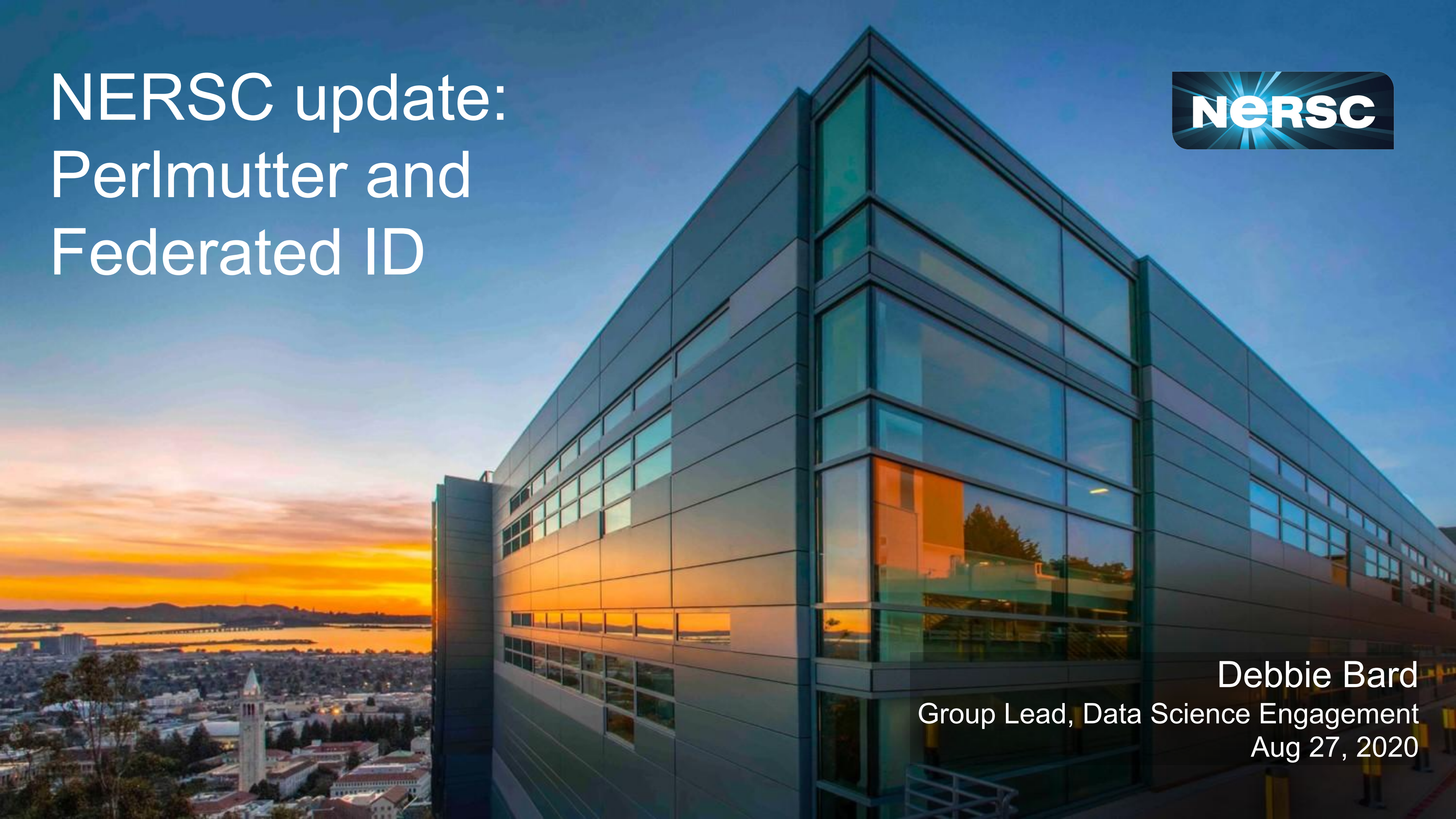
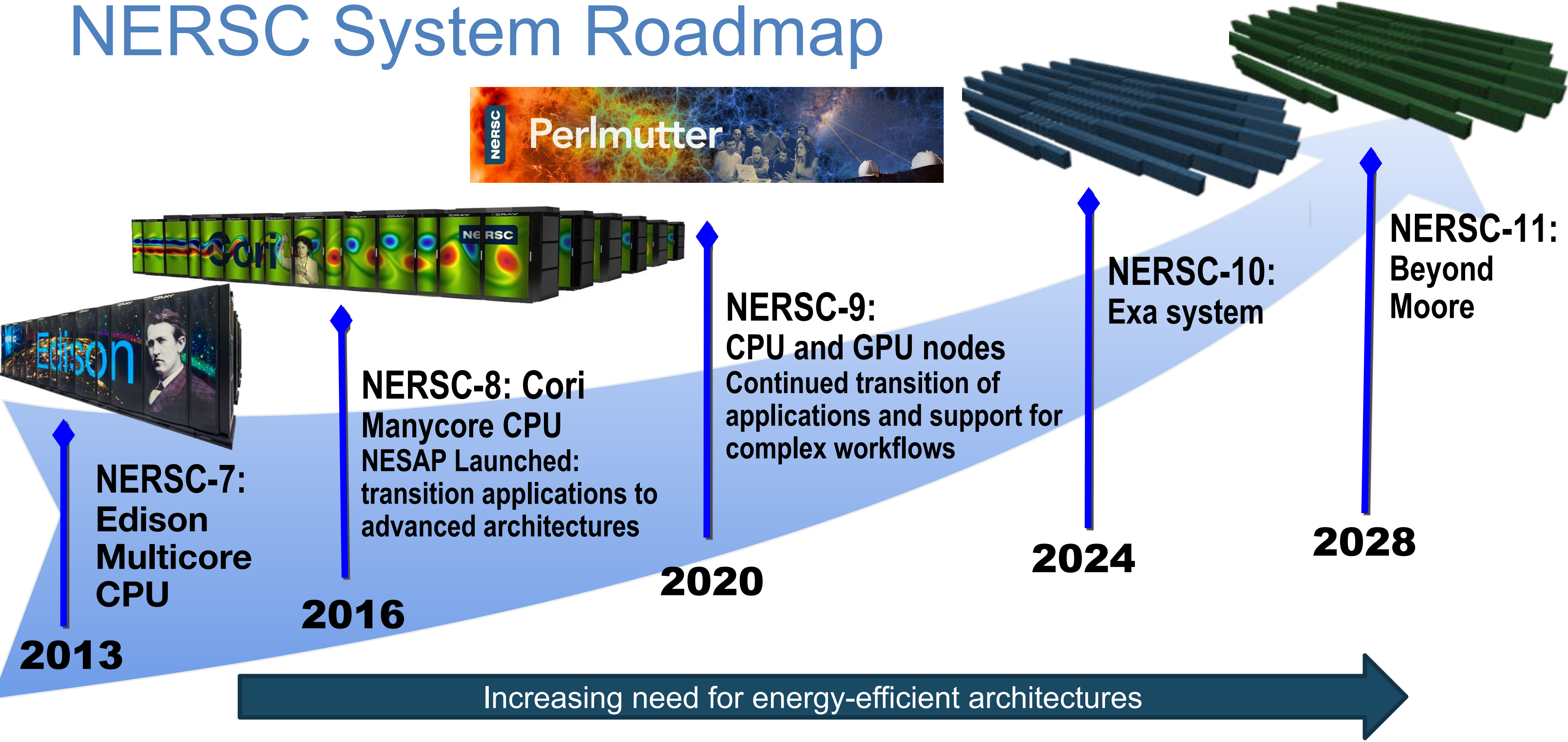
NERSC update: Perlmutter and Federated ID



Debbie Bard
Group Lead, Data Science Engagement
Aug 27, 2020

NERSC System 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

2020

NERSC-10: Exa system

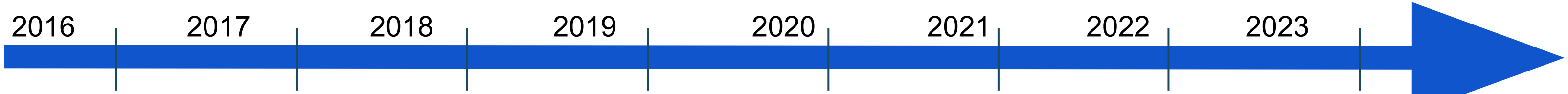
2024

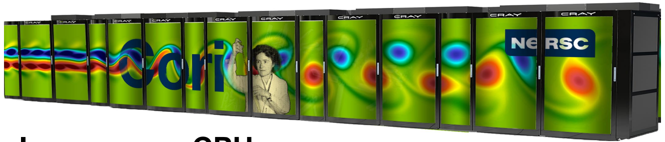
NERSC-11: Beyond Moore

2028

Increasing need for energy-efficient architectures

DOE HPC Roadmap - GPUs



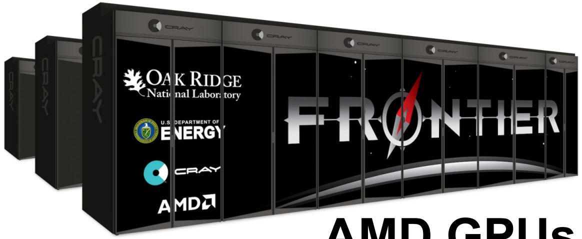


Intel many-core CPU

NVIDIA Ampere

Intel GPUs

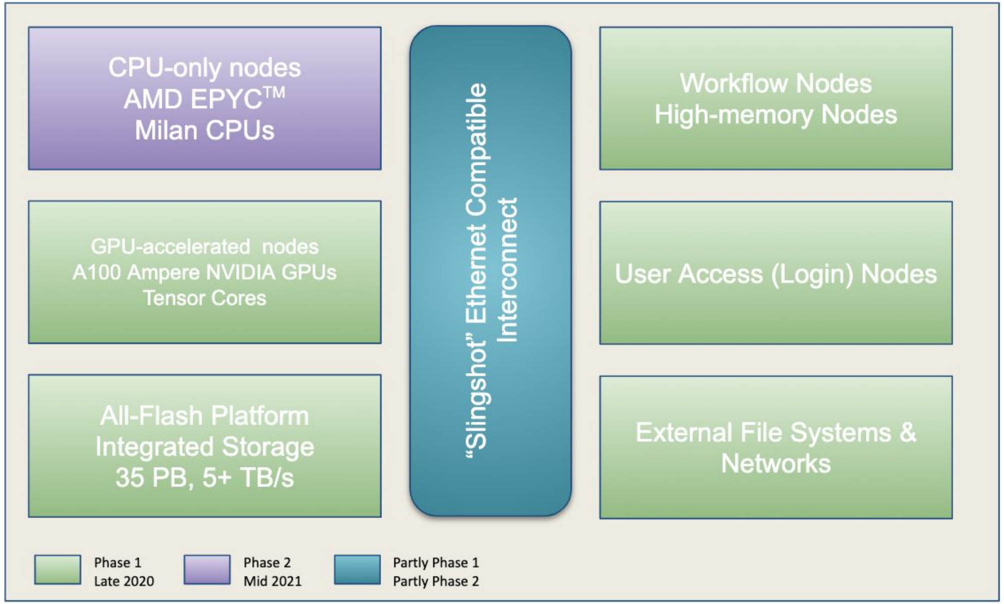
NVIDIA Volta GPUs



AMD GPUs

Perlmutter: a System Optimized for Science

- AMD/NVIDIA **A100-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
  - seamless connection between inside/outside the machine
- Single-tier **All-Flash Lustre** HPC file system, 6x Cori’s bandwidth
- Dedicated login and high memory nodes to support complex workflows



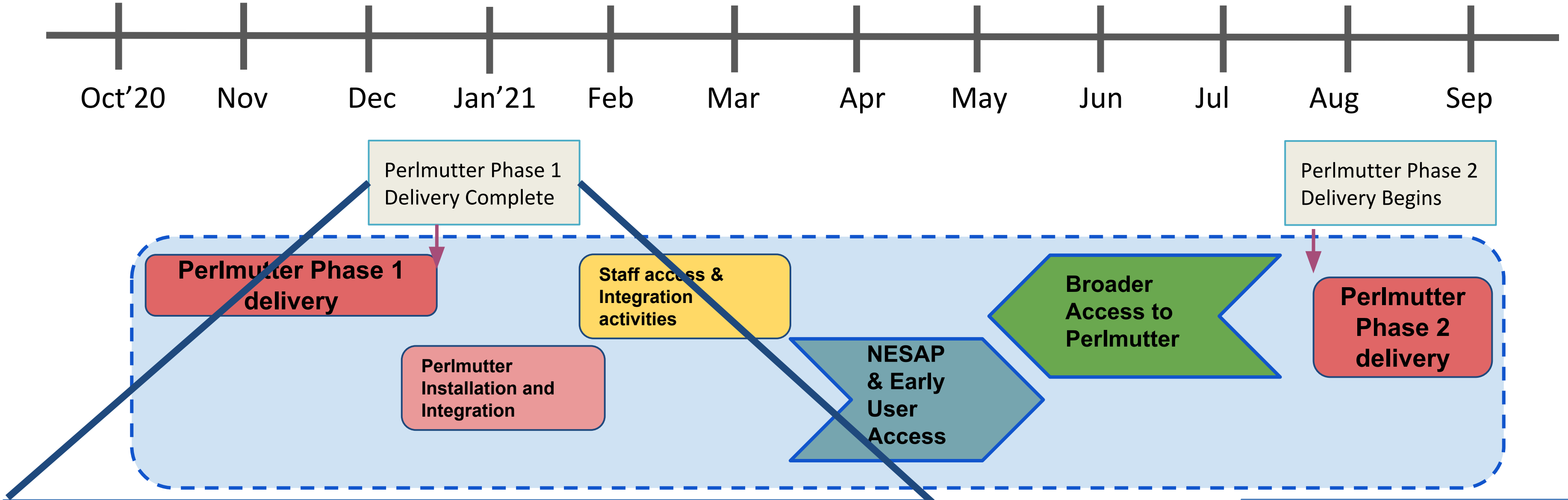
🟩 Phase 1 Late 2020
🟪 Phase 2 Mid 2021
🟦 Partly Phase 1 Partly Phase 2

NERSC

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

Perlmutter Phased Timeline



Oct'20 Nov Dec Jan'21 Feb Mar Apr May Jun Jul Aug Sep

Perlmutter Phase 1 Delivery Complete

Perlmutter Phase 2 Delivery Begins

Perlmutter Phase 1 delivery

Perlmutter Installation and Integration

Staff access & Integration activities

NESAP & Early User Access

Broader Access to Perlmutter

Perlmutter Phase 2 delivery

Phase 1

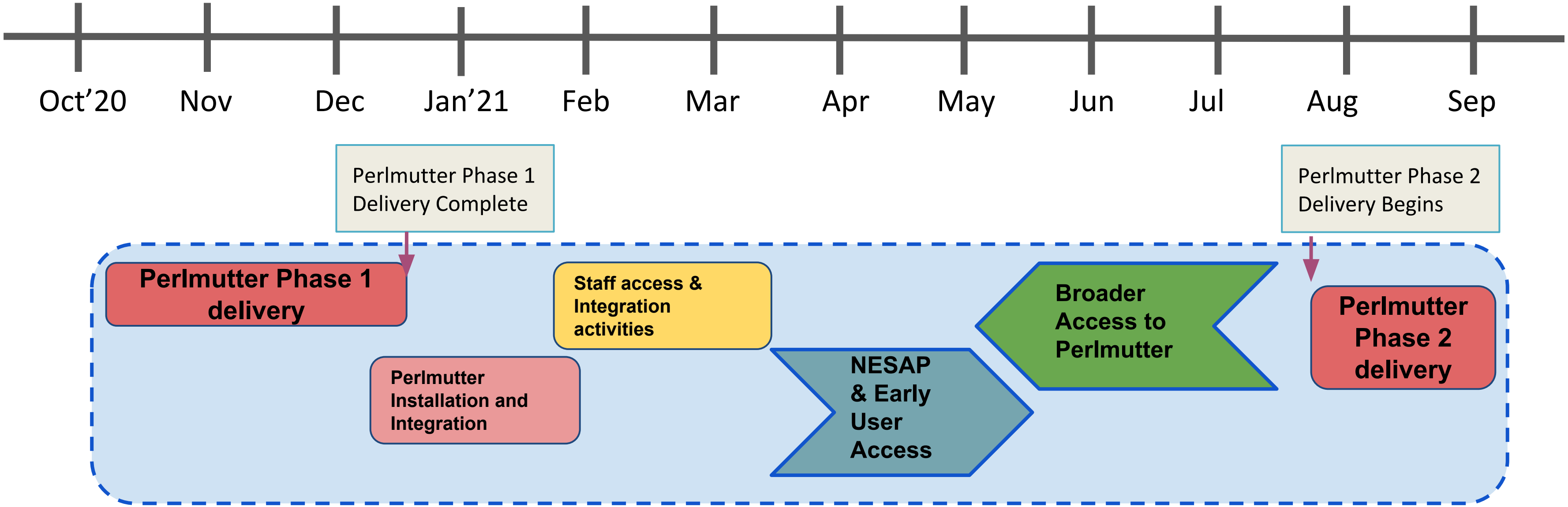
- 35PB Lustre storage
- Login and System configuration nodes (non-compute nodes)
  - Initially a subset of temporary Login nodes with Rome
- 12 GPU racks
- Slingshot 10

Phase 1a: Replace login nodes with Milan based nodes

Phase 2

- Slingshot 11
- 12 CPU racks

Perlmutter Phased Timeline


Oct’20
Nov
Dec
Jan’21
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Perlmutter Phase 1 Delivery Complete
Perlmutter Phase 2 Delivery Begins
Perlmutter Phase 1 delivery
Staff access & Integration activities
Perlmutter Installation and Integration
NESAP & Early User Access
Broader Access to Perlmutter
Perlmutter Phase 2 delivery

Phase 1

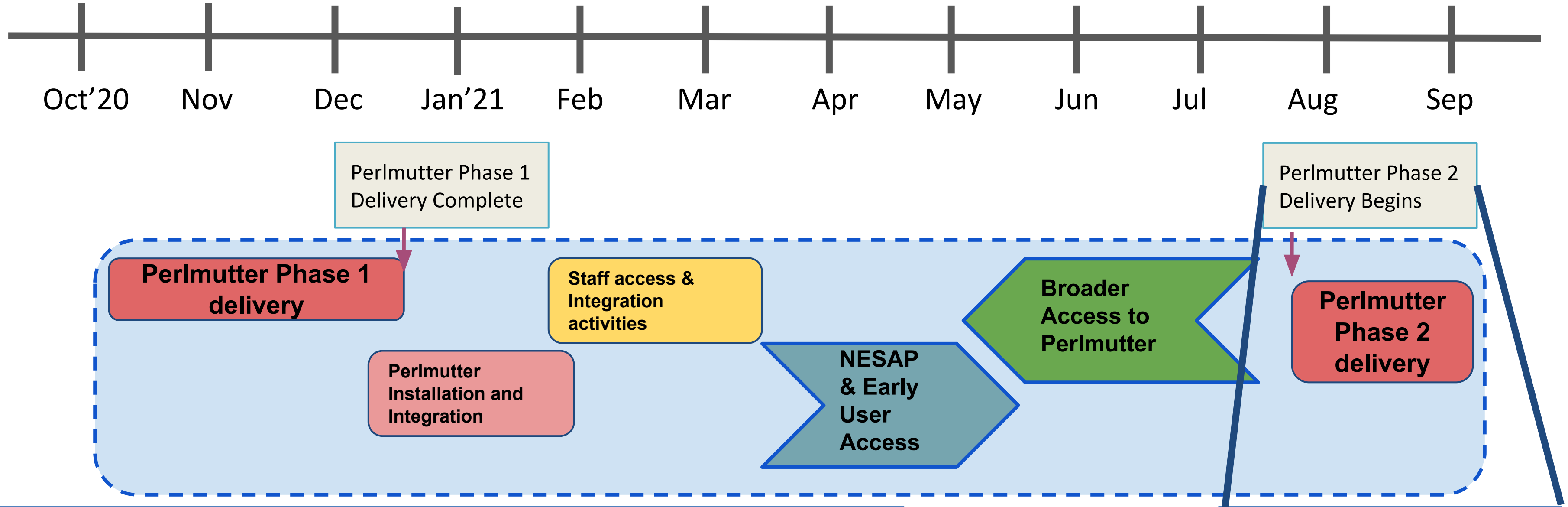
- 35TB Lustre storage
- Login and System configuration nodes (non-compute nodes)
    - Initially a subset of temporary Login nodes with Rome
- 12 GPU racks
- Slingshot 10

Phase 1a: Replace login nodes with Milan based nodes

Phase 2

- Slingshot 11
- 12 CPU racks

Perlmutter Phased Timeline



Oct’20 Nov Dec Jan’21 Feb Mar Apr May Jun Jul Aug Sep

Perlmutter Phase 1 Delivery Complete

Perlmutter Phase 2 Delivery Begins

Perlmutter Phase 1 delivery

Perlmutter Installation and Integration

Staff access & Integration activities

NESAP & Early User Access

Broader Access to Perlmutter

Perlmutter Phase 2 delivery

Phase 1

- 35PB Lustre storage
- Login and System configuration nodes (non-compute nodes)
  - Initially a subset of temporary Login nodes with Rome
- 12 GPU racks
- Slingshot 10

Phase 1a: Replace login nodes with Milan based nodes

Phase 2

- Slingshot 11
- 12 CPU racks

Perlmutter Compute Nodes

CPU-only nodes: AMD Milan CPU

- ~64 cores
- “ZEN 3” cores - 7nm+
- AVX2 SIMD (256 bit)
- 8 channels DDR memory
- *~ 1x Cori compute capacity*

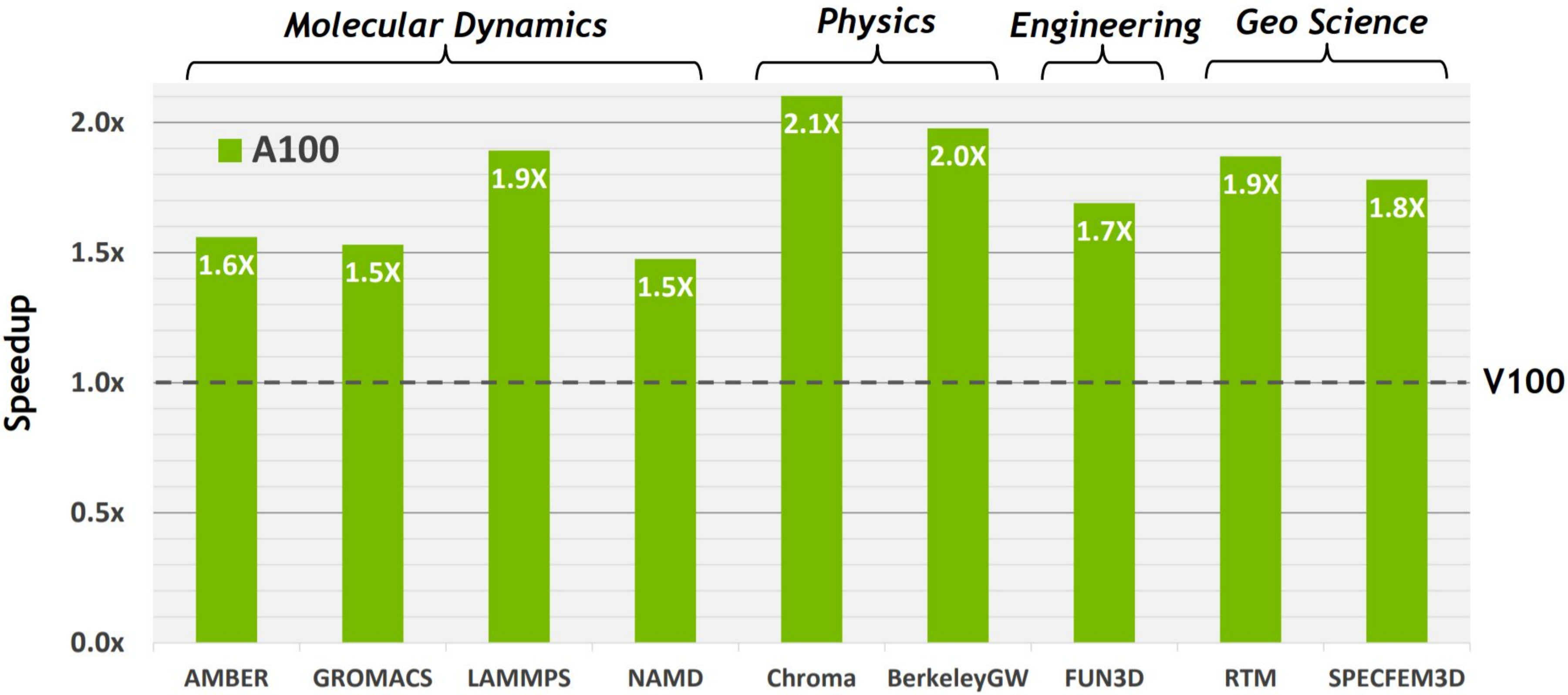
Rome specs

GPU+CPU nodes: 1x Milan + 4x Nvidia A100

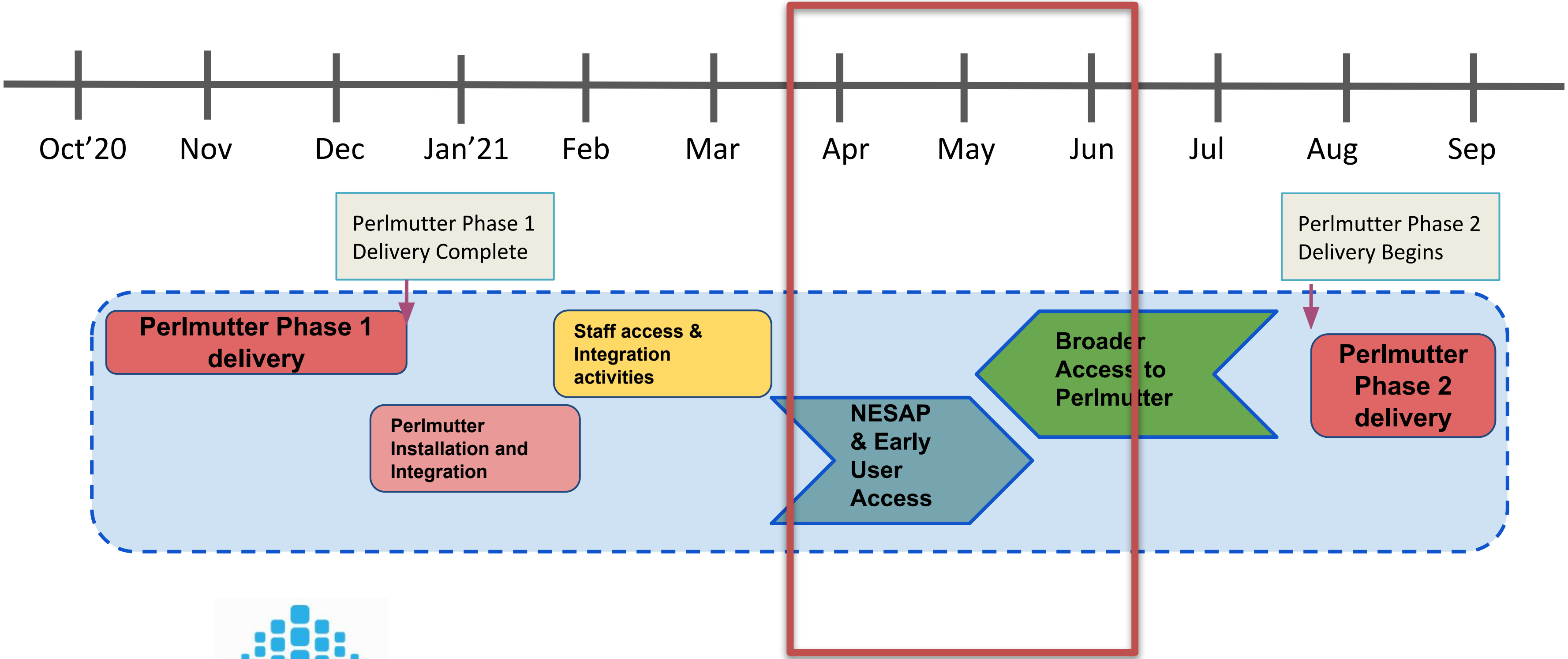
- NVLINK-3 (Between 4 GPUs)
- FP16, TF32, FP64 Tensor Cores
- GPU direct
- Multi-Instance GPU (MIG)
- *~3x Cori compute capacity*

	V100	A100
FP64 Peak	7.5 TF FMA	19.5 TF TC (9.7 TF FMA)
FP16 Peak	125 TF TC	312 TF TC
SMs	80	108
Memory BW	900 GB/s	1555 GB/s
Memory Size	16 GB	40 GB
L2 Cache	6 MB	40 MB
Shared Mem. / SM	96 KB	164 KB

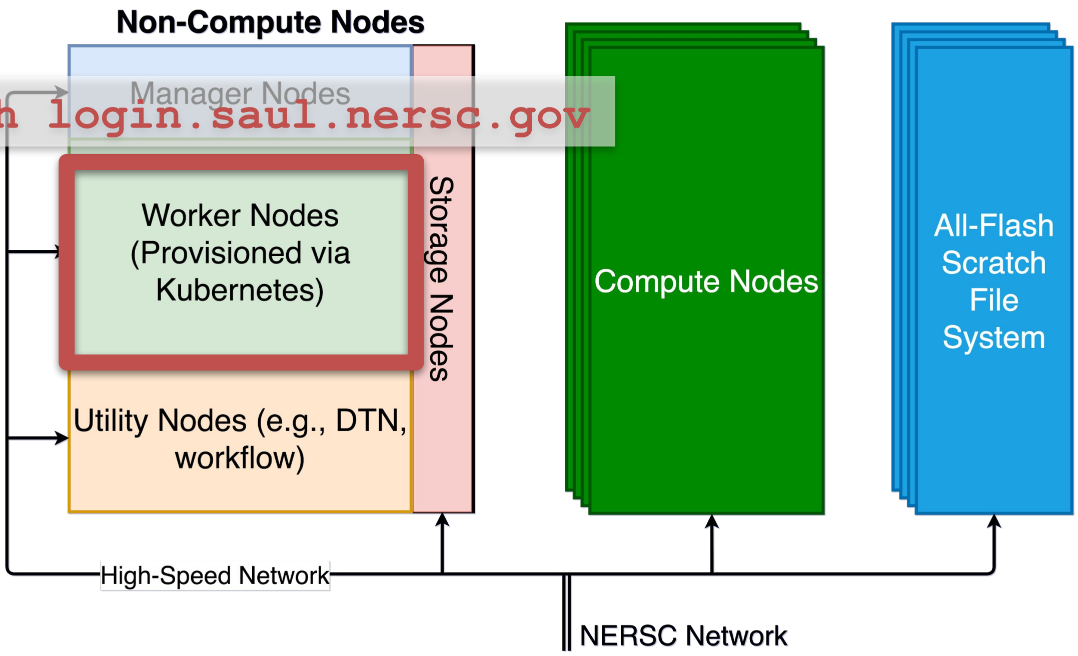
A100 vs V100


Molecular Dynamics
Physics
Engineering
Geo Science
🟩 A100
Speedup
2.0x
1.5x
1.0x
0.5x
0.0x
V100
1.6X
1.5X
1.9X
1.5X
2.1X
2.0X
1.7X
1.9X
1.8X
AMBER
GROMACS
LAMMPS
NAMD
Chroma
BerkeleyGW
FUN3D
RTM
SPECFEM3D

Perlmutter Phased Timeline


Oct'20
Nov
Dec
Jan'21
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Perlmutter Phase 1 Delivery Complete
Perlmutter Phase 2 Delivery Begins
Perlmutter Phase 1 delivery
Staff access & Integration activities
Perlmutter Installation and Integration
NESAP & Early User Access
Broader Access to Perlmutter
Perlmutter Phase 2 delivery

Perlmutter Architecture: Conceptual Overview



- Each user lands within a container orchestrated by Kubernetes
  - Spin up resources as needed
  - Insulation from other users' behavior
- Jupyter also for login
- Slurm for job scheduling
  - Additional flag for GPU
  - Queue policies TBD, but should resemble Cori

Programming Environment

	GPU Support	Fortran/ C/C++	OpenACC 2.x	OpenMP 5.x	CUDA	Kokkos / Raja	Cray MPI
PGI	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CCE		🟩				🟩	🟩
GNU	🟩	🟩	🟩	🟩	🟩	🟩	🟩
LLVM	🟦	🟦		🟦	🟦	🟦	🟦

Minerva Library	Vendor Supported?	GPU Enabled?
Python (Anaconda)	🟩	🟩
Spark	🟩	🟩
R	🟩	🟩
TensorFlow	🟩	🟩
Keras	🟩	🟩
Caffe	🟩	🟩
PyTorch	🟩	🟩

🟩 Vendor Supported

🟦 NERSC Supported

Allocations & Charging

- In Allocation Year (AY) 2021:
  - All Perlmutter time will be “free”
  - However, no commitment for hours available
  - *Expect that there will be outages for testing & stabilization*

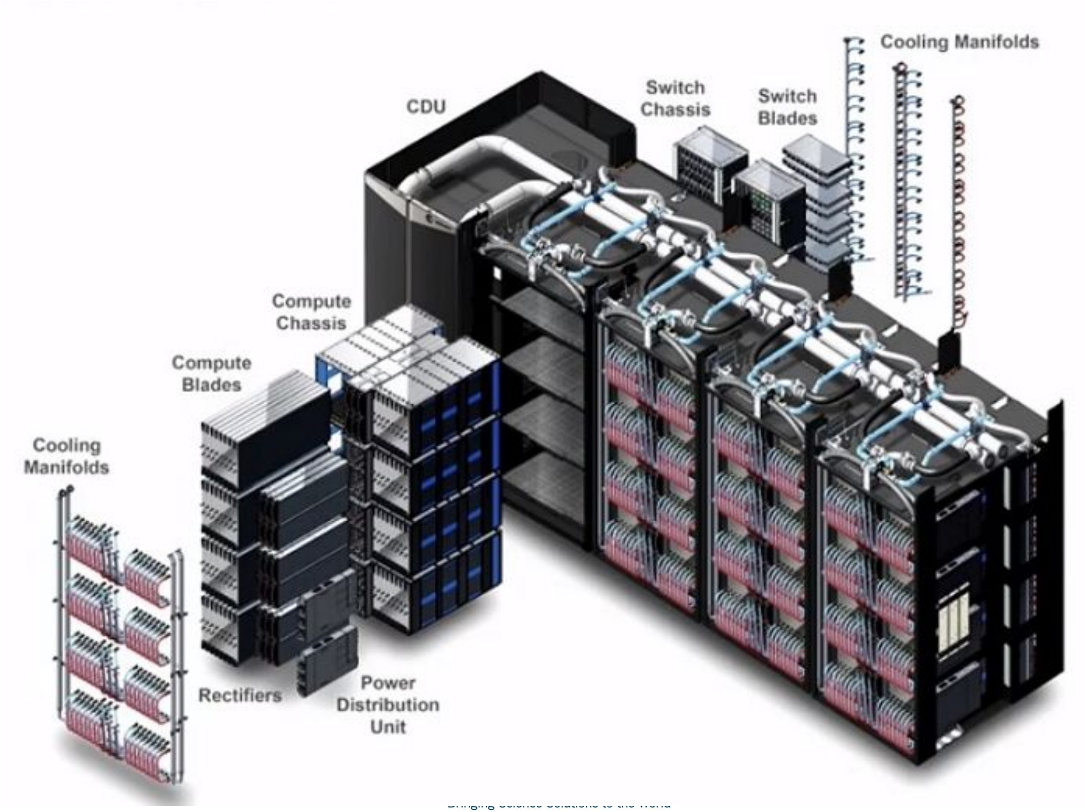
- ERCAP requests & charging will begin in AY 2022

- Allocation and Charging Units
  - **GPU Node Hours** for Perlmutter GPU accelerated nodes
  - **CPU Node Hours** for Perlmutter CPU, Cori Haswell and KNL nodes
  - **GPU** & **CPU** hours will not be interchangeable!
  - CPU Node Hour charges will incorporate performance scale factors for 3 types of nodes

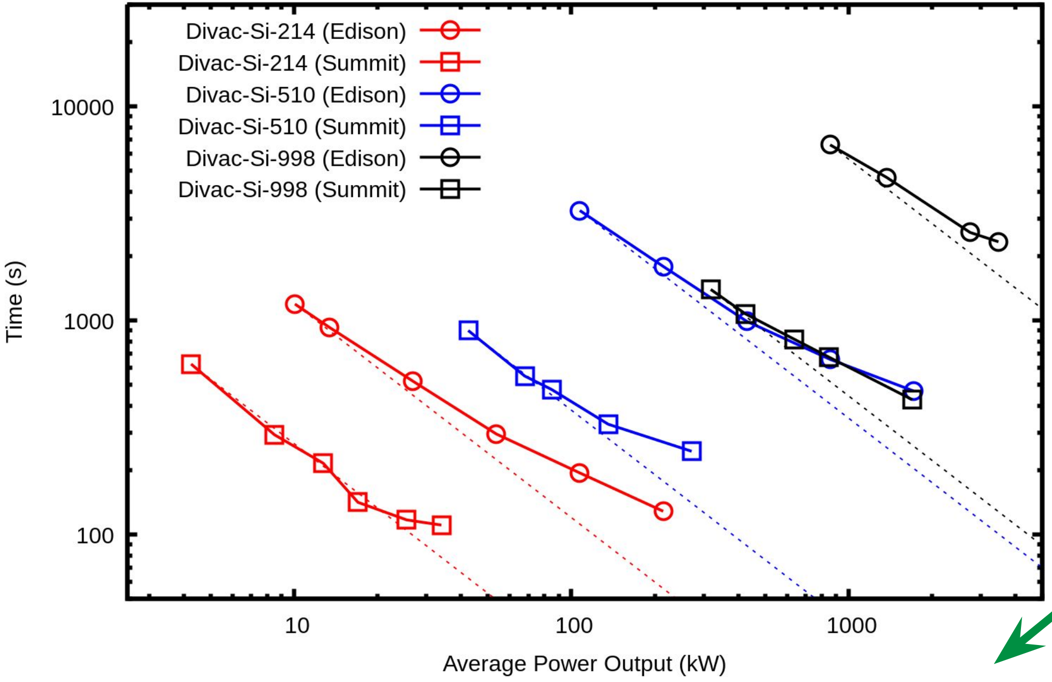
Perlmutter: Physical Integration

After moving cabinets into place, first step is physically connecting cabinets to power and water.

Then connections to the many networks that the system and NERSC need!



Why GPUs



Improving Energy Efficiency

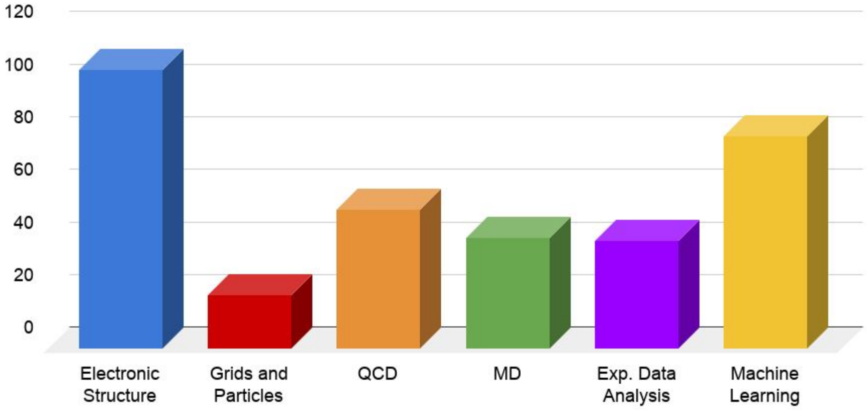
NESAP (NERSC Exascale Science Applications Program)

Goal:

Partner with Cray/NVIDIA and ~25 Teams (broad range across workload) at Deep Level to Prepare Apps for Perlmutter.

Disseminate Lessons Learned to NERSC Community Through Documentation, at Training Events and Community Hackathons

GPU Community Hackathons



Higher is Better

Projected Speedups on Perlmutter over Edison for Top NESAP Apps in Algorithmic Areas.

Includes Software Improvements from NESAP.

GPU For Science Days

User Access to Perlmutter Phase I (GPU Nodes)

~ April 2021:

- All users given accounts to Phase I for code development and small-scale testing
- Priority access for large-scale testing and project milestones
  - NESAP teams
  - Exascale Computing Project teams

~ May 2021:

- Large-scale scientific computing access for all GPU-capable projects
- GPU-readiness evaluation form required
- Key GPU-enabled community apps will be available

Perlmutter Environment will be largely familiar

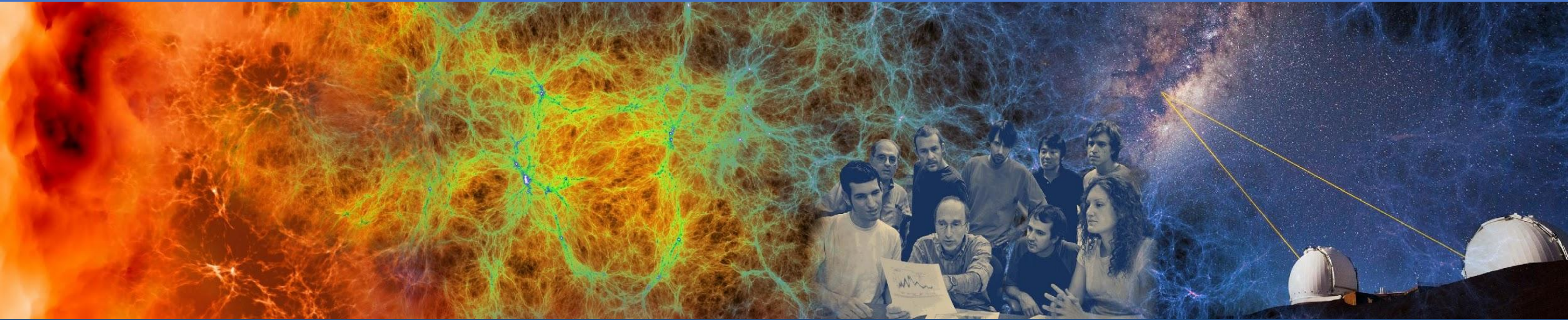
- Differences:
  - LMOD instead of modules (also on Cori in AY21*)
    - Hierarchical based modules, should be easier to find & load modules
  - Different programming environments
    - No Intel
- NERSC and/or Vendor will provide many libraries for users
- Users will be able to compile software not provided
  - User Spack instance, to draw upon pre-existing recipes
  - Help from NERSC consultants to install in user or project space
- New: Cray Minerva Data & Analytics Software Stack

*probably

Learning Opportunities: Training and hackathons

- NERSC will hold training sessions for diverse interests and levels of experience
  - How to use the system, How to compile codes, Performance optimization, NVIDIA A100 Architecture deep dive, Development and tools, Machine Learning, Chemistry / materials science applications and more!
- Hackathons come in a variety of forms, but generally:
  - Pair code teams with experienced mentors.
  - Give an opportunity learn new profiling techniques and tools.
  - Identify, explore and directly fix problems with your codes.
  - Research and learn about new coding strategies and methods.
  - Develop contacts for future collaboration, code development and support.
  - Virtual-only formats are being adopted, tested and expanded on now.

The Superfacility project: Federated ID @ NERSC

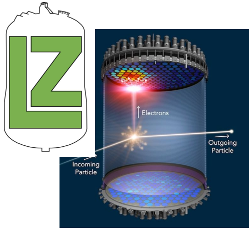


The CS Area Superfacility ‘project’ coordinates and tracks work to support experimental science at NERSC

Project Goal:

By the end of CY 2021, 3 (or more) of our 7 science application engagements will demonstrate automated pipelines that analyze data from remote facilities at large scale, without routine human intervention, using these capabilities:

- Real-time computing support
- Dynamic, high-performance networking
- Data management and movement tools
- API-driven automation
- Authentication using **Federated Identity**



ALS

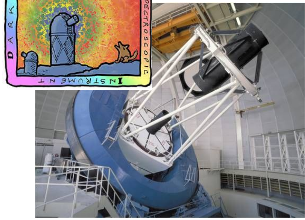
FICUS
Facilities Integrating Collaborations for User Science
JGI-EMSL Collaborative Science Call

LSST DESC
Dark Energy Science Collaboration

MOLECULAR FOUNDRY

LCLS

DARK ENERGY SPECTROSCOPIC INSTRUMENT

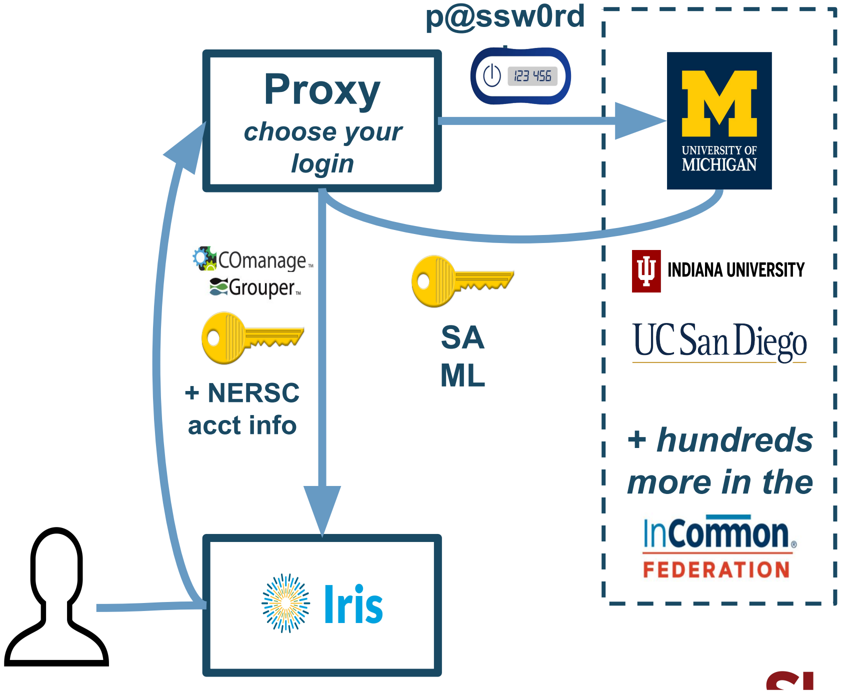


Find out more at our recent demo series:
https://www.nersc.gov/research-and-development/superfacility/

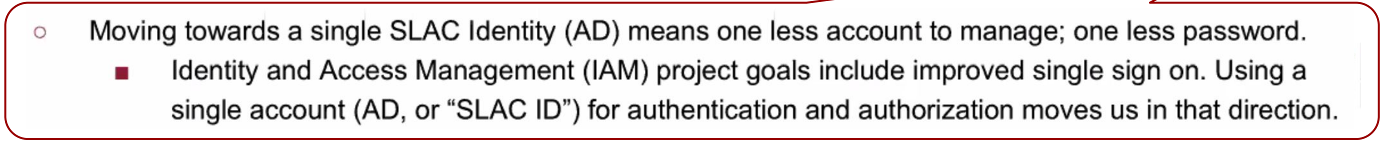
Federated Identity @ NERSC

Federated Identity (FedID) allows a person to use a **single digital identity across multiple organizations**

- Simplifies cross-facility workflows (Superfacility)
- Users benefit from fewer, more familiar, passwords and login pages
- NERSC benefits from fewer support tickets (eg, password resets)
- Home institution manages the user identity lifecycle
- NERSC still manages local authorization
- Core technology is well-established and mature



SLAC

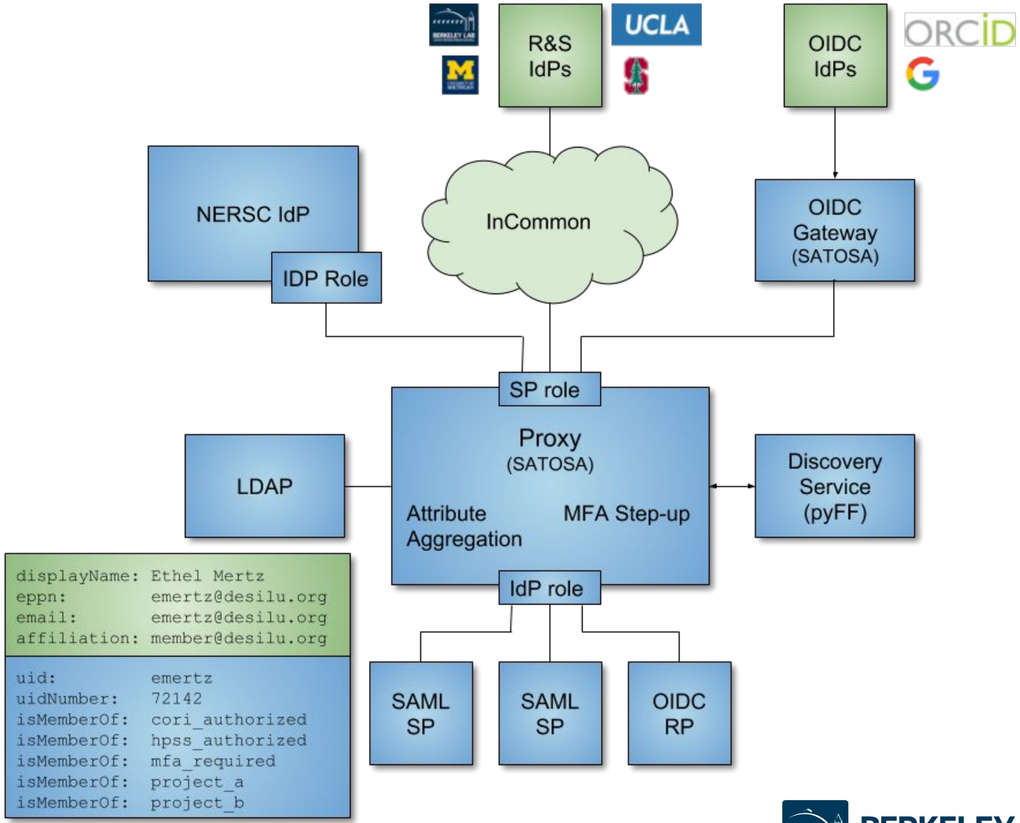


- Moving towards a single SLAC Identity (AD) means one less account to manage; one less password.
  - Identity and Access Management (IAM) project goals include improved single sign on. Using a single account (AD, or “SLAC ID”) for authentication and authorization moves us in that direction.

Federated ID @ NERSC

- Spent much of 2019 and 2020 putting the key technologies in place
  - Iris, IAM technical design review
    - “Choice of IAM service tools (e.g. Grouper, CoManage, Shibboleth IdP & SP) and web-development tools (e.g. React, Flask) are good choices and state-of-the art”
  - Iris Rollout (December 2019)
- Internal security review of policy (completed), technology (in-progress)
- Taking proposal for phased rollout to DOE program managers this Fall
  - Pre-Pilot: Berkeley Lab (1 month)
  - Pilot: DOE Labs + DOE OneID (1-2 months)
  - Full Deployment: Above + SIRTFI-Compliant InCommon Members (ongoing)
- Alignment with DOE’s Distributed Computing Data Ecosystem (DCDE) pilot
  - Same standard technology stack under consideration: COmanage / Grouper / Shibboleth / SATOSA, SAML authentication

Federated ID Login Flow



```
displayName: Ethel Mertz
eppn:        emertz@desilu.org
email:       emertz@desilu.org
affiliation: member@desilu.org

uid:         emertz
uidNumber:   72142
isMemberOf:  cori_authorized
isMemberOf:  hpss_authorized
isMemberOf:  mfa_required
isMemberOf:  project_a
isMemberOf:  project_b
```

Federated ID @ NERSC

- Plan to accept authentications from:
  - DOE providers
  - InCommon providers in Research and Scholarship category that assert compliance with SIRTFI security framework
- IdP of last resort will continue to be NERSC
- Future possible additions
  - More of InCommon
  - International providers (eg CERN)
  - Social providers (eg ORCID)

Thanks!

Catch up on NERSC news from the NUG Annual Meeting last week: https://www.nersc.gov/users/NUG/annual-meetings/nug-2020/