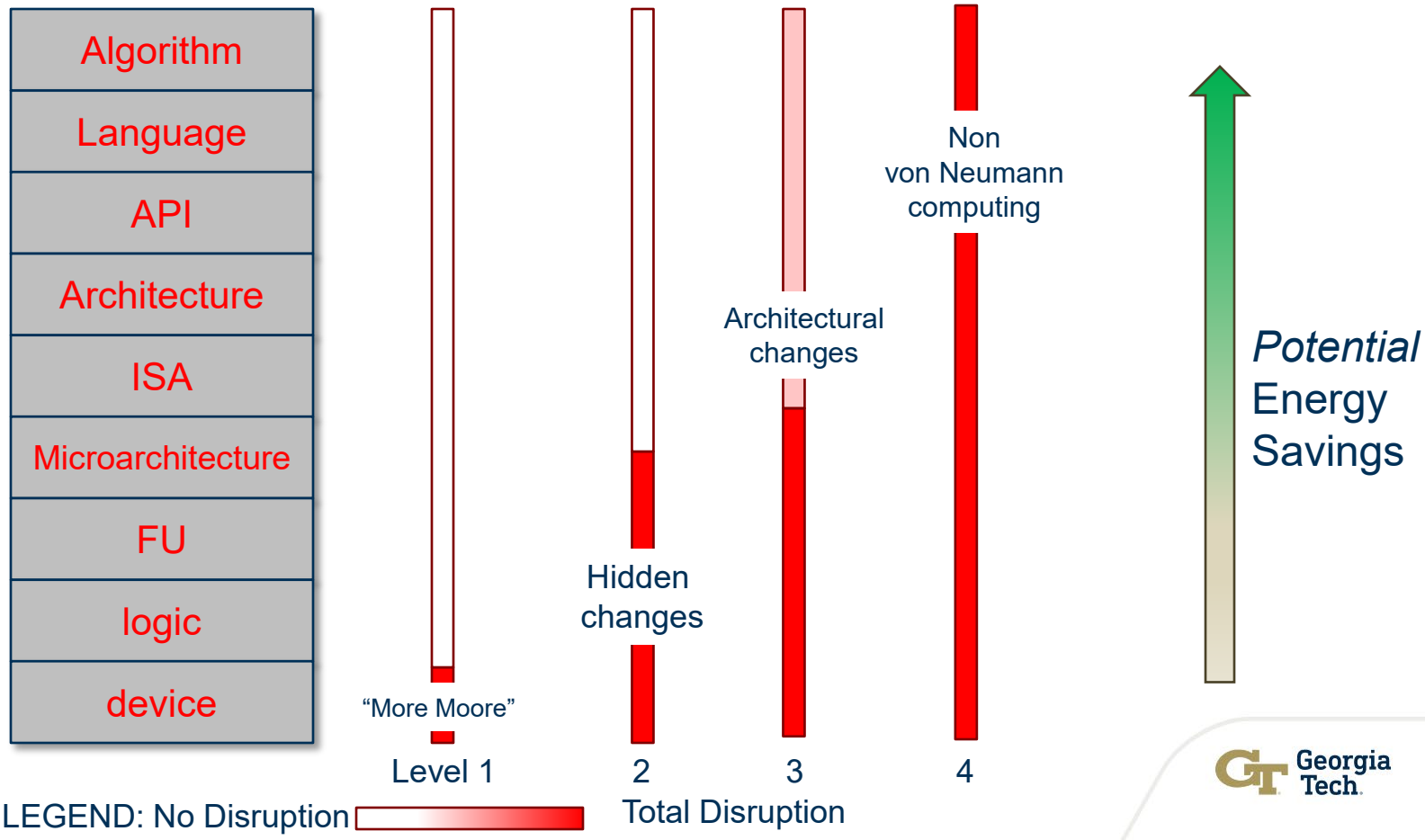
Hylton, T., Conte, T.M., Hill, M. D., A Vision to Compute like Nature: Thermodynamically, CACM, v.64, n.6, 1 June 2021

Georgia Tech

Level 4: Non-von Neumann

- (Very) immature technologies…
- …but HUGE potential for energy savings
- Low investment so far (except for QC!)
- Need new algorithms, programmability, demonstrators

Potential Approaches vs. Disruption in Computing Stack



Most significant bits

- **Need investment in experimental demonstrators**
  - ***Industry is too risk adverse*** to make this
  - **EDA needs to keep up** with experimental technologies
- **Software is highly important**
  - Level 2 is the most attractive, Level 3 somewhat less so
  - Level 4: These systems need to become easily programmable
- ***For my money:***
  - level 2: Superconducting Computing (Including Reversible)
  - level 3: Accelerate Everything *and* Move Compute To Data
  - level 4: Thermodynamic And Probabilistic Computing