Never forecast the future (unless you are willing to be wrong)

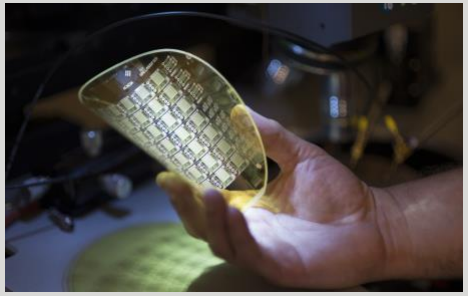
A view of the future of HPC

Outline

- Context
- Perspective
- Random thoughts

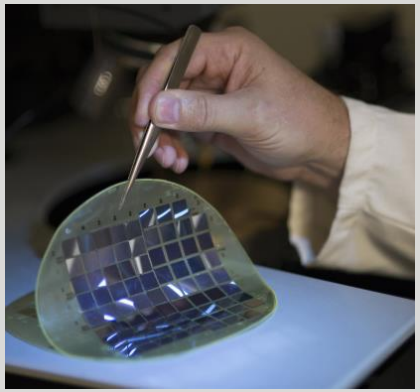
David J. Mountain, PhD
Senior Technical Director
NSA Advanced Computing Systems Research Program (ACS)
djmount@lps.gov

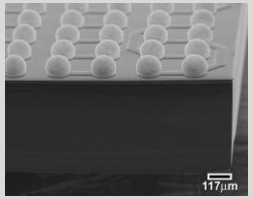
Who is Dave Mountain?

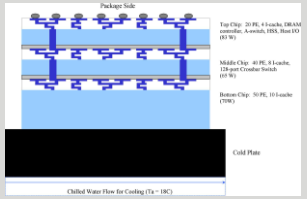


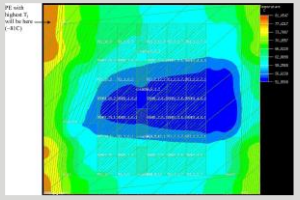
Dr. David J. Mountain received a BS in Electrical Engineering from the University of Notre Dame, an MS in Electrical Engineering from the University of Maryland, College Park, and a PhD in Computer Engineering from the University of Maryland, Baltimore County.

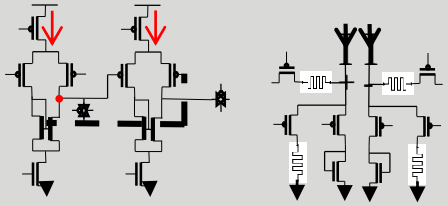
His personal research projects have included radiation effects studies, hot carrier reliability characterization, chip-on-flex process development utilizing ultra-thin circuits, and neuromorphic computing. He has been actively involved with 3D electronics research for 30 years.











He is the author of more than three dozen technical papers, has been awarded nine patents, and is a Senior Member of the IEEE.

What is ACS?

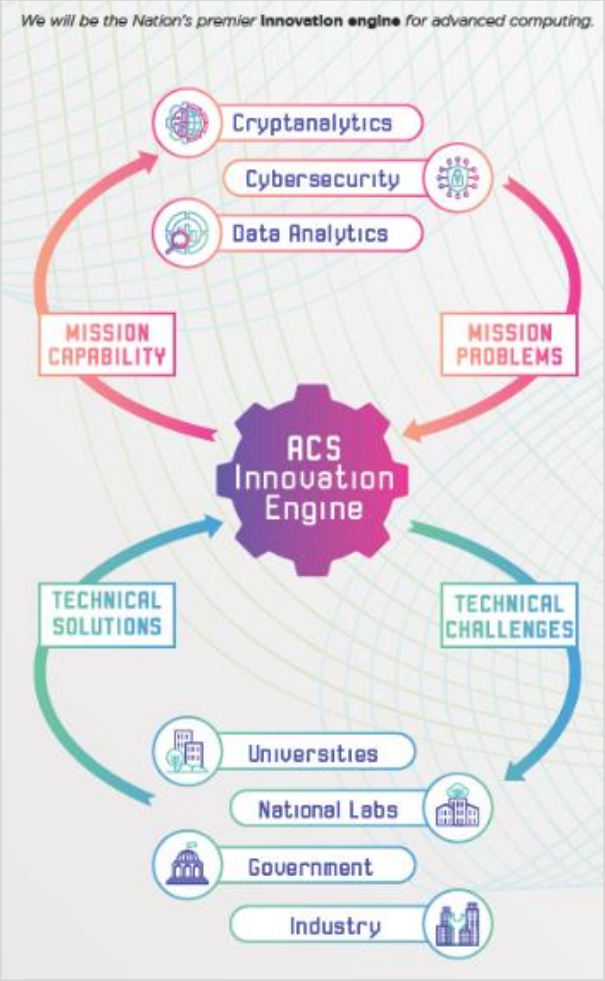
Advanced Computing Systems (ACS) Research Program

We will be recognized, both internally and externally, as the nation's premier **innovation engine** for advanced computing.

We conduct **exploratory research** that combines *algorithms, architectures and technologies* to demonstrate and/or develop advanced computing systems that provide **asymmetric advantage** for agency mission.

Our innovation engine is built upon **world class, mission oriented, participatory research.**

Resilience & Probabilistic Computing | **Productivity & Data Analytics** | **Modeling Simulation & Emulation** | **Computer Architecture & Engineering** | **Neuromorphic Computing** | **Energy Efficiency**



TODAY'S HPC ENVIRONMENT -- TECHNOLOGY

[Photo credit: iStock.com/danbaily/Jessica Marx]

It's not only merely dead, it's really most sincerely dead.
~Computer engineer who shall remain anonymous~

For 3^{+} decades we had huge, 'free' computing gains by transistor scaling – faster, lower energy, cheaper, consistent programming environment

1971: 4004 microprocessor -- 4 bit data path, 740 kilohertz clock, 10 µm
2005: Pentium -- 64 bit data path, 3.8 gigahertz clock, 90 nm

80,000x raw compute improvement (40 percent increase per year); primarily driven by a 12,000x reduction in transistor area.

Over the last 2 decades, we lost faster, lost a consistent programming environment, may have lost cheaper, and energy wins are greatly reduced

Significant ramifications for HPC

- Purpose built and co-design are necessary methodologies
- Accelerators everywhere and 'processors in everything' will happen

TODAY'S HPC ENVIRONMENT -- ECOSYSTEM

Large scale and high performance systems are dominated ($ spent) by cloud computing/hyperscalers

- What they want, vendors will provide
- They are not immune to transistor scaling ramifications[1,2]
  - In-house, purpose built designs and systems are proliferating;
    - both performance and $ wins for them

AI/ML is ubiquitous

- 'No programming' required
- Heavy on linear algebra – computationally dense
- Pushes large scale architectures in a particular direction

Vendor base is shrinking

- Advanced technology node fabs/foundries
- System integrators
- SW providers

cadence
SYNOPSYS
groq
cerebras
Esperanto Technologies
MYTHIC
natural INTELLIGENCE
parallella
brainchip
WAVE COMPUTING

1. H. Dattatraya et al, "Silent Data Corruptions at Scale," arXiv:2102.11245v1, Feb 2021.
2. P. Hochschild et al, "Cores that don't count," Workshop on Hot Topics in Operating Systems, 2021, Ann Arbor, MI.

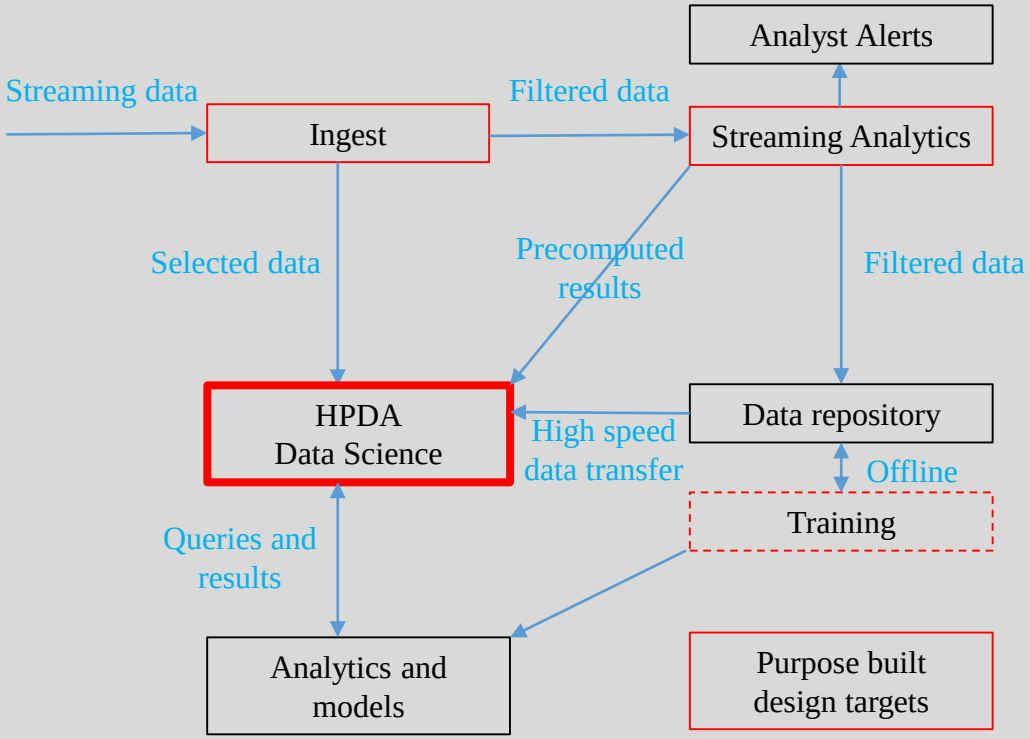
LOOKING AT HOW NEW IDEAS CAN CREATE INNOVATIVE HPC SYSTEMS

Primary factors[1]
- Ops/W
- Ops/$
- Risk (time/cost to design, integrate, and effectively use)

Primary constraints
- Ops needed for the mission
- Available power and cooling (maybe area)
- Mission value
- Programmer effort

Leverage what commercial entities need/do well
Focus our efforts on what we must do

Commercial → Purpose Built → Highly specialized



Overall data and work flow for an analytical mission

1. D. Mountain, "Analyzing the Value of Using Three-Dimensional Electronics for a High-Performance Computational System," IEEE Transactions on Advanced Packaging, February 2008.

LOOKING AT HOW NEW IDEAS CAN CREATE INNOVATIVE HPC SYSTEMS

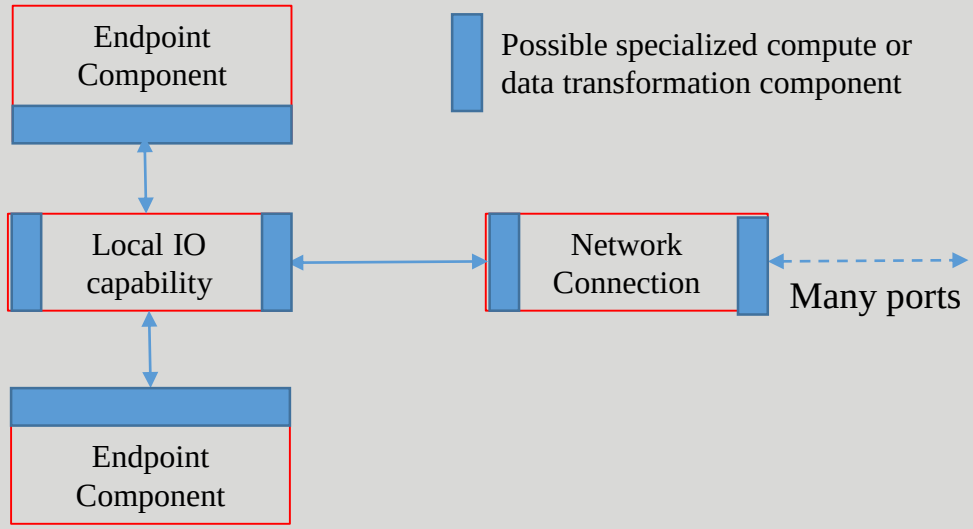
Advanced packaging

- 2.5D, 3D, WSI, and photonics tend to drive off-chip performance wins (@ constant power with uncertain $ implications)
  - Technical concern is trade-offs of power, data, cooling (two surfaces and three needs)[1]
  - System concern is the scaling limitation of the performance win (imbalance of memory BW/capacity)

- Chiplets tend to drive $ and energy wins (@ constant performance)
  - Real opportunity to reduce risk through re-use of validated IP
  - Real opportunity to enable purpose built designs
  - Real risk of poor choices constraining HPC opportunities

How does HPC leverage where commercial industry will take these?

- Encourage the 'right' modularization and standardization
  - Moving data – standard
  - Easy swap of electrical IO with optical IO
  - Transforming data – specialized modules for purpose built design
    - Standards on where they can be placed?
  - Test/diagnose/repair – possible point of convergence with industry
- Find 'one team, one mission' opportunities to modestly increase USG voice
  - Project 38[2]

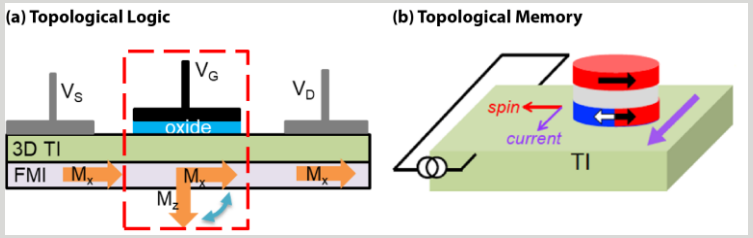


Notional purpose built architecture for an HPDA system

1. D. Mountain, "Opportunities and Challenges for 3D Technology in Advanced Computing Systems," VLSI Multilevel Interconnect Conference (VMIC 2008), Fremont, CA, October 2008.
2. https://www.nitrd.gov/nitrdgroups/images/b/b4/NSA_DOE_HPC_TechMeetingReport.pdf

SOME THINGS 'JUST' OVER THE HORIZON

New materials, devices, and functions require attention and discernment[1]



Neuromorphic computing is in its infancy

Silicon Brain not required

Image: zmescience.com

Quantum information science will drive certain technologies

Sensors, QC, SCE accelerators, cryo-computing and IO, advanced cryo-packaging

Heterogeneous integration will need to include multiple temperature domains

mK, 4K, 77K, 300K – modularity needed

1. S. Manipatruni et al, "Scalable energy-efficient magnetoelectric spin-orbit logic," Nature, Dec. 2018.

WAVING A MAGIC WAND …

Future Advanced Computing Ecosystem (FACE)[1]

People and skills

Research testbeds and prototypes for ‘playing around’ and applications/SW development

Access to leading edge technology ecosystems

Data sets and computational neuroscience

Security – HPC + Big Data, supply chain

1. The FACE strategic plan can be downloaded at https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf


Laboratory for Physical Sciences
LPS