Computing Beyond Moore's Law

John Shalf

Department Head for Computer Science

Lawrence Berkeley National Laboratory

CSSS Talk

July 14, 2020

jshalf@lbl.gov

Technology Scaling Trends

Exascale in 2021… and then what?

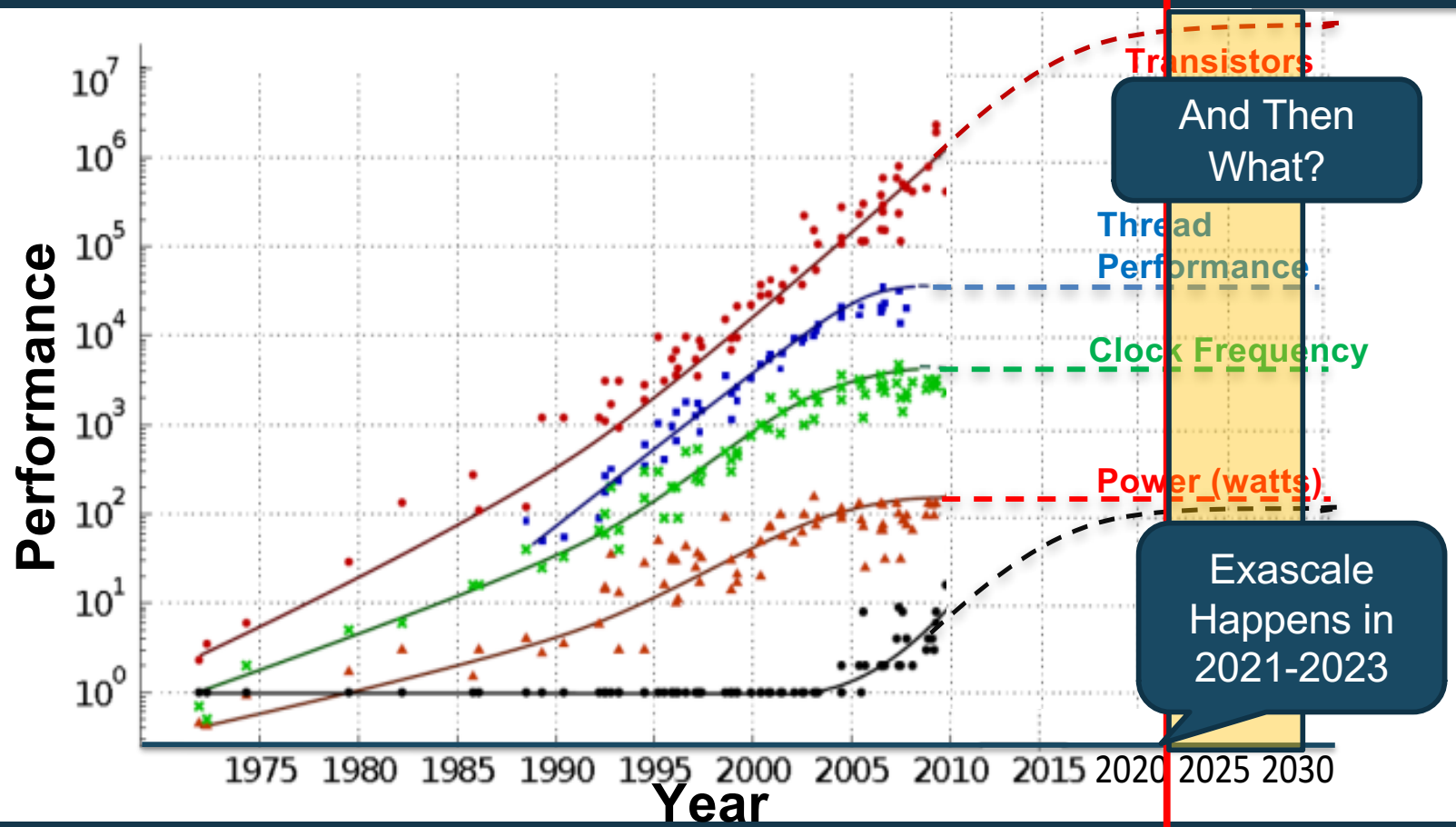
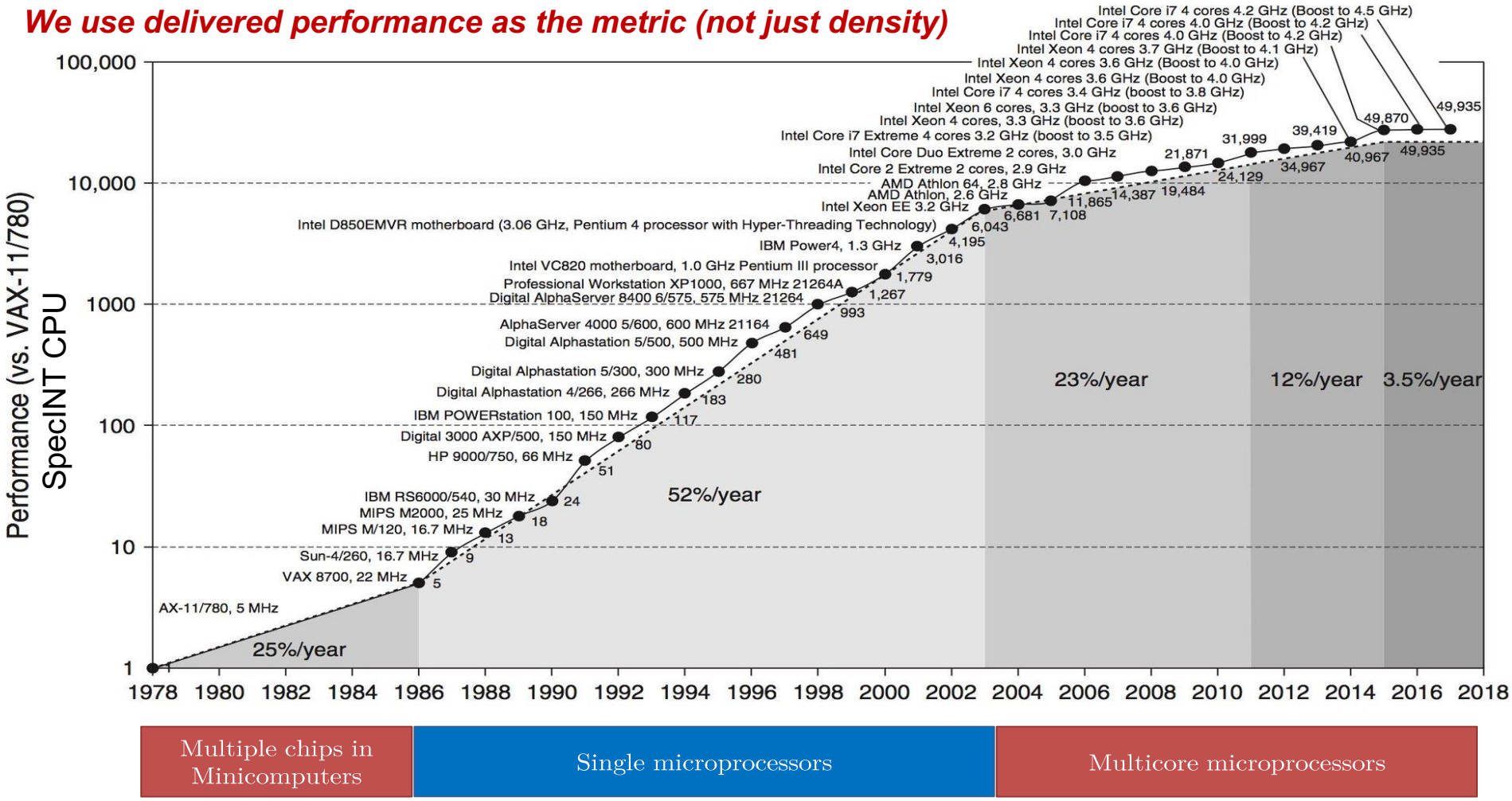


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Moore's Law IS Ending

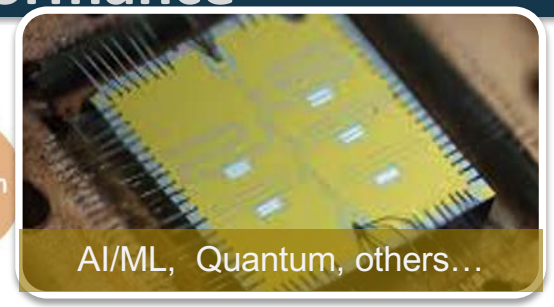
Hennessy / Patterson

We use delivered performance as the metric (not just density)



Multiple chips in Minicomputers	Single microprocessors	Multicore microprocessors

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
CMOS
General purpose

New models o computati

Neuromorphic
Quantum
Analog
Adabiatic reversible
aflow
oximate computing

New Models of Computation
Decades beyond exascale

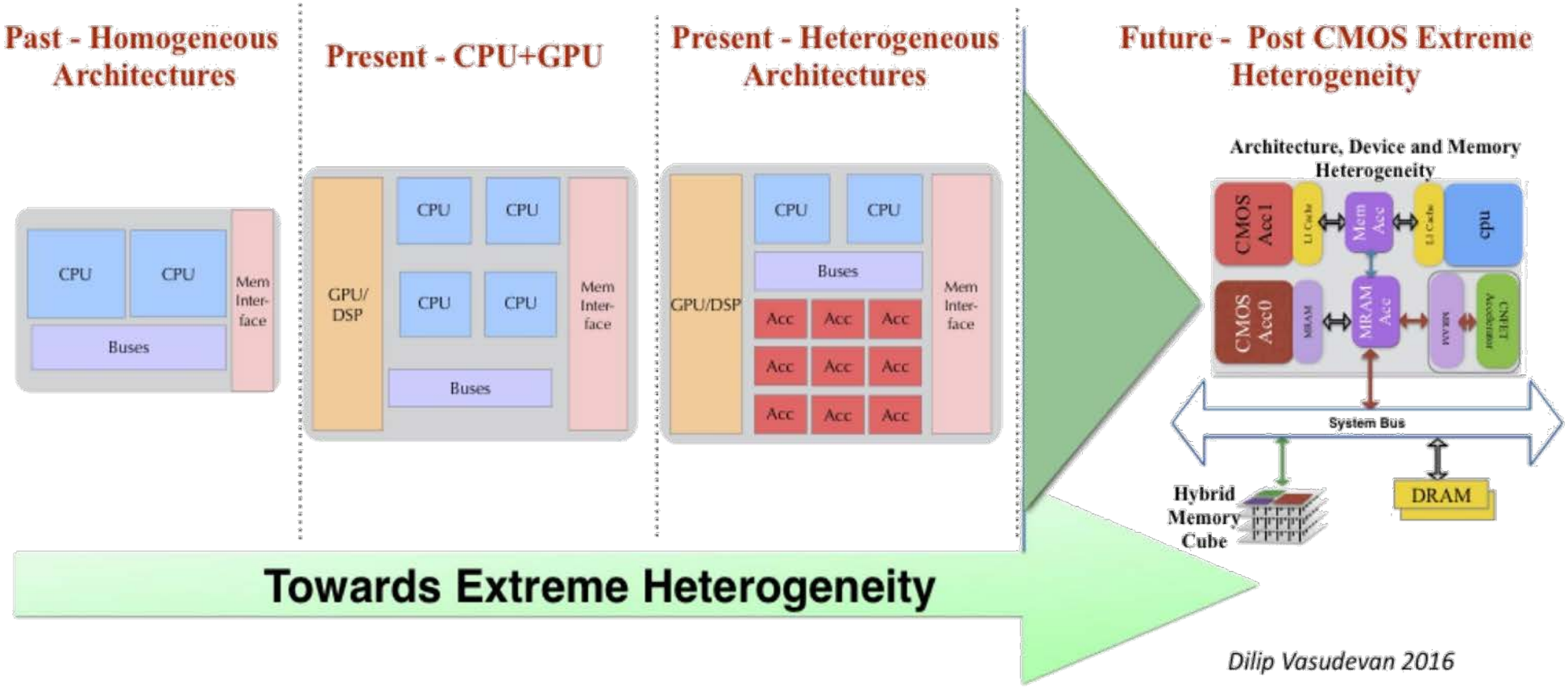
Superconducting
3D stacking, adv. packaging
Reconfigurable computing
Systems on chip
NTV
Dark silicon

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

The Future Direction for Post-Exascale Computing



Dilip Vasudevan 2016

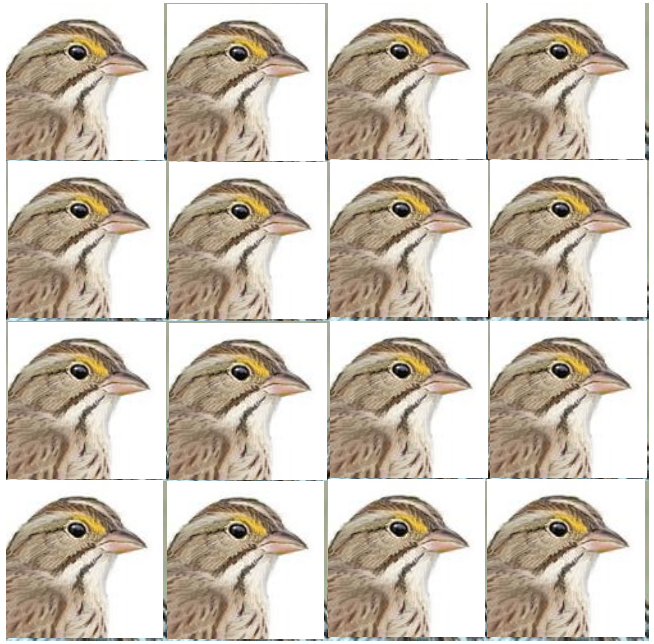
Specialization:

Natures way of Extracting More Performance in Resource Limited Environment

Powerful General Purpose

Xeon, Power

Many Lighter Weight
(post-Dennard scarcity)



KNL AMD, Cavium/Marvell, GPU

Many Different Specialized
(Post-Moore Scarcity)

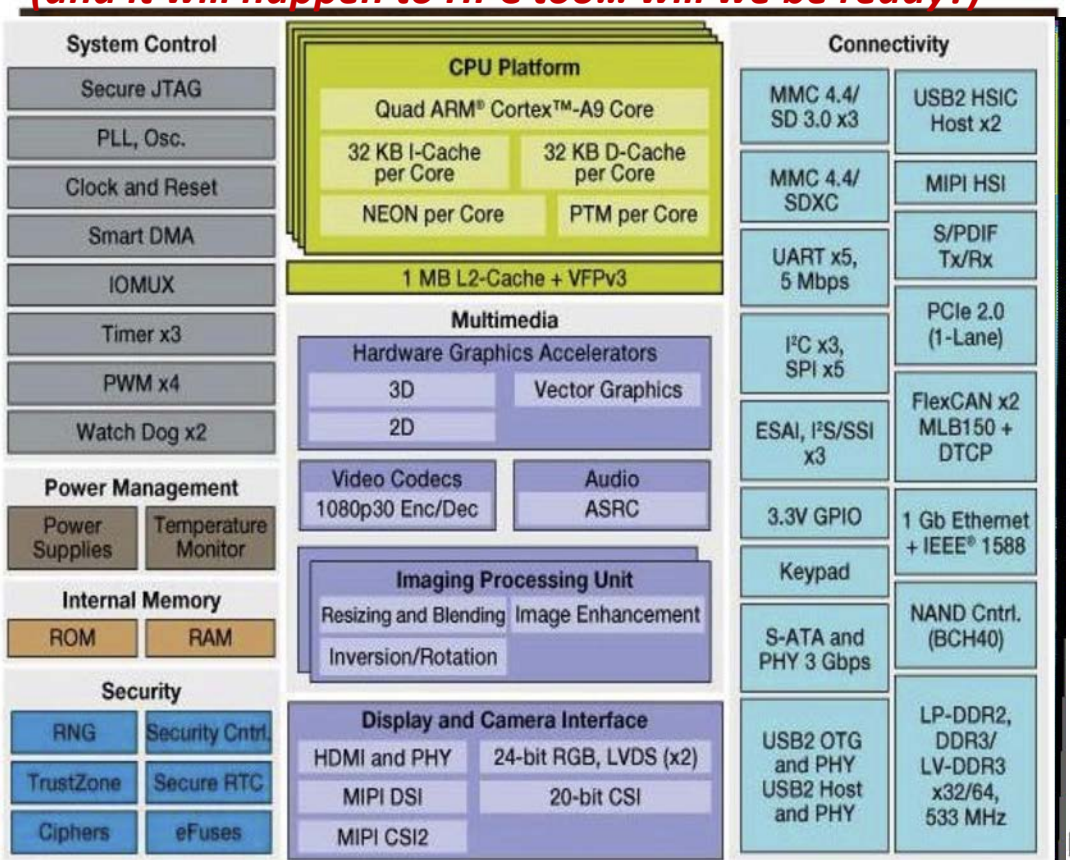
Apple, Google, Amazon

Extreme Hardware Specialization is Happening Now!

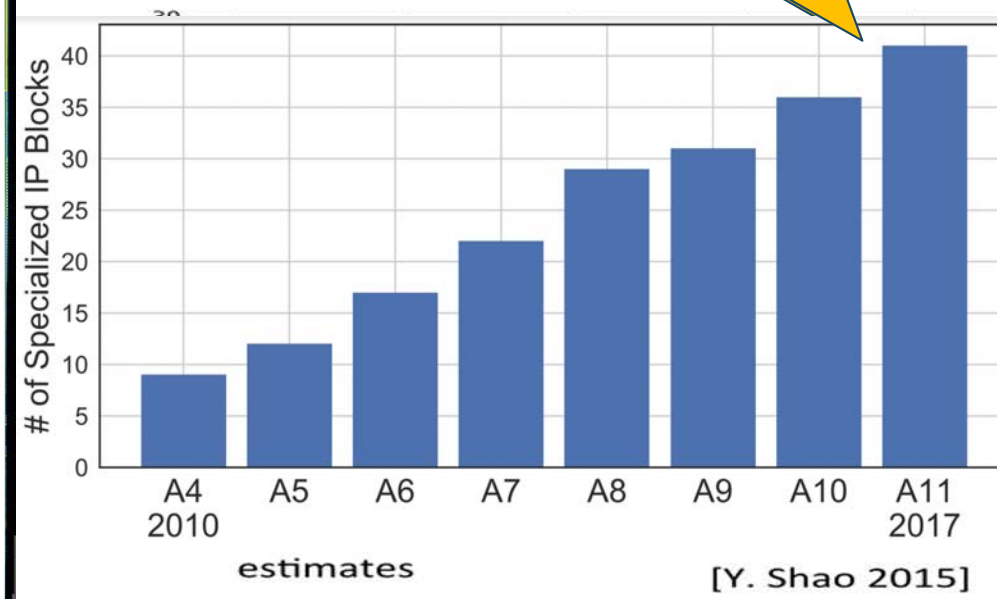
This trend is already well underway in broader electronics industry

Cell phones and even megadatacenters (Google TPU, Microsoft FPGAs…)

(and it will happen to HPC too… will we be ready?)



40+ different heterogeneous accelerators in Apple A11 (2019)

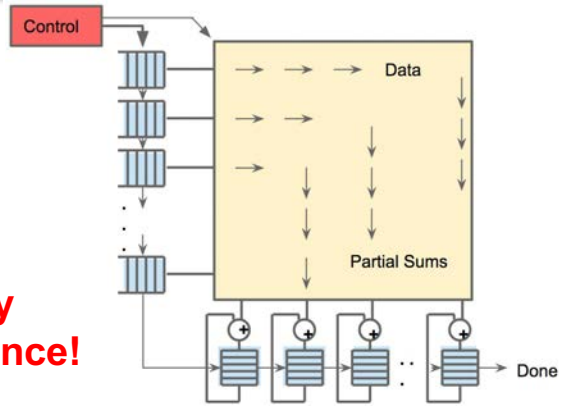


[www.anandtech.com/show/8562/chipworks-a8]

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.

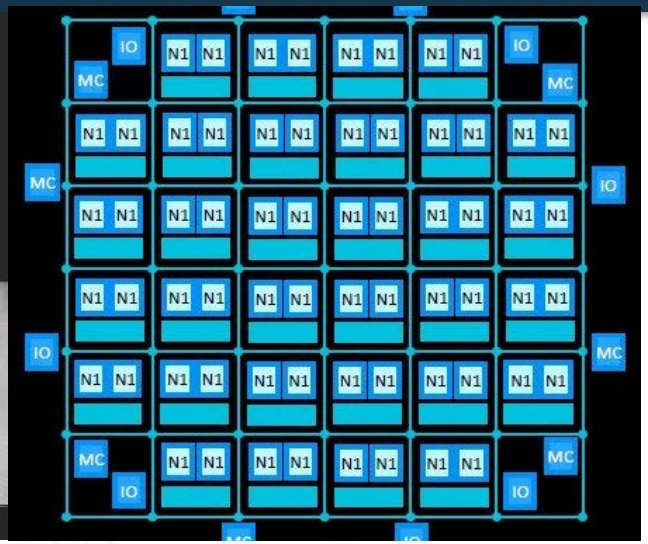
Model	MHz	Measured Watts		TOPS/s		GOPS/s /Watt		GB/s	On-Chip Memory
		Idle	Busy	8b	FP	8b	FP		
Haswell	2300	41	145	2.6	1.3	18	9	51	51 MiB
NVIDIA K80	560	24	98	--	2.8		29	160	8 MiB
TPU	700	28	40	92	--	2,300		34	28 MiB

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

Amazon AWS Graviton CustomARM SoC *(and others)*

AWS Graviton2 processor

- 4x the vCPUs
- 7x CPU performance
- ~2x performance/vCPU
- ~30 Billion transistors
- 7nm



AWS CEO Andy Jassy:

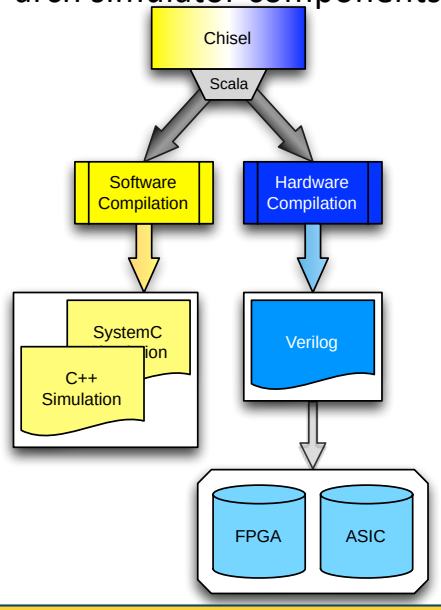
"AWS isn't going to wait for the tech supply chain to innovate for it and is making a statement with performance comparisons against an Intel Xeon-based instance. The EC2 team was clear that Graviton2 sends a message to vendors that they need to move faster and AWS is not going to hold back its cadence based on suppliers."

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

Co-Develop Hardware and Algorithm

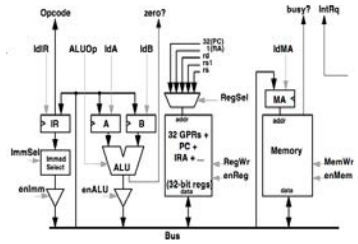
Chisel

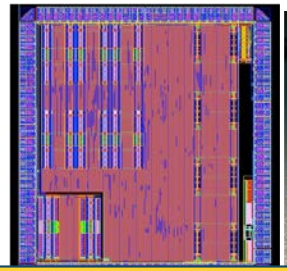
DSL for rapid prototyping of circuits, systems, and arch simulator components

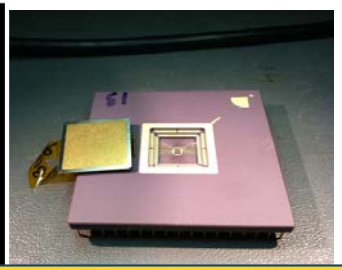


RISC-V

Open Source Extensible ISA/Cores

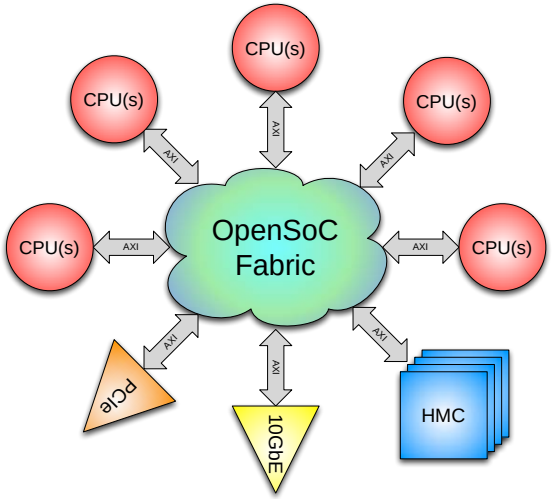




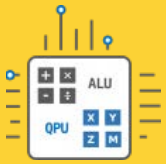


OpenSOC

Open Source fabric To integrate accelerators And logic into SOC

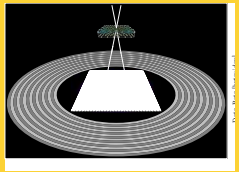


SuperTools Superconducting RISC-V



QUASAR Quantum ISA

Multiagency Architecture Exploration

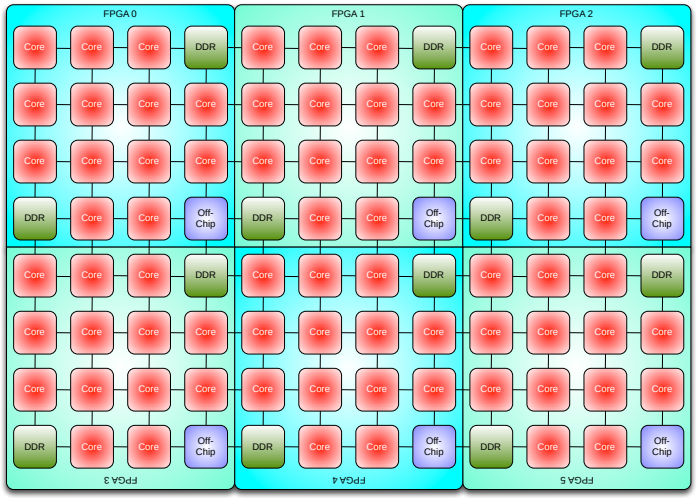


Active Sensors

Research platform: 96-core Tiled CPU on FPGA

SC2016 Demo (accidentally Sunway-like architecture emulation)

COMPUTER ARCHITECTURE LABORATORY
EXASCALE DESIGN SPACE EXPLORATION



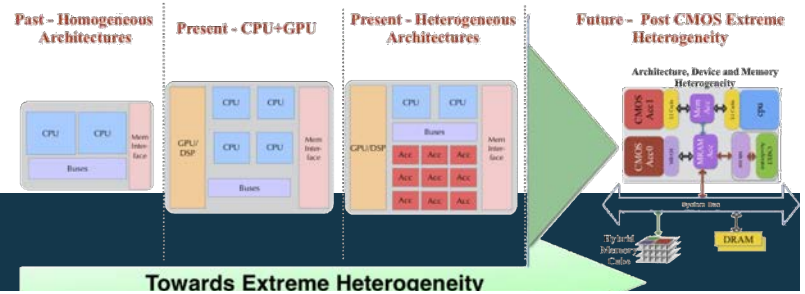
2 people spent 2 months to create

- **Z-Scale processors connected in a Concentrated Mesh**
- **4 Z-scale processors**
- **2x2 Concentrated mesh with 2 virtual channels**
- **Micron HMC Memory**

http://www.codexhpc.org/?p=367

Putting Architecture Specialization to work for

- **But what are the right specializations to include?**
- **What is the cost model (we know we cannot afford to spin our own chips from scratch)**
- **Leverage the Open Source and ARM IP Ecosystem:**
  - ***IP is the commodity (not the chip)!!!***
- **What is the right partnership/economic model for the future of HPC?**

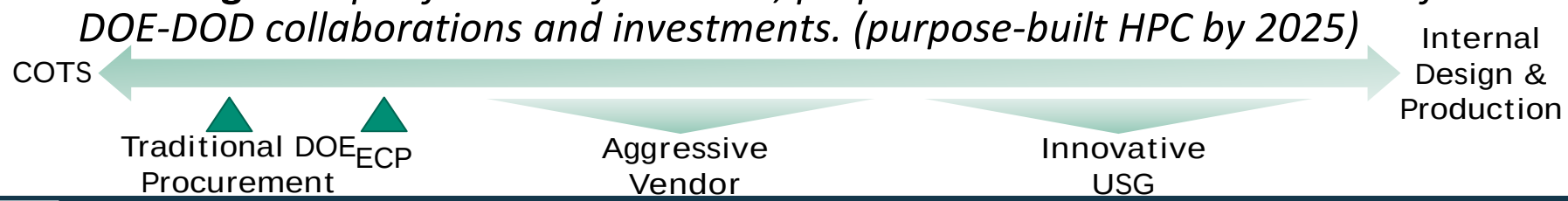


Project 38 -- Background

DOD and DOE recognize the imperative to develop new mechanisms for engagement with the vendor community, particularly on architectural innovations with strategic value to USG HPC.

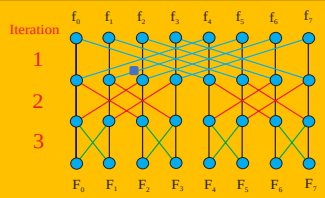
Project 38 (P38) is a set of vendor-agnostic architectural explorations involving DOD, the DOE Office of Science, and NNSA (these latter two organizations are referred to in this document as "DOE"). These explorations should accomplish the following:

- **Near-term goal:** *Quantify the performance value and identify the potential costs of specific architectural concepts against a limited set of applications of interest to both the DOE and DOD.*
- **Long-term goal:** *Develop an enduring capability for DOE and DOD to jointly explore architectural innovations and quantify their value.*
- **Stretch goal:** *Specification of a shared, purpose built architecture to drive future DOE-DOD collaborations and investments. (purpose-built HPC by 2025)*

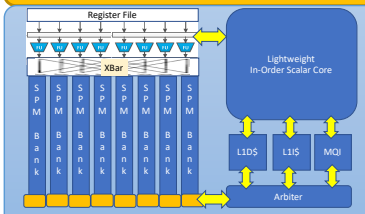


Recapping Key P38 Technology Features

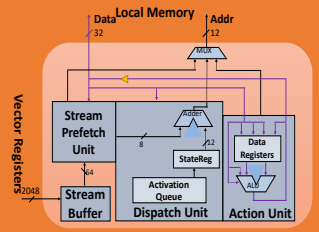
innovative USG



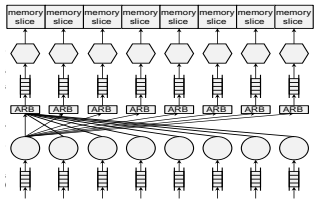
- **Fixed Function Accelerators & COTS IP *(Extreme Heterogeneity)***
  - RISC-V and ARM cores
  - Fixed function FFT (Generated by SPIRAL)



- **Word Granularity Scratchpad Memory (Gather Scatter):**
  - Gather-scatter within processor tile
  - more effective SIMD



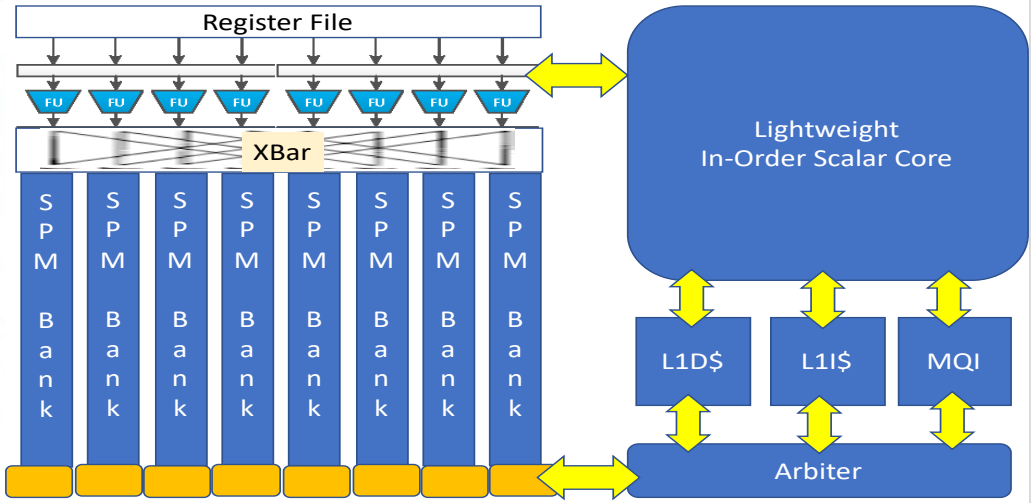
- **Recoding engine (Efficient programmable FSM & data reorg.)**
  - Sub-word granularity and high control irregularity
  - Handles branch-heavy code (avg. 20x improvement over processor core)
  - One lane is 1/100th the size of a x86 processor core

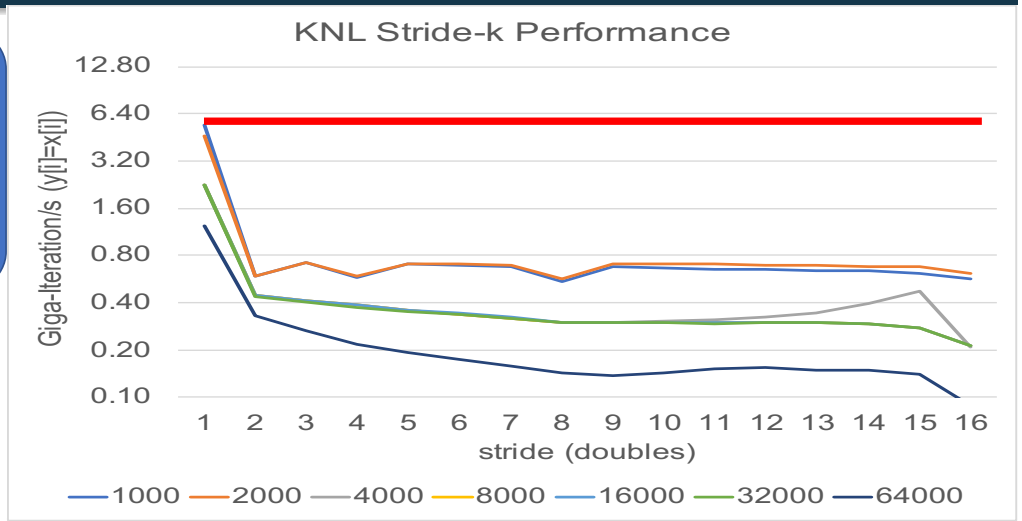


- **Hardware Message Queues (Lightweight Interprocessor Communication)**
  - Gather-scatter between processor tiles
  - Async between tiles to eliminate overhead of barriers

George Fann & Yuan Zheng

General-Purpose: Tensor Contractions on Word Granularity SPM

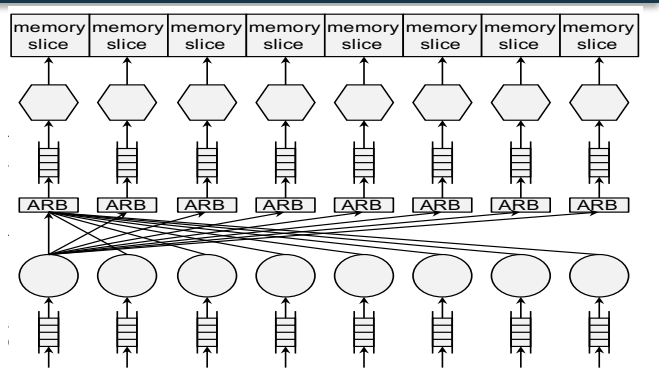




	number_of_particles	basis_size	number_of_blocks	nonzero_fraction	runs the contraction?	Number of SIMD lanes	Bandwidth waste for loading the t3 or v in inner loop	Bandwidth waste for the entire application
1	40	70	40	0.2	yes	8	55%	36%
2	60	70	40	0.2	yes	8	100%	65.4%
3	65	70	40	0.2	yes	8	700%	457.8%
4	40	70	40	0.1	yes	8	154%	100.7%
5	40	70	40	0.2	yes	16	166%	109%

Create Hardware Features to Accelerate Broadly used Numerical Algorithm Primitives

- **Accelerate commonly used primitives for interprocessor communication**
  - Queues & DAGs commonly used in pseudocode
  - Why not make them REAL? (in design library)



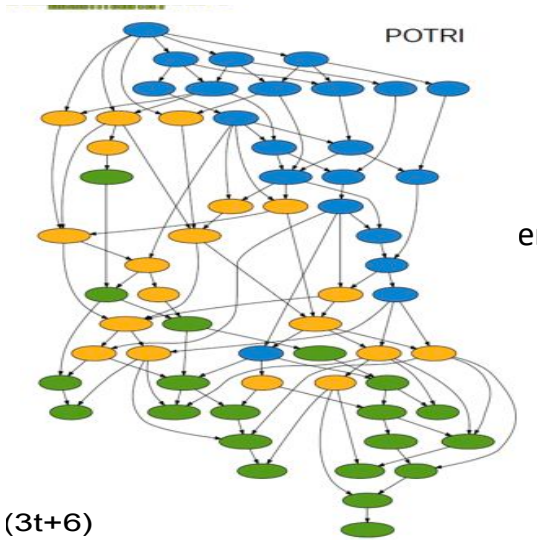
Example Pseudocode

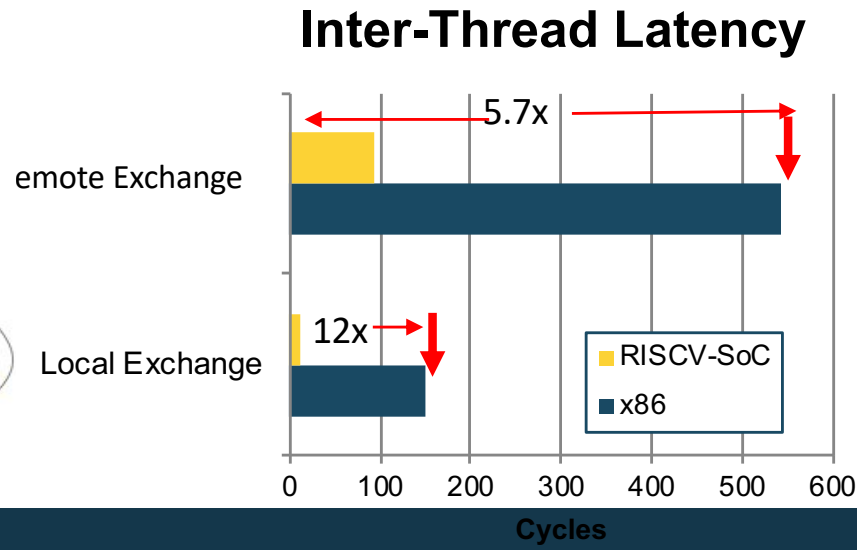
Algorithm: triangularSolve (Kale/Charm++)

Input: Row *myRows*[]
Output: Values *x*[]

```
if any DataMessage msg arrived then
        receiveDataMessage(msg)
end
for each Row r in independent rows do
        computeRow(r,0)
end
while there are pending rows do
        wait for DataMessage msg
        receiveDataMessage(msg)
end
```

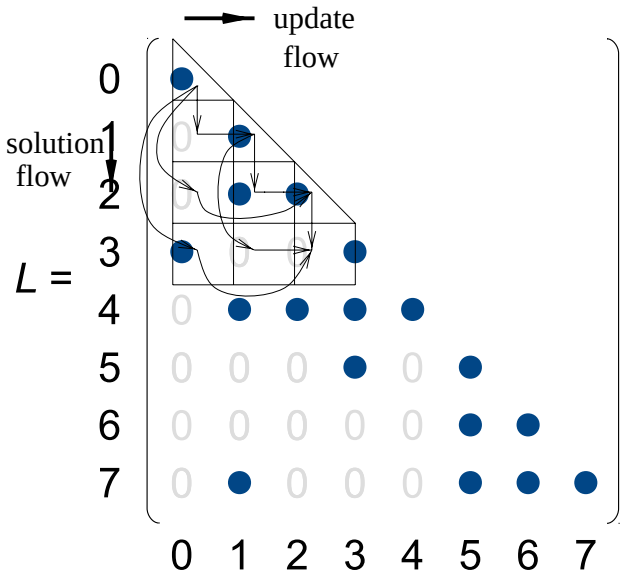
Algorithm 4: Local Triangular Solve



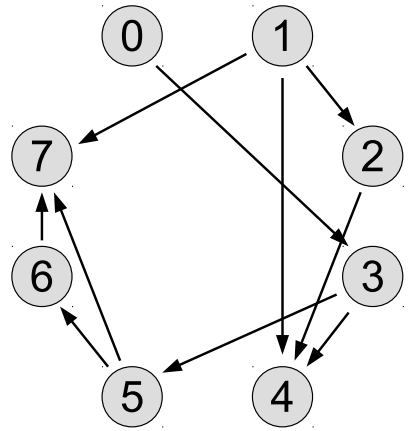


Sparse Matrix Trisolve (refresher)

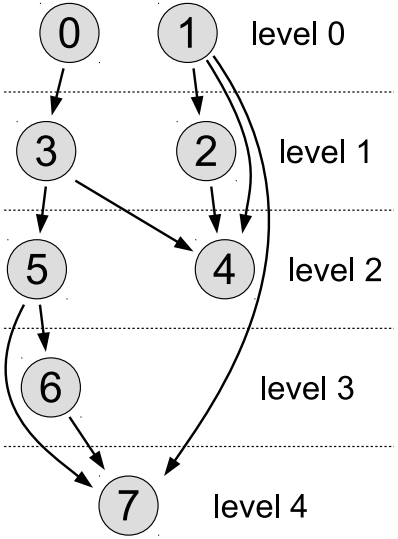
Currently Use OMP Atomic to track dependencies



(a) L's matrix form.

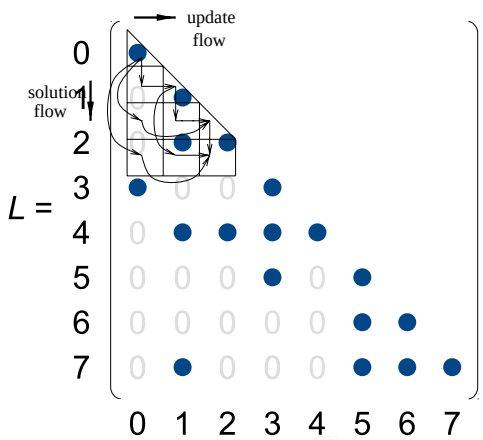


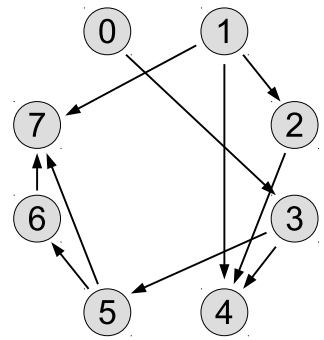
(b) L's graph form.

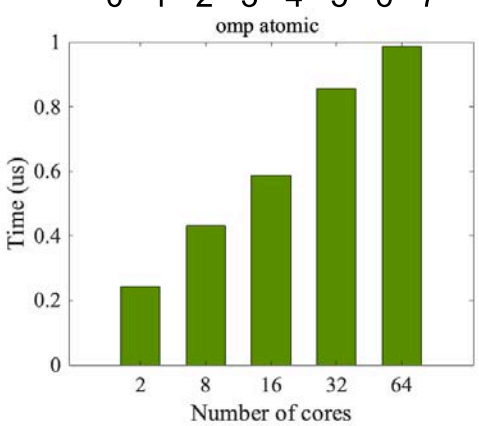


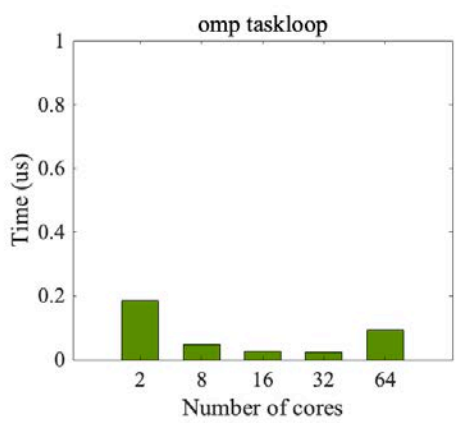
(c) Level-sets generated.

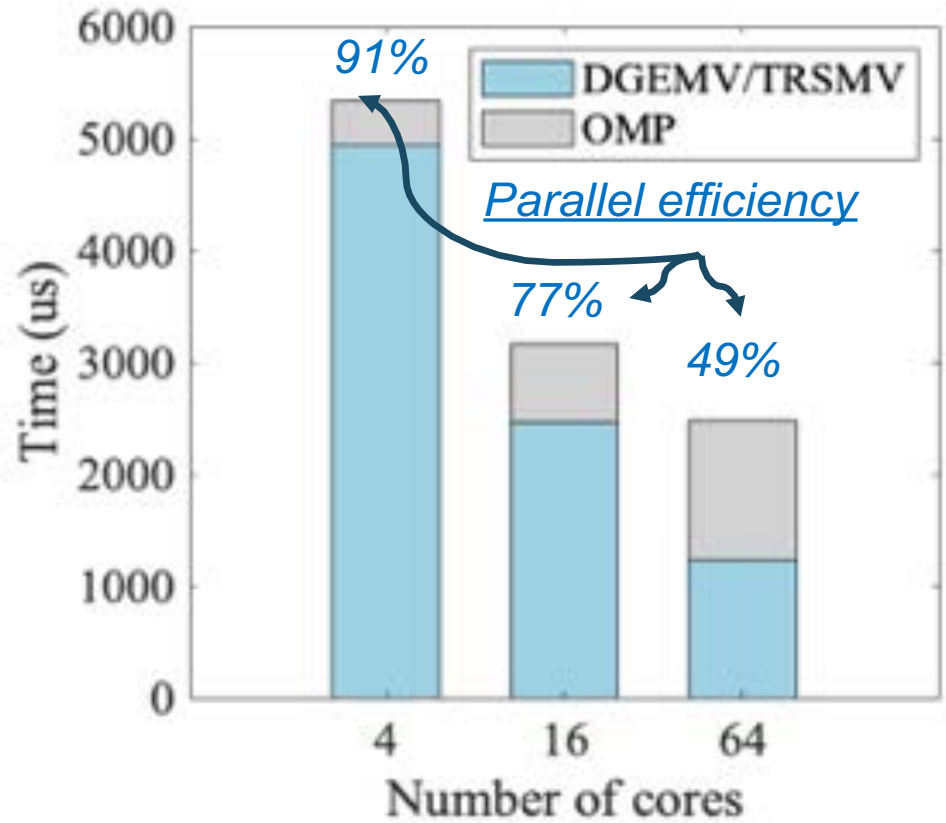
Example of CoDevelopment of Hardware and Software: SuperLU Dependency Tracking



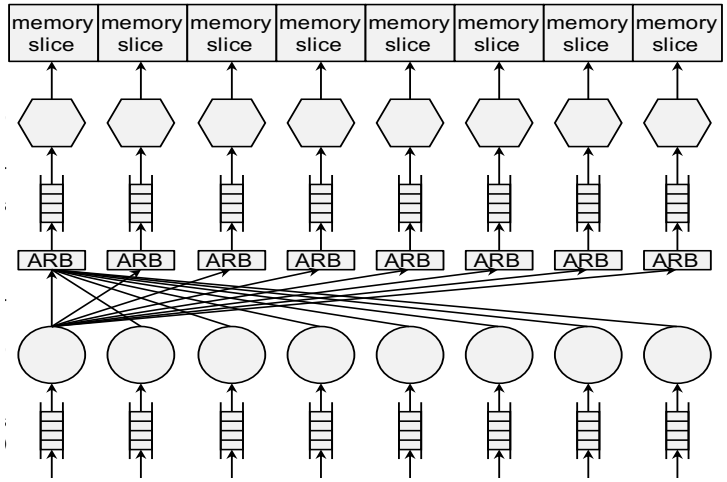








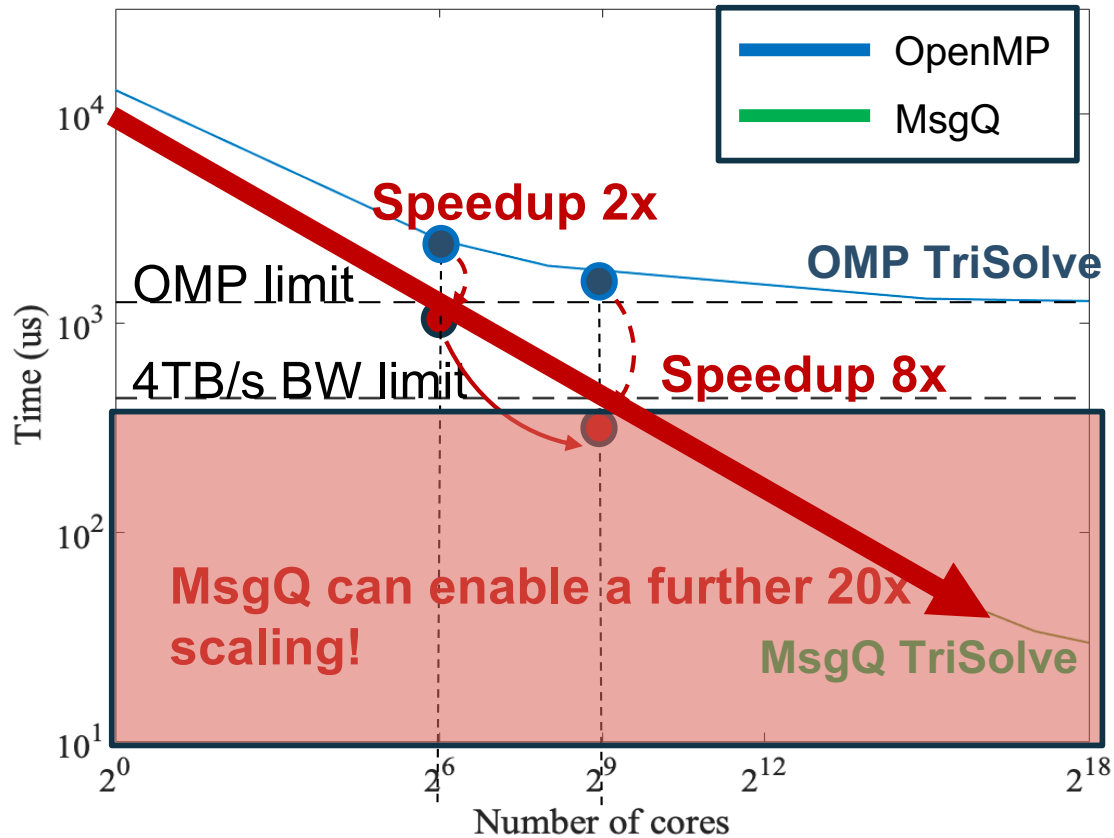
Benefit of MsgQ's on KNL-like architecture



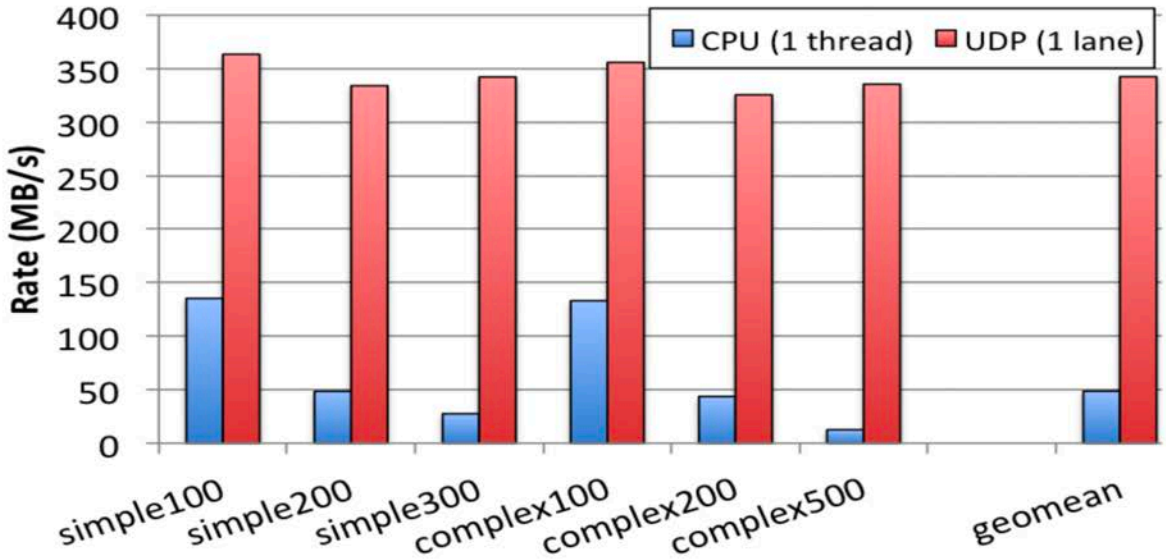
Algorithm: Redesign SuperLU algorithm to use MsgQ instead of atomics to track dependencies.

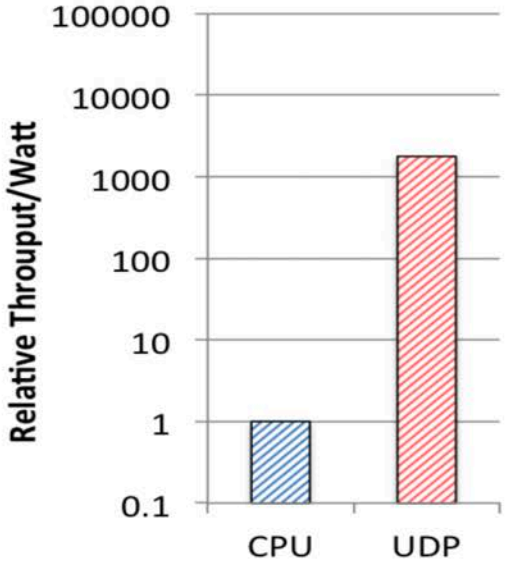
Performance:

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



Recode: *Regex 1-lane Performance and Energy Efficiency*

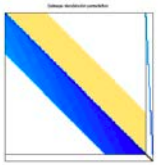
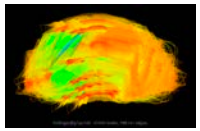
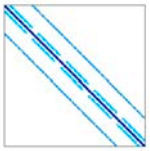
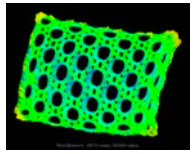
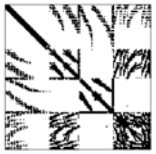
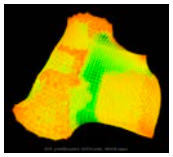
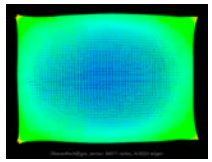
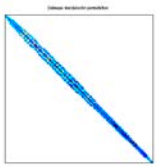
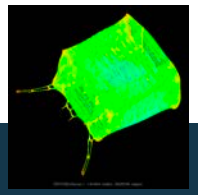


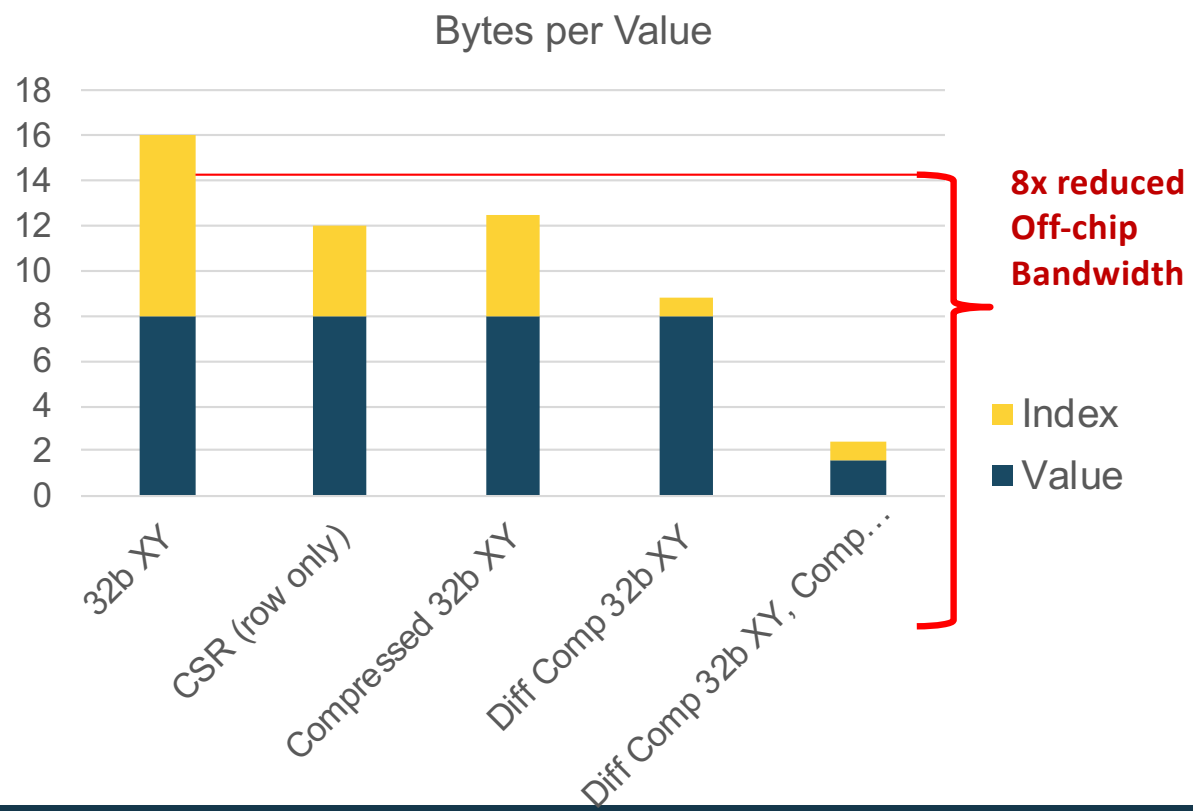


- **7x faster per lane than x86, 64 lanes => ~450x faster than single x86 thread**
- **Recode engine (UDP) scales to ~150 Gbps for a 64-lane Recode engine (<<1 watt total)**
- **128 tile chip could achieve 20 Tbps total line rate; 256 tiles => 40 Tbps**
- **Large pattern sets supported with NFA, and scale-out**

Extreme, Scalable Regex at 10-40 Tbps

SNAPPY: Sparse Matrix Compression Accelerator

Matrices	Spyplot	Visualization
g7jac160		
Xenon1		
Copter2		
Gas sensor		
Shipsec1		

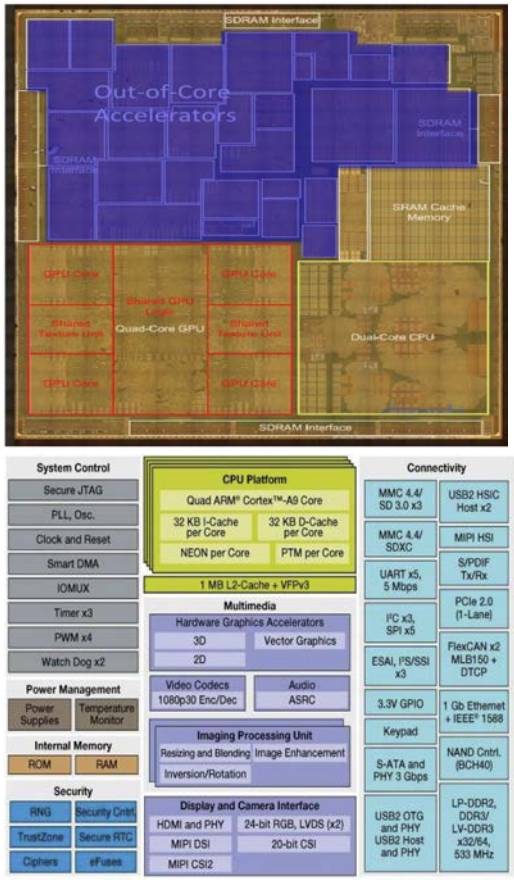


Fixed Function Accelerators Design Study

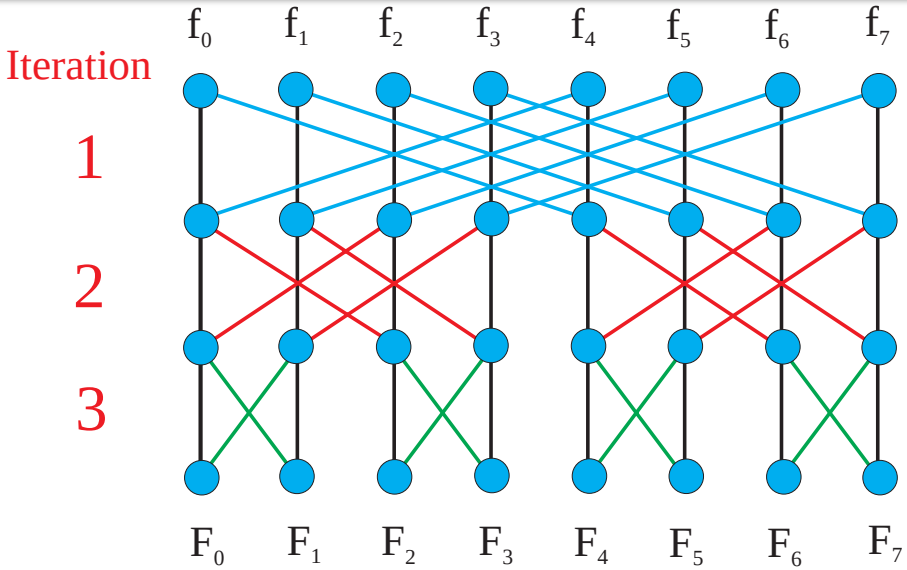
Dark Silicon

- **Adopt SmartPhone SoC Strategy --** *mix fixed-function accelerators with programmable cores*

- **Target commonly used scientific primitives/libraries**
  - BLAS (level 1,2,3)
  - FFT (FFTW or SPIRAL interface)



FFT Example *With FFTx (Francetti, Popovic, Canning)*



For FFT of size N

- Storage = N * operand_size
- Compute = 5/2 * N * log2(N) FLOPs
- Use Pseudo-2D algorithm for large FFTs

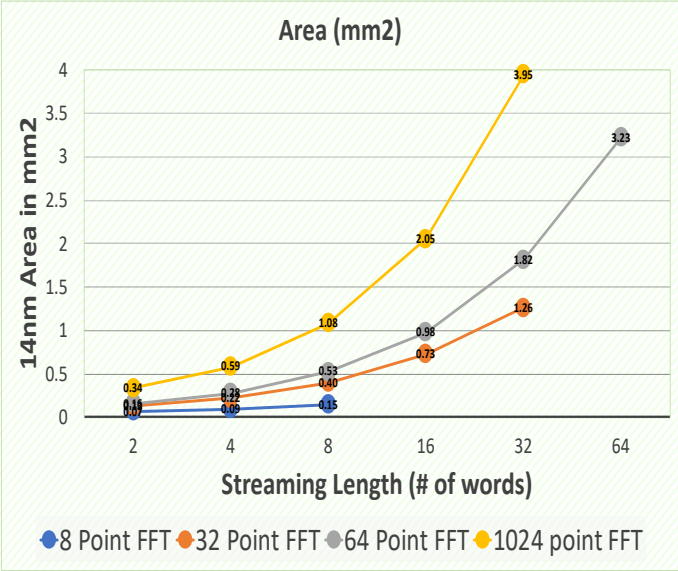
Single FFT Accelerator Resource

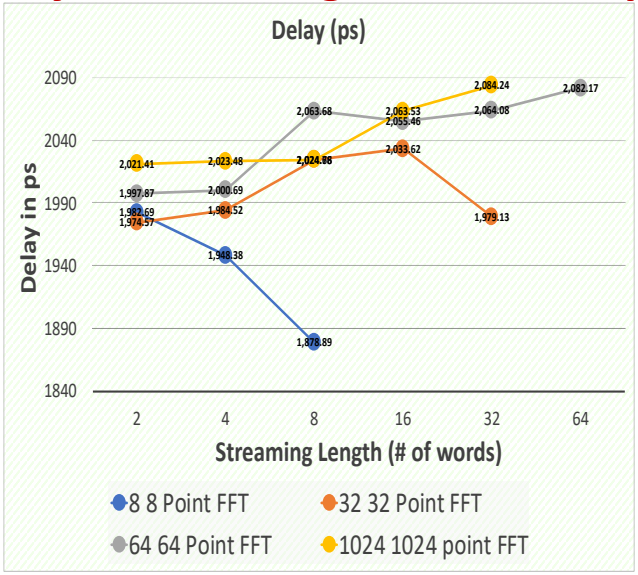
- **Assumptions: *Spiral HW Generator***
  - 1GHz @ 14nm technology node
  - 2M point transform (data off-chip)
  - HPC Challenge Benchmark: Single precision (Float32) complex, out-of-place
- **Limit: 100 GB/s off-chip memory**
  - 16k points on-chip engine
  - **Analytic model for FP limit ~1.5TFLOPs SP**
  - **4.5mm² area for compute @ 14nm**
- **Limit: 1TB/s off-chip memory**
  - *~10k MADD + ~5k add -> 15k FP@1GHz*

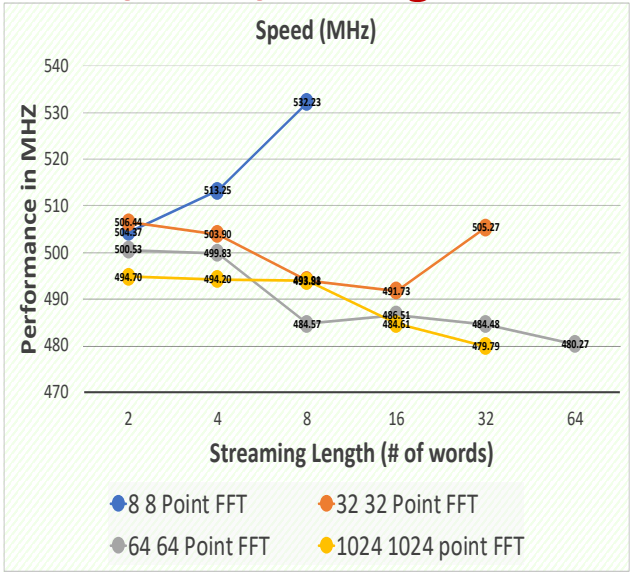
  Analytical model for FP limit ~15TFLOPs SP
  - **47mm² area for compute @14nm**

FFT Radix 2 RTL generated by SPIRAL – @14nm

Run RTL through synthesis to get accurate power/area/timing







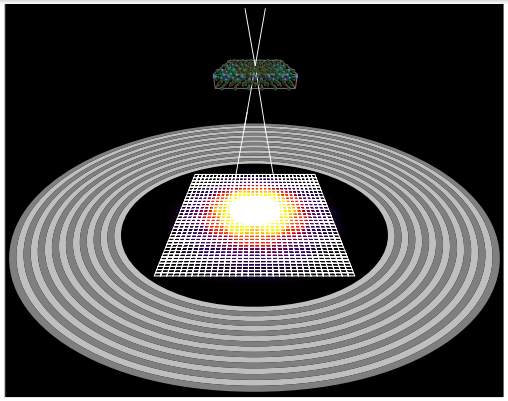
Chip-layout at 14nm using Mentor Design Synthesis Flow

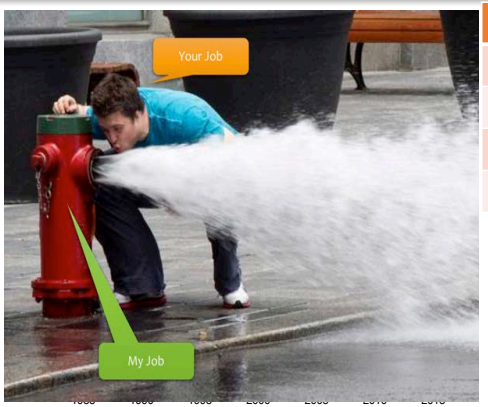
- **Shows 2x improved density improvement over analytic model, but 2x lower clock**
- **Floating point multiplier is the Critical path around 1900 ps leading to**
  - 500 MHz design for standard cell based synthesis
  - Improved StdCell library (better than OpenSDK) could result in further improvements

BERKELEY LAB

Results for RISC-V FFT Accelerator for CryoEM

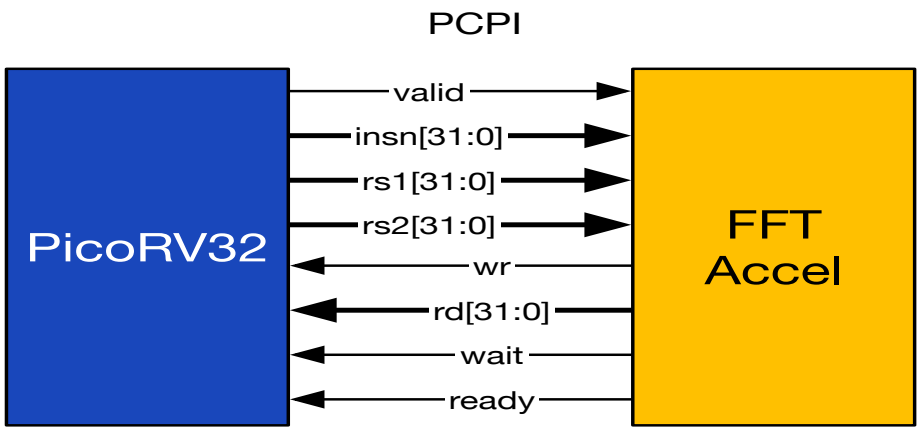
Benchmarking FFT Accelerator for image analysis ***(Donofrio, Fard)***



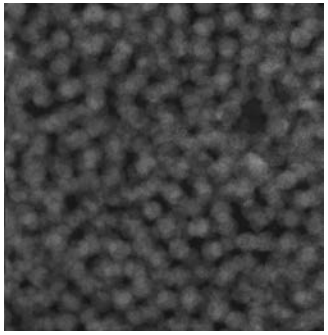


Detector / Microscope Installation Year

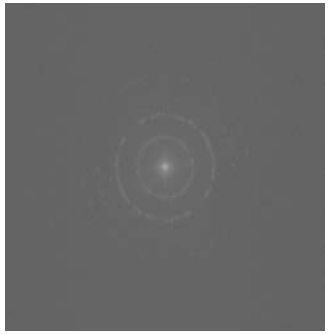
Instruction	opcode[3:2]	Description
fft_config	10b	Configures FFT parameters
fft_status	01b	Reads FFTAccel status registers
fft_start	11b	Starts FFT processing
fft_stop	00b	Stops FFT processing



Original Image



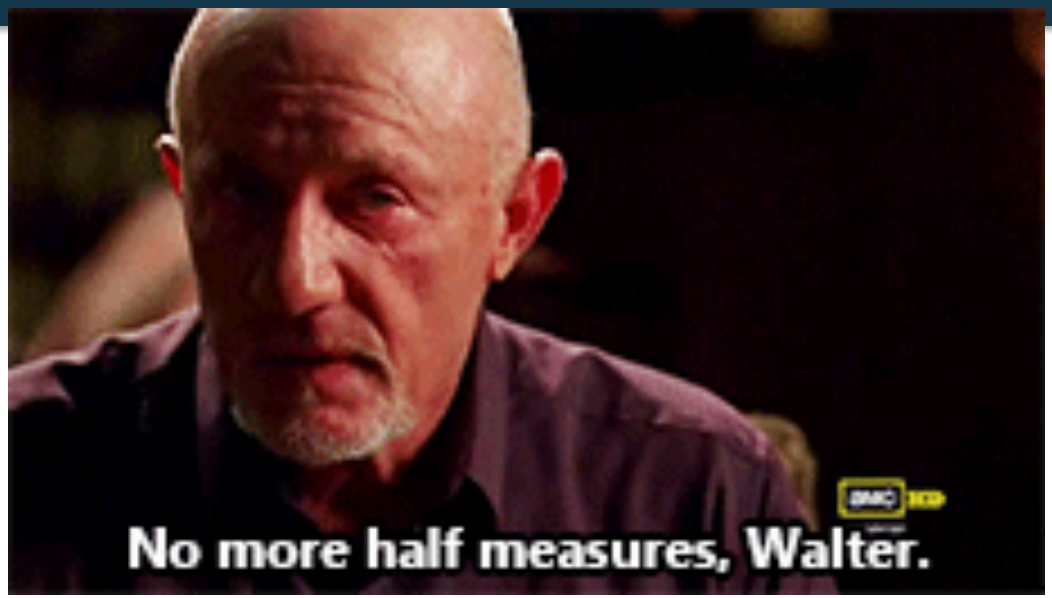
FFT



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**



Full Measure

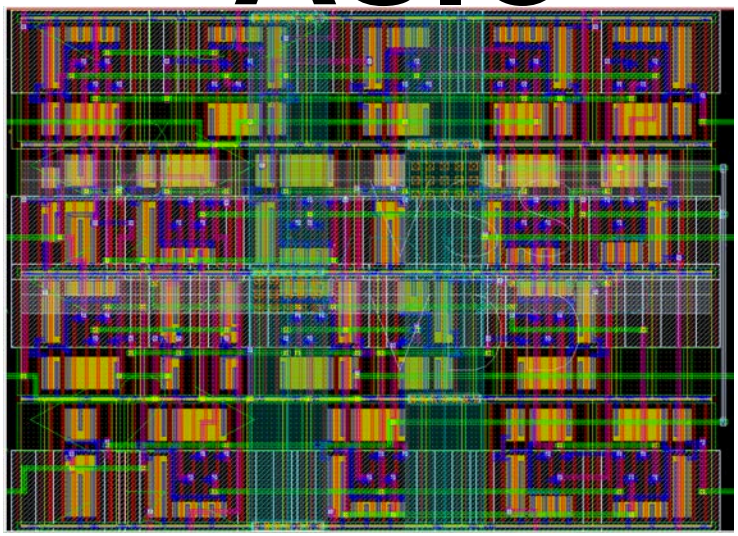
Full Custom Acceleration for Targeted Science
(Industrializing use of Anton or GRAPE-like technology)

FPGA vs. ASIC

FPGA

Cost for first FPGA (NRE):	$2,500-$7,500
Cost for 20,000th :	$2,500-$7,500
Clock Rate:	0.1-0.3Ghz

ASIC

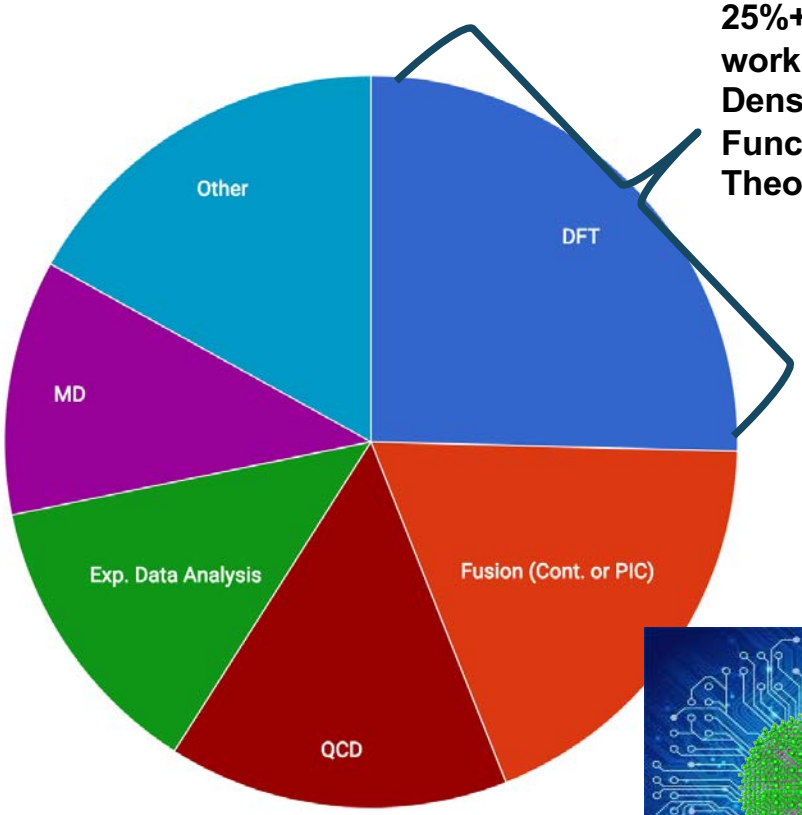


Cost for first ASIC (NRE):	$2M-$15M
Cost for 20,000th :	$150-$250
Clock Rate:	1-2 Ghz (10x)
Area Efficiency:	10x FPGA
Energy Efficiency :	10x-100x FPGA

Example Algorithm-Driven Design of Hardware Accelerators

Example: LS3DF/Density Functional Theory (DFT)

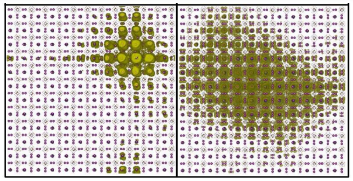
- **What: Design the hardware accelerator around the target algorithm/application**
  - Purpose-built acceleration
  - Lab-led reference design
- **Why: Huge opportunities to improve performance density and efficiency**
  - FFT hardware accelerator 50x-100x higher performance density than GPU or CPU+SIMD (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)

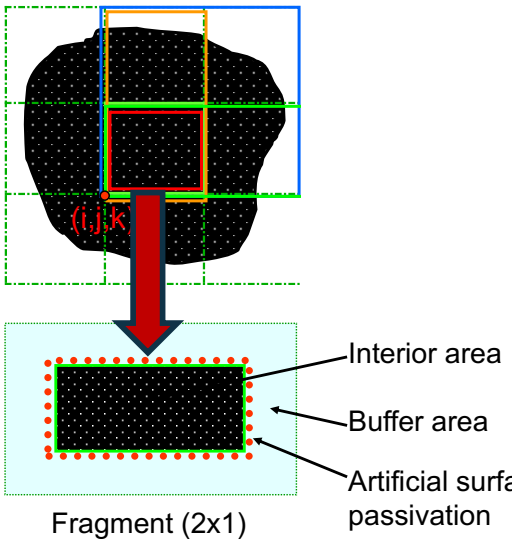


25%+ of DOE workload is Density Functional Theory (DFT)

The DFT kernel for each fragment

Communication Avoiding LS3DF Formulation – Scales O(N)





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

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

Sub_diag, *

Hpsi, *

Precond. CG step

Projection, *

Line minimiz.

Orth., *

Sub_diag, *

O(N² Log(N))

Comm bound if non-local

3D parallel FFT

TSQR & Choelesky

ZGEMM

O(N³)

Compute-bound

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

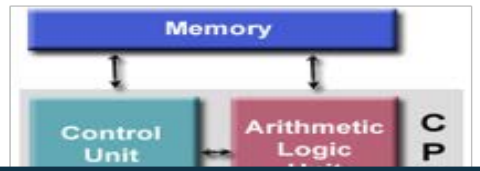
One patch per FPGA
400 bands/patch

Compute Intensive Kernels
Targeted for HW Specialization

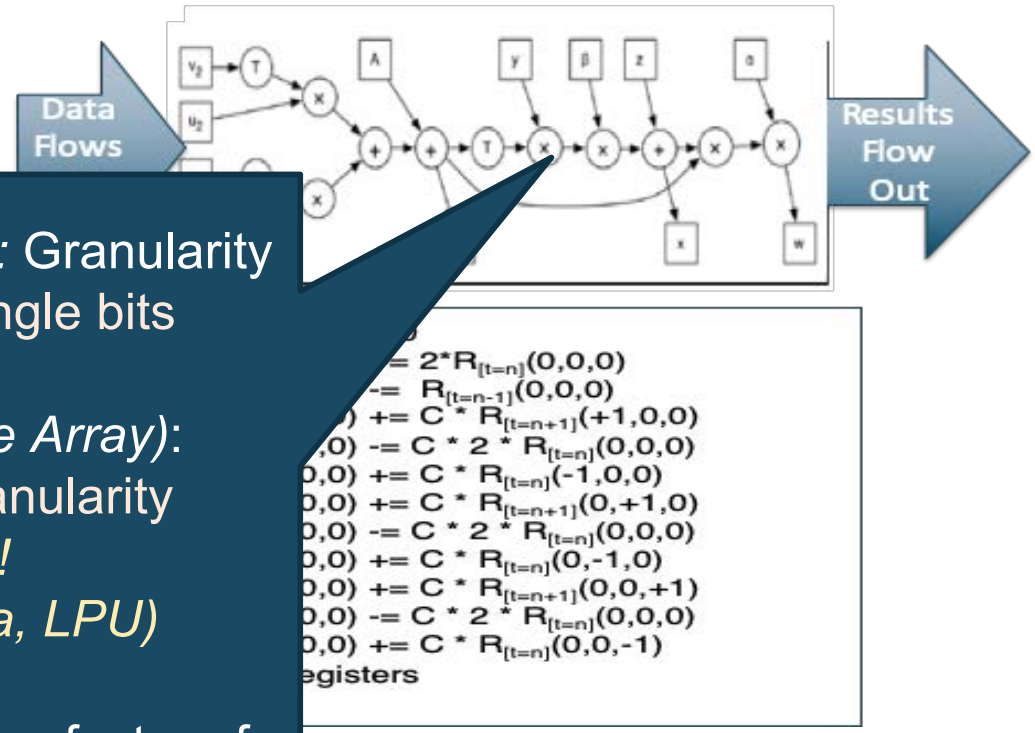
Von-Neumann Instruction Processors vs. Hardware Circuits

(must redesign for static dataflow and deep flow-through pipelines)

Von Neumann CPU



Dataflow (FPGA, GraphCore etc.)



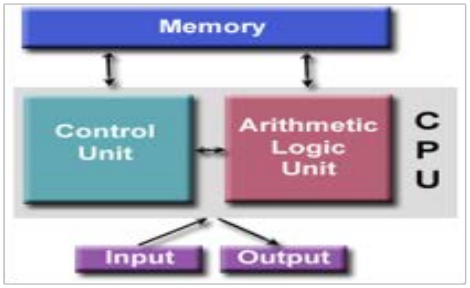
FPGA (*Field Programmable Gate Array):* Granularity of these operations and wires are single bits

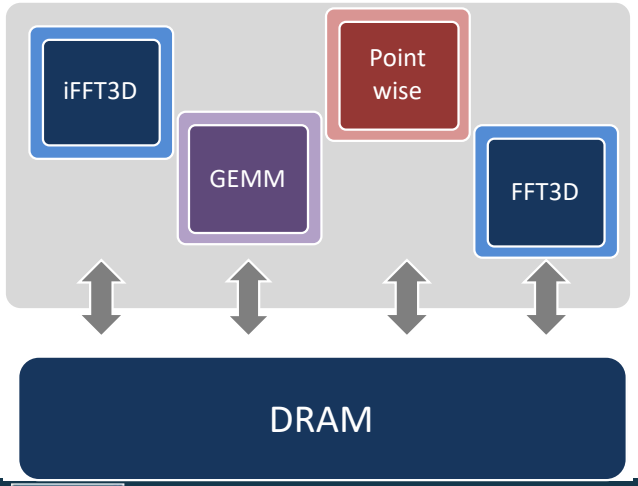
CGRA *(Coarse Grain Reconfigurable Array)*: Programmability & ALUs at word granularity *improves speed and density!!* *(Cerebras, GraphCore, SambaNova, LPU)*

ASIC or Chiplet (*custom circuit):* Another factor of 10x on density and energy efficiency.

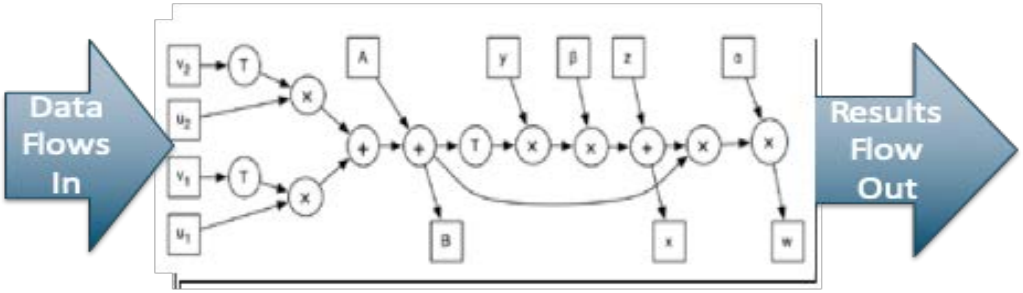
Algorithm Reformulated as Custom Circuit

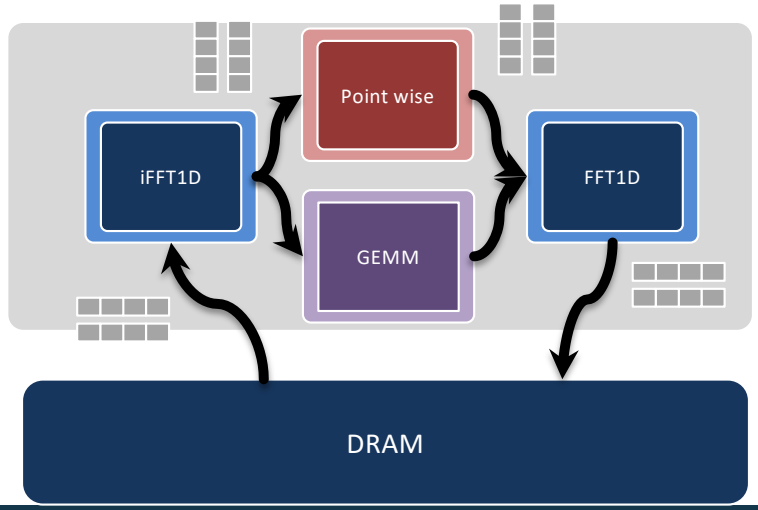
Von Neumann CPU





Dataflow (FPGA, GraphCore etc.)

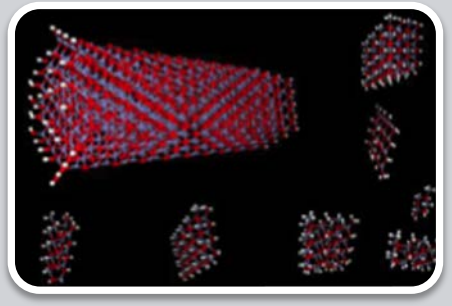


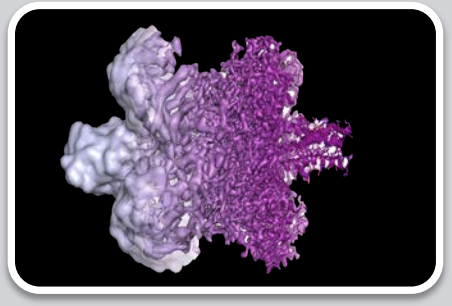


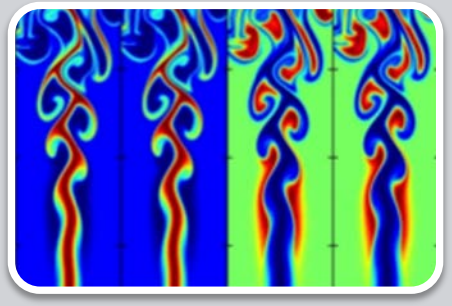
See Also Torsten Hoefler “StreamBLAS” for FPGA

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 solution

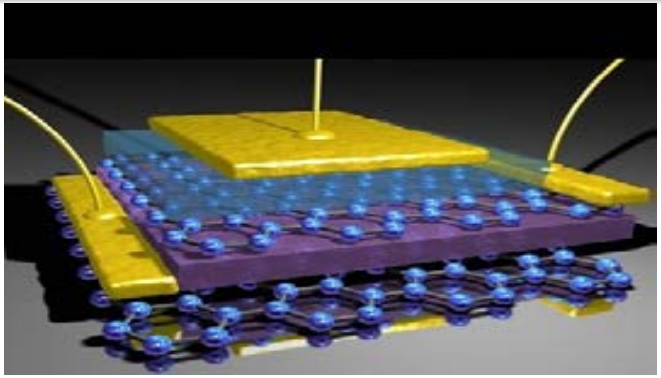
Digital fluid Accelerator

3D integration

Petascale *chip*

1024-layers

General / special HPC solution



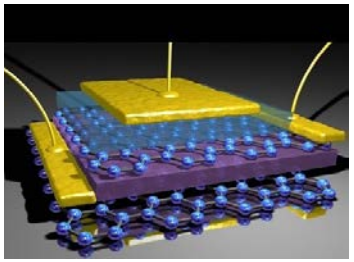
Post CMOS Device Technology

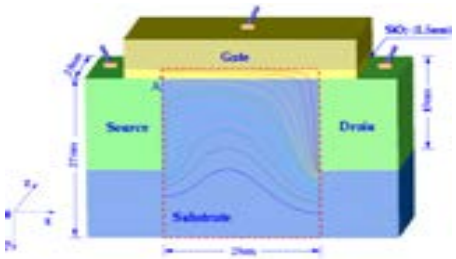
Accelerating the pace for discovery
for the future of Microelectronics

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**





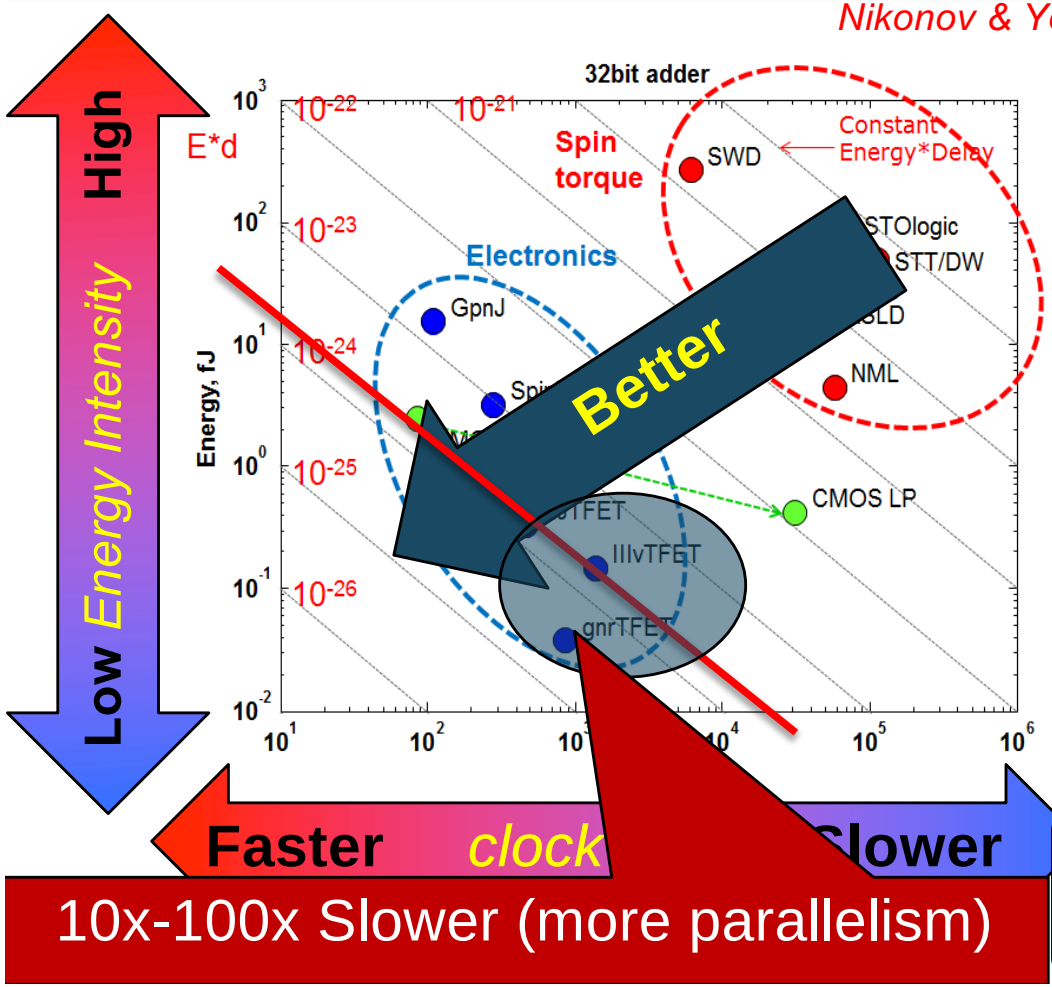
OSTP Report 2015: John Shalf
Robert Leland and Shekhar Borkar

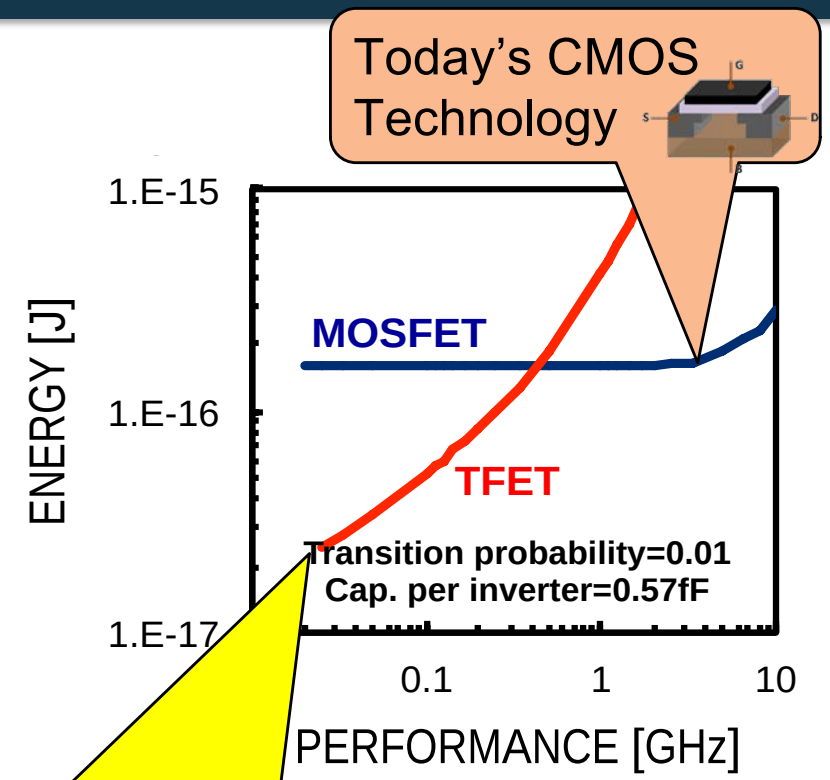
TABLE 1. Summary of techology options for extending digital electronics.

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

Nikonov & Young



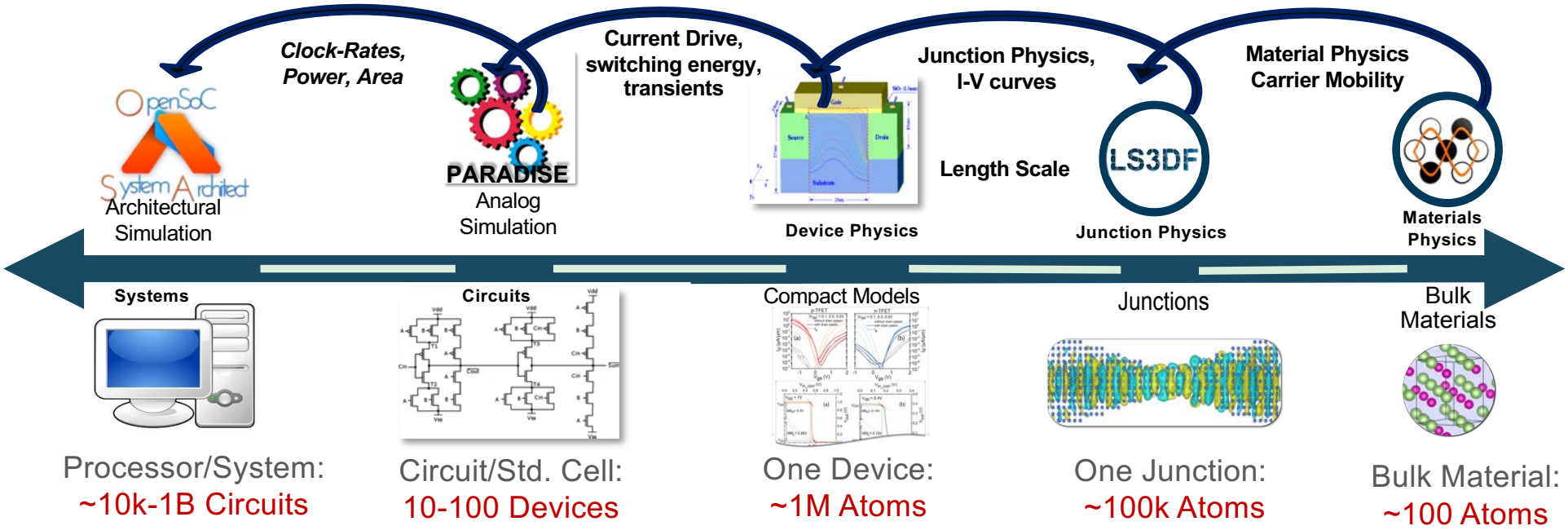


10x-100x Slower (more parallelism)

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

Multiscale Modeling to ccelerate Post-CMOS Development

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

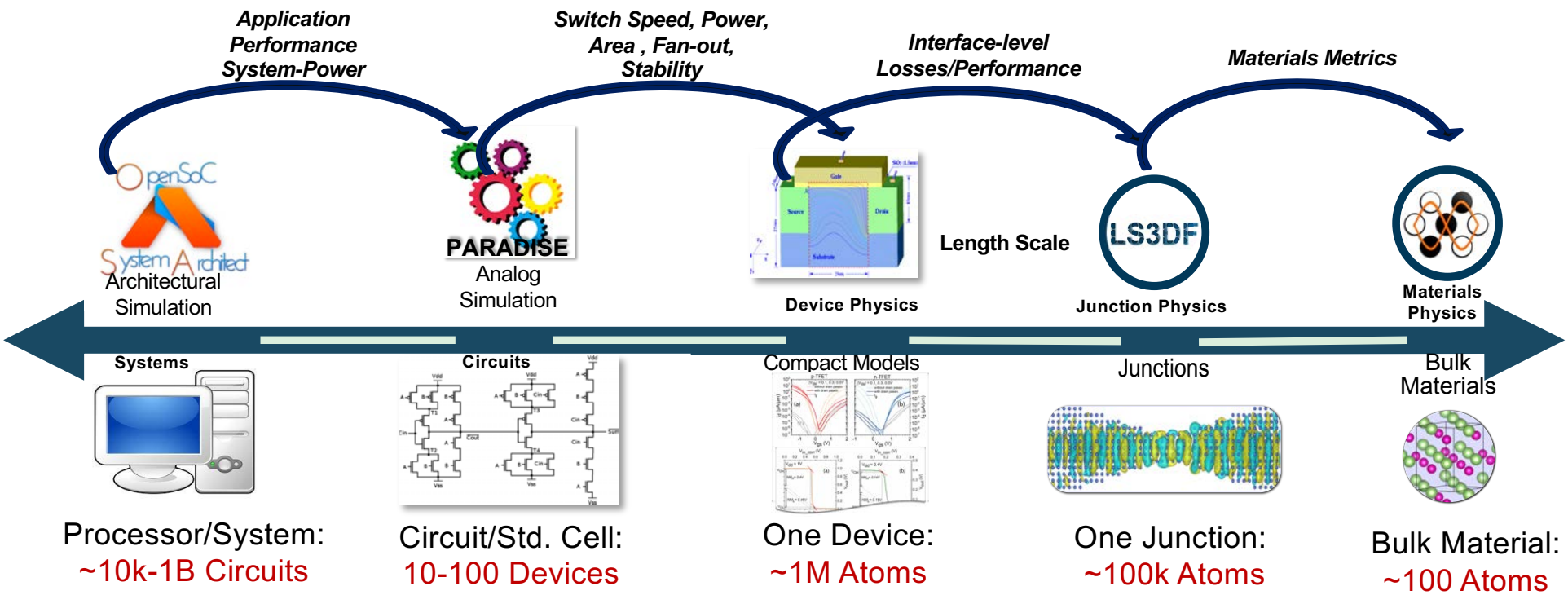


A holistic end-to-end modeling approach is required

BERKELEY LAB

Gap: Connecting and Scaling

Accelerated feedback path to focus device and material discovery process

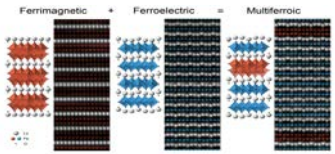


Length Scales

Integrated Plan to Accelerate Microelectronics Discovery

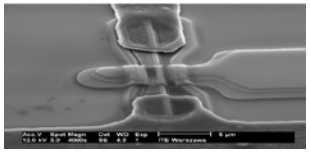
End-to-End Acceleration of Discovery and Evaluation of New Devices

Materials Discovery



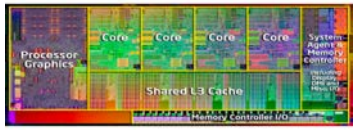
Computational Design
Synthesis
Characterization

ME Transistor



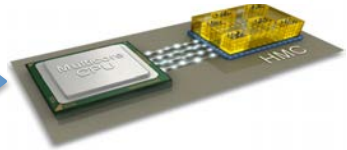
Device Design
Fabrication
Parametrics

Architecture



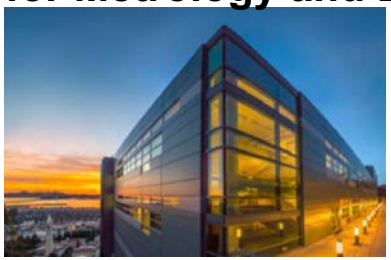
RTL/Gate Simulator
Power
Delay

System



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

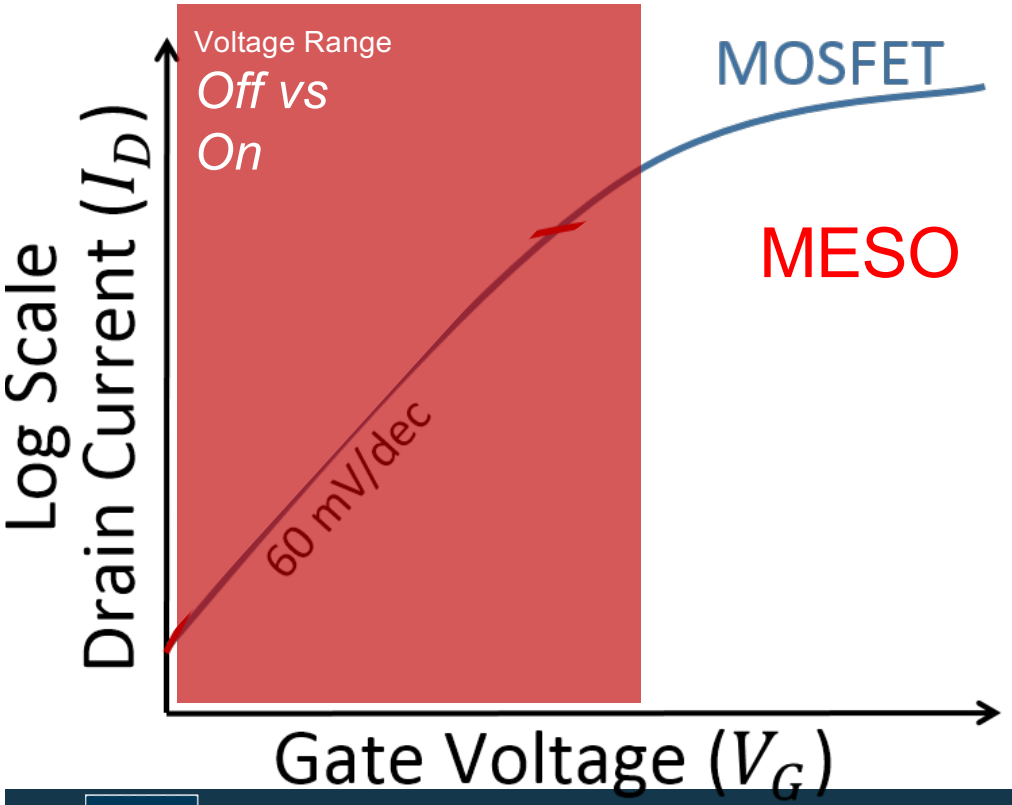
EUREKA CXRO

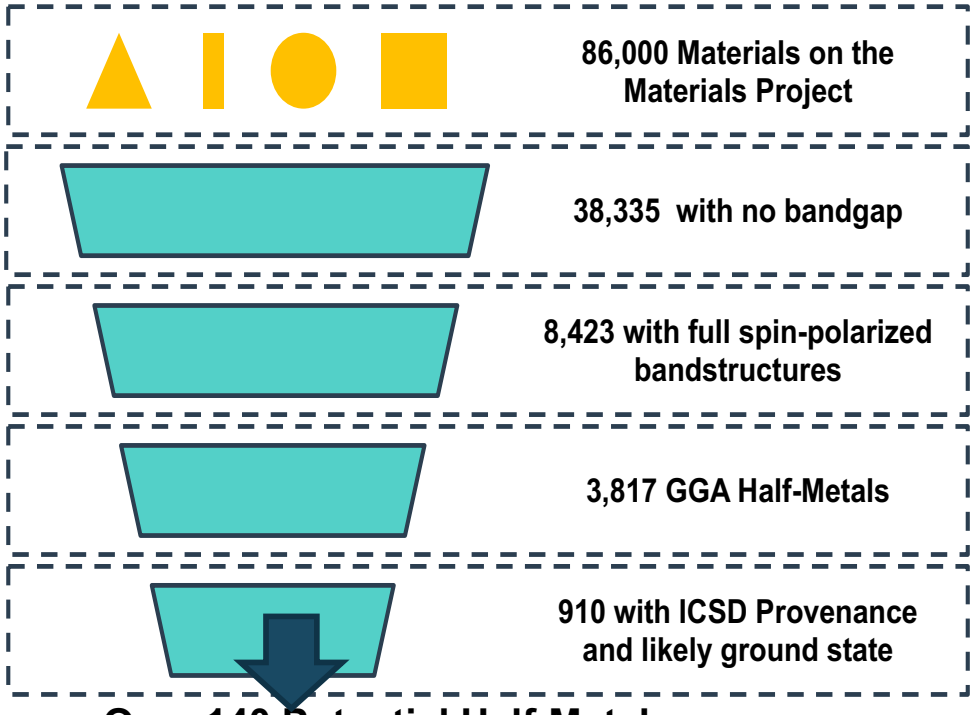
Physical, Chemical, Materials and Computer Sciences

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

PARADISE: Post-Moore Architecture and Accelerator Design Space Exploration

George Michelogiannakis & Dilip Vasudevan

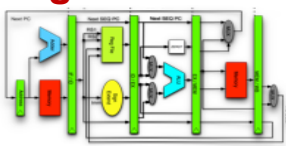
- **Multiple devices, memories, and other "post Moore" technologies in development**
- **Evaluating each in isolation misses big picture**
  - Devices can be better designed with high-level metrics
  - Architects can evaluate how exploit new technologies

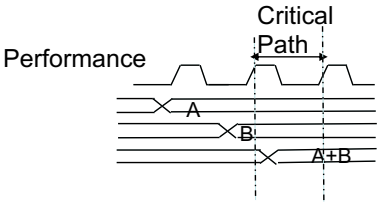
Until now, we lacked the tools to do so systematically and rapidly for many technologies

(PARADISE addresses that gap)

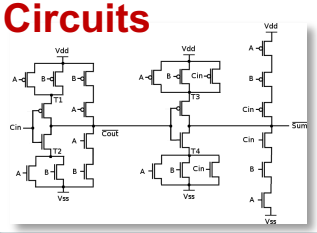
Systems

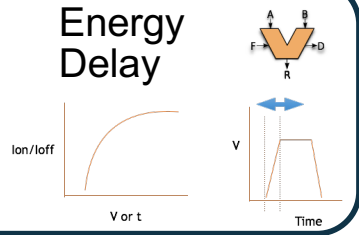
Logic Blocks



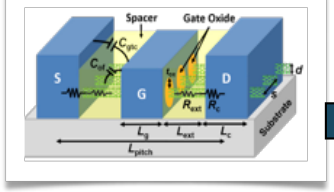


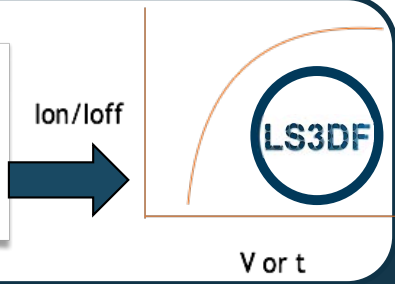
Circuits



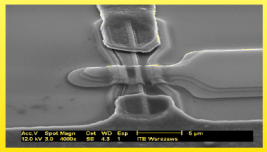


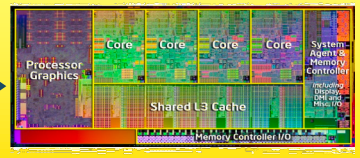
Devices

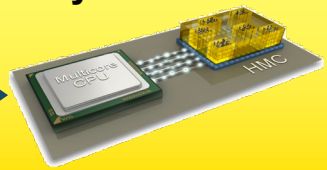




Transistor/Devices → **Architectures** → **Systems**

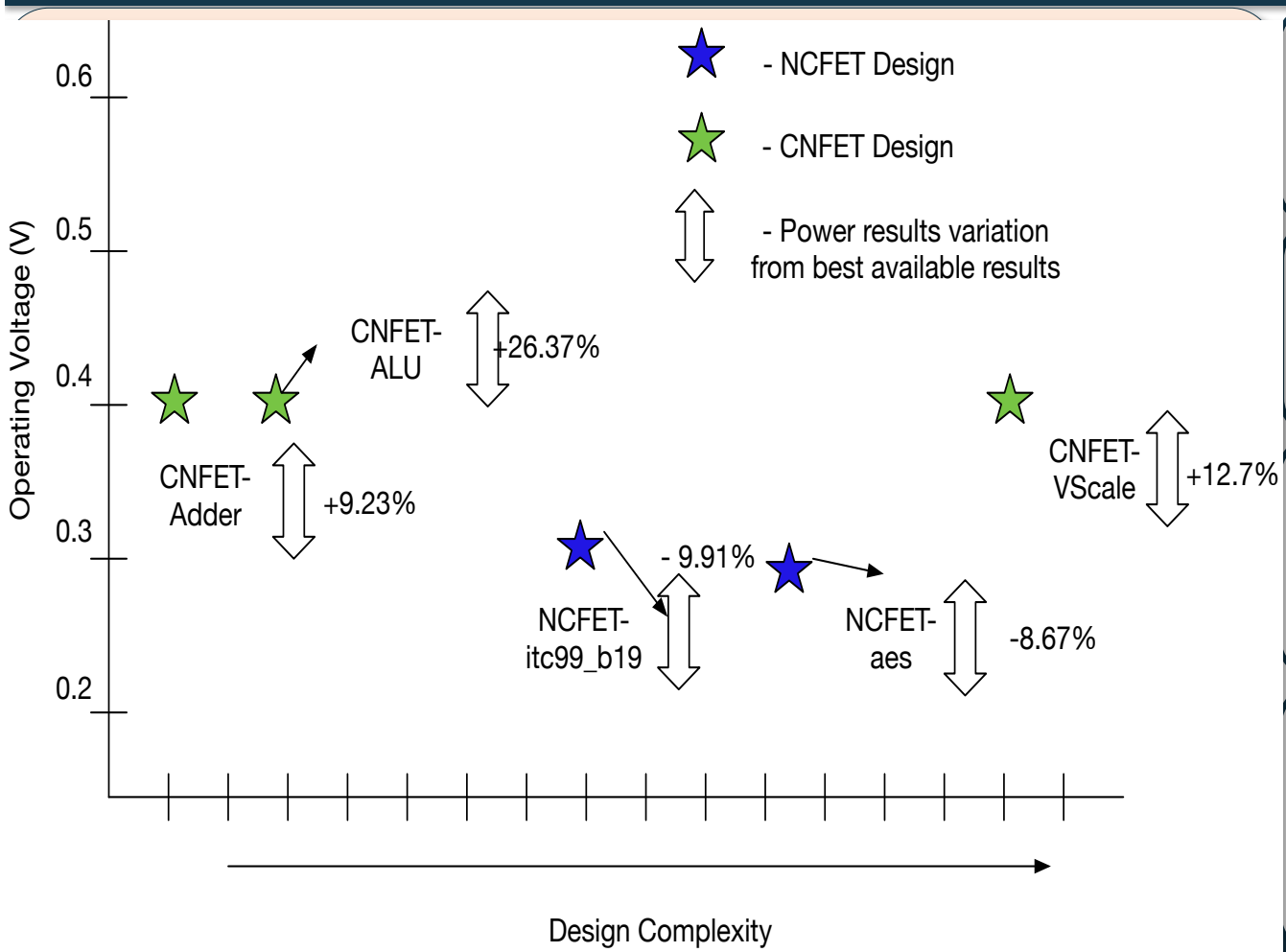


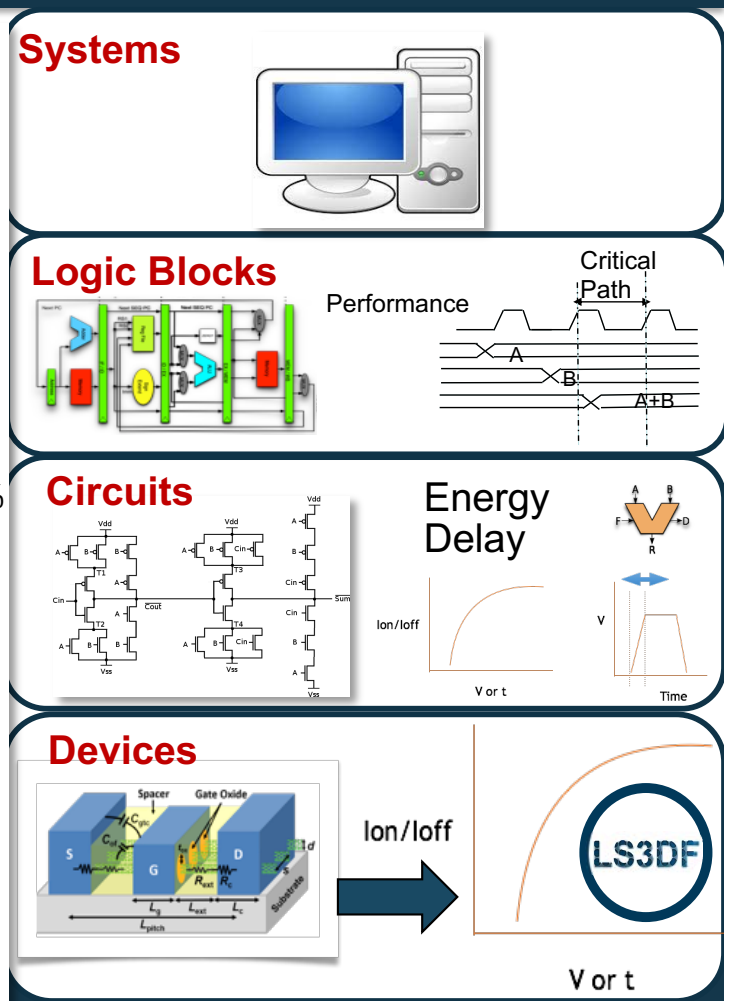




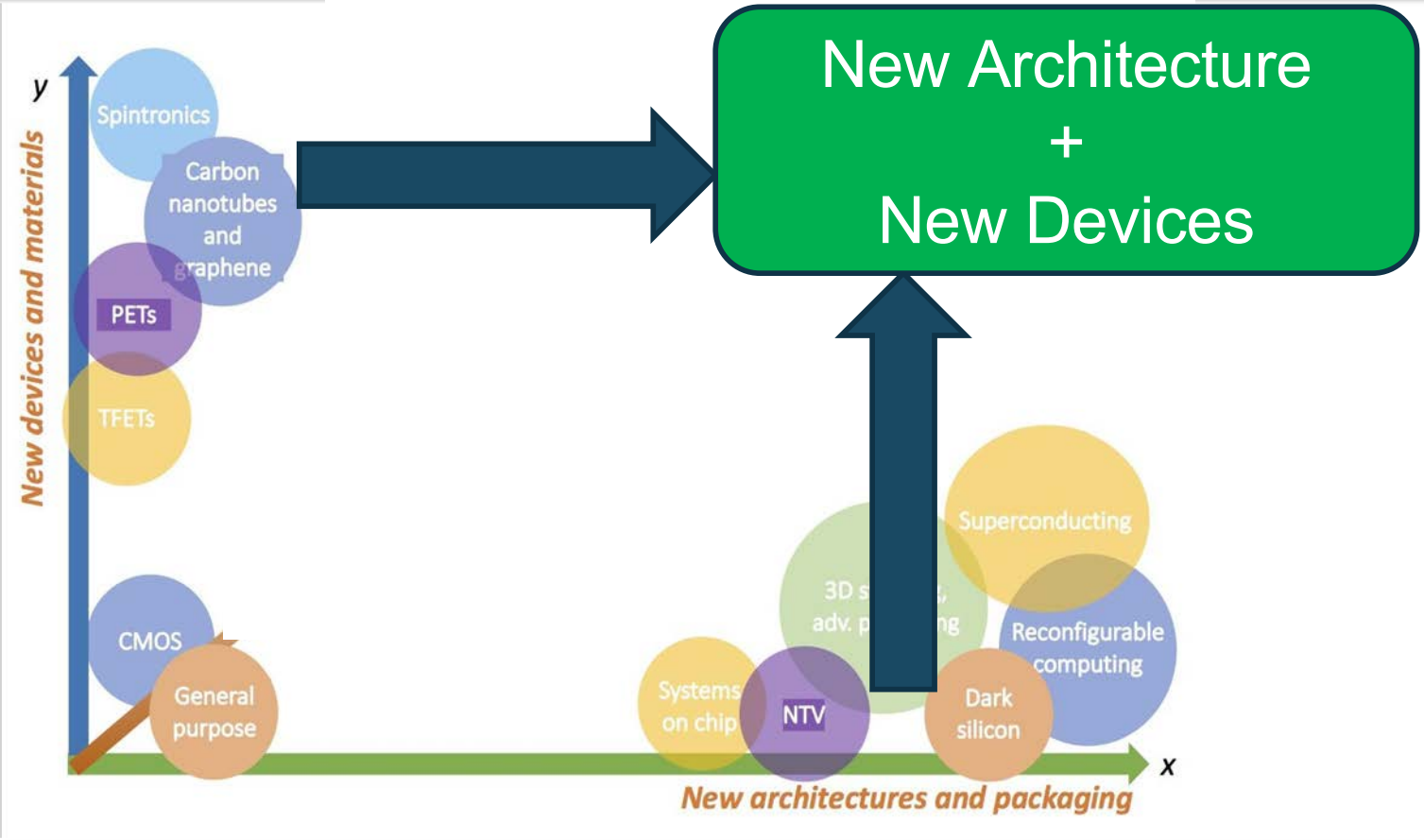
PARADISE: Post-Moore Architecture and Accelerator Design Space Exploration

George Michelogiannakis & Dilip Vasudevan



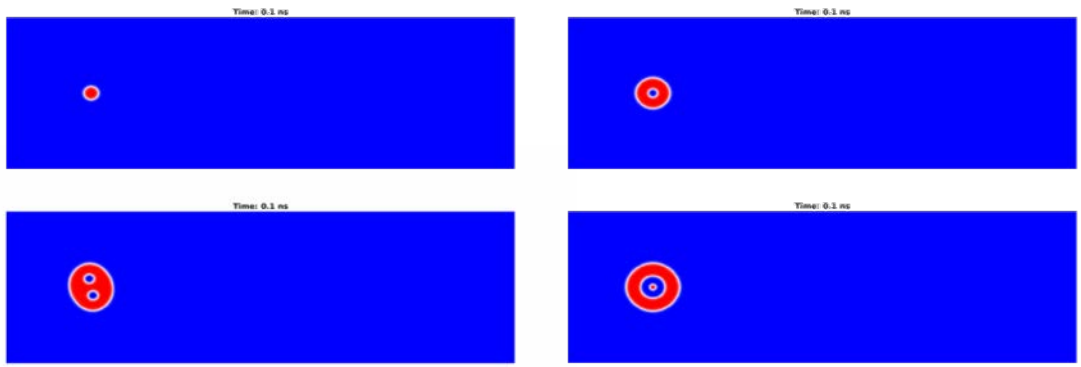


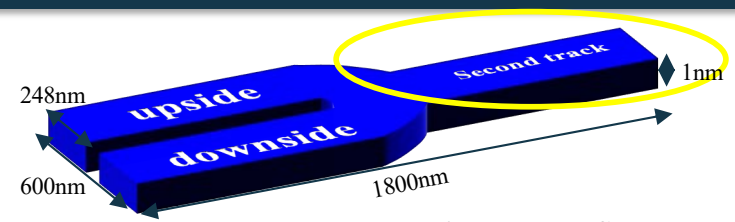
The Sum of the Parts is Greater than the Whole



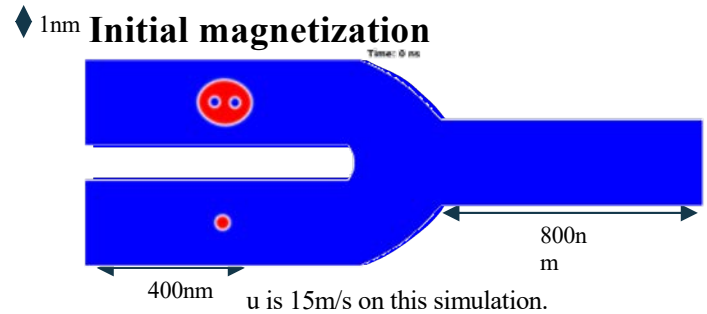
Skyrmions “bags” for Multi-Valued Logic

Four type of skyrmion bags moving by STT to check skyrmion Hall effect.



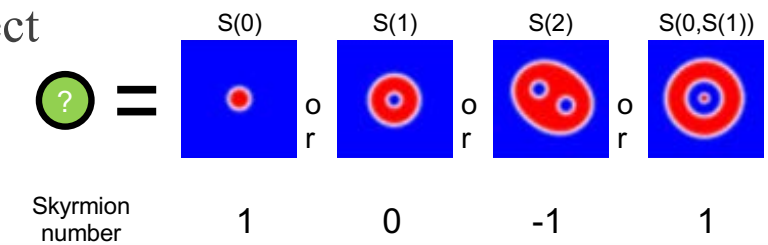


We considered only STT

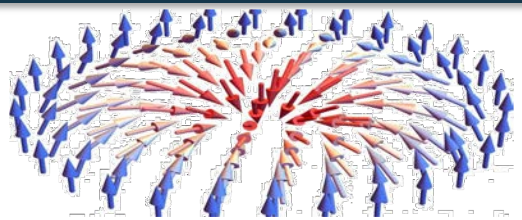


u is 15m/s on this simulation.

From this results, we can check velocity while Hall effect dominant case and edge effect dominant case.



Skyrmion-based Spiking Neural Networks



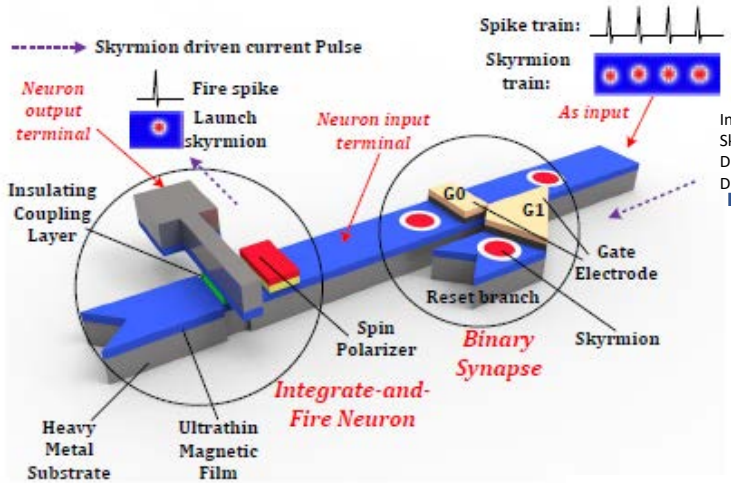
1 A:0, B:0, Y:0

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

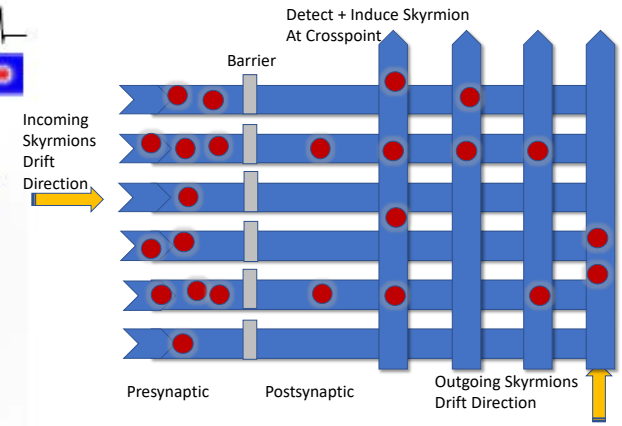
A:0, B:1, Y:0

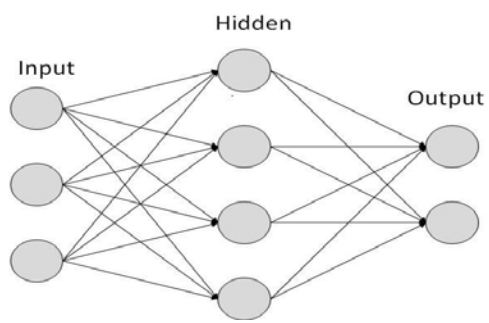
3 A:1, B:1, Y:1

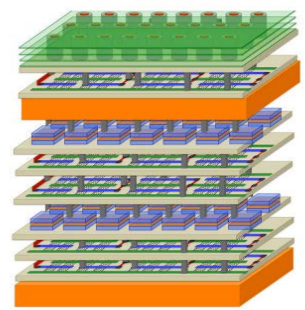
Z. He et al., 1705.02995v1 (2017)



Dilip Vasudevan & Mi Young Im



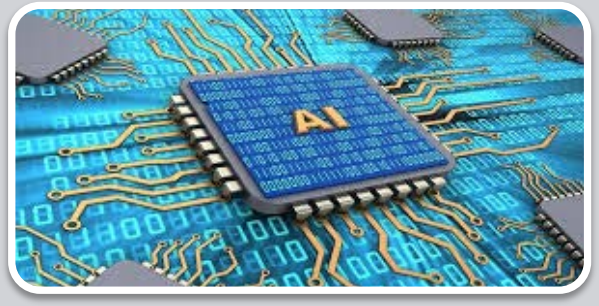


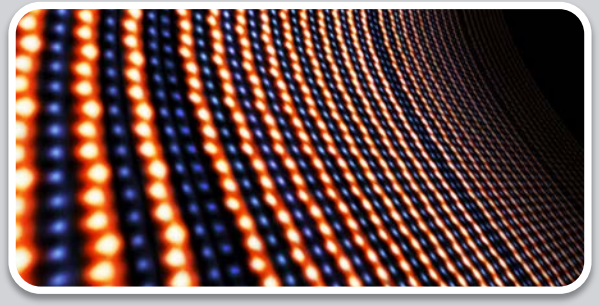


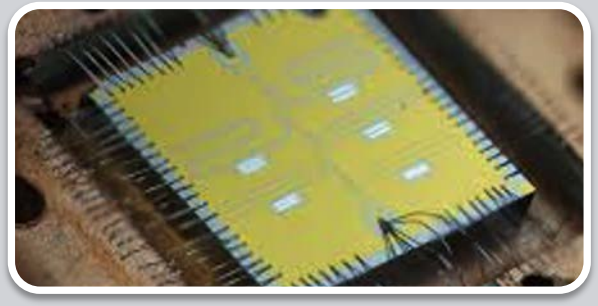
Conclusions

- **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**

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

Data Movement Challenge

Photonics and Advanced Packaging

http://www.padalworkshop.org/

Data Movement Costs:

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

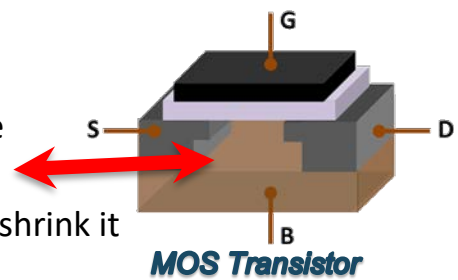
Energy Efficiency of copper wire:

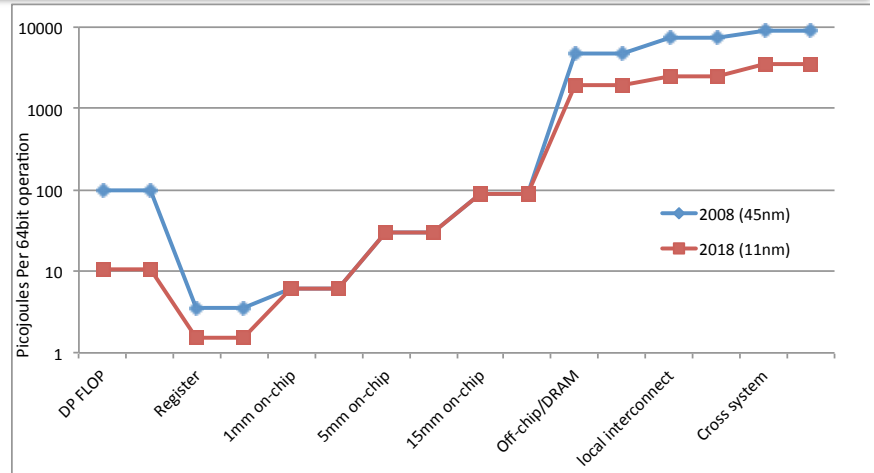
- **Power = Frequency * Length / cross-section-area**

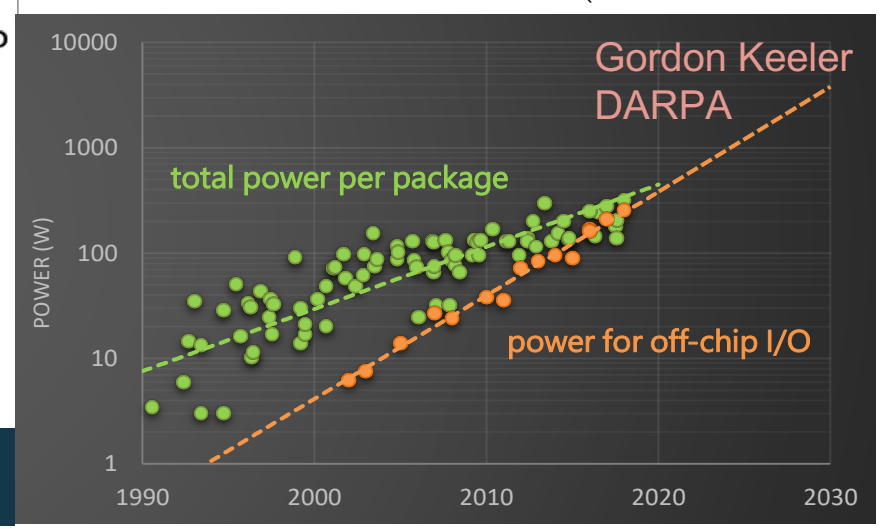
- Wire efficiency *does not improve* as feature size shrinks

Energy Efficiency of a Transistor:

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







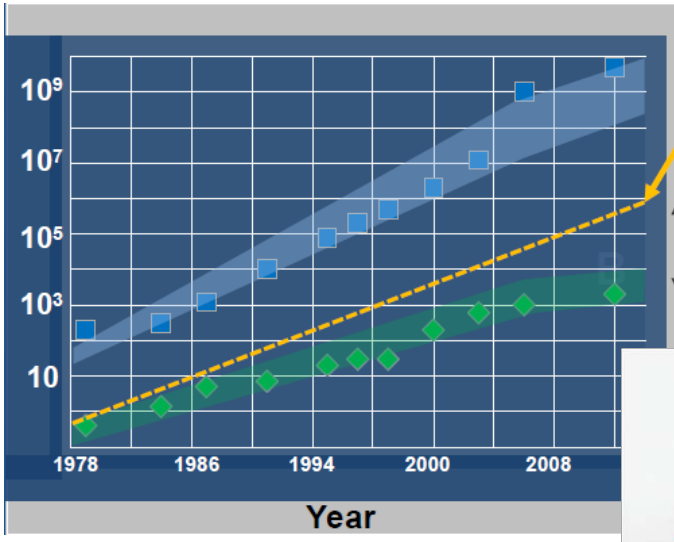
Net result is that moving data on wires is starting to cost more energy than computing on said data (interest in Silicon Photonics)

Package Performance is Pin Limited

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

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

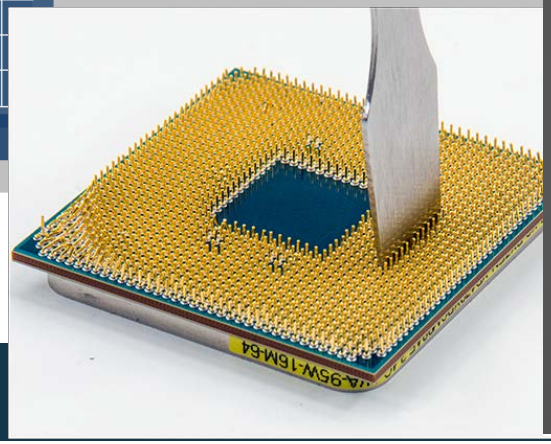
K = 0.82, a = 0.45 for early Microprocessors

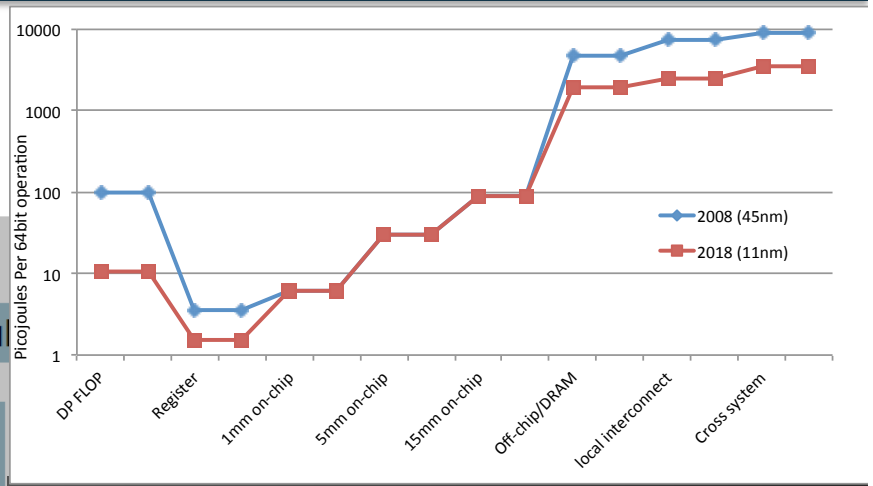


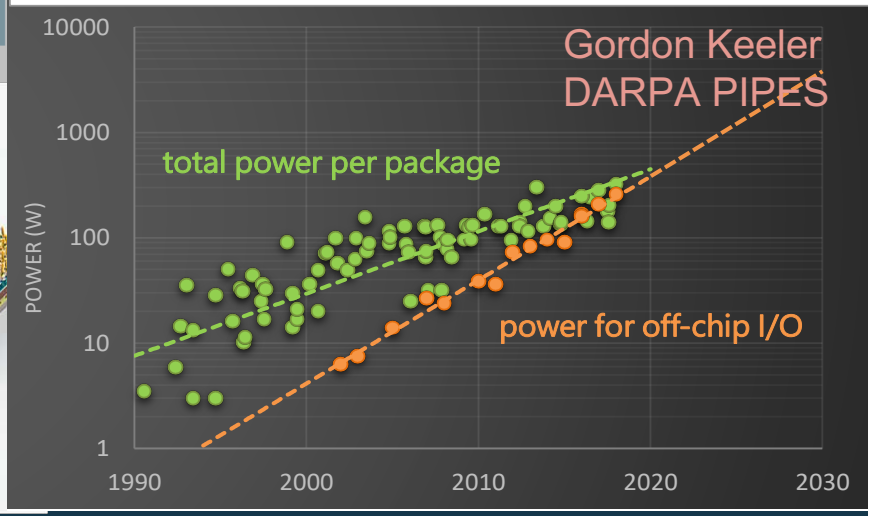
Pins x GHz from Rent's Ru

Bandwidth Gap:
~500 x and growing!

High SERDES rates run counter to end of Dennard Scaling





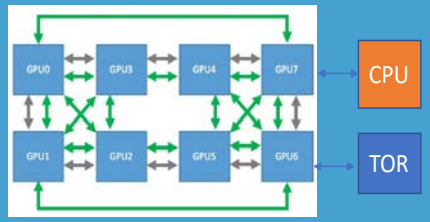


BERKELEY LAB

Diverse Node Configurations for Datacenter Workloads

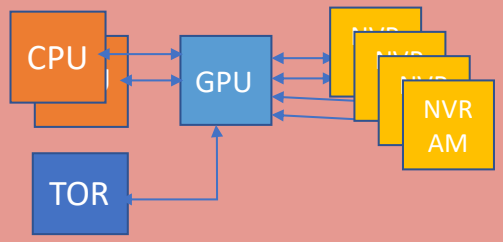
Training

- 8 connections: GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)



Data Mining

- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)

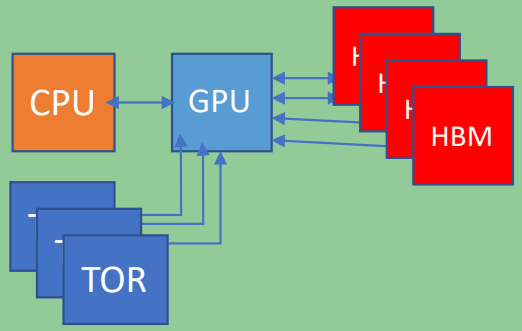


Inference

- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU

Graph Analytics

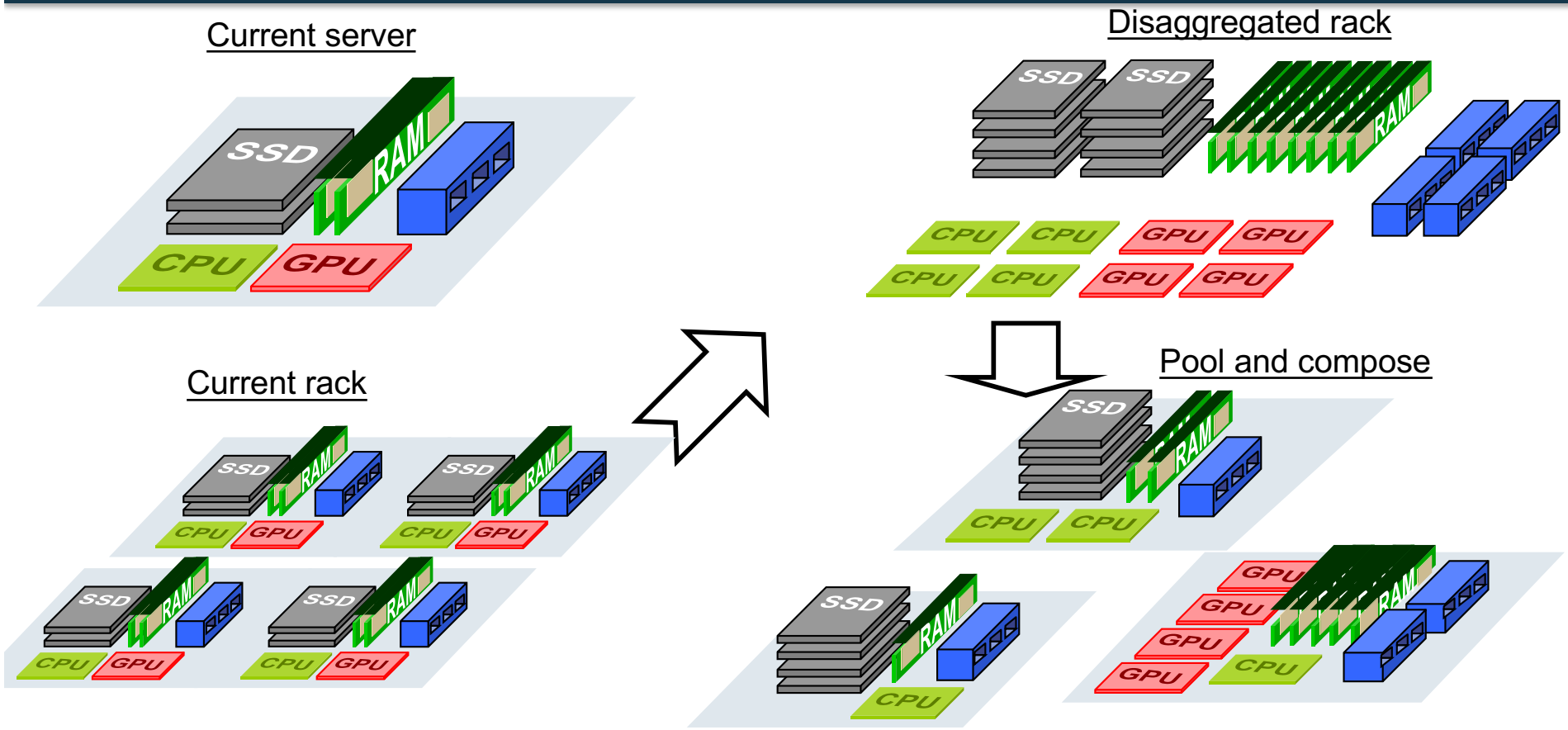
- 16 links HBM
- 8 links TOR
- 1 Link CPU



BERKELEY LAB

TOR | NVRAM | GPU | HBM | CPU

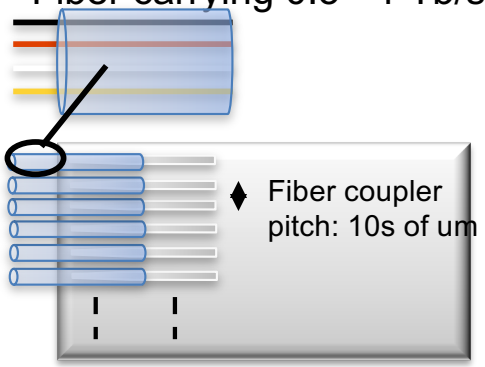
Disaggregated Node/Rack Architecture



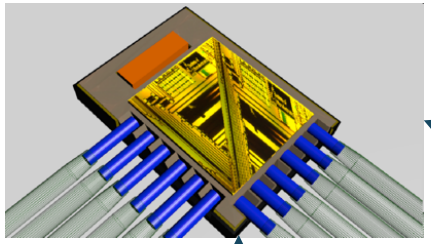
Most solutions current disaggregation solutions use Interconnect bandwidth (1 – 10 GB/s)
But this is significantly inferior to RAM bandwidth (100 GB/s – 1 TB/s)

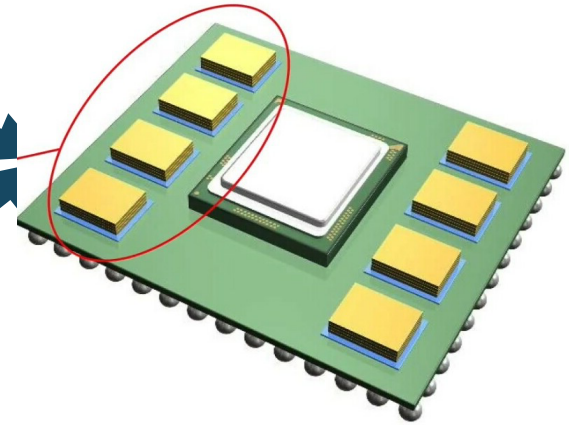
Photonic MCM (Multi-Chip Module)

Fiber carrying 0.5 - 1 Tb/s

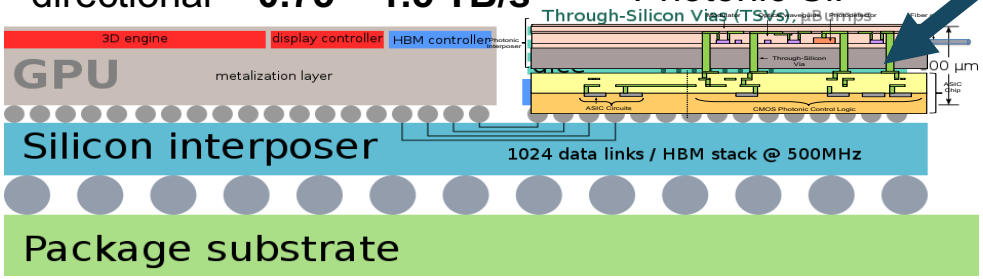


High-Density **fiber coupling array** with 24 fibers = 6-12 Tb/s bi-directional = **0.75 – 1.5 TB/s**





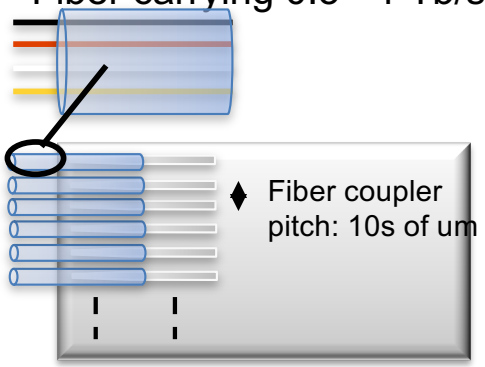
Photonic SiP



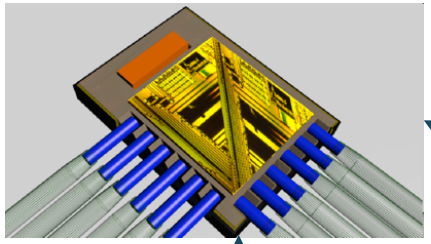
Optical switch

Photonic MCM (Multi-Chip Module)

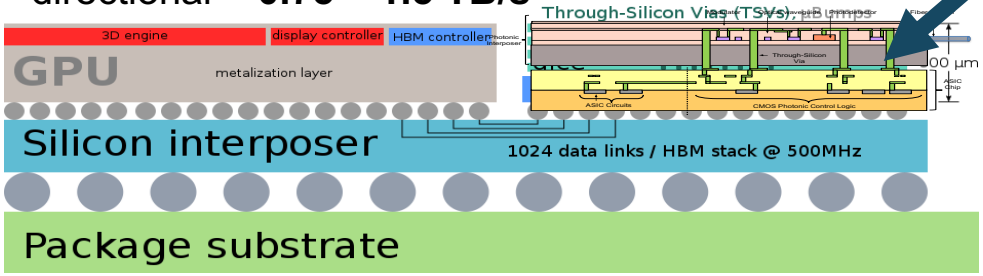
Fiber carrying 0.5 - 1 Tb/s

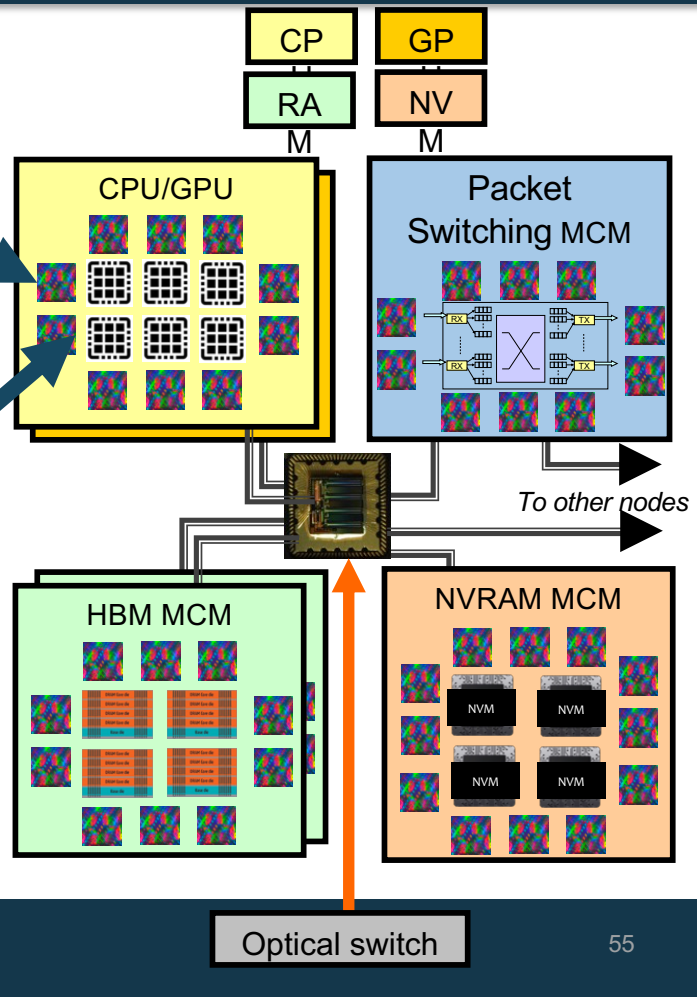


High-Density **fiber coupling array** with 24 fibers = 6-12 Tb/s bi-directional = **0.75 – 1.5 TB/s**

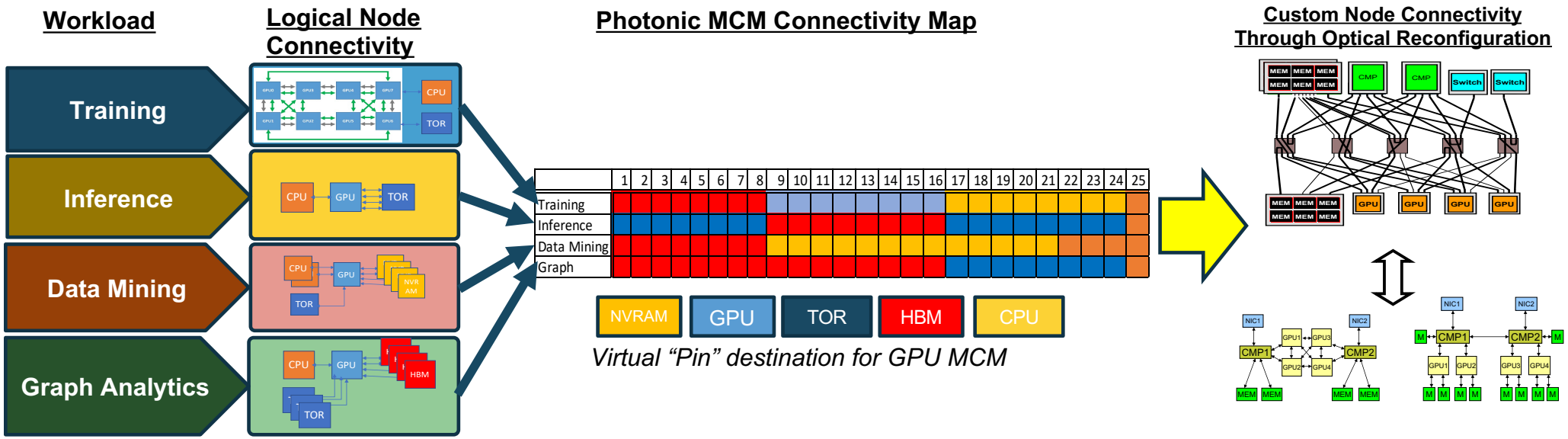


Photonic SiP



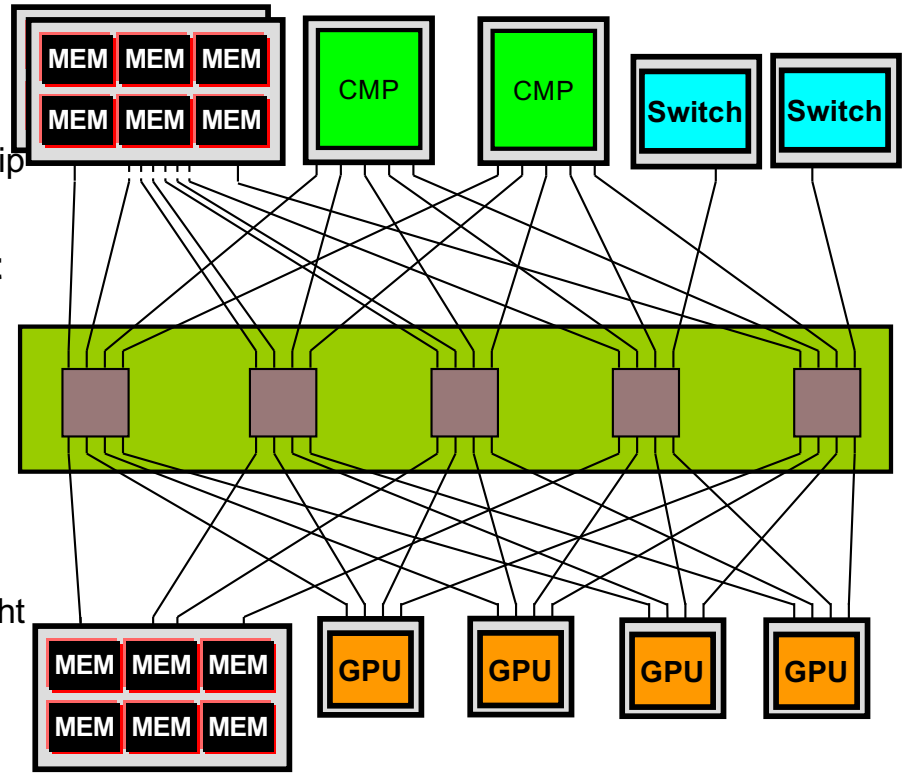


Case for Disaggregation from a Workload Perspective

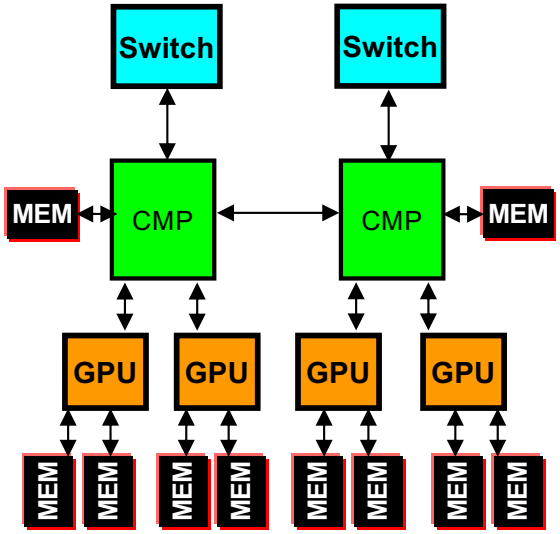


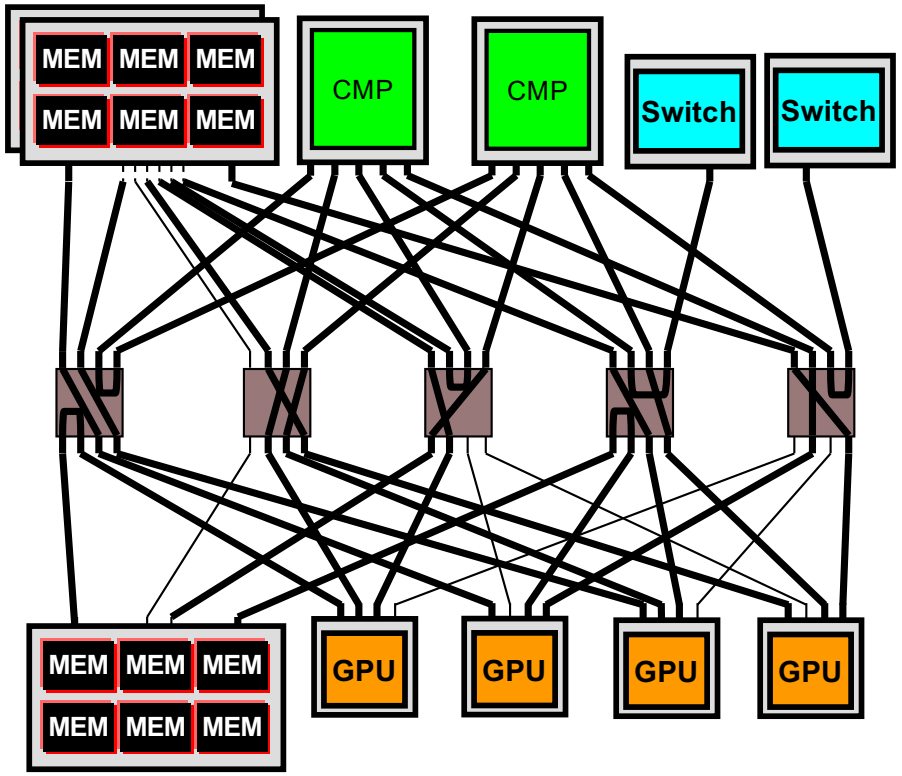
Intra-node bandwidth steering

- **Introduce low-radix optical circuit switches to the OC-MCM topology**
  - 4x4 to 8x8 realizable with today's technology
  - Tens of switches can be collocated on a single chip
- **Slower reconfiguration compared to packet switching**
  - Reconfiguration takes microseconds
  - ***But traffic patterns are persistent for long periods (minutes to hours!)***
- **But transparent for packets**
  - No buffering for point-to-point means Time-of-Flight latencies
  - Extremely energy efficient to reconfigure
  - **Minimize marooned resources**

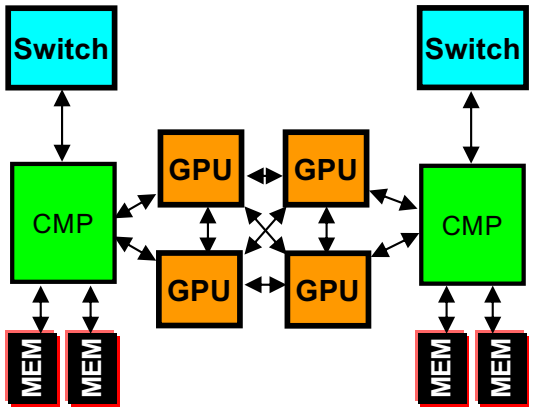


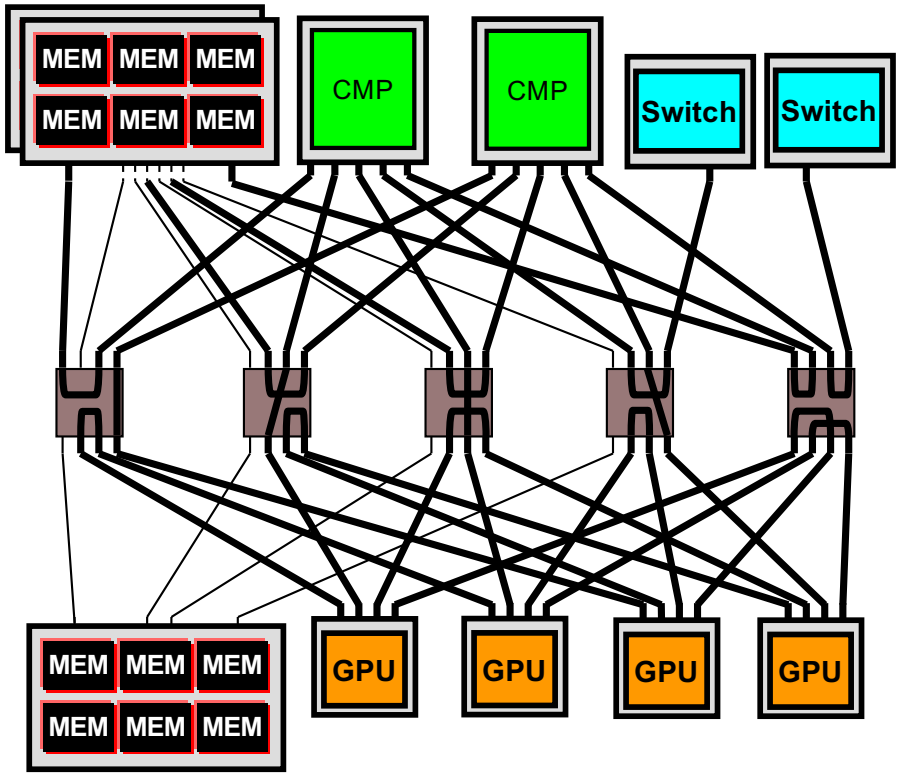
ML : Inference Configuration





ML : Training Configuration

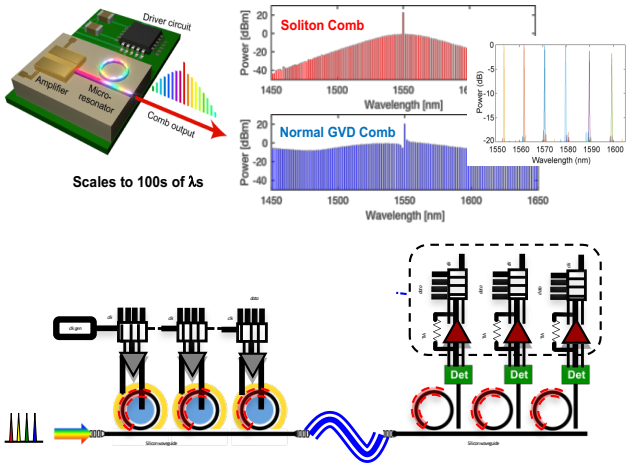




PINE: Photonic Integrated Networked Energy Efficient Datacenters

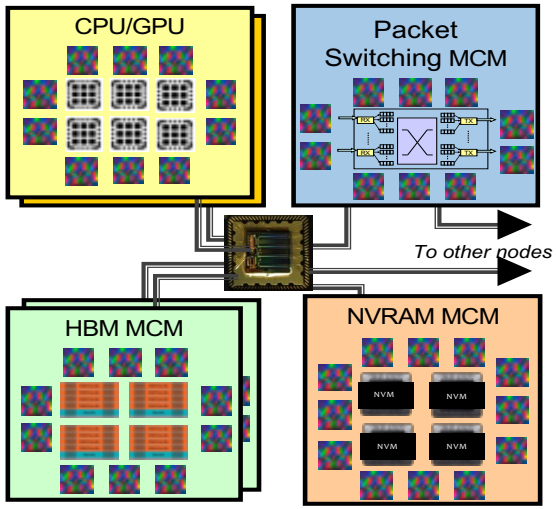
Resource Disaggregation to custom-assemble diverse accelerators for diverse workload requirements

1) Energy-bandwidth optimized optical links

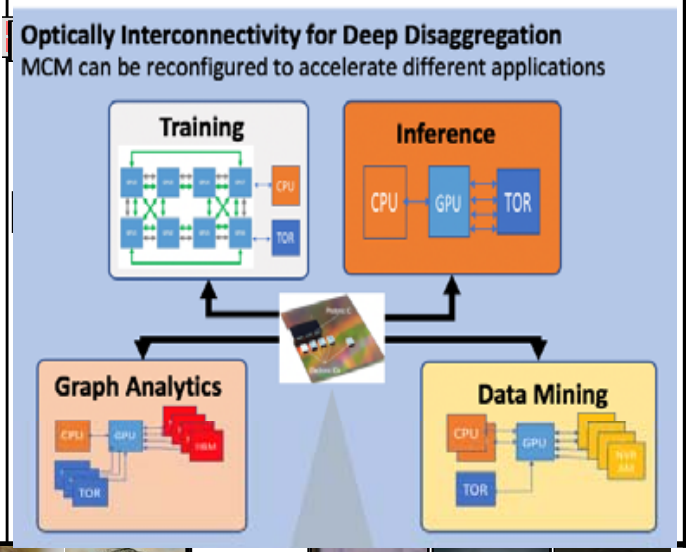


1 Tb/second per fiber

2) Embedded silicon photonics into OC-MCMs



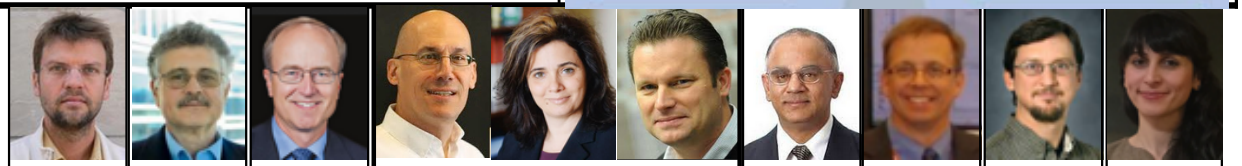
3) Bandwidth steering for Custom Node Connectivity



ENLITENED

arpa·e
CHANGING WHAT'S POSSIBLE

Bergman



Johansson Coolbaugh Bowers Gaeta Lipson Kinget Patel Dennison Shalf Ghobadi

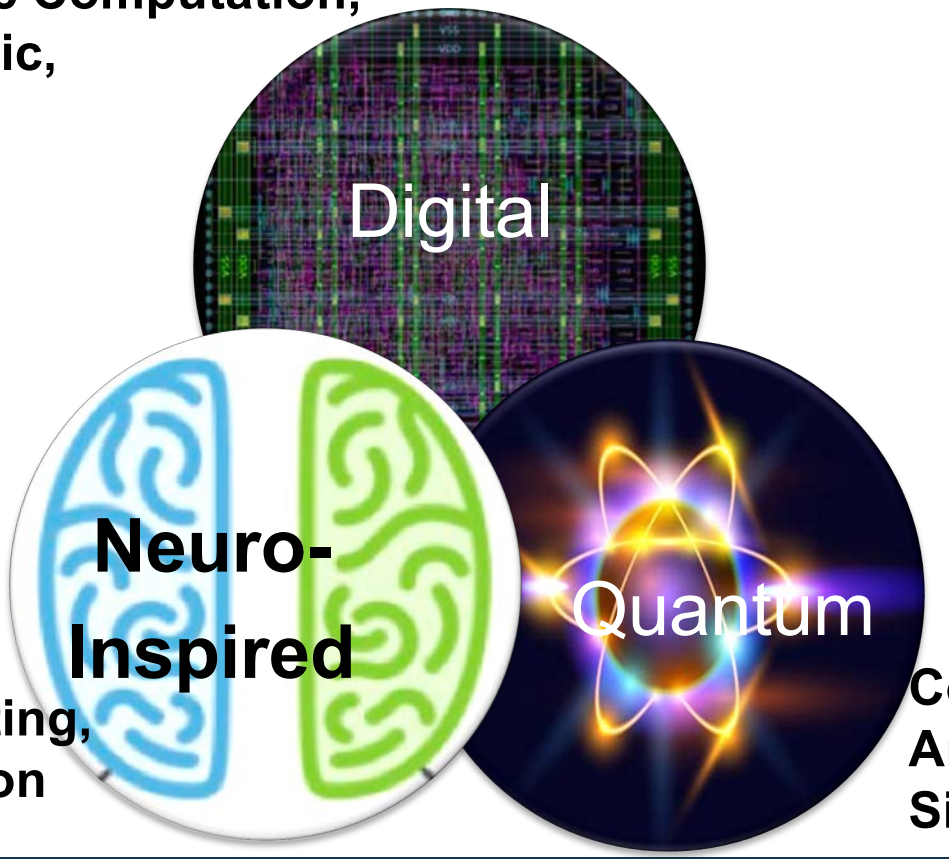
FREEDOM PHOTONICS · SUNY POLYTECHNIC INSTITUTE · UCSB · CISCO · NVIDIA · BERKELEY LAB · Microsoft

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**

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