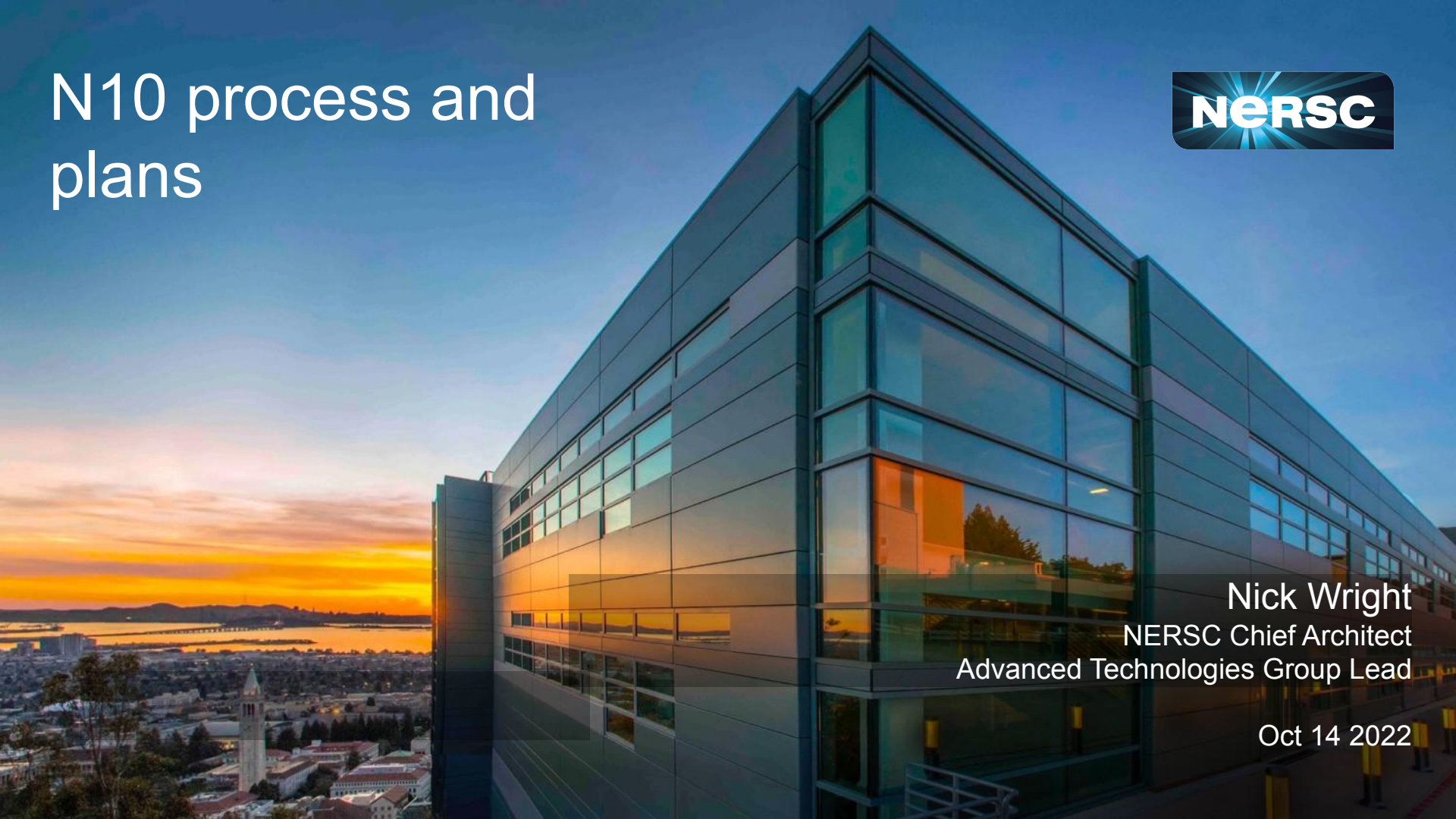
N10 process and plans



Nick Wright
NERSC Chief Architect
Advanced Technologies Group Lead

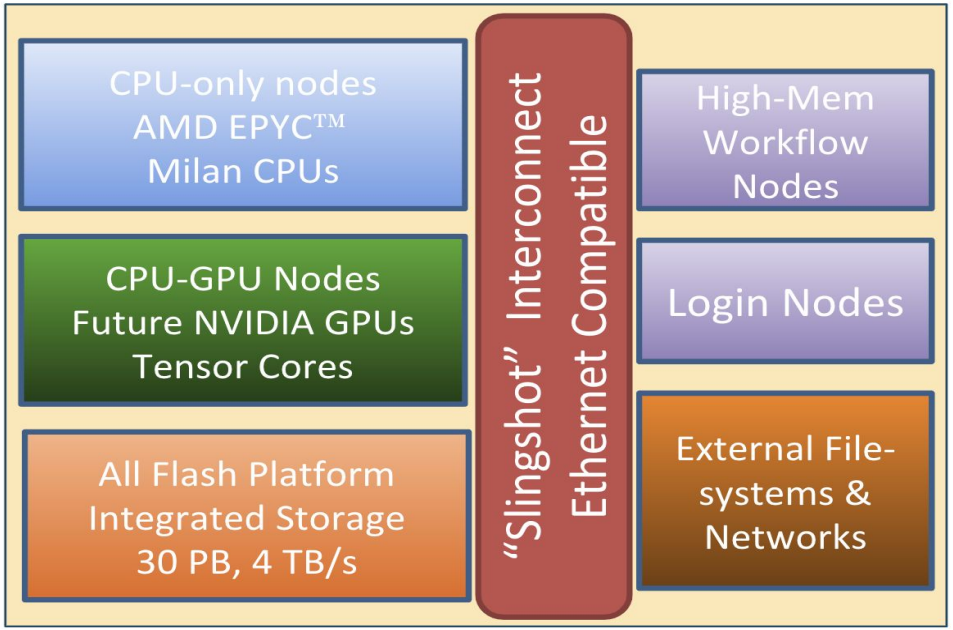
Oct 14 2022

NERSC has a dual mission to advance science and the state-of-the-art in supercomputing

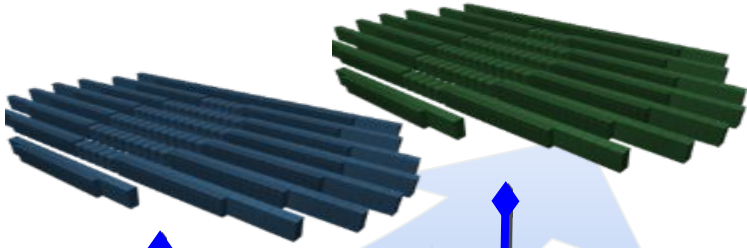
- We collaborate with computer companies years before a system's delivery to deploy advanced systems with new capabilities at large scale
- We provide a highly customized software and programming environment for science applications
- We are tightly coupled with the workflows of DOE's experimental and observational facilities – ingesting tens of terabytes of data each day
- Our staff provide advanced application and system performance expertise to users

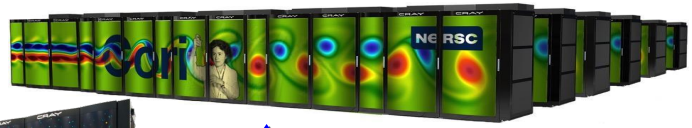
Perlmutter: A System Optimized for Science

- GPU-accelerated and CPU-only nodes meet the needs of large scale simulation and data analysis from experimental facilities
- Cray “Slingshot” - High-performance, scalable, low-latency Ethernet-compatible network
- Single-tier All-Flash Lustre based HPC file system, >6x Cori’s bandwidth
- Dedicated login and high memory nodes to support complex workflows



NERSC Systems Roadmap





NERSC-7: Edison Multicore CPU

2013

NERSC-8: Cori Manycore CPU
NESAP Launched: transition applications to advanced architectures

2016

NERSC-9: CPU and GPU nodes
Continued transition of applications and support for complex workflows

2021

NERSC-10: Exa system

2026

NERSC-11: Beyond Moore

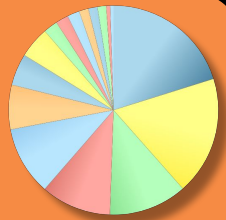
2030

Increasing need for energy-efficient architectures

NERSC's approach to strategic planning

User Requirements

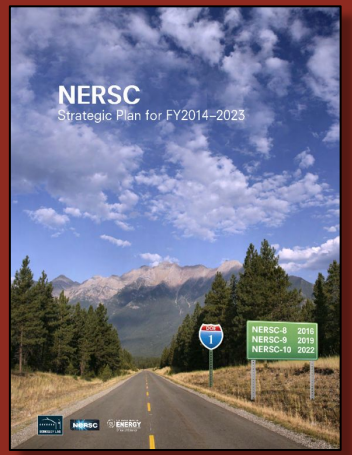
Workload Analysis



Technology Trends

HGST a Western Digital brand
DDN STORAGE
SAMSUNG
Micron
intel
NVIDIA
IBM
DELL
Hewlett Packard Enterprise
SEAGATE

NERSC Strategy



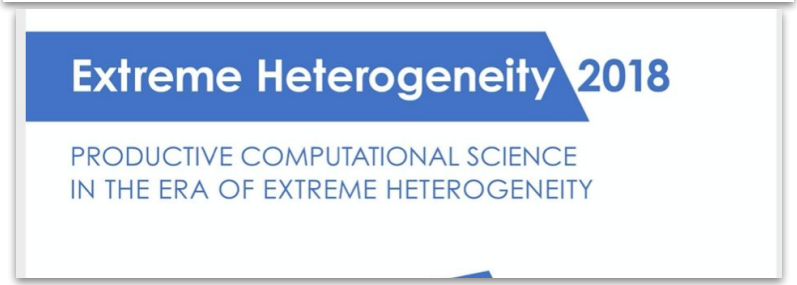
End of Moore's Law ?


DOI:10.1145/3282307

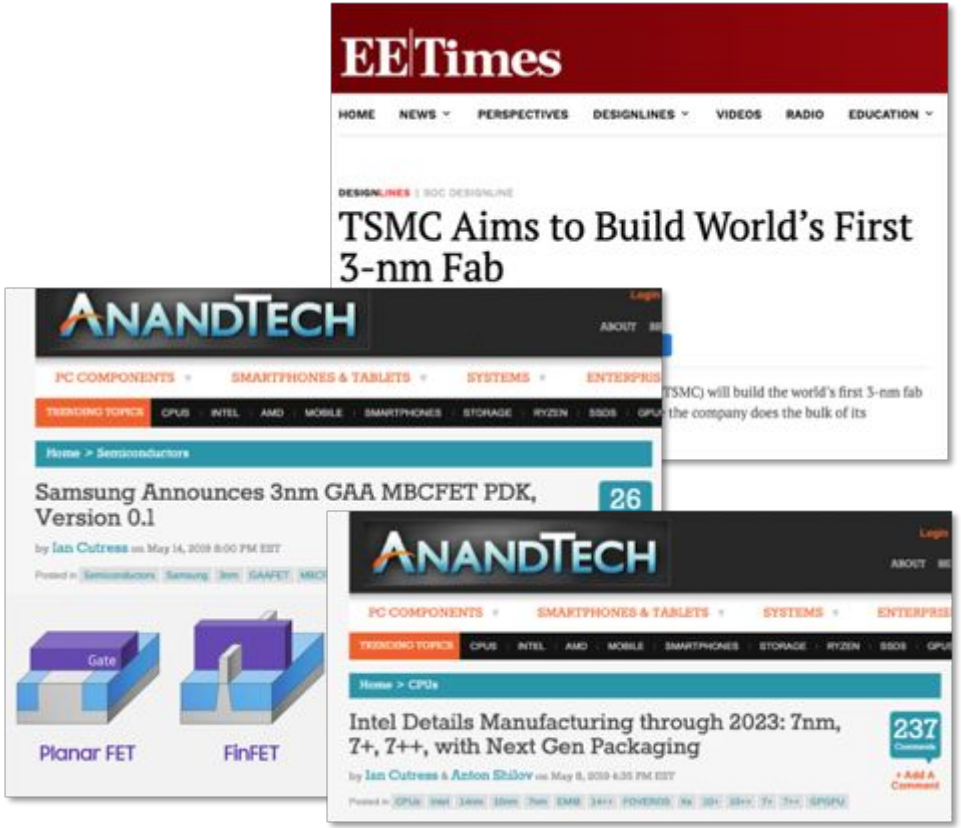
Innovations like domain-specific hardware, enhanced security, open instruction sets, and agile chip development will lead the way.

BY JOHN L. HENNESSY AND DAVID A. PATTERSON

A New Golden Age for Computer Architecture


Extreme Heterogeneity 2018

PRODUCTIVE COMPUTATIONAL SCIENCE
IN THE ERA OF EXTREME HETEROGENEITY



Other Bookmarks

CNBC

Search quotes, news & videos

WATCHLIST | SIGN IN

MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV INVESTING CLUB PRO

TECH

Intel says Moore's Law is still alive and well. Nvidia says it's ended.

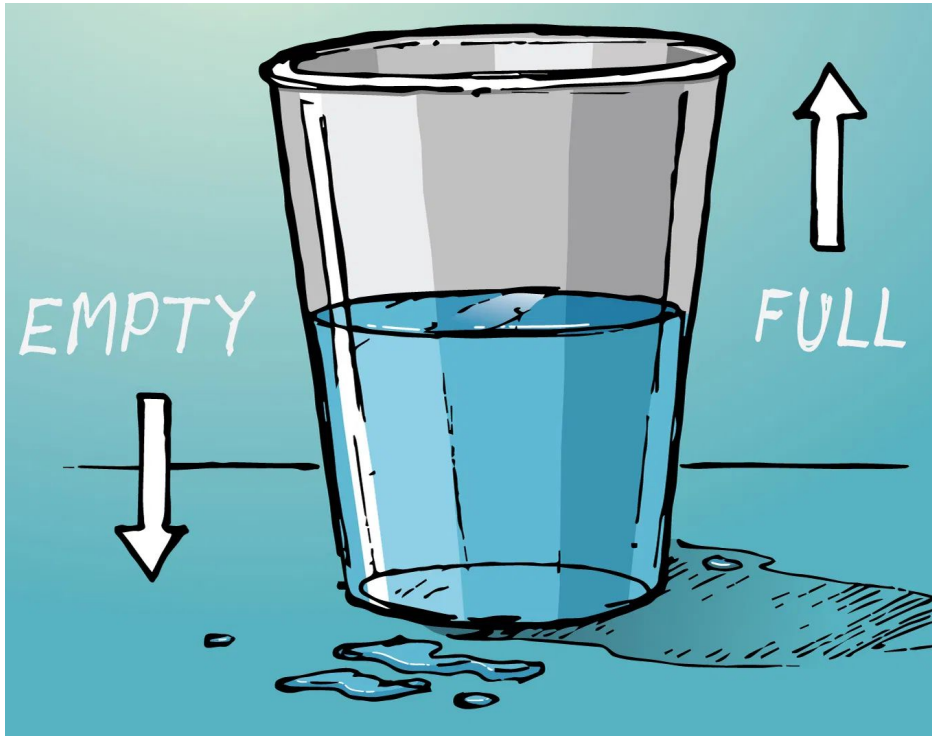
PUBLISHED TUE, SEP 27 2022·3:26 PM EDT

Kif Leswing
@KIFLESWING

WATCH LIVE

KEY POINTS

- Intel CEO Pat Gelsinger said on Tuesday at a company launch event that Moore's Law is "alive and well."
- Nvidia CEO Jensen Huang said last week Moore's Law has ended.
- Intel has committed to continue manufacturing some of its chips, while Nvidia relies entirely on third-party foundries for its production.



Introducing Intel's new node naming

10nm SuperFin

- In high-volume production

Intel 7

Previously referred to as Enhanced SuperFin

- 10-15% perf/watt gain
- FinFET transistor optimizations
- Now in volume production

Intel 4

Previously referred to as 7nm

- 20% perf/watt gain
- Full use of EUV lithography
- Meteor Lake for client tape in Q2 2021
- Granite Rapids compute tile for data center

Intel 3

Power and area improvements

- 18% perf/watt gain
- Denser HP library
- Increased intrinsic drive current
- Reduced via resistance
- Increased EUV use
- Manufacturing products 2H 2023

Intel 20A

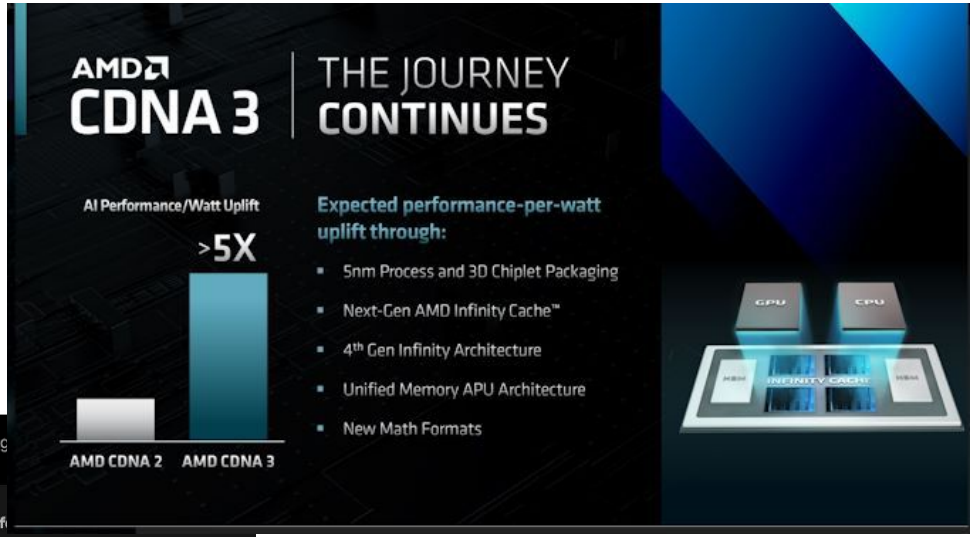
The angstrom era of semiconductors

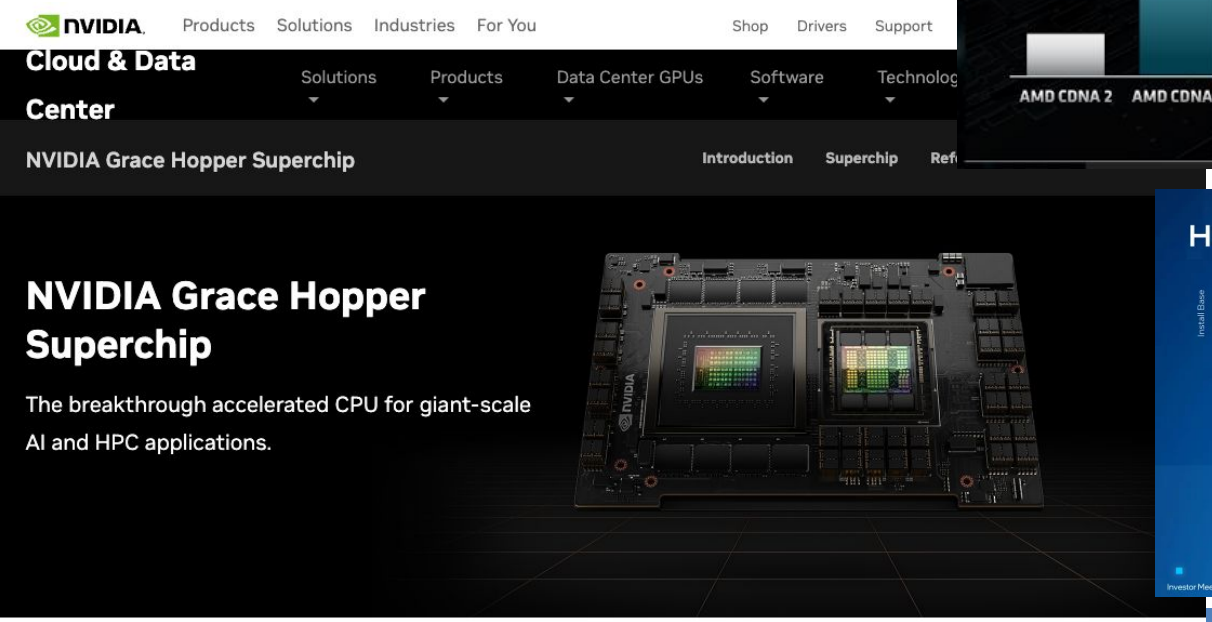
- Breakthrough innovations in 1H 2024
- RibbonFET – new transistor architecture
- PowerVia – interconnect innovation

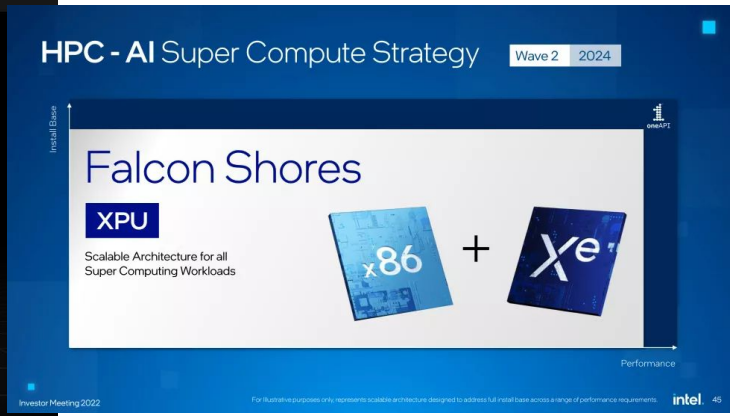
Embargoed until Monday, July 26, 2 p.m. PT

intel
accelerated

More Tightly Coupled CPU-GPU







Technology Trends Summary

- No more increases in clock speed for CPUs & GPUs
  - More & more cores
- Increases in performance will primarily be obtained through power increases
  - At the socket & the system level
- Tighter & Tighter CPU-GPU integration
  - Grace-Hopper from NVIDIA
  - MI-300 from AMD
- Flash Storage will continue to increase in capacity and eat into HDD space

What do we expect N10 to look like?

	Perlmutter	NERSC-10 Improvement
Aggregate Performance	1	10x
Peak Power	~6 MW	~3x
CPU	64 cores	~2 x
GPU	~20 TF	2-3 x
Interconnect	25 GB/s/link	2 - 4x
Number GPUs per node	4	1 ?
Number of Nodes	1,536 GPU + 3,072 CPU	> 10x
Storage	35 PB, >5 TB/s Lustre FS	> 5x Capacity spread over Lustre & reconfigurable storage

What are the implications for NERSC users?

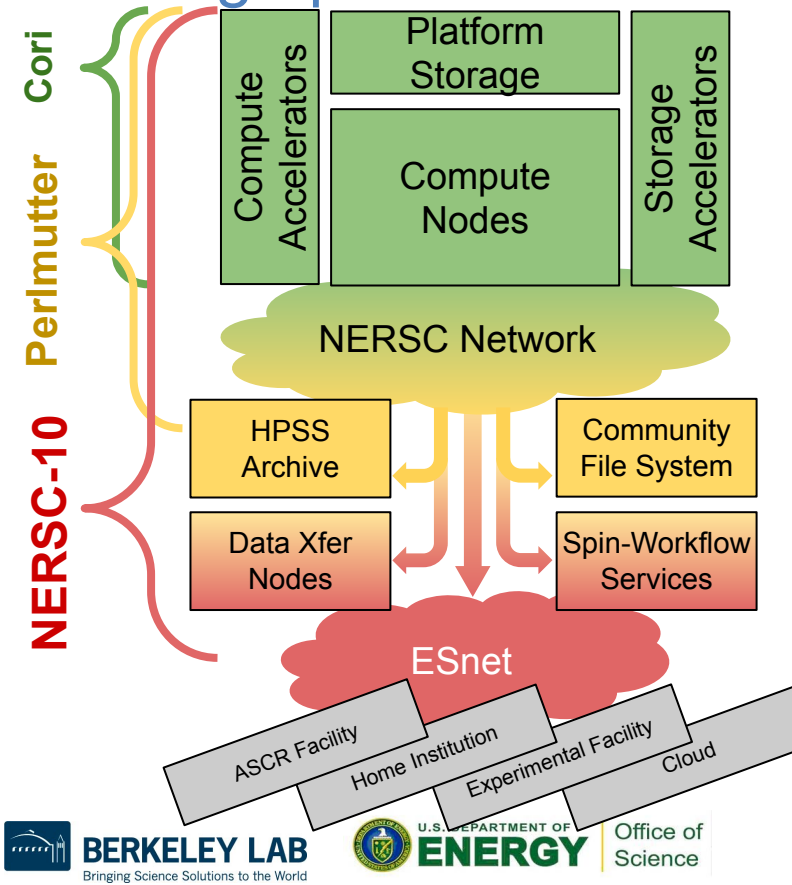
- Applications that don't use GPU's should try to !
- Applications that use GPUs on Perlmutter will run on N10 with little to no modifications
- Will need to express more parallelism
  - ~2x per CPU
  - ~2-4x per GPU
  - Will need (at least) 4x more MPI ranks to use the same fraction of the machine
- If you can - consider modifying your application/algorithm to exploit lower precision
-

NERSC-10 Architecture: Designed to support complex simulation and data analysis workflows at high performance

NERSC-10 will provide on-demand, dynamically composable, and resilient workflows across heterogeneous elements within NERSC and extending to the edge of experimental facilities and other user endpoints

Complexity and heterogeneity managed using complementary technologies

- **Programmable infrastructure**: avoid downfalls of one-size-fits-all, monolithic architecture
- **AI and automation**: sensible selection of default behaviours to reduce complexity for users



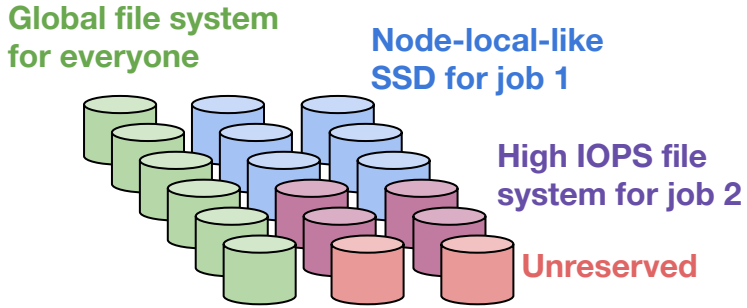
Reconfigurable storage tailors performance to each workflow's characteristics and needs

NERSC-10 will be programmable to optimize for each workflow

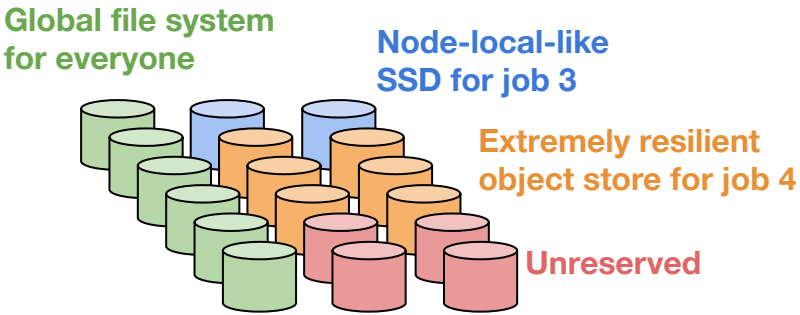
1. User requests hardware resources, connections between them, and data placement
2. System schedules CPU, accelerators, storage, networking, and data movement
3. Same resources are later reconfigured to adapt to new requirements

NERSC-10 will achieve this by embracing technology trends

- Disaggregated, software-defined infrastructure to connect heterogeneous components
- AI and automation to manage
  - complexity of scheduling and operations
  - data movement between reconfigurations
  - complexity for users - sensible defaults



Later that day...



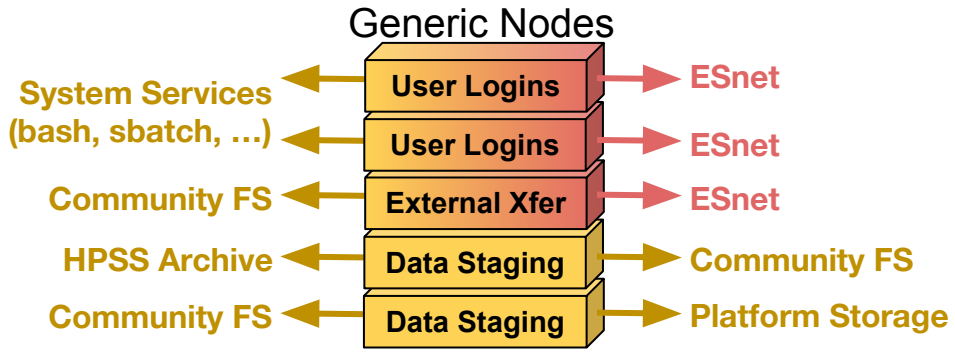
Pools of nodes and bandwidth can be reconfigured to support different SC workflows

Software-defined networking redirects bandwidth to paths that need it

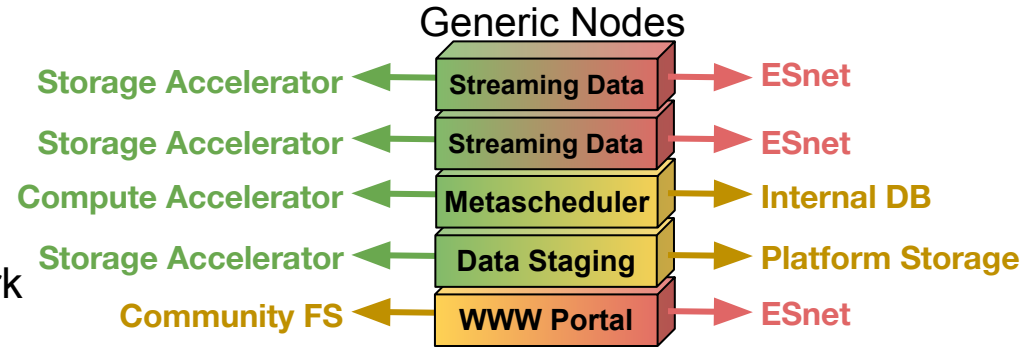
Microservices allow services that utilize bandwidth to scale up/down

One hardware pool configurable to...

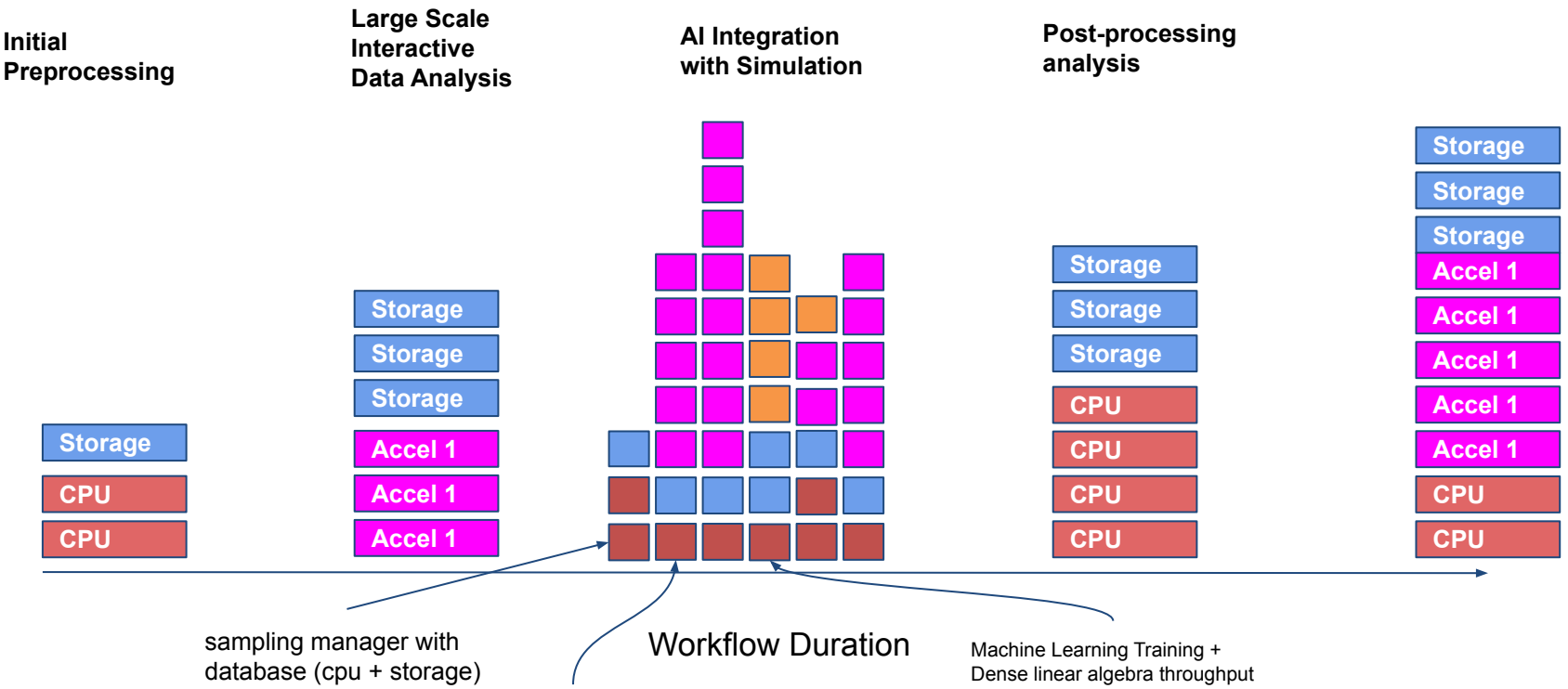
- **DTNs** - file transfer from external facilities
- **Routers** - stream data directly to compute
- **Movers** - file transfer between storage tiers
- **Metaschedulers** - dispatch units of work to compute



Later that day...

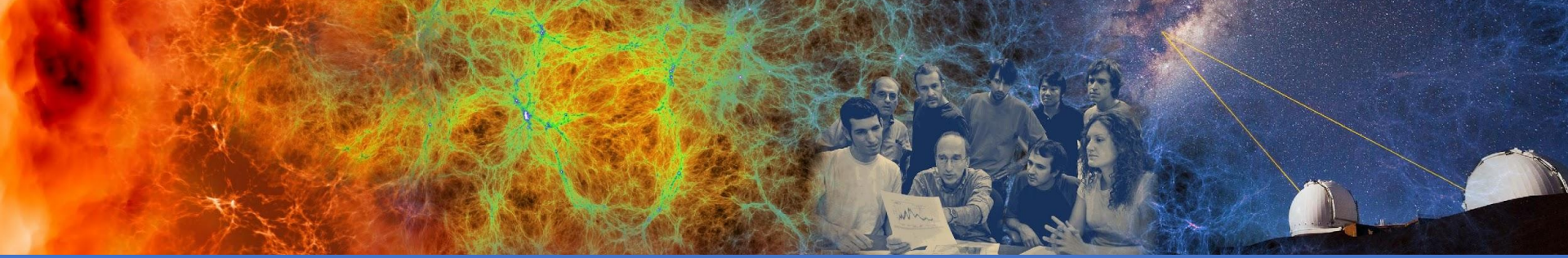


Flexible and dynamic scheduling of compute, storage and bandwidth enables a workflow to reserve various resources at different times.



Summary

- NERSC-10 will be 10x the performance of Perlmutter
- GPU-enabled applications should have minimal issues in porting/running their applications
- Currently NERSC is planning to release the NERSC-10 RFP in CY-23 for delivery in 2026
- If you are not running on GPUs yet let us know why !
- We are always interested in hearing from users !
  - Fill out the user survey !! What can we do better?



Questions ?