# Computation and Communication in a Post Moore's Law Era

**Dr. George Michelogiannakis**

Research scientist

Lawrence Berkeley National Laboratory

Adjunct professor

Stanford University

# Overview and Outline

- Traditional device scaling is ending
- We have to preserve computation performance scaling with a variety of emerging technologies
- Meeting future goals cannot happen without a multi-layer approach
  - Need tools and methodologies
- If we succeed, communication will become the bottleneck
  - We can no longer overdesign networks
- This calls for a grand strategy
- This talk is meant to be thought-provoking: Lots of ongoing work

# Poll: What Did Dr. Moore Say

- Transistor density will increase by 2x every **12** months
- Transistor density will increase by 2x every **18** months
- Transistor density will increase by 2x every **24** months

(may have multiple answers)

# Poll: What Did Dr. Moore Say

- **Transistor density will increase by 2x every 12 months**
  - In 1965 [1]
- Transistor density will increase by 2x every 18 months
  - Average of the two
  - Actual doubling rate around 1975
- **Transistor density will increase by 2x every 24 months**
  - In 1975 [2]

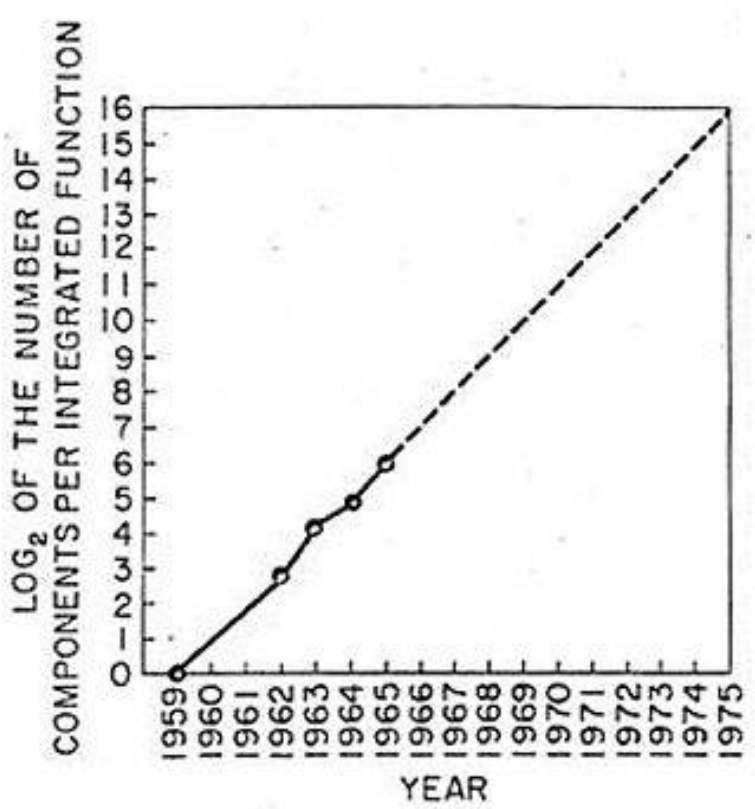



Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

Dr. Moore's 1965 paper [1]


[1] G. E. Moore, "Cramming More Components onto Integrated Circuits," Electronics, Vol. 38, No. 8, 1965, pp. 114-117.
[2] G. E. Moore, "Progress In Digital Integrated Electronics," International Electron Devices Meeting, IEEE, 1975, pp. 11-13.

# Technology Scaling Trends

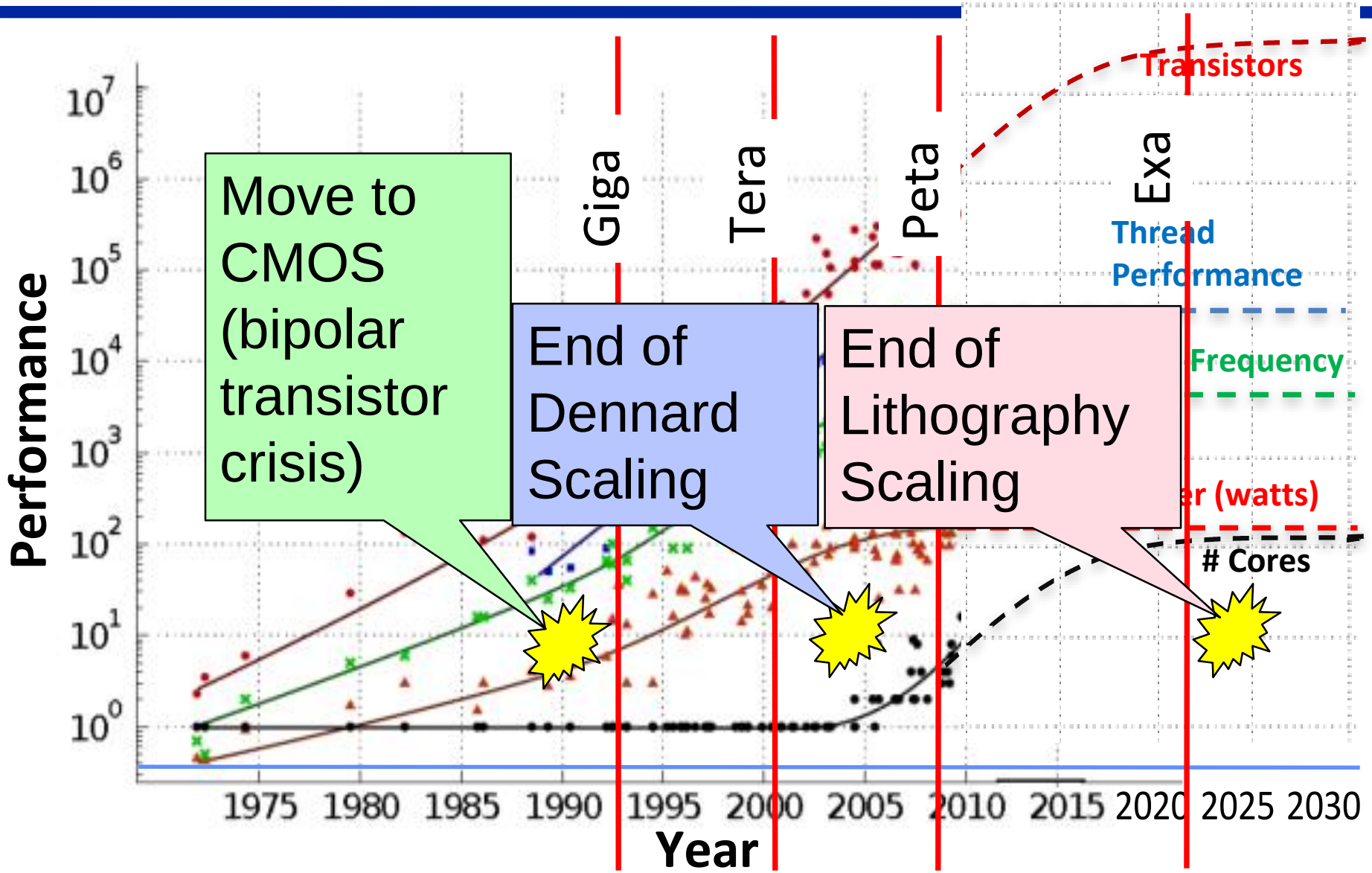
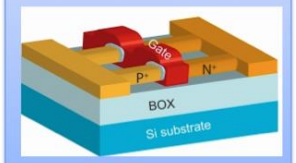



Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, and John Shalf

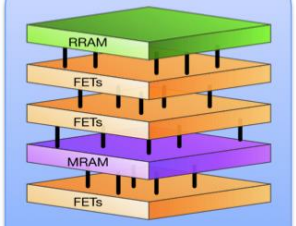
# Moore's Law of Documentation

# Computation Challenge: Preserve Performance Scaling With Emerging Technologies

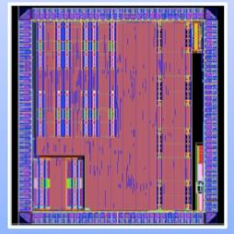



Emerging transistors

Emerging memories




3D integration



Specialized architectures

Performance

## Now – 2025

**Moore's Law continues through ~5nm -- beyond which diminishing returns are expected.**

2016

2016-2025

**End of Moore's Law**
***2025-2030?***

## Post Moore Scaling

**New materials and devices introduced to enable continued scaling of electronics performance and efficiency.**

Performance

2025+

# Energy Challenge: HPC System Trends

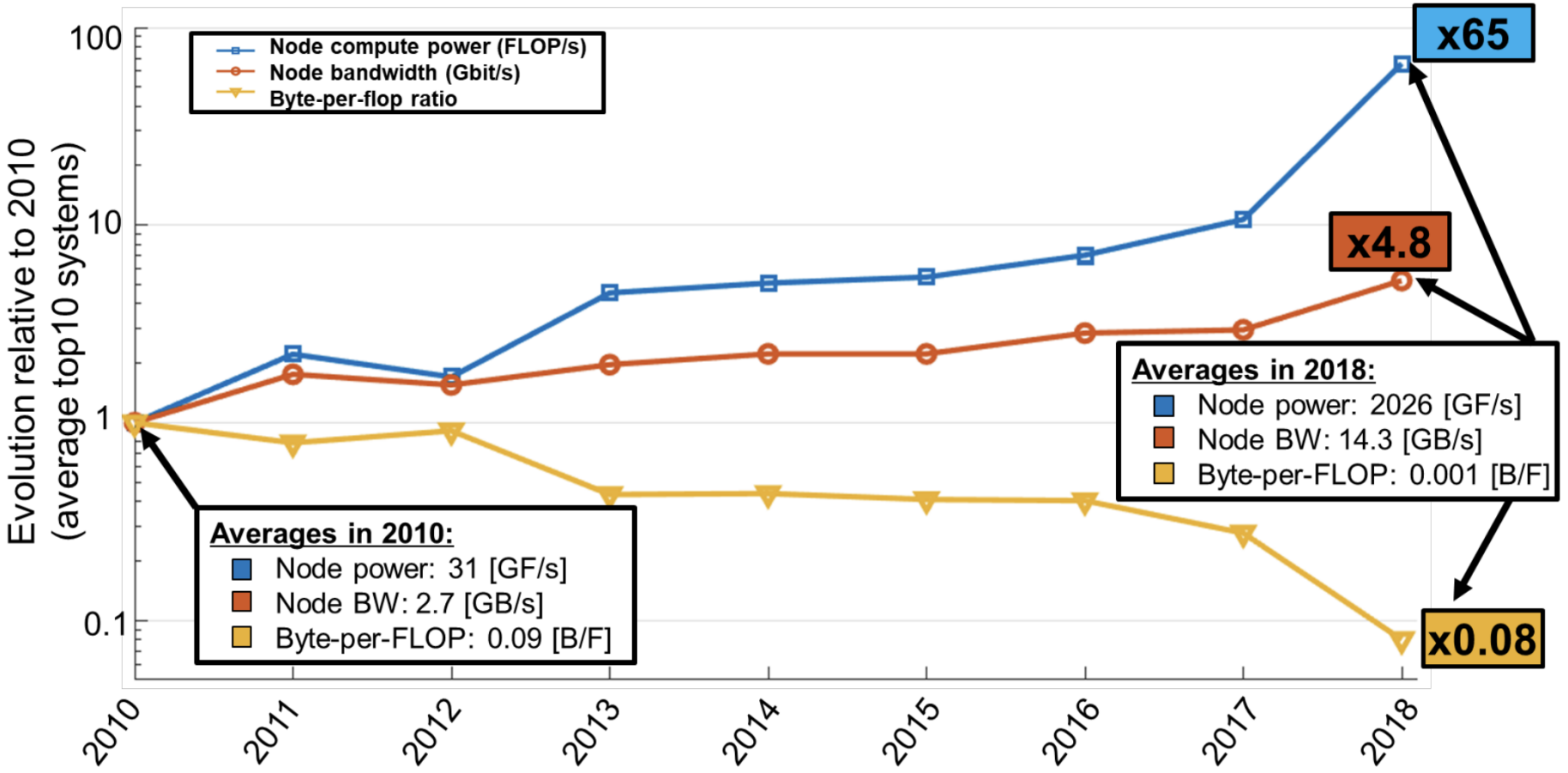
- Summit supercomputer at ORNL
  - Top performance in Linpack (top500.org results) with 122.3 PetaFLOPS
  - 13 MW ⇨ **13.9 GFLOPs / Watt**
  - 6 GPUs per node. 2 CPUs


- Next challenge: Exascale computing within 20 MW
  - 50 GLOPs / Watt

TOP500 The List.

# Communication Challenge: Top 10 System Trends

## Performance/Communications Trends for Top 10 (2010-2018)




Sunway TaihuLight (Nov 2017) B/F = 0.004; Summit HPC (June 2018) B/F = 0.0005 → **8X decrease**

Keren Bergman, "Empowering Flexible and Scalable High Performance Architectures with Embedded Photonics", IPDPS 2018

# Communication Energy Challenge

- 14 GFLOPs / Watt (Summit) ⇨ 72 pJ / FLOP
  - 0.36 pJ / bit
- Exascale target: 50 GLOPs / Watt ⇨ 20 pJ / FLOP
  - 0.1 pJ / bit
- **Total communication budget**

- The above assume 200 bits / FLOP

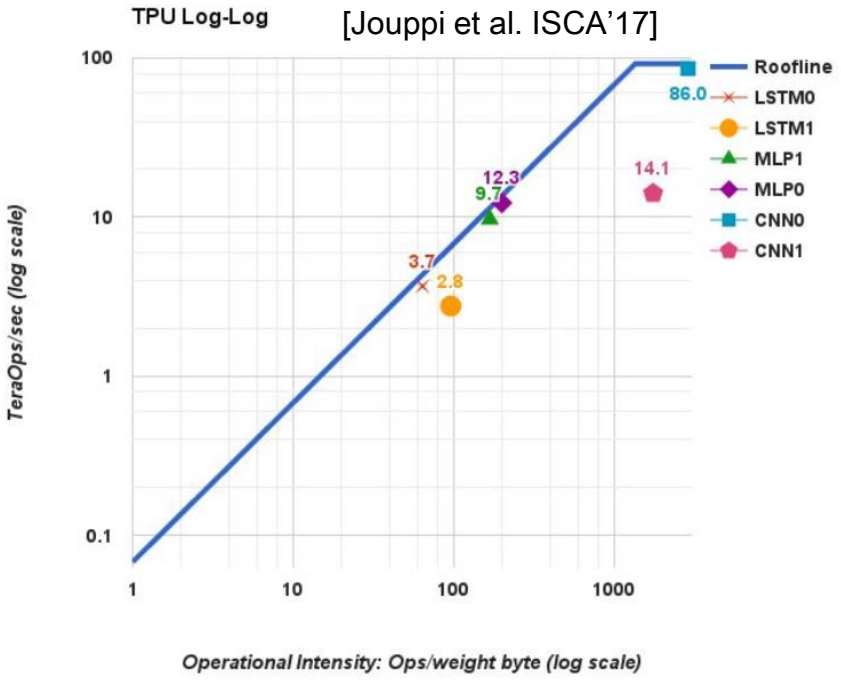
| Data Movement Energy: | |
|---|---|
| – Access SRAM | O(10fJ/bit) |
| – Access DRAM cell | O(1 pJ/bit) |
| – Movement to HBM/MCDRAM (few mm) | O(10 pJ/bit) |
| – Movement to DDR3 off-chip (few cm) | O(100 pJ/bit) |

# Result: Specialization May Be Limited By IO
# Google's TPU as an Example

- Dedicated hardware for DNNs
  - Peak compute capacity: 92 TOPS/s (8-bit precision)
  - Peak bandwidth: 34 GB/s
- Must reuse a byte **2706** times to fully exploit compute capacity
  - Operational intensity: 2.7KOPs/byte, hit rate: 99.96%, 0.003 bit/OP
- Only **two** operations have high operational intensity: CNN0 and CNN1
- Operational intensity of others (e.g., translate and Rankbrain which are **90%** of the applications) are **1 – 1.5** orders of magnitude smaller
- LSTM0 would require **40x** more bandwidth to (theoretically) allow full TPU utilization

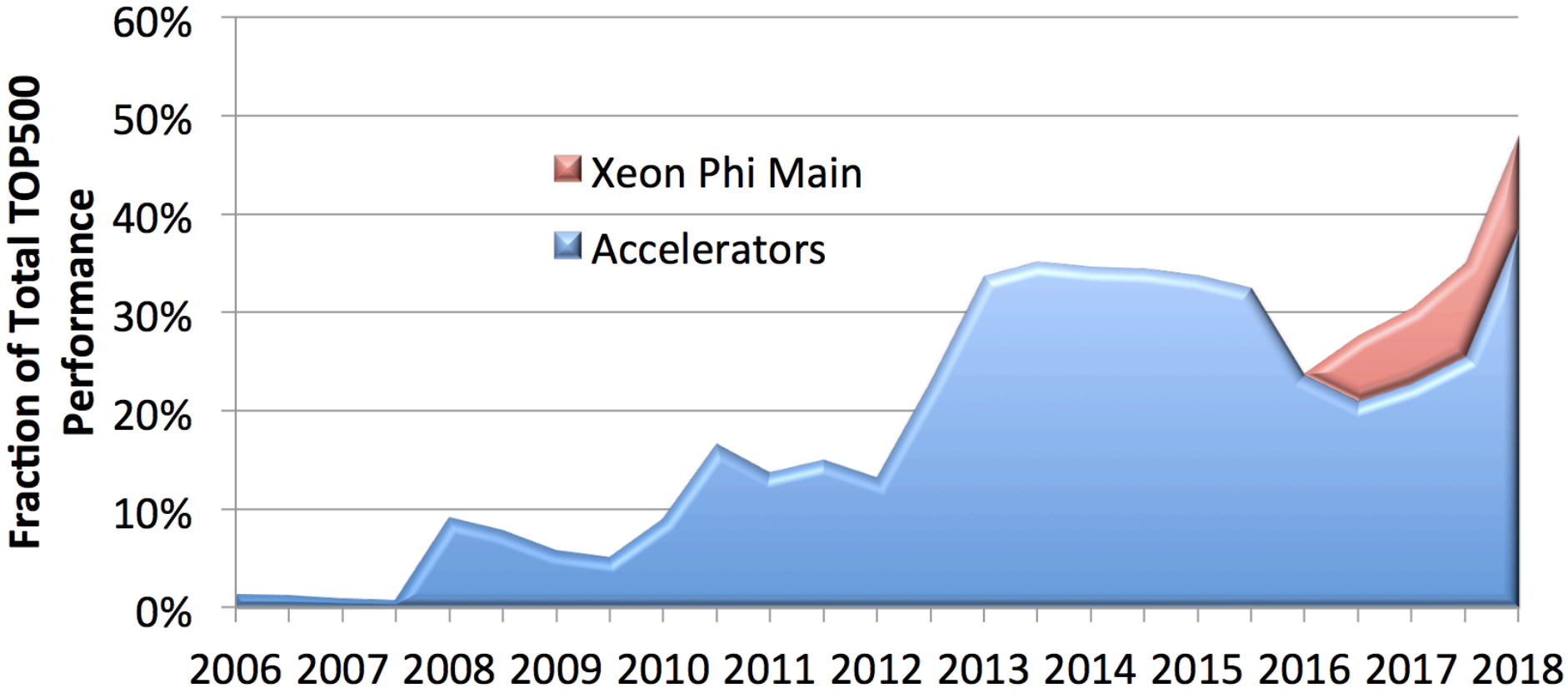
[Google cloud]

[Jouppi et al. ISCA'17]




[Keren Bergman]

# Specialization is Increasing

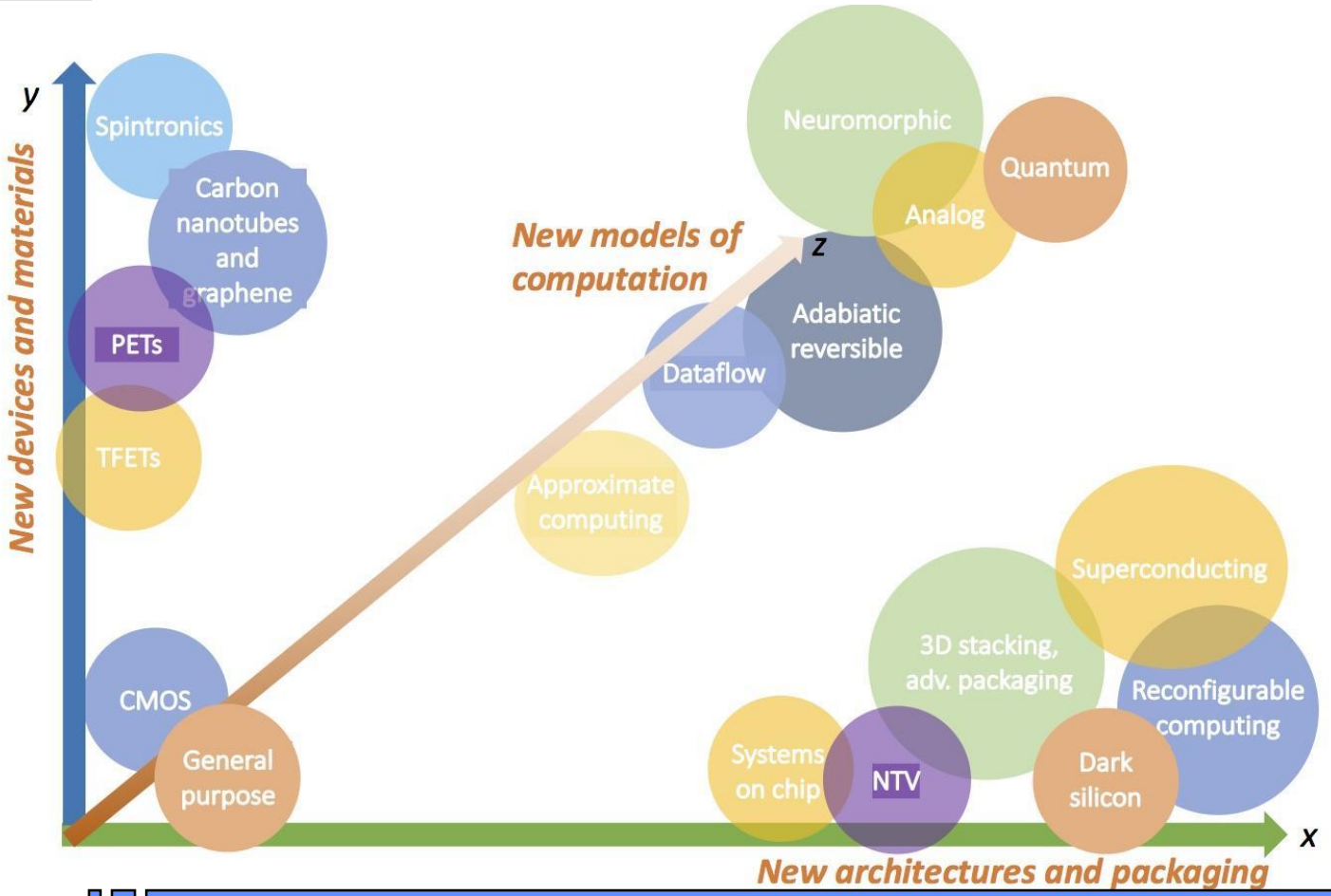



TOP 500
The List.


# Preserve Computational Performance Scaling

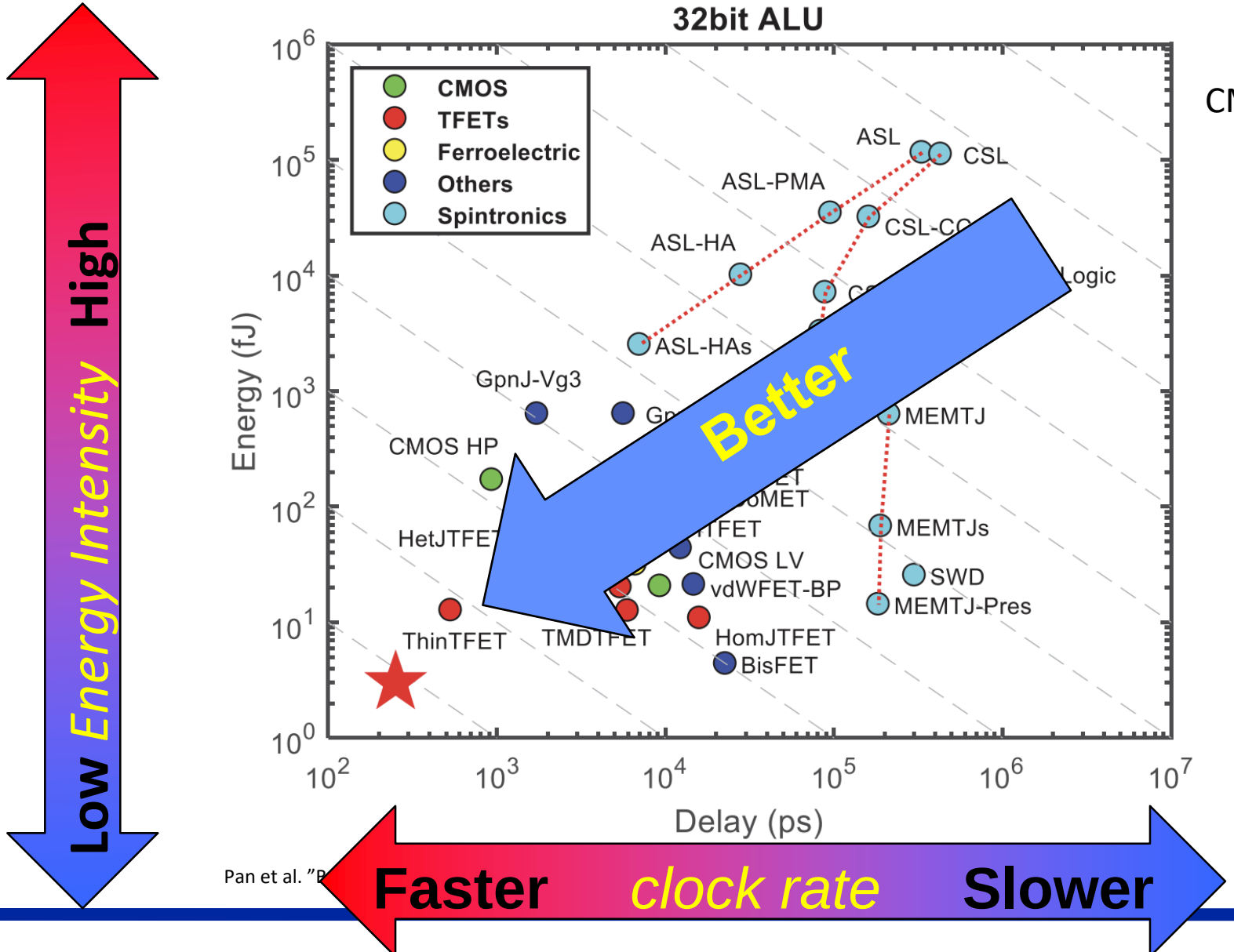
# Long- and Short-Term Solutions

**New Materials and Devices**
20+ years (10 year lead time)




**More Efficient Architectures and Packaging**
The next 10 years after exascale

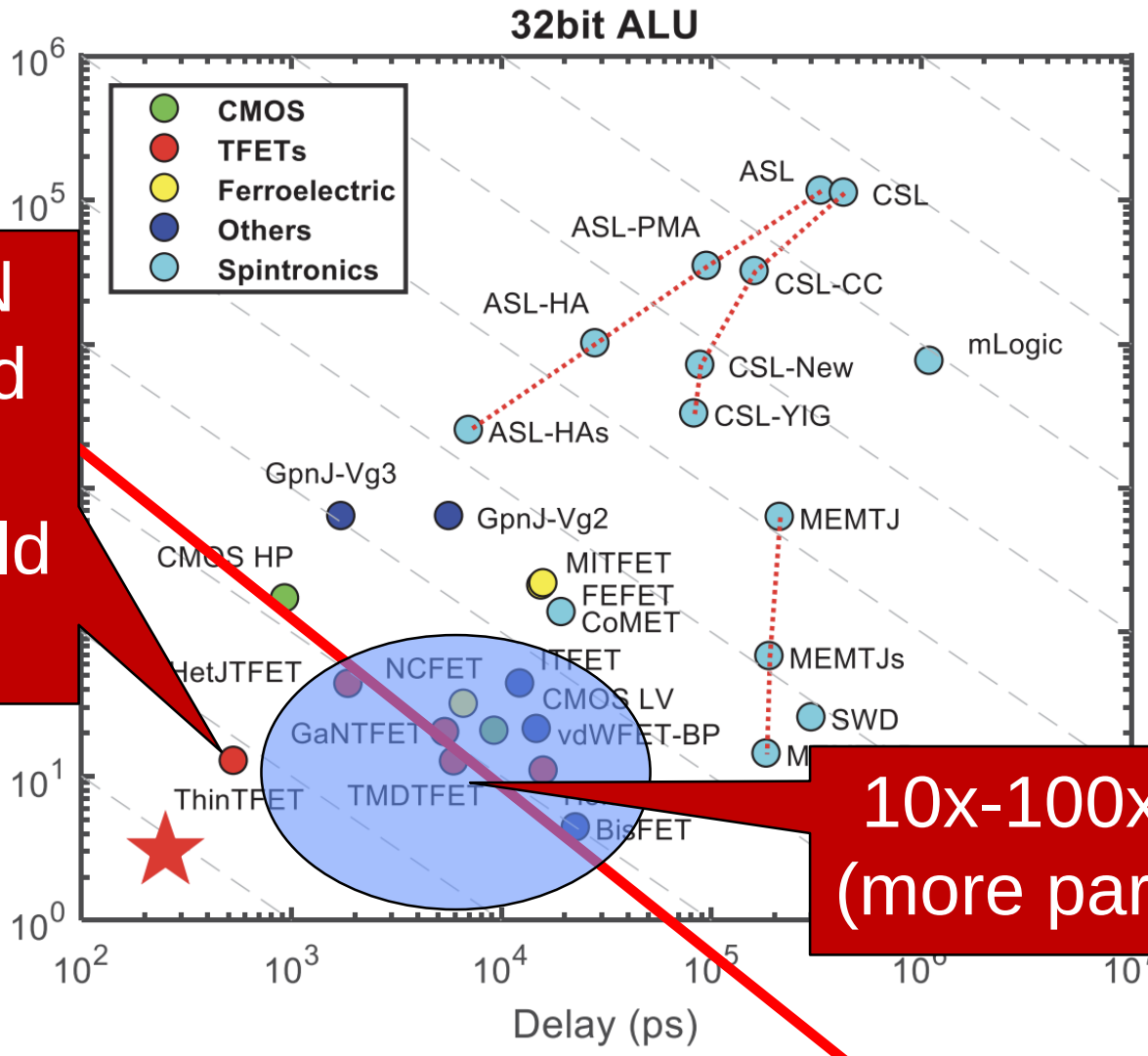
# Comparing CMOS Alternatives




CMOS is 15nm (ITRS)

Pan et al. "P

# Have to Adapt to New Devices



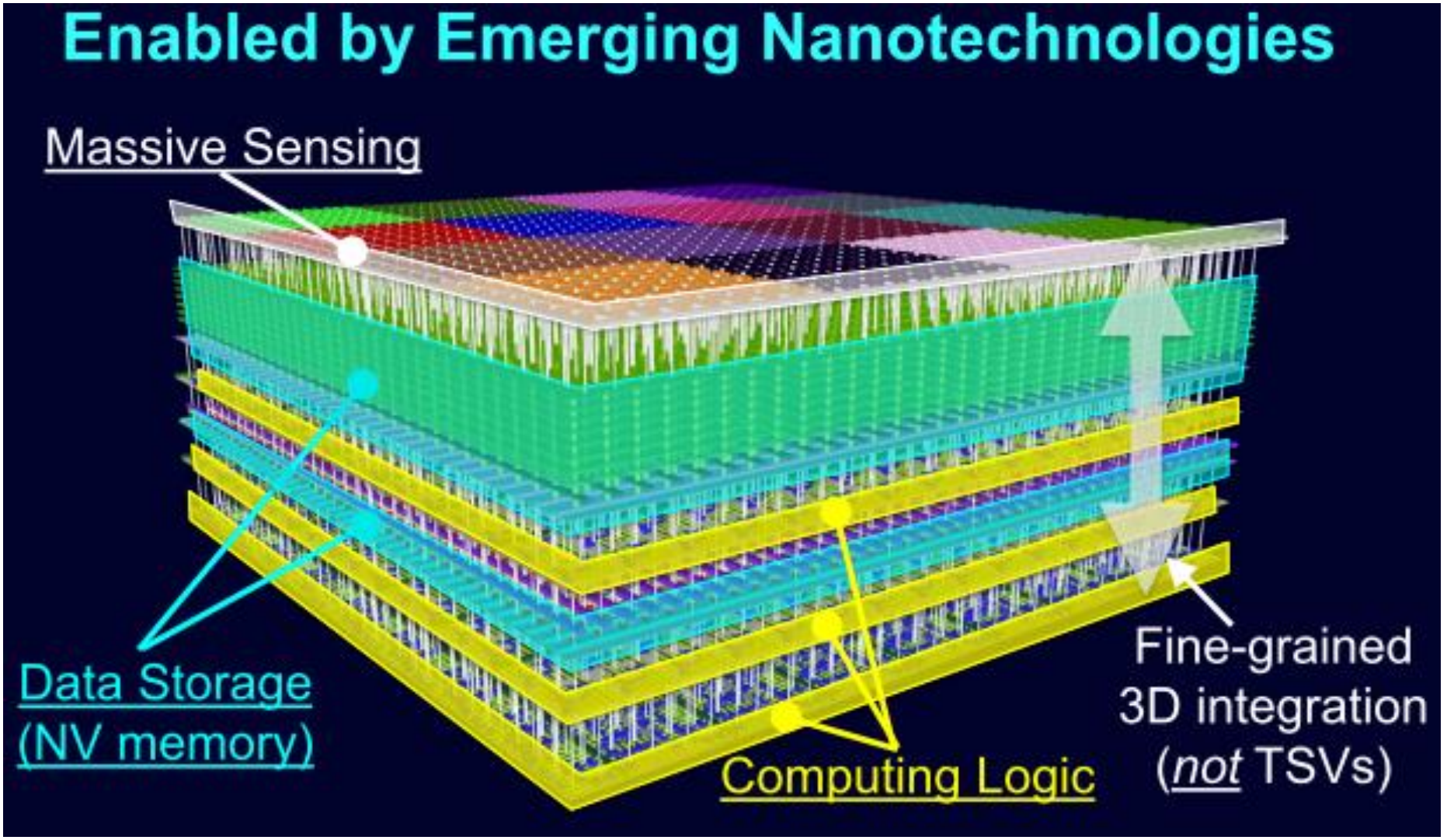

Strong ON current and steep subthreshold slope

CMOS is 15nm (ITRS)

10x-100x slower (more parallelism)

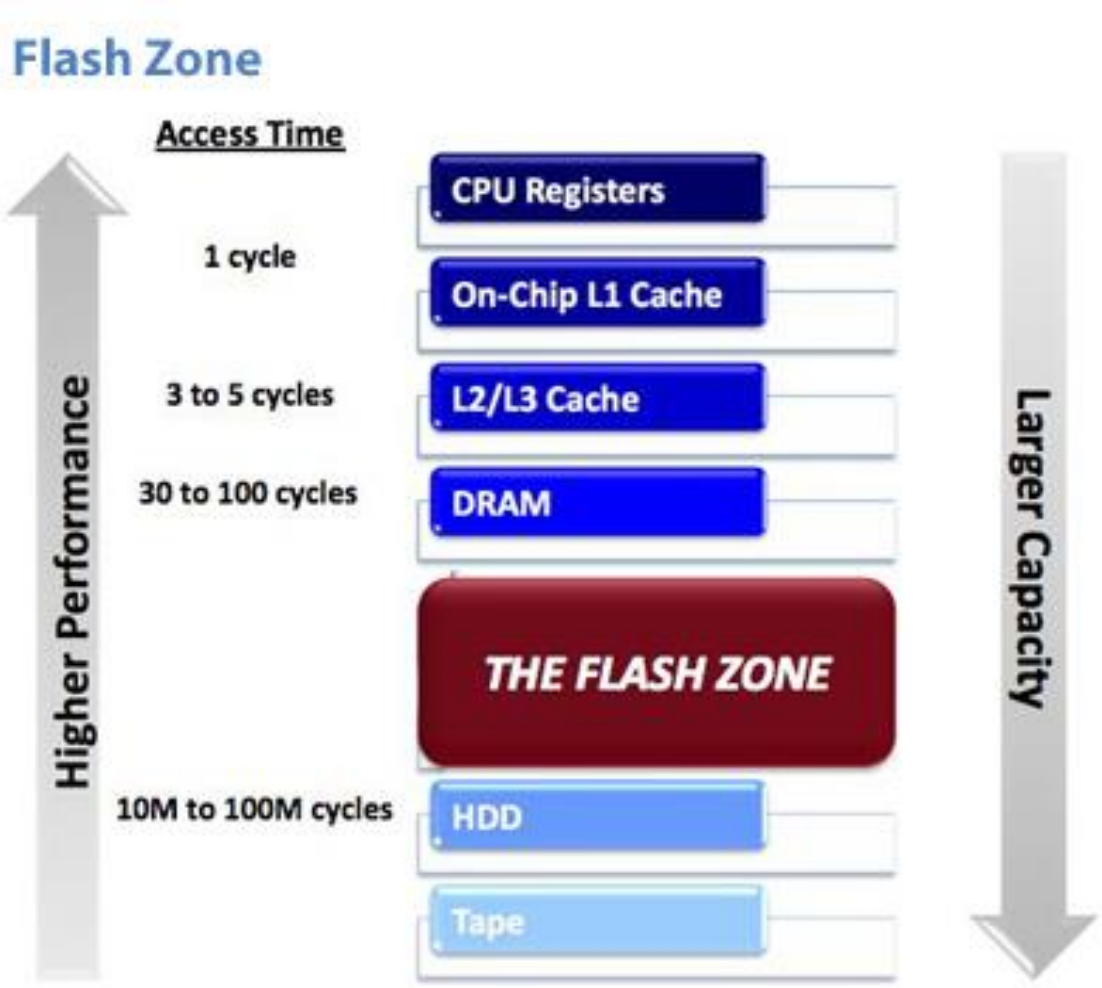
Pan et al. "Beyond-CMOS device Benchmarking for Boolean and Non-Boolean Logic Applications.," Arxiv, 2017.

# 3D Integration of Tomorrow



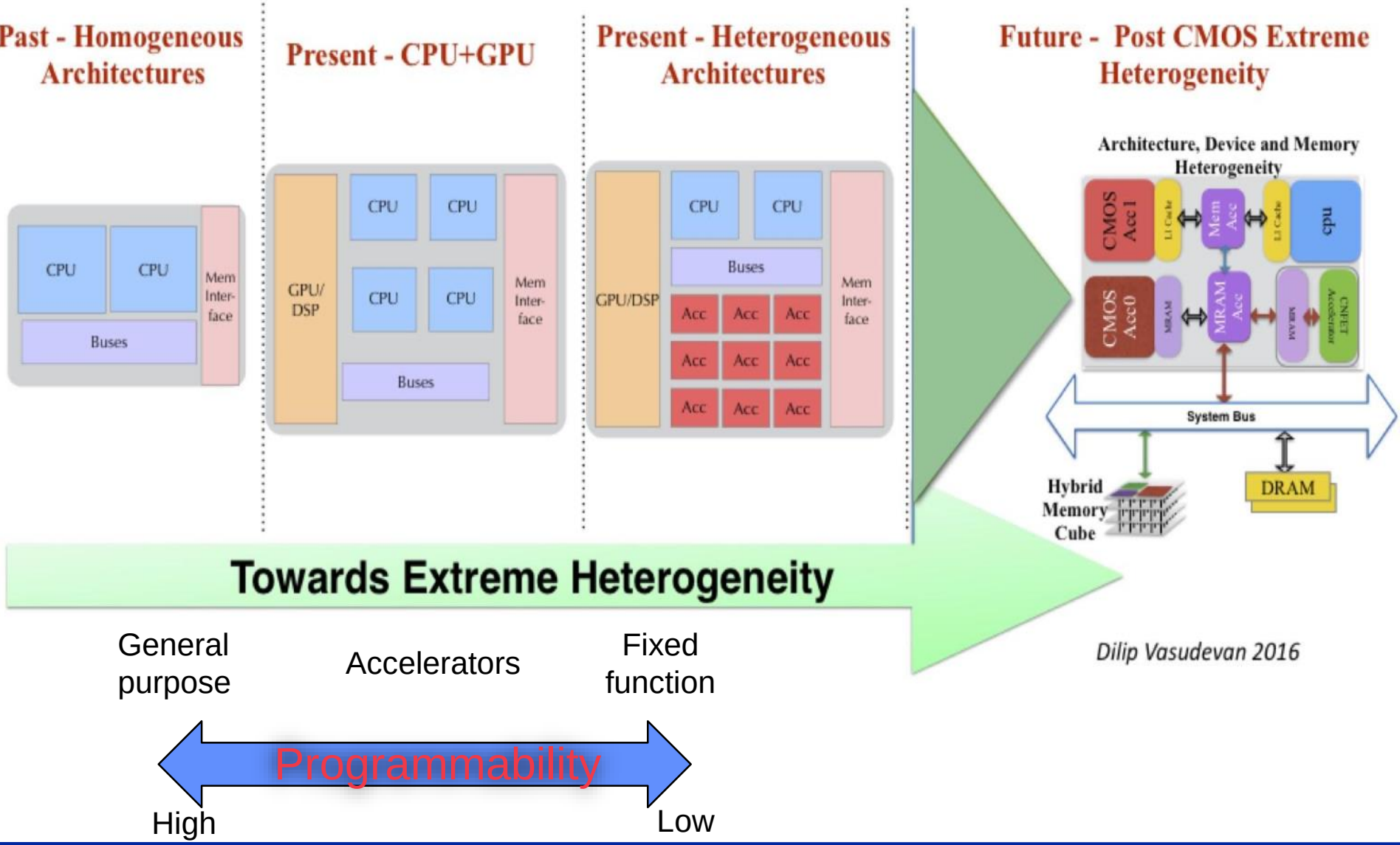
# What About Memory Hierarchy?

- Non-volatility higher at the hierarchy
  - Challenge assumption that non-volatile storage is slow and distant
- New memory hierarchy likely different




AGIGARAM “The Flash Zone”

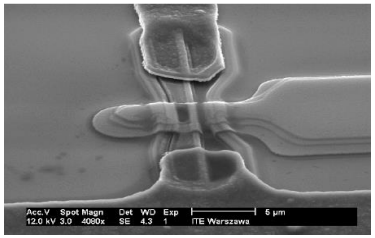
# Towards Diverse Accelerators



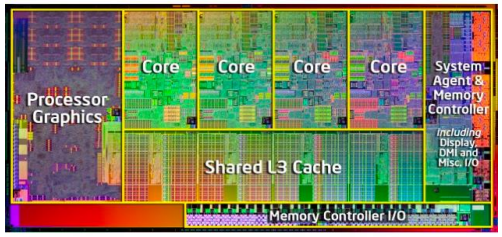
# Problem Statement: Evaluate At Architectural Level

- Evaluating each option in isolation misses the big picture
  - Devices can be better designed with high-level metrics
  - Architects can figure out how to best use new technologies
  - Software experts can assess impact to programmability and compilers

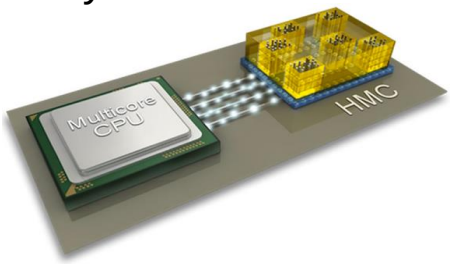
Transistor/Devices



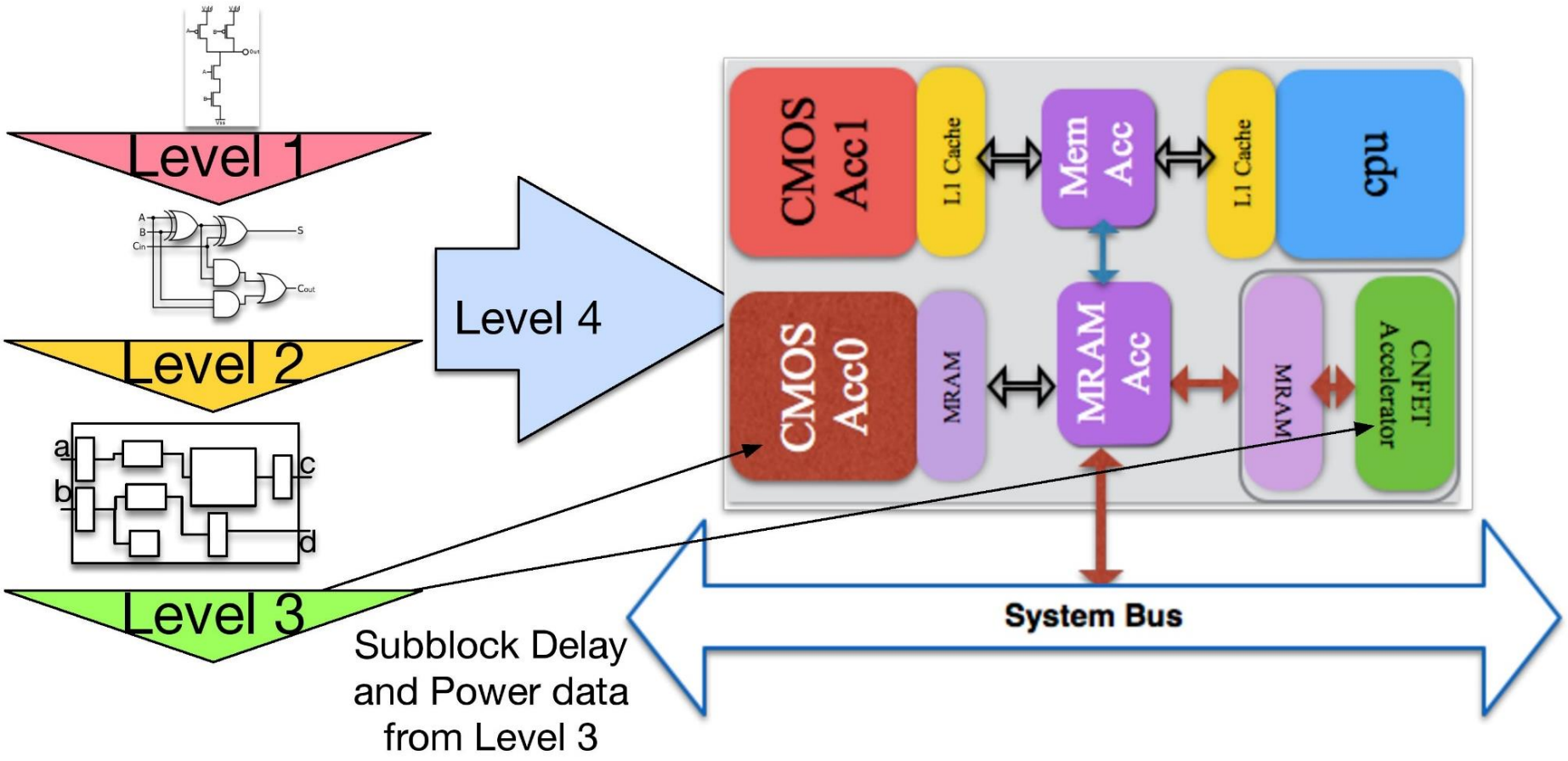
Architecture



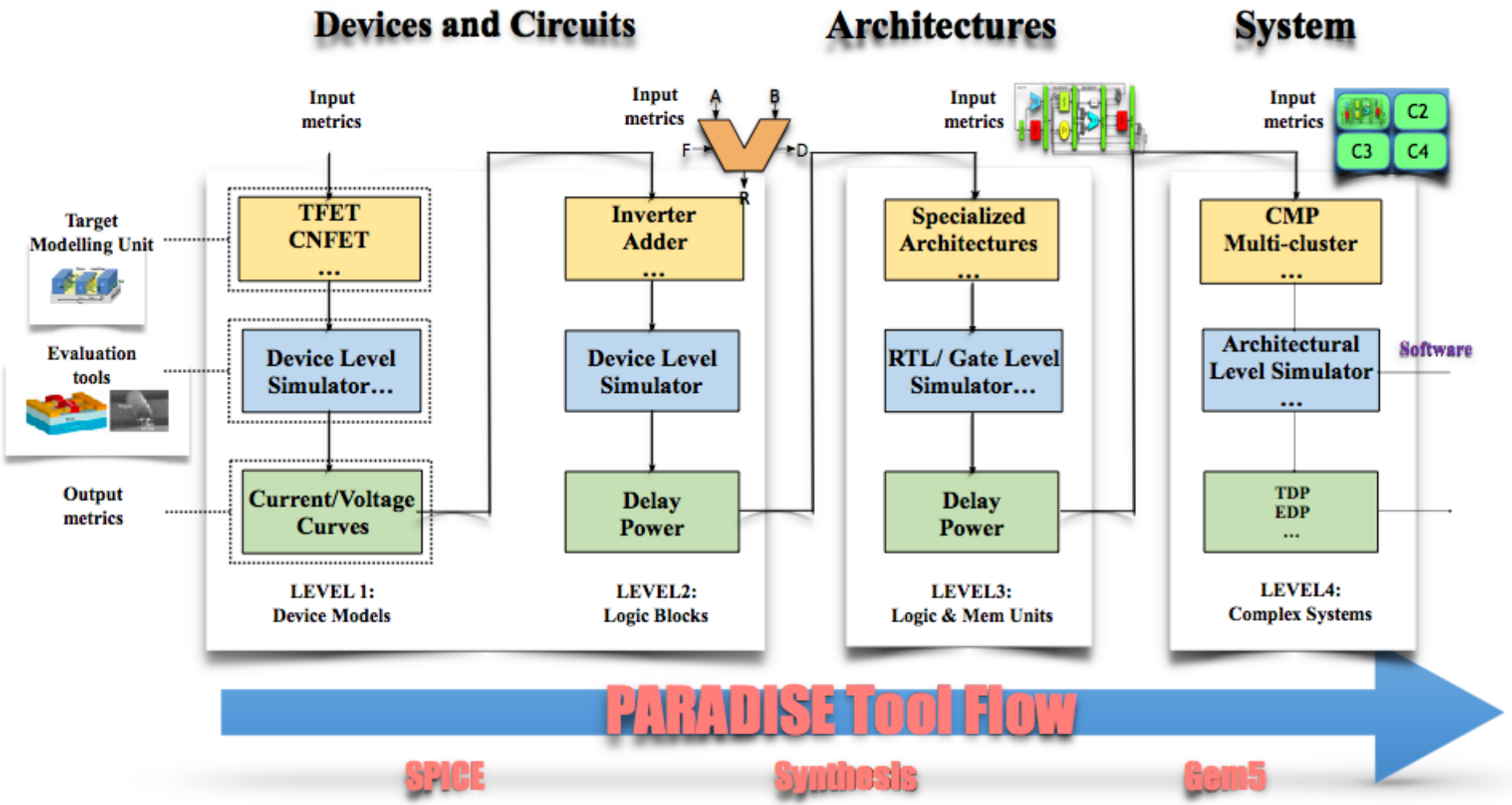

System



# Multi-Level Architectural Simulation

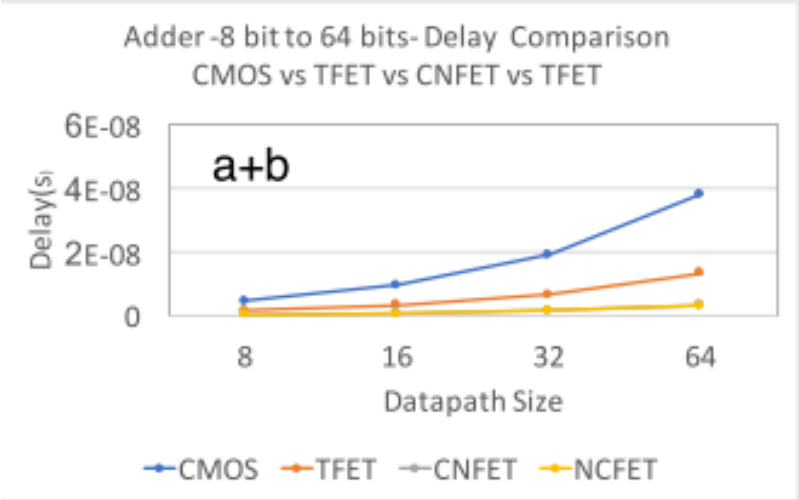


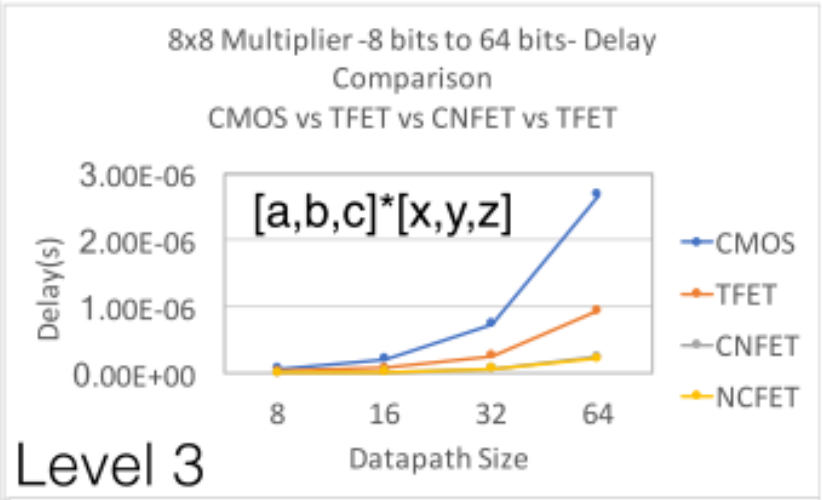
# PARADISE End-To-End Tool Flow



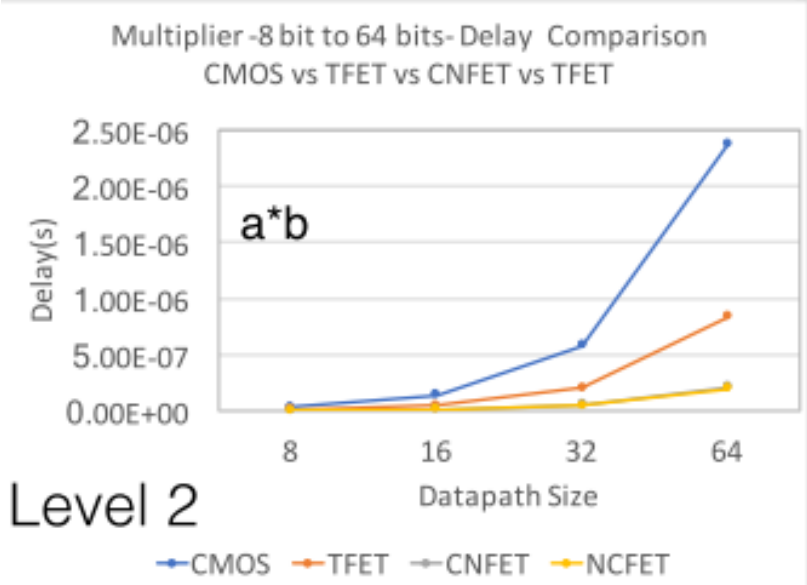
# Comparison Studies (PARADISE generated)

## Level 1

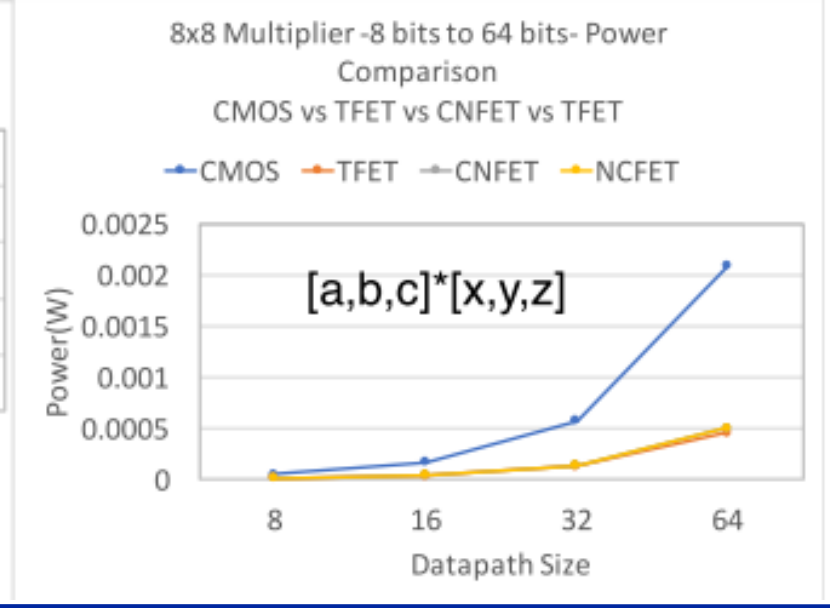


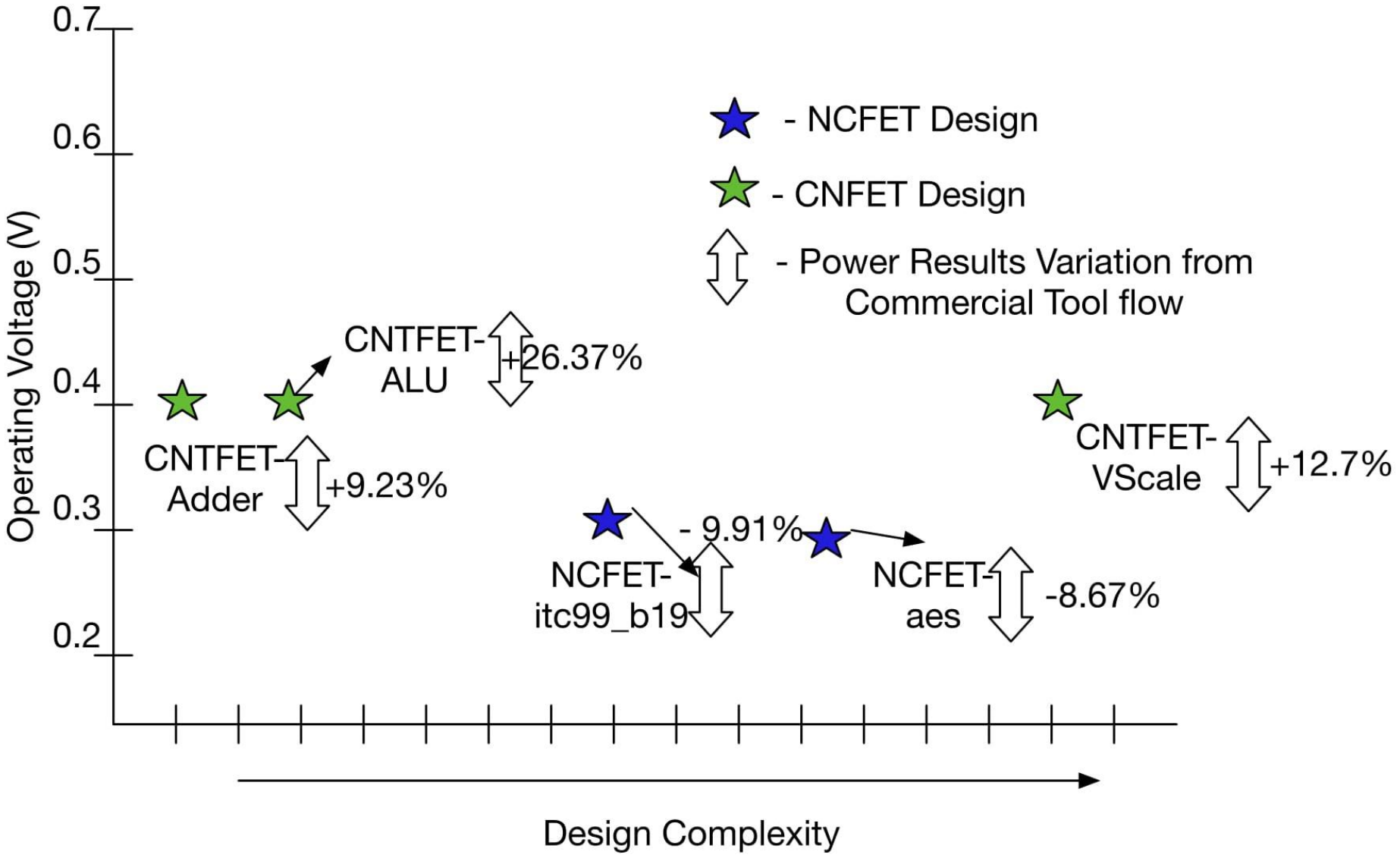

## Level 3



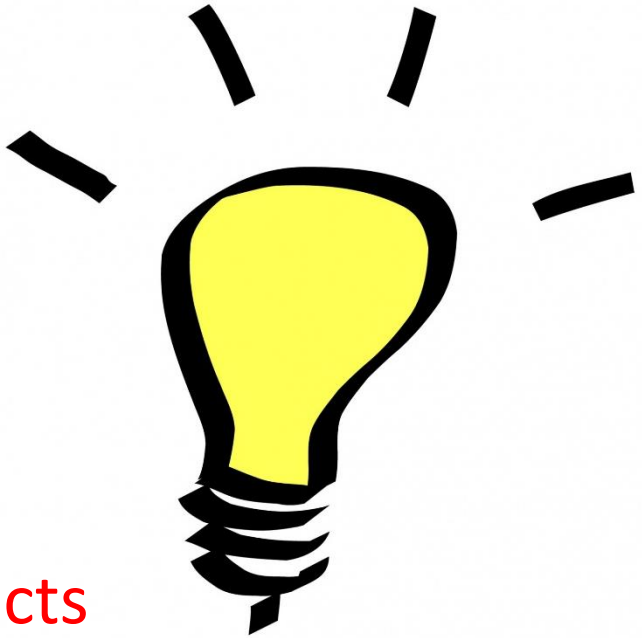

## Level 2



# Design Space Exploration at RTL Level



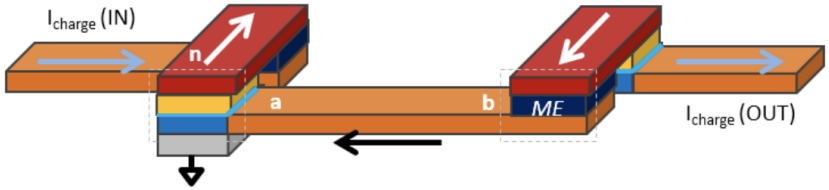
## How To Use These Tools?

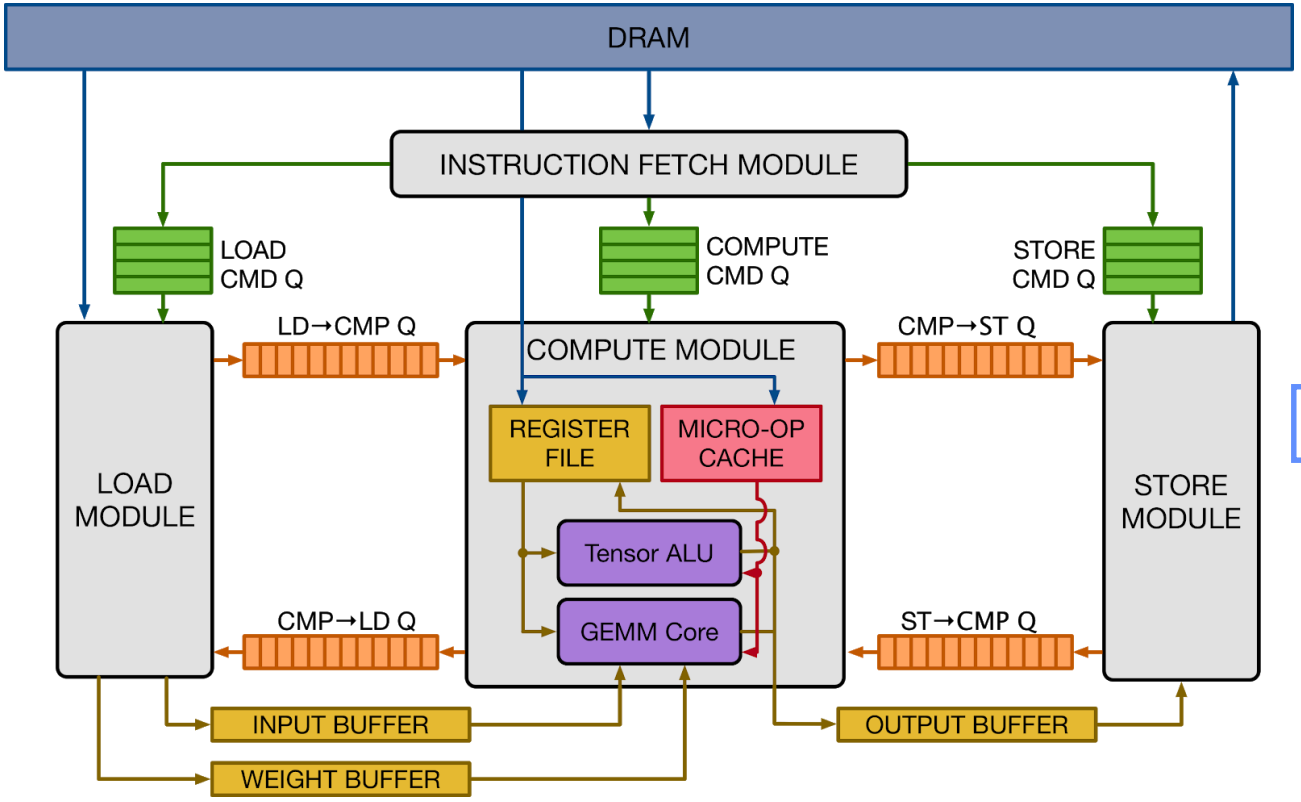


Three ongoing projects

# VTA Core + MESO Deep Learning

- Deep learning acceleration with a magneto-electric spin-orbit (MESO) logic device




210 TOPS/W

MESO: 10x to 30x lower switching energy
5x higher logic density

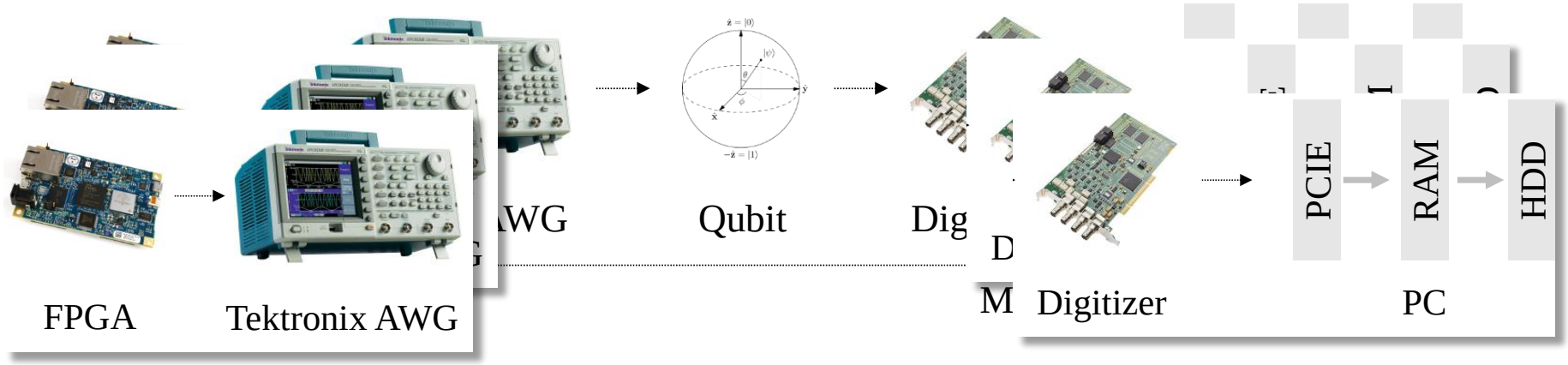
Moreau et al, "VTA: An Open Hardware-Software Stack for Deep Learning". Cornell University, 2018
Manipatruni et al, "Scalable Energy-Efficient Magnetoelectric Spin-Orbit Logic", Nature, 2019

# (2) Quantum Control Processor

✶ $Quantum\ Computer = Quantum\ PU + Control\ Hardware$

Off the shelf and high cost

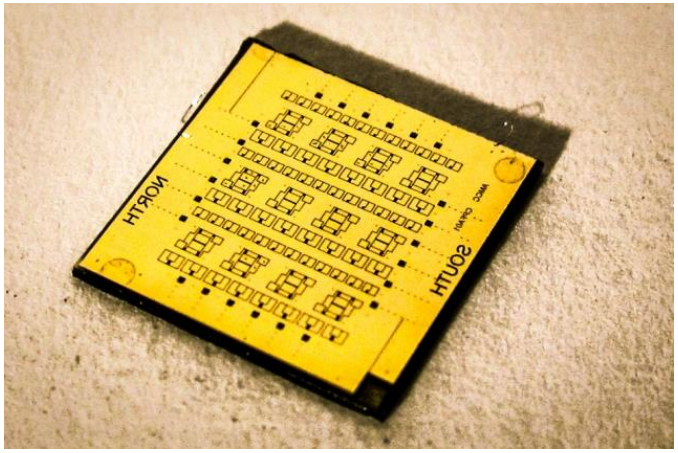
Large amount of data and slow speed



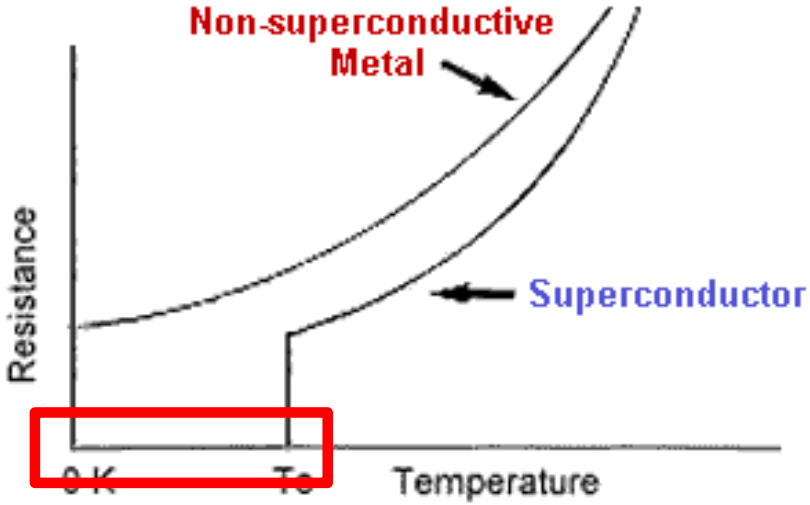

*1000 qubits,*
*gate time 10ns,*
*3 ops/qubit*
***300 billion ops per second***

# (3) Superconducting Logic

- Resistance drops to zero
  - Tc approx 4 Kelvin
- 100's of Gigahertz
  - Deep pipelines
- Memory is a grand challenge
- Can measure architecture impact and synergy with memory technologies



MIT News



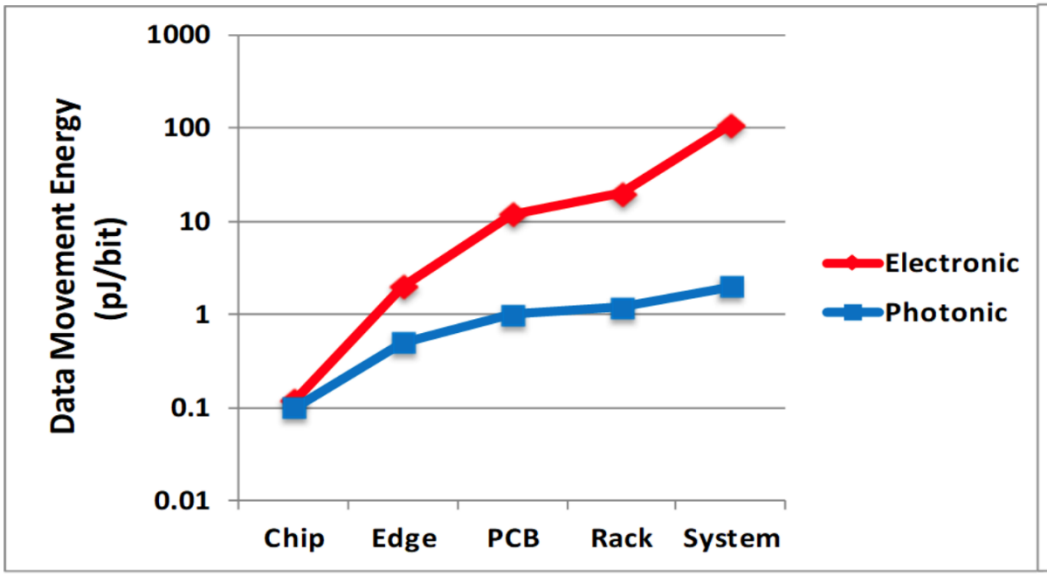

Gallardo et al, "Superconductivity observation in a (CuInTe 2 ) 1-x (NbTe) x alloy with x=0.5"

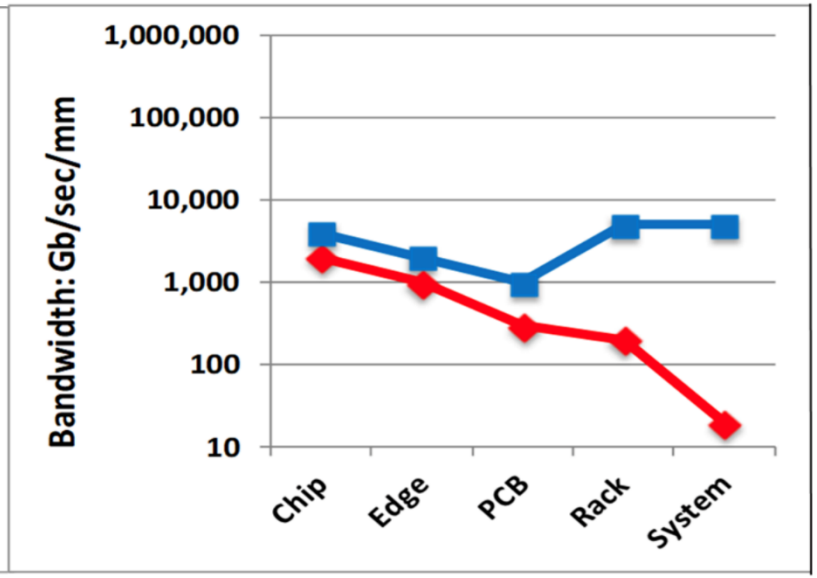

## Preserve Communication Scaling

To avoid making it the limiting factor

# The Photonic Opportunity

## The Photonic Opportunity for Data Movement




**Reduce Energy Consumption**

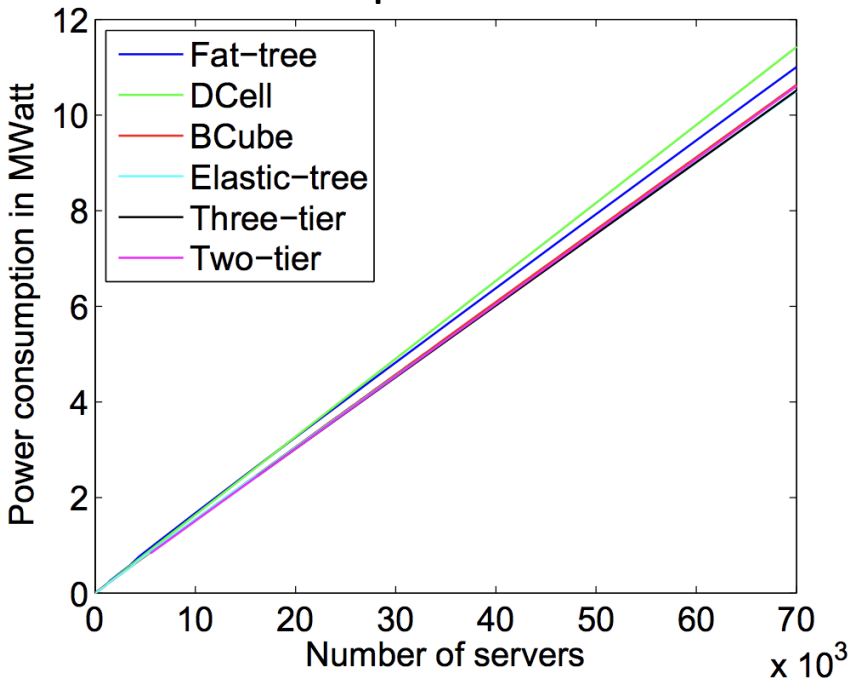
**Eliminate Bandwidth Taper**

R. Lucas et al., "Top ten exascale research challenges," DOE ASCAC subcommittee Report, 2014

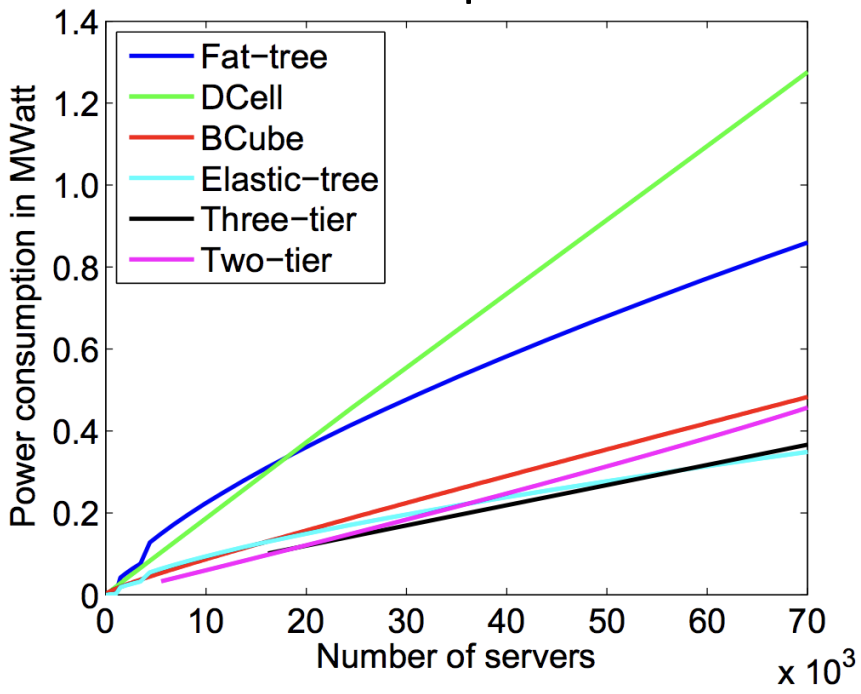
# Drop-In Replacements Not Enough

- Even if we have a network that consumes no energy, we cannot reach a **2x** improvement
  - Only **4%** to **12%** of total power is in the network
- Key: use emerging photonic components to change the architecture

Total power




Network power

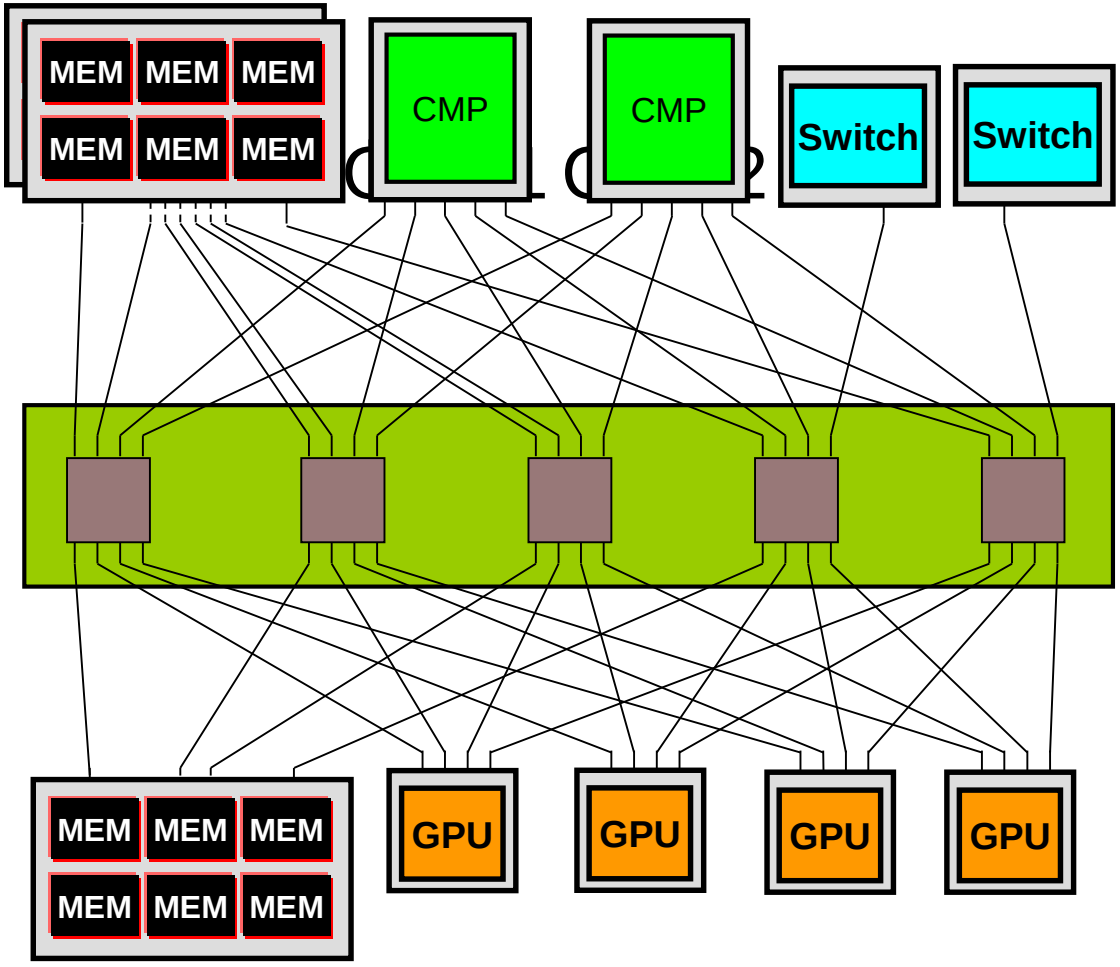



Rastin Pires et al, "Power Consumption Analysis of Data Center Architectures", GreeNets 2011

# Reconfigurability

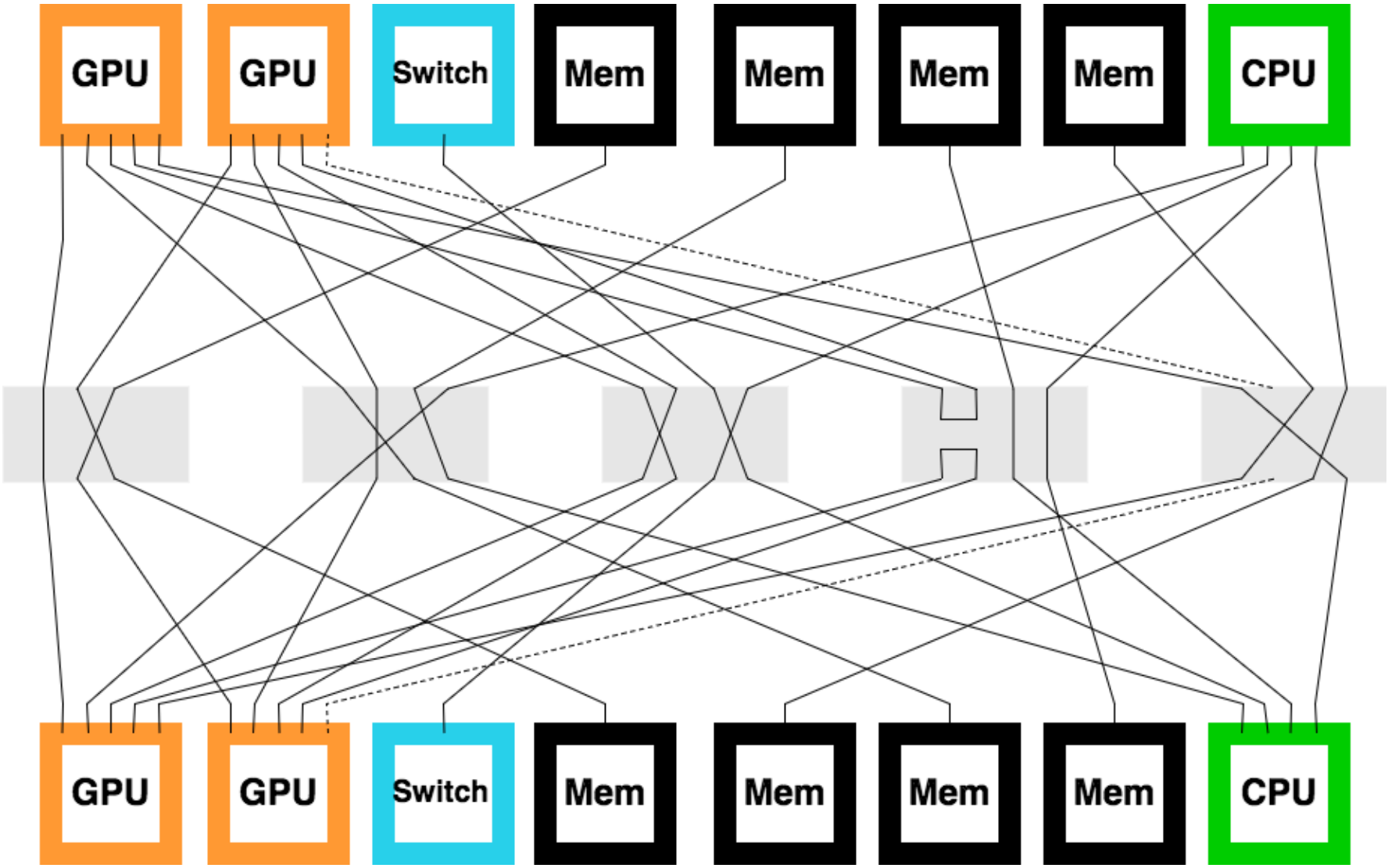
- Use capabilities of photonics to change the architecture
- Intra node
  - Resource disaggregation
- System-wide
  - Bandwidth steering

# Optical Switches on Nodes

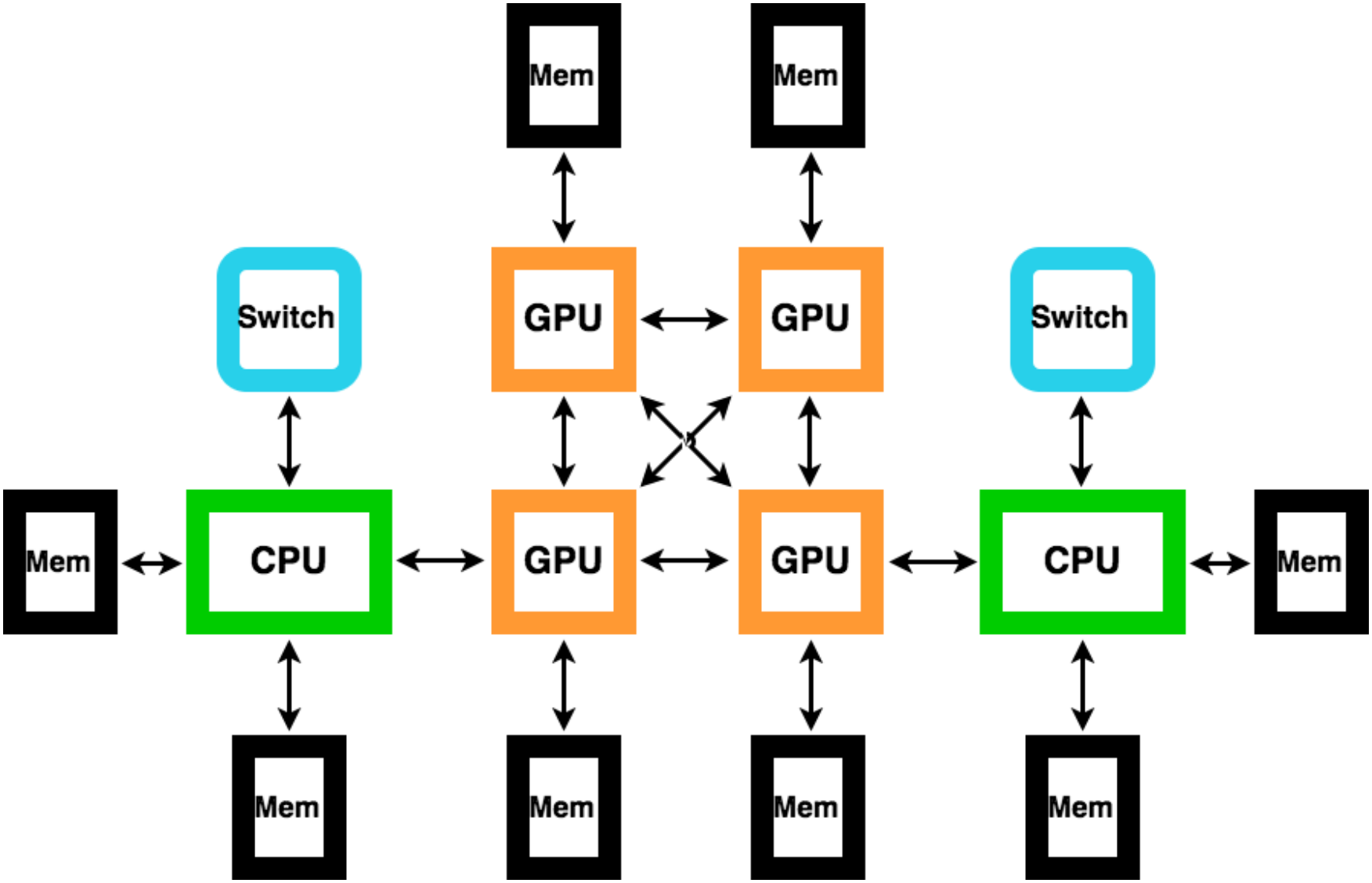



Keren Bergman, “PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability”, OI 2018

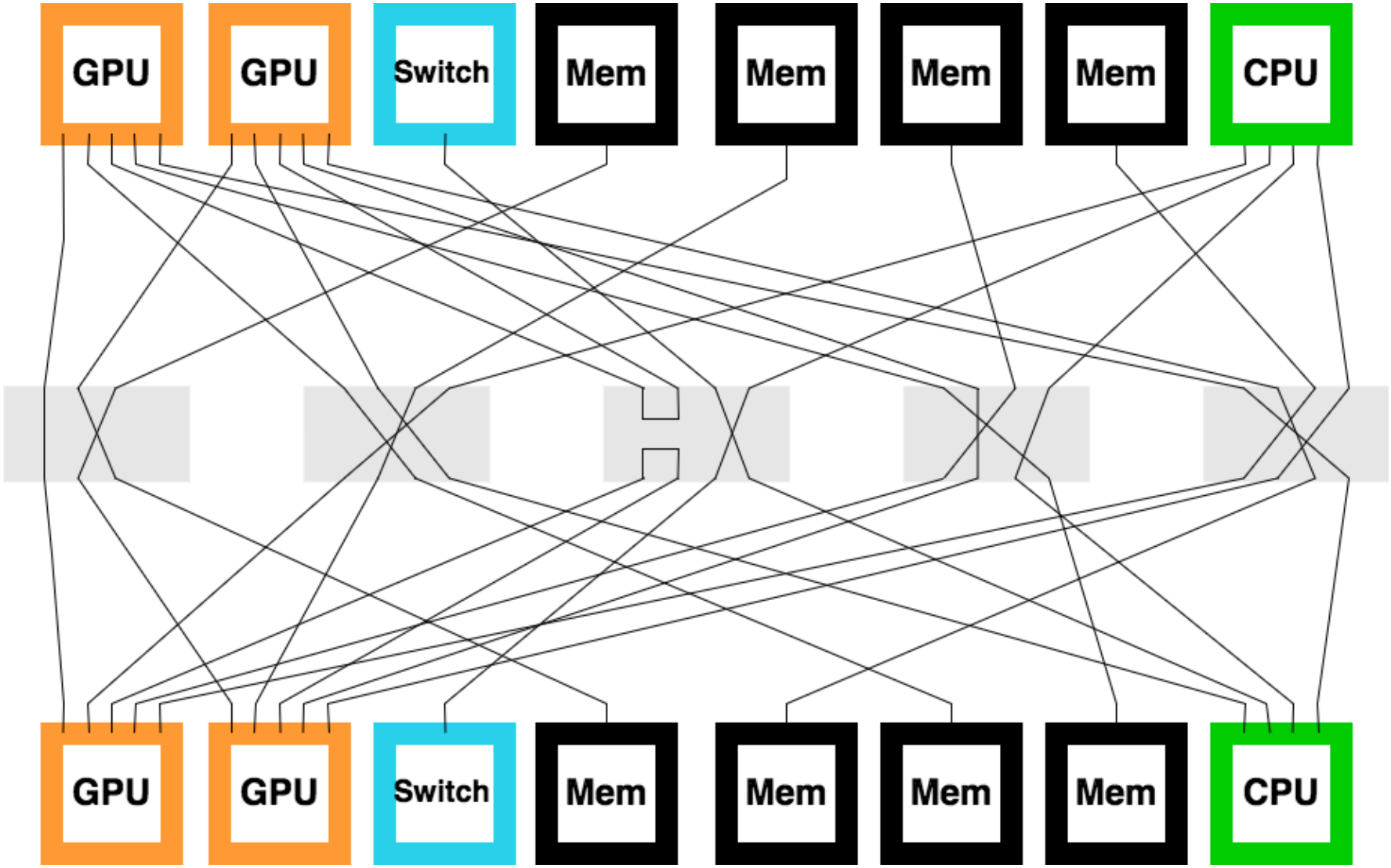
# Intra-Node Reconfigurability



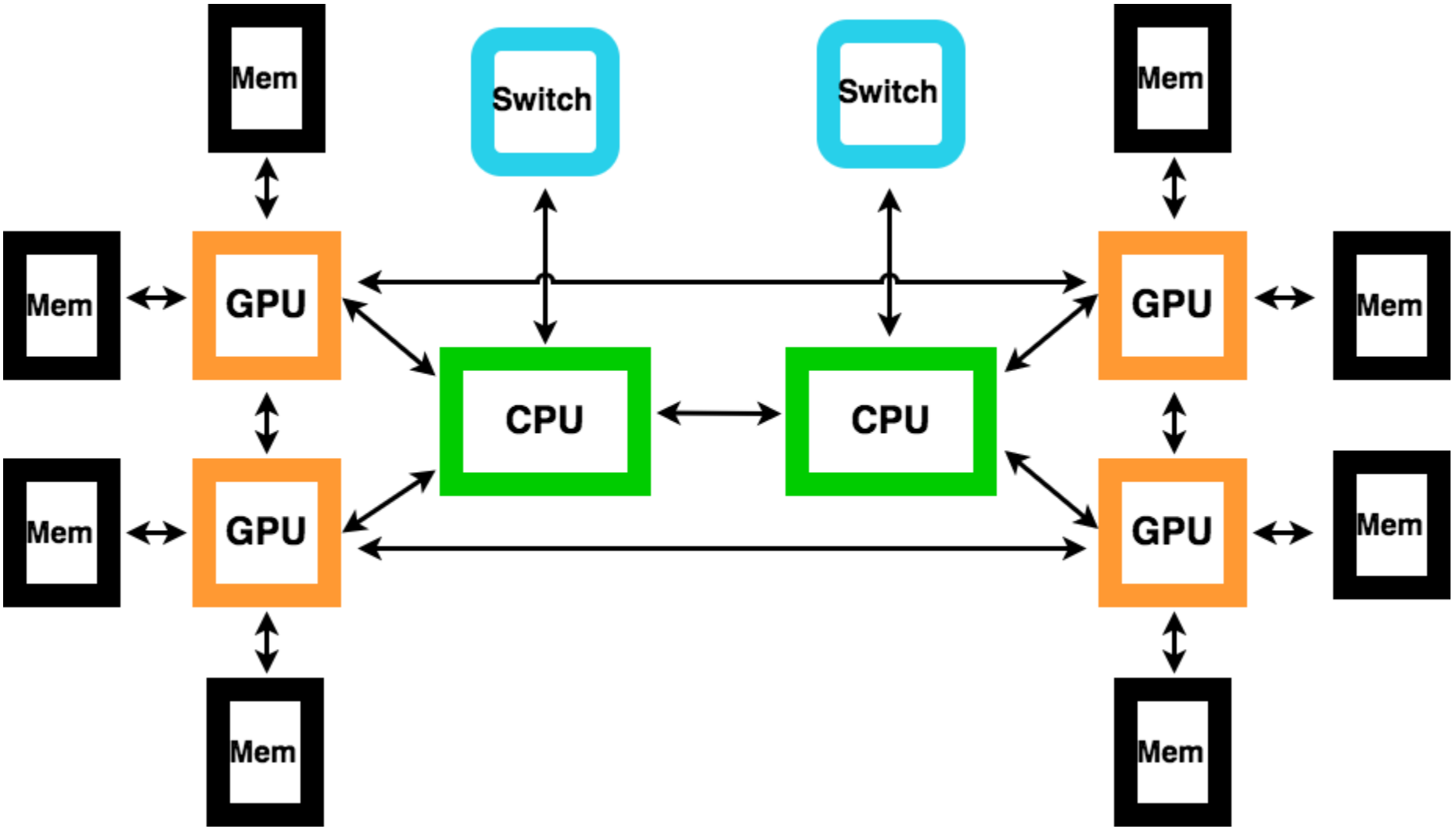
# Intra-Node Reconfigurability



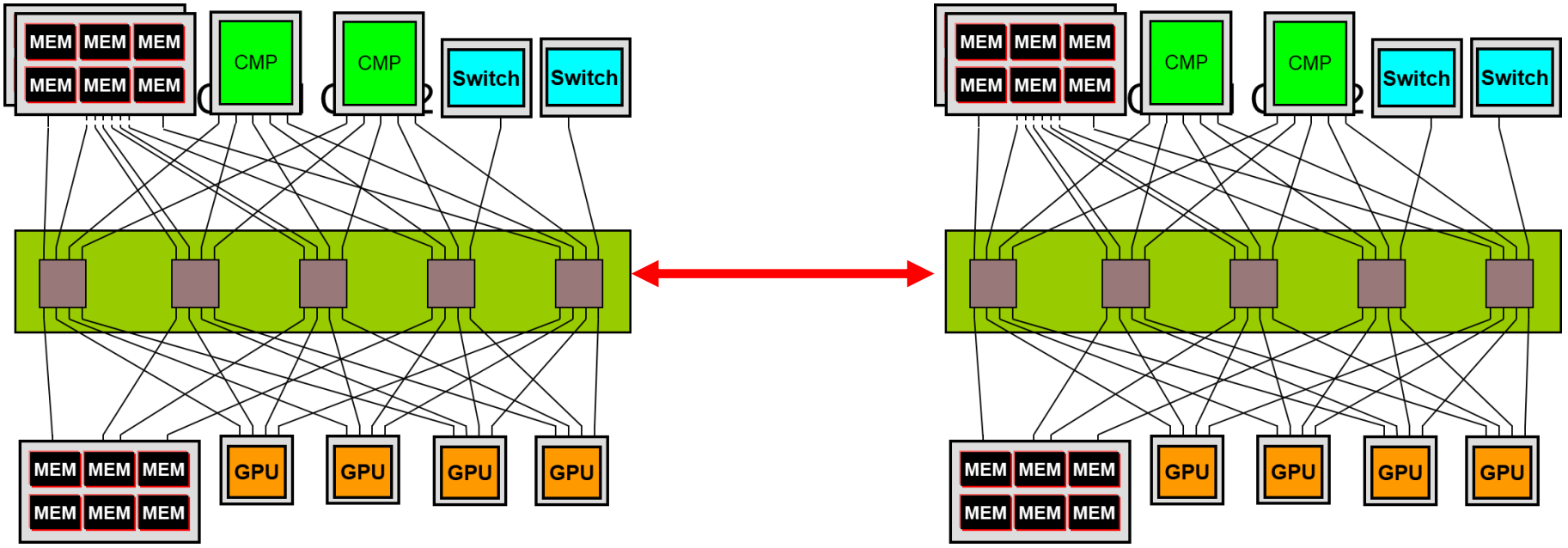
# Intra-Node Reconfigurability



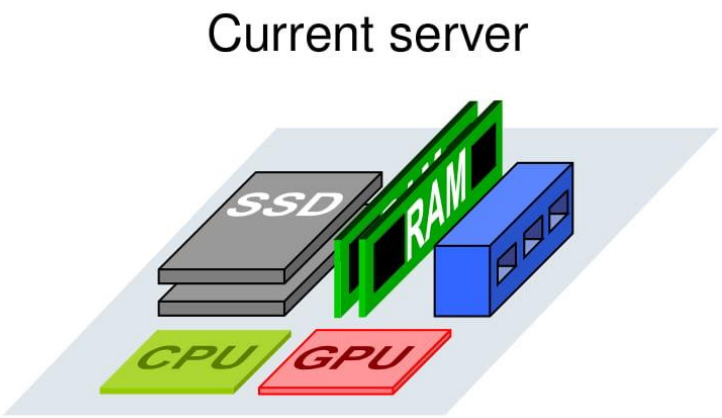
# Intra-Node Reconfigurability

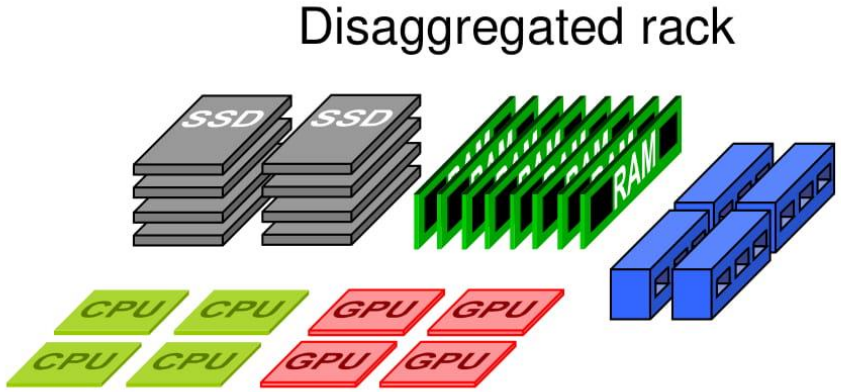


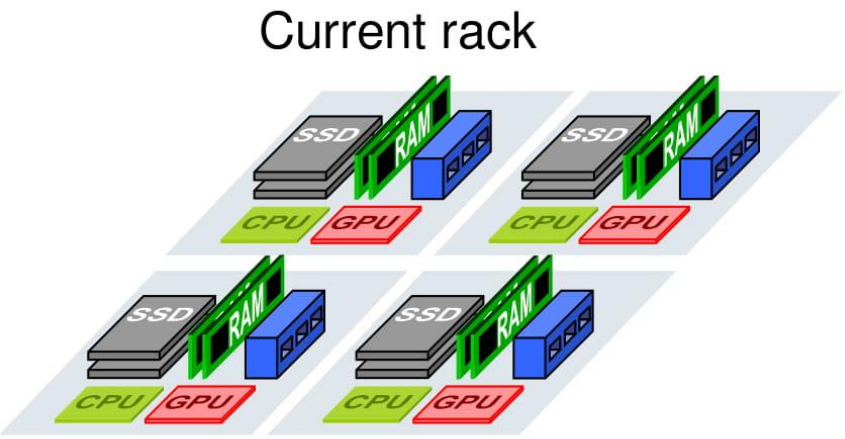
# If Connections Span Nodes

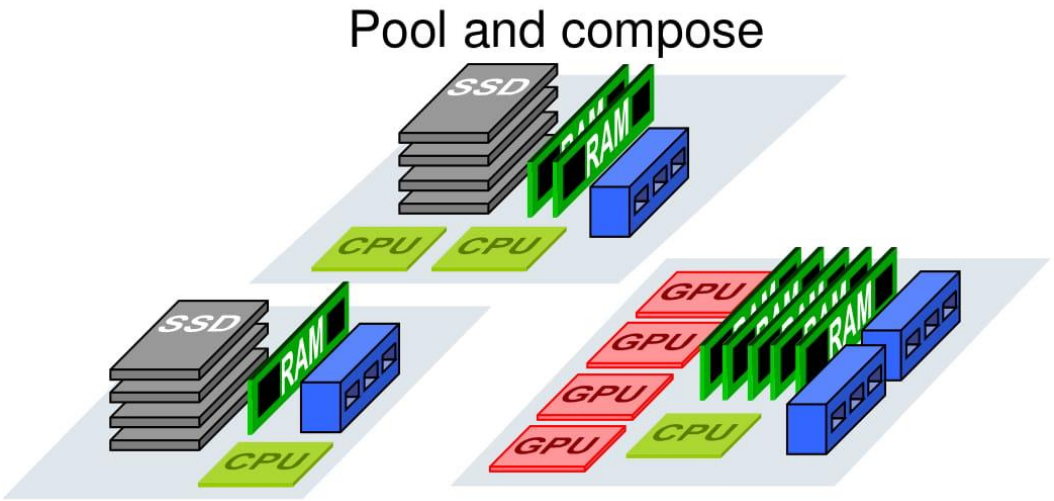


# Aggregate Remote Resources



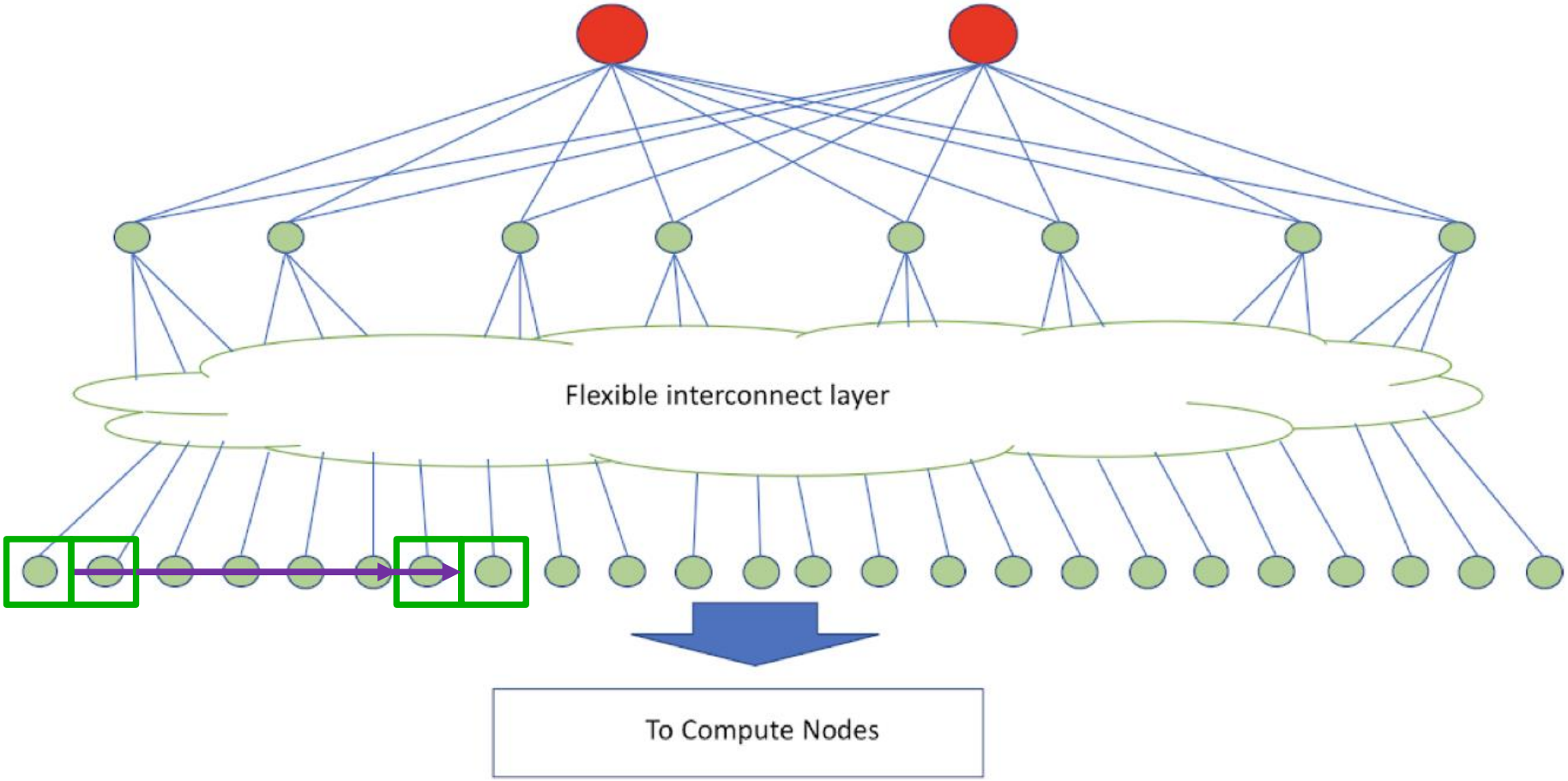

Keren Bergman, “PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability”, OI 2018

# Node Reconfigurability Challenges

- Photonic switches with sufficient radix
- Efficient conversion to optics
  - In package?
- Algorithm to decide node configuration
- How changing node configuration affects network traffic, scheduling, and system management [1]

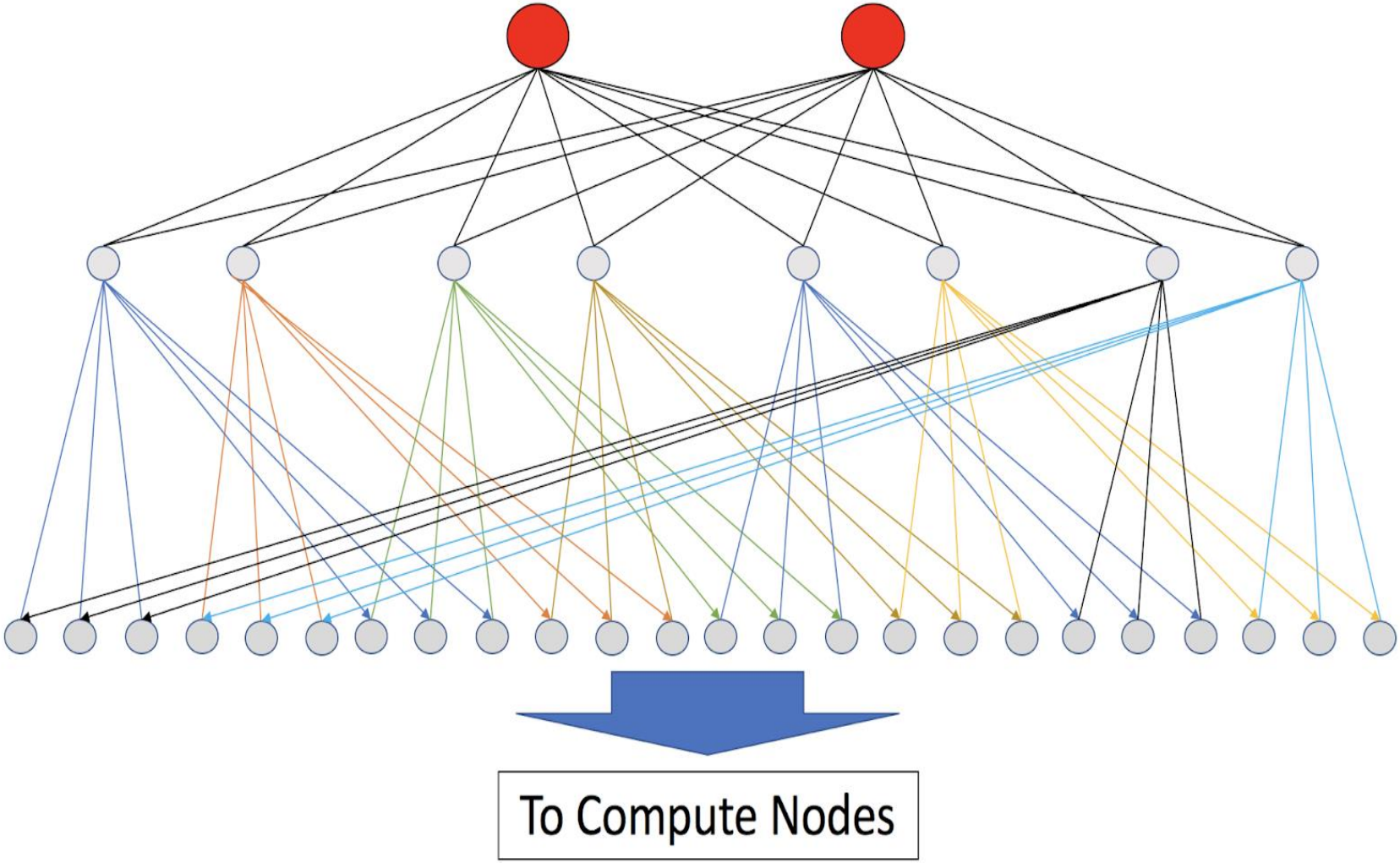
[1] D. Z. Tootaghaj et al., “Evaluating the combined impact of node architecture and cloud workload characteristics on network traffic and performance/cost,”, 2015 IEEE International Symposium on Workload Characterization.

# Use Optics for Efficient B/W Steering




[Min Yee (Jason) The]

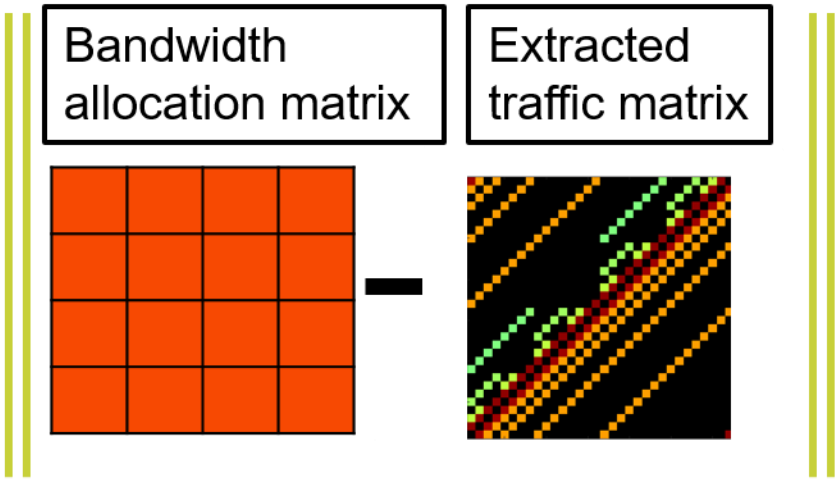
# Bandwidth Steered



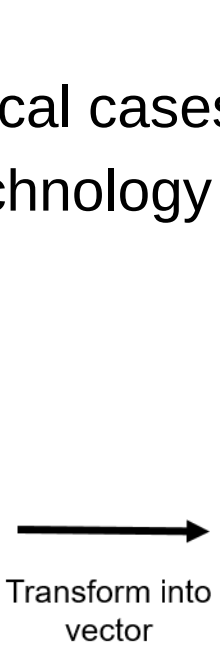

[Min Yee (Jason) The]

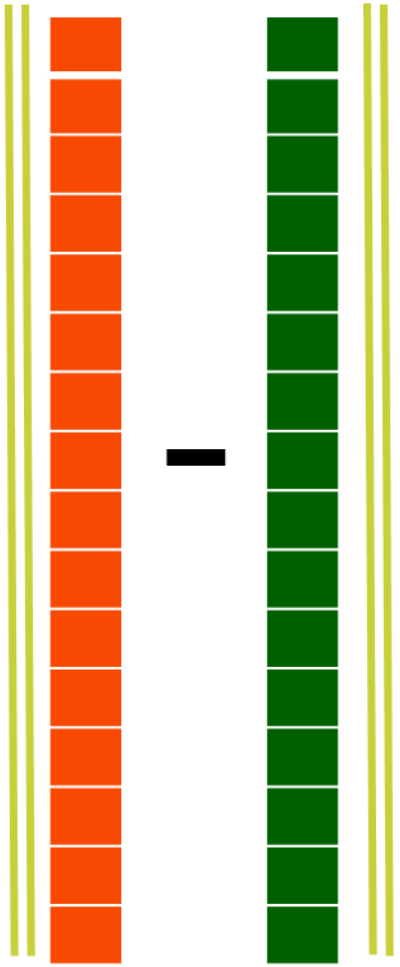
# Algorithmically Challenging

- NP-hard optimally
- Respect physical limitations
- Understand implications in pathological cases
- Solid models of underlying optics technology
  - Cost of reconfiguration




Transform matrix form into vector form



[Min Yee (Jason) The]

# Conclusion

- It's an exciting time to be an architect
- It's hard to predict how digital computing will look like in 20 years
- Likely more diversified by application domain and even specific algorithm
- We should focus on a **grand strategy** to best make use of our available options
  - To include computation and communication

# Questions