

Flash Talks – Session 1

Energy-Efficient Computing for Science Workshop

September 9-12, 2024

Bethesda, MD



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Opportunity and Potential Impact

- Visualization is key to gaining insights from large data but can be energy-inefficient.
- Main idea: Use uncertainty visualization as a mechanism to achieve maximal energy efficiency and accuracy.
- Impact: Uncertainty-informed visualization will help scientists to reliably analyze large data with minimal energy consumption.

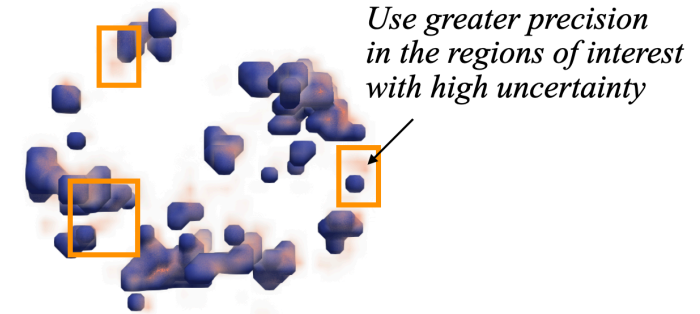
State of the Art and Challenges

- Data approximations: ↑ energy efficiency, ↓ vis accuracy.
- Challenge: Without knowledge of uncertainty, a tradeoff between energy and precision cannot be understood.
- Understanding how data approximations (low precision, dimensionality reduction, interpolation) affect uncertainty and energy will need theoretical research, which is currently lacking.

Supernova dataset



Original data visualization
(Energy = 0.39 watt-hours)



Reduced data with spatial
uncertainty visualization in red
(Energy = 0.006 watt-hours)

64X more energy-efficient
holistic view

Execution and Timeline

- Research thrusts: (1) How large data can be best approximated while capturing uncertainty (2) How uncertainty can be leveraged for energy efficiency.
- Barriers: Developing energy-efficient uncertainty vis algorithms will need rigorous statistical derivations and deriving novel visual mappings.
- Benchmarks: Original large data (e.g., see image above), alternative data approximations and uncertainty vis.

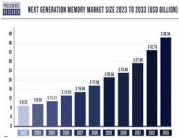
Opportunity and Potential Impact

Proposal: Develop and integrate CryoASIC technology in various emerging technology (Quantum computing, HPC, HEP, Space exploration)



Why Now?

Increasing demand for energy-efficient computing driven by exponential growth in data processing



Breakthrough?



- Enable energy savings,
- increase computational power,
- reduce heat dissipation,
- higher clock speed
- smaller area

Execution and Timeline

Key Step: → Material research → Device Fabrication → System Integration → Application and adaptation

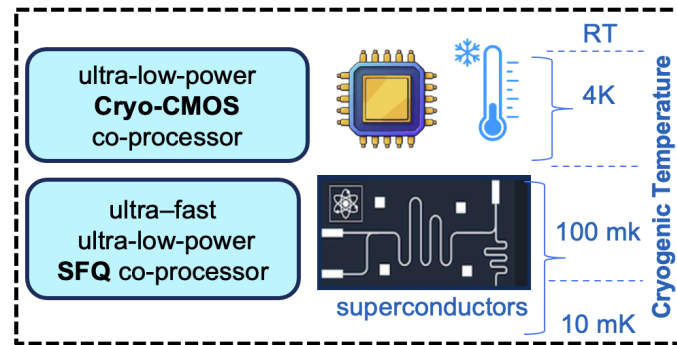
From TODAY:

5 years : Focus on Performance metrics and feasibility

10 years : Integration into Specialized System (e.g. QC, HPC, HEP)

15 years : Widespread adaptation across various industries

ASIC (Application Specific IC's) are custom-designed chips for specific functions.



- [1] Cryo-CMOS controller at 4K → achieved 80 mW power, 20% improved nominal voltage, ~10x better noise figure
- [2] SFQ-based Processing unit for Data Centers → 200% power saving
- [3] C-SQUID for TES → achieved 4.5nW per pixel at 100K at 100 GHz

State of the Art and challenges

Shortcomings: Current approaches primarily rely on **scaling down transistor sizes** and improving **architectural efficiency**. These methods are **reaching their limits**, and further improvements are becoming increasingly **challenging due to physical constraints**.

Why lower temperature helps ?

$S_{TH} \sim kT/q$

Temperature goes down
↓
Subthreshold goes down
↓
Voltage can be scaled
↓
gate oxide scaled
↓
gate length scaled

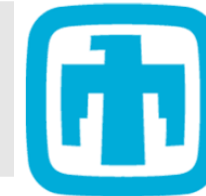
Technical hurdles: Developing CryoASICs requires overcoming technical challenges such as **materials compatibility, thermal management, thermal cycling, and device reliability**.

Integration: Requires careful consideration of **compatibility and performance trade-offs**.

Resources Needed:

- * **Research Funding**
- * **Infrastructure** → Access to clean room, cryogenic facilities, specialized equipment.
- * **Talents** → Develop workforce, team of skilled researchers, engineers and scientist

[1] Bishnu Patra, et al. Cryo-cmos circuits and systems for quantum computing applications. IEEE Journal of Solid-State Circuits, 53(1):309–321, 2017.
[2] Anna Herr and Quentin Herr. A data center in a shoebox: Imec's plan to use superconductors to shrink computers. IEEE Spectrum, 61(6):37–41, 2024.
[3] Steven W Leman, et. al Integrated superconducting transition-edge-sensor energy readout (ister). IEEE TAS, 33(5):1–7, 2023.



Opportunity and Proposed Research Direction

- 1. Low and mixed precision.
 - Do parts of the computations in lower precision?
- 2. Denser numerical models, higher arithmetic intensity
 - Revisit boundary elements, MFS, RBF, etc.
- 3. Krylov methods: inexact Krylov, inexact inner product
Faster linear (KKT) system and optimization solves

Execution and Timeline

- First show proof-of-concept on current hardware
- Thrusts 1-3 can be pursued in parallel by different teams
- Timeline: ~5 years for basic research, then ~5 years co-design, followed by ~5 years of refactoring software
- Risks: Lack of convergence for ill-conditioned systems

State of the Art and Challenges

- 1. Typically double precision all the way
- 2. Sparse FEM is currently popular, very flexible, but low arithmetic intensity
- 3. Standard Krylov methods are commonly used but require high precision/accuracy
- Challenge: Adapt methods and prove convergence

Potential Impact

- (1) is the lower hanging fruit, but will likely have modest impact
- (2) and (3) are harder and would require substantial changes in algorithms/libraries/applications, but potentially higher pay-off
- Could develop and run benchmarks on existing hardware

Toward Cross-Layer Energy Optimizations in AI Systems

Mosharaf Chowdhury, University of Michigan (w/ Jae-Won Chung & Nishil Talati)



Opportunity and Proposed Research Direction

- Energy is the gating factor for AI, and layer-by-layer optimizations are close to their limits that cannot be breached without additional information and control.
- We propose a narrow waist for energy optimization to enable auto-tuned cross-layer optimizations across applications (APP), algorithms (ALGO), software (SW), and hardware (HW) without tightly coupling them.

Execution and Timeline

- Introducing interfaces to expose energy–performance Pareto frontiers in a top-down fashion.
- Developing optimization algos that utilize such interfaces.
- Resistance and inertia from HW vendors is a key concern.
- For success, we need large-scale heterogeneous testbeds and detailed performance models/sims of HW.

State of the Art and Challenges

- Energy optimizations are siloed within layers without any comprehensive model. Each layer makes independent optimizations that can even contradict each other. While some cross-layer solutions exist, they are limited to narrow deployment scenarios.
- Information/control flow across layers w/o losing generality requires a broad cross-community approach.

Potential Impact

- Non-experts will be able to deploy large, general-purpose AI systems that are 2-10X more energy-efficient than what experts can do today.
- Adoption will be facilitated with high quality open-source AI energy stack, e.g., Zeus (<https://ml.energy/zeus>).
- We will measure success by the extent of overall energy efficiency gains across the stack and in each layer.

Hilary Egan (NREL) with Juliane Mueller (NREL) and Dejan Milojicic (HPE)

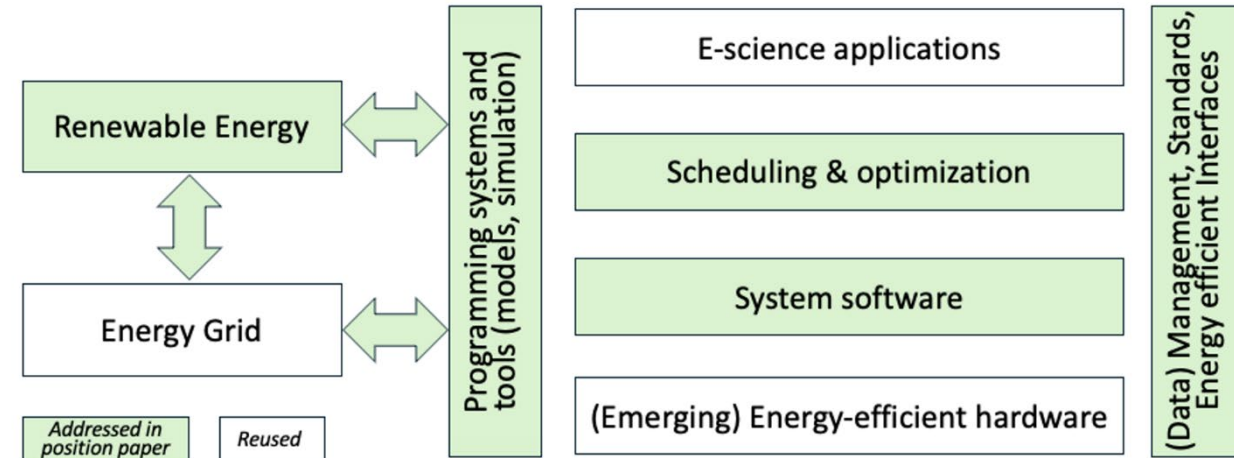
Optimizing Systems for Energy-Efficient Scientific Computing

Opportunity and Potential Impact

- Data centers are predicted to consume 9% of the nation's electricity by 2030
- Renewable energy sources are penetrating grids now and adoption is only accelerating, requiring proactive solutions to accommodate powering data centers
- **We must use renewable energy to power both newly built and existing data centers**

State of the Art and Challenges

- Renewables are intermittent/less reliable than conventional power supply and not integrated at a wide-scale
- Varying DER availability combined with varying load creates a potentially very large-scale stochastic optimization problem complicated by inexistent end-to-end view and agreed upon metrics
- Required infrastructure, grid integration upgrades, and installation costs present a high upfront cost



Execution and Timeline

- Agreed-upon *standards and metrics* to be able to assess the utility of investment decisions now and in the future
- *Scheduling* must be improved vertically across system stacks and horizontally across multiple data centers
- New *energy-efficient tools for modeling, simulation, and federated digital twins*, will be required to enable the assessment of energy efficiency and sustainability of scaling up experiments.

Rafael Ferreira da Silva, Oak Ridge National Lab

Eco-Driven AI-HPC: Optimizing Energy Efficiency in Distributed Scientific Workflows

Opportunity and Proposed Research Direction

- Optimize energy efficiency in distributed scientific **workflows across Edge-Cloud-HPC continuum**
- Opportunity: Increasing data volumes, AI integration, and technological advancements

Execution and Timeline

- Key steps: Develop energy-aware **scheduling algorithms, high-fidelity simulators, and data transfer optimization**
- 5-10-15 year outlook: Implement initial optimizations, **refine models, achieve full integration across facilities**
- Barriers: Coordination across facilities, technology integration, **data management complexity**
- Resources needed: Multi-facility testbeds, **comprehensive energy consumption data, advanced simulation tools**

State of the Art and Challenges

- Current approaches lack energy-aware scheduling for geographically dispersed resources
- Challenges: I/O-intensive workflows, **heterogeneous resources, balancing performance and energy use**

Potential Impact

- Aim to **reduce energy consumption by 10-20%** in exascale computing
- Success metrics: Energy reduction per unit of computing, **increased scientific output per unit of energy**

Patricia Gonzalez-Guerrero, Lawrence Berkeley National Laboratory

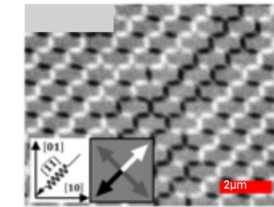
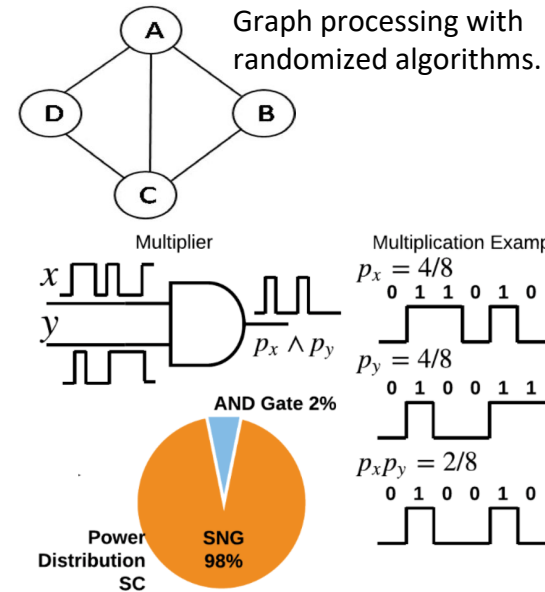
Hardware-Software Codesign for Energy-Efficient Randomized Algorithms

Opportunity and Potential Impact

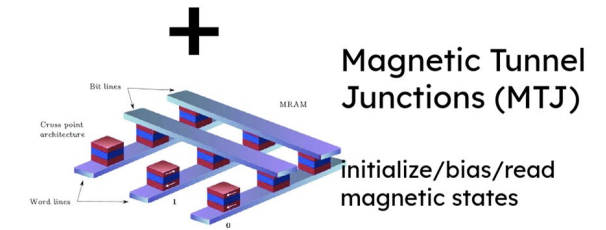
Randomized computing, *i.e. Ising machines, stochastic computing, Randomized Numerical Linear Algebra (RandNLA)*, has proven useful to accelerate time to solution, power consumption and reliability for complex algorithms. We propose the investigation of specialized hardware architectures that can leverage the random behavior of CMOS and emergent technologies for ultra energy efficient computing.

State of the Art and Challenges

Hardware specialization is already present in HPC and commercial systems (*GPU, TPU...*) demonstrating performance acceleration at the cost of energy consumption. Some work has explored computing with stochastic behavior (*probabilistic computing*). However, there is not a clear link between algorithms (RandNLA), hardware accelerators and technologies either conventional (digital) or emergent (MTJs, ASI)



Artificial Spin Ice
P-Bit algorithms expressed in the Ising model.



Magnetic Tunnel Junctions (MTJ)
Initialize/bias/read magnetic states

Fast and parallel calculation on the Spin Ice Lattice using dipolar interaction

Execution and Timeline

- Clearly identifying applications that benefit from randomized algorithms.
- Specialized hardware targeting randomized algorithms, *i.e.* RandNLA. Technology agnostic, high level modeling to meet real application requirements
- Evaluation of classical (CMOS) and emergent technologies (ASI, MTJ...)
- Accurate Hardware evaluation and prototyping (FPGAs, LBNL ALS)

Opportunity and Potential Impact

- Novel, high-performance neuromorphic compiler
- We have large spiking neuromorphic systems today ... and will have larger systems in the future
- We will be able to fully utilize these large systems
- Run larger benchmarks and workloads!

State of the Art and Challenges

- Compilation and optimization of neuromorphic algorithms is a manual and intensive process
- Many potential optimizations involve solving hard problems (like graph partitioning)
- Some specialized optimizations; few general optimizations

Compiler feature	Potential Neuromorphic Equivalent
Data-flow and Control-flow analysis	Spike behavior analysis – predict spiking behavior (not the algorithm output)
Constant folding and propagation	Replace neurons that start with potentials higher than their threshold with an input signal
Register allocation	Partition neurons based on their connectivity
Inline expansion	Replace synapses with large delay values with a chain of neurons
Dead-code elimination	Remove neurons that don't impact the rest of the network

Execution and Timeline

- Steps:
 - Identify key optimizations that can be done
 - Develop novel algorithms or utilize existing solutions (chance for new theory)
 - Test methods on state-of-the-art platforms
- Requires platforms, datasets, and workloads

Toby Isaac, Argonne National Laboratory

Assessing the energy efficiency of AI for scientific applications

Energy Efficiency Challenge of AI4S: Tradeoff Analysis

- If AI4S is different from general AI, it should be more **rigorously quantifiable**.
 - Positivity bias, weak baselines, asymmetric comparisons (e.g. runtimes of models with significantly different accuracy) work against good science [Good example: arxiv:2407.07218]
 - Models must be compared/understood through their **tradeoffs** between time, energy & accuracy

Execution

- **Machine-and-algorithm-independent figures of merit** as in ECP for AI4S projects
- **Instrumentation** of both models and training algorithms
- **Publication and reporting standards** for both training and trained model performance

Energy Efficiency Challenge of AI4S: Embodied Energy

- Model efficiency \neq energy per model evaluation: training energy must be amortized over model evaluations
 - Foundation model training from tokens is approaching years of exascale compute
 - Generating tokens from numerical simulations will be a significant additional cost of AI4S vs. other LLMs
 - # model evaluations for DOE-scale models (e.g. climate model surrogate) significantly smaller vs. general applications

Potential Impact

- **Better information for decision makers:** reported embodied energy and model accuracy give more confidence when adopting new models
- **Better utilization of computing resources**

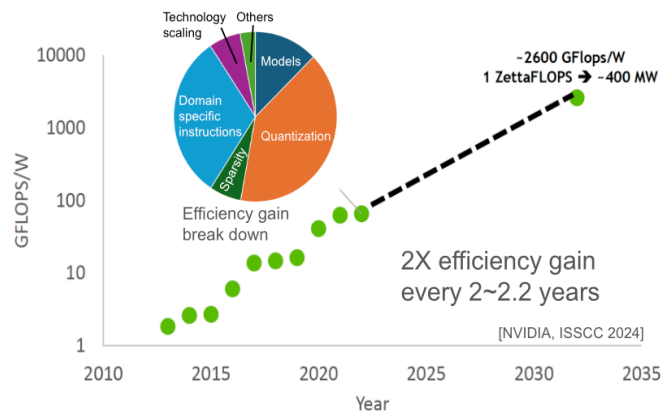
Jaydeep P. Kulkarni, The University of Texas at Austin

Advancing Energy-Efficient Domain-Specific Computing Through Memory-Centric Architectures And Distributed In-Network-Computing

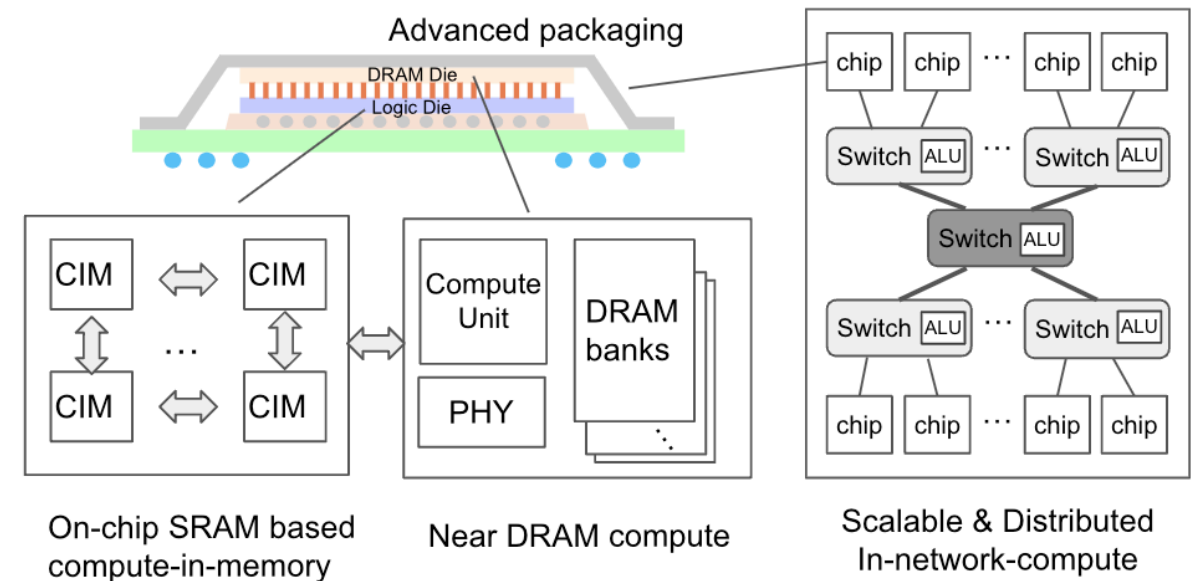
Opportunity and Potential Impact

- AI / Scientific computing scales up rapidly with huge data
- Data movement occupies major chunk and sources of efficiency gain of AI/Scientific HW are unsustainable
- **More systematic:** Distributed Memory-Centric Computing
- Over order magnitude efficiency gain with ultra scalability

State of the Art and Challenges



- Domain-specific optimization dominate the efficiency trajectory
- Hard to keep the efficiency scaling for the near future



Execution and Timeline

- We have in-house on-chip memory-centric demos
- **Future:**
 - In 5 years: custom 3D packaged DRAM; compute-capable networking; memory-centric software frameworks
 - In 10 years: chiplet or network-based scaled system
- **Key barriers:** DRAM process access; compiler and software support; funding support

Opportunity and Proposed Research Direction

- **ML-enabled modeling**
 - From microelectronic device to system
 - Integrated performance and energy models
 - Learning from execution traces for accuracy
- **ML-enabled optimization**
 - Gradient-based optimization
 - End-to-end/component level

Execution and Timeline

- **Model development**
 - CPU/GPU/accelerator/memory/network models
 - Performance/energy/thermal models
- **Training data**
 - Testbeds/simulators to generate data
- **Optimization technology**
 - Gradient-based optimizer enhancement

State of the Art and Challenges

- **Energy efficient computing trends**
 - Domain-specific computing
 - Novel computing paradigm and technology
- **Increasing heterogeneity and co-design space complexity**
- **Modeling and optimization technology**
 - Modeling: simulators/emulators
 - Optimization: RL, Bayesian optimization, ...
 - Challenge: limited speed and scale

Potential Impact

- An ML-enable modeling and optimization framework
- Significantly accelerate the modeling and optimization of future HPC architectures and systems
- Enable faster and deeper co-design

Denis Mamaluy, Sandia National Laboratories

Predictive Simulations for CMOS and Beyond-CMOS Devices for Energy Efficient Computing

Opportunity and Proposed Research Direction

- Develop a universal, first-principles (without fitting parameters) Device Simulator that reveals physics and predicts electrical characteristics of CMOS and/or beyond-CMOS devices
- New numerical methods empowered by AI/ML now enable sufficiently fast and predictive first principles simulations

Execution and Timeline

- Outcome: A framework that allows to share/validate device parameters and characteristics (component Digital Twins)
- Validation steps: 1) gather/update novel device parameters/characteristics; 2) conduct simulations; 3) include necessary physics; 4) validate results
- ~5 years: accurately predict electrical characteristics on *any* CMOS or beyond-CMOS device
- ~10 years enable the whole-system co-design framework

State of the Art and Challenges

- Drift-diffusion TCAD tools are no longer predictive for nano-scale devices (quantum effects dominate)
- Commercial/academic first principles quantum transport codes are still too slow to model realistic devices, scale as $O(V^3)$ with the volume V
- The dimensions of state-of-the-art (GAAFETs/TFETs/van-der-Waals material) transistors, $V \sim (50\text{nm})^3$, are still too large for existing first principle codes

Potential Impact

- Universal Predictive Device Simulator enables:
 - Co-Design on the Material-Device-Circuit level without fabrication
 - Whole-system Energy Efficiency (or SWaP) optimization
 - Accurate assessment and optimization of beyond-CMOS technologies
- The success is directly measured by the ability to predict outcomes of experimental measurements without fabrication of actual devices.

George Michelogiannakis, LBNL

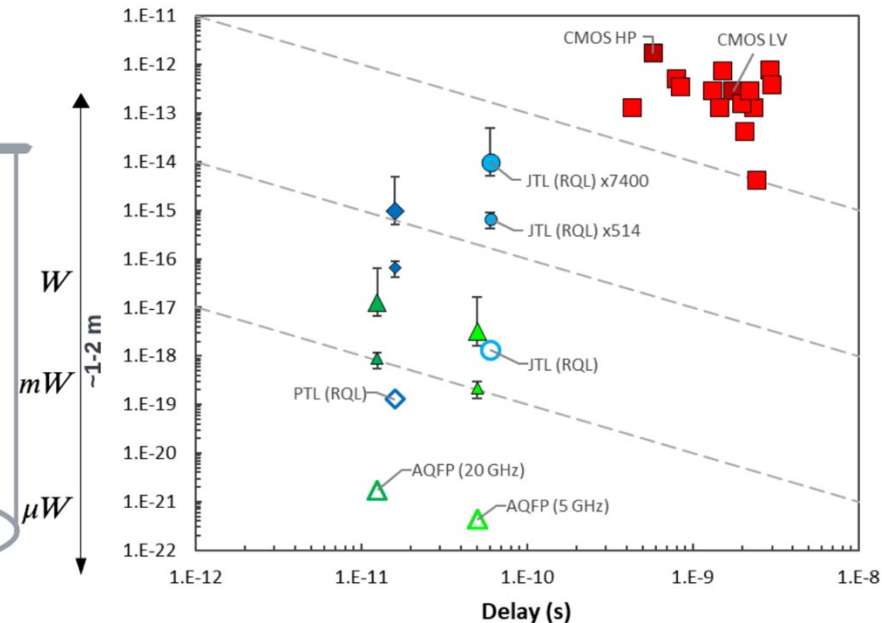
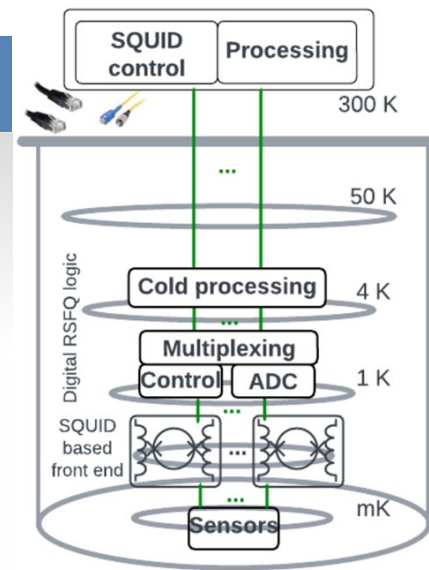
Superconducting Robust Digital Computing for HPC

Opportunity and Potential Impact

- What: Alleviate superconducting digital computing's limitations and adapting to scientific applications
- Now: HPC and sensor applications are reaching a wall
- Impact: Orders of magnitude higher EDP and increase reliability, response times, and cost for cryogenic sensors
- Success: Comparable BER with 100x to 1000x lower EDP

State of the Art and Challenges

- Superconducting digital computing (SDC) faces low device density, and bit error rates in cables
- Compute methods and memory must adapt to SDC instead of copy paste from CMOS
 - Such as “free” on-chip data movement (figure)
- Requires expanding our metrics and re-designing H/W



Execution and Timeline

- Key steps: Develop SDC accelerators for key HPC applications and processing for cryogenic sensors
- Timeline: Initial investments in manufacturing and devices with parallel efforts in hardware design. Then joint funding to prototype circuits. Requires testbeds
- Barriers: SDC circuit error evaluation methodology, prototyping and testing in labs, HPC system integration

Tom Peterka, Argonne National Laboratory

Priority Research Directions for an Energy-Efficient Data Framework

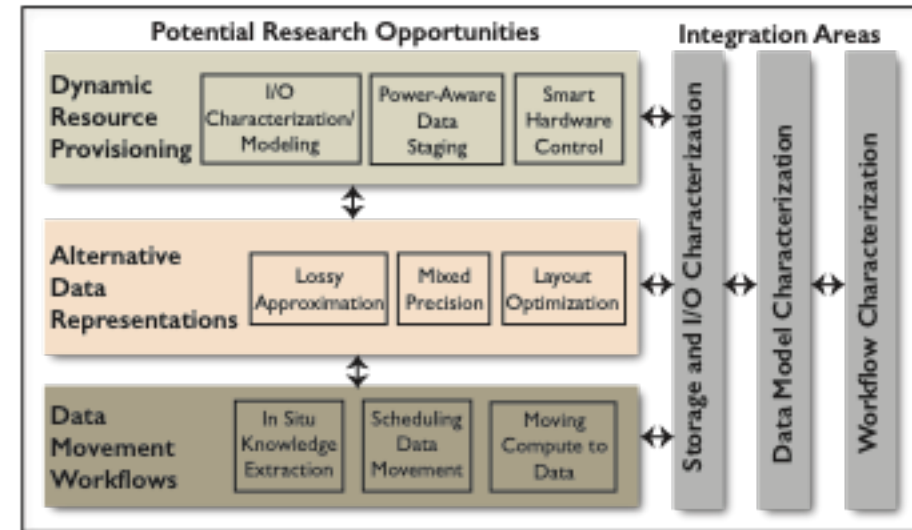
Opportunity and Potential Impact

- Control energy cost of storage systems, data representations, and data movement in workflows by exposing h/w to s/w layers, monitoring, user interfaces.
- DOE science is data-intensive; ML and AI are data-driven; data movement is expensive. Recent ASCR projects provide foundations for energy optimization.
- Increase energy efficiency, improve machine utilization, potential new advances in system software, scientific computation, and data analysis.

- Characterization needed to measure success.

State of the Art and Challenges

- Today black boxes w/ little to no energy info/control, focus on computation.
- Need fine-grain energy characterization and control, focus on data management and movement.



Potential research opportunities include three main research areas and integration with other potential topics.

Execution and Timeline

- Research characterization of energy usage for storage, data models, workflows.
- Research energy optimization in dynamic resource provisioning, data representations, workflows.
- Sustained funding, disruptive hardware, market drivers
- Hardware testbeds, system software, synthetic benchmarks, test applications

David A. Roberts Ph.D., Micron Technology Inc.

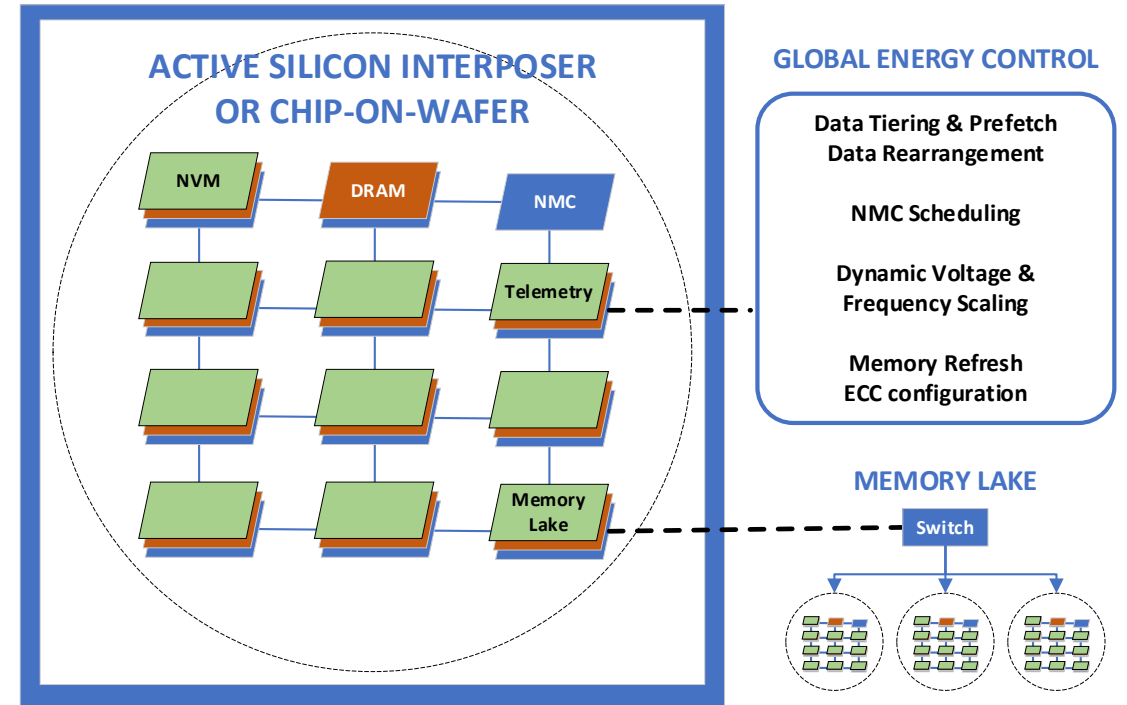
FEMTO: Fusion of Energy-efficient Memories with Telemetry-driven Optimization

Opportunity and Potential Impact

- Tightly couple 3D-integrated Near-Memory Compute (NMC) processors with tall, dense 3D memory stacks to reduce data movement. Interconnected Memory Lake for scale-out capacity
- Distributed telemetry drives global energy-aware scheduler that controls task & data movement, DVFS, ECC & Refresh parameters
- Breakthroughs in memory density & control algorithms increase performance per Watt utilizing ML-capable NMC

State of the Art and Challenges

- New customizable systems need discovery of topology, telemetry, control parameters and their impact. Disparate non-standard interfaces, hundreds of proprietary control knobs today
- In some systems, system optimizer with energy cost function needs to quickly learn dynamic input-output relationships
- Telemetry and NMC that are optimized for advanced ML-based scheduling and prefetching will take time to research



Execution and Timeline

- Near-term requirements are abstracted System Design Space Exploration models for workload, compute, memory & network
- Must have fidelity to model energy impact of all control knobs. Accurate component parameters and fast models for large scale
- For success we need representative, summarized workloads that capture essence of behaviors that affect energy

Fred Suter, Oak Ridge National Lab

Comprehensive Digital Twins of Leadership Computing Facilities

Opportunity and Proposed Research Direction

- Comprehensive Digital Twin: virtual prototyping capabilities to **explore "what-if" scenarios**
 - From the applications/workflows to the power and cooling of the facility
- Several **complementary simulation tools** exist
 - Different component of the full system
 - Different scales

Execution and Timeline

- Key steps: Combine simulation tools, design **multi-scale models**, consider **representative workloads**
- 5-10-15 years outlook: Obtain **full insight** on the energy efficiency of LCFs and their applications
- Barriers: Mastering scale and complexity
- Resources needed: Ground truth data including **app. fingerprints**, workflow **benchmarks**, and **telemetry** data

State of the Art and Challenges

- Different tools for different views but **without a full picture, estimations may deviate from reality**
- Building a monolithic framework is hardly possible
- Combining tools in a **modular** way is a challenging task
- **Calibration** and **multi-scale models** are key to realism

Potential Impact

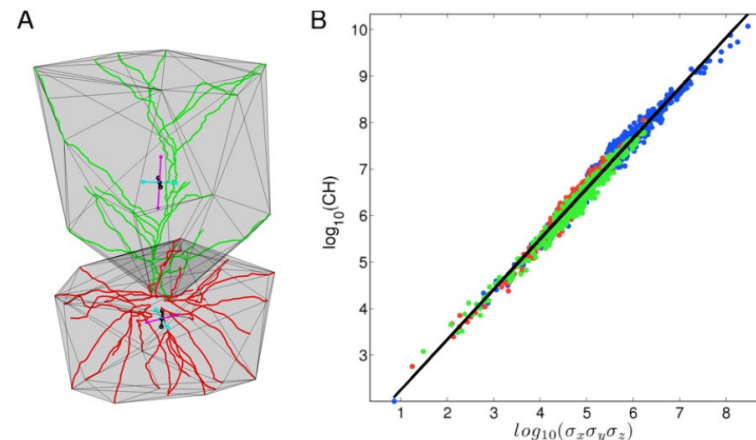
- Helping the procurement of future Leadership Class Facilities with energy efficiency in mind
- Success metrics: Performance vs. Energy studies at scale, exploration of alternate designs with objective classification indicators

Corinne Teeter, Sandia National Laboratories

Neural-Inspired Architectures and Learning Rules for Energy Efficient Computing

Opportunity and Potential Impact

- Problem: Current AI extraordinarily power intensive
- Proposal: Use brain as a proof of principle to guide creation of energy efficient AI
 - Two high risk/reward proposed research areas
 - Brain-Inspired 3D circuit architectures
 - Local learning algorithms
- Impact: 100's W \rightarrow 20 W



Teeter 2010, Teeter & Stevens 2011

All neural arbors spread themselves out over space according to a scaling law that covers 7 orders of magnitude.

Likely to optimize connectivity and dissipate power.

State of the Art and Challenges

Current AI training:

- Back propagation through time
- Slow and power intensive
- Requires many forward and backward passes
- Principled algorithm that mostly works well—with recent computational power/memory improvements it became possible to implement it. Now we are stuck here.

Current hardware

- 2D is established paradigm
- Cannot reach brain-like connectivity
- Majority of power goes to connections

Execution and Timeline

Local learning

- To date researchers have not found a way to achieve useful learning using local rules in deep neural networks.
- Lack of theory
- Large computing platforms
- Potentially specialized hardware and codesign

Brain-Inspired 3D hardware

- Software for 3D physical hardware simulation
- Advanced tools for new fabrication techniques
- Fabrication is expensive

Xingfu Wu, Argonne National Laboratory

A Software Codesign Framework for Energy Efficient Scientific Applications

Opportunity and Proposed Research Direction

- End-to-end autotuning for energy efficiency
 - Huge tunable parameter space, our ML-based autotuning effort
- Leveraging LLMs to build performance and power models and to tackle automated code generation and translation for energy efficiency
 - Recent advance in LLMs for code generation and our effort

Execution and Timeline

Tackle three critical research areas through a software codesign approach:

- Application-system characterization and modeling
- Cross-layer software codesign for energy efficiency
- Energy efficient code generation and translation by leveraging LLMs.

State of the Art and Challenges

- Achieving energy efficiency becomes a challenge because of the complexity of heterogeneous HPC ecosystems
- LLMs have become a paradigm-changing innovation
- State-of-the-art code-centric LLMs still lack sufficient training data for energy efficient scientific codes

Potential Impact

- Automated generation of energy efficient scientific codes for different HPC systems
- Boosting gains in power efficiency by tapping into interoperability among multiple layers of HPC ecosystems
- Saving energy and cost and reducing carbon footprint for HPC centers

Hui Zhou, Argonne National Laboratory

Energy Efficient Runtimes Require Flexibility and Interoperability

Opportunity and Proposed Research Direction

- Drastic improvement takes novel solutions (e.g. dynamic meshing, mixed-precision, adaptive algorithms)
- Novel solutions are hindered by the runtime support.
- Flexibility – non-dominant patterns, low performance penalties
- Interoperability – hybrid runtimes focus and leverage each other, avoid working around and reinventing wheels.

Execution and Timeline

- Example research: hybrid/dynamic composable launch, MPI stream, MPI async, Integrated progress engine, Unified MPI+OpenMP parallel region, and so on.
- Vision: less complaint from users who explore novel solutions, more options, more collaborations, more solutions.
- Barrier: the chicken and egg conundrum.
- Necessary for success: a fertile and stable talent base.

State of the Art and Challenges

- MPI focuses on bulk-synchronous patterns, neglects dynamic processes, progress management, hybrid execution contexts.
- OpenMP broadens beyond its original scope, yet still struggles to get along with MPI.
- New runtimes repeating the same silo stories.
- **Synchronizations** are the bottlenecks of efficiency.

Potential Impact

- Bootstrapping novel solutions
- Avoid duplicate investments
- Leverage, complementing, rather than impeding
- More innovative experiments, less complaining on complexity and support