Specialized Processing for Edge Computing

Bringing the computation to the data

David Donofrio
Tactical Computing Laboratories

ARM Research Summit September 2019

Post Moore Technology Curve

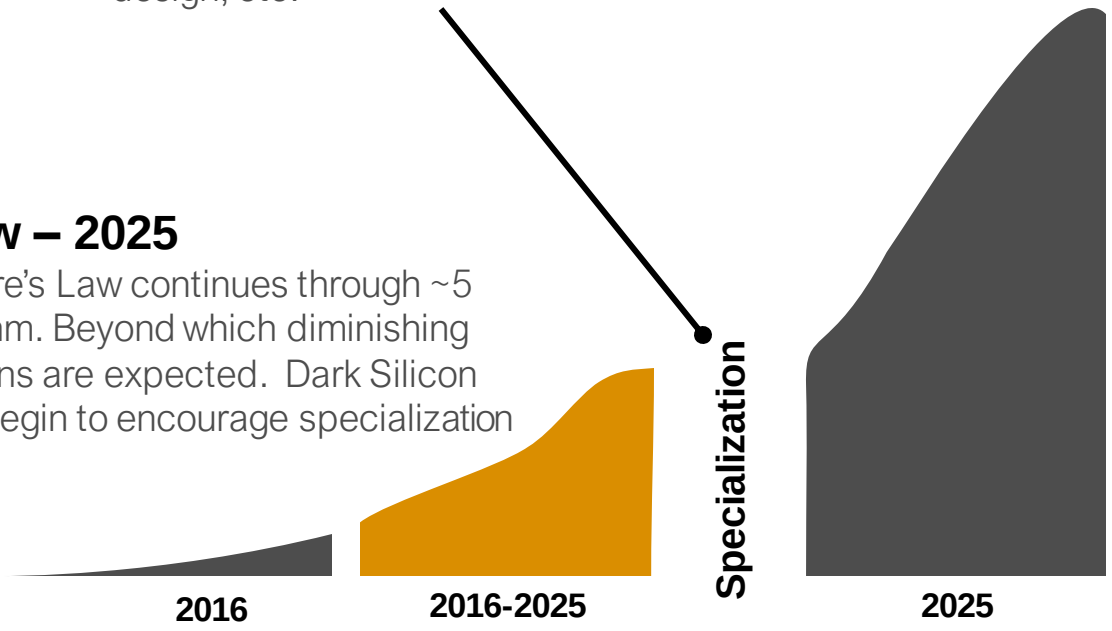
Great *opportunities* exist for innovation through the end of Moore's Law

End of Moore's Law

End of Moore's Law requires a different set of optimizations to continue performance scaling. Opportunities for additional specialization, reconfigurable computing, hardware / software co-design, etc.

Throughout…

Continued increases in parallelism and heterogeneity will require advanced runtimes, programming environments and compiler optimizations in order to take full advantage of these new architectures



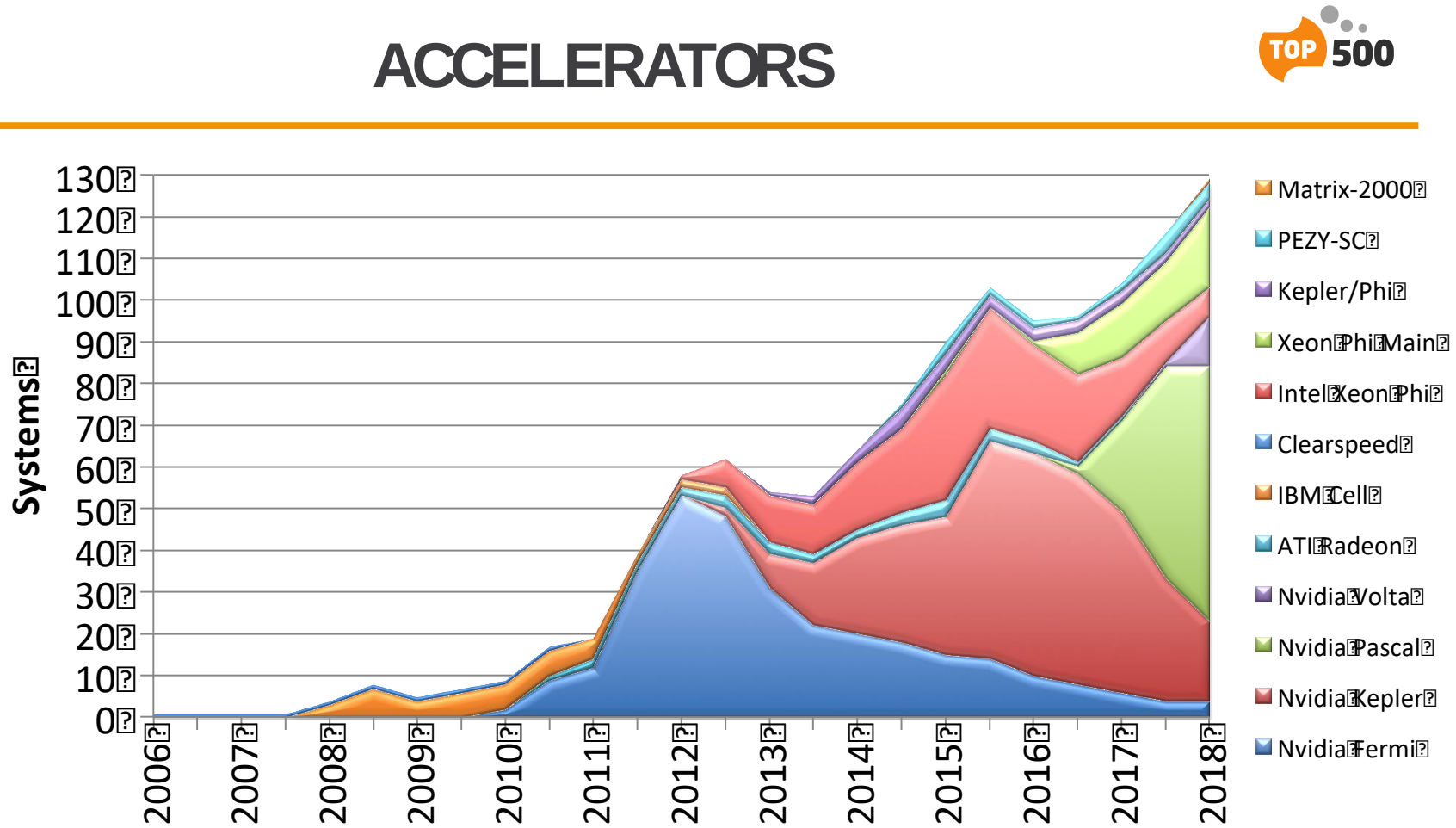
Now – 2025

Moore's Law continues through ~5 or 7nm. Beyond which diminishing returns are expected. Dark Silicon will begin to encourage specialization

Post Moore Scaling

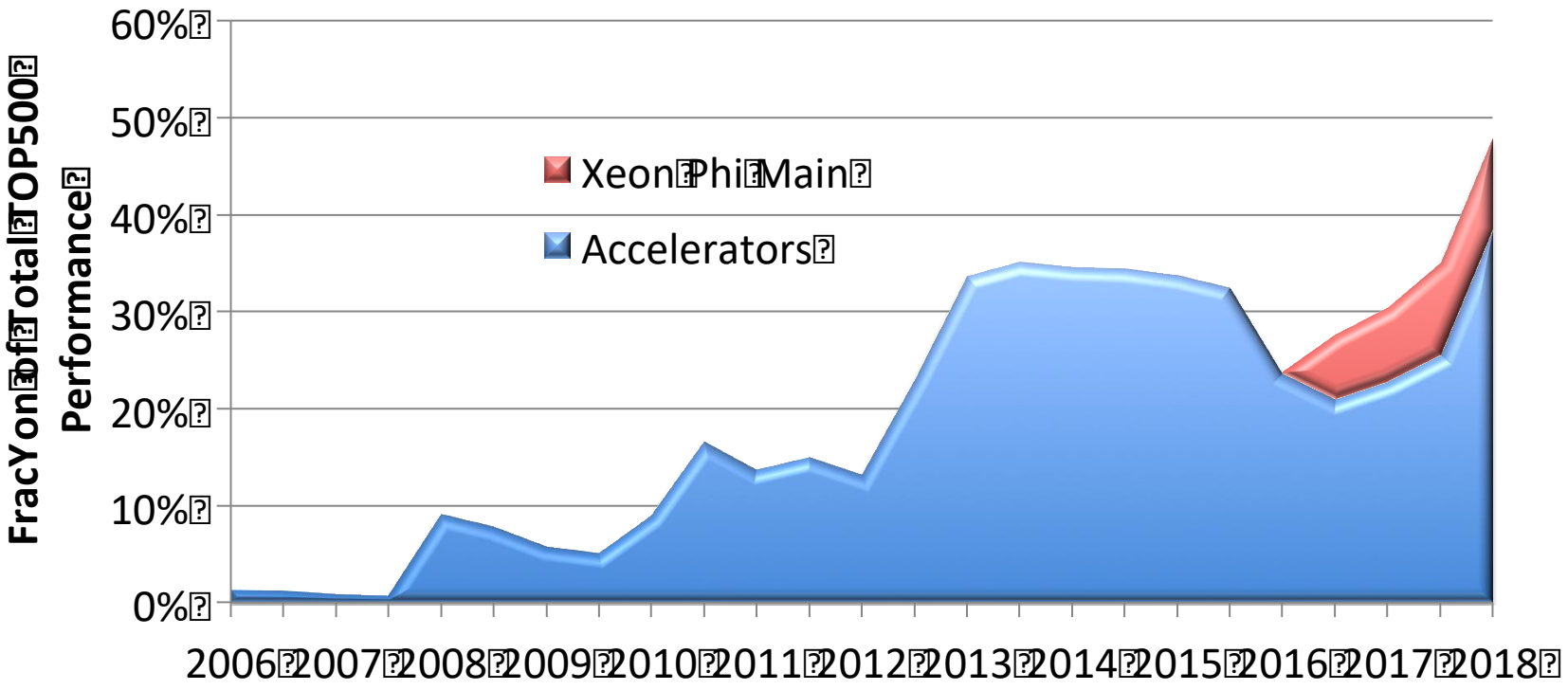
New materials possibly introduced to allow continued process and performance scaling.

Accelerators already driving performance gains



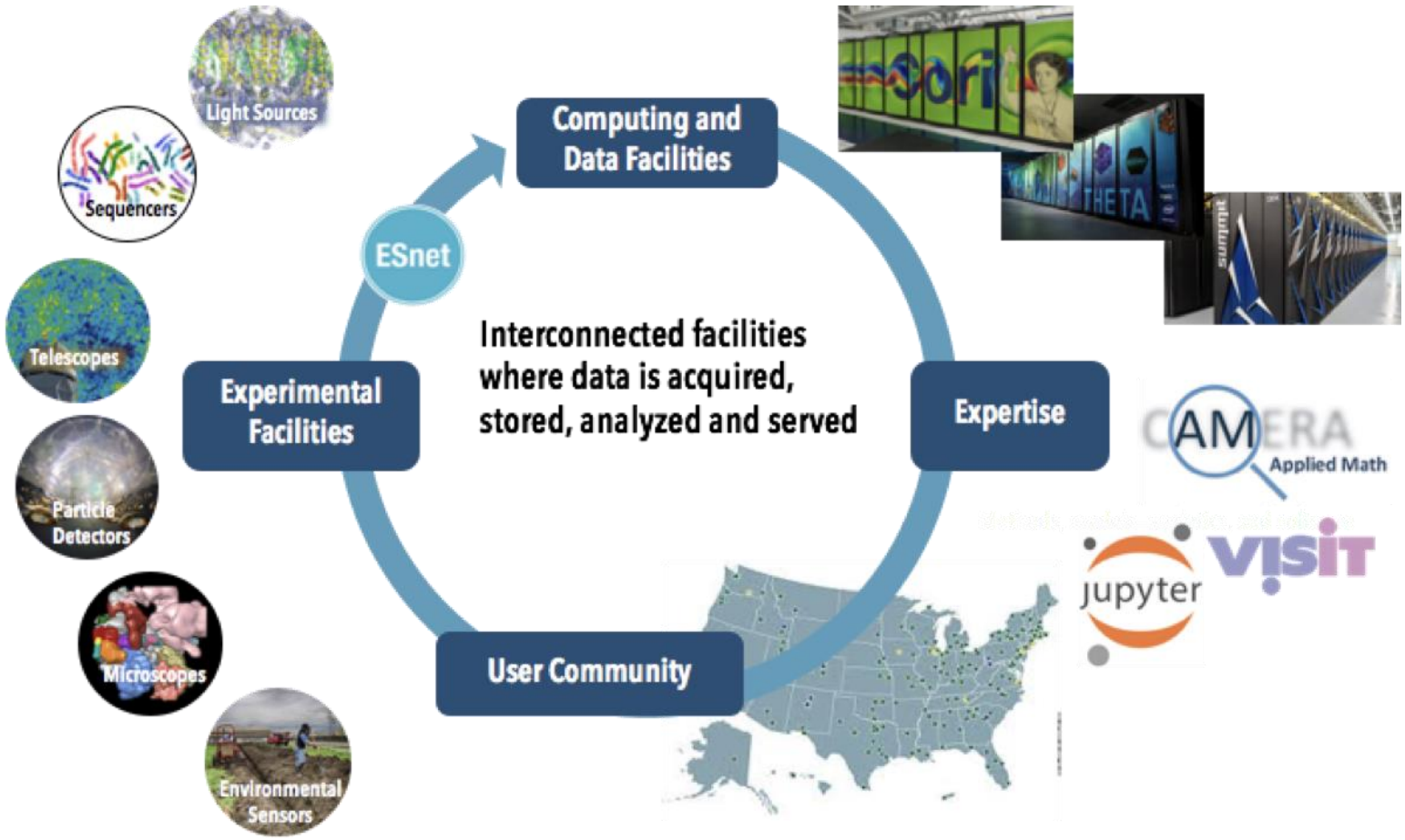
Accelerators already driving performance gains

PERFORMANCE SHARE OF ACCELERATORS



Coupling of Data and Experiments

Opportunities for specialized processing throughout the Superfacility Model



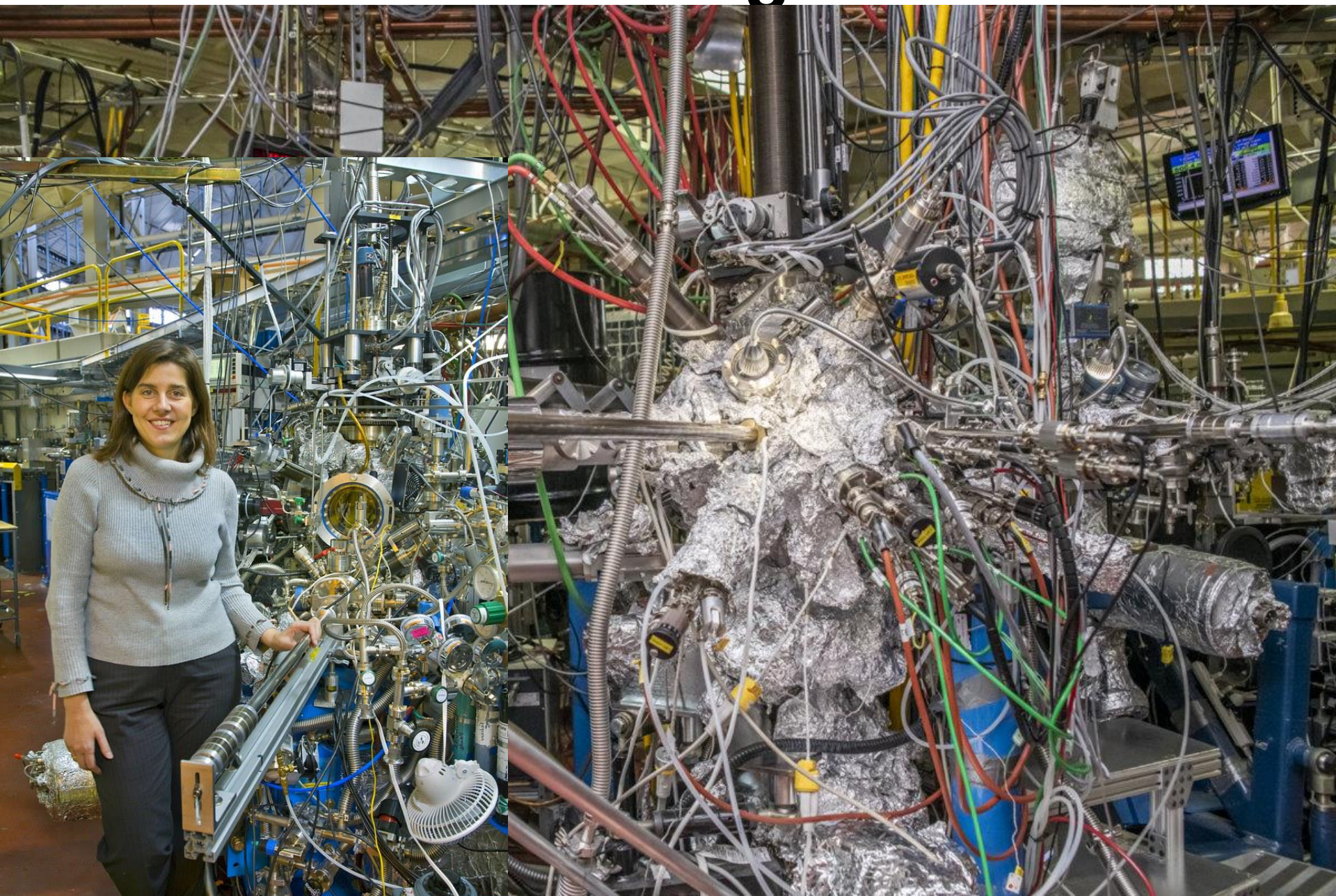
Experiments Each have Unique Needs

Experimental equipment rarely available as COTS

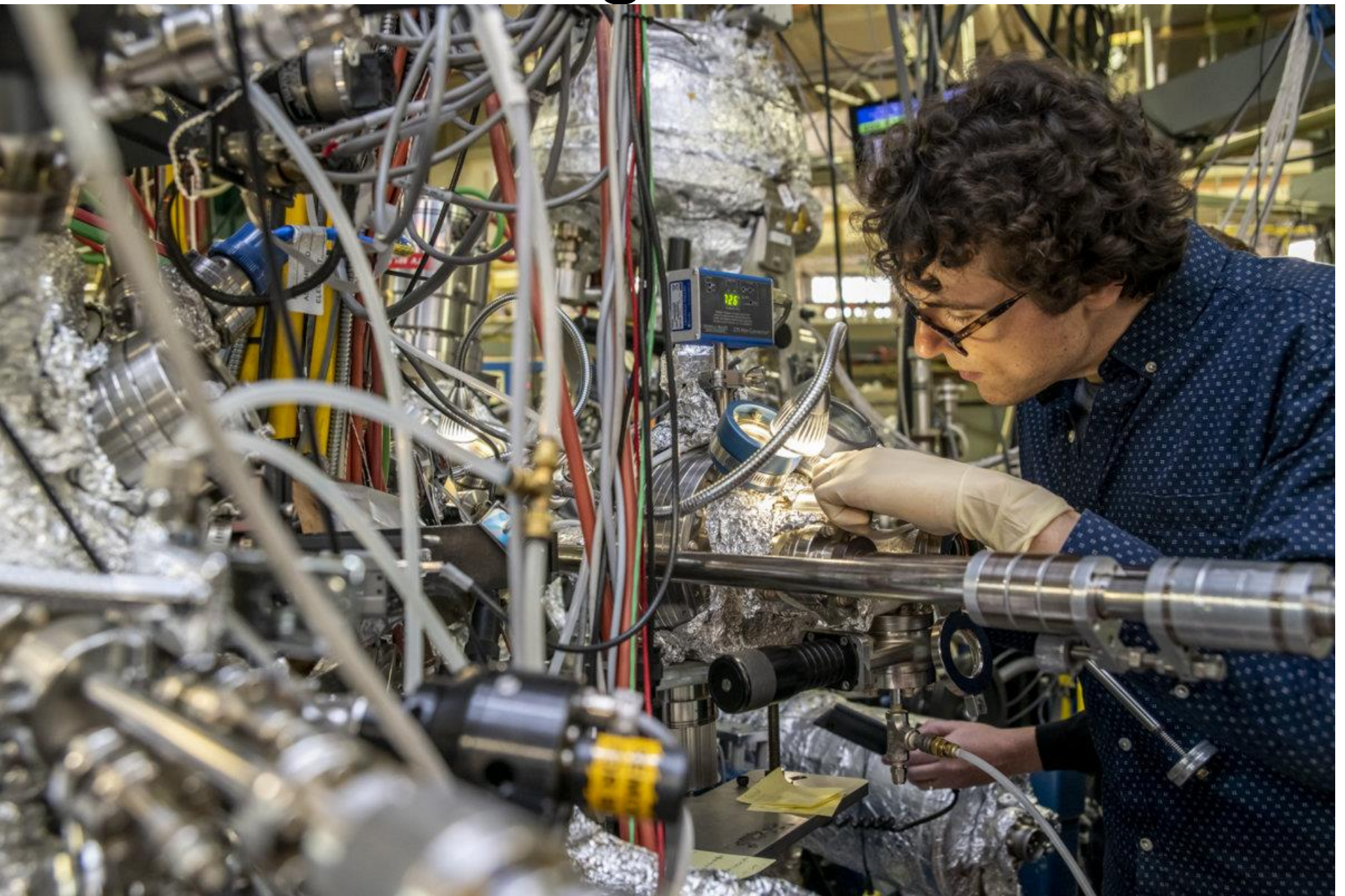
- **Large scale experiments have embraced specialization out of necessity for years**
- **Convergence of HPC and Edge/IoT requirements**
  - Rapid design cycle
  - Low Volume
  - Cost sensitivity
- **Edge devices increasingly have higher performance demands**
- **May lead the way in architectural diversity**

Can the trend towards greater specialization in HPC be applied towards edge computing?

This is What the Edge Looks Like



This is “The Edge”



Bringing the Processing to the Data: Hardware for Data Analysis and Reduction

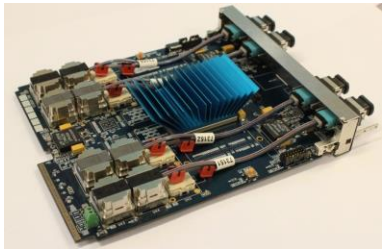
Augment HPC facilities to handle increases in EOS

On Sensor / Field Deployable Processing

Leverage our architectural exploration and design tools to design custom, programmable logic to be integrated on existing sensors or to act as independent, field deployable computing.

Can be used in a facility or act as standalone, low-power field deployed unit

Near Sensor and Real Time Processing



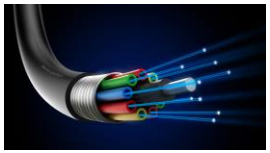
Custom logic FPGAs or ASIC) can be placed near the sensor to analyze and reduce data in real time as it is produced.

The same logic can be used for automated, real time control of instruments

Example: LHC Triggers or control processor for Quantum Processor

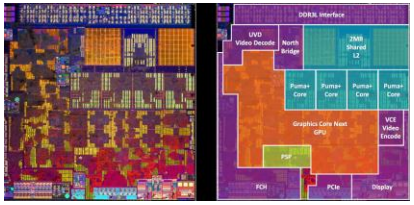
ESnet

Smart HPC Interconnects



Embedding hardware within the network may allow better utilization of existing HPC interconnects

HPC Specialized Accelerators



Specialized accelerators incorporated into HPC systems. Could be integrated onto an SoC or discrete compute elements

Require advances in programming and execution models

Bringing the Processing to the Data: Hardware for Data Analysis and Reduction

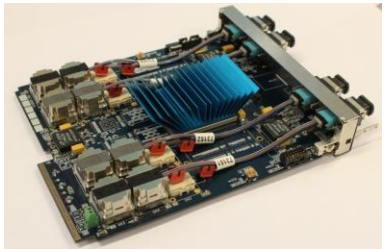
Augment HPC facilities to handle increases in EOS

On Sensor / Field Deployable Processing

Leverage our architectural exploration and design tools to design custom, programmable logic to be integrated on existing sensors or to act as independent, field deployable computing.

Can be used in a facility or act as standalone, low-power field deployed unit

Near Sensor and Real Time Processing



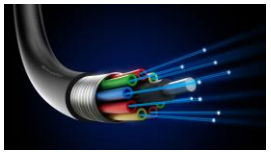
Custom logic FPGAs or ASIC) can be placed near the sensor to analyze and reduce data in real time as it is produced.

The same logic can be used for automated, real time control of instruments

Example: LHC Triggers or control processor for Quantum Processor

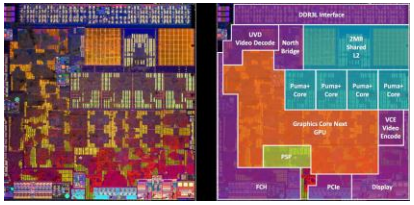
ESnet

Smart HPC Interconnects



Embedding hardware within the network may allow better utilization of existing HPC interconnects

HPC Specialized Accelerators

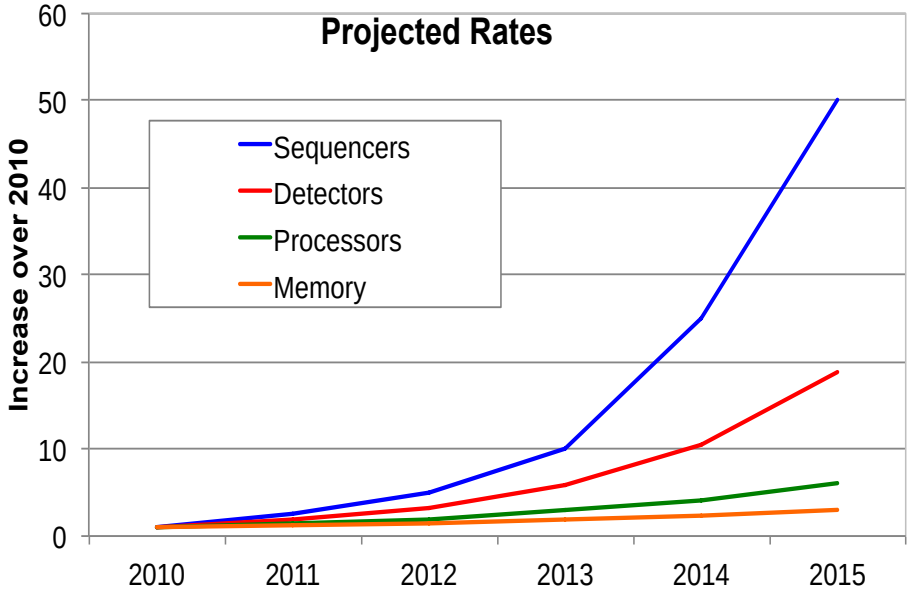


Specialized accelerators incorporated into HPC systems. Could be integrated onto an SoC or discrete compute elements

Require advances in programming and execution models

On-detector processing

Putting hardware specialization to work to augment existing HPC resources

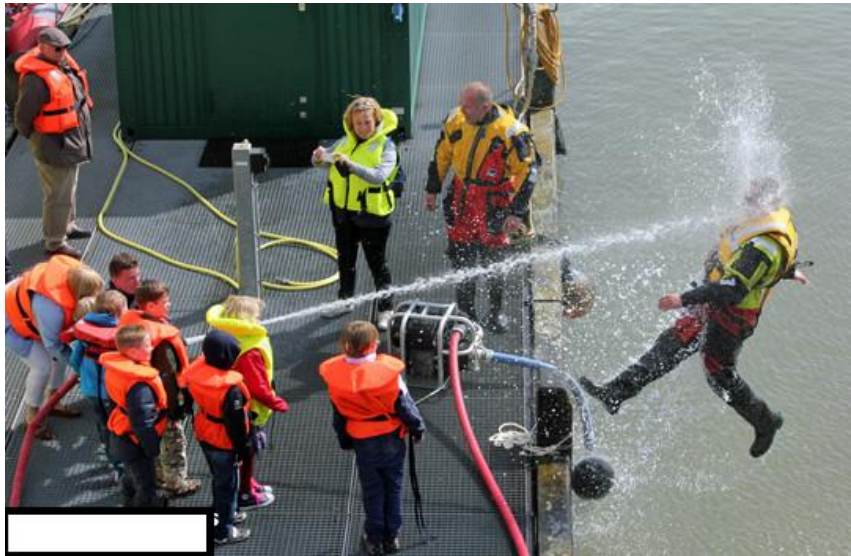


The Problem:

- Future detectors threaten to overwhelm data transfer and computing capabilities w/ data rates exceeding 1 Tb/s
- Data processing experiment driven

Proposed solution:

- Process the data *before it leaves the sensor*
- Application-tailored, programmable processing allows data reduction to occur on-sensor
- Programmability allows data reduction techniques to be tailored to the experiment – even *after* the sensor is built!



FPGAs for Near Sensor Processing

Flexible architectures well suited for streaming data

- **FPGAs well suited to streaming processing**
  - IO Bandwidth > 400Gb/s; 1.5TFlops per device – HBM at 460 GB/s
- **ARM + FPGA Fabric a powerful combination**
  - Programmability w/ custom accelerator solutions
- **Experimentalists often think / work in MATLAB or NumPy**
  - ISA matters less… users want higher abstraction
- **Customization is good… but one-off solutions are less desirable**
  - Are there common building blocks needed across multiple experiments?
- **Need Agile RTL methods to rapidly design and deploy hardware blocks**

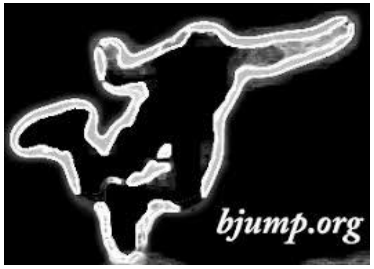
What tools do we have available?

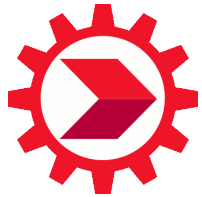
Hardware generation as a methodology

Agile RTL Methods

Increasing diversity in new hardware DSLs

- **Chisel**
- **PyMTL**
- **PyRTL**
- **PyRope**
- **BaseJump RTL**
- **FIRRTL**
- **…**
- **Can new HDLs abstract away the HW?**

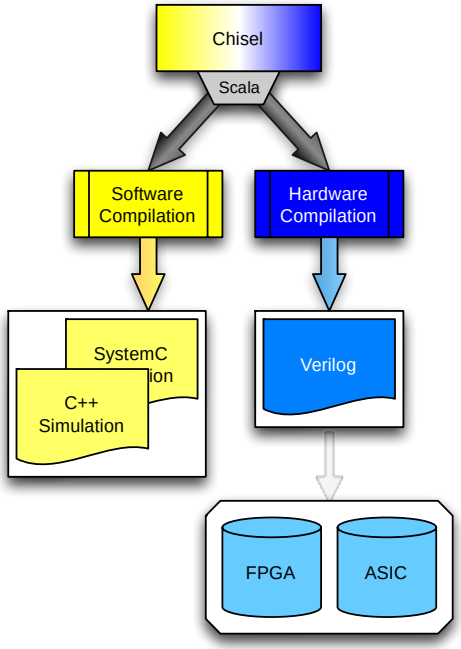




Hardware Generators: *Enabling Technology for Exploring Design Space Together with Close Collaborations with Applied Mathematics*

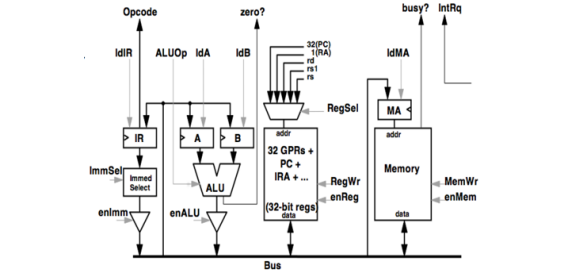
Chisel

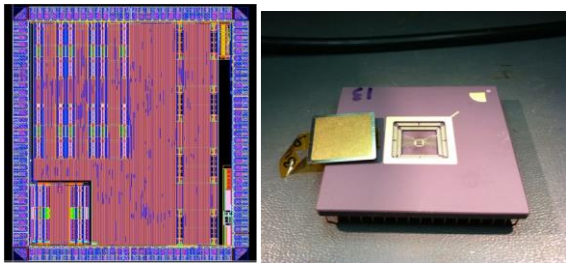
DSL for rapid prototyping of circuits, systems, and arch simulator components



RISC-V

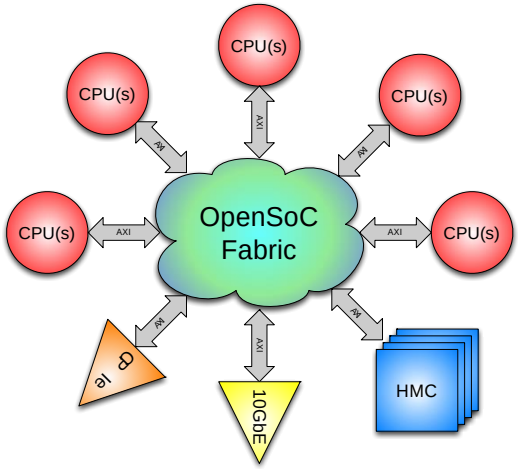
Open Source Extensible ISA/Cores





OpenSOC

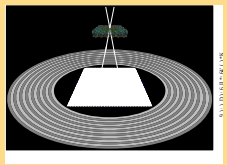
Open Source fabric
To integrate accelerators
And logic into SOC



SuperTools
Superconducting
RISC-V

QUASAR
Quantum
ISA

Multiagency
Architecture
Exploration

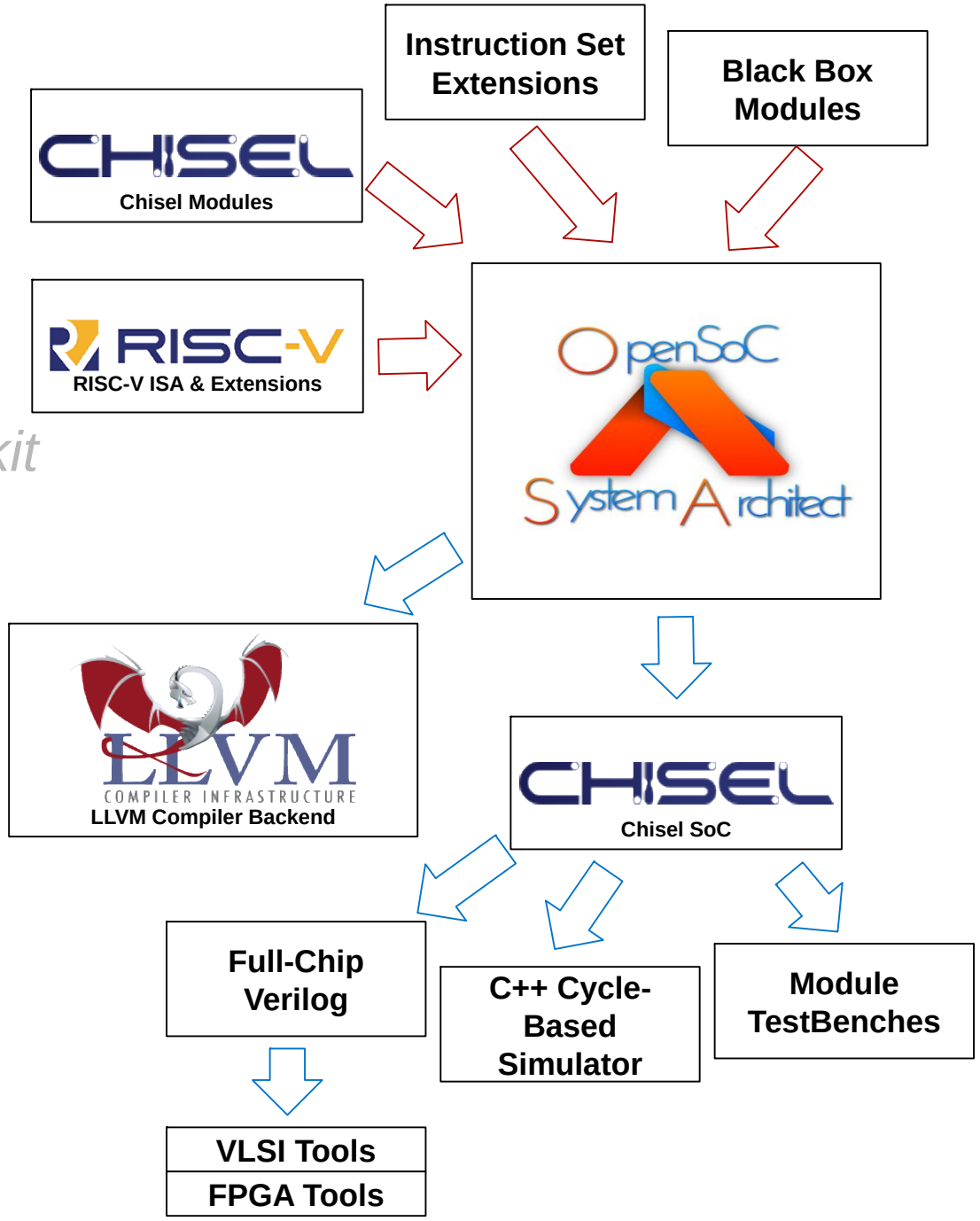


Active
Sensors
(CryoEM)

Existing projects using LBL Open Source Hardware Generators

OpenSoC System Architect

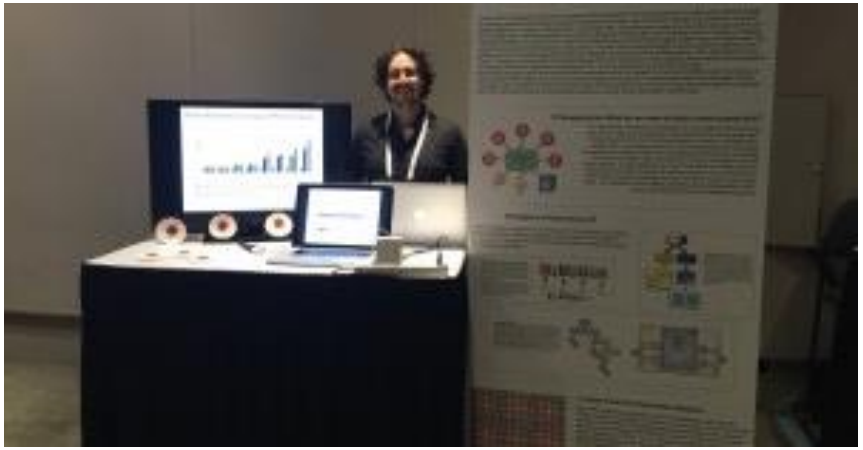
A complete hardware and software development toolkit

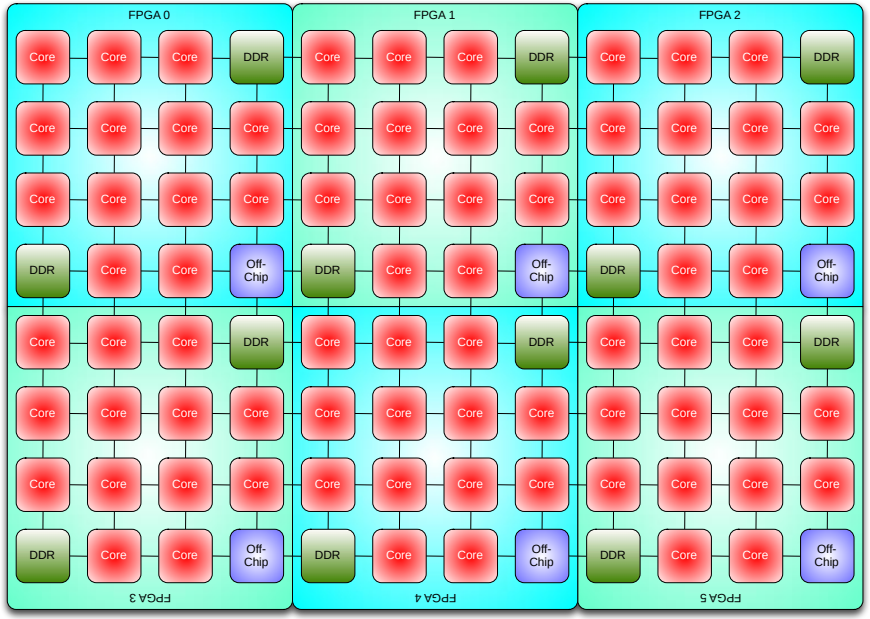


Demo: 96 Core SoC Design for HPC

COMPUTER ARCHITECTURE LABORATORY
EXASCALE DESIGN SPACE EXPLORATION

SC2016 Demo





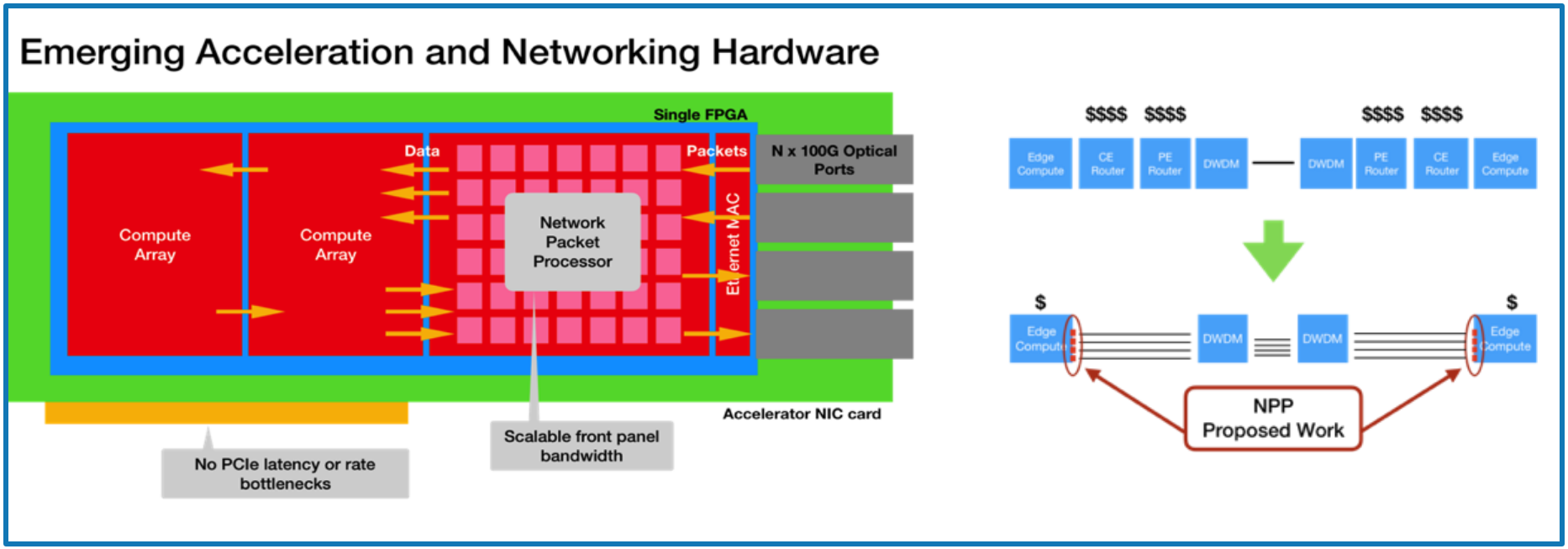
2 people spent 2 months to create

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

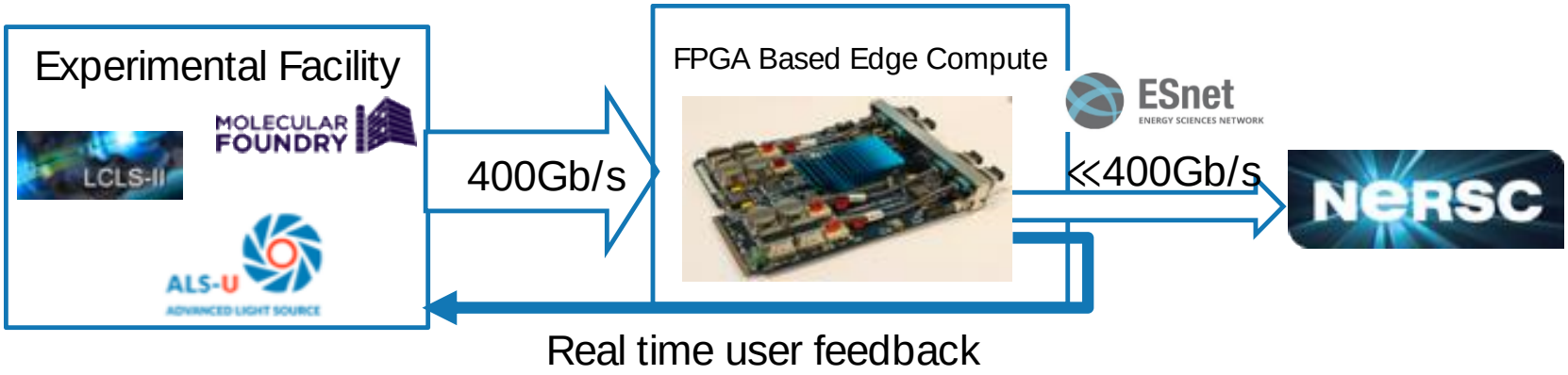
http://www.codexhpc.org/?p=367

In-Transit Data Processing with FPGAs

Projects currently in progress…

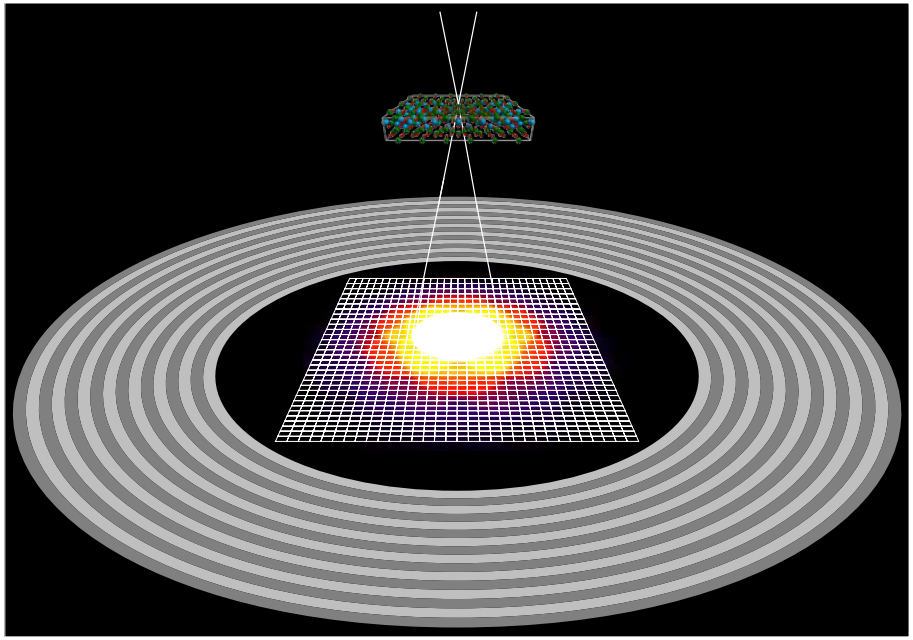


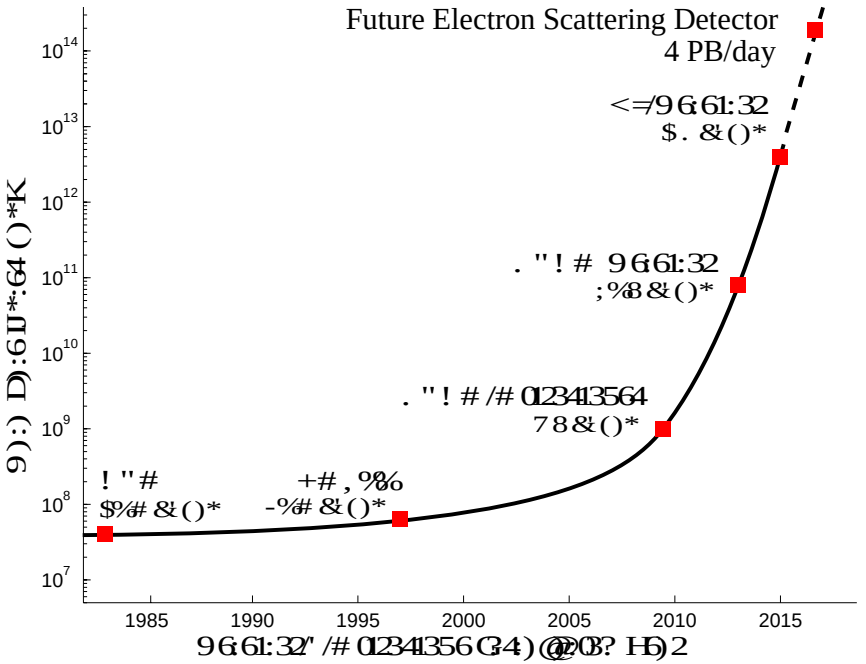
Real-time data reduction for NCEM



National Center for Electron Microscopy (NCEM)

Future Electron Scattering Detector



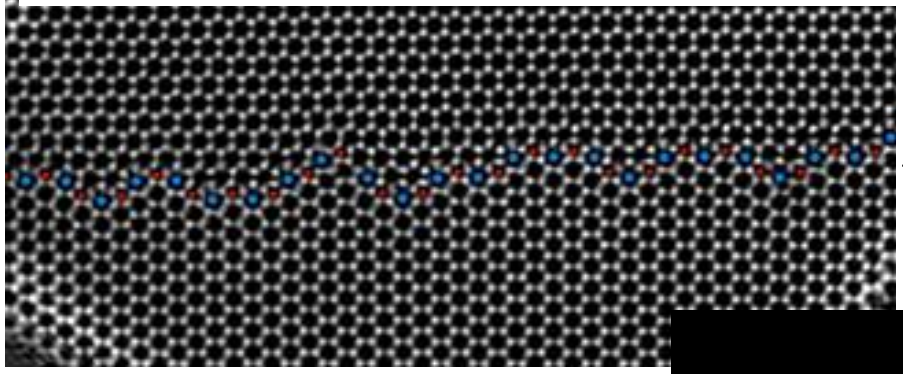


- 100,000 fps pixel detector
  - 576 x 576 x 10 #m
- Segmented silicon HAADF
- Fabricate detectors Q4CY2016

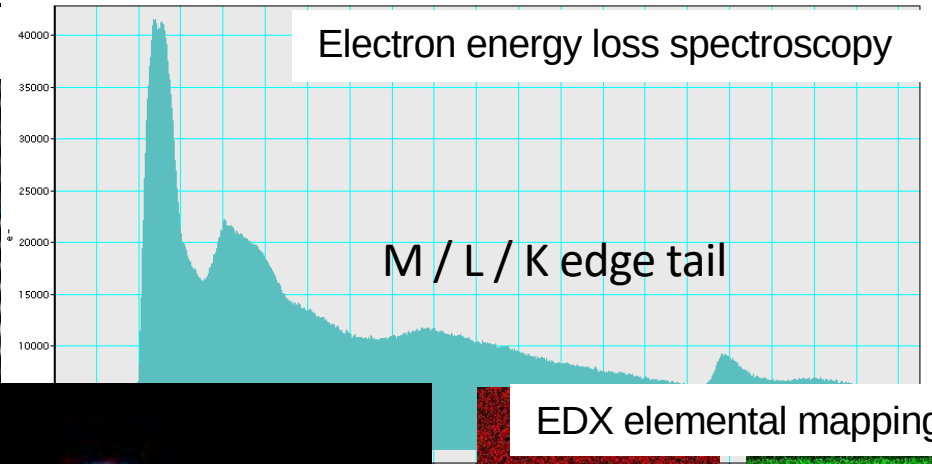
- Dedicated (donated) 400 Gbs link to NERSC
  - Link testing underway
- Stream events to processors on Cori
- Future goal: firmware processing (reduce data rate)

Data Diversity makes National Center for Electron Microscopy (NCEM) perfect use case for pilot study

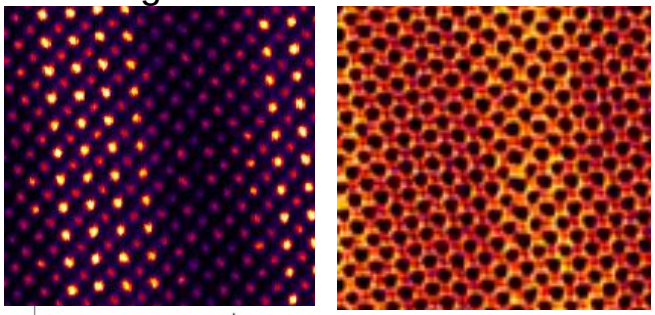
High resolution transmission electron microscopy



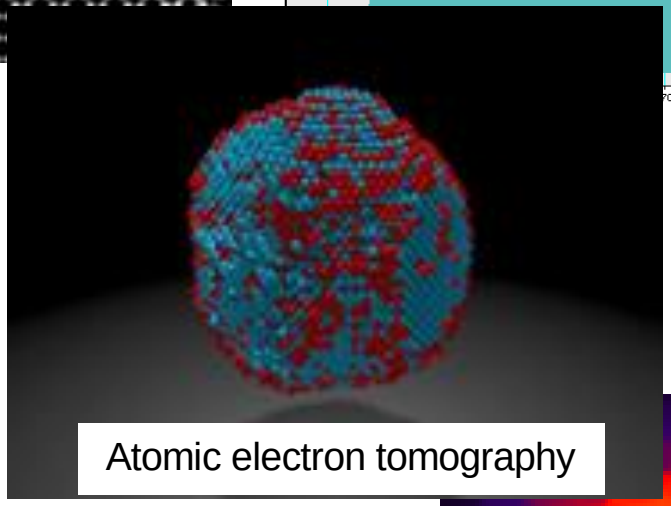
Electron energy loss spectroscopy



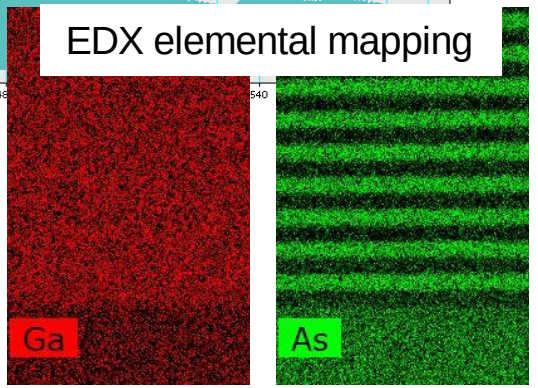
dark field bright field
scanning transmission electron micro.



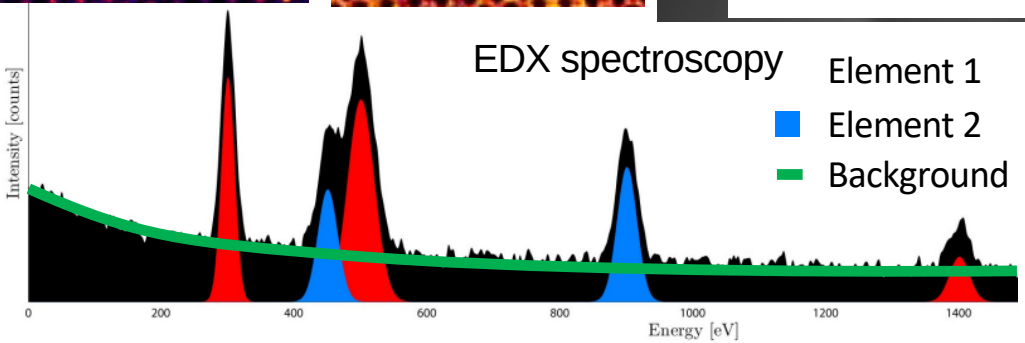
Atomic electron tomography



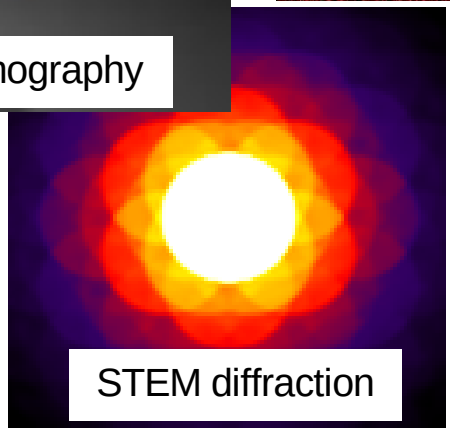
EDX elemental mapping



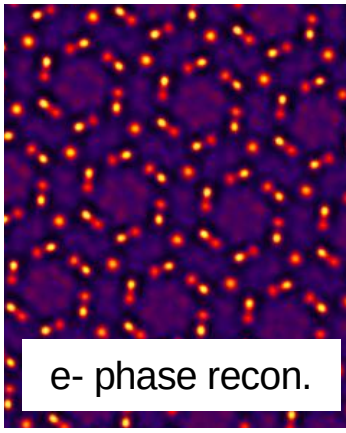
EDX spectroscopy



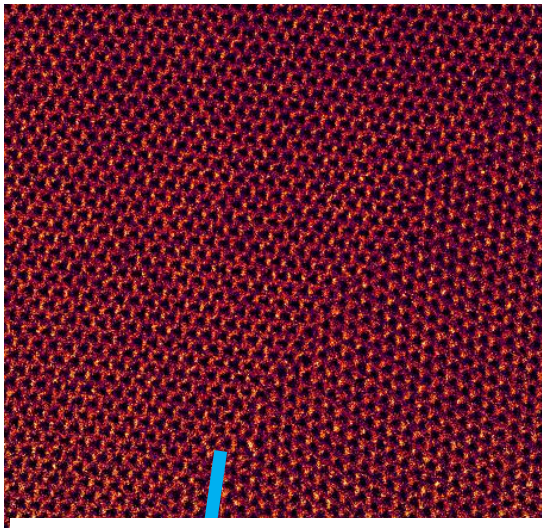
STEM diffraction

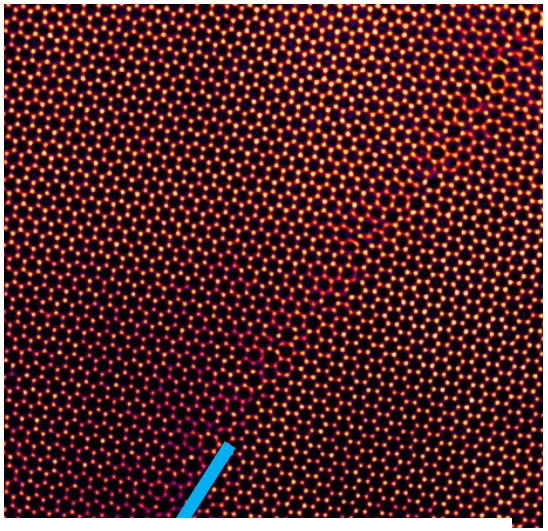


e- phase recon.

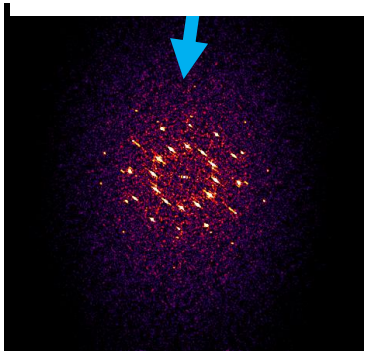


Online data quality metrics and high sparsity of Fourier Transform





Can an algorithm tell the image on the right is better?
Yes! Take Fourier transforms:



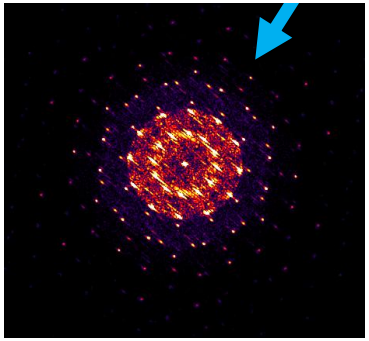
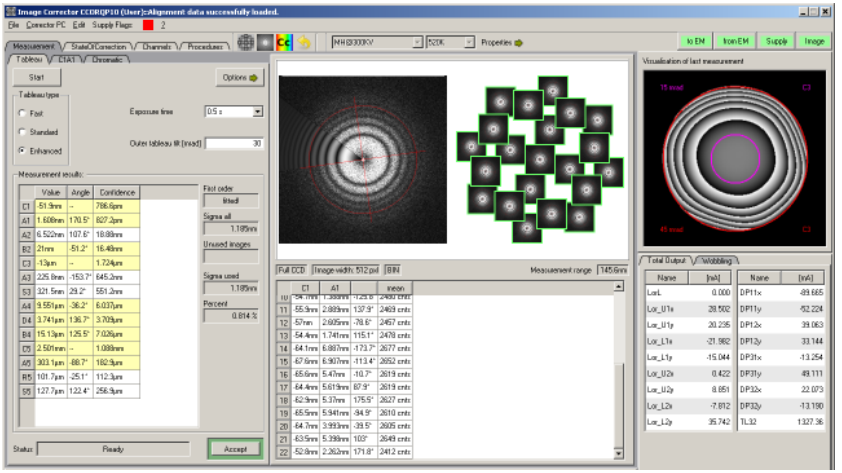


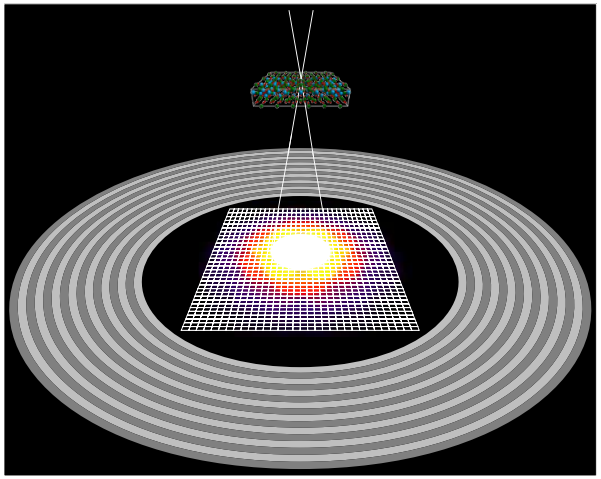
Image on the left has a lot more noise in background, less spots.

Below is an example of how the hardware aberration correction software works – can evaluate defocus, astigmatism, etc. but far too slowly to be used live.



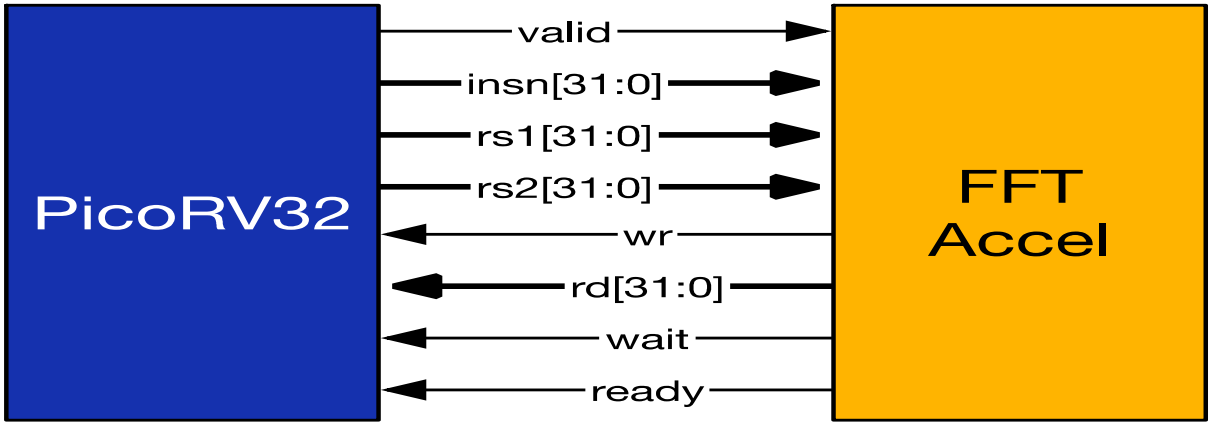
Results for RISC-V FFT Accelerator for CryoEM

Benchmarking FFT Accelerator for image analysis

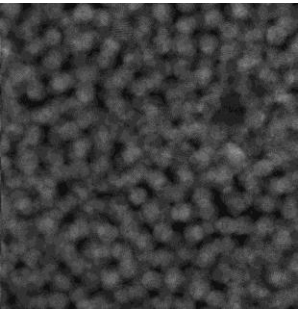


- 100,000 fps pixel detector
  - 576 x 576 x 10 #m
- Segmented silicon HAADF
- Fabricate detectors Q4CY2016

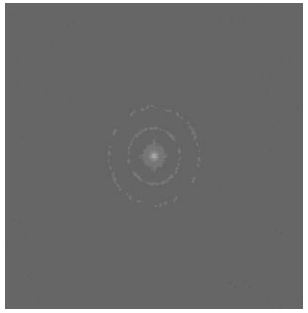
Instruction	opcode [3:2]	Description
fft_config	10b	Configures FFT parameters
fft_status	01b	Reads FFTAccel status registers
fft_start	11b	Starts FFT processing
fft_stop	00b	Stops FFT processing



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**

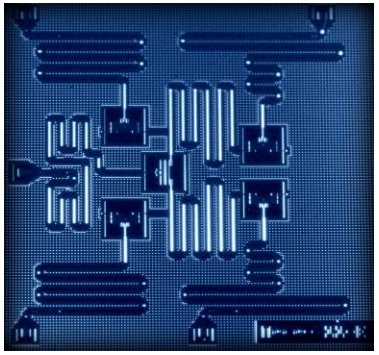
Edge computing is about more than just data…

Quantum Control Processor

How do we control an increasing number of qbits?

$$Quantum\ Computer \neq Quantum\ Processing\ Unit$$

≠

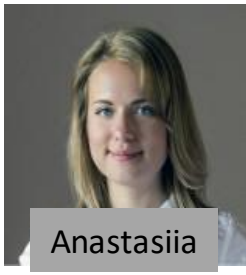


Quantum Computing Infrastructure: Challenges / Opportunities

- Control System / Test Electronics Complexity
- Control response time limits QC performance
- Complexity of Circuits/ Programs to implement quantum algorithms

Quantum Instruction Set Architecture (QUASAR)

PIs: Carter, Siddiqi, DuBois, Advanced Quantum-Enabled Simulation Testbed, July, 2017
Anastasiia Butko and Dubois: QUASAR Developers


Anastasiia

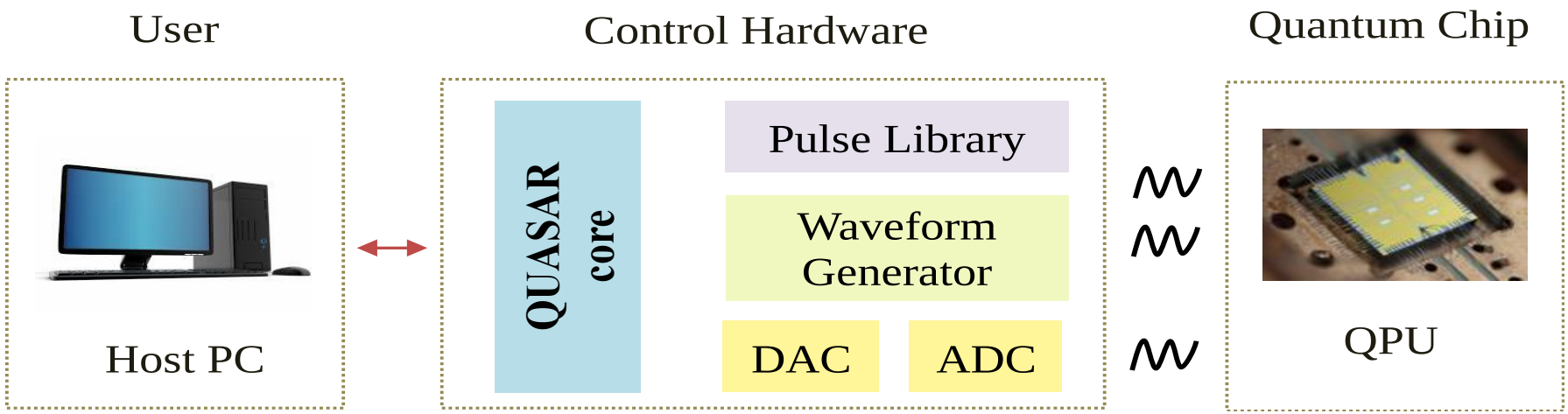
New Concept

We have developed the Quantum instruction set architecture (QUASAR) to provide a compact and efficient method to program complex sequences of operations for the interface to a quantum computer. We have also developed a QUASAR microcontroller core that enables fast in-situ programming and control of qubits for a quantum computer.

Significance and Impact

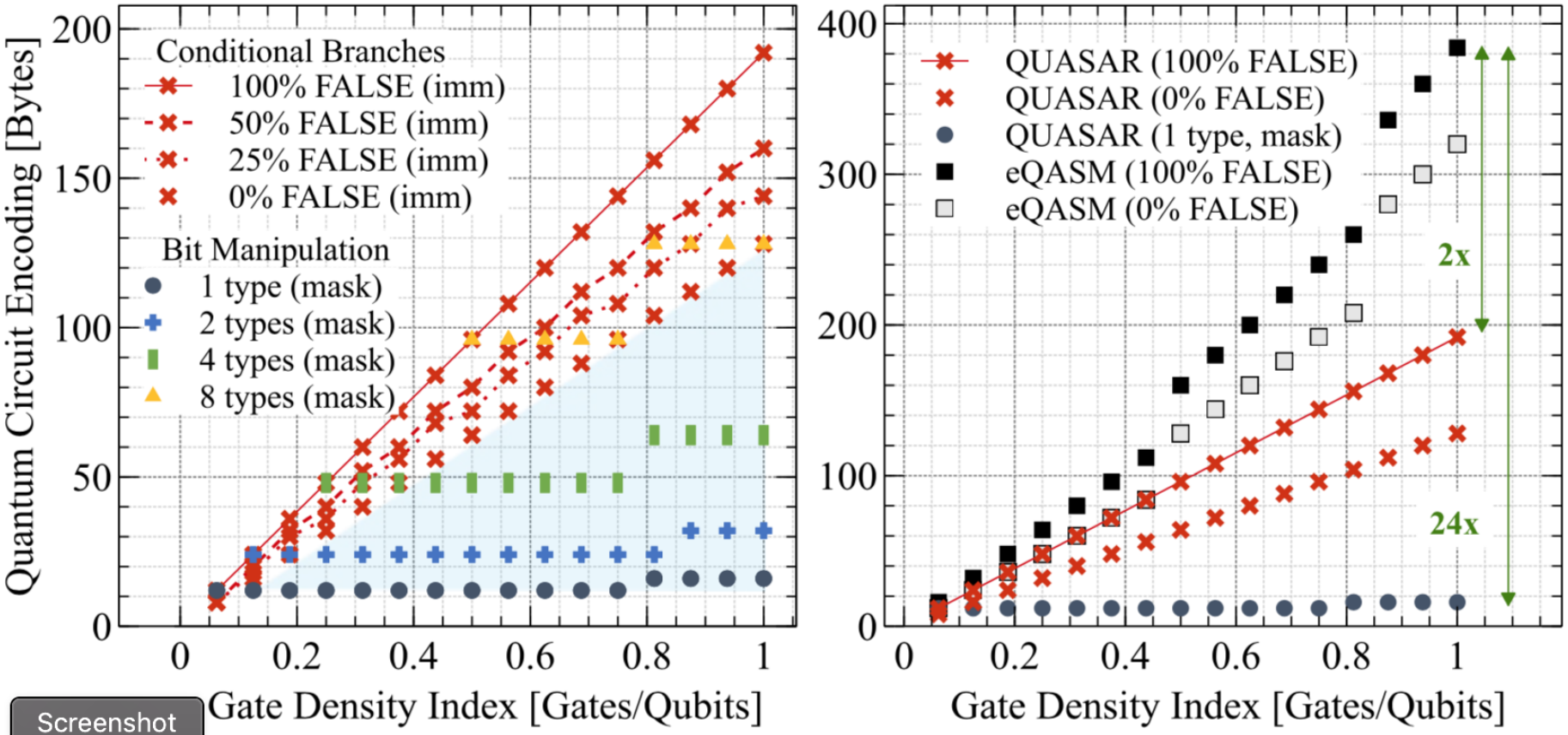
The QUASAR ISA significantly improves quantum computer programmability thereby facilitating their integration into future extremely heterogeneous systems. The QUASAR core (operating in-situ with the qubits using superconducting logic) reduces signal propagation and decision-making delays enabling more efficient usage of limited qubit lifetimes.

Research Details



Software-to-hardware quantum system : (from left to right) user host computer loads quantum program into QUASAR core memory. QUASAR core executes program generating commands to low-level control hardware, i.e. waveform generator, analog-digital converter (ADC), etc. Low-level analog devices communicate with quantum chip via waveforms. Quasar Microcontroller is implemented using the Chisel Hardware Description Language to enable future optimization and technology integration.

QUASAR Performance Results

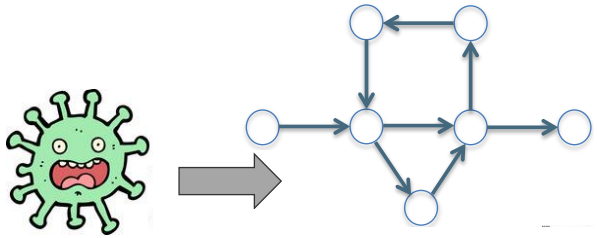


Screenshot

(a) QUASAR: conditional branches vs. bit manipulation (b) QUASAR (this work) vs. eQASM [10]

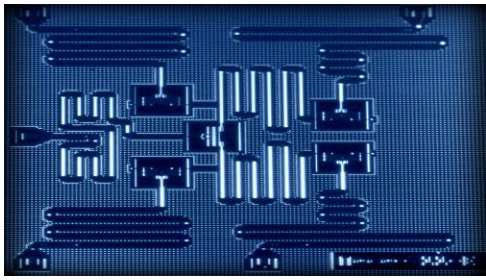
More opportunities for specialization

Custom computing could be applied and incorporated in many areas



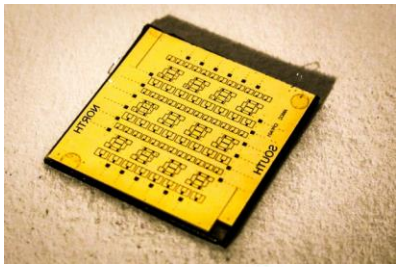
JGI – Accelerating sequence alignment

- Traditional processors not optimal for bioinformatics codes
- FPGAs provide potential solution



Quantum – Control Processor

- Future Quantum accelerators limited by not just number of qubits
- Leverage our work with superconducting logic design tools



LSST – Transient detection

- Power constrained on-site, custom, low-power computing could reduce need for lossy compression or high-bandwidth, long-haul networks

Bringing the Processing to the Data: Hardware for Data Analysis and Reduction

Augment HPC facilities to handle increases in EOS

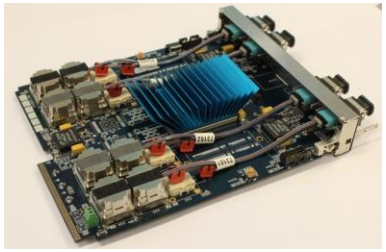
On Sensor / Field Deployable Processing

Leverage our architectural exploration and design tools to design custom, programmable logic to be integrated on existing sensors or to act as independent, field deployable computing.

Can be used in a facility or act as standalone, low-power field deployed unit

Near Sensor and Real Time Processing

ESnet

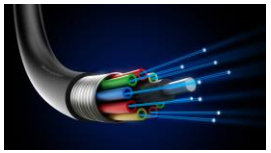


Custom logic FPGAs or ASIC) can be placed near the sensor to analyze and reduce data in real time as it is produced.

The same logic can be used for automated, real time control of instruments

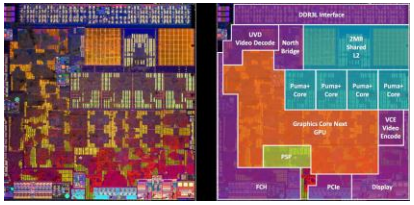
Example: LHC Triggers or control processor for Quantum Processor

Smart HPC Interconnects



Embedding hardware within the network may allow better utilization of existing HPC interconnects

HPC Specialized Accelerators



Specialized accelerators incorporated into HPC systems. Could be integrated onto an SoC or discrete compute elements

Require advances in programming and execution models

Questions?