NSF Plans and Roadmaps for HPC

John Towns
Deputy Director, NCSA
University of Illinois at Urbana-Champaign
jtowns@illinois.edu

["C Team" presenter; some content stolen from Dan Stanzione and Katie Antypas]

Disclaimers

I am not Katie Antypas (NSF Office Director, Office of Advanced Cyberinfrastructure)

- I do not represent NSF's views; did steal a few slides from her (with permission)

I am not Dan Stanione (TACC Director)

- [though we are sometimes confused… mostly because of the hair]
- I did steal some slides from Dan (with permission)

The situation in the United States is... "fluid"

- What I say may be changing as I talk…

Opinions expressed are my own and not those of the University of Illinois, NSF, or US Federal Government

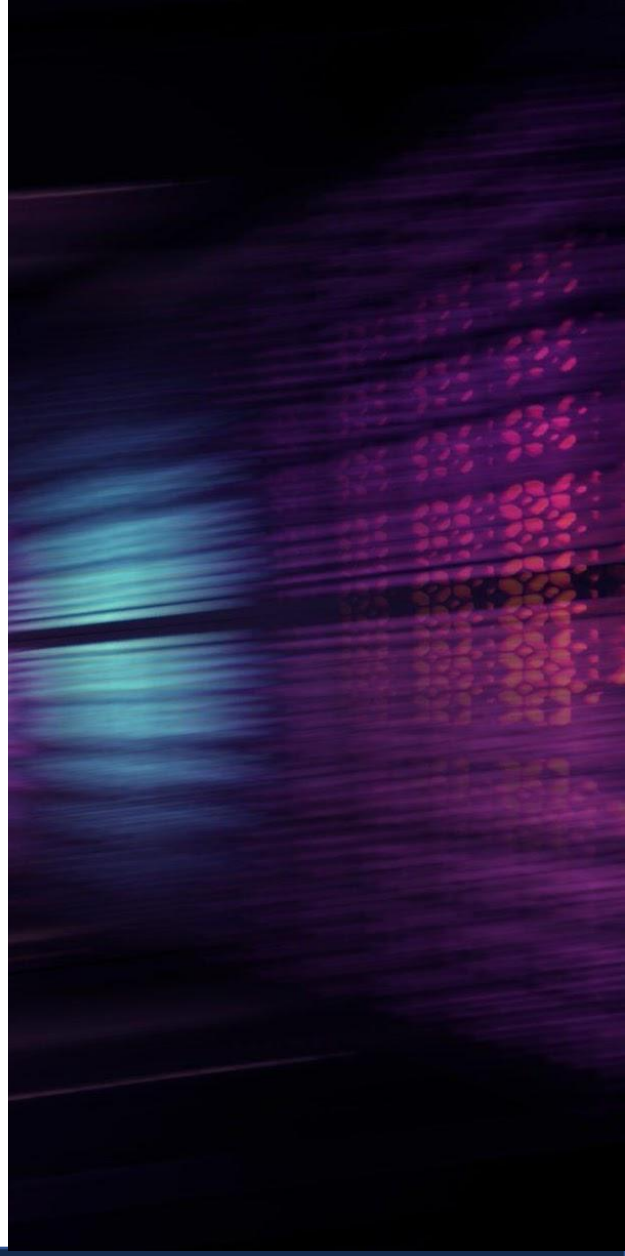
Did I mention that things are fluid?

US Federal Government has:

- issued many Executive Orders and related missives
- induced federal agencies to cut grant funding and indirect cost rates aggressively
- cracked down on free speech at universities
- created lots of chaos

US academic institutions:

- being targeted by Federal Government
- have had funding cut or suspended
  - many individual grants have been eliminated or have had their budgets cut
- largely do not know how to act in current circumstances



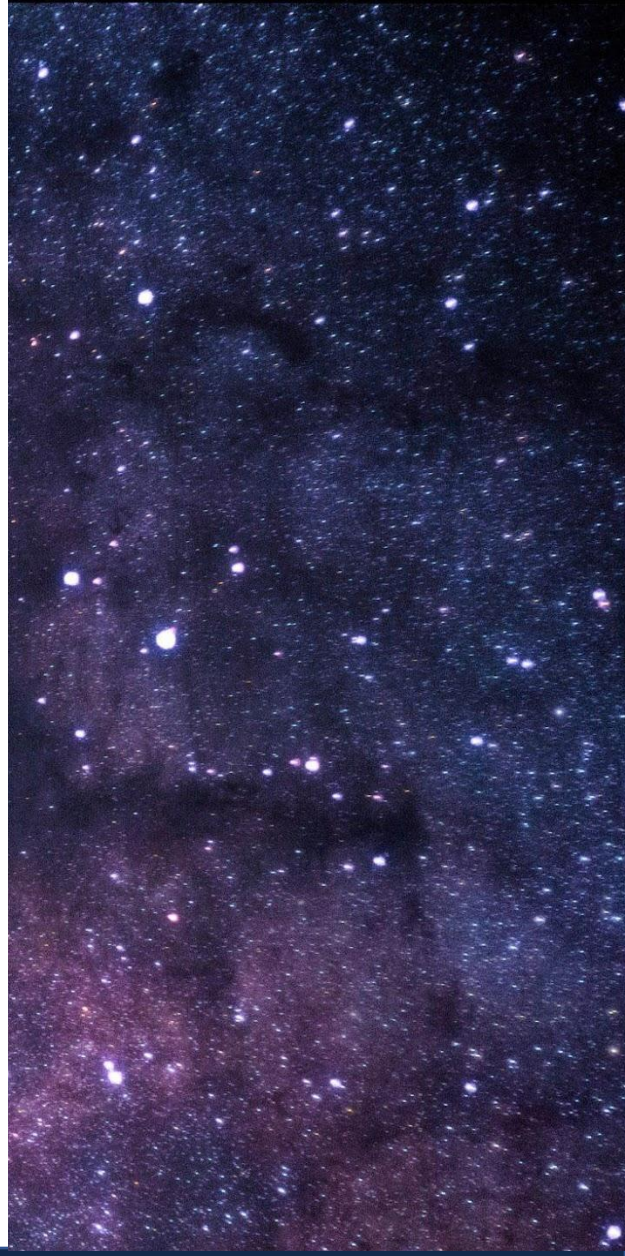
NSF Approach

NSF is not a mission agency

- attempts to support very broad range of scholarly effort

NSF struggling with current environment (as are most agencies)

- budget uncertainties
- rapid emergence of AI everywhere
- rapid evolution of computing technologies; quantum computing in particular
- science applications struggling to adapt quickly enough
    - thus ongoing need for more traditional resources



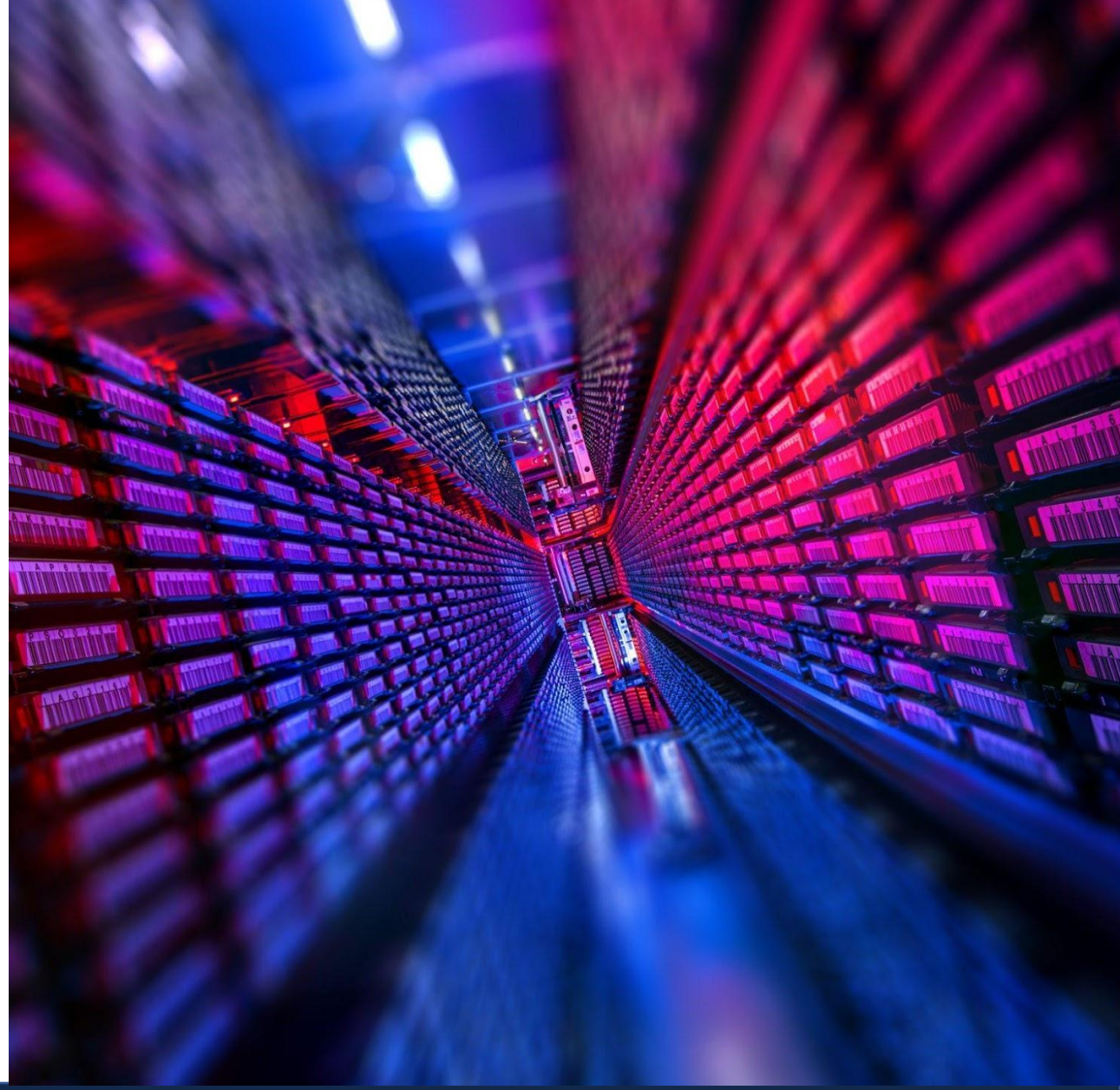
NSF Computing Effort

Various programs

- one major compute system
  - Leadership Class Computing Facility (LCCF); under construction
- ACCESS program and associated compute resources
- emerging quantum computing investments
- bunch of other stuff

Balancing these is a challenge

- budget uncertainties
- policy makers are only thinking about AI and quantum
- vendors have largely the same focus in mostly segregated groups



NSF Program Areas in HPC+

LCCF

- NSF's closest thing to a large-scale resource
- dwarfed by DoE resources
- [more in a moment]

ACCESS and associated resources

- successor to TeraGrid/XSEDE
- provides access to a variety of capacity systems with a shared allocations process
  - US$5-20M acquisition costs + operations
  - https://access-ci.org/
- year long allocations, large allocations reviewed twice per year
- currently 17 hardware Resource Providers
  - CPUs, GPUs, and a mix of other things

Quantum is an effort from Mathematical and Physical Sciences directorate

- National Quantum Virtual Laboratory awards

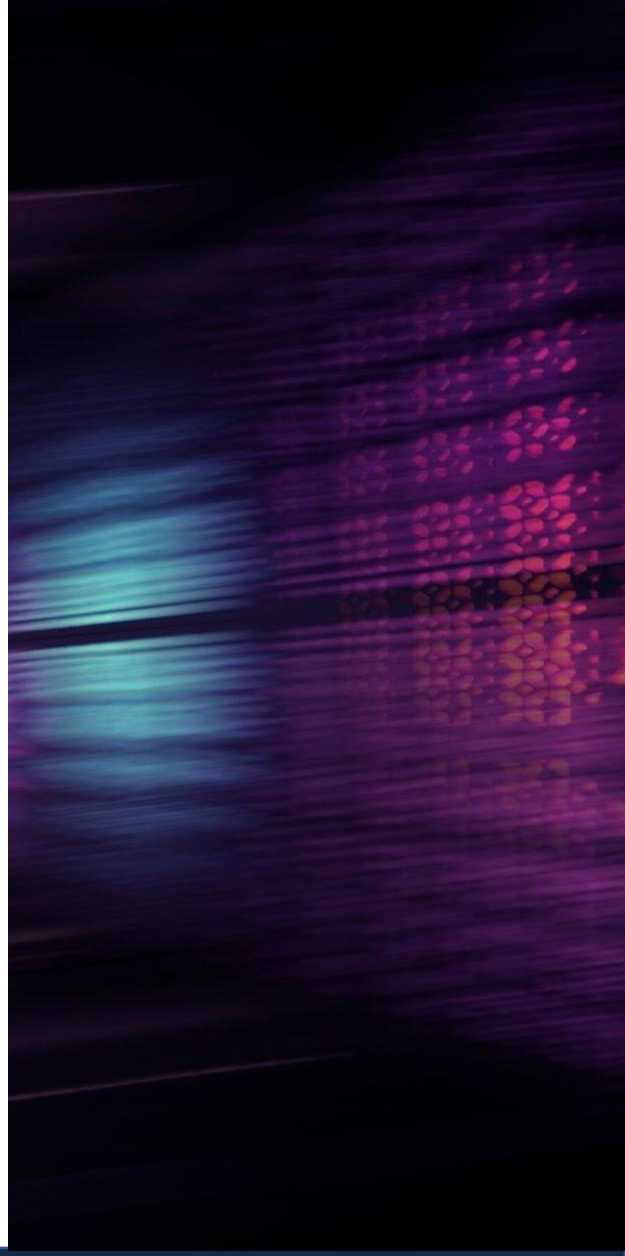
Other NSF Programs

National Research Platform
- distributed network of small (~8 node) GPU/CPU clusters with distributed data storage, across dozens of sites.

Cloud Testbed Program
- Cloudlab, Chameleon
  - computing research/cloud research platforms

CloudBank
- shared Portal for access to commercial providers, with NSF-allocated credits.

NCAR
- computing for climate/weather research



LEADERSHIP-CLASS COMPUTING FACILITY

- The LCCF (Leadership Class Computing Facility) elevates computing at NSF for the first time to be on a par with other facilities.
  - While open to all open science users, other large facilities or a core audience!
  - Construction phase began in July.
- The first system, Horizon, will become available in early 2026
  - Datacenter completion in October 2025, rack delivery will begin then.
- The project includes, buildings, systems, and importantly, *people* to support them. , including software optimization and tuning.

TACC
TEXAS ADVANCED COMPUTING CENTER

NSF

DISTRIBUTED CENTERS

- The LCCF Hardware (and staffing) will not only be at TACC, but also at four other sites around the country. (Through construction and operations).

- NCSA --
  Focus on applications using accelerators and Quantum
- SDSC --
  High throughput, and HT Inference for large scale scientific Instruments

PSC –
Focus on storage systems (and data rep site)

AUCC --
Accessibility, Workforce, interactive systems

A LITTLE MORE ON HORIZON

- Primary compute capability will be :
  - ~400 PF double precision (10x Frontera)
  - >10 ExaFlops bfloat16 precision.
- Solid-state storage capacity:
  - ~0.5 Exabytes. == 25x bandwidth of Frontera (8TB/s Write; 16TB/s Read)
- Roughly a million cores of CPU, roughly 4k GPUs.
  - NVIDIA Grace-Blackwell 2,000 nodes, 800Gbps Infiniband
  - NVIDIA Vera CPU, 4,750 nodes, 400Gbps Infiniband
- Some interesting opportunities in data at this scale – the storage controllers alone will have >75k additional cores directly attached to the flash drives we could do analytics on.
- Additional nodes for:
  - Interactive computing (e.g. Jupyter)
  - Persistent services (Inference services, Gateways, API instances, "serverless" functions, etc.).

What is ACCESS?

NSF program to

- help researchers and educators utilize advanced computing systems and services
- support science applications that requires more than a desktop or laptop
  - many domains
  - most now incorporating AI in some way

ACCESS: Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support

- US$52 million award for five years to five lead institutions and their sub-awardees to facilitate the ACCESS program
- does NOT include funding for hardware resources and direct support of them

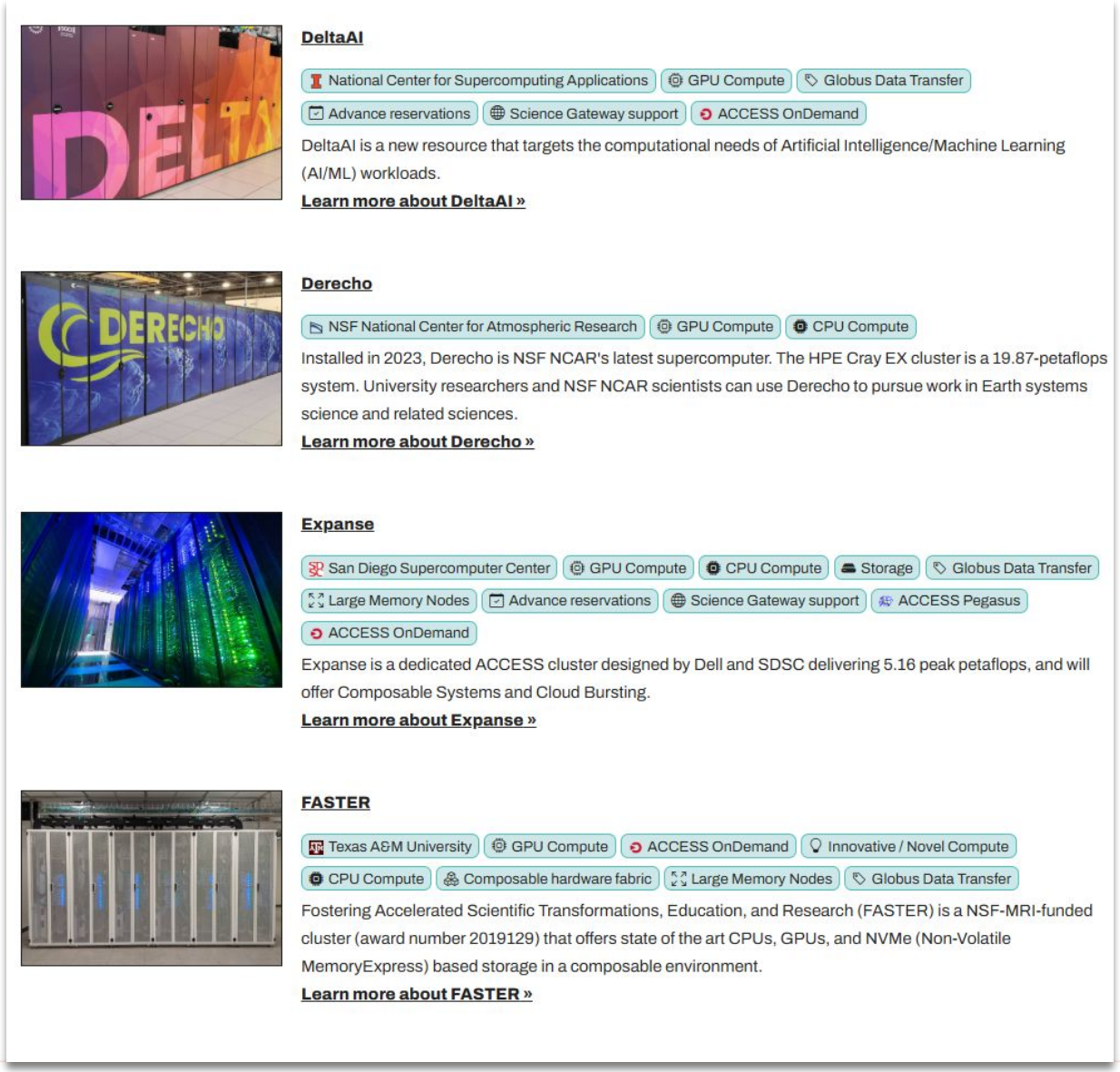
https://access-ci.org/about/organization/

ACCESS Resources

Resource list at:
https://allocations.access-ci.org/resources

Major resources:

- Expanse at SDSC
- Bridges2 at PSC
- Delta/DeltaAI at NCSA
- Stampede3 at TACC
- JetStream2 at IU/TACC
- Anvil at Purdue



NAIRR Vision: A national research infrastructure to drive US AI innovation, discovery and national competitiveness

National goals

- Accelerate AI and AI-powered discovery and innovation.
- Expand the US AI R&D workforce and train the next generation of AI researchers and educators.
- Increase integration and use of world-class public and commercial AI resources.
- Advance public trust in AI

Envisioned NAIRR Architecture