Computing Beyond Moore's Law

John Shalf

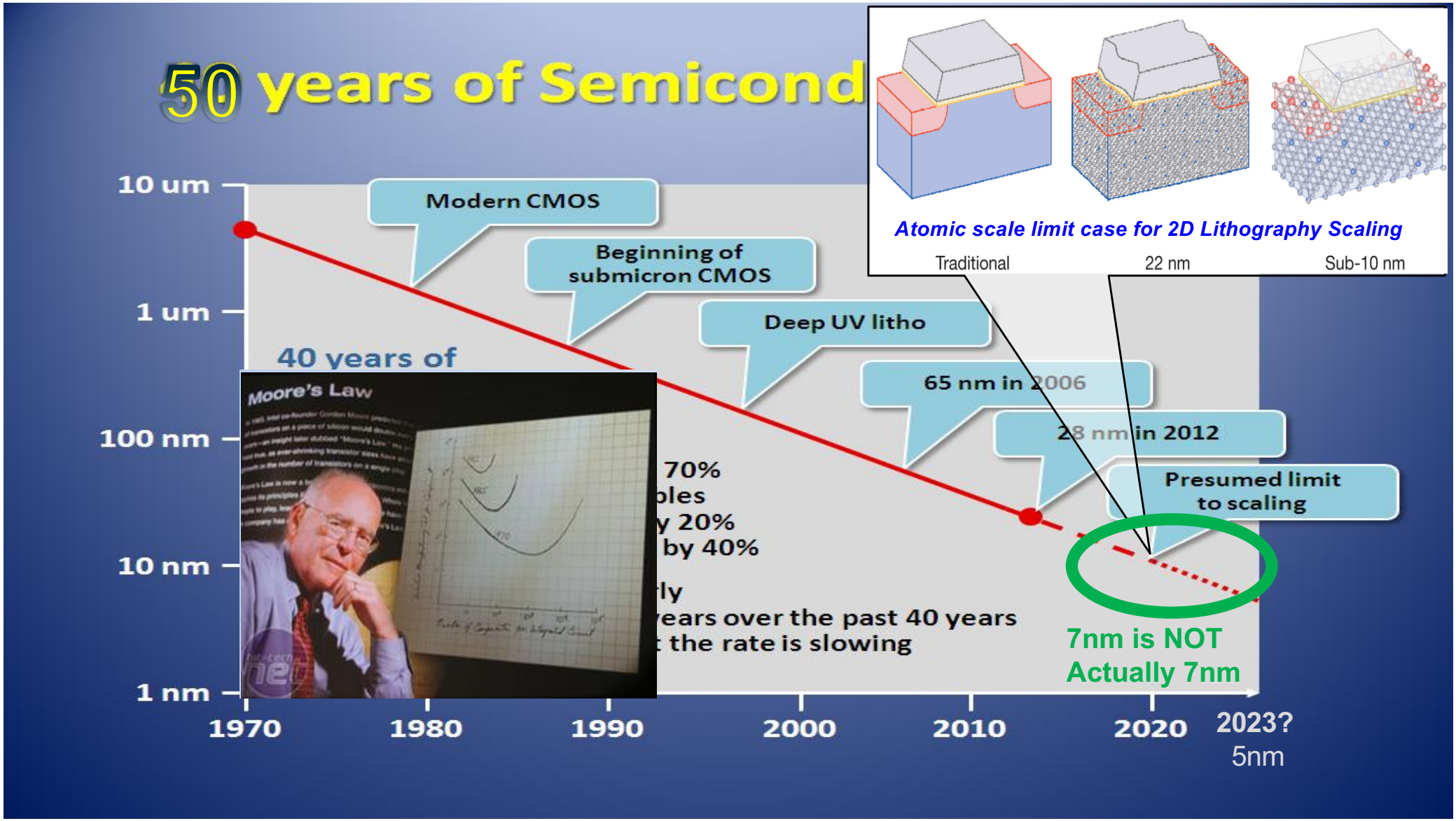
Department Head for Computer Science

Lawrence Berkeley National Laboratory

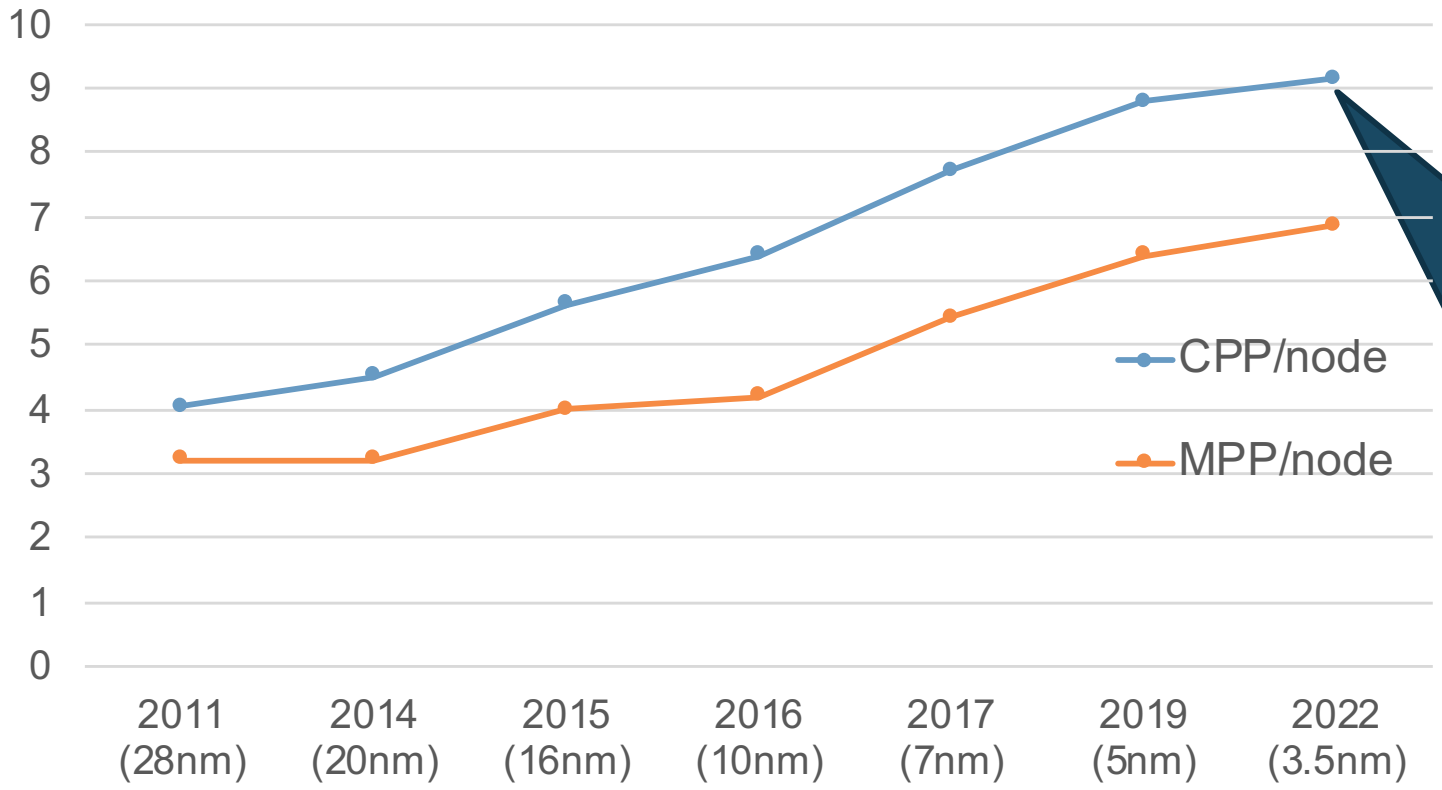
The International Supercomputing Conference

June 18, 2019

jshalf@lbl.gov


50 years of Semicond
10 um
1 um
100 nm
10 nm
1 nm
Modern CMOS
Beginning of submicron CMOS
Deep UV litho
65 nm in 2006
28 nm in 2012
Presumed limit to scaling
40 years of
Moore's Law
70%
les
y 20%
by 40%
ly
ears over the past 40 years
the rate is slowing
7nm is NOT
Actually 7nm
1970
1980
1990
2000
2010
2020
2023?
5nm
Atomic scale limit case for 2D Lithography Scaling
Traditional
22 nm
Sub-10 nm

Actual Half-Pitch vs. Technology Node (nm)



Data Source: *ASML*
IC Knowledge

At 3.5nm node

Minimum contact feature size=32nm
(9x worse than node)

Minimum Metal feature size = 24nm
(6x worse than node)

Technology Scaling Trends

Exascale in 2021… and then what?

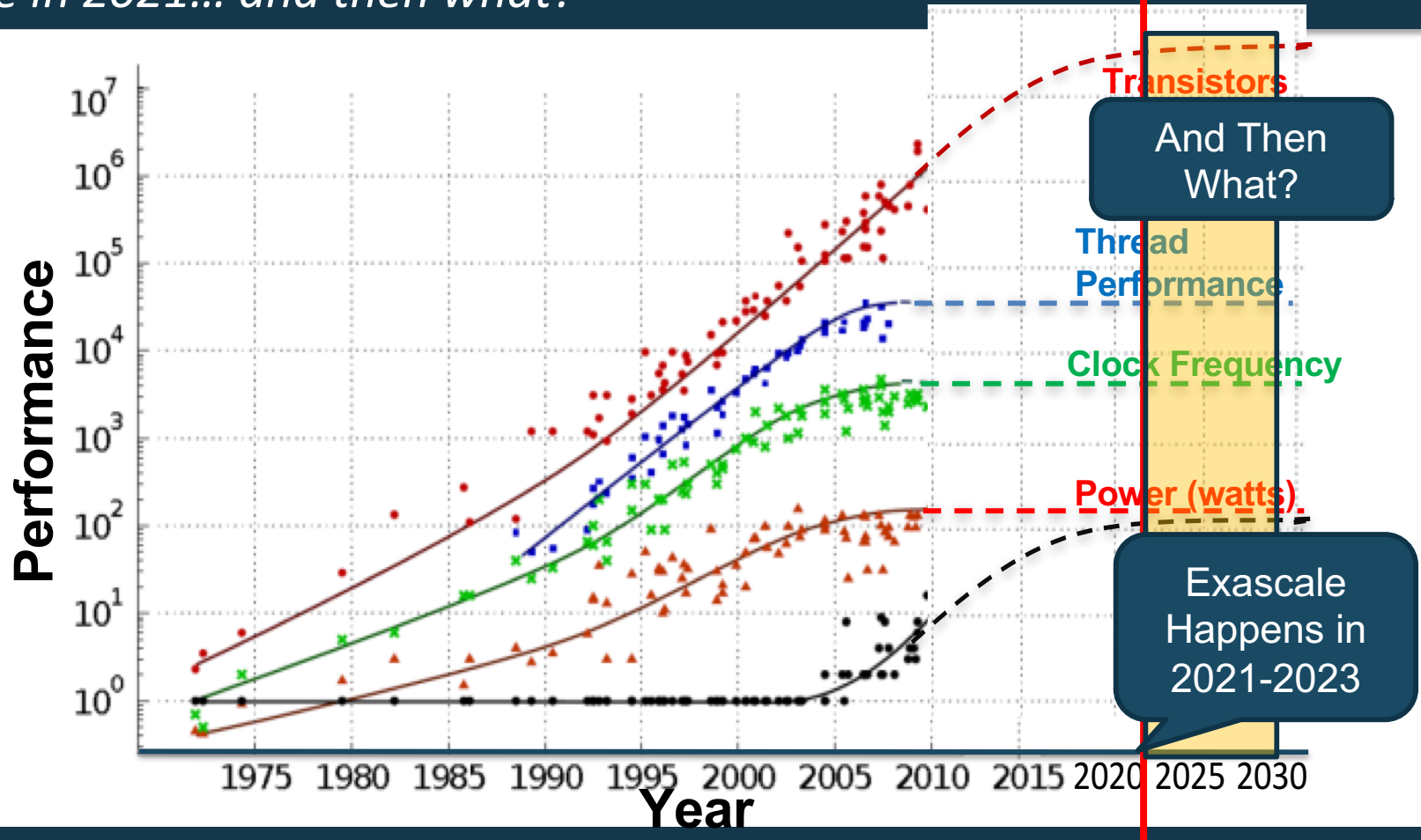


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Numerous Opportunities Exist to
tinue Scaling of Computing Performance

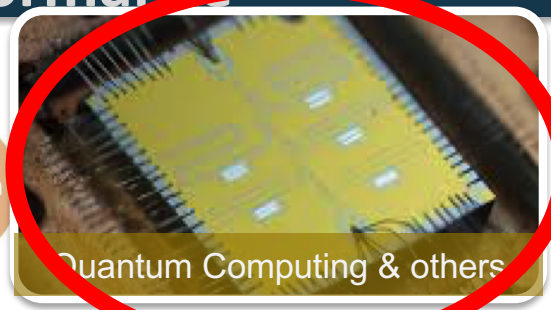
New Materials and Devices
20+ years (10 year lead time)

intronics
Carbon nanotubes and graphene
PETs
TFETs

New models of computati

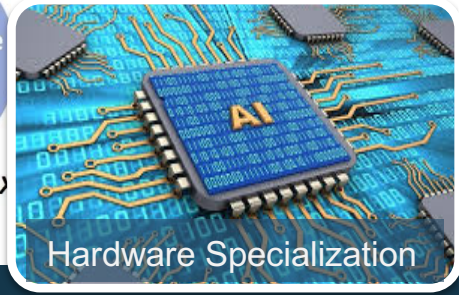
New Models of Computation
Decades beyond exascale

Neuromorphic
Quantum
Analog
Adabiatic reversible
aflow
oximate computing



CMOS
General purpose
Systems on chip
NTV
3D stacking, adv. packaging
Superconducting
Reconfigurable computing
Dark silicon

More Efficient Architectures and Packaging
The next 10 years after exascale

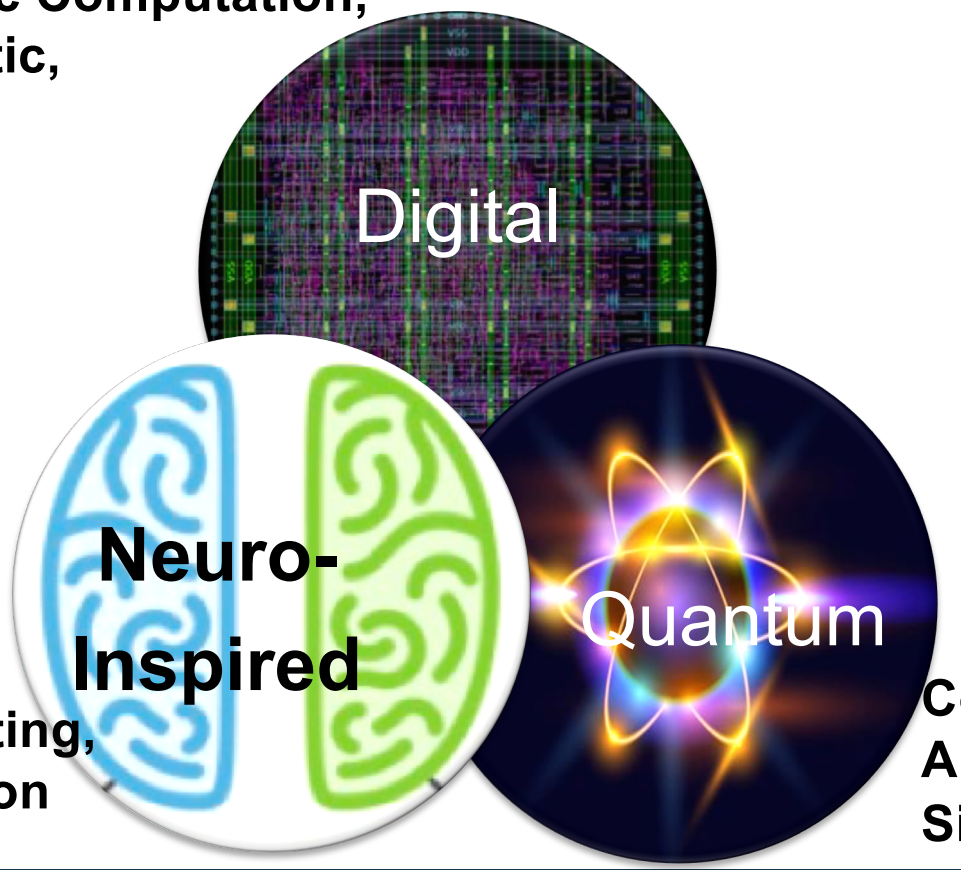


Many unproven candidates yet to be invested at scale. Most are disruptive to our current ecosystem.

BERKELEY LAB

Beyond Moore Computing Taxonomy

Symbolic Computation,
Arithmetic,
Logic



Cognitive Computing,
Pattern Recognition

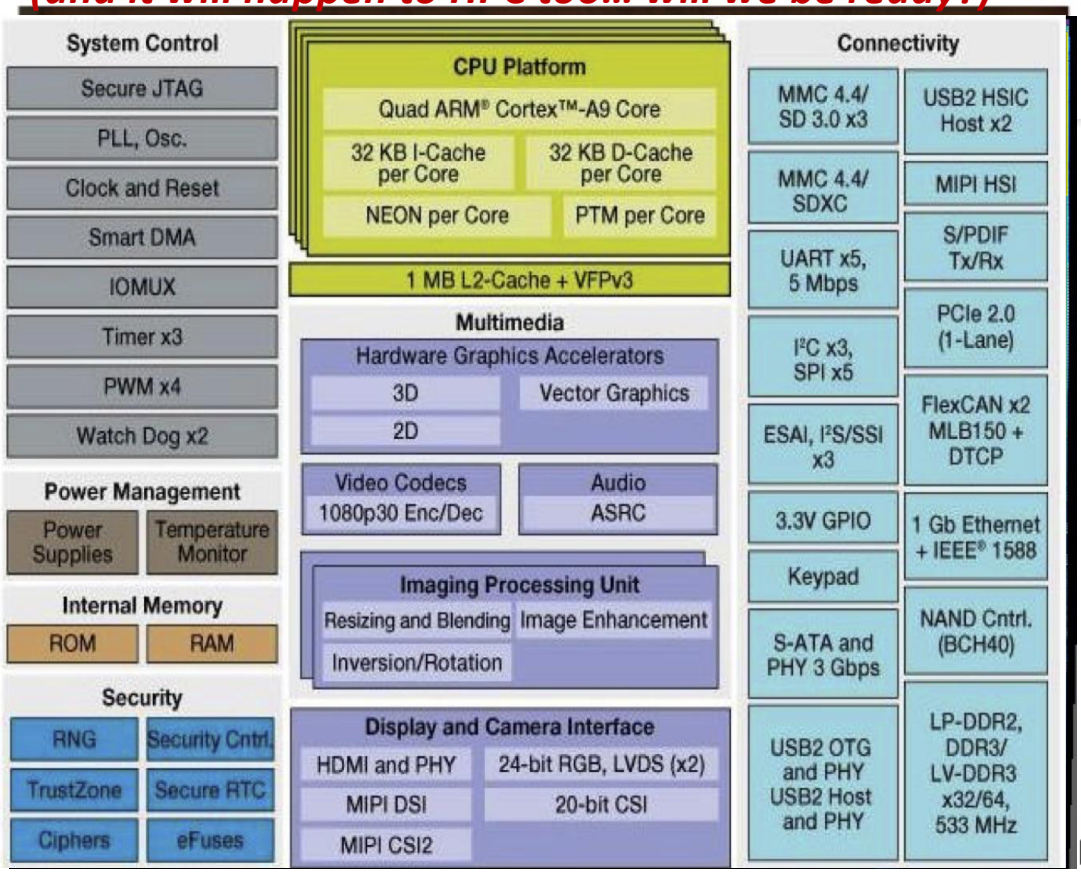
Combinatorial/NP,
Annealing/Optimization,
Simulated Atoms

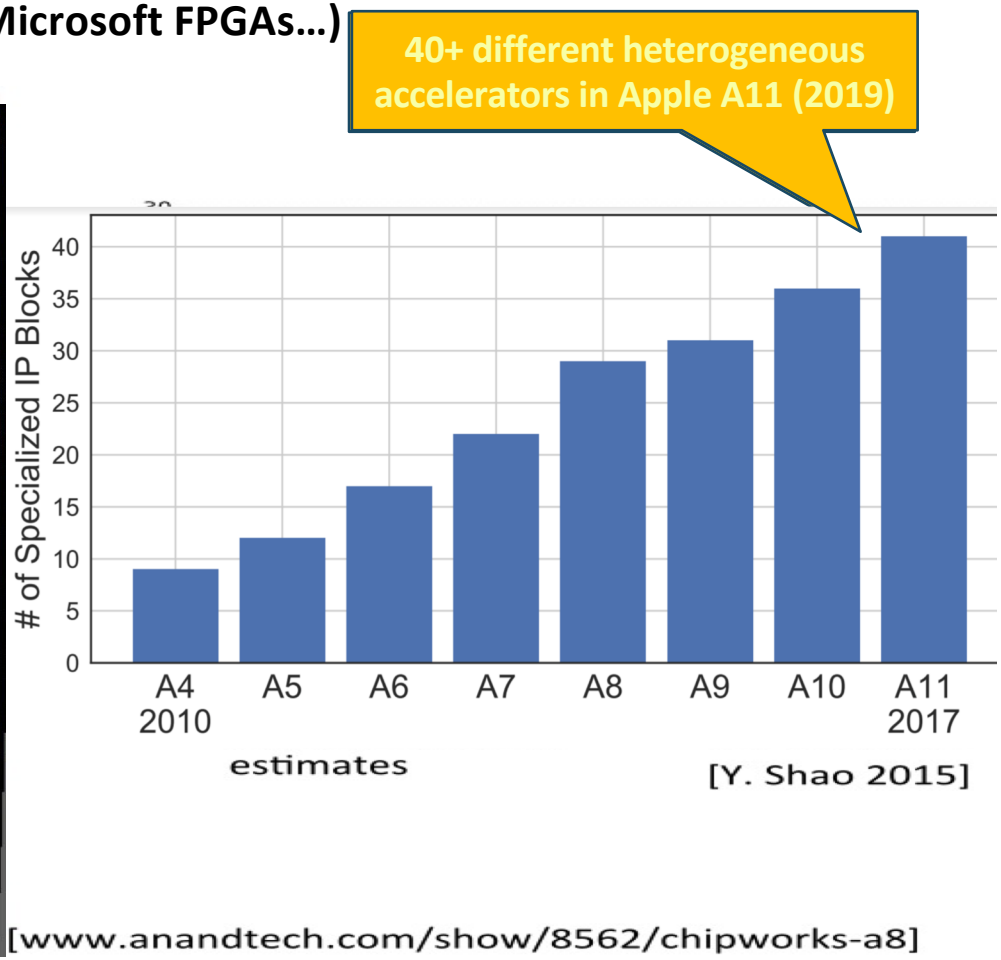
Hardware Specialization and the Move Towards Extreme Heterogenous Acceleration

Make Heterogeneous Acceleration Productive for Science

Extreme Hardware Specialization is Happening Now!

This trend is already well underway in broader electronics industry
Cell phones and even megadatacenters (Google TPU, Microsoft FPGAs…)
(and it will happen to HPC too… will we be ready?)



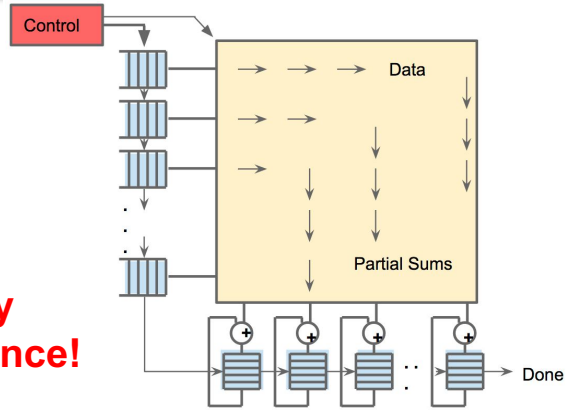


[www.anandtech.com/show/8562/chipworks-a8]

Large Scale Datacenters also Moving to Specialized Acceleration
The Google TPU

Deployed in Google datacenters since 2015

- **"Purpose Built" actually works** - *Only hard to use if accelerators was designed for something else*
- **Could we use TPU-like ideas for HPC?**
- **Specialization will be necessary to meet energy-efficiency and performance requirements for the future of DOE science!**

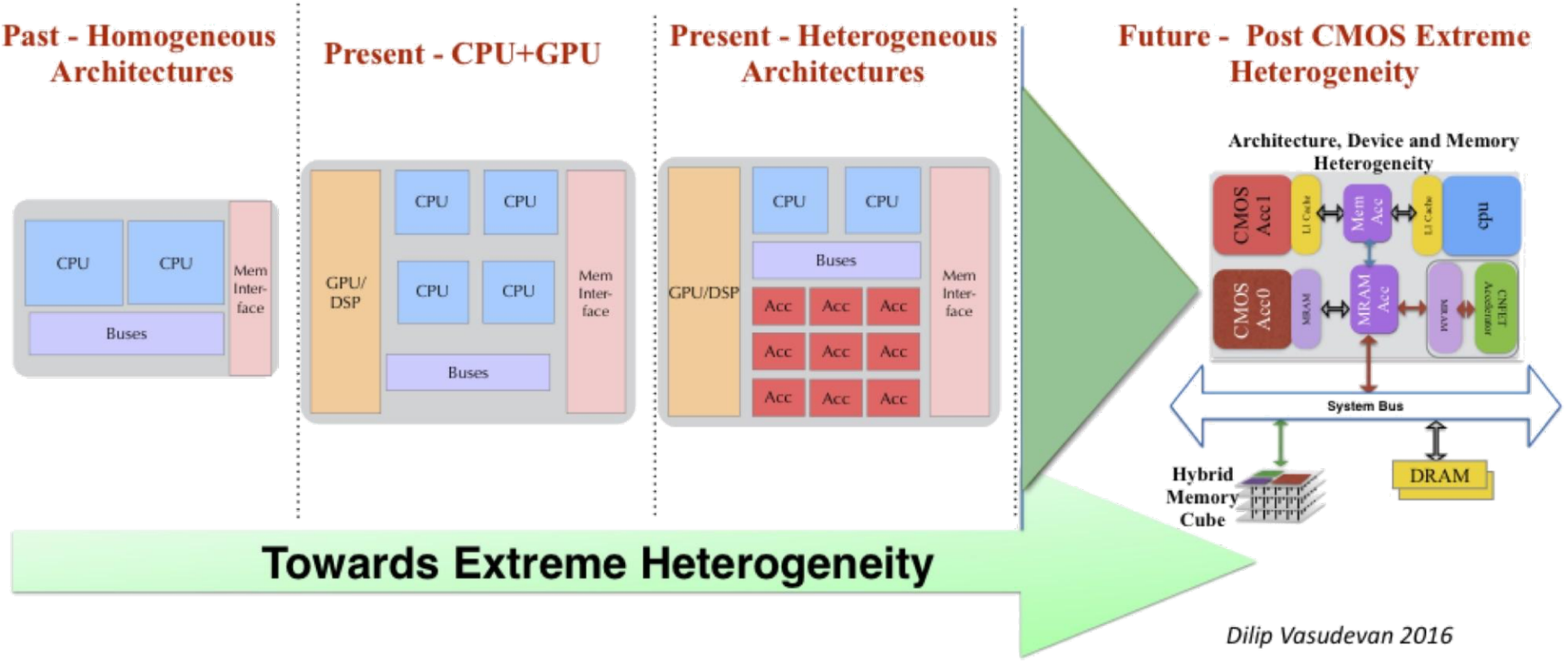


of the Matrix Multiply Unit. Software
B input is read at once, and they instantly
f 256 accumulator RAMs.

Model	MHz	Measured Watts		TOPS/s		GOPS/s /Watt		GB/s	On-Chip Memory
		Idle	Busy	8b	FP	8b	FP		
Haswell	2300	41	145	2.6	1.3	18	9	51	51 MiB
NVIDIA K80	560	24	98	--	2.8		29	160	8 MiB
TPU	700	28	40	92	--	2,300		34	28 MiB

Notional exascale system:
2,300 GOPS/W →? 288 GF/W (dp) → a 3.5 MW Exaflop system!

The Future Direction for Post-Exascale Computing



3-Pronged Strategy for Post-Exascale Performance Scaling

Deep knowledge of the mathematics is required to utilize and to design effective accelerators for science

Performance portability is necessary but insufficient…

1. Hardware Driven Algorithm Design

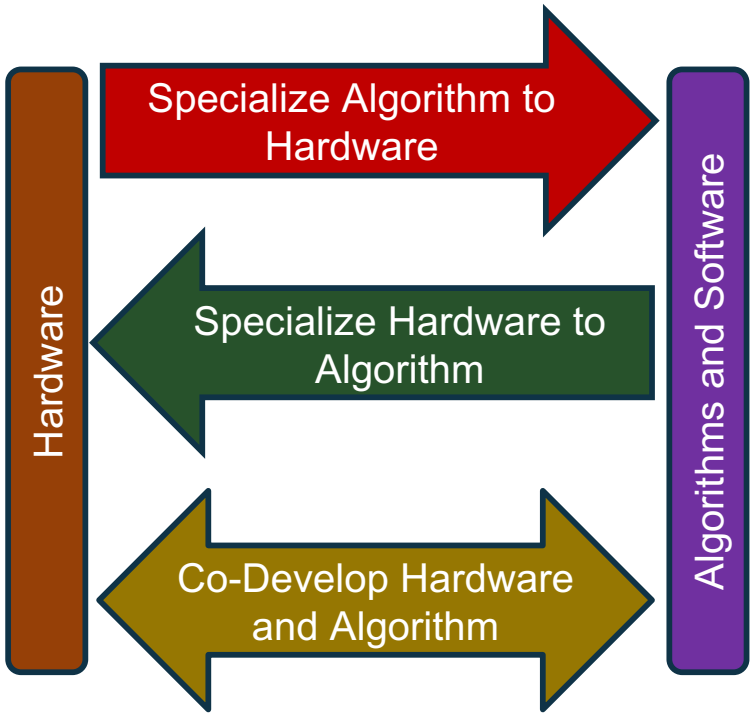
- Evaluate emerging accelerators in context of workload
- Modify algorithms to take full advantage of new accelerators

2. Algorithm Driven Hardware Design

- Design accelerators based on algorithm or application requirements
- Use FPGAs for implementation or Lab-led reference design

3. Co-Develop Hardware and Algorithms

- Design algorithms and hardware together
- Cooperative design with selected industry partner



Example: Redesigning Algorithms for Emerging Accelerators

David Keyes has a nice graph that demonstrates this is historically true!

Specialize Algorithm to Hardware

Example: SuperLU optimization paths *(Sherry Li)*

	Optimization	CPU	+ GPU offload
Baseline	Recompile	0.6x KNL slowdown	Doesn't work
Code Refactoring	Vectorization, CUDA	**1.8x** KNL speedup	**0.9 – 3x**
Algorithm Redesign	From 2D to 3D Comm-Avoiding	**0.6 - 27x**	**2.5 – 27x**

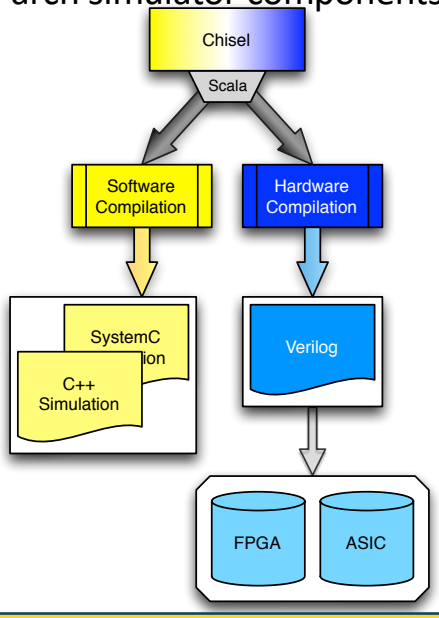
- **Why:** optimization alone is insufficient
  - Limits to what we can accomplish with code refactoring
- **How:** Algorithm redesign is necessary
- **Observation:** Performance portability only addresses refactoring; larger performance gains rely on algorithm redesign

Hardware Generators: *Enabling Technology for Exploring Design Space Together with Close Collaborations with Applied Math & Applications*

Co-Develop Hardware and Algorithm

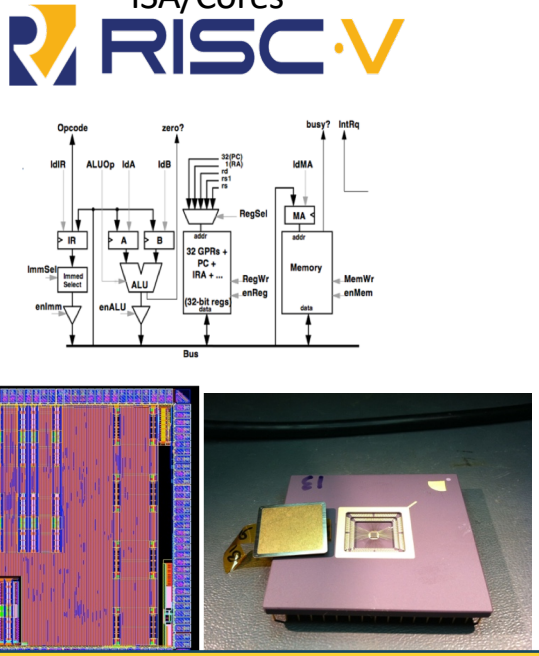
Chisel

DSL for rapid prototyping of circuits, systems, and arch simulator components



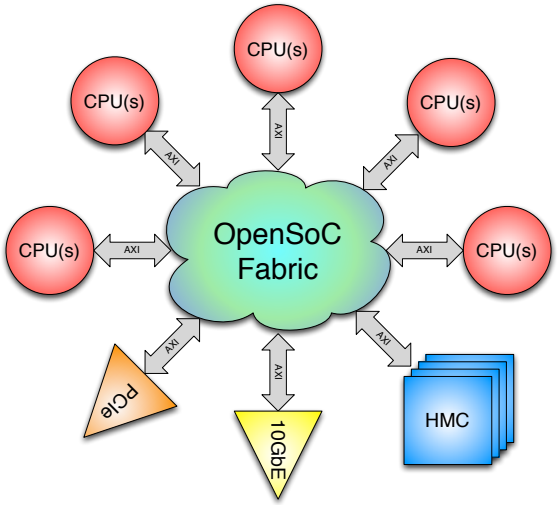
RISC-V

Open Source Extensible ISA/Cores



OpenSOC

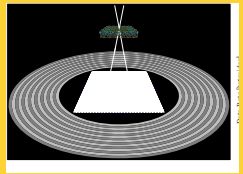
Open Source fabric To integrate accelerators And logic into SOC



SuperTools Superconducting RISC-V

QUASAR Quantum ISA

Multiagency Architecture Exploration

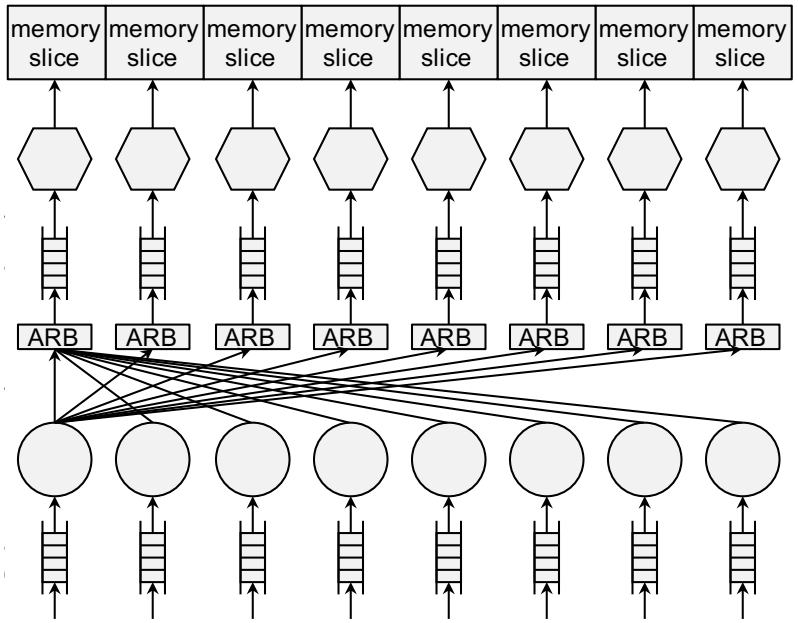


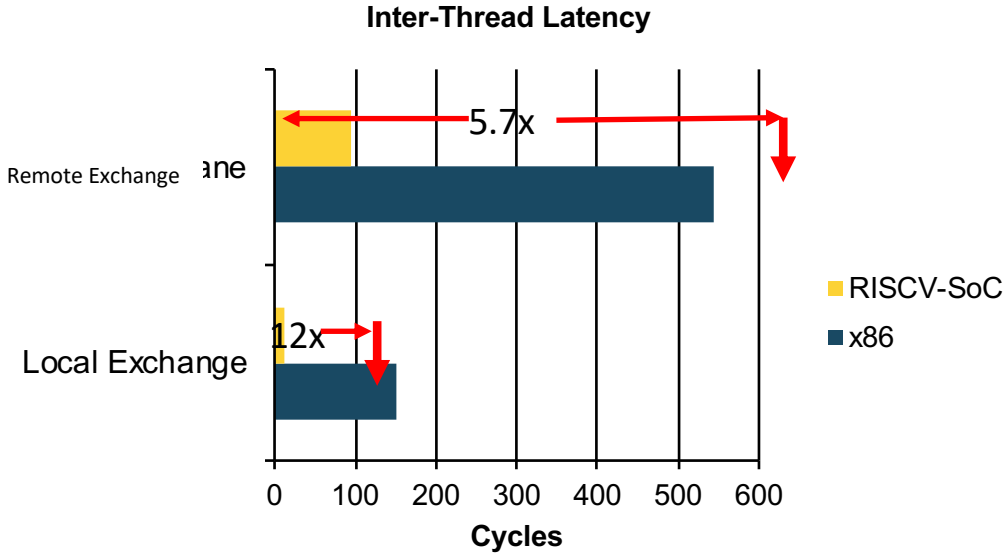
Active Sensors

Create Hardware Features to Accelerate Broadly used Numerical Algorithm Primitives

Co-Develop Hardware and Algorithm

Message Queues for low-overhead inter-core communication



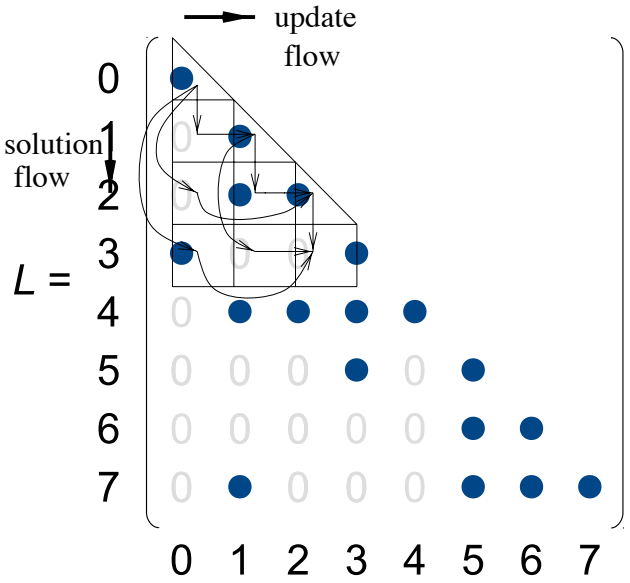


- **Tensilica TIE Queues**
- **Green Flash/Green Wave**
- **OpenSOC Fabric**

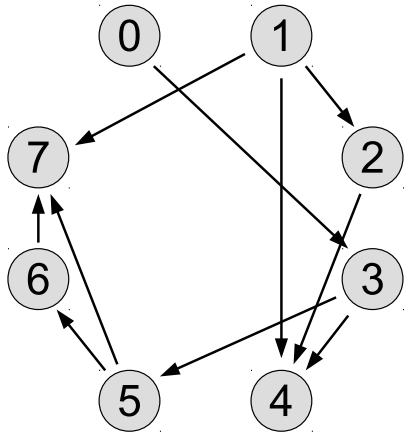
Sparse Matrix Trisolve (refresher)

Currently Use OMP Atomic to track dependencies

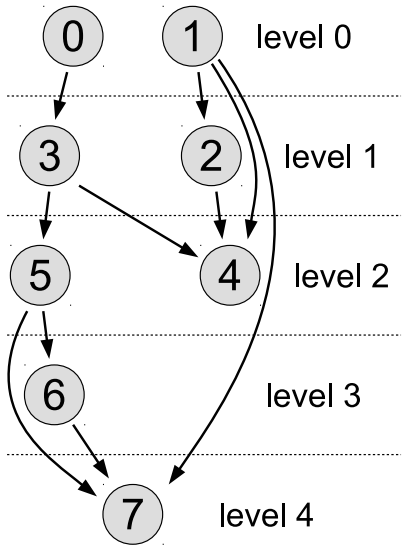
Co-Develop Hardware and Algorithm



(a) L's matrix form.



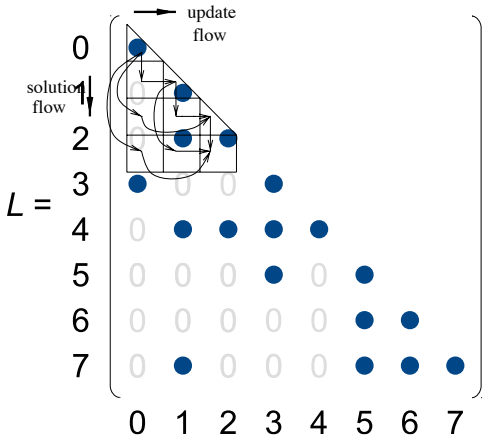
(b) L's graph form.

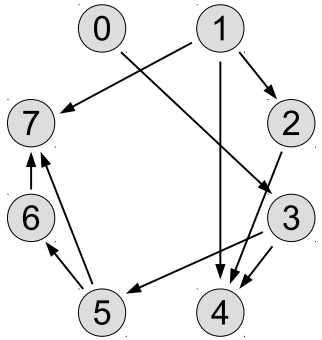


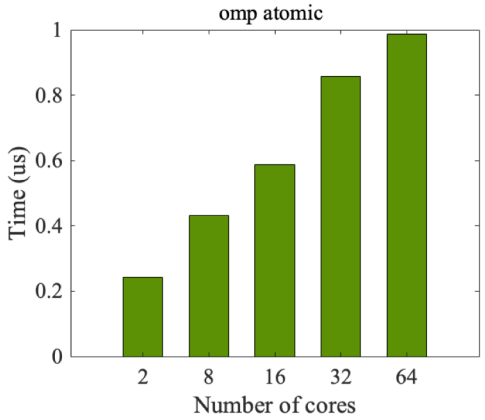
(c) Level-sets generated.

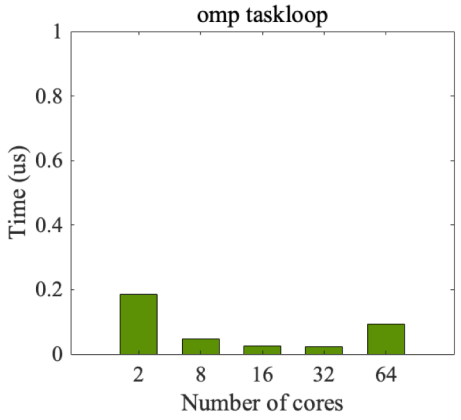
Example of CoDevelopment of Hardware and Software: SuperLU Dependency Tracking

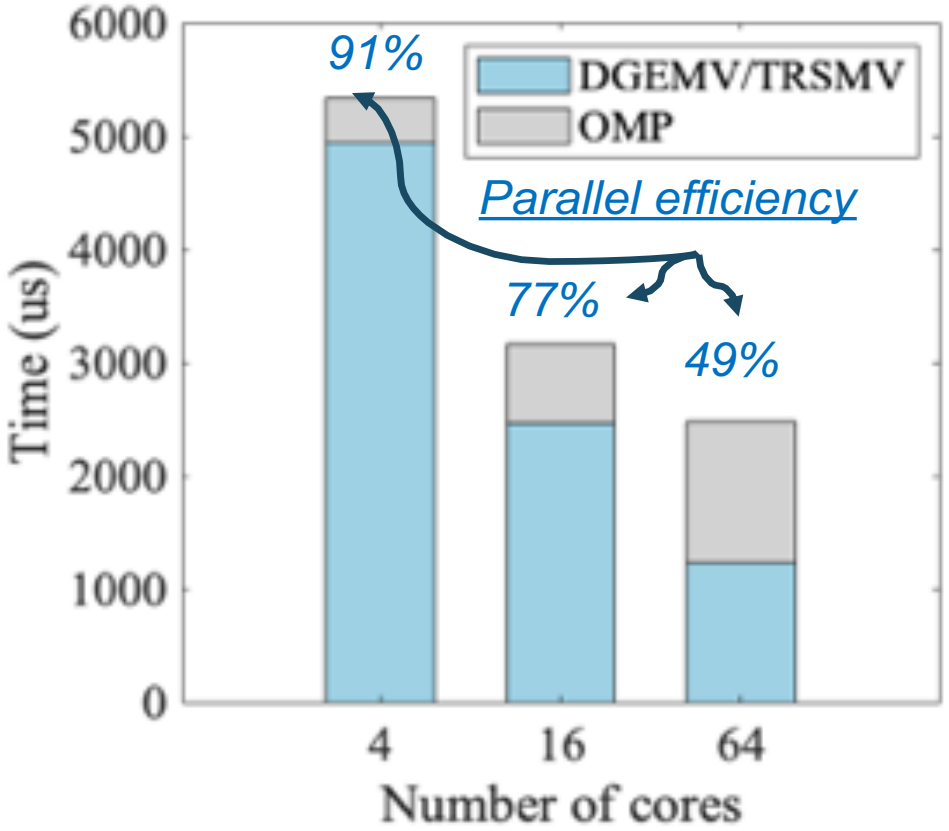
Co-Develop Hardware and Algorithm











Benefit of MsgQ's on KNL-like architecture

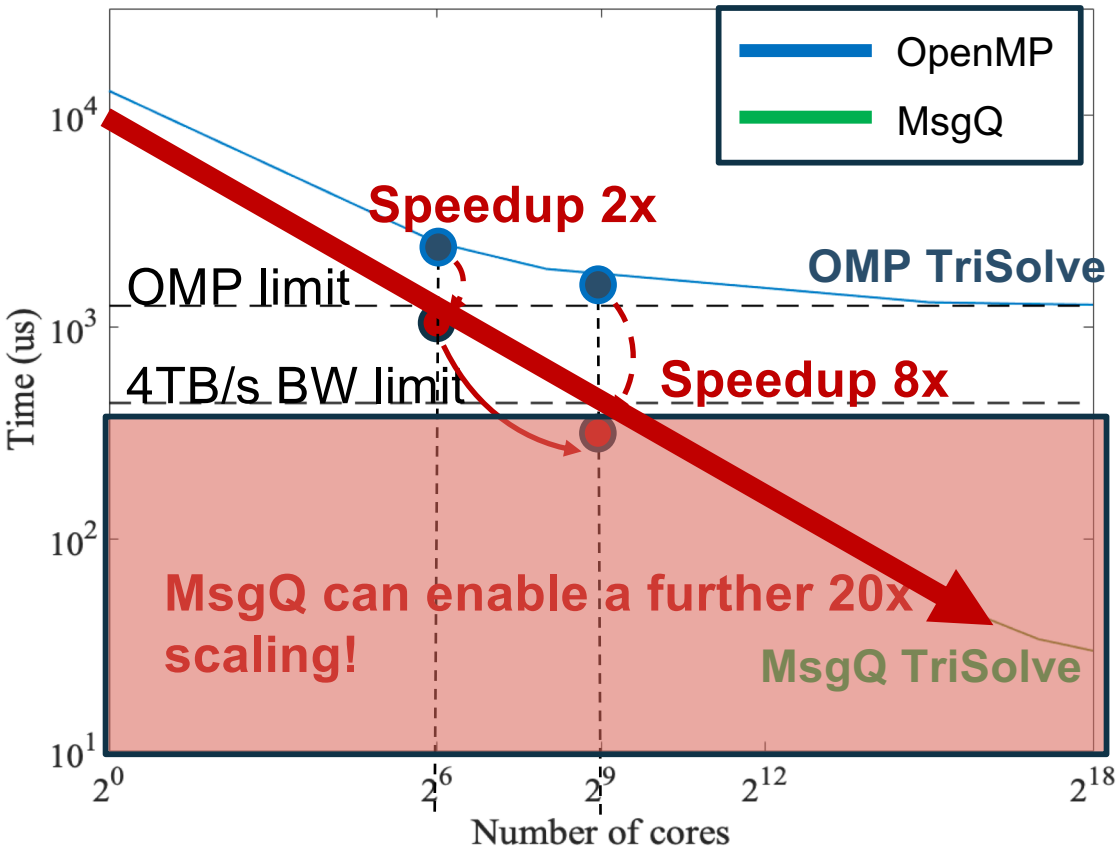
Sherry Li, Sam Williams, Yao Yang, Nan Ding)

Co-Develop Hardware and Algorithm

Algorithm: Redesign SuperLU algorithm to use MsgQ instead of atomics to track dependencies

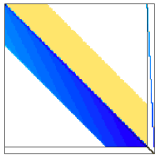
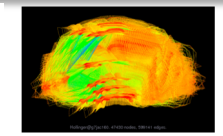
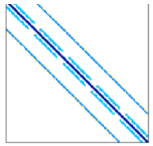
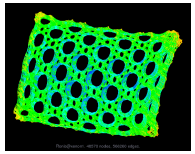
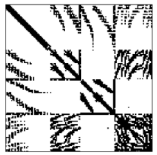
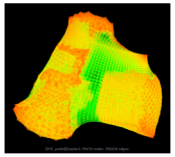
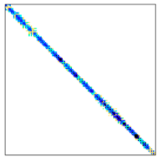
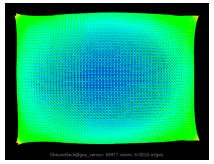
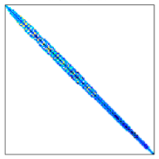
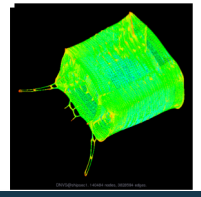
Performance:

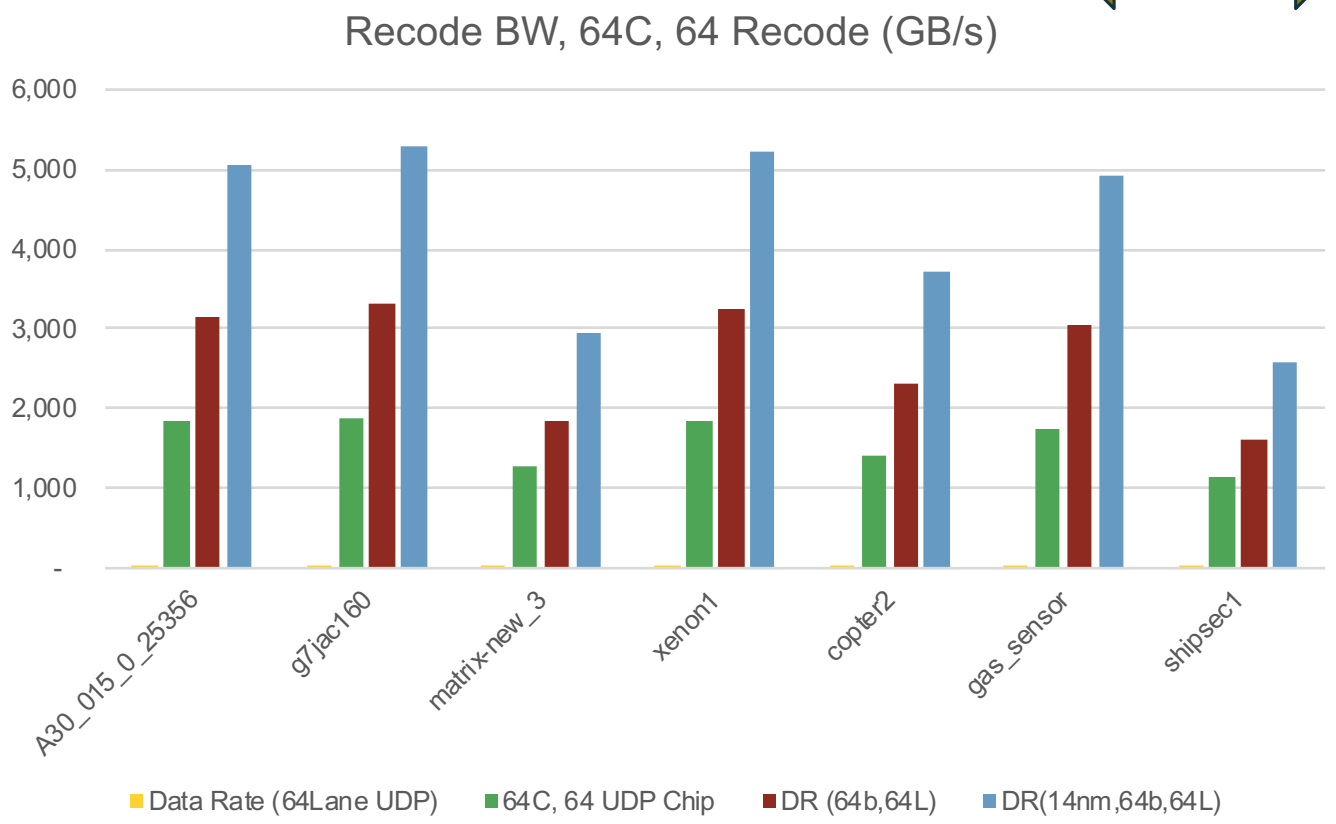
- 12x lower overhead per message than OpenMP
- 4x faster than OpenMP for 64cores
- Potential for 8x-20x further scaling



SNAPPY: Sparse Matrix Compression Accelerator

Co-Develop Hardware and Algorithm

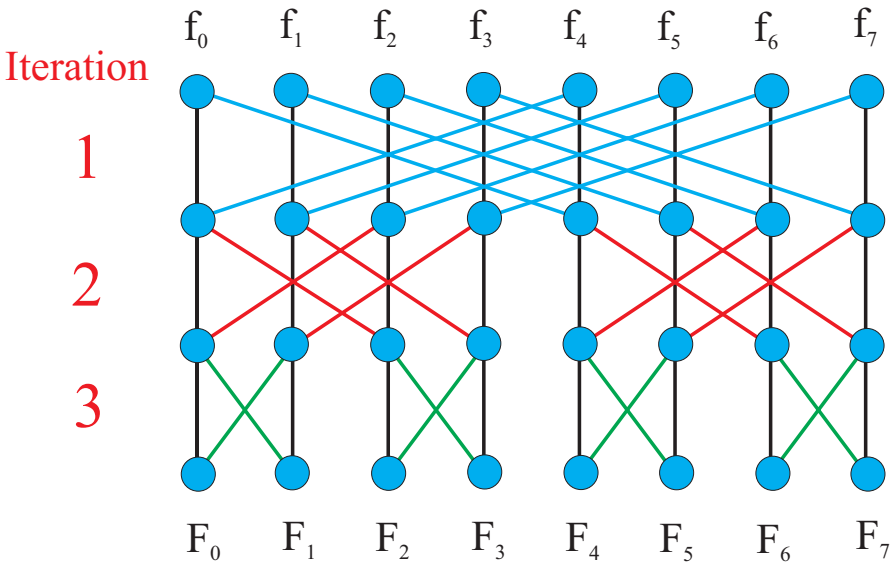
Matrices	Spyplot	Visualization
g7jac160		
Xenon1		
Copter2		
Gas sensor		
Shipsec1		



FFT Discrete Accelerator Hardware Design Study

With FFTx (Francetti, Popovic, Canning)

Co-Develop Hardware and Algorithm



- **Assumptions**
  - HPC Challenge Benchmark: Single precision complex, out-of-place
  - FFT logic generated by SPIRAL
  - 14nm technology node (1Ghz)
  - Synthesized w/Mentor toolflow

100 GB/s off-chip memory BW

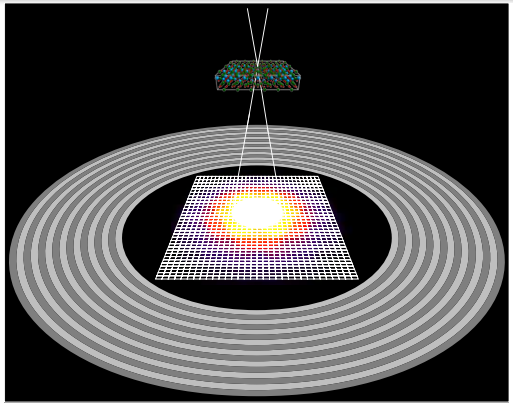
- **FP limit ~1.5TFLOPs SP**
- **4.5mm² area for compute logic**

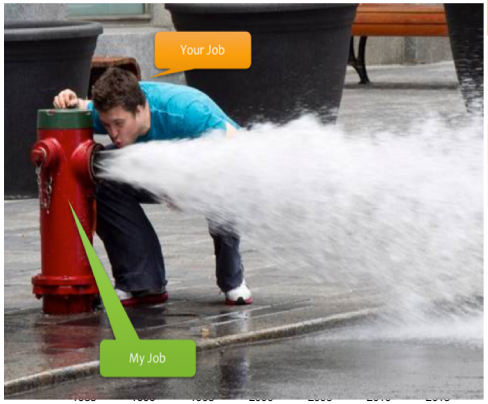
1TB/s off-chip memory BW

- **FP limit ~15TFLOPs SP**
- **36mm² area for compute logic**

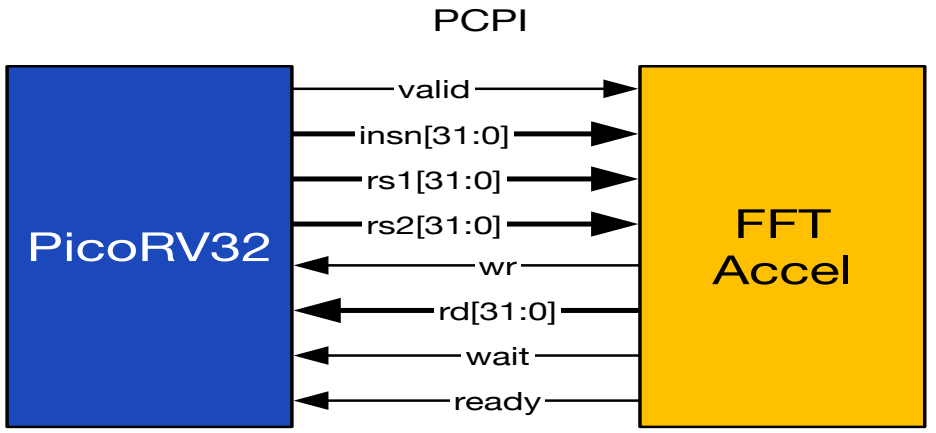
Results for RISC-V FFT Accelerator for CryoEM

Benchmarking FFT Accelerator for image analysis

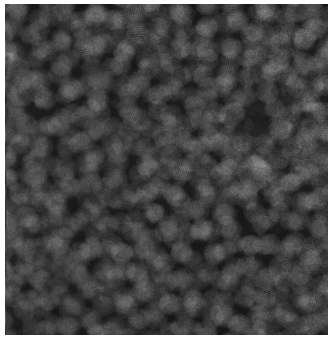




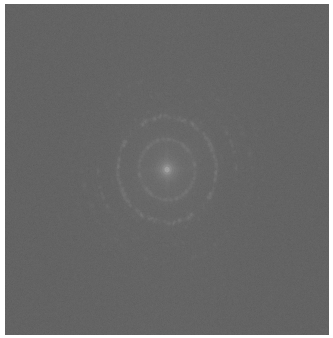
Instruction	opcode[3:2]	Description
fft_config	10b	Configures FFT parameters
fft_status	01b	Reads FFTAccel status registers
fft_start	11b	Starts FFT processing
fft_stop	00b	Stops FFT processing



Original Image



FFT



Created RISC-V Core with FFT ISA Extension

RISC-V+FFT Accel 126x faster than x86 host

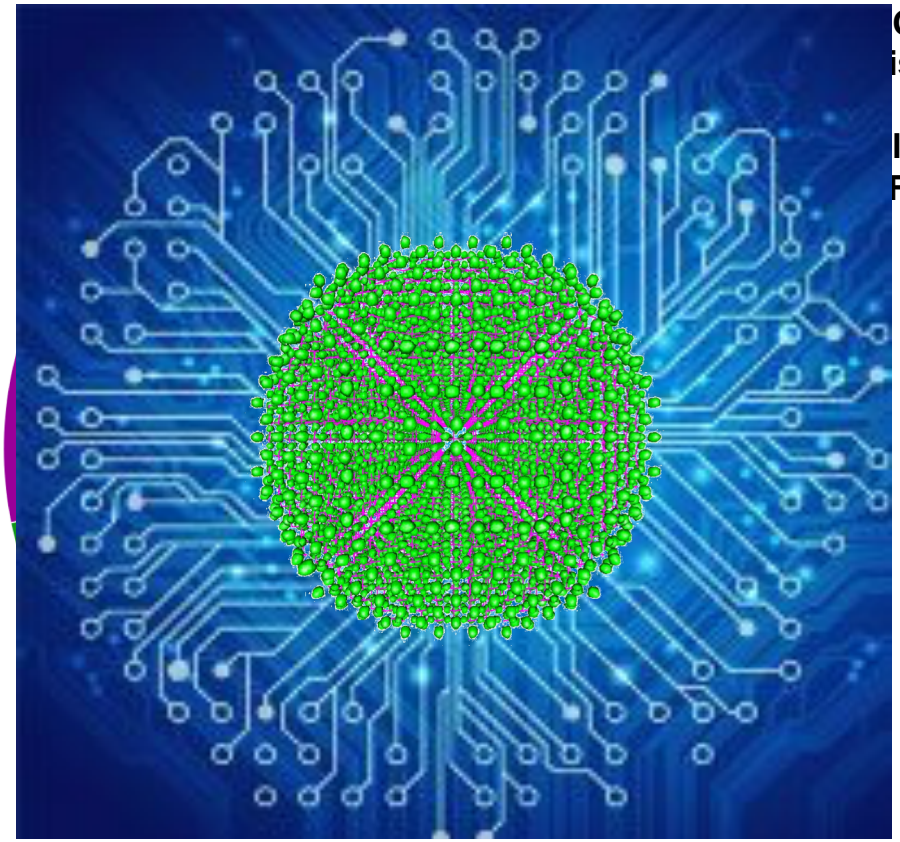
- FFT on Intel Core i7-5930K @ 3.50GHz: **~265ms**
- FFTAccel (Floating): **~2.10ms**

Example Algorithm-Driven Design of Hardware Accelerators

Specialize Hardware to Algorithm

Example: LS3DF/Density Functional Theory (DFT)

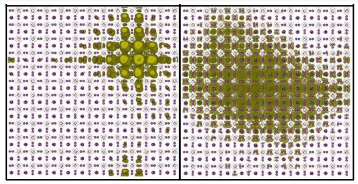
- **What: Design the hardware accelerator around the target algorithm/application**
  - Purpose-built acceleration
  - Lab-led reference design
- **Why: Huge opportunities to improve performance density and efficiency**
  - FFT hardware accelerator 50x-100x higher performance density than GPU or CPU+SIMD (using SPIRAL generator)
- **How: Use Density Functional Theory (DFT) as the target for this experiment**
  1. Large fraction of the DOE workload
  2. Mature code base and algorithm
  3. LS3DF formulation minimizes off-chip communication and scales O(N)

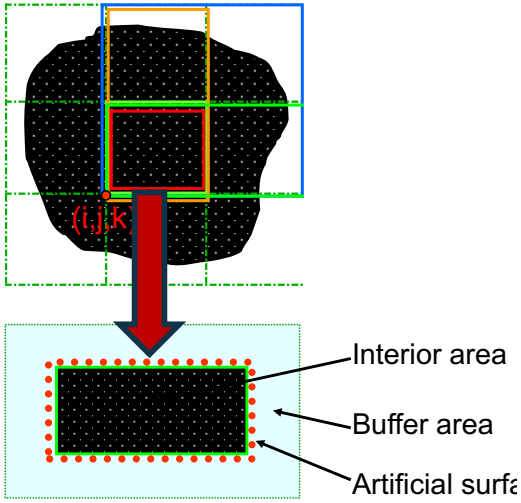


OE is I FT)

The DFT kernel for each fragment

Communication Avoiding LS3DF Formulation – Scales O(N)





$h(i,j) = \langle \psi_i | H | \psi_j \rangle$

Sub_diag, *

Hpsi, *

Precond. CG step

Projection, *

Line minimiz.

Orth., *

$h(i,j) = \langle \psi_i | H | \psi_j \rangle$

Sub_diag, *

FPGA

$O(N^2 Log(N))$

Comm bound if non-local

3D parallel FFT

TSQR & Choelesky
ZGEMM

$O(N^3)$

Compute-bound

LS3DF O(N) Algorithm Formulation
Minimizes off-chip Communication

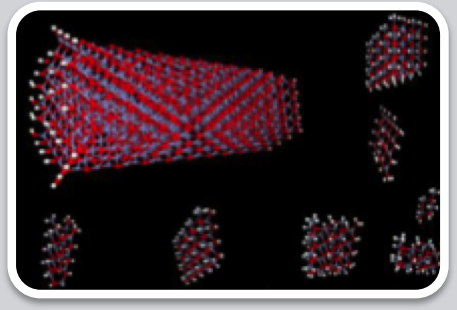
One patch per FPGA
400 bands/patch

Compute Intensive Kernels
Targeted for HW Specialization

BERKELEY LAB

Architecture Specialization for Science

(hardware is design around the algorithms) can't design effective hardware without math



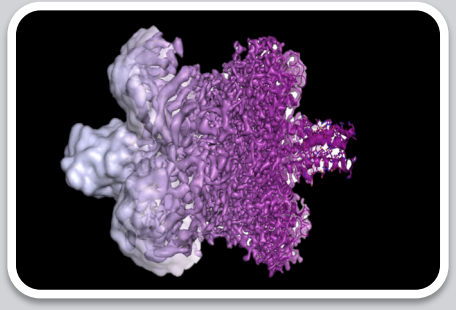
Materials

Density Functional Theory (DFT)

Use O(n) algorithm

Dominated by FFTs

FPGA or ASIC



CryoEM Accelerator

LBNL detector

750 GB / sec

Custom ASIC near detector

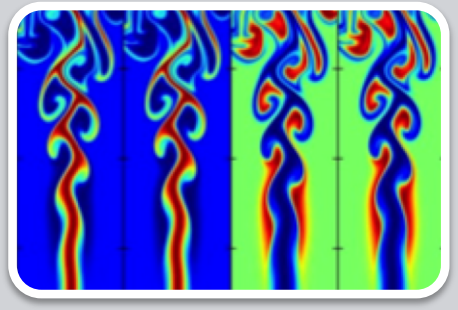
Genomics Accelerator

String matching

Hashing

2-8bit (ACTG)

FPGA



Digital fluid Accelerator

3D integration

Petascale *chip*

1024-layers

General / special HPC solution

What is a sustainable economic model for the HPC Ecosystem? *(what can we learn from the rest of industry?)*

- **Sustained Partnerships (BlueGene, RedStorm)**
  - must be > 4 years (challenging with current procurement rules)
- **Full Custom for Targeted Workload (Google TPU)**
  - Full Custom for targeted workload
- **General Purpose/ARM with Accelerators for Common Functions (Amazon)**
  - Charge per event activation rather than per hour of VM time
  - Prototype on FPGA (F1 instances) move to ARM+accelerators ASIC (e.g. Graviton)
- **Rack Disaggregation (Google and Facebook)**
  - Buy accelerators, memory and resources a-la-carte and config. custom nodes on-the-fly
- **Inverted Procurement Model (Open Compute Project: *Facebook/Microsoft/Amazon)***
  - Create open reference design specification and have integrators bid against it (using their own proprietary innovations if they want to)
  - Ensures a consistent software interface for customer and multiple suppliers

Data Movement Challenge

Photonics and Advanced Packaging

Data Movement Costs:

Energy to move data proportional to distance. *Power is near chip thermal limits*

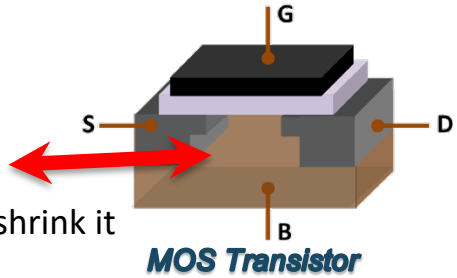
Energy Efficiency of copper wire:

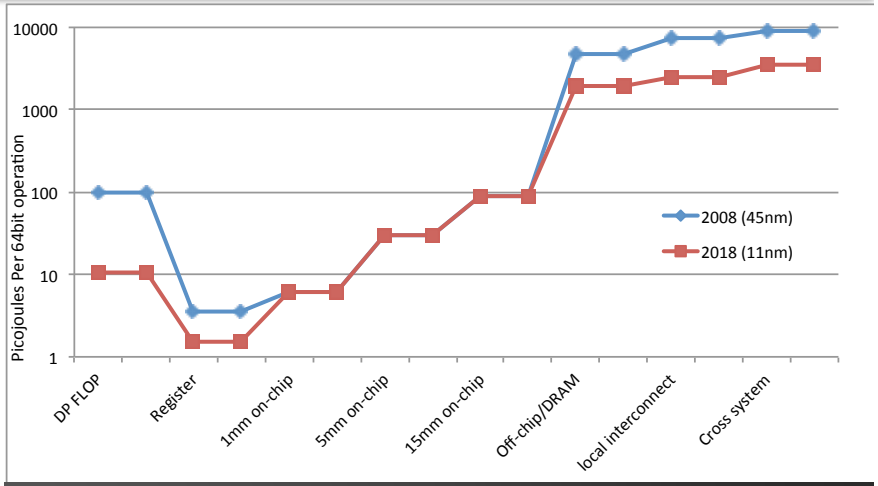
- **Power = Frequency* Length / cross-section-area**

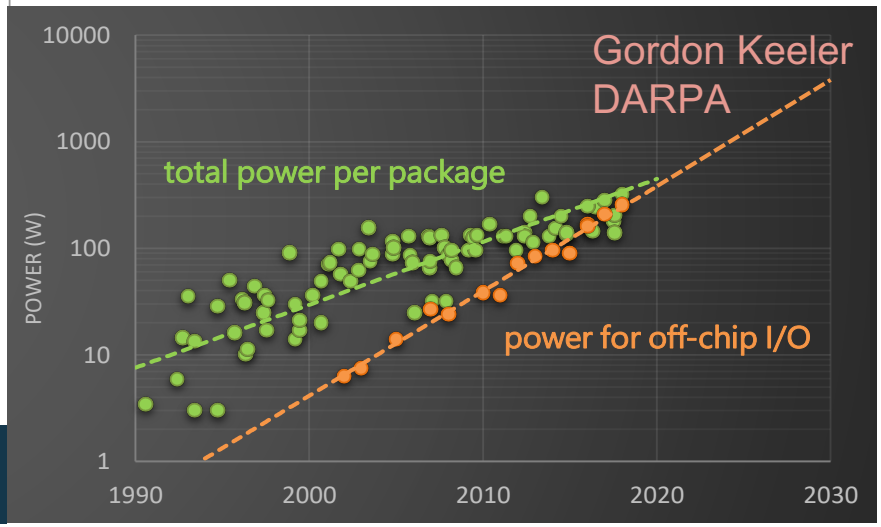
- Wire efficiency *does not improve* as feature size shrinks

Energy Efficiency of a Transistor:

- Power = V^2 * frequency * Capacitance
- Capacitance ~= Area of Transistor
- Transistor efficiency improves as you shrink it







Net result is that moving data on wires is starting to cost more energy than computing on said data (interest in Silicon Photonics)

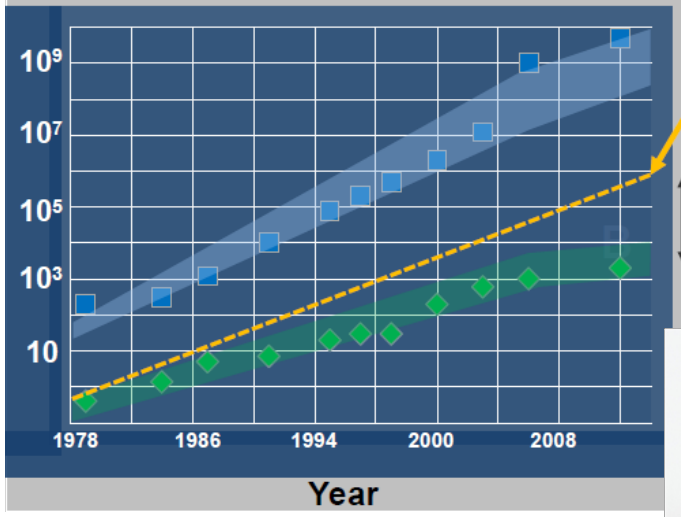
BERKELEY LAB

Package Performance is Pin Limited

Rent's Rule: *J. Poulton: NVIDIA*

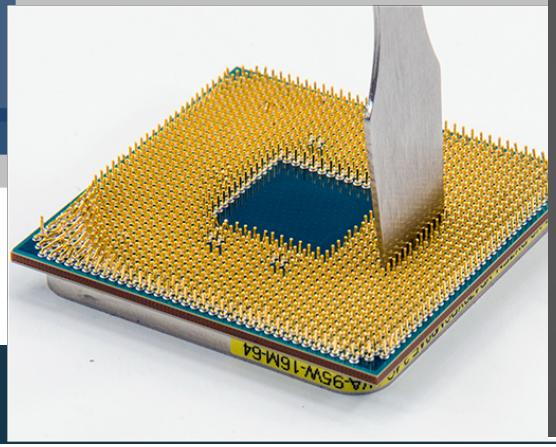
Number of pins = K x Gates^a (IBM, 1960)

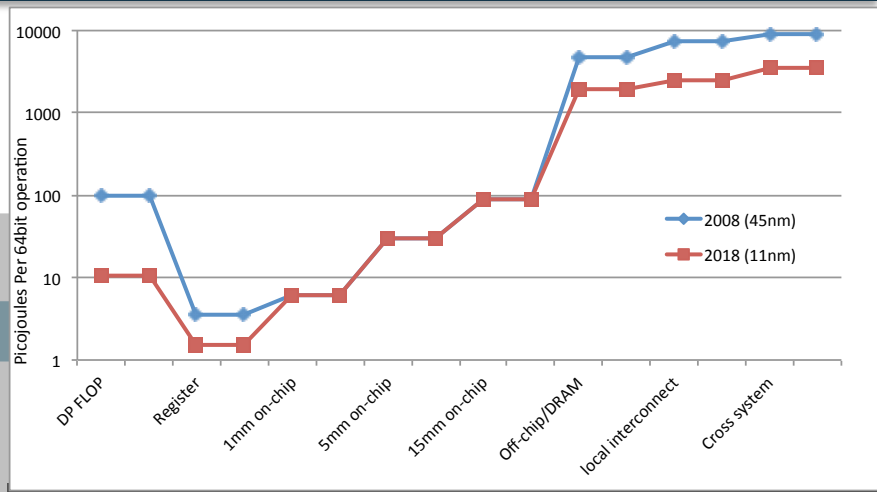
K = 0.82, a = 0.45 for early Microprocessors

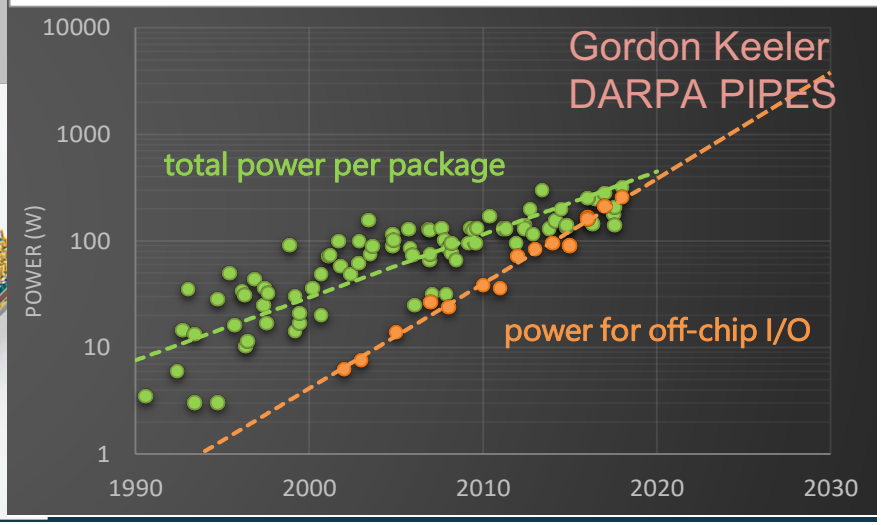


Pins x GHz from Rent's Rule

Bandwidth Gap: ~500 x and growing!







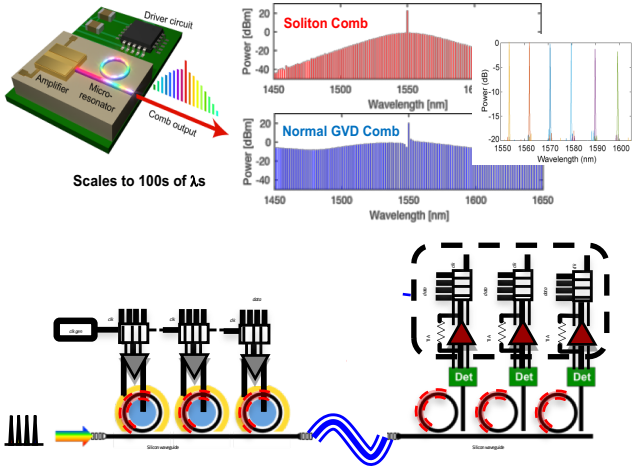
High SERDES rates run counter to end of Dennard Scaling

BERKELEY LAB

PINE: Photonic Integrated Networked Energy Efficient Datacenters

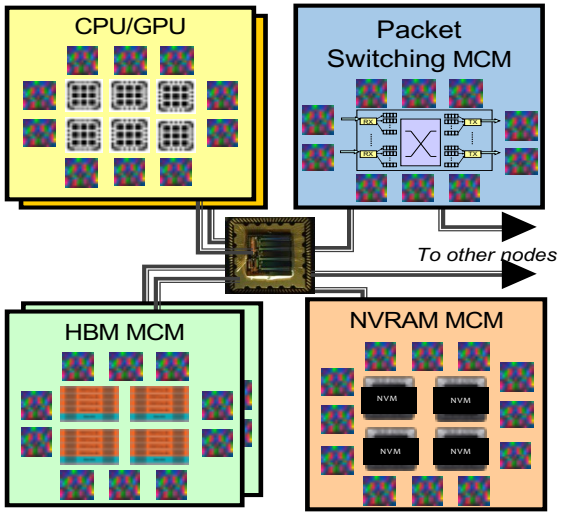
Resource Disaggregation to custom-assemble diverse accelerators for diverse workload requirements

1) Energy-bandwidth optimized optical links

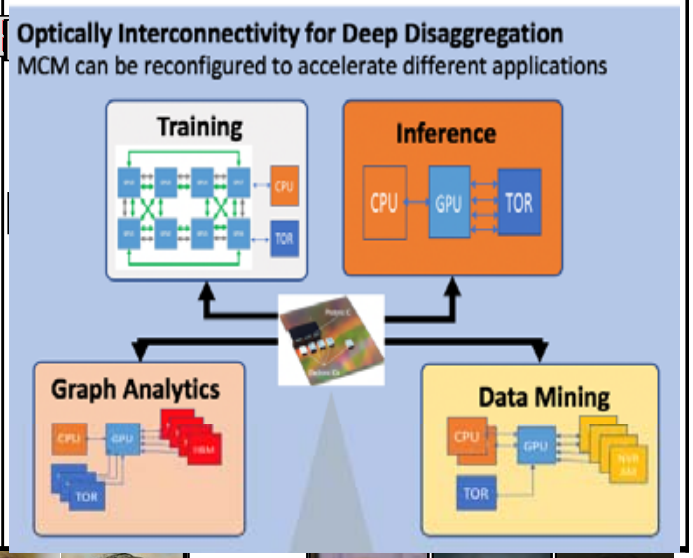


1 Tb/second per fiber

2) Embedded silicon photonics into OC-MCMs



3) Bandwidth steering for Custom Node Connectivity

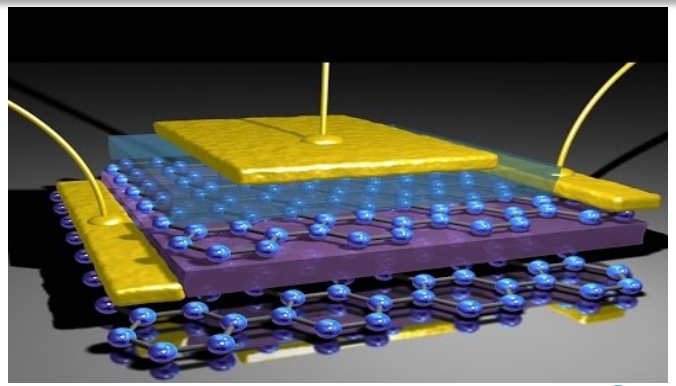


ENLITENED

arpa·e
CHANGING WHAT'S POSSIBLE

Bergman

Johansson — FREEDOM PHOTONICS
Coolbaugh — SUNY Polytechnic Institute
Bowers — UCSB
Gaeta
Lipson
Kinget
Patel — CISCO
Dennison — NVIDIA
Shalf — Berkeley Lab
Ghobadi — Microsoft



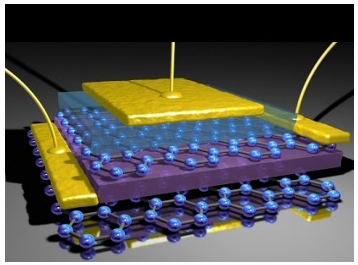
Post CMOS Device Technology

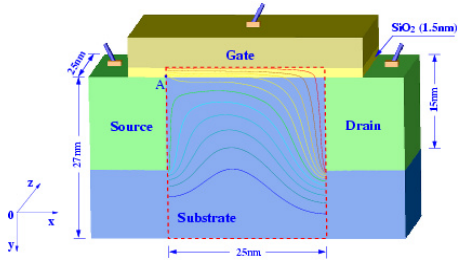
Accelerating the pace for discovery
for the future of Microelectronics

Many Options for New Device Technology

but few satisfy Borkar-Shalf Criteria (2013-2015 viewpoint)

1. **Gain**
2. **Signal to Noise**
3. **Scalability**
4. **Manufacturability**



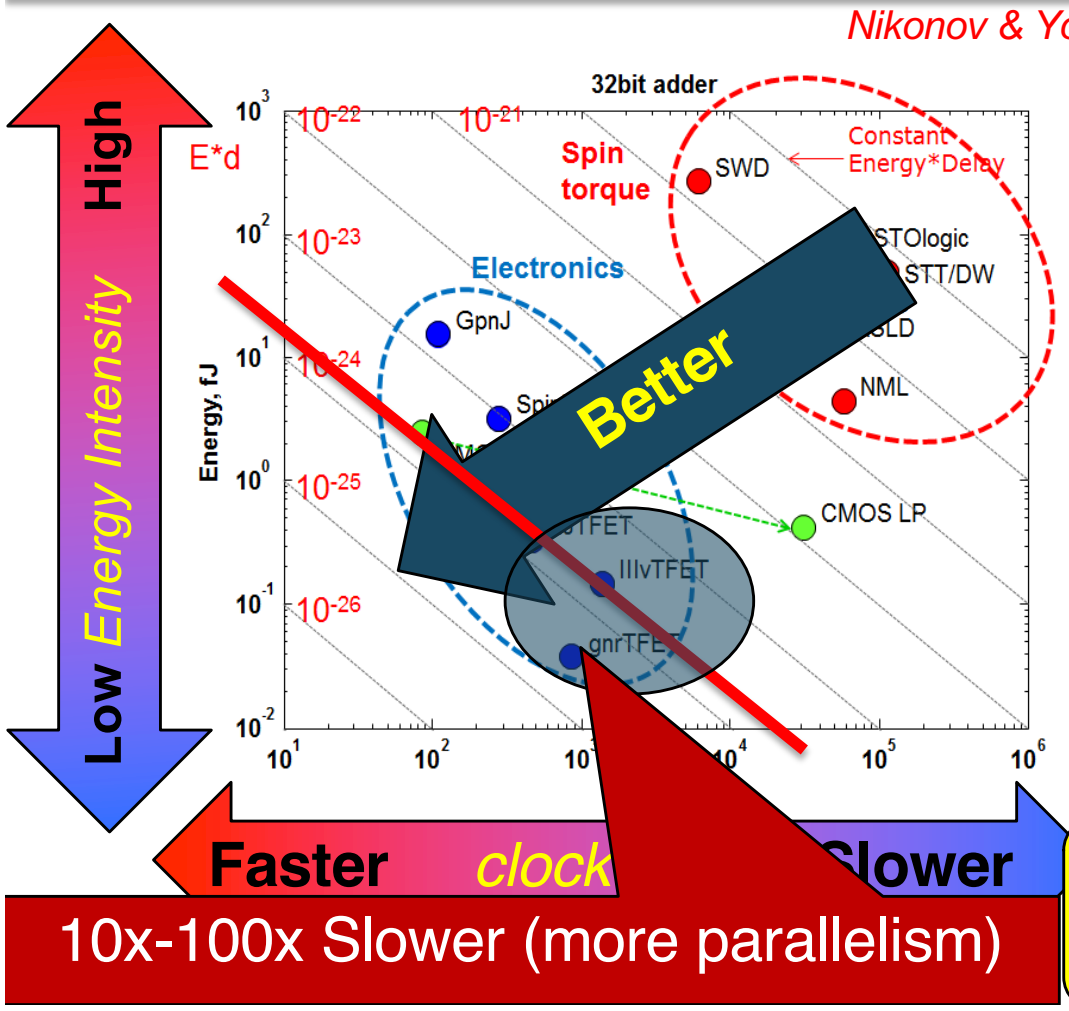


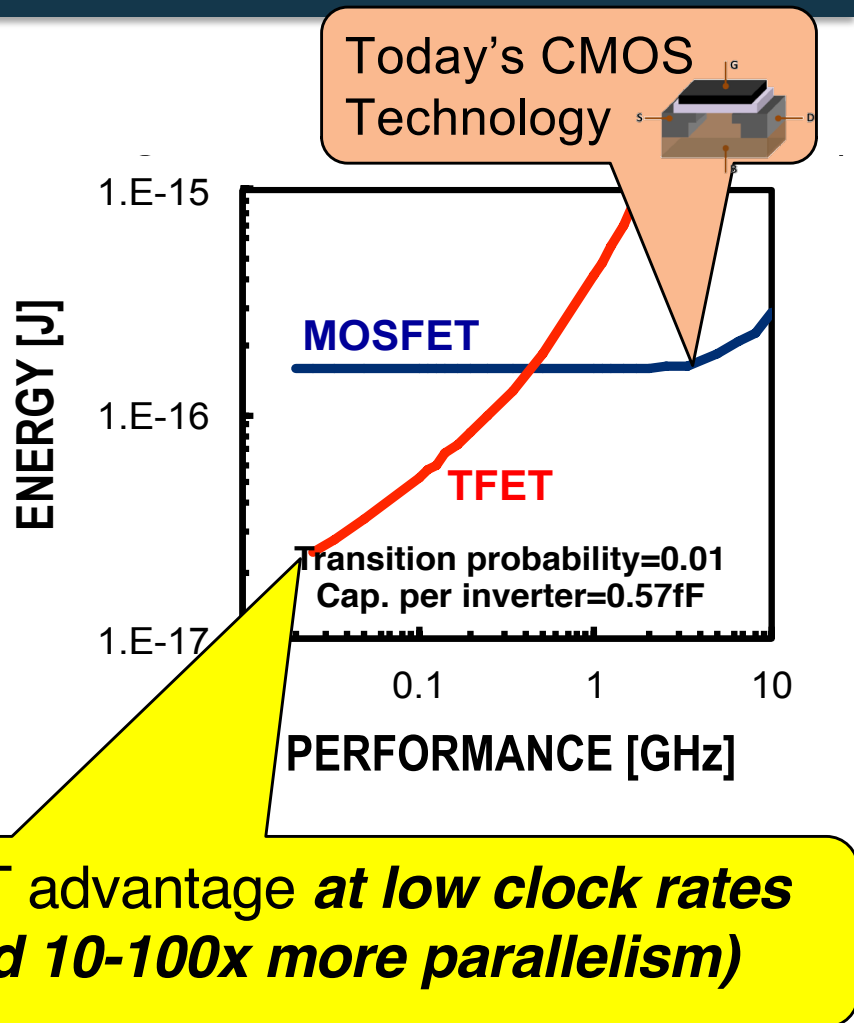
OSTP Report 2015: John Shalf
Robert Leland and Shekhar Borkar

TABLE 1. Summary of techology options for extending digital electronics.

Improvement Class	Technology	Timescale	Complexity	Risk	Opportunity
Architecture and software advances	Advanced energy management	Near-Term	Medium	Low	Low
	Advanced circuit design	Near-Term	High	Low	Medium
	System-on-chip specialization	Near-Term	Low	Low	Medium
	Logic specialization/dark silicon	Mid-Term	High	High	High
	Near threshold voltage (NTV) operation	Near-Term	Medium	High	High
3D integration and packaging	Chip stacking in 3D using thru-silicon vias (TSVs)	Near-Term	Medium	Low	Medium
	Metal layers	Mid-Term	Medium	Medium	Medium
	Active layers (epitaxial or other)	Mid-Term	High	Medium	High
Resistance reduction	Superconductors	Far-Term	High	Medium	High
	Crystaline metals	Far-Term	Unknown	Low	Medium
Millivolt switches (a better transistor)	Tunnel field-effect transistors (TFETs)	Mid-Term	Medium	Medium	High
	Heterogeneous semiconductors/strained silicon	Mid-Term	Medium	Medium	Medium
	Carbon nanotubes and graphene	Far-Term	High	High	High
	Piezo-electric transistors (PFETs)	Far-Term	High	High	High
Beyond transistors (new logic paradigms)	Spintronics	Far-Term	Medium	High	High
	Topological insulators	Far-Term	Medium	High	High
	Nanophotonics	Near/Far-Term	Medium	Medium	High
	Biological and chemical computing	Far-Term	High	High	High

Comparing CMOS Technology Alternatives

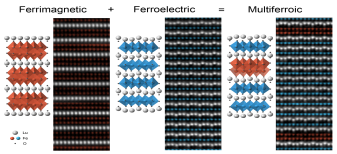
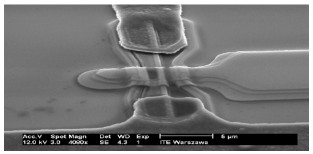
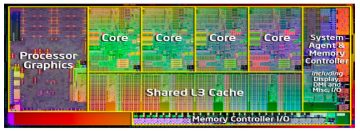
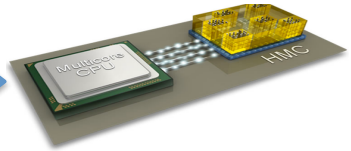




TFET advantage ***at low clock rates (need 10-100x more parallelism)***

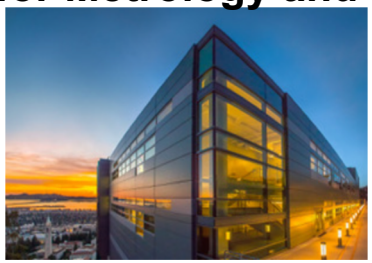
Integrated Plan to Accelerate Microelectronics Discovery

End-to-End Acceleration of Discovery and Evaluation of New Devices

Materials Discovery	ME Transistor	Architecture	System
			
Computational Design Synthesis Characterization	Device Design Fabrication Parametrics	RTL/Gate Simulator Power Delay	Arch. Level Simulator TDP, EDP

National User Facilities for Metrology and Experimental Validation





ADVANCED LIGHT SOURCE

MOLECULAR FOUNDRY

NERSC National Energy Research Scientific Computing Center

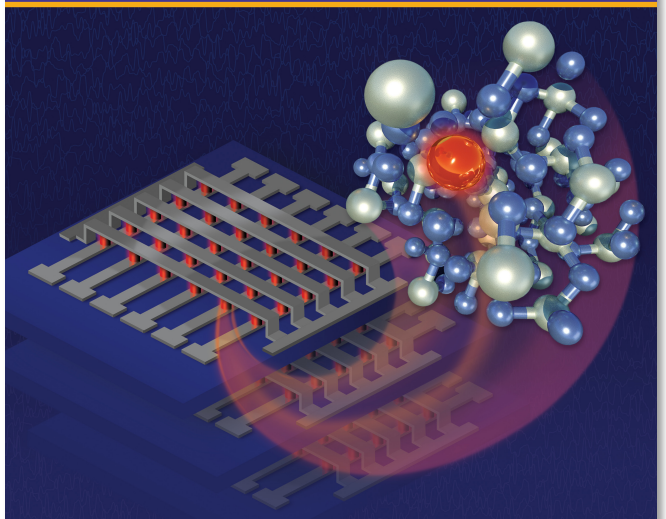
Berkeley UNIVERSITY OF CALIFORNIA

EUREKA CXRO

Physical, Chemical, Materials and Computer Sciences

DOE Microelectronics BRN (BES, HEP, ASCR)

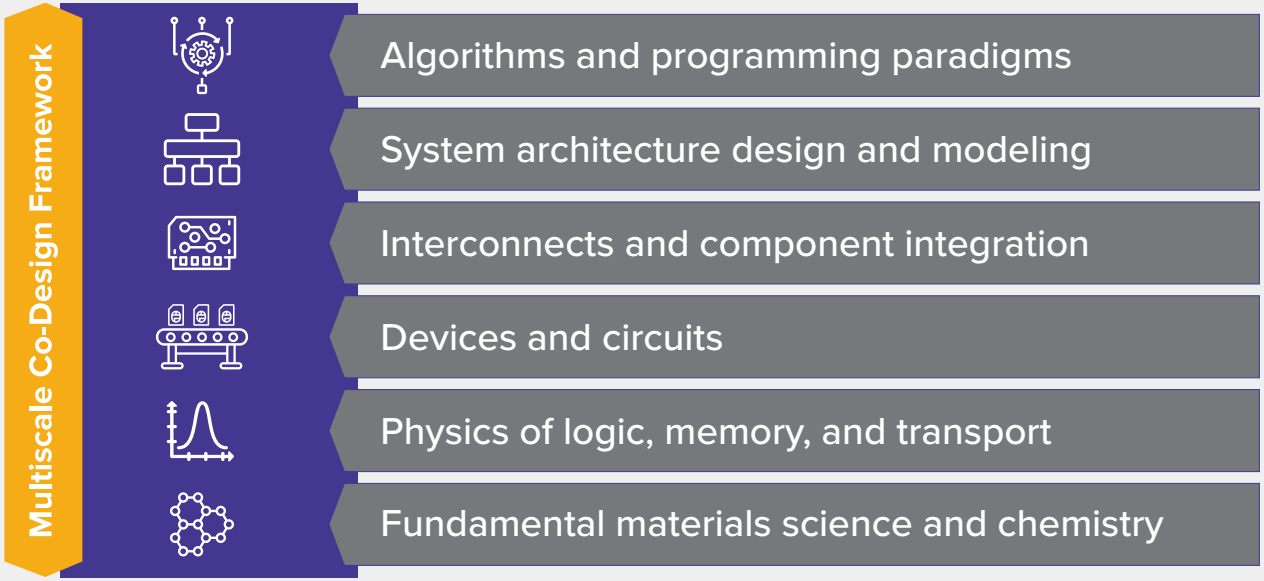
Basic Research Needs for
Microelectronics



Discovery science to revolutionize microelectronics beyond today's roadmaps

We need to accelerate the pace of discovery by orders of magnitude

Deep Microelectronics CoDesign Framework

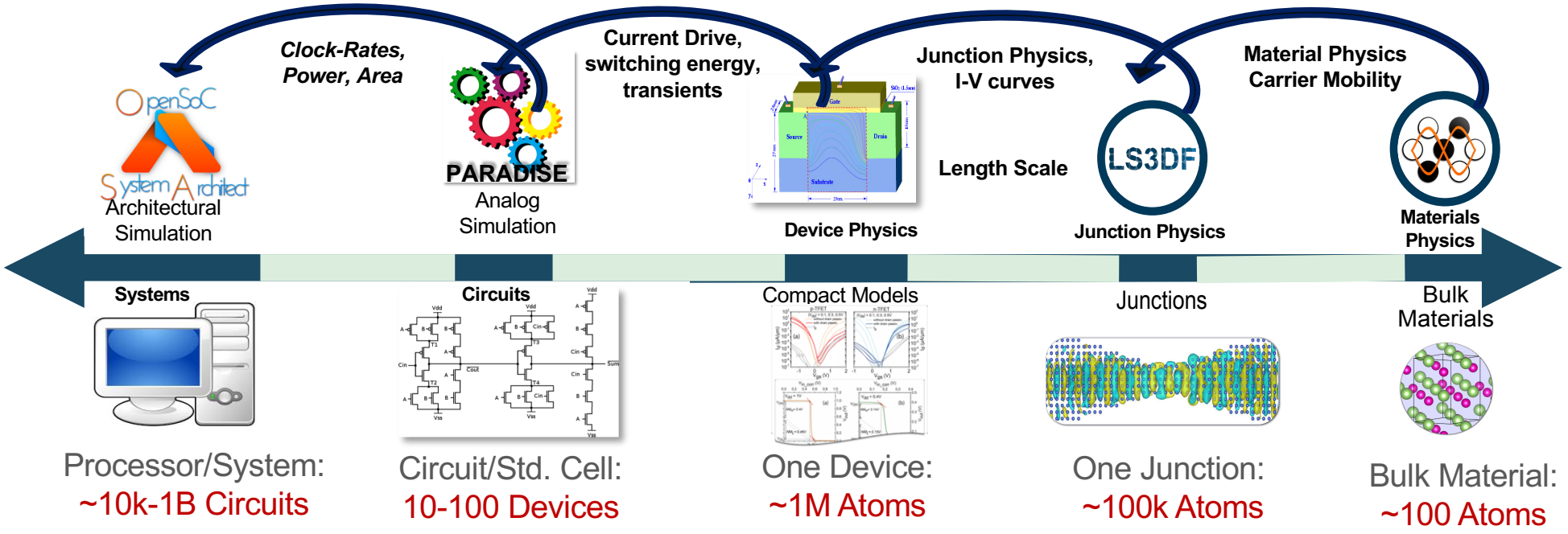


Co-design involves multi-disciplinary collaboration that takes into account the interdependencies among materials discovery, device physics, architectures, and the software stack for developing information processing systems of the future. Such systems will address future DOE needs in computing, power grid management, and science facility workloads.

BERKELEY LAB · Argonne NATIONAL LABORATORY · OAK RIDGE National Laboratory · Los Alamos NATIONAL LABORATORY EST. 1943 · Lawrence Livermore National Laboratory · Pacific Northwest NATIONAL LABORATORY · Sandia National Laboratories

Multiscale Modeling to cceleRate Post-CMOS Development

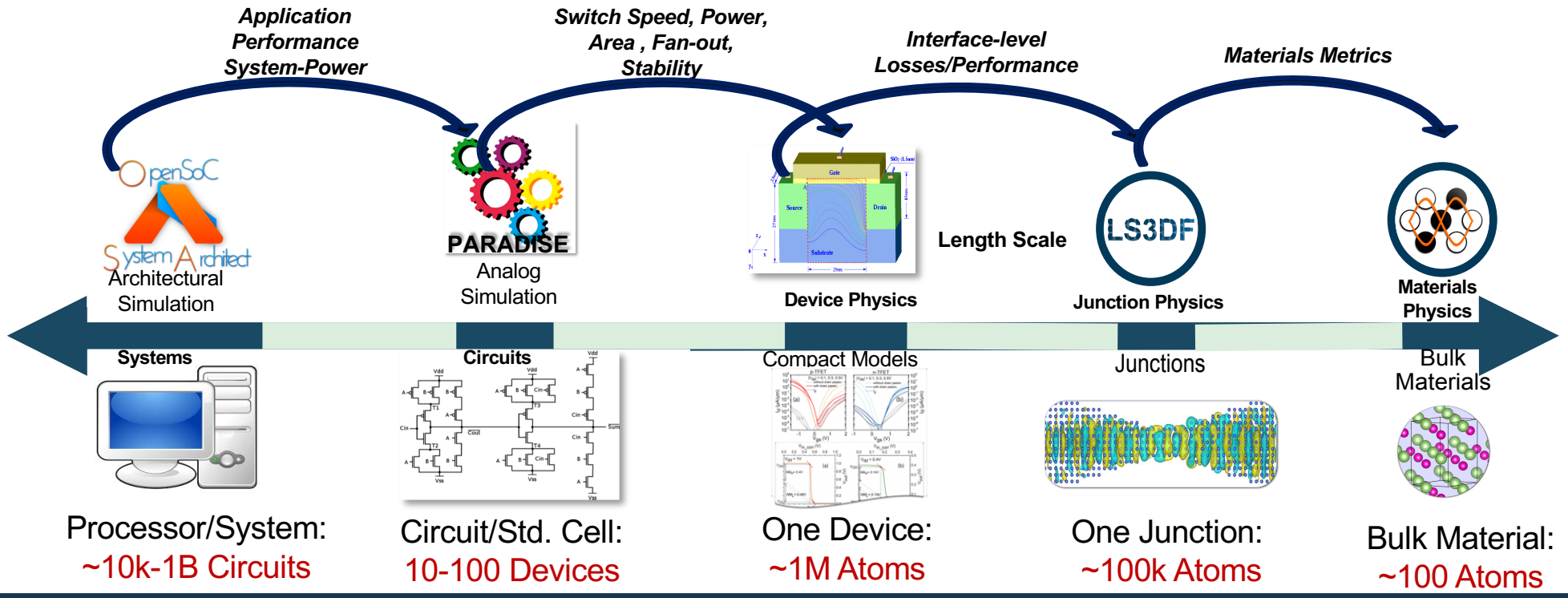
Characterizing materials, analyzing devices, understanding impacts on circuits, architectures, systems and applications.



A holistic end-to-end modeling approach is required

Gap: Connecting and Scaling

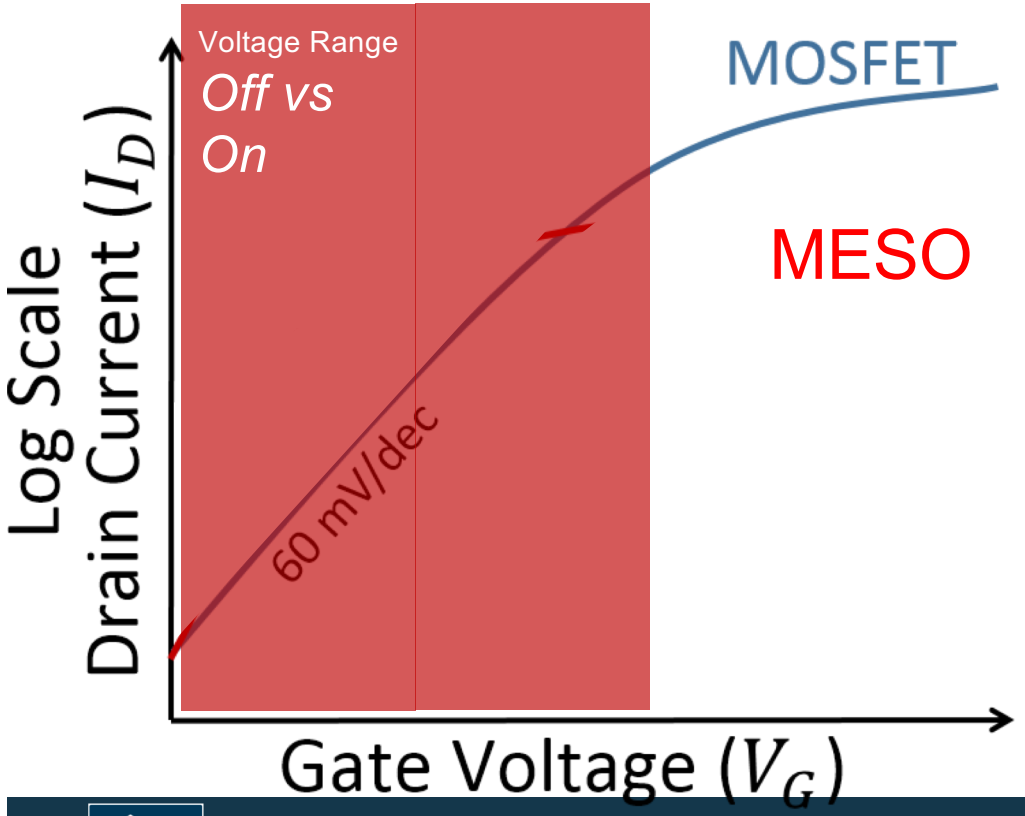
Accelerated feedback path to focus device and material discovery process

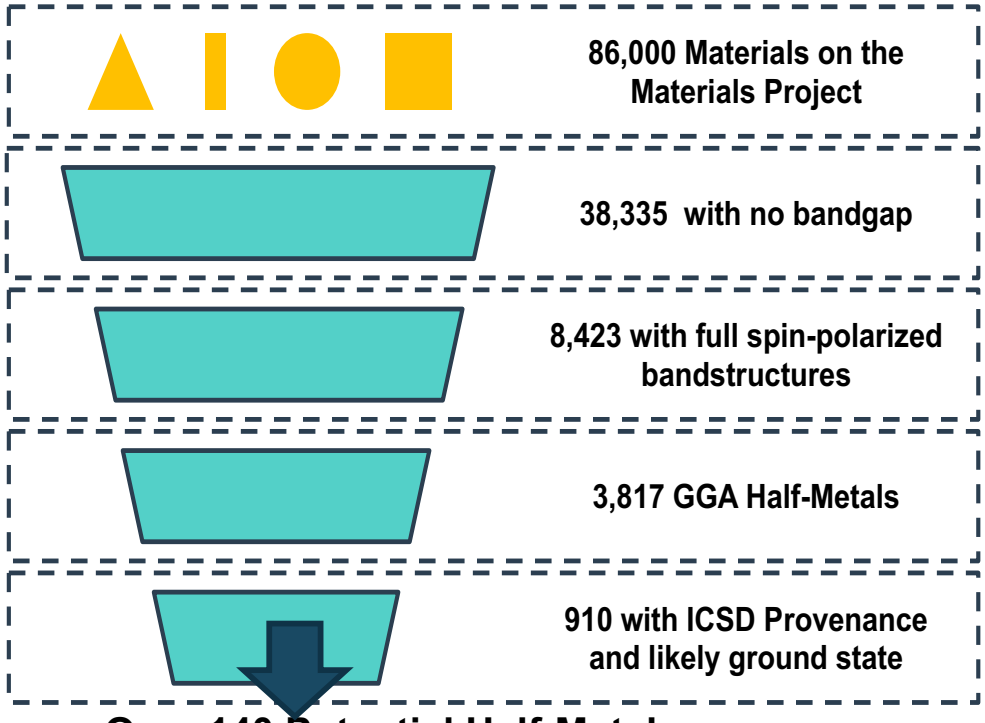


New Breakthroughs in Transistor Technology Require Fundamentally New Principles of Operation

A More sensitive switch: MESO Magneto-Electric Switch

Modulated by Inverse Spin Hall Effect instead of Thermionic Emission





Over 140 Potential Half-Metals for Experimental Investigation

BERKELEY LAB

PARADISE: Post-Moore Architecture and Accelerator Design Space Exploration

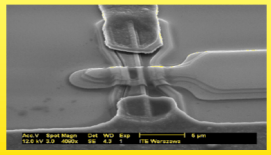
George Michelogiannakis & Dilip Vasudevan

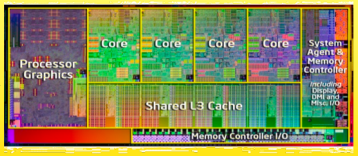
- **Multiple devices, memories, and other "post Moore" technologies in development**
- **Evaluating each in isolation misses big picture**
  - Devices can be better designed with high-level metrics
  - Architects can evaluate how exploit new technologies

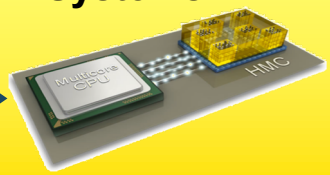
Until now, we lacked the tools to do so systematically and rapidly for many technologies

(PARADISE addresses that gap)

Transistor/Devices → **Architectures** → **Systems**

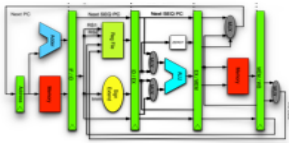


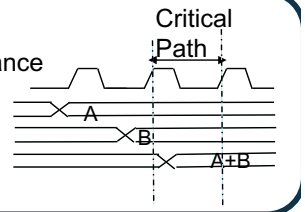




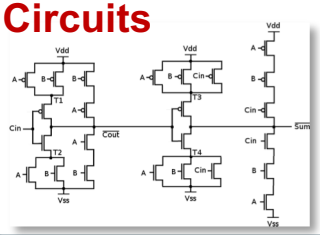
Systems

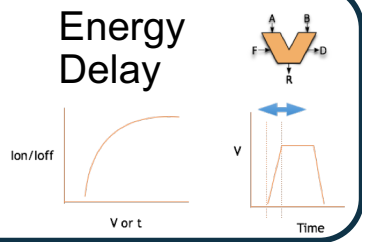
Logic Blocks



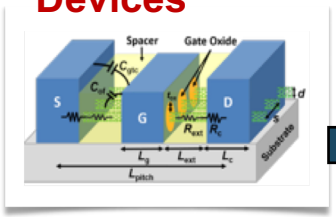


Circuits





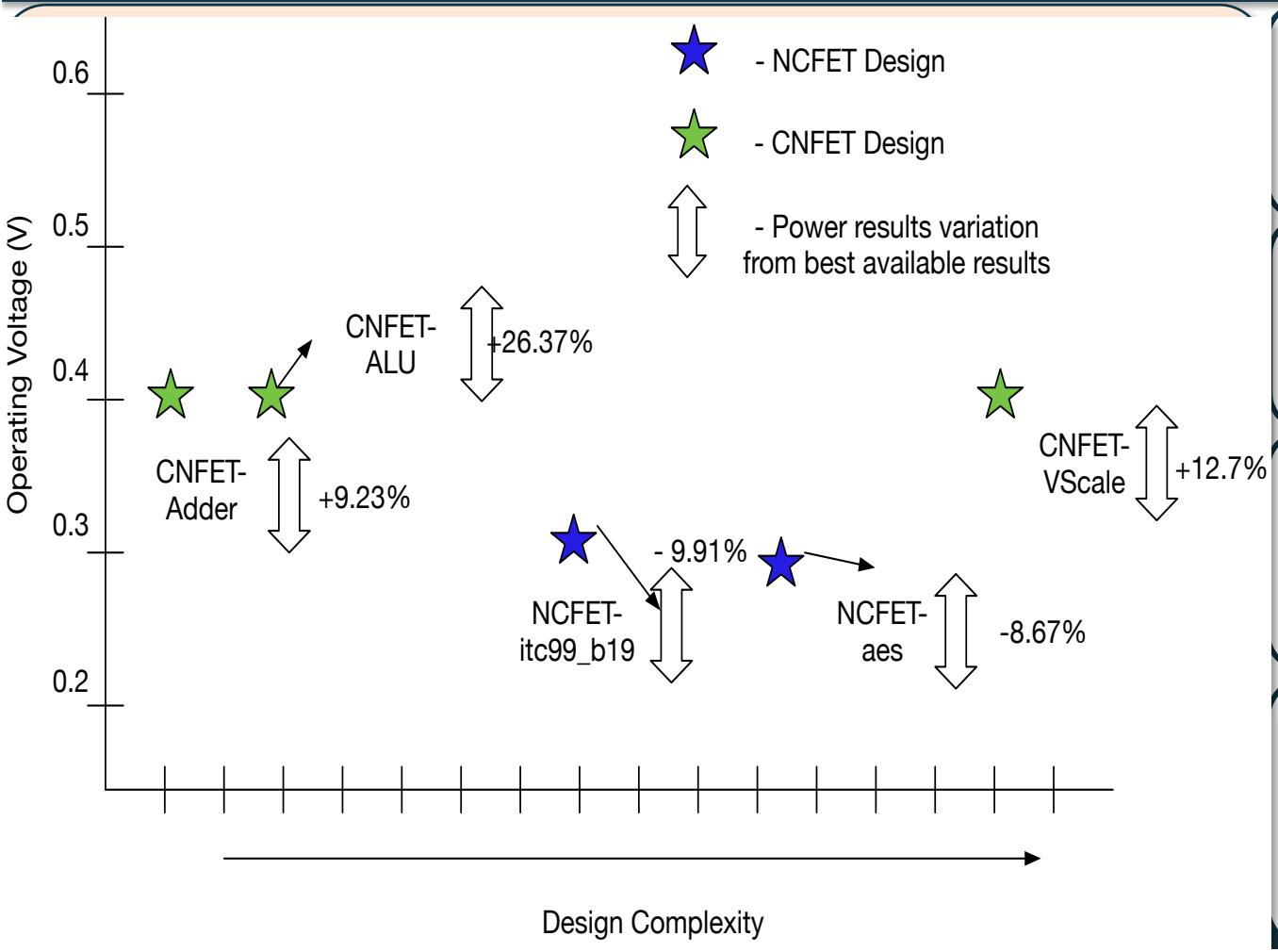
Devices

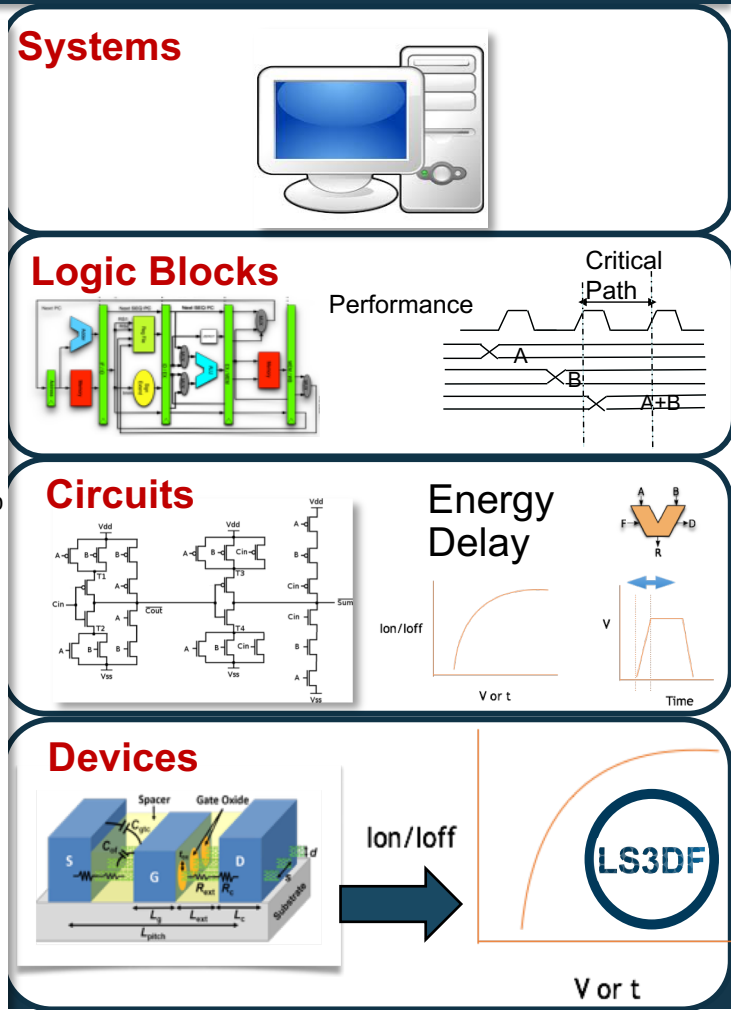


Ion/Ioff

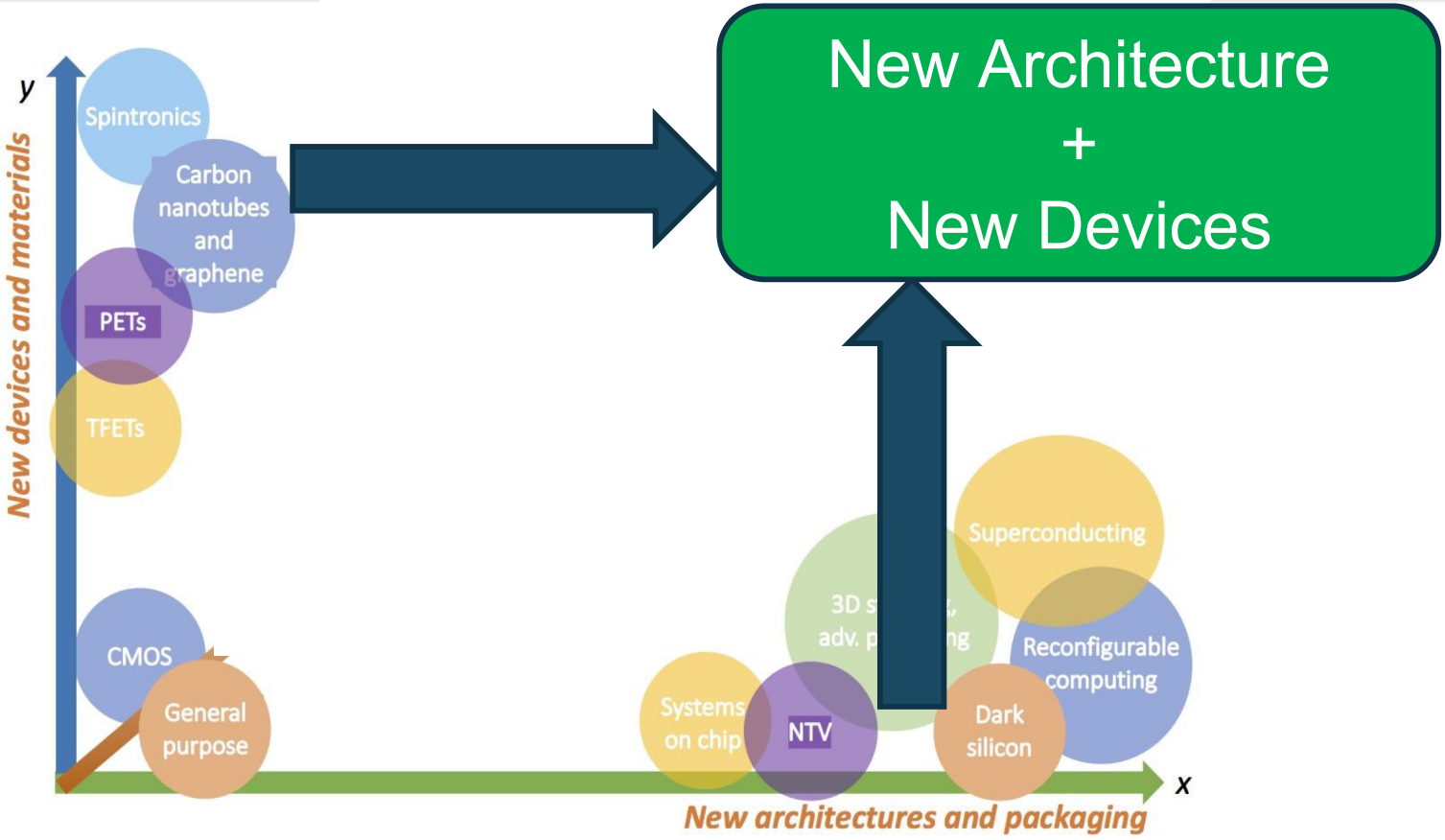
PARADISE: Post-Moore Architecture and Accelerator Design Space Exploration

George Michelogiannakis & Dilip Vasudevan

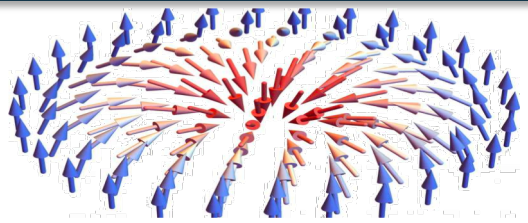




The Sum of the Parts is Greater than the Whole



Skyrmion-based Spiking Neural Networks



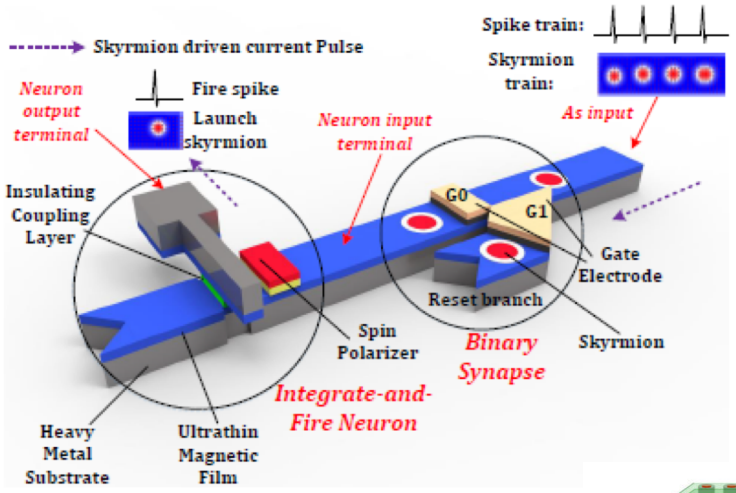
1 A:0, B:0, Y:0

2 A:1, B:0, Y:0

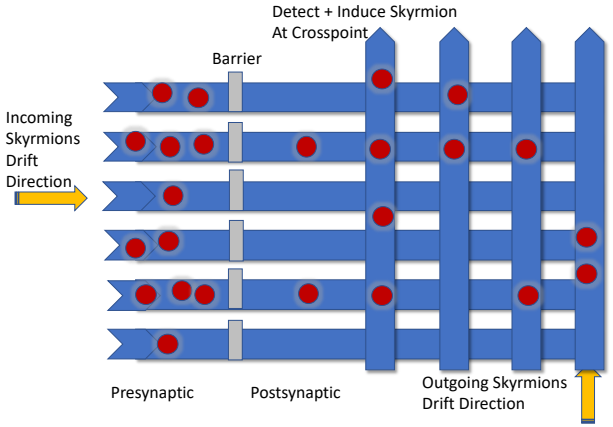
A:0, B:1, Y:0

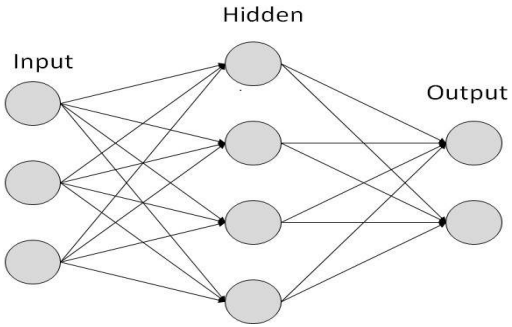
3 A:1, B:1, Y:1

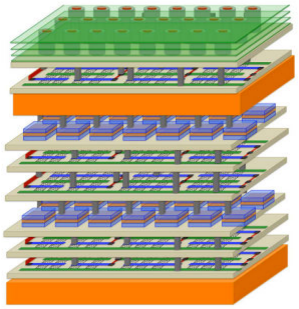
Z. He et al., 1705.02995v1 (2017)



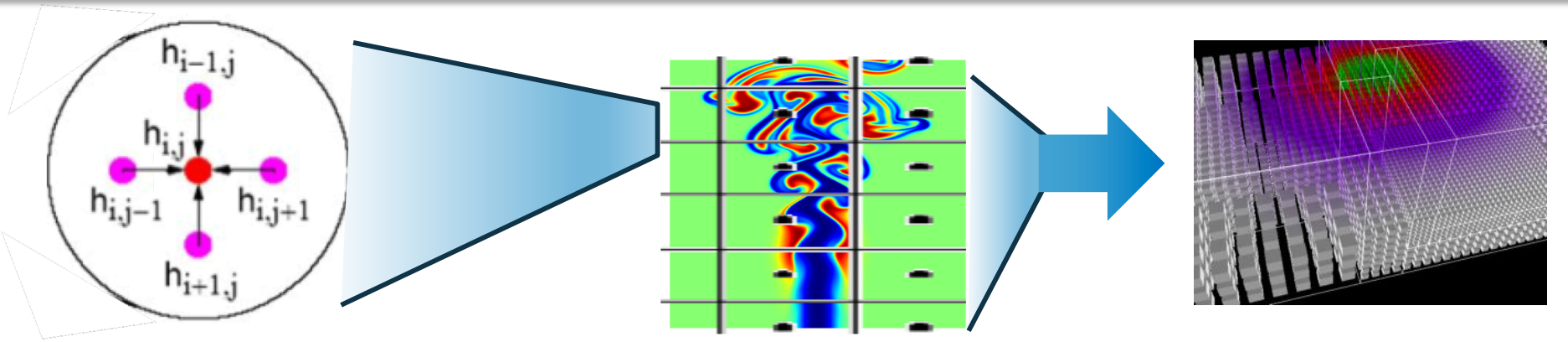
Dilip Vasudevan & Mi Young Im

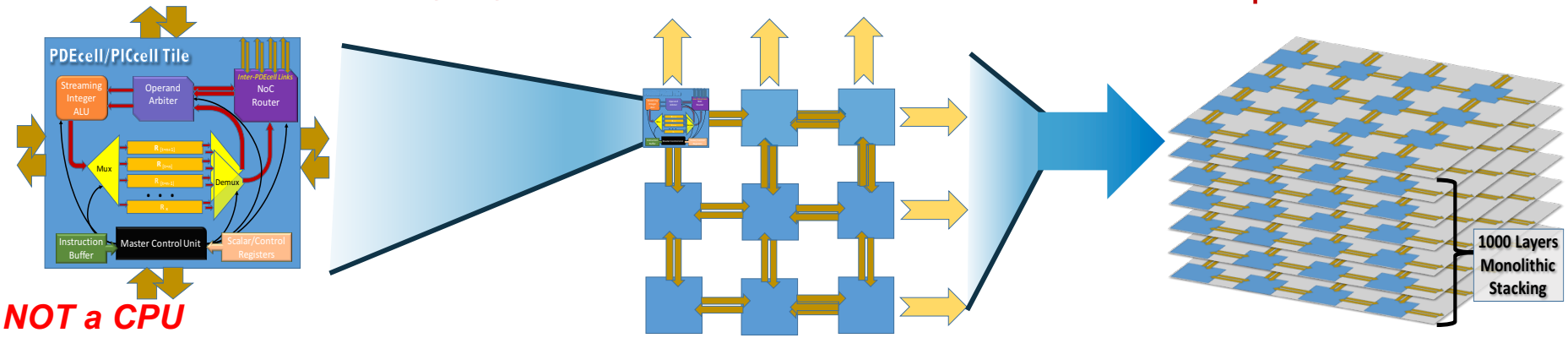






PDE Solvers & Particle Methods in ”Solid State Digital Fluid”





NOT a CPU
Just PDE ALU + Memory Cell

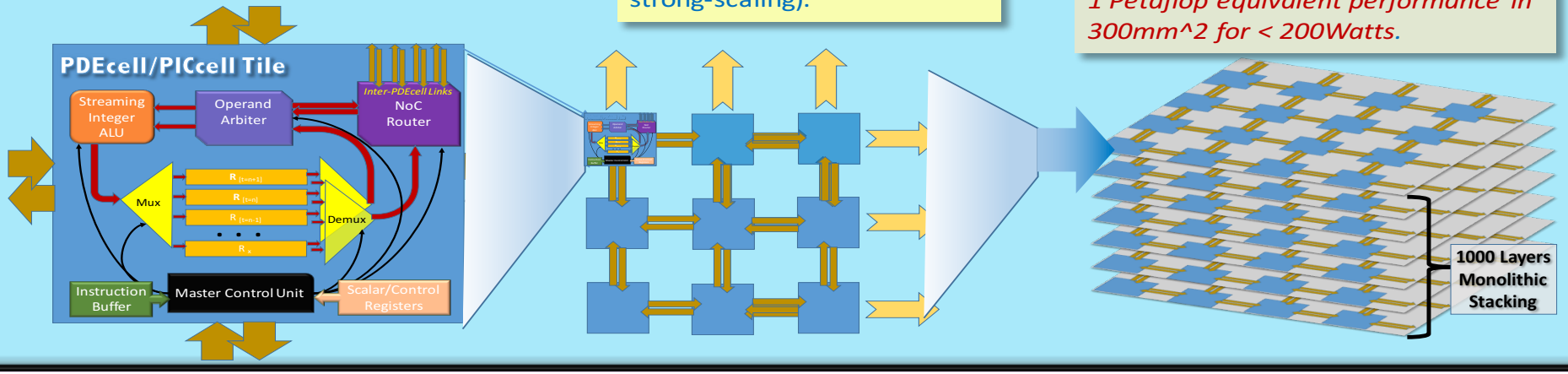
Concept: Solid State Virtual Fluid

Extreme (spatial) Specialization + New Devices + New programming models

PDEcell / PICcell: Ultra-simple compute engine (50k gates) calculates finite-difference updates, and particle forces from neighbors. Microinstructions specify the PDE equation, stencil, and PIC operators.
Novel features: *variable length streaming integer arithmetic and novel PIC particle virtualization scheme.*

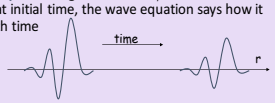
Computational Lattice: PDECells are tiles in a lattice/array on each 2D planar chip layer. Target 120x120 tiles per mm² @28nm lithography. Novel Features: each tile represents single cell of computational domain (pushes to limit of strong-scaling).

Monolithic 3D Integration: Integrate layers of compute elements using emerging monolithic 3D chip stacking.
Novel Features: *1000 layer stacking (20x more than current practice). Area efficient inter-layer connectivity and new energy efficient transistor logic (ncFET).*
1 Petaflop equivalent performance in 300mm^2 for < 200Watts.



Scalar waves in 3D are solutions of the hyperbolic wave equation: $-\phi_{,tt} + \phi_{,xx} + \phi_{,yy} + \phi_{,zz} = 0$

Initial value problem: given data for ϕ and its first time derivative at initial time, the wave equation says how it evolves with time



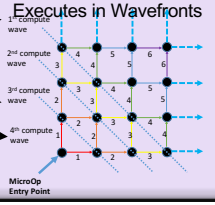
Discretized PDE Representation in DSL

$$\phi^{n+1}_{i,j,k} = 2\phi^{n}_{i,j,k} - \phi^{n-1}_{i,j,k}$$
$$+ \Delta t^2/\Delta x^2(\phi^{n}_{i+1,j,k} - 2\phi^{n}_{i,j,k} + \phi^{n}_{i-1,j,k})$$
$$+ \Delta t^2/\Delta y^2(\phi^{n}_{i,j+1,k} - 2\phi^{n}_{i,j,k} + \phi^{n}_{i,j-1,k})$$
$$+ \Delta t^2/\Delta z^2(\phi^{n}_{i,j,k+1} - 2\phi^{n}_{i,j,k} + \phi^{n}_{i,j,k-1})$$

Compiles to MicroOps

```
R[n+1](0,0,0) = 0
R[n+1](0,0,0) += 2*R[n](0,0,0)
R[n+1](0,0,0) -= R[n-1](0,0,0)
R[n+1](0,0,0) += C * R[n+1](+1,0,0)
R[n+1](0,0,0) -= C * 2 * R[n](0,0,0)
R[n+1](0,0,0) += C * R[n](-1,0,0)
R[n+1](0,0,0) += C * R[n+1](0,+1,0)
....
```

Executes in Wavefronts



Software Strategy for Accelerators

- **Challenge:** ***Current languages are too prescriptive***
  - Unable to accommodate diverse underlying hardware
  - Fundamentally designed for "instruction processors"
- **Top-Down Strategy :** ***higher level language abstractions***
  - Raise the level of abstraction for describing algorithms (less prescriptive)
  - DSLs (Proto) or hierarchical abstractions in libraries (SuperLU, GraphBLAS) or frameworks (such as AMReX)
  - **Impact**: One description of the algorithm can target multiple accelerators or systems
- **Bottom-Up Strategy :** ***Application/Algorithm Targeted accelerators***
  - Create specialized accelerators that target specific algorithms or methods
  - Target relatively fixed interfaces and APIs such as FFT, BLAS, or comm.
  - **Impact**: Same accelerator can target multiple applications

Aligning Technology Investments with Real Physics Challenges

Aligning Technology Investments with Real Physics Challenges

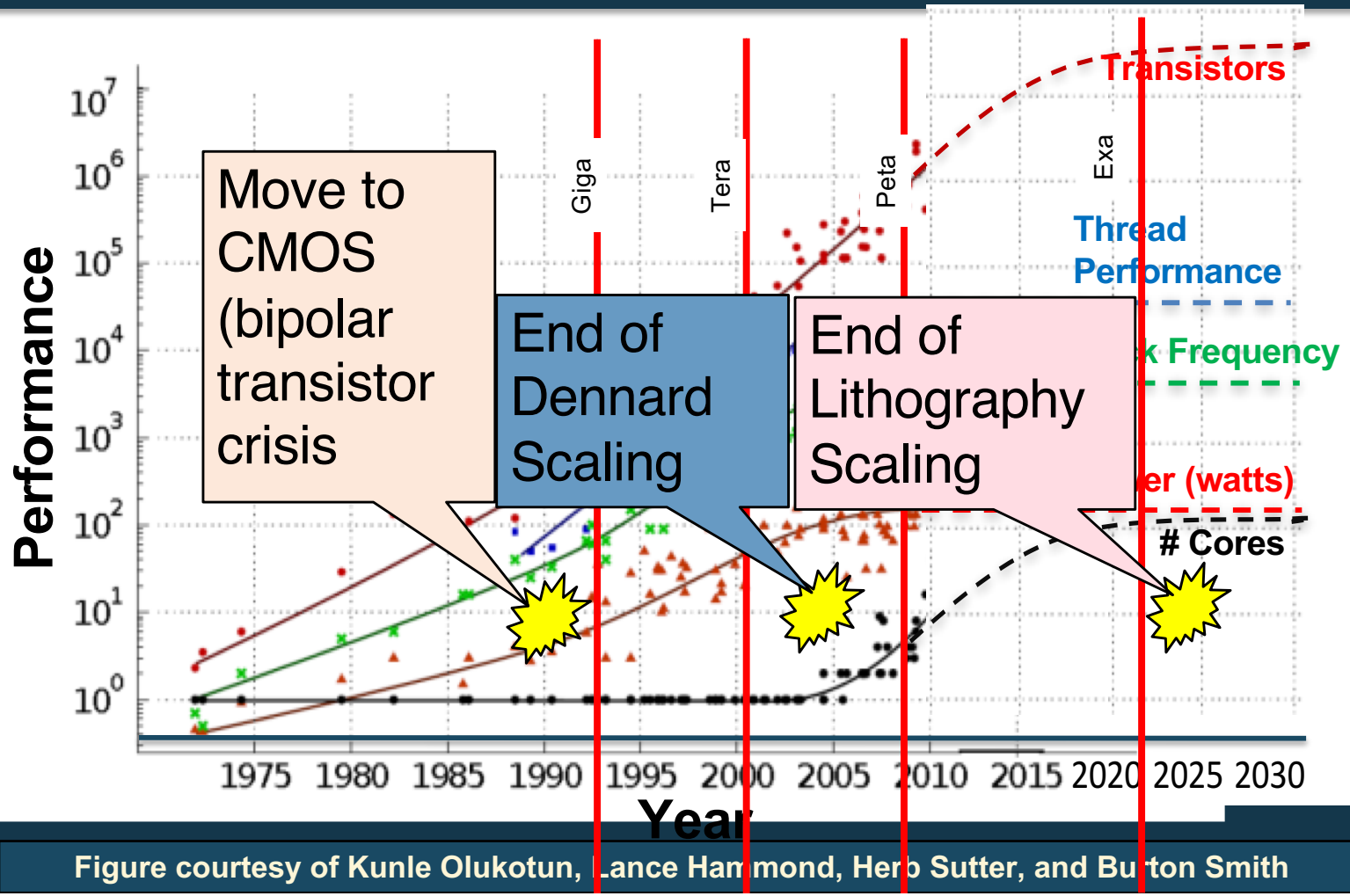


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Conclusion

- **The end of lithography scaling as we know it is coming within a decade (*close to when Exascale is done*)**

- **Consequence is that parallel processing will be even more central to our future, but has NEW challenges**
  - Diverse Hardware Specialization (*many in single node*)
  - Data Centric Computing (*Spatial Computing*)
  - Non-Bulk Synchronous execution models

- **Cannot ignore this until after exascale**
  - ***Focus on 1000x milestones is not aligned with underlying physics challenges***

Conclusions

- **Think more seriously about how to put specialization productively to use for science**
  - Requires deep understanding of applied mathematics and the underlying algorithms to be successful

- **Reevaluate the business/economic model for the design and acquisition of HPC systems**

- **Accelerate the development of materials, devices, and systems for post-CMOS electronics**

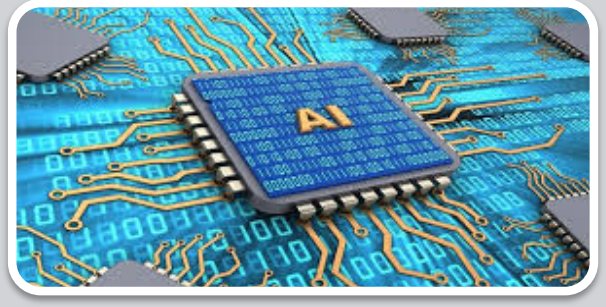
Extra

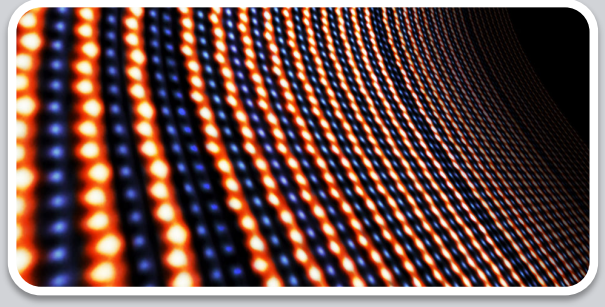
The End

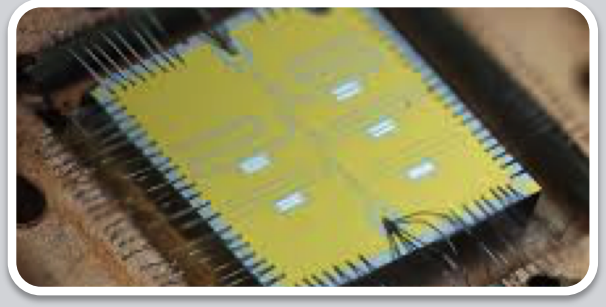
Multiscale Modeling of Post-CMOS Devices and Systems

Accelerate the discovery of new electronic materials and post-CMOS devices

Beyond-Moore Computing Directions







Heterogeneous Architectures

Specialized accelerators for performance / energy

Post CMOS Devices/Materials

Evaluate new devices using simulation across scales

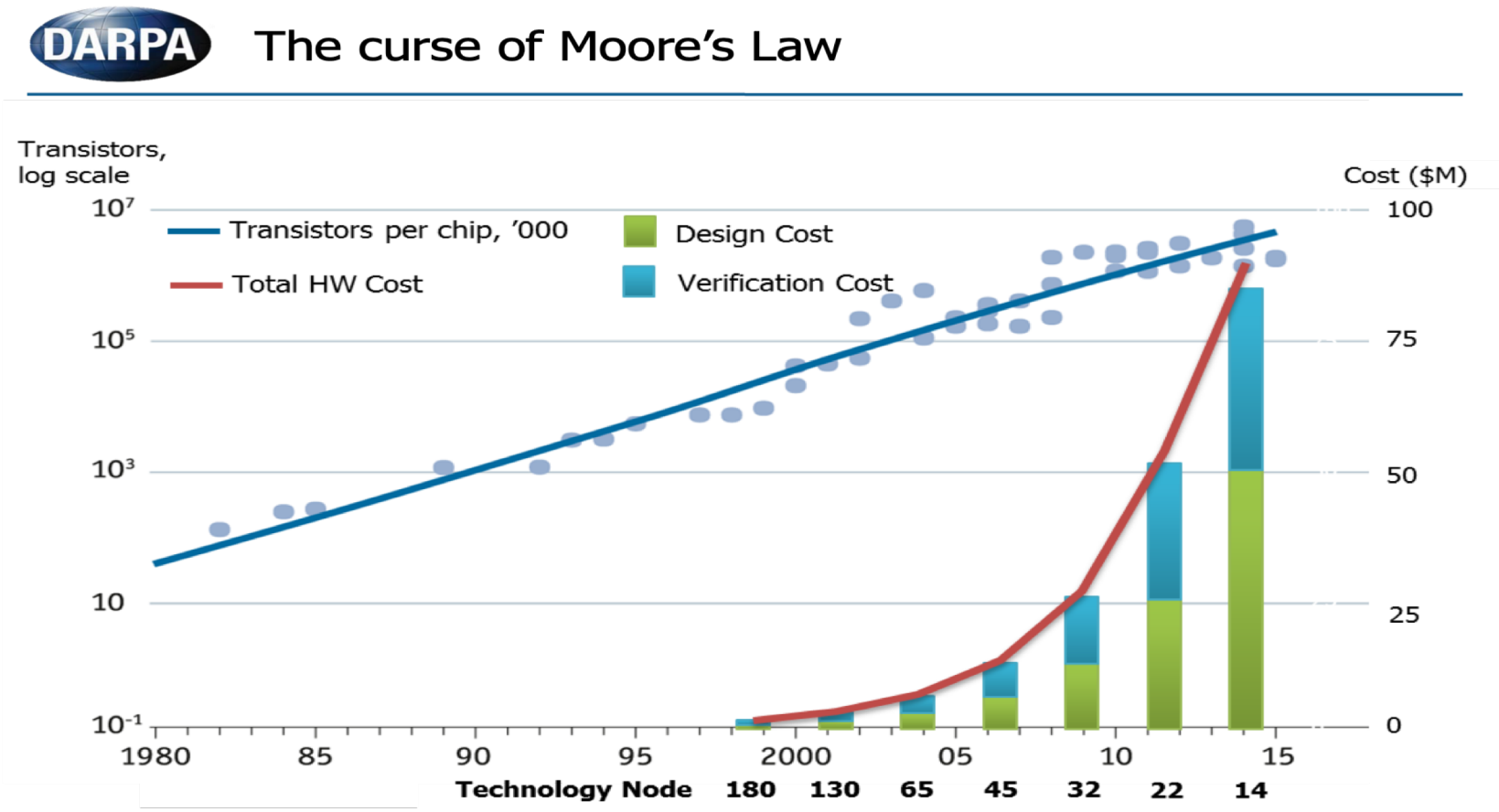
New Models of Computation

Quantum algorithms, tools and testbeds, for science applications

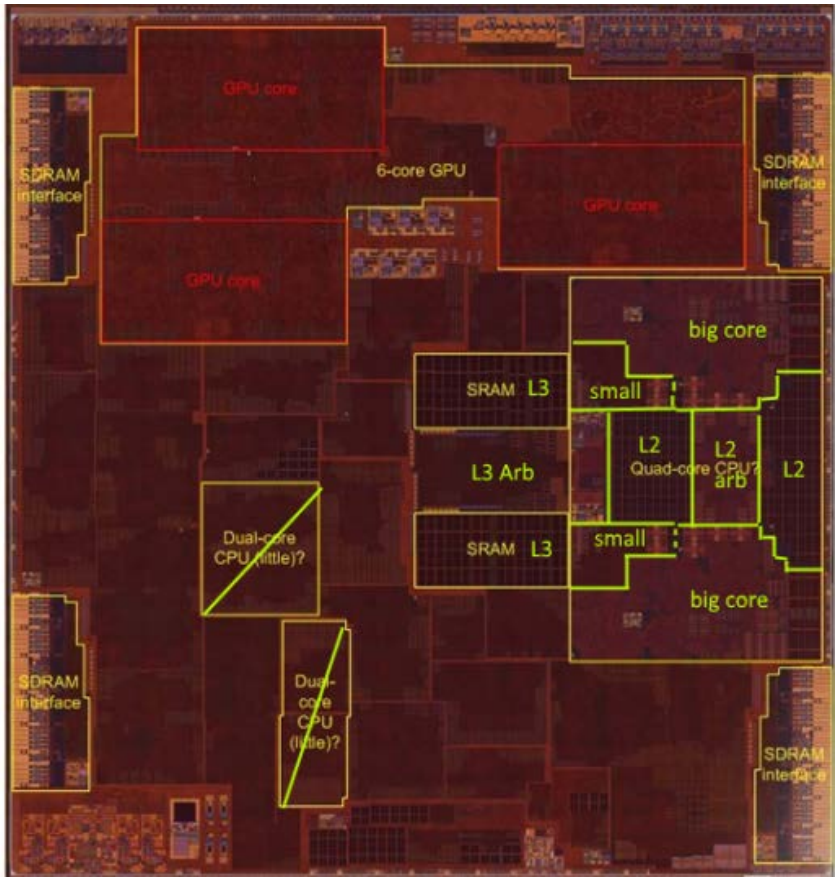
Workload Analysis, Testbeds, Deployment

Backup slides

Why do chips cost so much?



Fab and Mask Costs are NOT the Problem



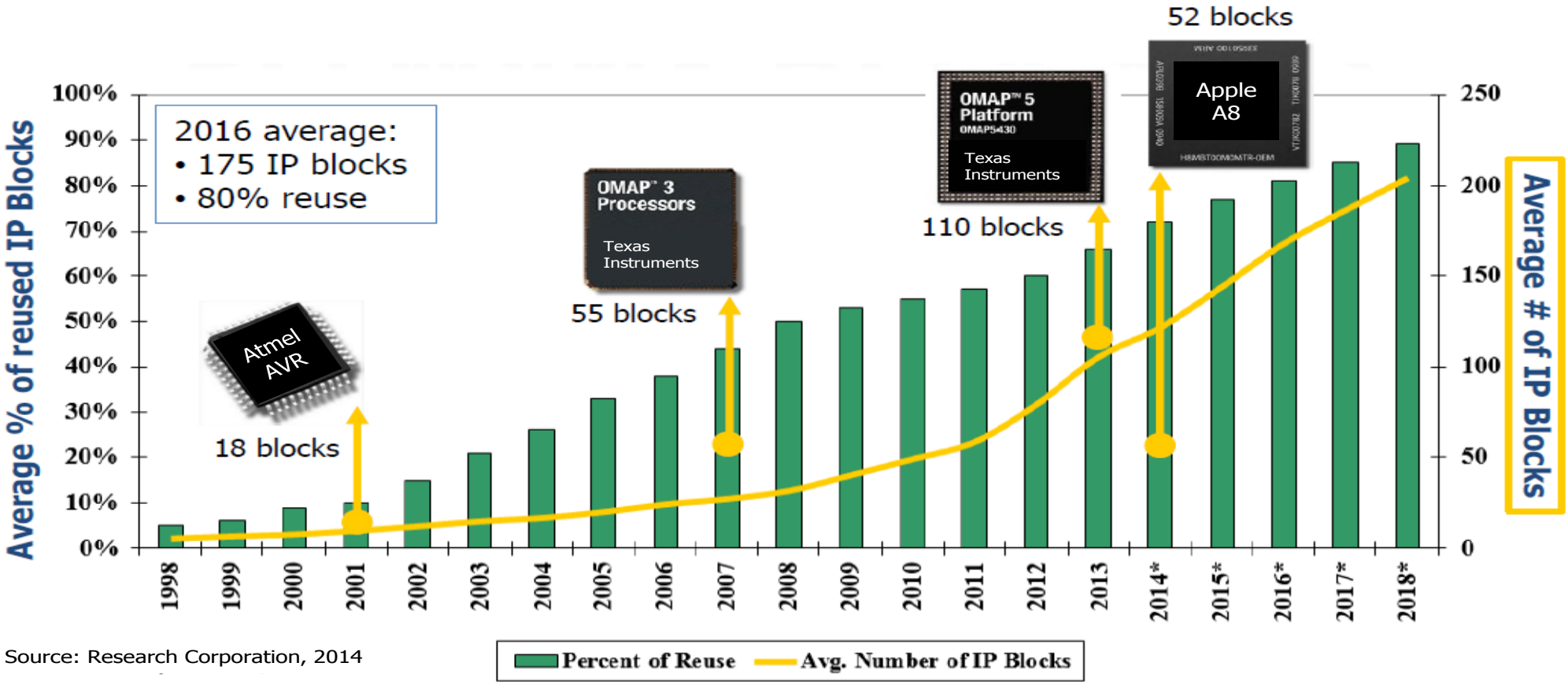
Source: TechInsights/Chipworks

Q: How much does it cost to produce 10,000 125mm^2 SoCs at 16nm?

A: <$2M

Small (1mm x 1mm) chip tapeouts at 28nm are $14K(US)

Getting to Critical Mass for IP Reuse for Accelerators



Source: Research Corporation, 2014