BERKELEY LAB COMPUTATIONAL RESEARCH
LAWRENCE BERKELEY NATIONAL LABORATORY

An Architect's Point of View of the Post Moore Era

Dr. George Michelogiannakis

Research scientist
Computer architecture group
Lawrence Berkeley National Laboratory

Work with Dr. Dilip Vasudevan

These are not DOE's or LBNL's official views

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
- Transistor density will increase by 2x every 18 months
- **Transistor density will increase by 2x every 24 months**
  - In 1975

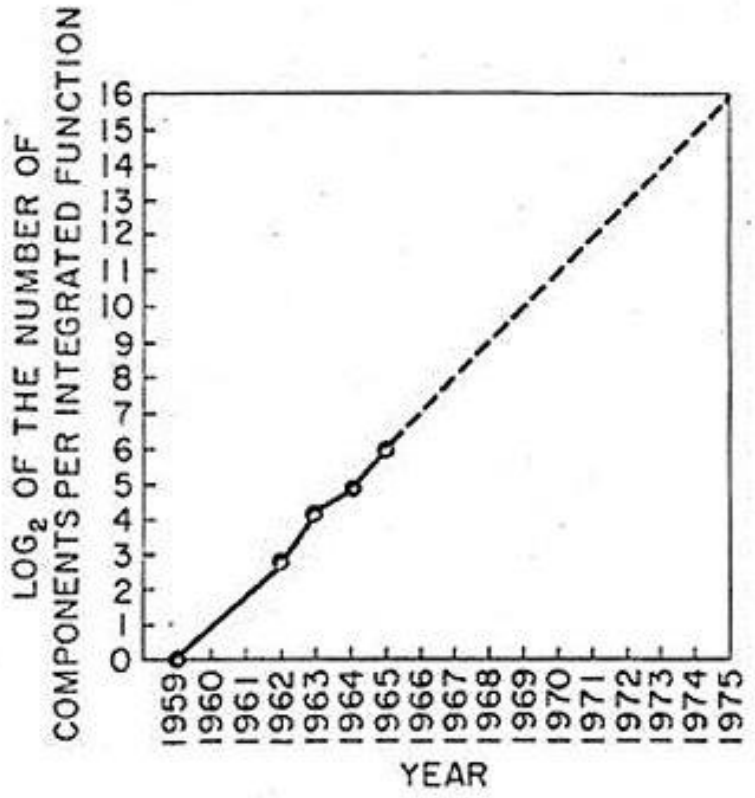
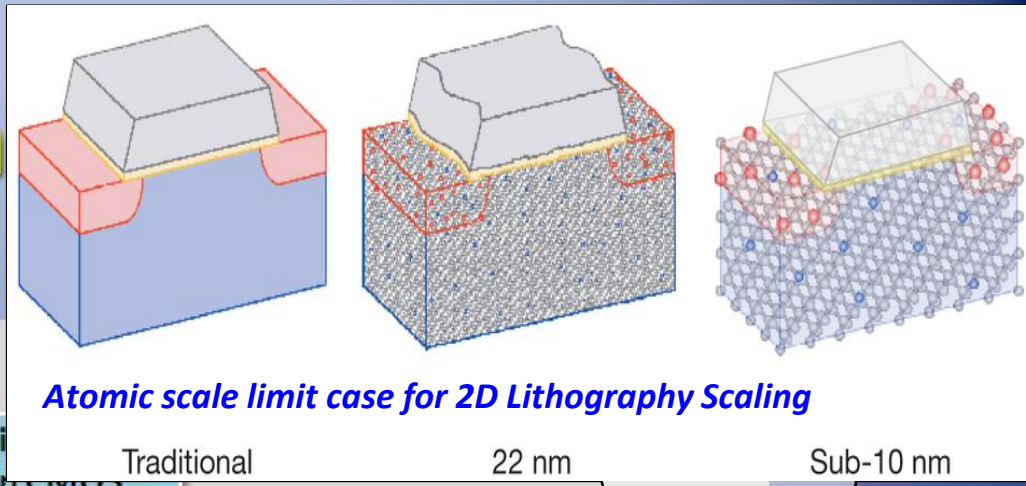


Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

Dr. Moore's 1965 paper

50 years of Sem



Atomic scale limit case for 2D Lithography Scaling

10 um
1 um
100 nm
10 nm
1 nm

Modern CMOS
Beginni
submicron CMOS
Deep UV litho
65 nm in 2006
28 nm in 2012
Presumed limit to scaling

40 years of scaling history

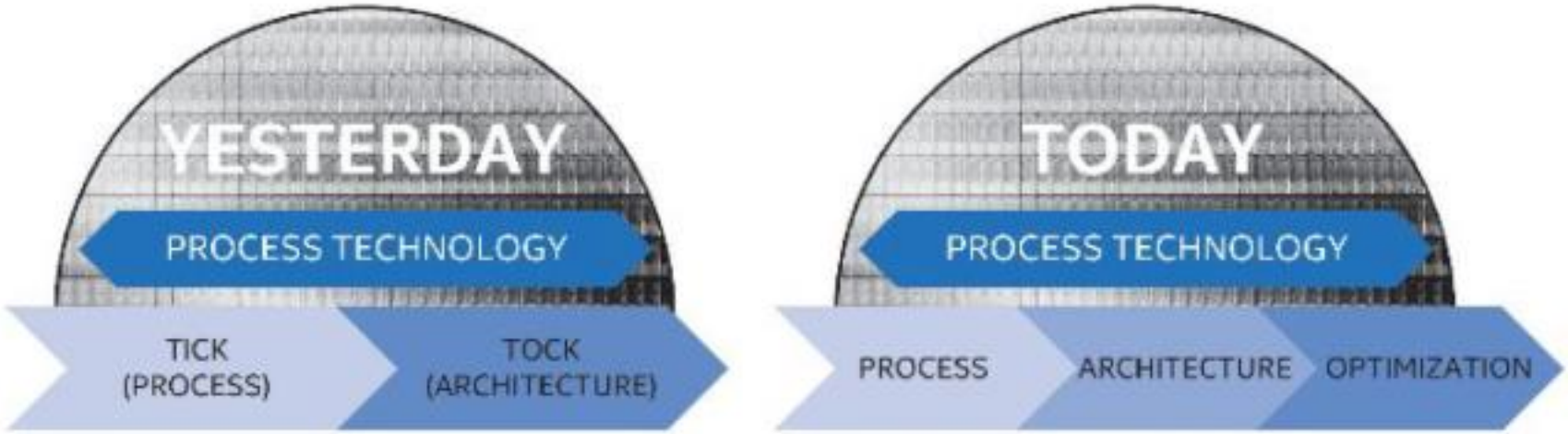
- Every generation
  - Feature size shrinks by 70%
  - Transistor density doubles
  - Wafer cost increases by 20%
  - Chip cost comes down by 40%
- Generations occur regularly
  - On average every 2.9 years over the past 40 years
  - Recently two years but the rate is slowing

1970 1980 1990 2000 2010 2020 2027?

5nm

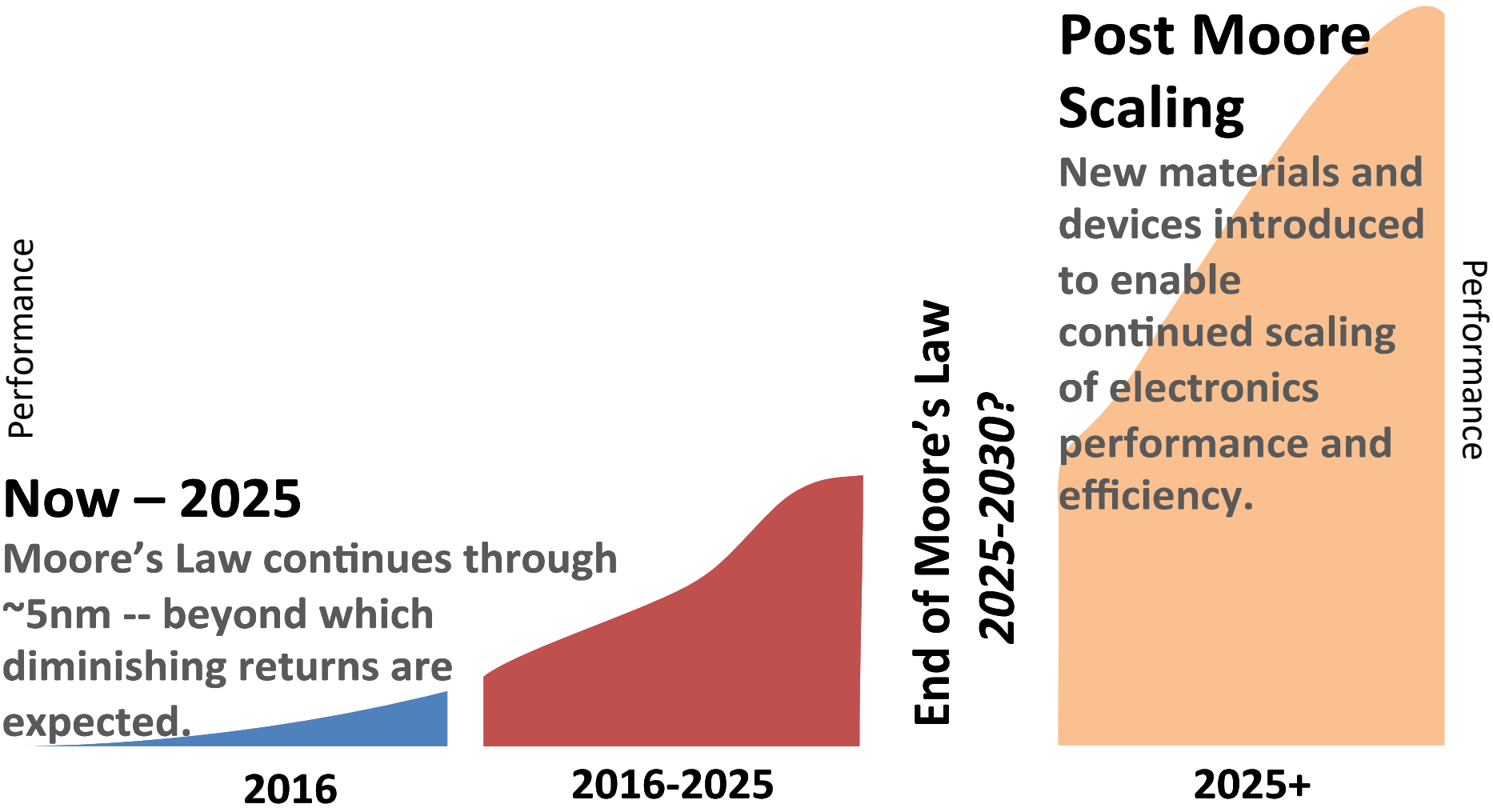
Moore's Law of Documentation

Scaling Already Slowing Down

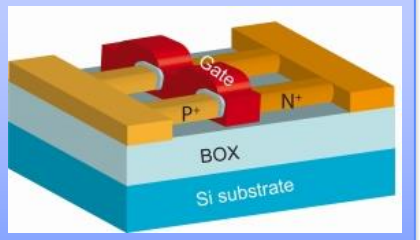


Peter Bright “Intel retires “tick-tock” development model, extending the life of each process“, 2016

Preserve Performance Scaling With Emerging Technologies

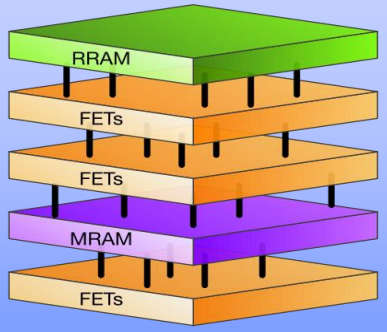


Emerging Technologies

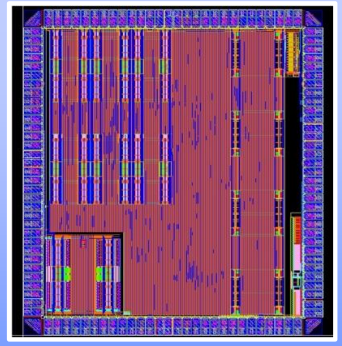


Emerging transistors

Emerging memories



3D integration

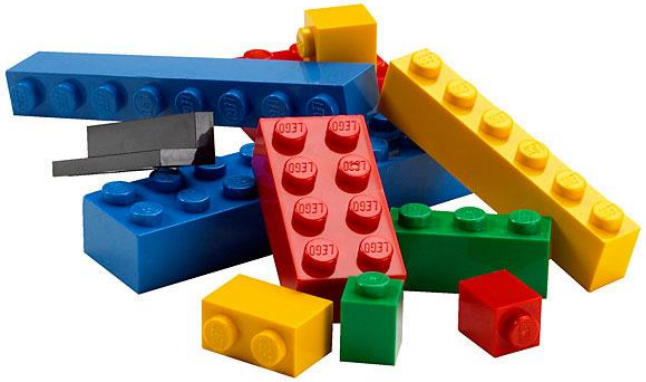


Specialized architectures

+ others

An Architect's Point of View

An Architect's Job

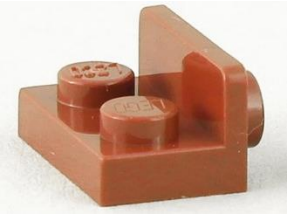


Lego Designs Have Been Getting Larger

New Lego Pieces

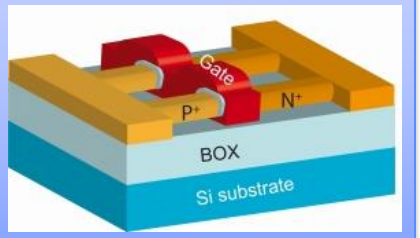
- Old designs can no longer become smaller with same strength

- Lego came up with new pieces:

- Which ones do we use?
- How does each one change the optimal design?
- How does each piece interact with others?
- What feedback can we provide Lego to refine each piece?

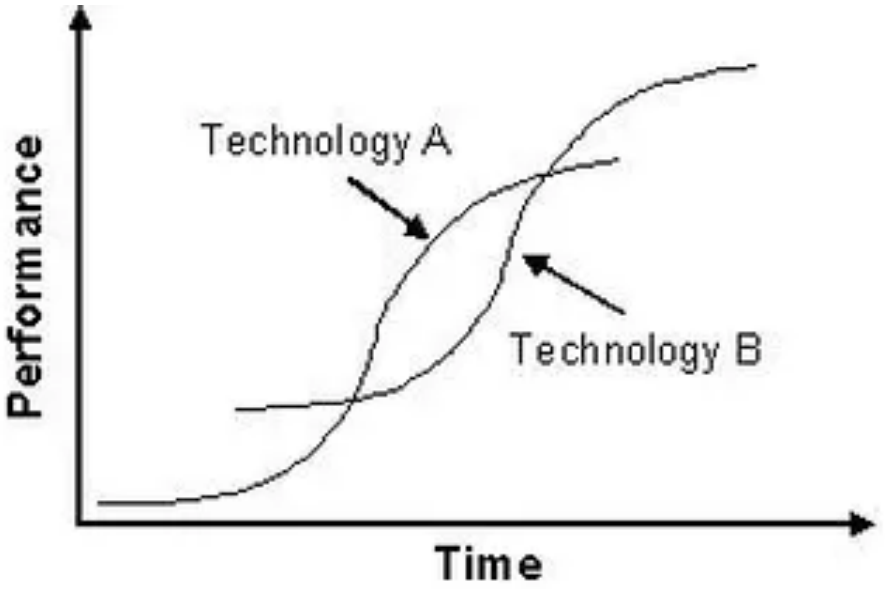
Emerging Transistors



Emerging transistors

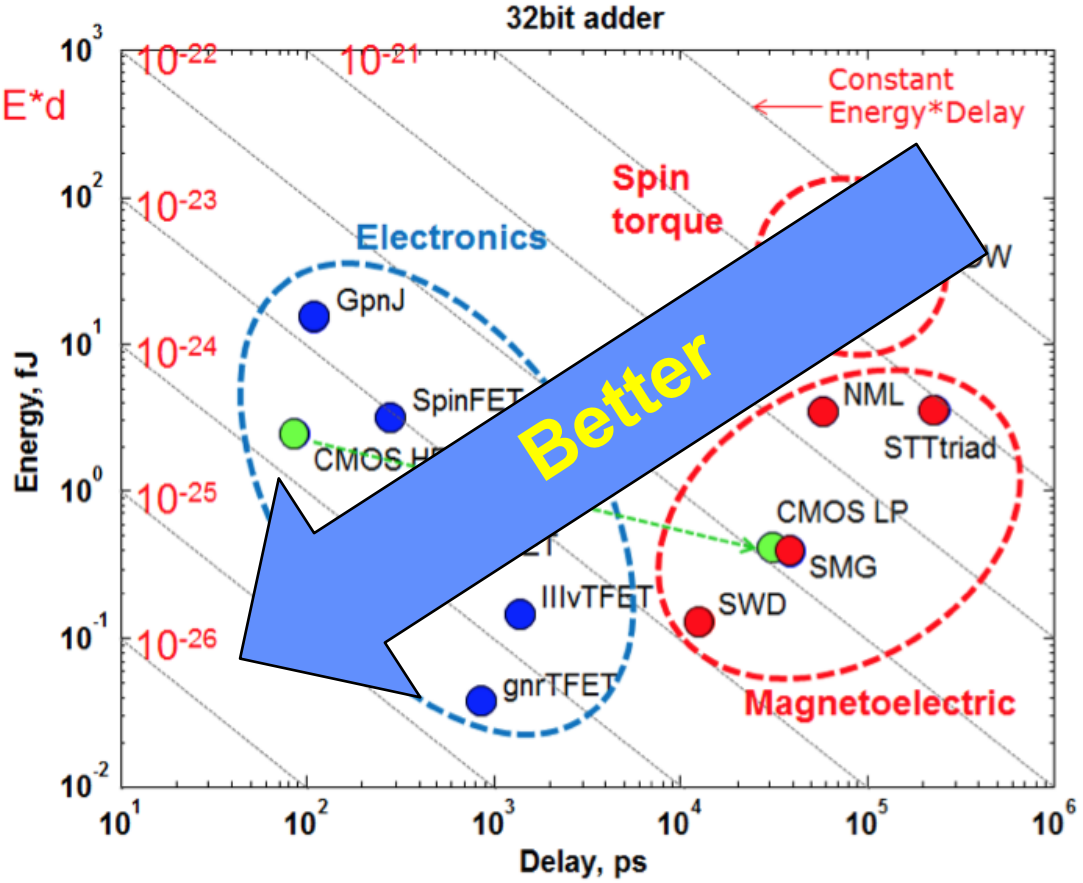
New Devices

- New devices need time to show their potential
- Two broad categories:
  - New designs
  - New materials
- Maybe not a single replacement for MOSFETs



Rick Lindquist “3 Steps for Constructive Disruption”

Many More



Nikonov and Young, "Benchmarking of Beyond-**CMOS Exploratory** Devices for Logic Integrated Circuits", 2015

Each dot is a moving target. We have to judge the potential

Emerging Memories

Emerging memories

Many Memories As Well

- Some of these are non-volatile

	SRAM	DRAM	eDRAM	2D NAND Flash	3D NAND Flash	PCRAM	STTRAM	2D ReRAM	3D ReRAM
Data Retention	N	N	N	Y	Y	Y	Y	Y	Y
Cell Size (F^2)	50-200	4-6	19-26	2-5	<1	4-10	8-40	4	<1
Minimum F demonstrated (nm)	14	25	22	16	64	20	28	27	24
Read Time (ns)	< 1	30	5	10^4	10^4	10-50	3-10	10-50	10-50
Write Time (ns)	< 1	50	5	10^5	10^5	100-300	3-10	10-50	10-50
Number of Rewrites	10^{16}	10^{16}	10^{16}	10^4-10^5	10^4-10^5	10^8-10^{10}	10^{15}	10^8-10^{12}	10^8-10^{12}
Read Power	Low	Low	Low	High	High	Low	Medium	Medium	Medium
Write Power	Low	Low	Low	High	High	High	Medium	Medium	Medium
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak	Sneak
Maturity									

J.S. Vetter and S. Mittal, "Opportunities for Nonvolatile Memory Systems in Extreme-Scale High Performance Computing," CiSE, 17(2):73-82, 2015.

Not good

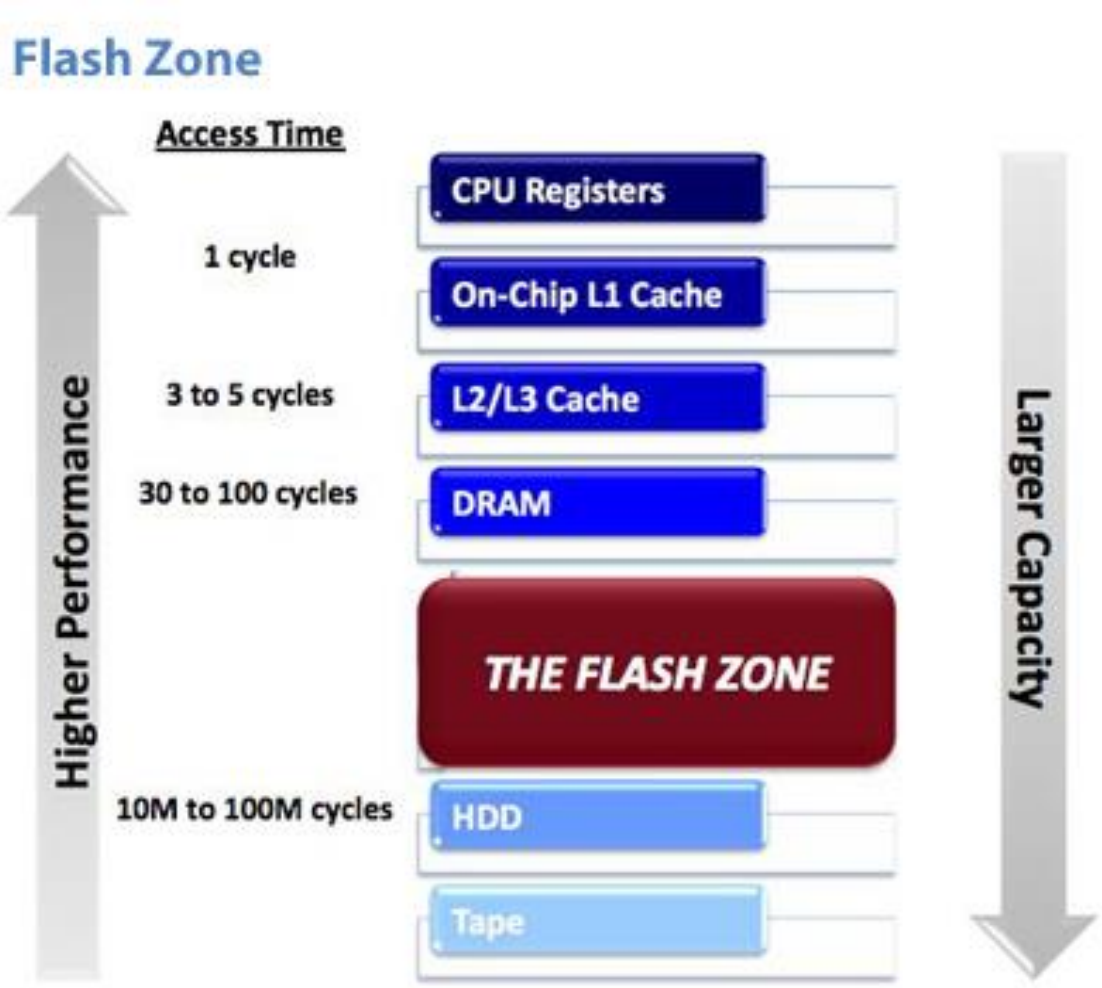
Very good

Good

Excellent

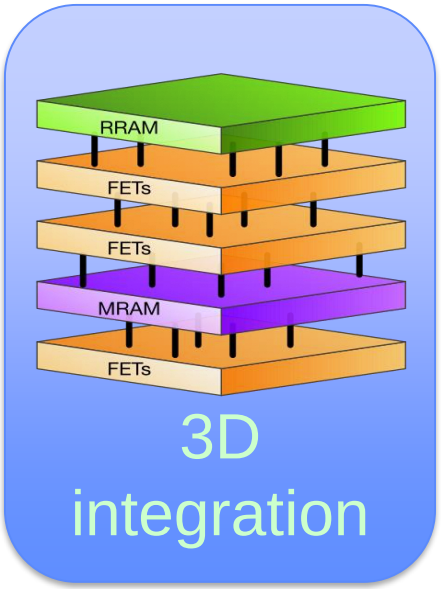
What About Memory Hierarchy?

- Non-volatility higher at the hierarchy
  - Challenge assumption that non-volatile storage is slow and distant
- New memories have different read, write, reliability constraints
- New memory hierarchy likely different

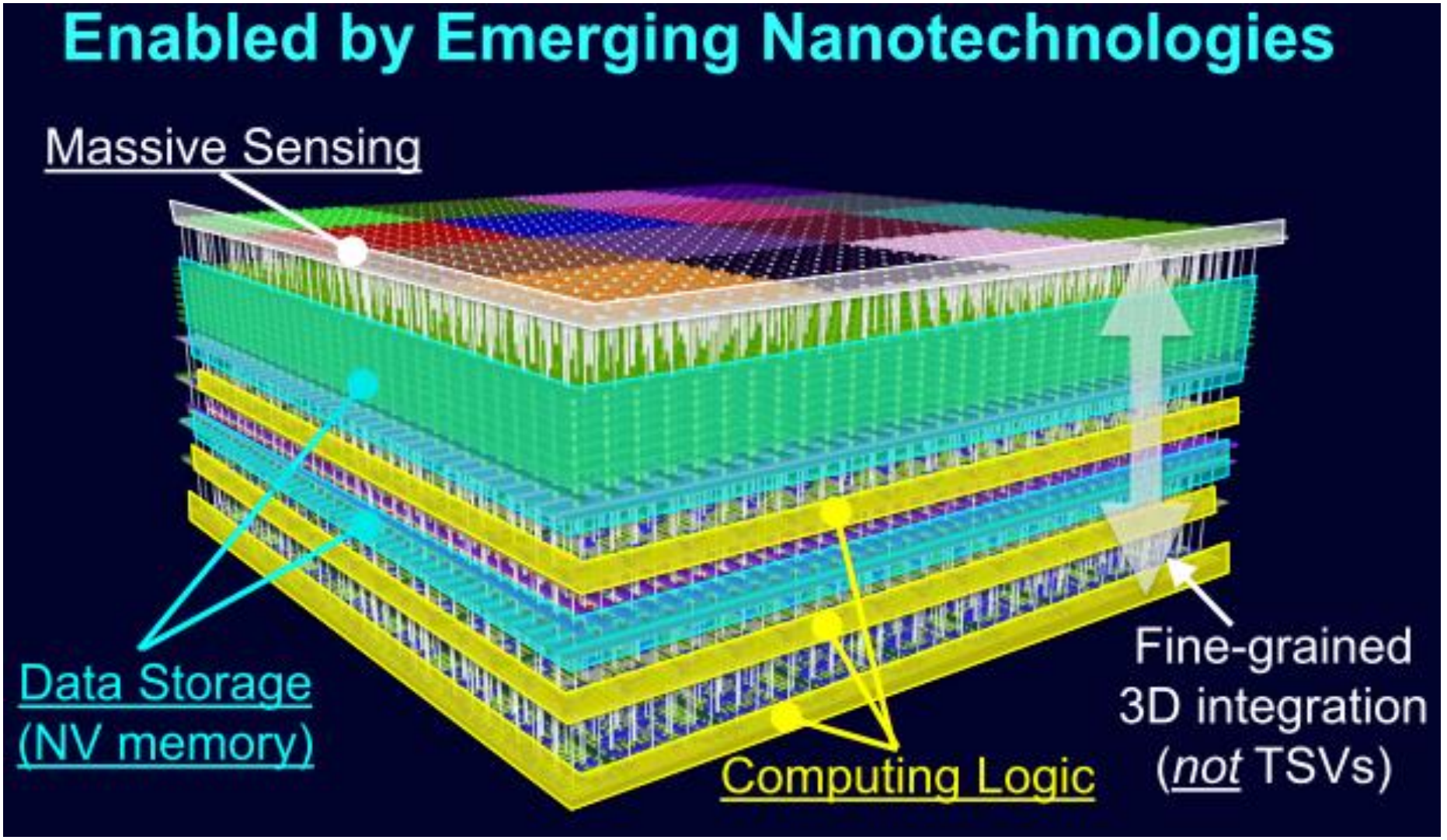


AGIGARAM “The Flash Zone”

3D Integration



3D Integration of Tomorrow



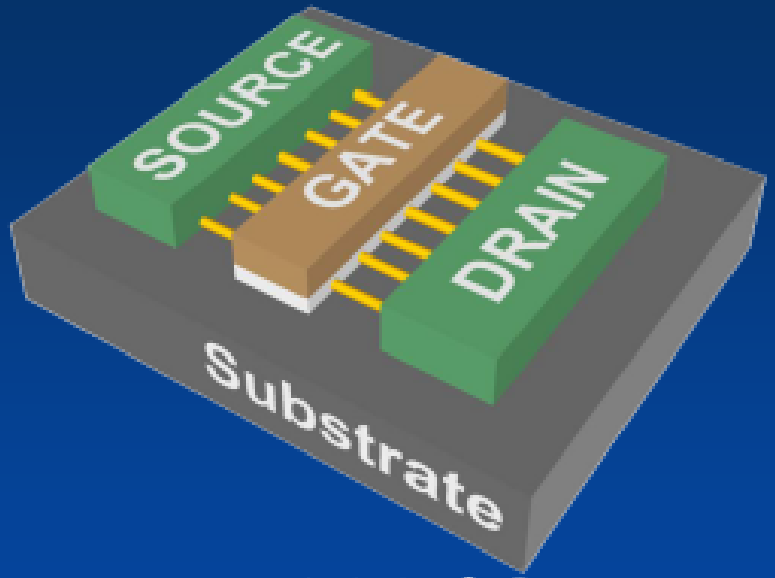
Shulaker "Transforming Emerging Technologies into Working Systems"

Technology Requirements

- Low-temperature fabrication: <400 °C

Logic

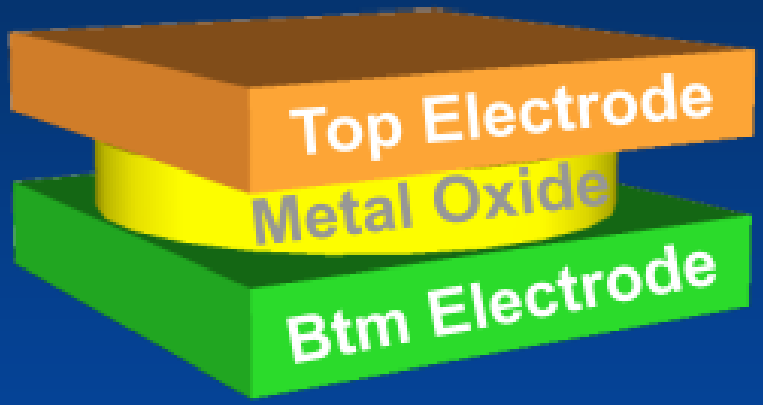
Carbon Nanotubes



<200 °C

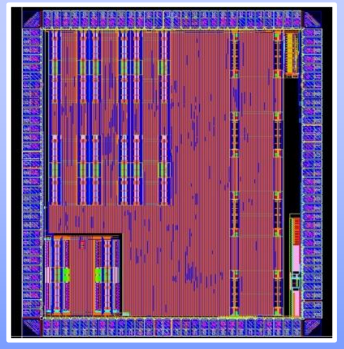
Memory

Resistive RAM



<200 °C

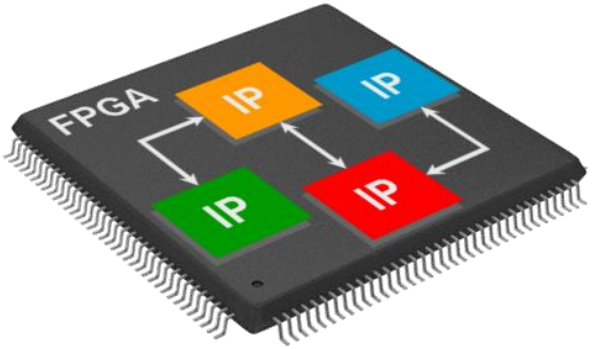
Specialization

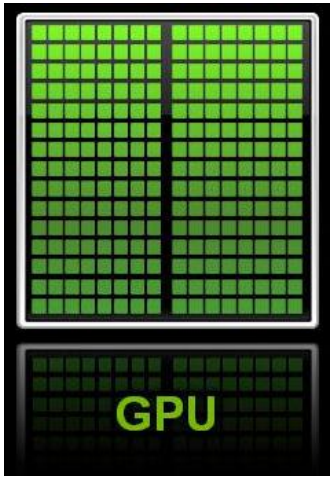


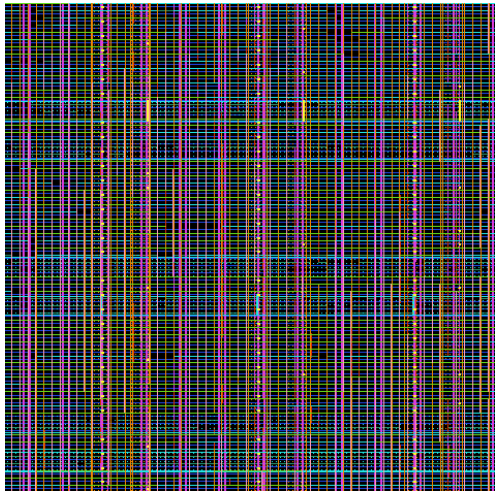
Specialized architectures

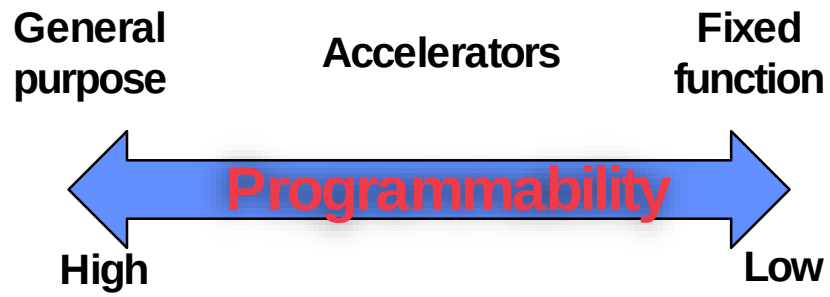
Specialization

- Hardware that is more suited for specific kinds of computation
  - Can also have accelerators for data transfer



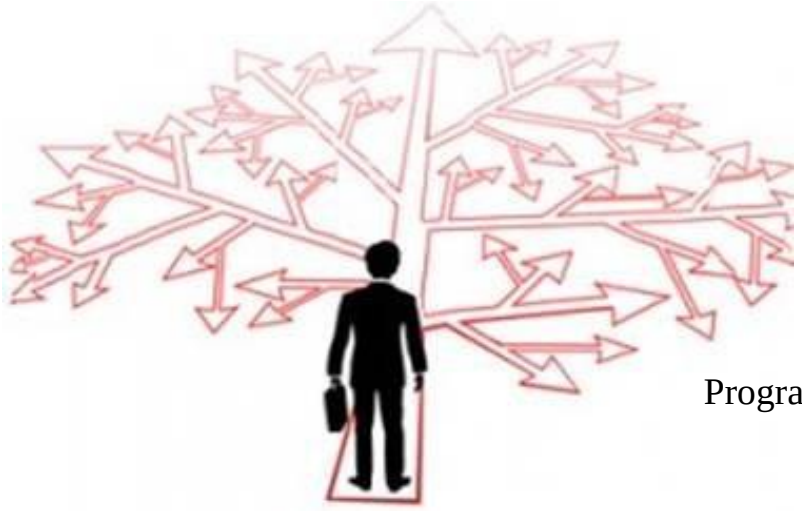


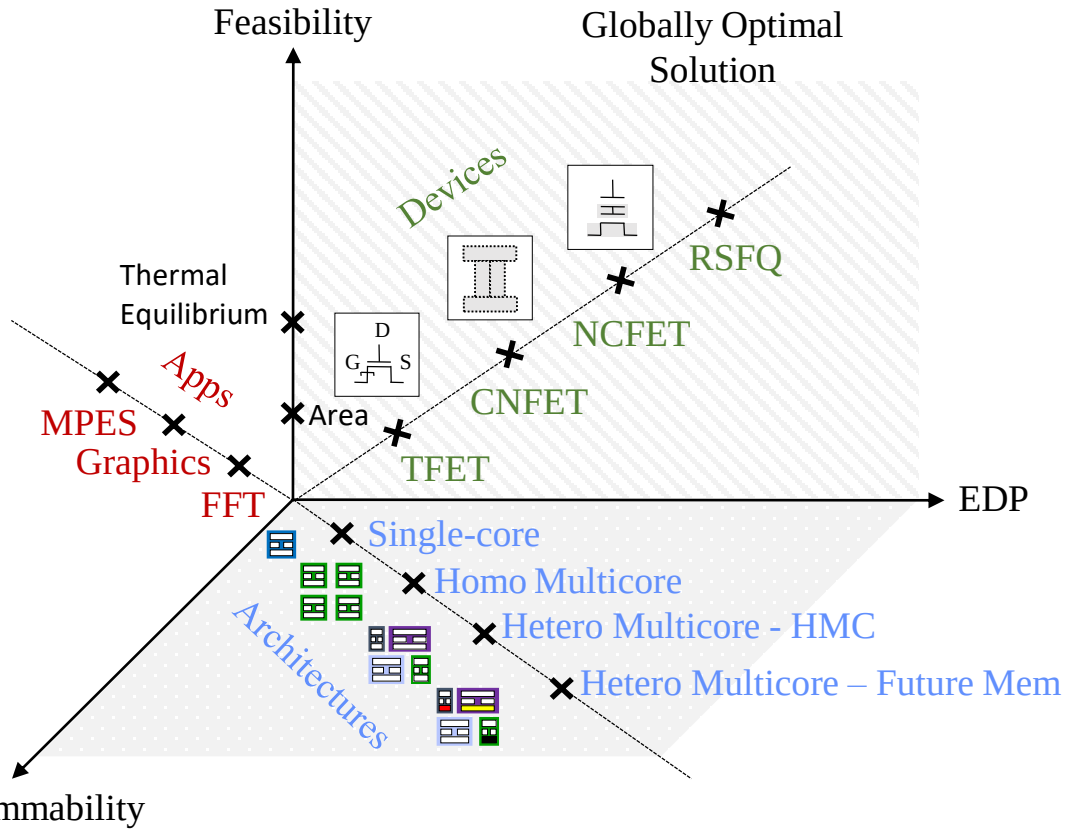




Overall: The Variety of Choices Is Overwhelming

- The vast number of choices is a problem by itself
  - It makes finding a good design harder, especially when designing manually

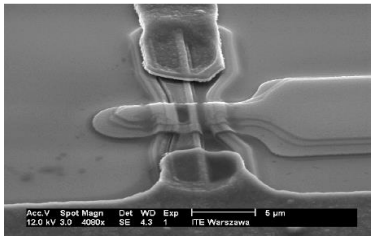




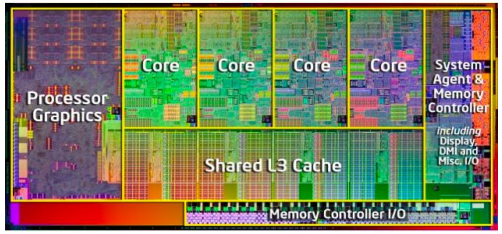
Evaluate At Architectural Level

- 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

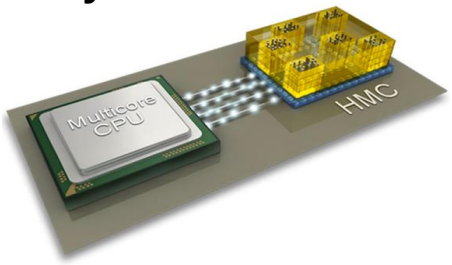
Transistor/Devices



Architecture



System

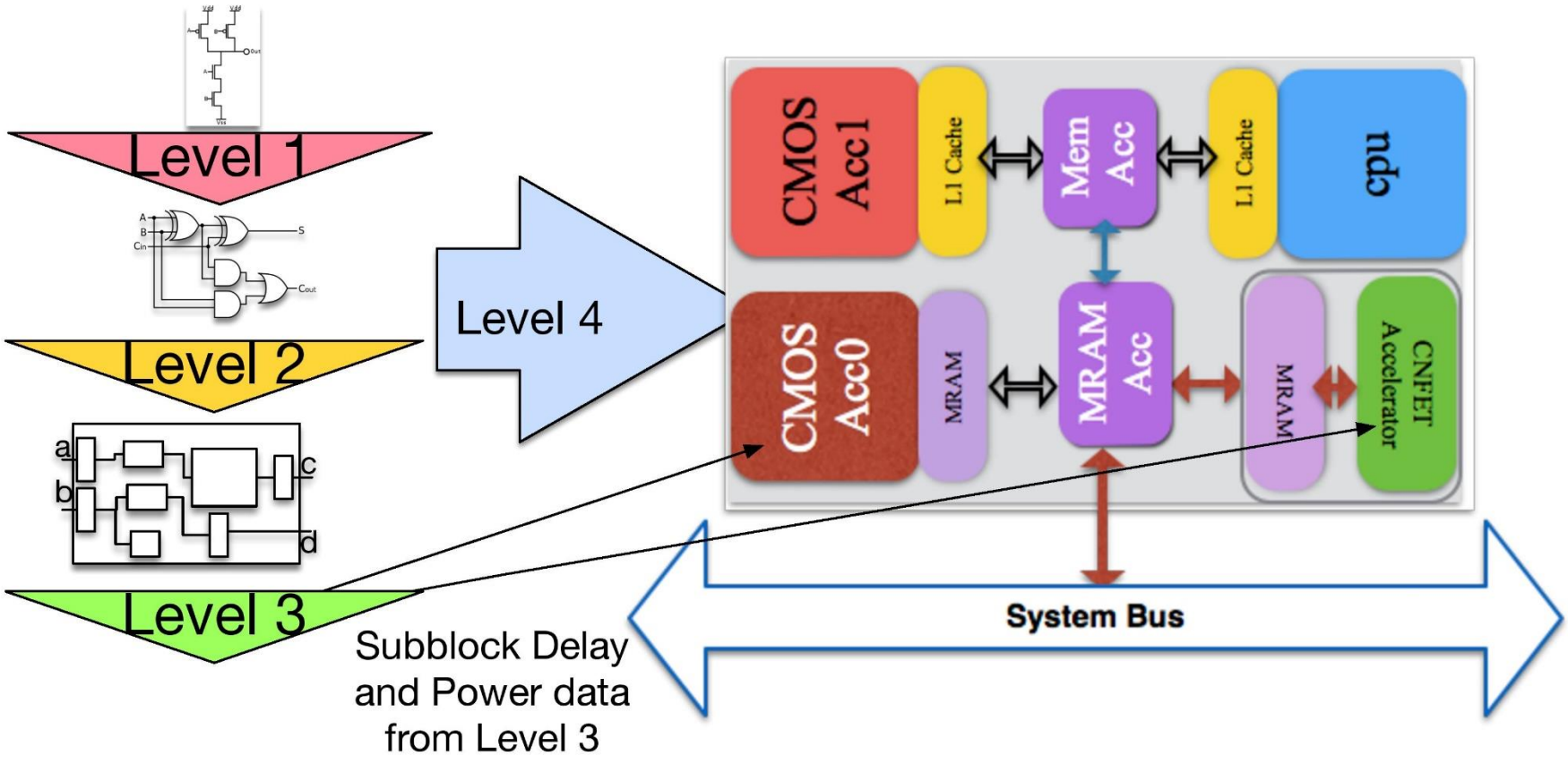


- But we lack the tools to do so systematically for many technologies

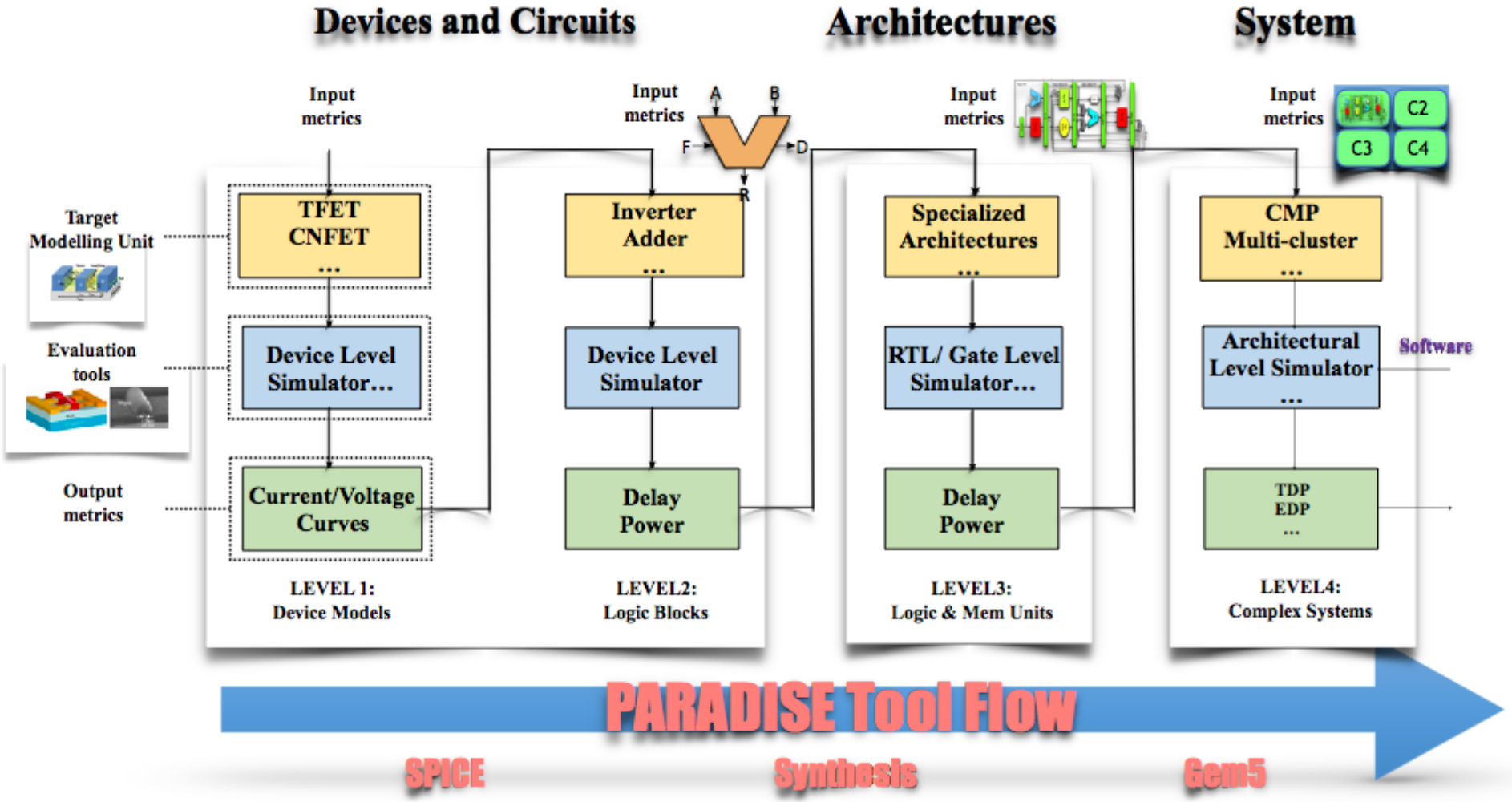
How To Make An Architect's Job Easier?

Multi-Level Architectural Simulation

Beyond Moore Architectural Simulation from Emerging Device and Gate Level analysis



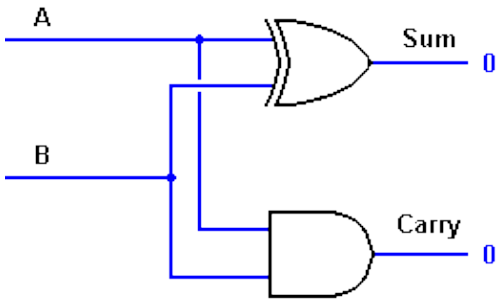
PARADISE End-To-End Tool Flow

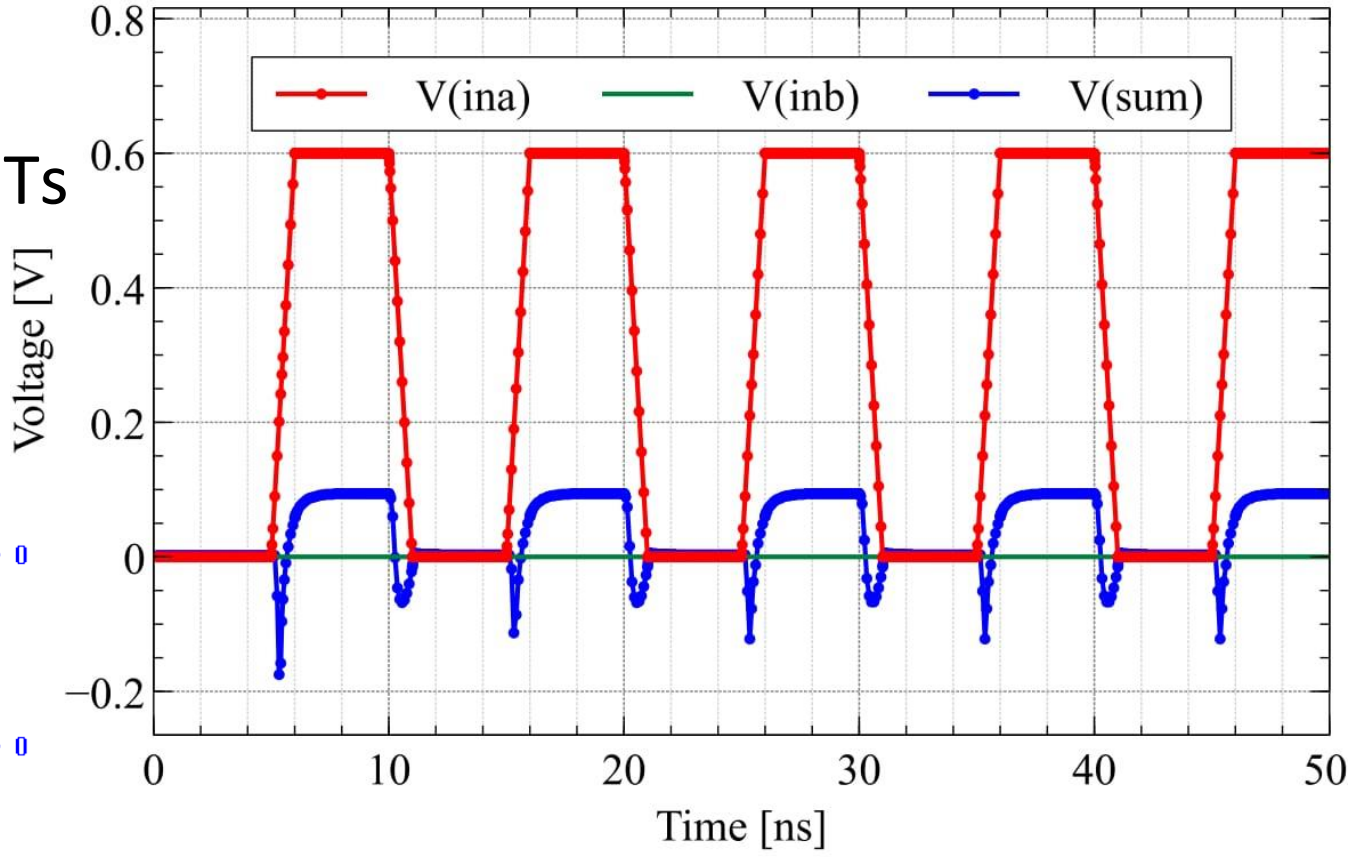


Levels 1 and 2 Physical Simulation

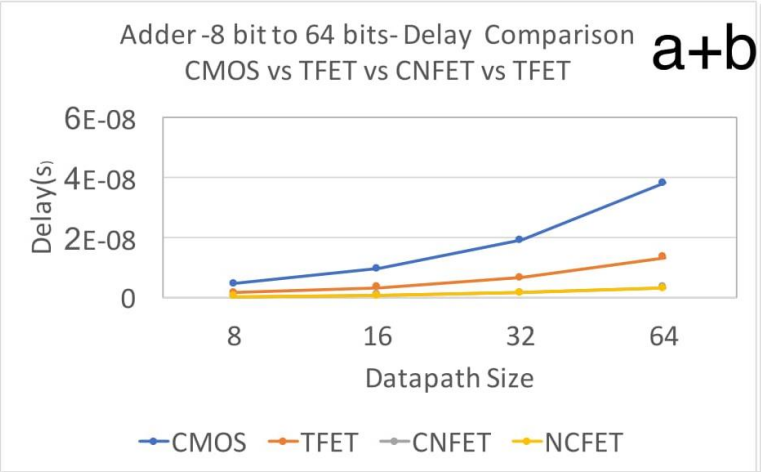
- Level 1 is the input for devices
- Xyce: open source parallel SPICE client

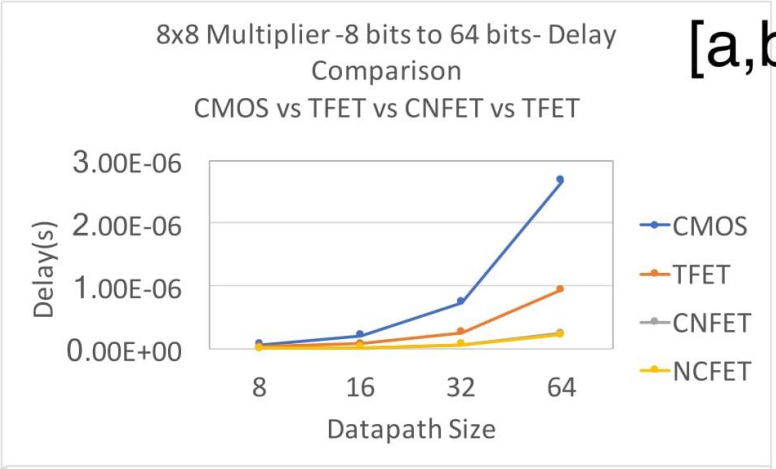
Adder using TFETs

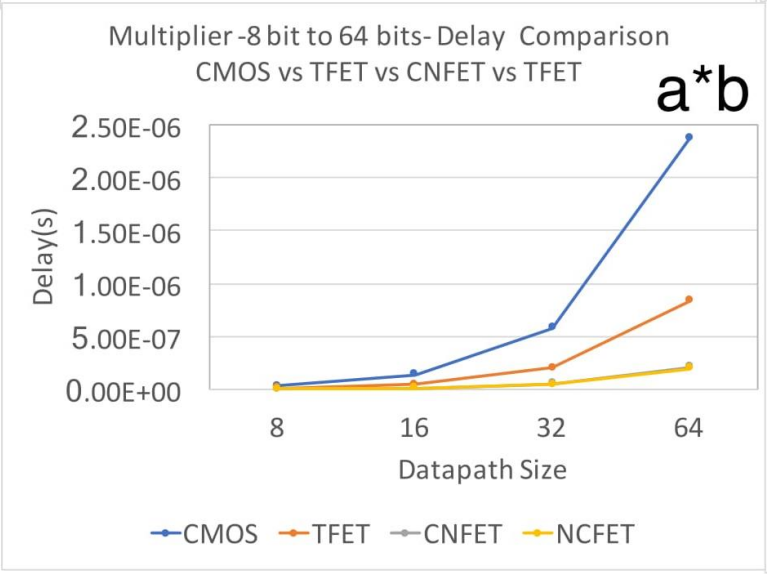


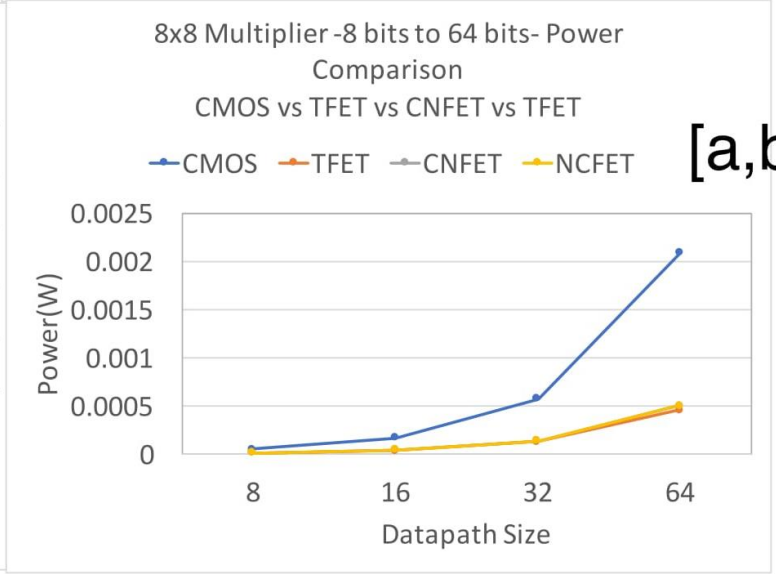


Comparison Studies (PARADISE generated)







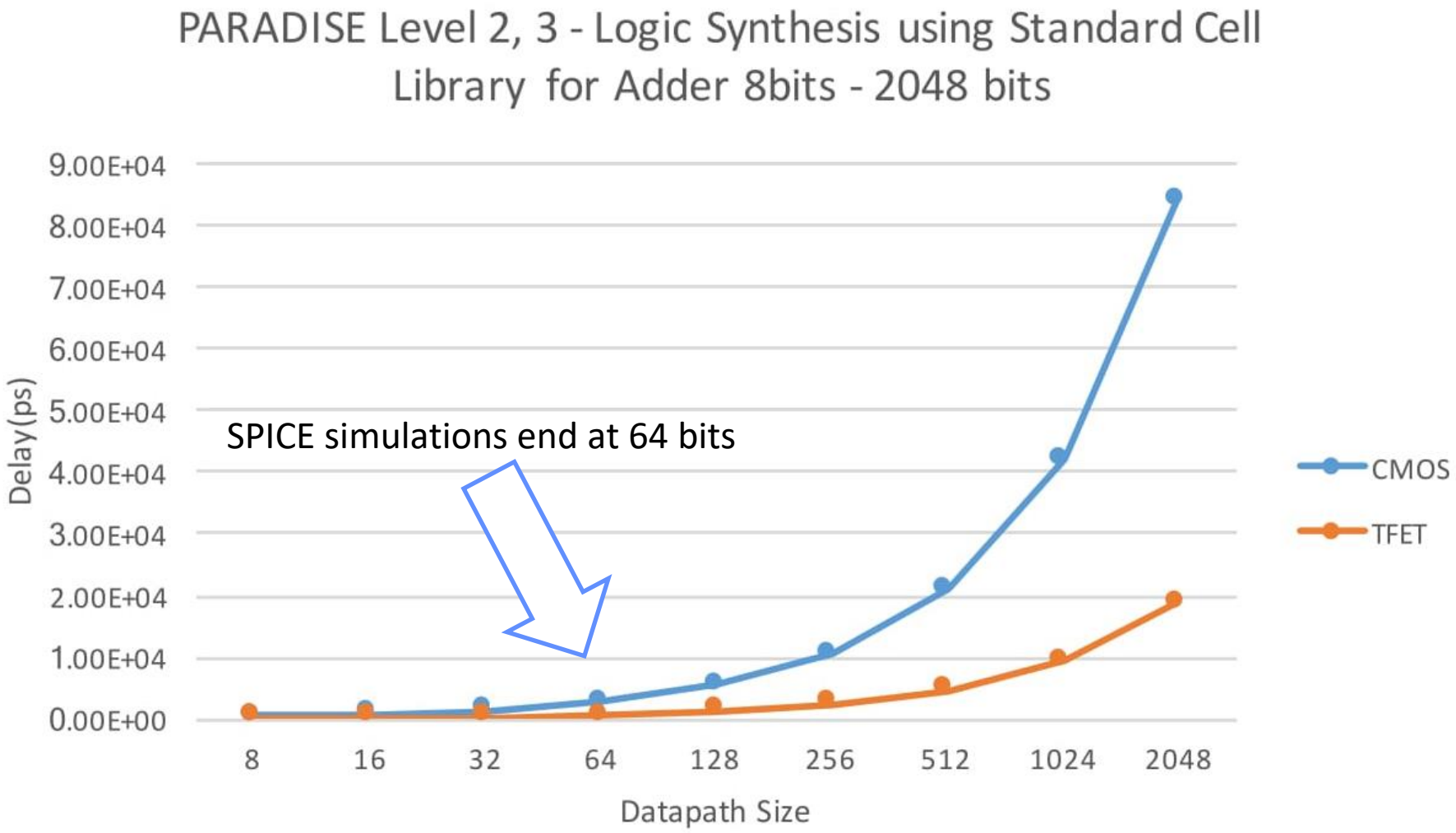


Delay comparison for a multiplier

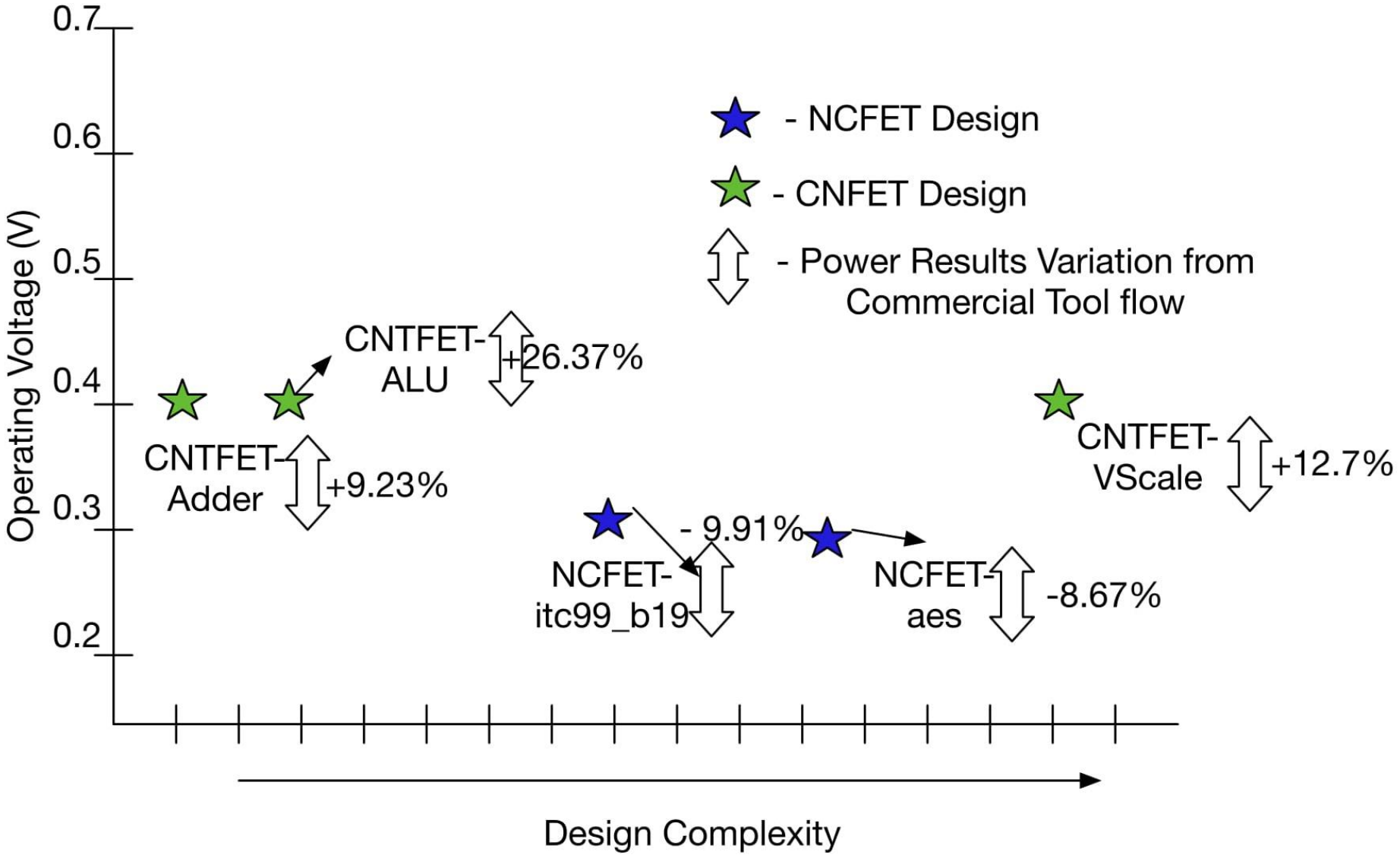
Power comparison for a multiplier

Level 3: RTL Synthesis

- Synthesis using Yosys and our own extension for power estimation

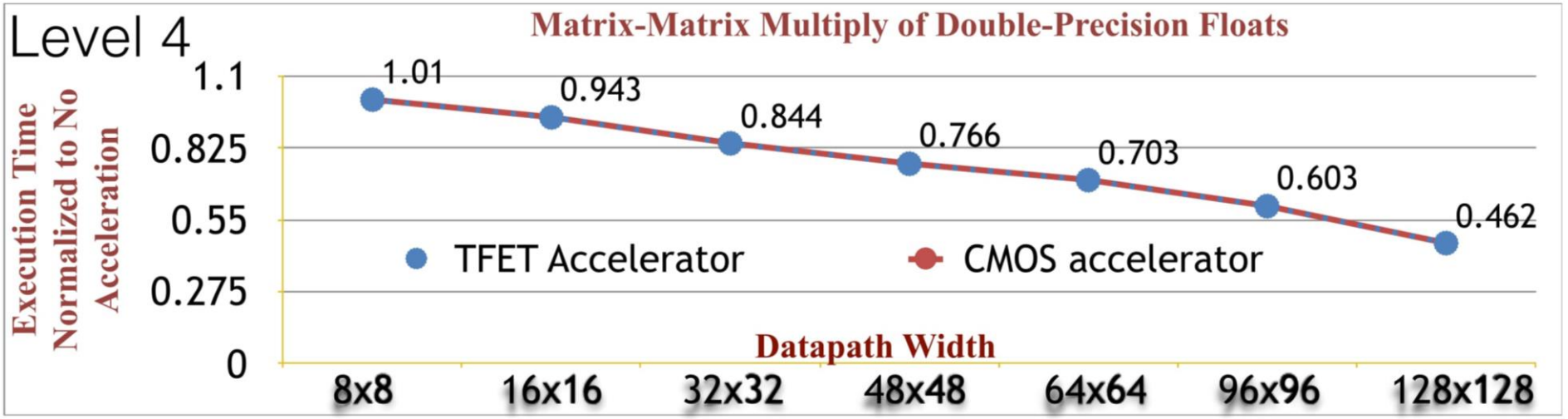


Design Space Exploration at RTL Level

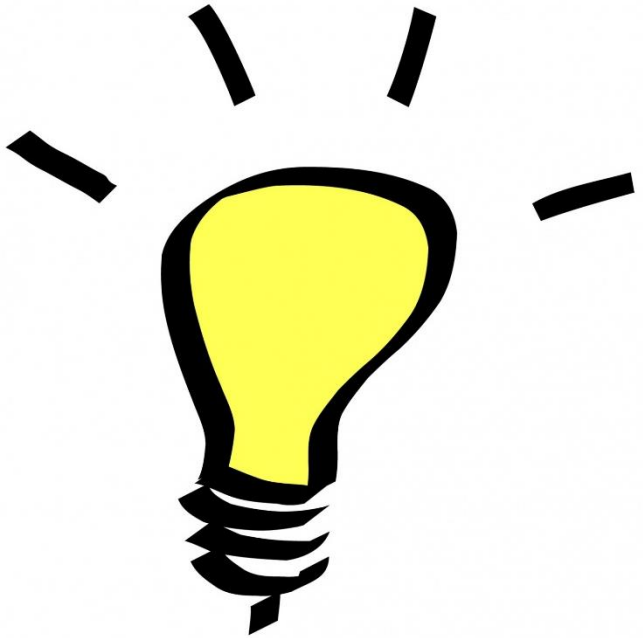


Level 4: Architectural Level

- Gem5 with Aladdin
- With small accelerators small delay differences do not have a significant application impact due to other overhead

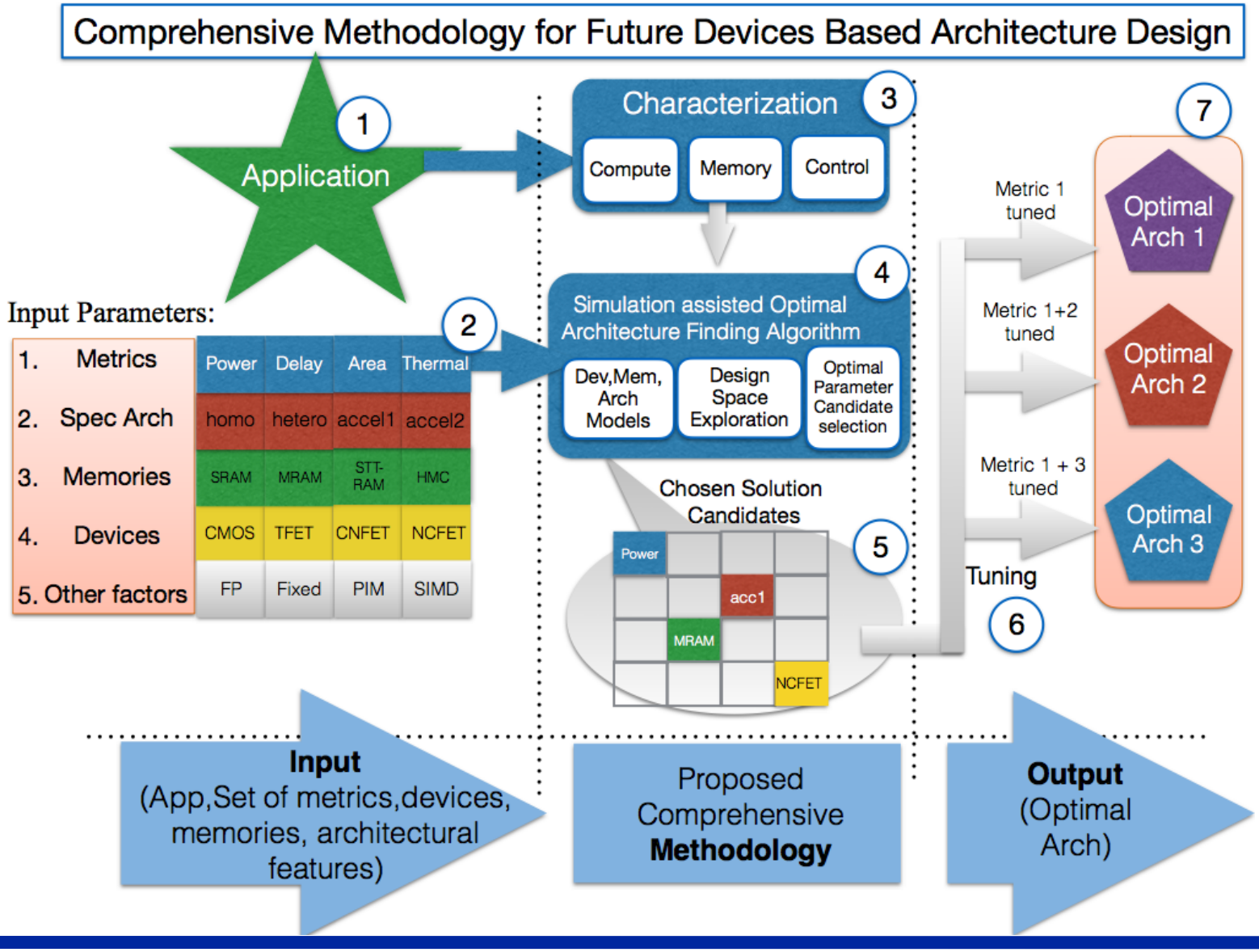


How To Use These Tools?



4 examples

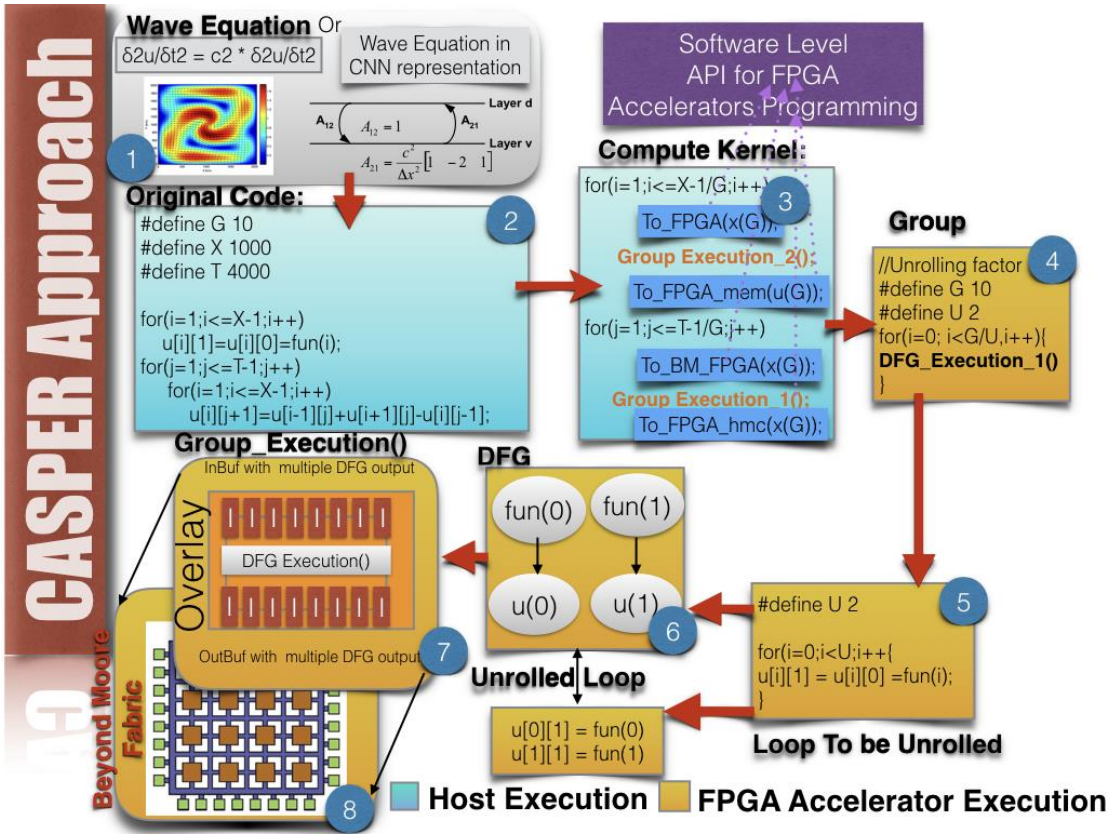
(1) Architecture Design Methodology



(2) CASPER

- FPGAs can be heterogeneous too
- Overlay step understands available FPGA hardware and maps IPs accordingly

End-to-End Open Source Reconfigurable DSE Methodology/Tool Flow for Beyond Moore FPGAs



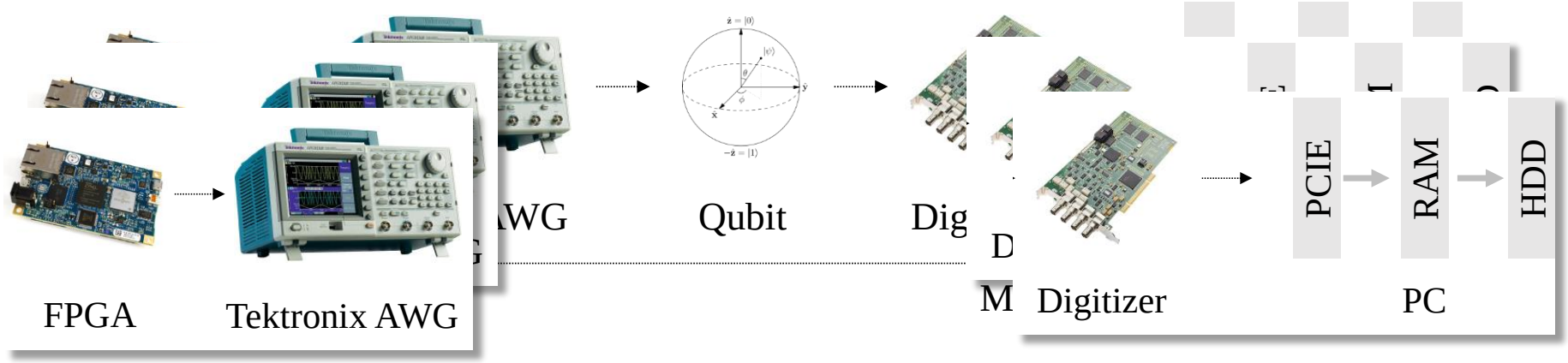
D. Vasudevan et al, "CASPER — Configurable design space exploration of programmable architectures for machine learning using beyond moore devices," 2017

(3) Quantum Control Processor

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

Off the shelf and high cost

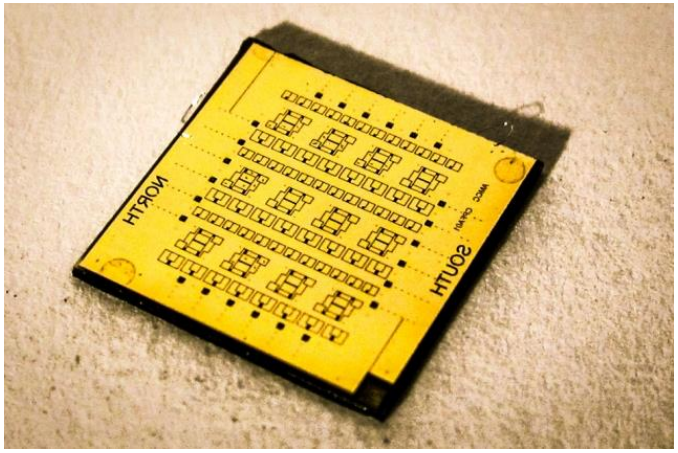
Large amount of data and slow speed



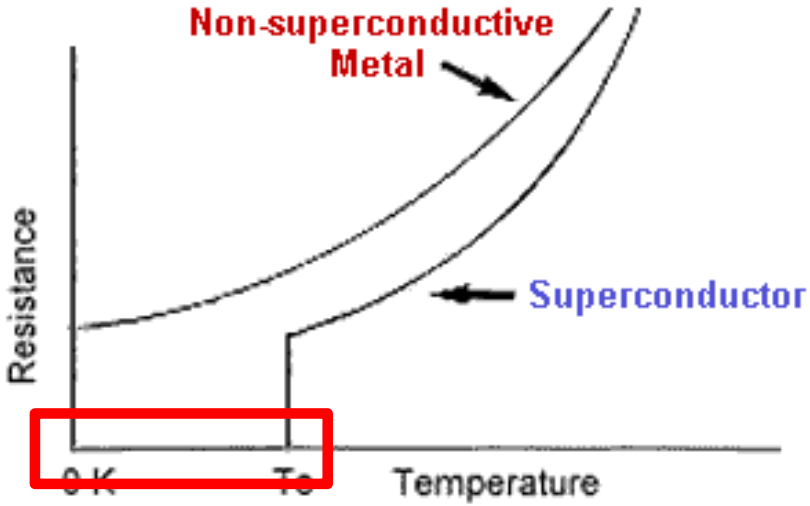
1000 qubits,
gate time 10ns,
3 ops/qubit
300 billion ops per second

(4) 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"

Looking for a PhD Thesis Topic? More Questions to Answer

- Which device technology will dominate?
  - For what domains, and with what side effects
- How does architecture change with device technology?
- How can we best take advantage of deep 3D?
  - With alternating logic and memory layers
- How large or distant do we make accelerators?
- How does the memory hierarchy change?
- How heterogeneous do architectures need to be?

Forewarn Programmers

- Build an architectural simulation tool that can be used by software developers
- What is the impact of challenging the far and expensive memory assumption?
  - Also non-volatile
- What about a heterogeneous memory hierarchy?
- How can we use reconfigurable accelerators?
- How to deal will reduced reliability?
  - Approximate computing may see a boost

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

Questions