LLNL capabilities and interests

Algorithms and Software; Materials/Processes Modeling and Simulation

BMC Workshop

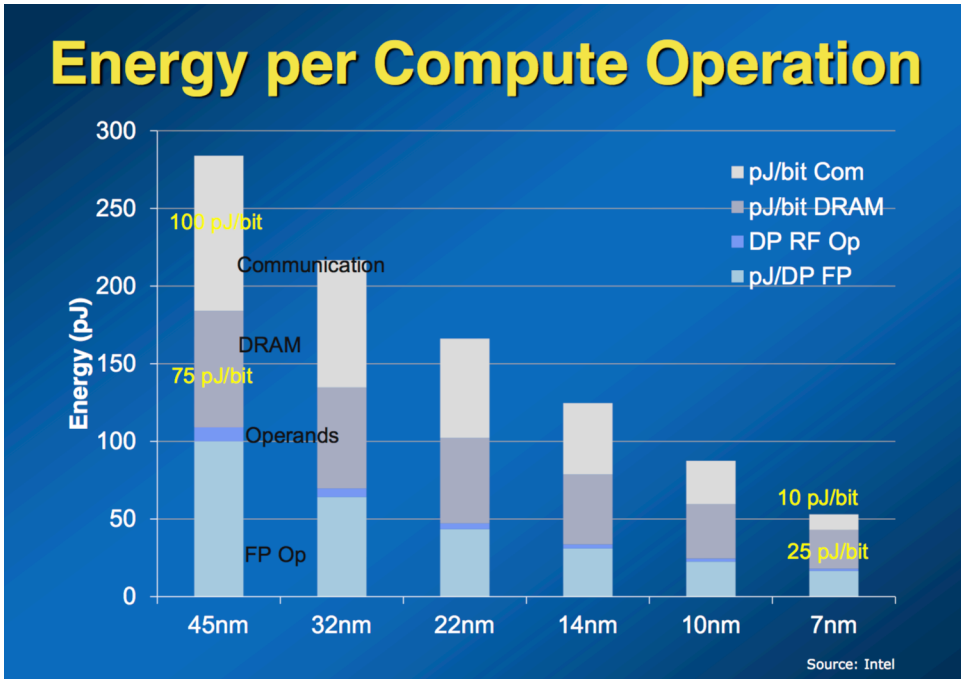
Maya Gokhale
Matthew Horsley
Peg Folta

July 27-28, 2016

LLNL-PRES-XXXXXX
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Algorithms and software: focus on memory



Source: Borkar IPDPS 2013

- Energy cost to access data in memory hierarchy are projected to exceed compute cost
- Packaging and pin limitations constrain bandwidth
  - "have to turn off a quarter of the cores"
- Advances in resistive memory technology offers alternative to CMOS

LLNL R&D over past 4 years on architecture, programming models, systems software, evaluation of memory-centric computing.

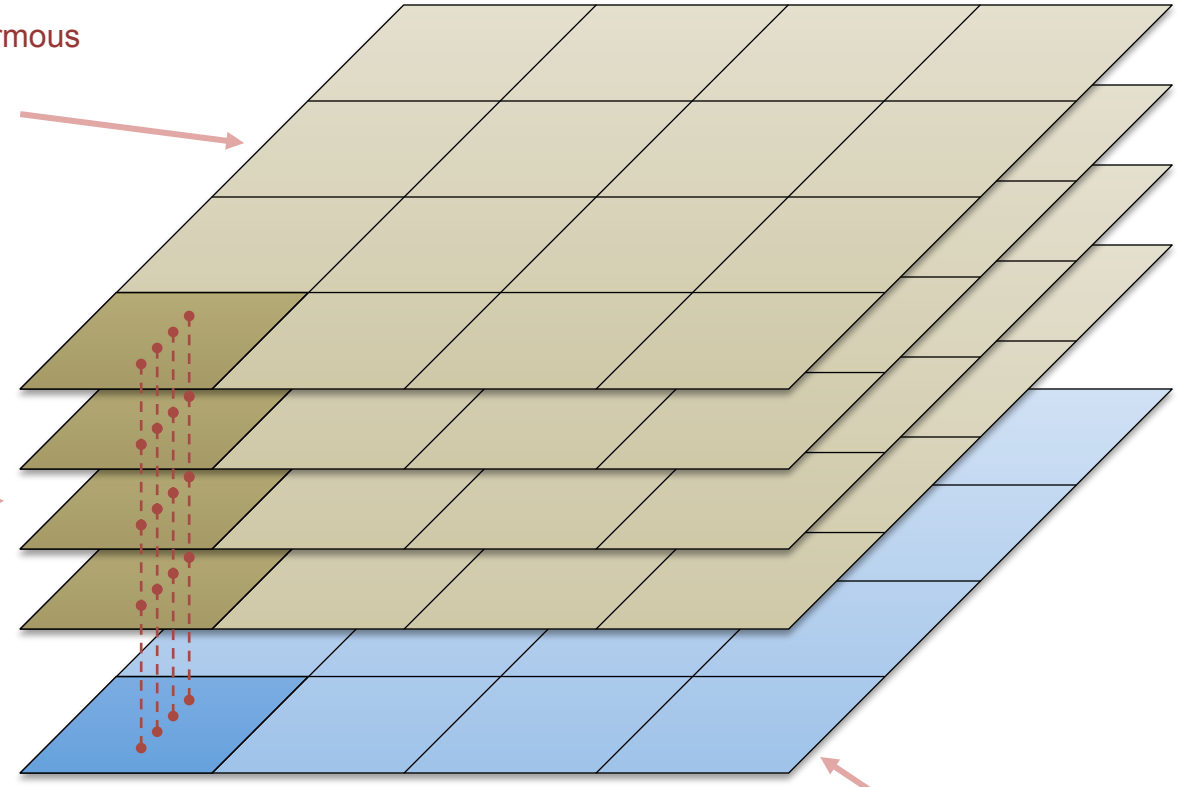
Memory integrated computing

- Exploit advances in 3D stacked memory/logic packages
  - Logic is on a separate layer in a different technology than the memory
  - Through silicon vias connect memory layers, logic
  - Can accommodate different sorts of memory
  - DRAM, NVRAM
  - Don't have to fab logic circuits on the memory die
  - Can co-locate processing engines with (persistent) memory
- Integrate compute and memory at finer granularity
  - Micron Automata processor
    - NFA logic embedded in the DRAM
    - Many NFAs running in parallel
    - String matching applications
  - RRAM with TCAM
    - Ternary Content Addressable Memory using memristor as building block
    - Pattern match and search applications

Enablers for processing near memory

Multiple channels yield enormous bandwidth in the package

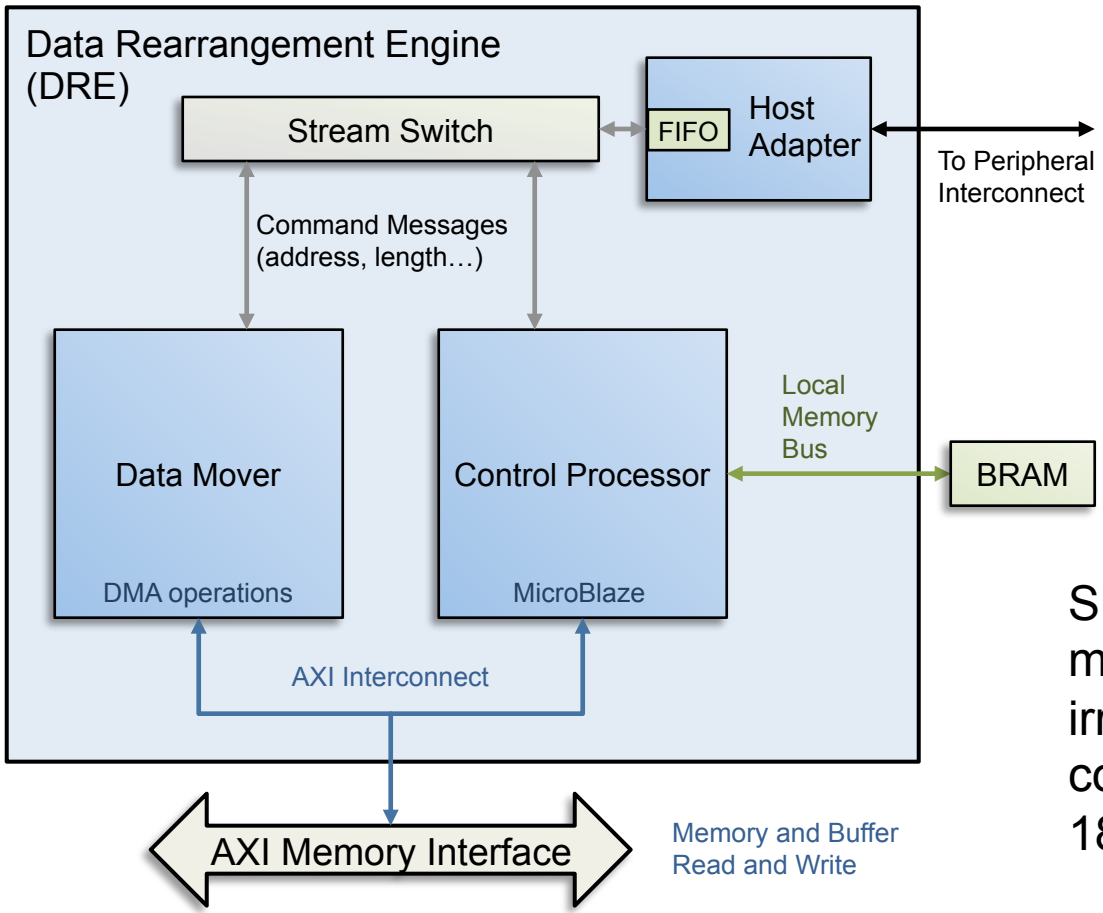
Stacked 3D or 2.5D memory packaging technology with TSVs to form a memory channel



Fast logic layer with low latency access for local memory controllers and customized processors

Fabrication cost requires compelling commercial drivers

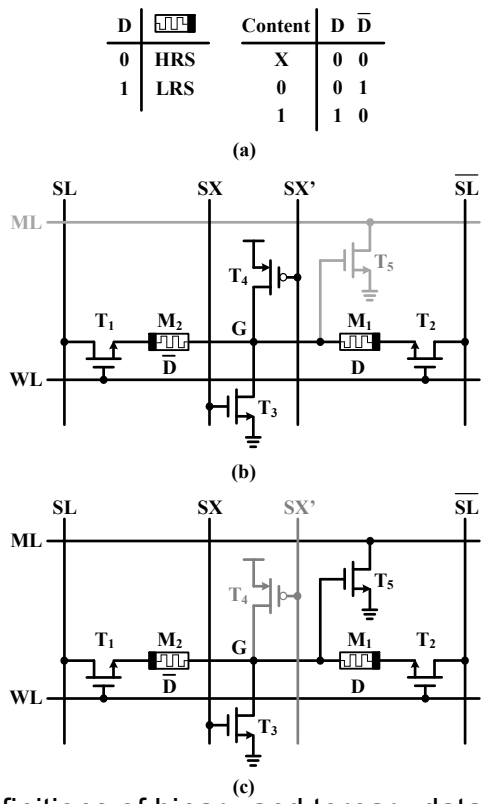
Proposed Data Rearrangement Engine Architecture



S. Lloyd and M. Gokhale, "In-memory data rearrangement for irregular, data intensive computing," *IEEE Computer*, pp. 18–25, 2015.

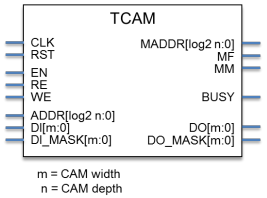
LDRD 2013-15

TCAM based on memristor

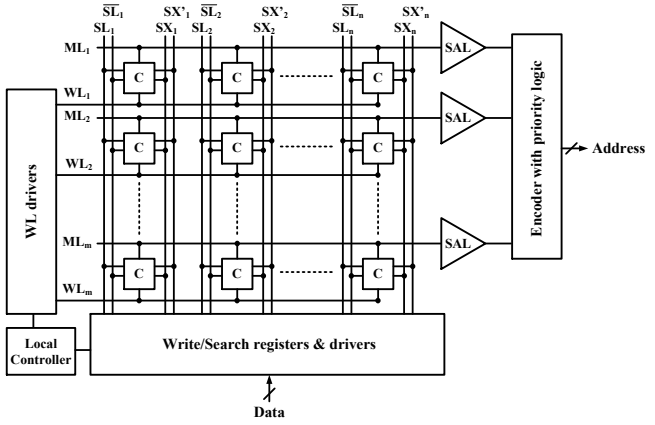


(a) Definitions of binary and ternary data, (b) RRAM-based TCAM cell in write mode, and (c) RRAM-based TCAM cell in search mode.

. (UCSC/LLNL Lab Fees project)



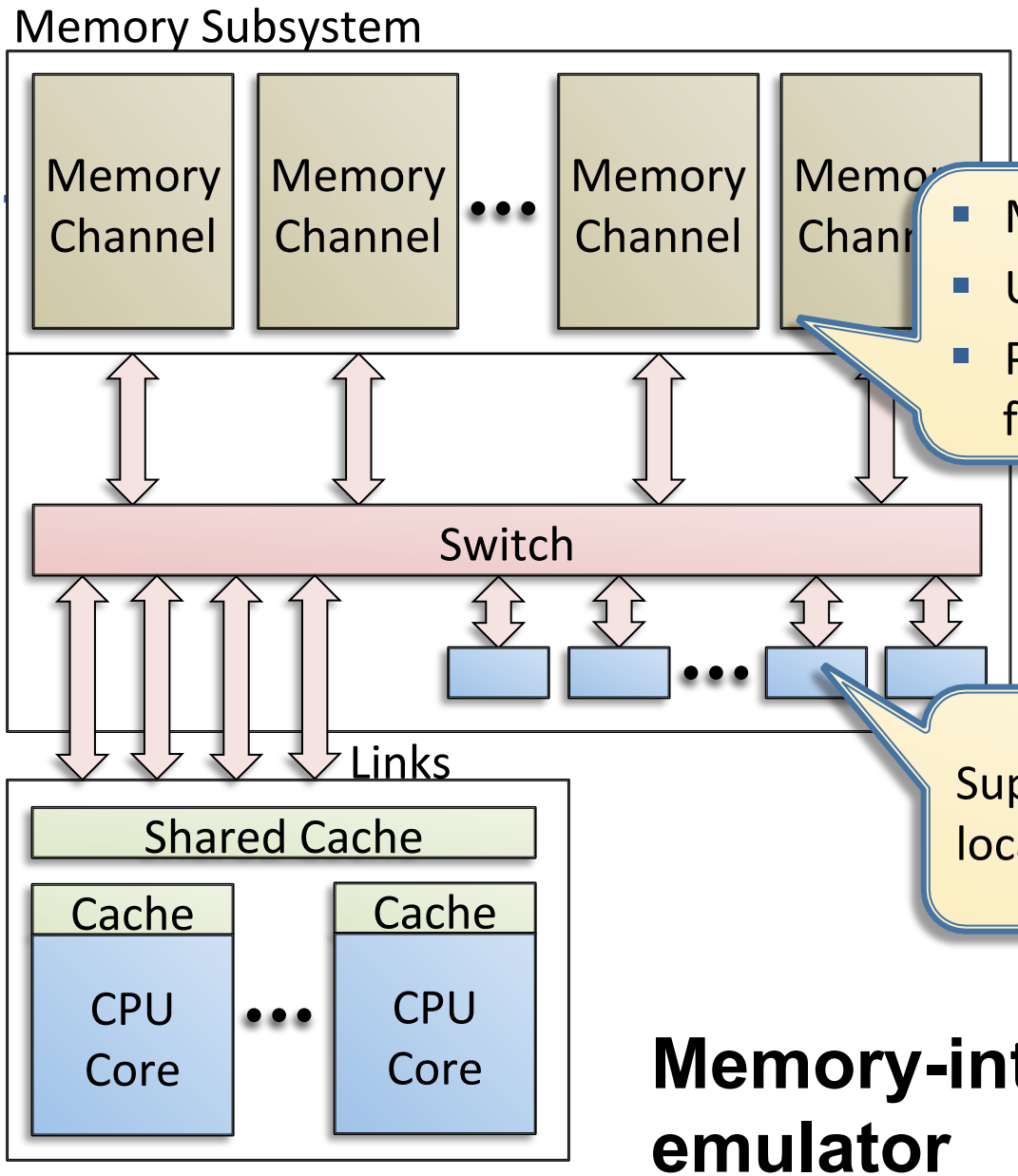
Name	Dir	Description
CLK	I	Clock
RST	I	Reset
EN	I	Enable, searches if not RE or WE
RE	I	Read enable
WE	I	Write enable
ADDR	I	Address for data read or write operation
DI	I	Data in
DI_MASK	I	Data in mask
DO	O	Data out
DO_MASK	O	Data out mask
MADDR	O	Match address
MF	O	Match flag
MM	O	Multiple Match
BUSY	O	Busy



Le Zheng, Sangho Shin, Scott Lloyd and Maya Gokhale, “RRAM-Based TCAMs for Pattern Search”, 2016 IEEE Int’l Symposium on Circuits and Systems, Montreal, Canada, May 2016

Emulation of memory integrated computing

- We have implemented an FPGA-based emulator to study future memory integrated computing architectures
  - Scott Lloyd lloyd23@llnl.gov
  - Open source with software, hardware, benchmarks
  - http://bitbucket.org/perma/emulator_st/src
- Used to evaluate data rearrangement engine, future persistent memories, TCAM modules
- Gives emulated performance of the benchmark running on CPU and using memory integrated computing resources
- Enables design and evaluation of hardware modules, API, application

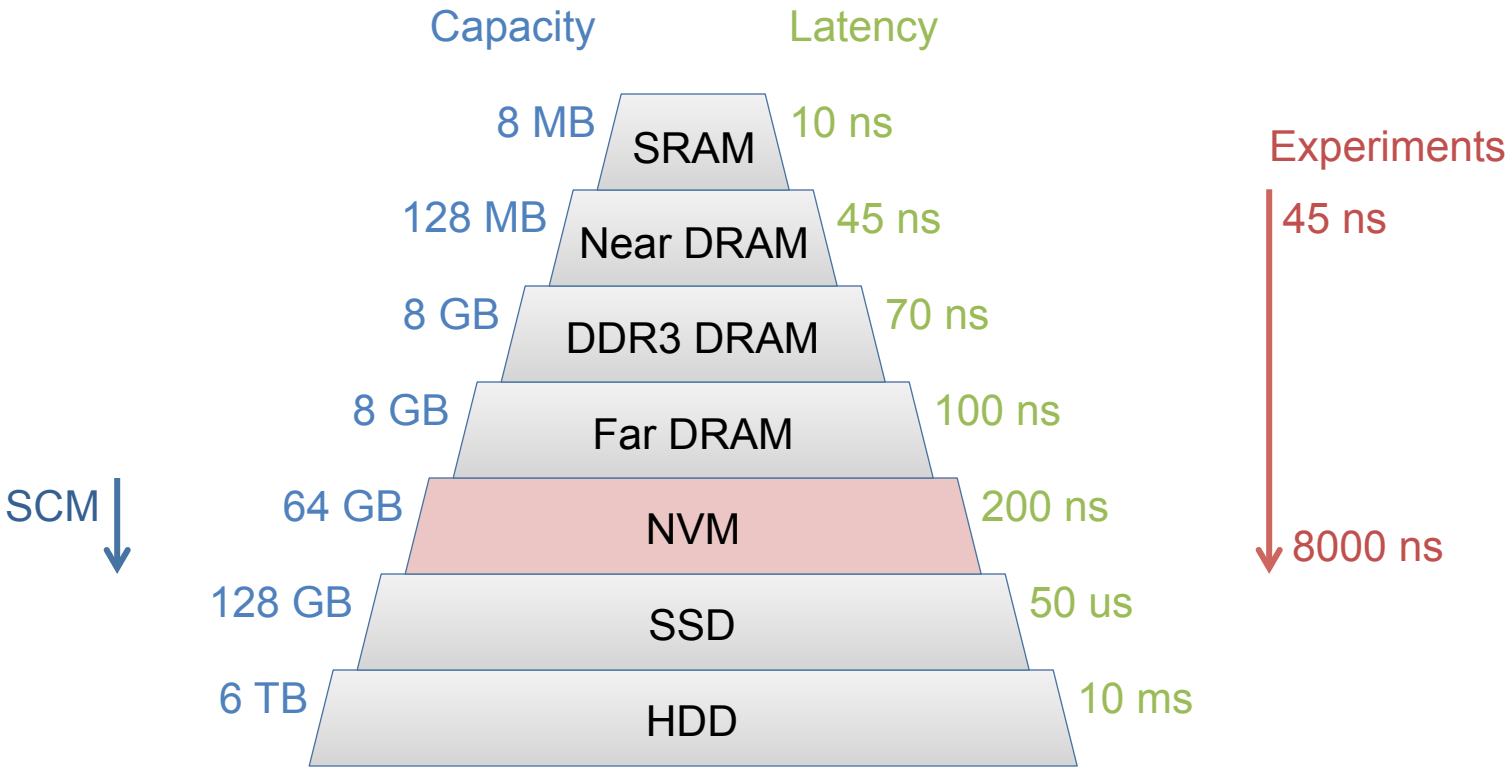


- Multiple Memory Channels
- Up to 16 concurrent memory requests
- Programmable memory access time from 8 ns to 262 us

Supports custom hardware modules located in the Memory Subsystem

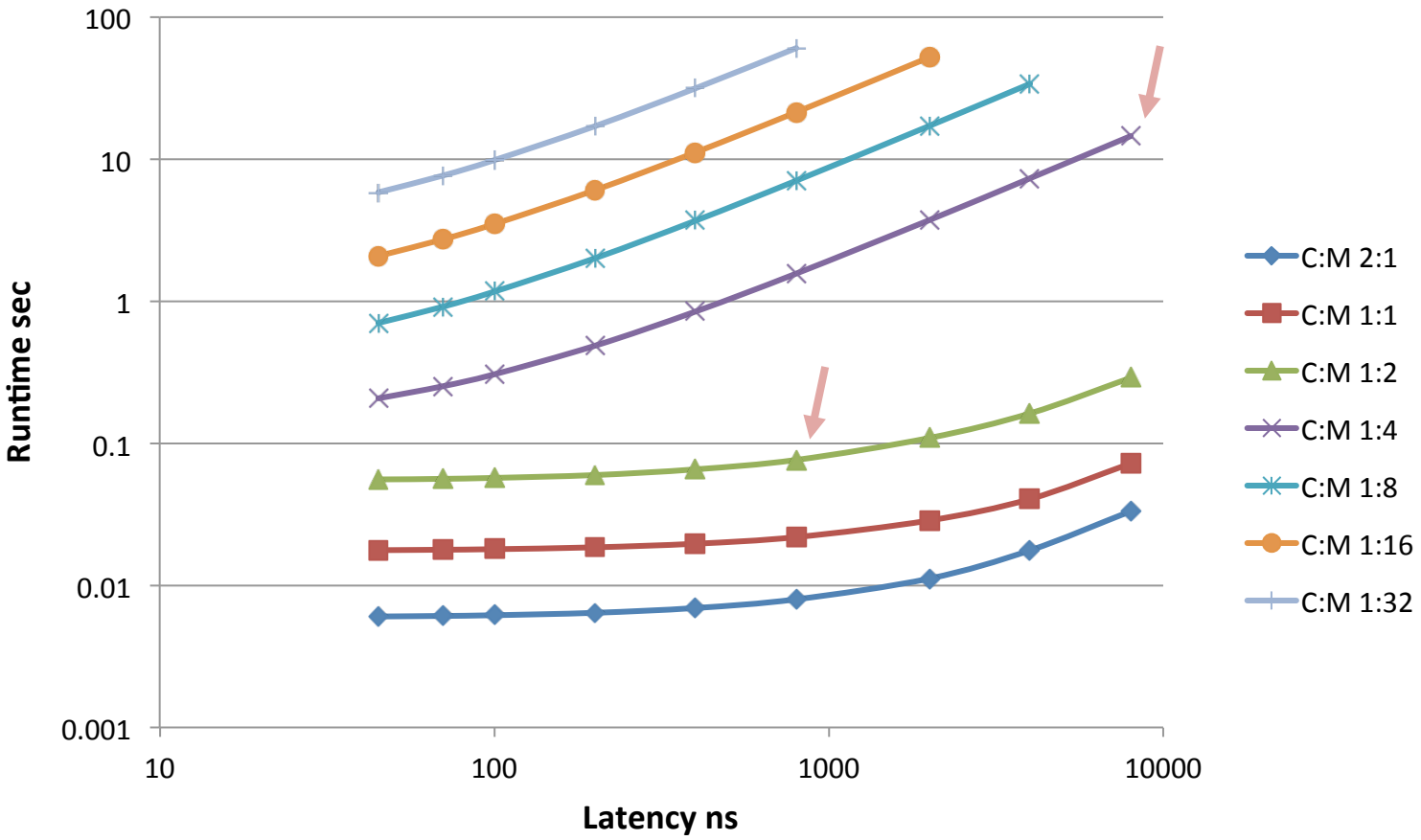
Memory-integrated computing emulator

Memory/Storage Hierarchy



DGEMM

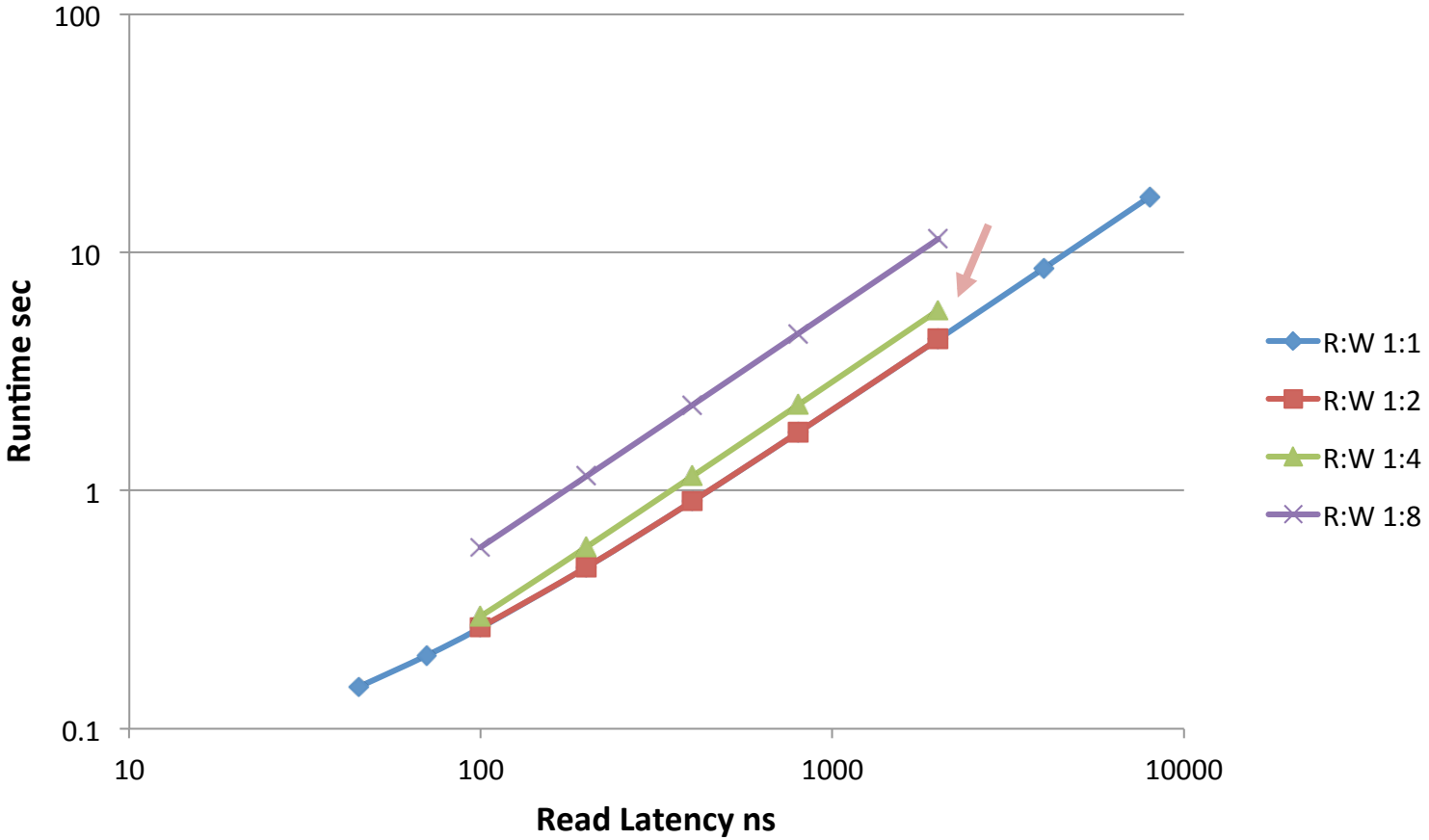
R:W 1:1, Regular Access Pattern



- At cache to memory ratios of 1:2 and lower, latency up to 800 ns can be tolerated
- At cache to memory ratios of 1:4 and higher, runtime increases linearly with latency
- Matrix multiply kernels for small (perhaps blocked) matrices can tolerate SCM latency

RandomAccess

C:M 1:1024, Irregular Access Pattern



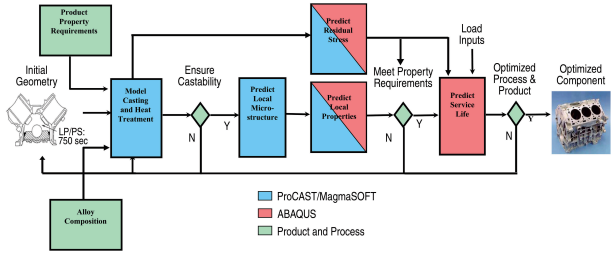
- A read to write latency ratio up to 1:4 has little impact on performance
- Direct linear correlation between memory latency and runtime
- Concurrent threads could in aggregate compensate for longer SCM latency

Memory-integrated computing summary

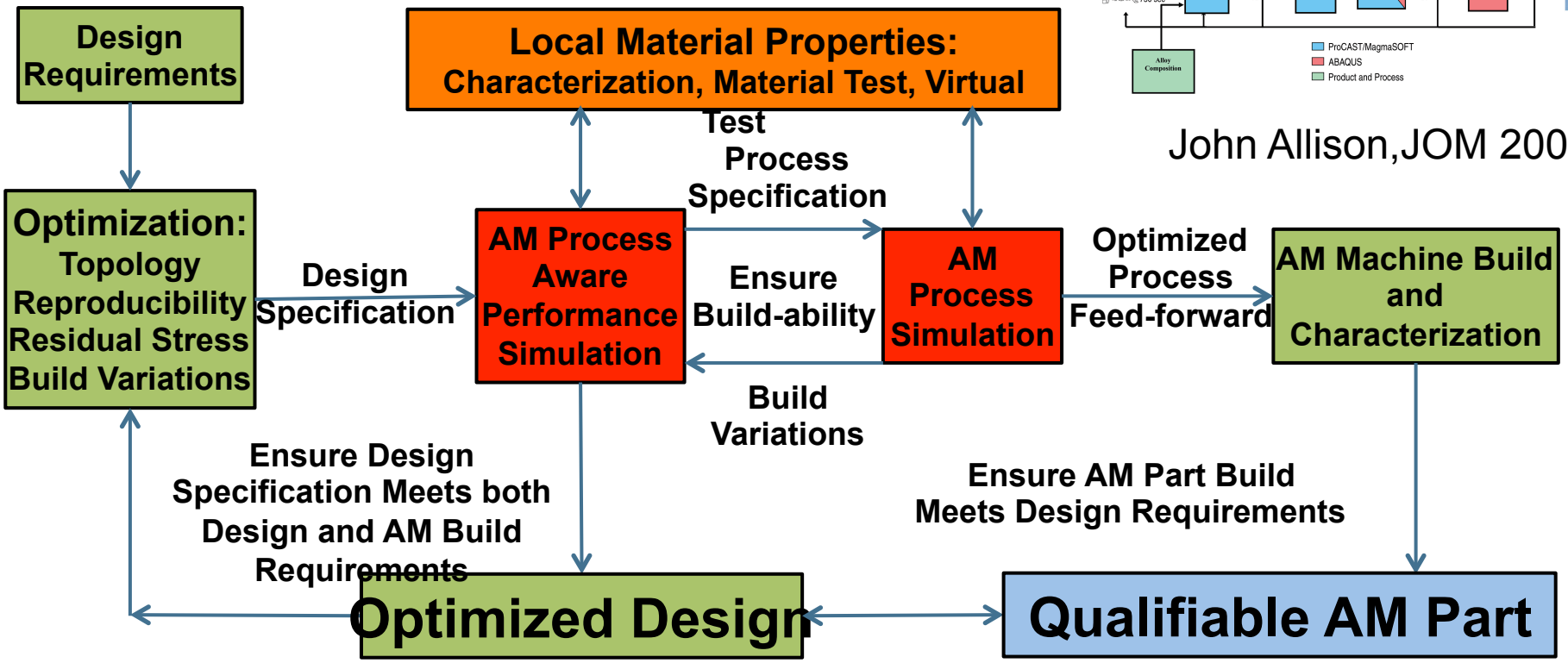
- Expertise
  - Design, prototype logic functions incorporated into memory
    - On the fly data rearrangement for irregular memory access
    - Pattern match based on Ternary Content Addressable Memory
  - Programming models, API, heterogeneous computing perspective
  - Exploit memory based logic functions from application
  - Prototype functionality with FGPA System on Chip
    - Multicore ARM processors + programmable logic
  - Measure emulated performance at 10ns – us granularity to sweep parameter space of (resistive) RAMS
  - LDRD, DoD funded since 2013
- Differentiating factors: Substantial investment and capability in
  - FPGA-based emulator,
  - heterogeneous programming models,
  - application benchmarks using memory-integrated computing logic

Materials modeling and simulation

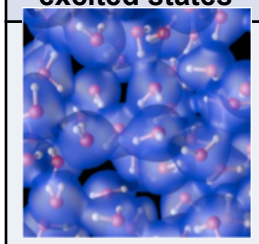
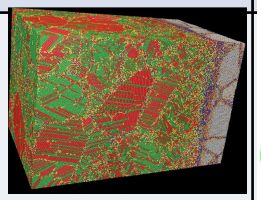
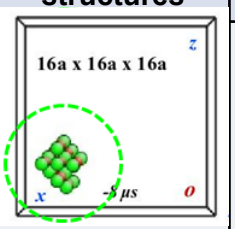
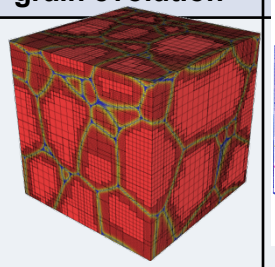
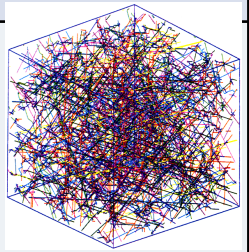
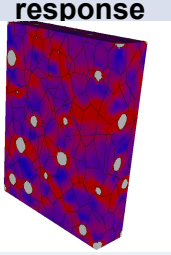
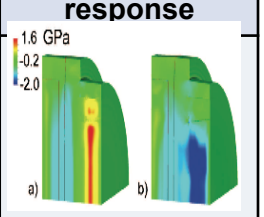
Workflow for Design Optimization of Additively Manufactured Parts (Integrated Computational Engineering)



John Allison,JOM 2006



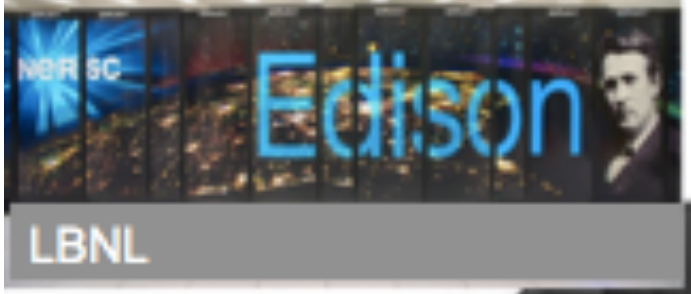
Seven Pillars of Computational Materials Science

Ab-initio	Atoms	Long-time	Microstructure	Dislocation	Crystal	Continuum
Inter-atomic forces, EOS, excited states	**Defects and interfaces, nucleation**	**Defects and defect structures**	**Meso-scale multi-phase, multi-grain evolution**	**Meso-scale strength**	**Meso-scale material response**	**Macro-scale material response**
						
LATTE Motif: Particles and wavefunctions, plane wave DFT, ScaLAPACK, BLACS, and custom parallel 3D FFTs Prog. Model: MPI + CUBLAS/ CUDA	ddcMD/CoMD Motif: Particles, explicit time integration, neighbor and linked lists, dynamic load balancing, parity error recovery, and *in situ* visualization Prog. Model: MPI + Threads	SEAKMC Motif: Particles and defects, explicit time integration, neighbor and linked lists, and *in situ* visualization Prog. Model: MPI + Threads	Motif: Regular and adaptive grids, implicit time integration, real-space and spectral methods, complex order parameter Prog. Model: MPI	Motif: "segments" Regular mesh, implicit time integration, fast multipole method Prog. Model: MPI	Motif: Regular grids, tensor arithmetic, meshless image processing, implicit time integration, 3D FFTs. Prog. Model: MPI + Threads	LULESH Motif: Regular and irregular grids, explicit and implicit time integration. Prog. Model: MPI + Threads

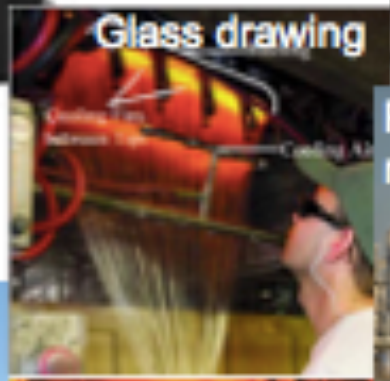
Jim Belak, ExMatEx

LLNL HPC4Manufacturing perspective

AMO established the HPC4Mfg Program to bridge US manufacturing needs and HPC capabilities at the national labs



HPC4Mfg



HPC4Mfg

- Guide design
- Optimize processes
- Pre-qualify parts
- Reduce testing
- Increase innovation

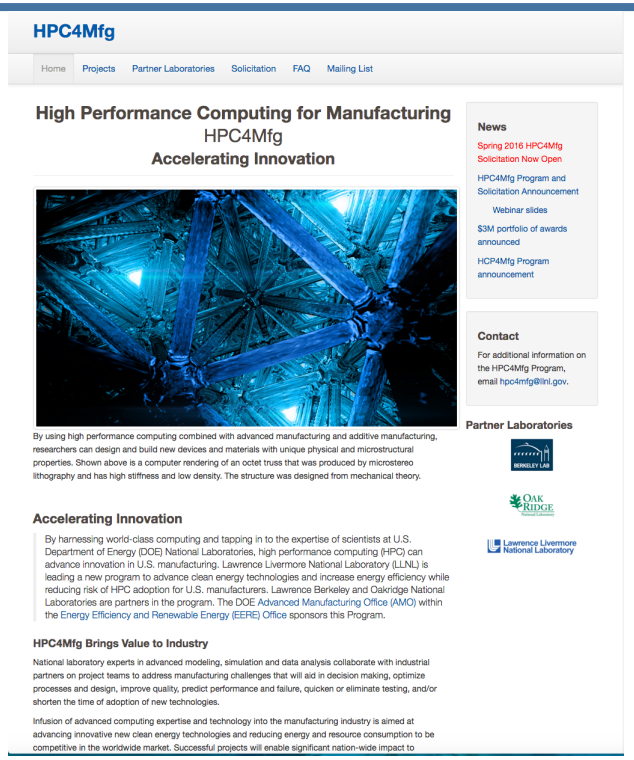
HPC4Mfg can contribute to the challenges of continuing Post Moore's Law… and beyond

- Identify and address manufacturing and applied energy challenges that push current computing capabilities and broaden the HPC user base
  - Multi-scale physics based simulation of metal processing
  - Qualification and characterization of additive manufacturing for aviation, automotive and defense
  - High fidelity grid-scale simulations
- Target projects aimed at device, component and system architecture and testing in the race to the end of Moore's Law and beyond

HPC4Mfg creates an ecosystem that enables the national labs to advance industry innovation

HPC4 MANUFACTURING

- Solicit challenges directly from US manufacturing industry that HPC can address
- Identify HPC expertise at the national labs to meet the industrial challenge; secure cycles
- Execute contracts to demonstrate HPC impact on industry timelines



Increase US economic competitiveness by advancing innovation and growing an HPC manufacturing workforce

ENERGY.GOV
Office of Energy Efficiency & Renewable Energy

The HPC4Mfg Program has had significant engagement from industry

HPC 4 MANUFACTURING

- Bi-annual nation-wide solicitations
- $8.7 M technical portfolio
  - Executing on 16 projects with 14 industry partners and 3 labs
  - Expect to announce ~10 new projects in August
- Established summer internships
- Annual Industry Day to identify industry-wide needs

EATON Powering Business Worldwide

GLOBALFOUNDRIES

United Technologies Research Center

EWI We Manufacture Innovation

SORAA

P&G

lift

PPG

AGENDA 2020 TECHNOLOGY ALLIANCE

Carbontec Energy Corporation

ZoomEssence CREATING THE NEXT WAVE OF INNOVATION

BERKELEY LAB Bringing Science Solutions to the World

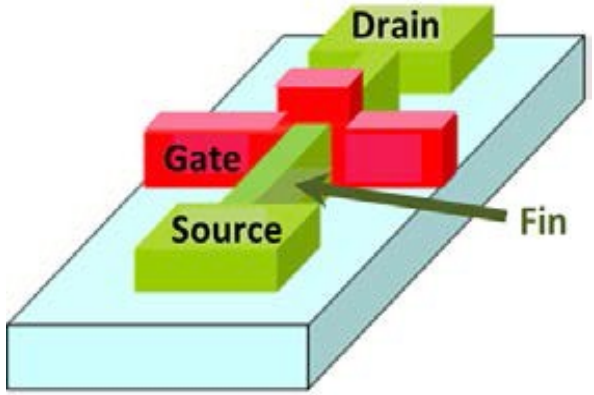
Lawrence Livermore National Laboratory

OAK RIDGE National Laboratory

Computational design and optimization of ultra-low power device architectures reduce cost

HPC 4 MANUFACTURING

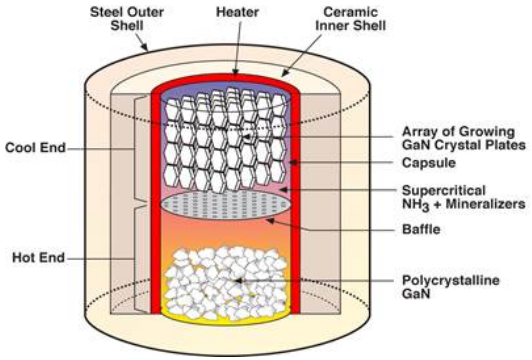
- **Industry partner: Global Foundries**
- Optimizing the design of transistors can result in lower design cost and energy consumption
- **Atomic scale device design modeling** of the heat loss at wire/semiconductor connections can results in the reduction of thermal loss in electronics.
- Device-scale first principles simulations will accelerate materials discovery. Million atom-scale predictions will enable accurate materials properties and device performances.

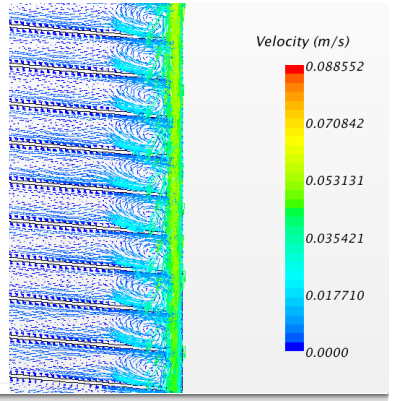


Scaling up a new GaN process could yield 20% cheaper LED lighting and new power electroni

HPC 4 MANUFACTURING

- **Industry partner: SORAA**
- **High fidelity simulations of chemical processes** in the ammonothermal crystal growth process show more complicated flow structure than previous simulations, improving predictions of the local temperatures and flow velocities within the reactor.
- This new information is being used to optimize uniform growth of GaN.
- Insight will save years to facilitate large-scale commercial production.





Microstructure prediction in additively manufactured metal parts can help qualify parts

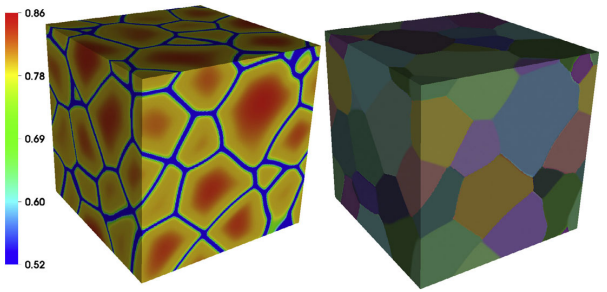
HPC 4 MANUFACTURING

- **Industry partner: Eaton**

- Additively manufactured parts can be used to reduce weight in aircraft and automobiles parts saving transportation fuel.

- **Modeling the material microstructure of parts made using the laser-melt process of powder-bed additive manufacture** (3D printing). Material properties can be predicted from the microstructure helping to ensure that the parts will meet engineering requirements such as strength and fatigue life.

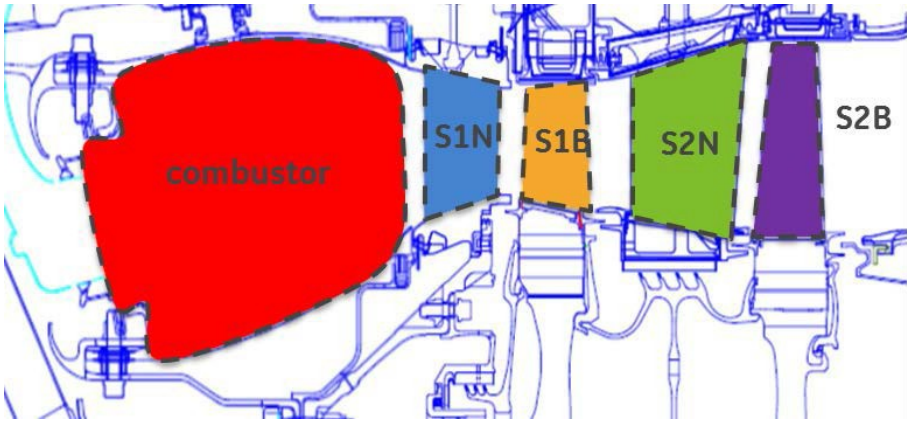
- Qualification of as-built parts is critical for meeting product reliability standards and hasten adoption of the new technology.

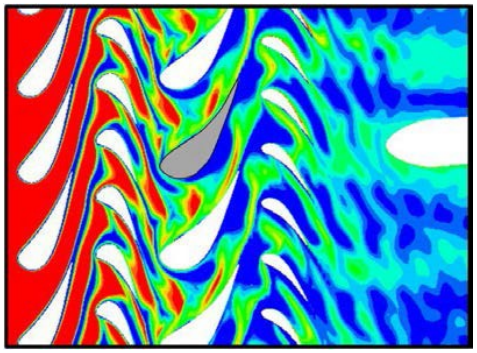


Improved engine designs can save fuel, reduce costs, and increase engine lifetime

HPC4 MANUFACTURING

- **Industry partner: GE Aviation**
- Over 35 billion gallons of fuel are used each year to transport airline passengers worldwide.
- New high-efficiency jet aircraft engines will be designed using **advanced turbulence calculations and models of the complete combustor/vane systems**

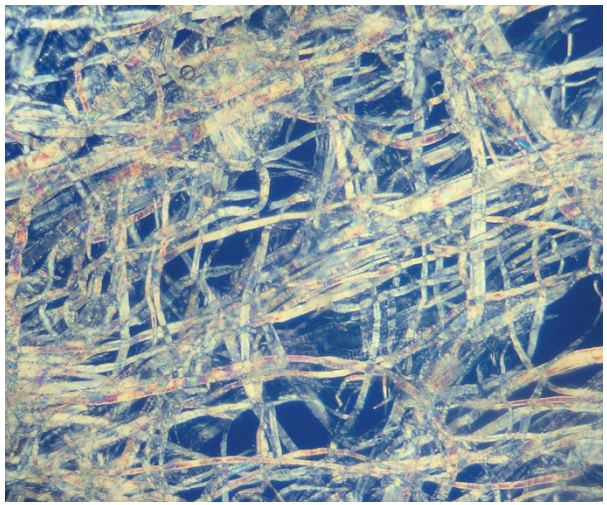




Modeling the microstructure of paper products can reduce paper pulp use by 20%

HPC 4 MANUFACTURING

- **Industry partner: Procter & Gamble**
- Papermaking is one of the most energy intensive industries due in large part to the energy needed to dry paper pulp.
- **Model microstructure of paper** for fiber orientation and texture to predict macro-properties for product optimization.
- Reducing paper pulp use by 20%, reducing energy consumption significantly.



LLNL Capabilities and Interests Summary

Primary Expertise & Interest Areas

- Computing architecture and programming models
- Materials modeling and experimental validation
- Application of HPC expertise into industry (HPC4Mfg)

Most Differentiating Factor

- Detailed approaches for memory integrated computing
- Co-design of advanced computing technology applied to materials modeling
- HPC4Mfg Program leadership

Main Contribution/Role

- Design, prototyping through emulation, evaluation of memory integrated computing structures into RRAM and other BMC devices
- Modeling of materials at multiple scales on advanced architectures
- Expanded framework to support broad industry engagement with national lab expertise (HPC4…)

Lawrence Livermore
National Laboratory