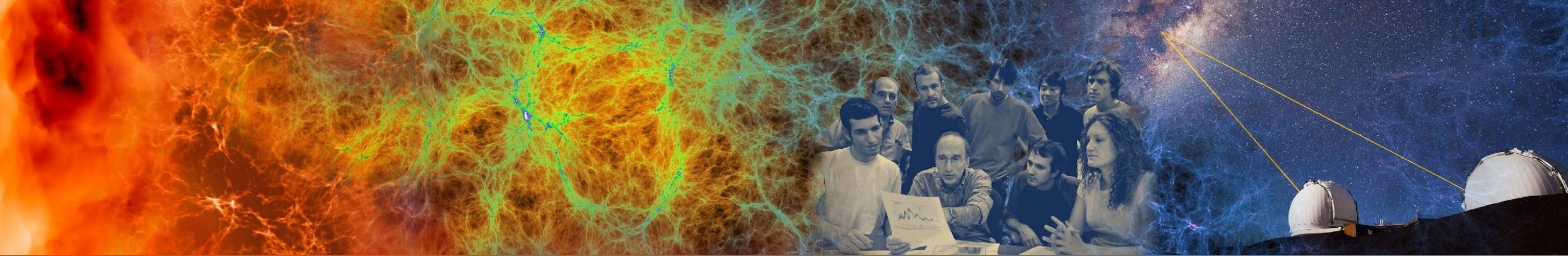
Triton + NERSC

SONIC Workshop March 2024

Andrew Naylor
NESAP Postdoc
March 1, 2024

Who am I?

- NERSC NESAP Postdoc
- Work within Data & AI Services group
- Focus on AI inference
- Avid drummer

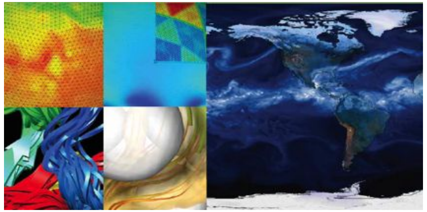


TLDR:

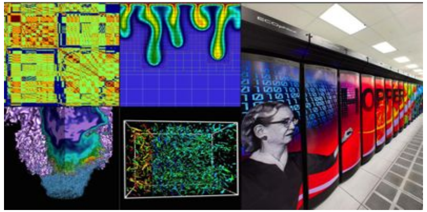
Currently unable to run SONIC but ExaTrkX is running Triton on NERSC and NERSC is interested in AI inference-as-a-service

National Energy Research Scientific Computing Center

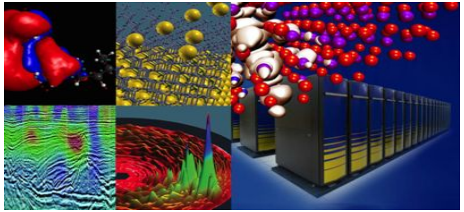
- NERSC (at LBNL) is the *mission* High Performance Computing and Data facility for the DOE Office of Science
- Celebrating 50 Years
- ~10,000 Users, 800+ Projects, ~2000 NERSC citations per year



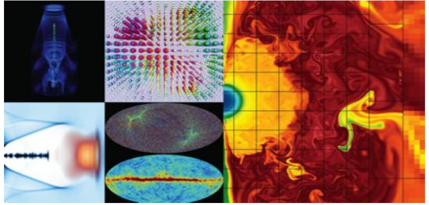
Bio Energy, Environment



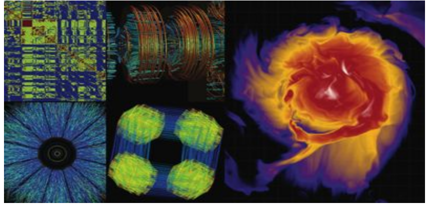
Computing



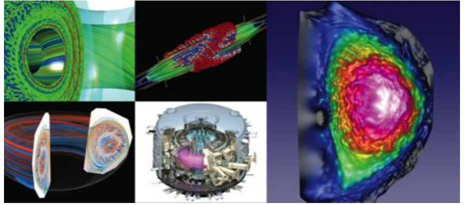
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



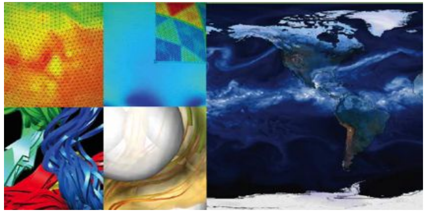
Nuclear Physics



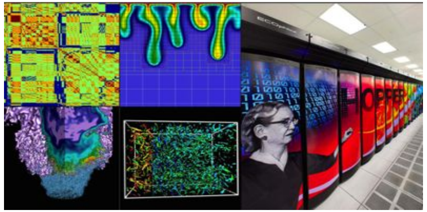
Fusion Energy, Plasma Physics

National Energy Research Scientific Computing Center

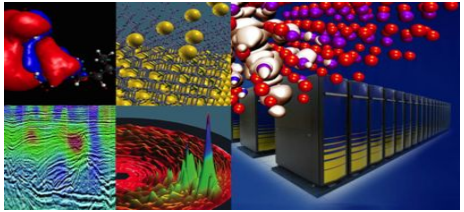
- We deploy supercomputer systems for cutting edge simulations and data analytics at scale
- NERSC Science Acceleration Program (NESAP) is a collaboration with partners to prepare for advanced architectures and new systems.



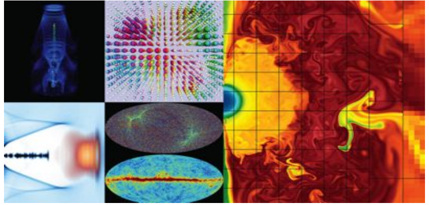
Bio Energy, Environment



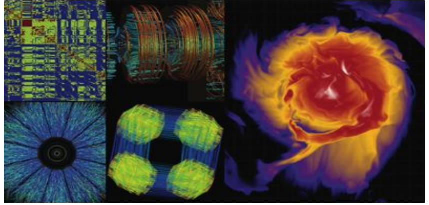
Computing



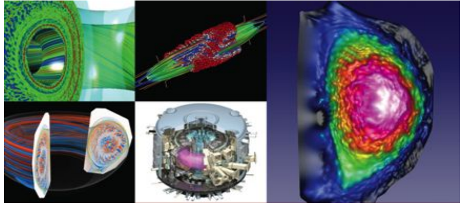
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics

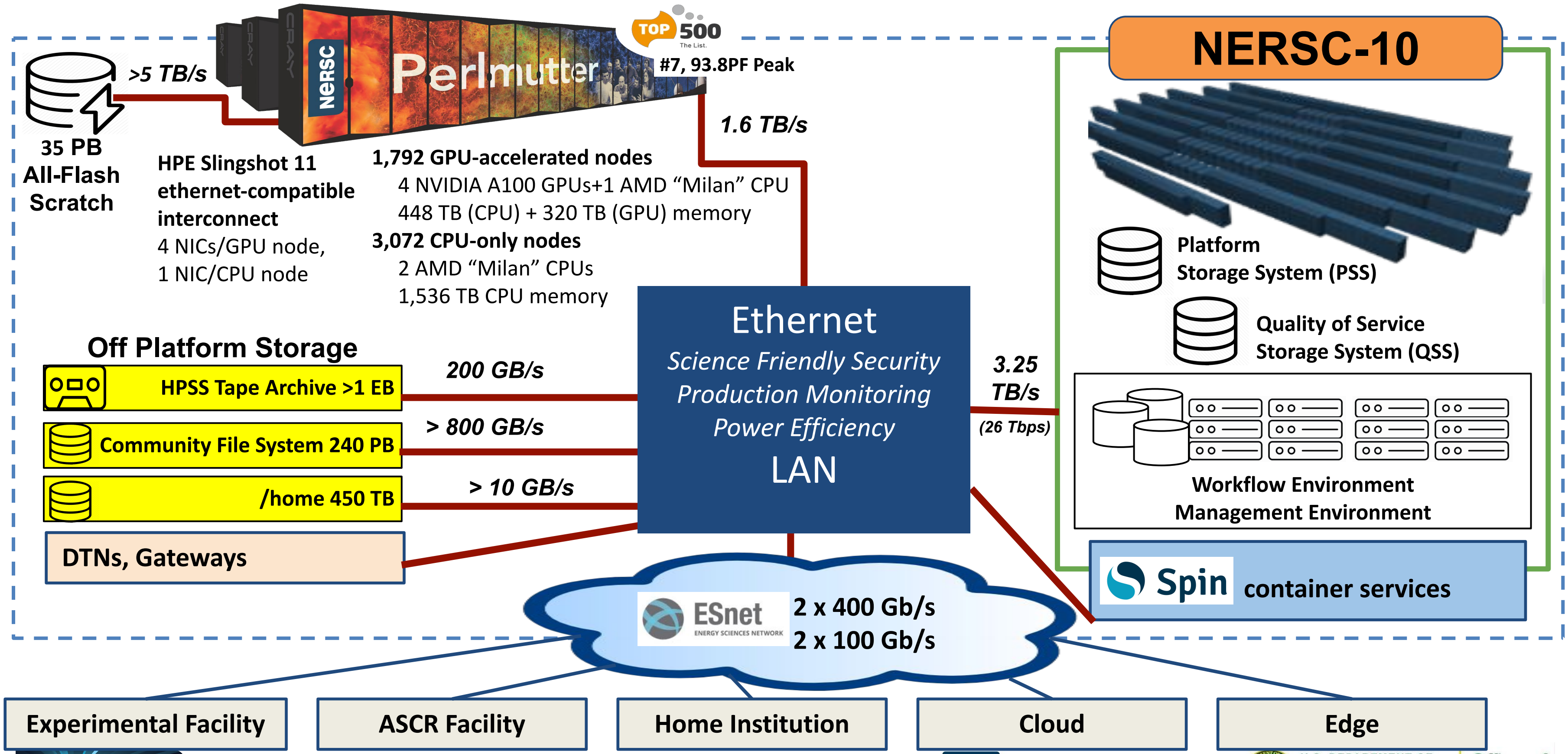


Nuclear Physics



Fusion Energy, Plasma Physics

NERSC Center Architecture



>5 TB/s

35 PB All-Flash Scratch

Perlmutter

TOP 500

#7, 93.8PF Peak

1.6 TB/s

HPE Slingshot 11 ethernet-compatible interconnect
4 NICs/GPU node,
1 NIC/CPU node

1,792 GPU-accelerated nodes
4 NVIDIA A100 GPUs+1 AMD “Milan” CPU
448 TB (CPU) + 320 TB (GPU) memory

3,072 CPU-only nodes
2 AMD “Milan” CPUs
1,536 TB CPU memory

Off Platform Storage

HPSS Tape Archive >1 EB	200 GB/s
Community File System 240 PB	> 800 GB/s
/home 450 TB	> 10 GB/s
DTNs, Gateways	

Ethernet
Science Friendly Security
Production Monitoring
Power Efficiency
LAN

3.25 TB/s
(26 Tbps)

NERSC-10

Platform Storage System (PSS)

Quality of Service Storage System (QSS)

Workflow Environment Management Environment

Spin container services

ESnet ENERGY SCIENCES NETWORK

2 x 400 Gb/s
2 x 100 Gb/s

Experimental Facility | ASCR Facility | Home Institution | Cloud | Edge

NERSC

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science

Containers at NERSC

Containers are valuable to our scientific computing users

- Encapsulation, isolation, reproducibility, portability, and even scalability

NERSC supports user container workloads via Shifter

- Developed at NERSC to address security concerns of docker
- Enables scalability on HPC systems
- Users can build their images with docker, then easily convert to shifter with a simple pull command

NERSC also supports podman-hpc

- NERSC built wrapper for podman (open source tool)
- All the benefits of shifter, but using OCI standard runtime
- Users can build images at NERSC

Spin: Container Services for Science

Spin

Many projects need more than HPC.

Spin is a platform for services.

Users deploy their **science gateways, workflow managers, databases, and other network services** with Docker containers.

- *Access HPC file systems and networks*
- *Use public or custom software images*
- *Orchestrate complex workflows*
- *Secure, scalable, and managed*

docker RANCHER kubernetes

Some projects using Spin:

DARK ENERGY SPECTROSCOPIC INSTRUMENT	Track and compare analyses of nightly sky surveys	science gateway
ESS-DIVE	Classify and store reusable earth sciences data	data repository
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY	Manage production genomic workflows and data at scale	science gateway
LZ	Process real-time events for dark matter detection	workflow manager
The Materials Project materialsproject.org	Explore materials properties or build simulated materials	science gateway

Workflow capabilities

Kubernetes (in the future?)

- Currently cannot provide k8s for our GPU compute resources in a cloud-like way
- Maybe user deployment of usernetes (theoretically)
- We are looking into approaches for the NERSC-10 system (~2026) as well as pilot collaborations with SchedMD on k8s+slurm integration

Realtime queues are available to enable on-demand HPC resources

- e.g., for experiments that need realtime processing during data collection
- Available by special request: Resource Usage Policies - NERSC Documentation

Superfacility API **(https://api.nersc.gov/api/v1.2/)**

- An API for interacting with NERSC supercomputers
- Vision: all NERSC interactions are callable; backend tools assist large or complex operations.
- Able to submit jobs, create reservations, move + upload files, etc…

NERSC looking ahead

We want to enable and support all major types of ML workflows

- We have so far focused on supporting training workloads, which are well suited to HPC
- As AI4Science matures, inference workloads become more important
- We are interested in supporting GPUaaS-like workflows

We want to provide a rich platform/ecosystem for MLOps for science

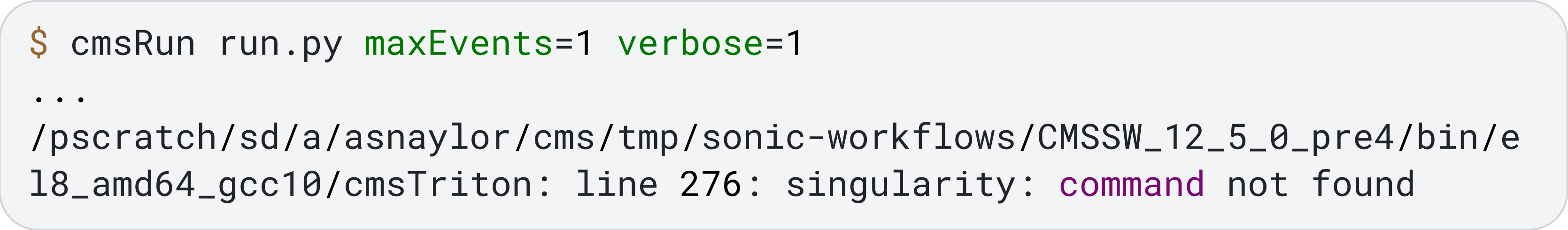
- Productive and performant interfaces to deploy distributed ML workloads, search hyperparameters, track experiments, share models, etc.

NERSC-10 (2026) will be designed with “workflows” heavily in mind

- will further enable complex research workflows for experimental science

Experience with SONIC

- Unable to run SONIC right now
- Following sonic-workflows
- After challenges to setup CMSSW on Perlmutter. Fallback server fails:



```
$ cmsRun run.py maxEvents=1 verbose=1
...
/pscratch/sd/a/asnaylor/cms/tmp/sonic-workflows/CMSSW_12_5_0_pre4/bin/e
l8_amd64_gcc10/cmsTriton: line 276: singularity: command not found
```

- NERSC does not fully support singularity
  - ○ Maybe get singularity from cvmfs or user deployed binary (performance untested)
  - ○ Should work with podman-hpc

Experience with SONIC

- Modified `cmsTriton` script to use podman-hpc
    - Still unsuccessful, requires deep-dive
- Was able to deploy the Triton Server via podman-hpc

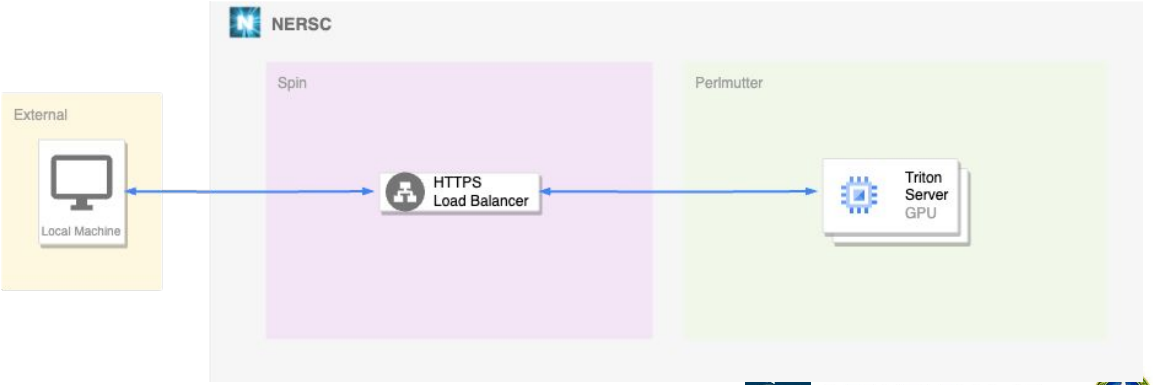
```
$ podman-hpc run ... --gpu fastml/triton-torchgeo:22.07-py3-geometric
tritonserver ... --model-repository=/data/models/
```

- However, it was missing a model:

```
$ cmsRun run.py maxEvents=1 verbose=1 address=$ADDRESS
An exception of category 'MissingModel' occurred while
   [0] Calling beginStream for module
DeepMETSonicProducer/'deepMETsResolutionTune'
```

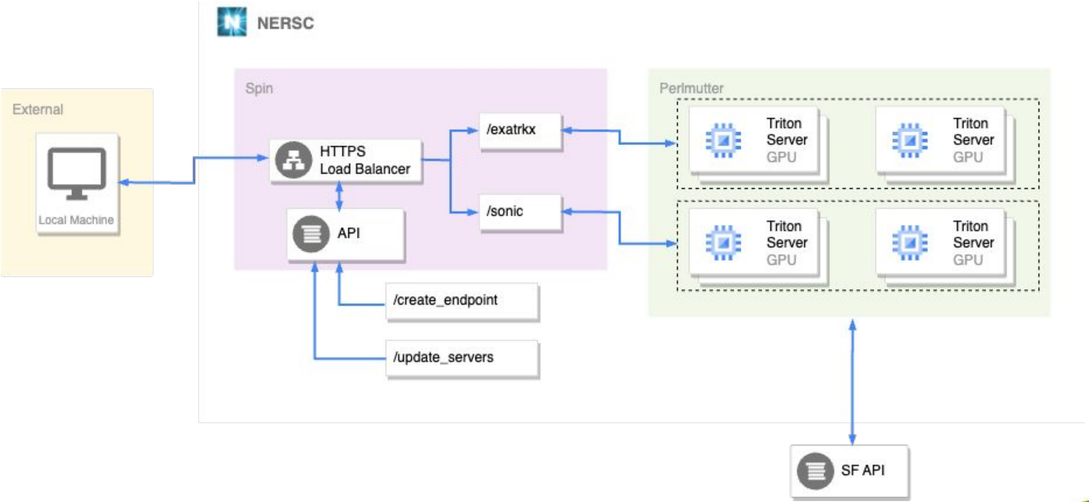
Experience with Triton @NERSC

- GNN-based track finding (ExaTrkX) as-a-service tested on Perlmutter
  - ○ Triton ensemble model & custom backend
  - ○ Seen positive results in as-a-service testing
- Tested Triton server with Resnet50 (PyTorch backend) through Spin LB
  - ○ Performance was x5 slower than Triton on Perlmutter
  - ○ Issues and performance at NERSC are not yet ironed out as early days



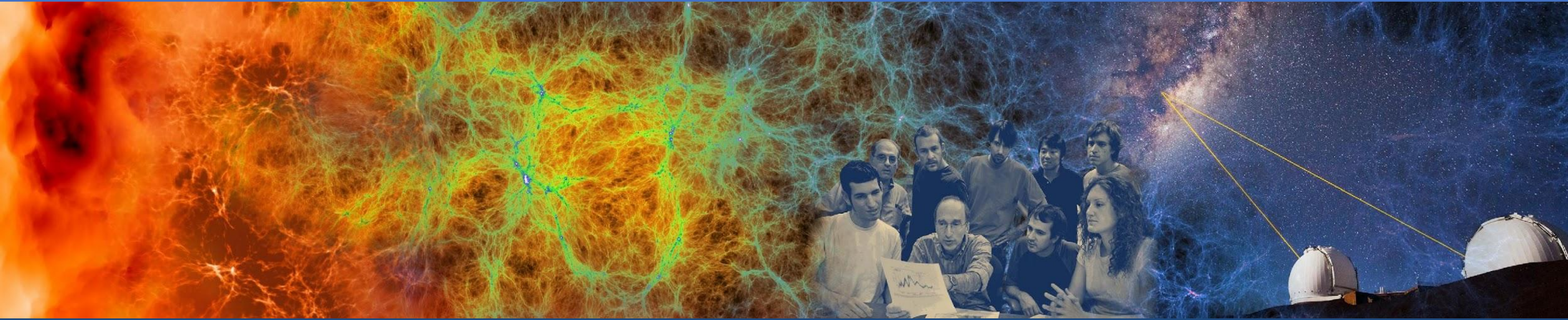
Potential Future Plans

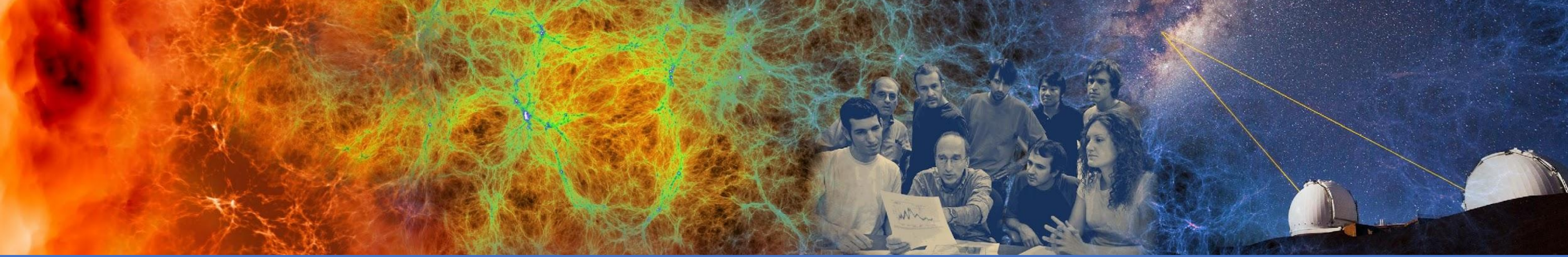
- Exploring idea of NERSC providing inference load balancer on Spin
    - ○ Users provide inference servers and we connect it to the internet
- Create endpoints within nersc through an API
- Control Triton server deployment through SF API



Thank you for listening. Any Questions?

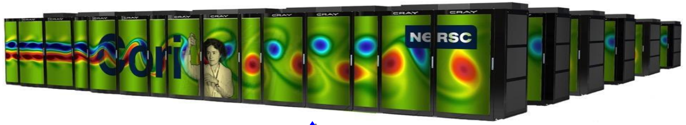
Andrew Naylor
anaylor@lbl.gov
FastML Slack

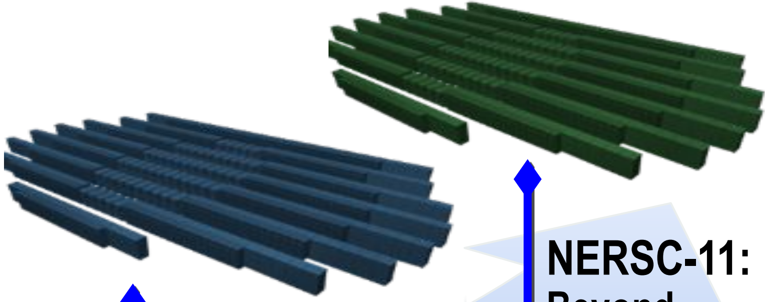




Backup slides

NERSC Systems Roadmap





2013

NERSC-7: Edison
Multicore CPU

2016

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

2020

NERSC-9: Perlmutter
CPU and GPU nodes
NESAP Expanded Simulation, Learning & Data: Continued transition of applications and support for complex workflows

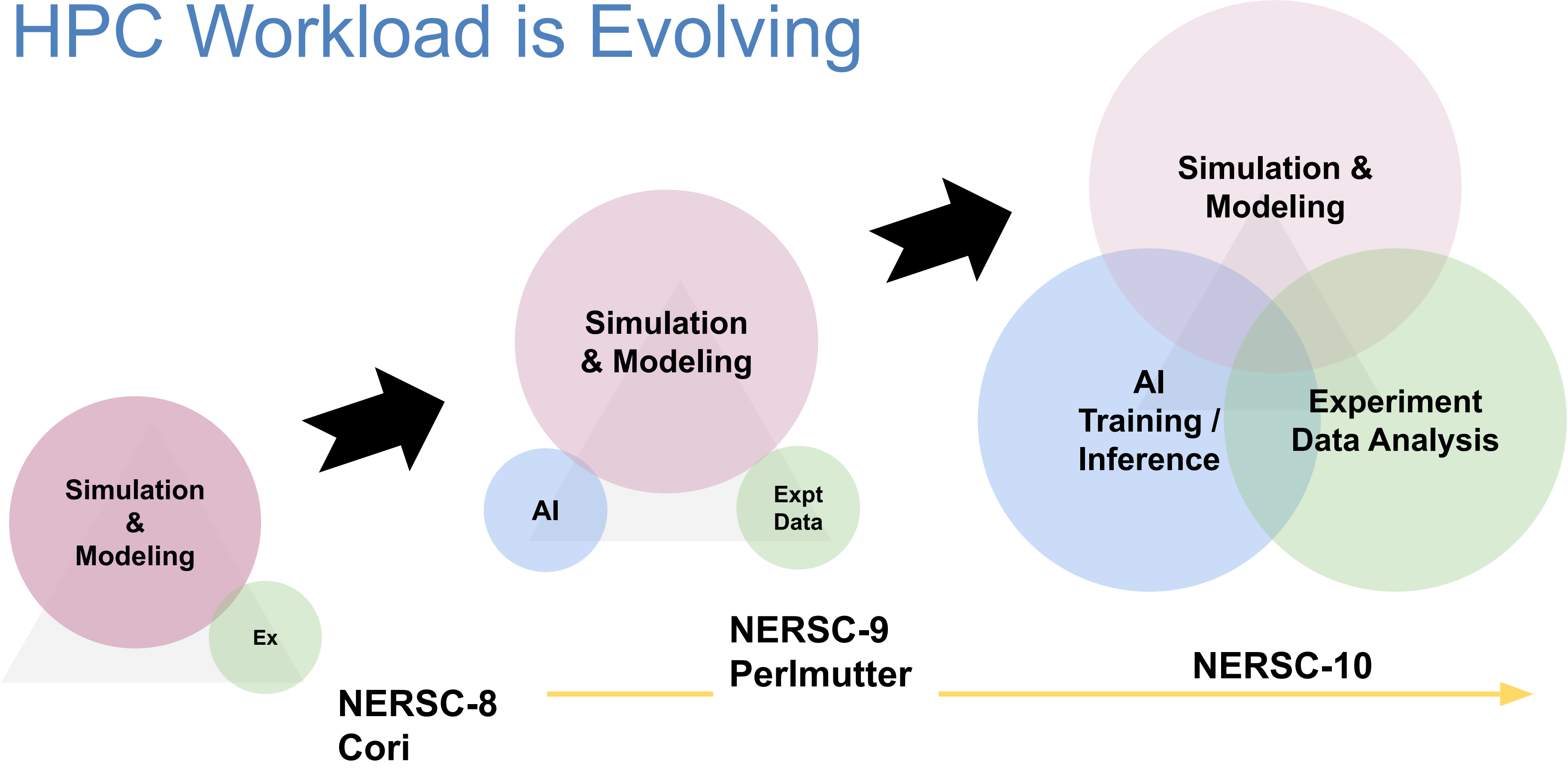
2026

NERSC-10:
Exa system
NESAP Workflows: Accelerating end-to-end workflows with technology integration

2030+

NERSC-11:
Beyond Moore

HPC Workload is Evolving



NERSC has a rich data ecosystem

globus online

jupyter

data transfer and access

netCDF

HDF

mongoDB®

MySQL™

data management

julia

data analytics

PyTorch

scikit learn

machine learning

visit

ParaView
Parallel Visualization Application

visualization

SHIFTER

Spin

containers

GNUparallel

Parsl

papermill

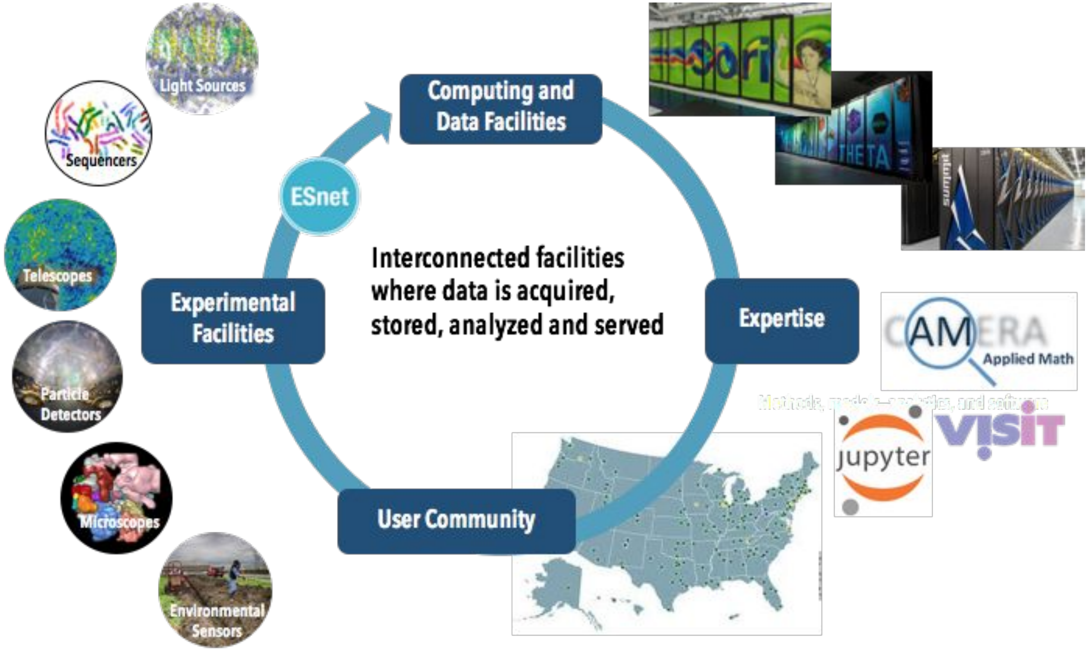
FireWorks

workflows

The Superfacility Model: an ecosystem of connected facilities, software and expertise to enable new modes of discovery

Superfacility@LBNL: NERSC, ESnet, AMCR, & SDD working together to support experimental science

- A model to integrate experimental, computational and networking facilities for reproducible science
- Enabling new discoveries by coupling experimental science with large scale data analysis and simulations

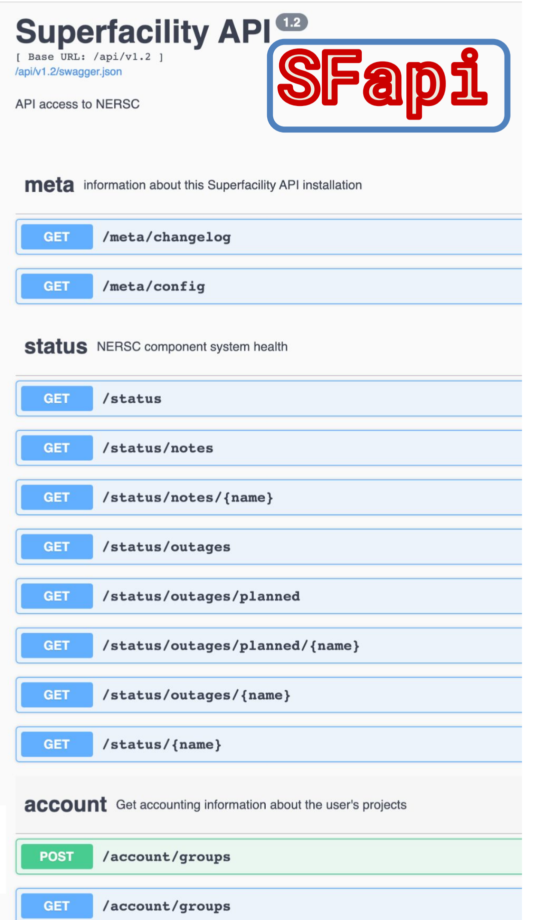


Machine-readable supercomputers: the Superfacility API

Vision: all NERSC interactions are callable;
backend tools assist large or complex operations.

Endpoints currently deployed:

`/meta`	information about this Superfacility API installation
`/status`	NERSC component system health
`/account`	Get accounting information about the user's projects
`/utilities`	basic file browsing, upload and download of small files to and from NERSC
`/storage`	Transfer files between Globus endpoints.
`/compute`	Run commands and manage batch jobs on NERSC compute
`/tasks`	Get information about your pending or completed tasks
`/reservations`	submit and manage future compute reservations



Superfacility API 1.2

[Base URL: /api/v1.2]
/api/v1.2/swagger.json

API access to NERSC

SFapi

meta information about this Superfacility API installation

GET /meta/changelog
GET /meta/config

status NERSC component system health

GET /status
GET /status/notes
GET /status/notes/{name}
GET /status/outages
GET /status/outages/planned
GET /status/outages/planned/{name}
GET /status/outages/{name}
GET /status/{name}

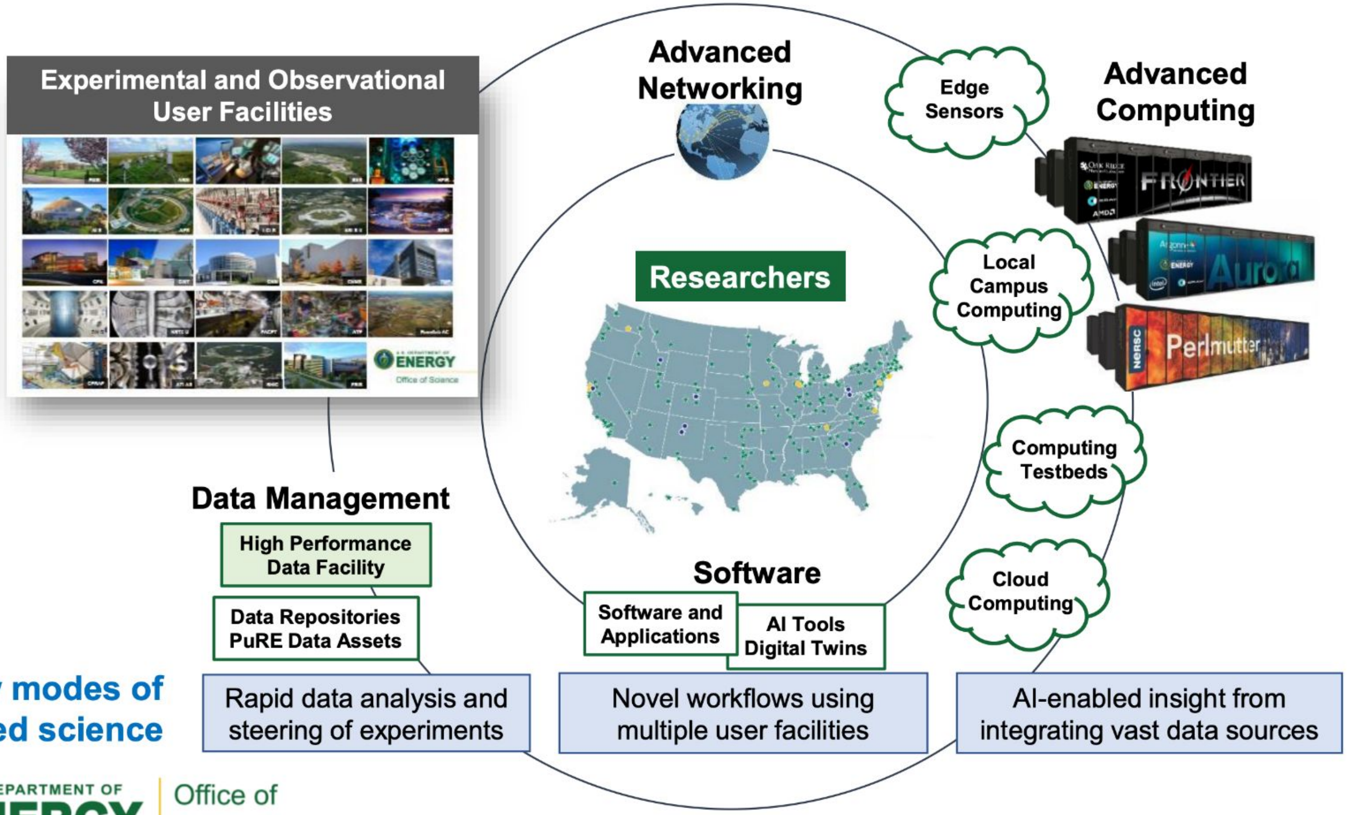
account Get accounting information about the user's projects

POST /account/groups
GET /account/groups

https://api.nersc.gov/

DOE's Integrated Research Infrastructure (IRI) Vision:

To empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation



Experimental and Observational User Facilities

Advanced Networking

Edge Sensors

Advanced Computing

Researchers

Local Campus Computing

Computing Testbeds

Cloud Computing

Data Management

High Performance Data Facility

Data Repositories PuRE Data Assets

Software

Software and Applications

AI Tools Digital Twins

New modes of integrated science

Rapid data analysis and steering of experiments

Novel workflows using multiple user facilities

AI-enabled insight from integrating vast data sources

U.S. DEPARTMENT OF ENERGY | Office of Science

Slide adapted from Ben Brown, ASCR

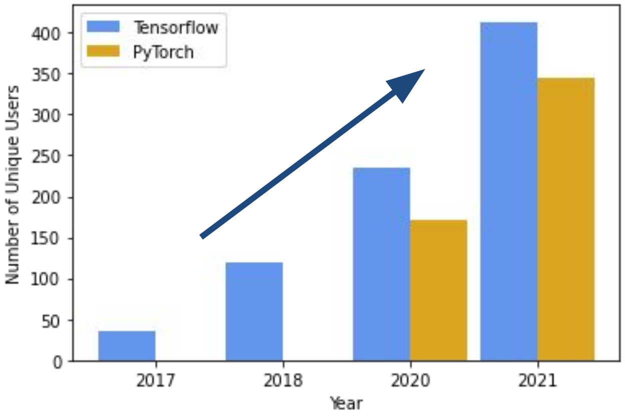
Growing scientific AI workload at NERSC

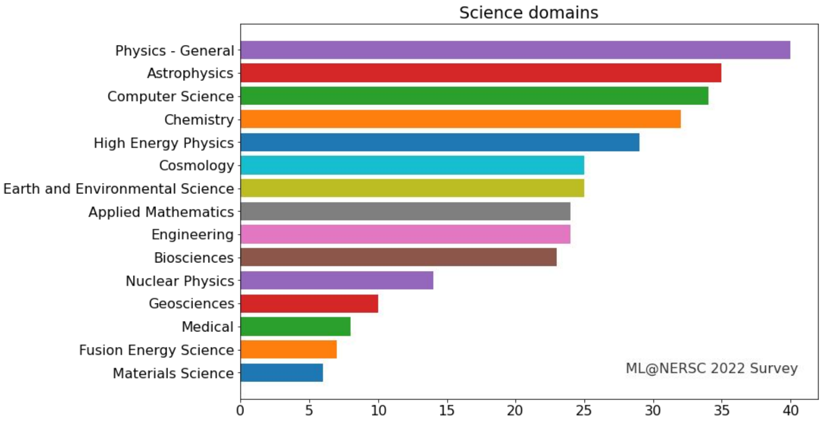
We track ML software usage

- Module loads and python imports
- Users of DL frameworks increased more than 6x from 2018 to 2021

We track ML trends through 2-yearly survey

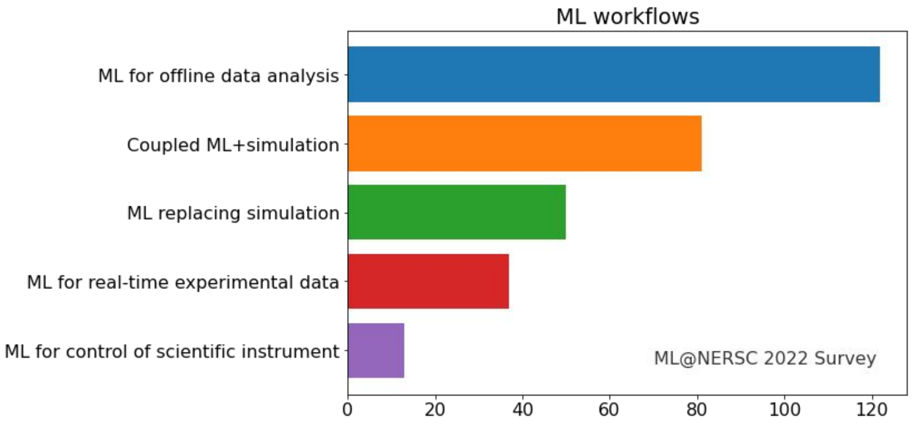
- Targets scientific communities potentially using HPC resources (not just NERSC)
- Covers problem type, workload, model architectures, scaling, hardware, software, and usage of NERSC software/resources

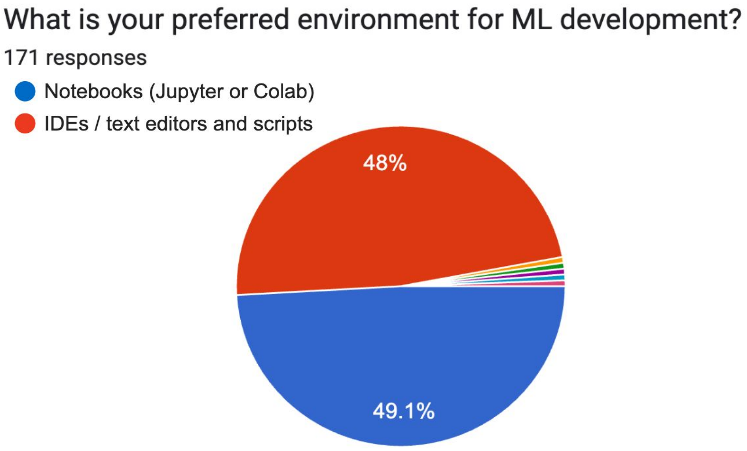


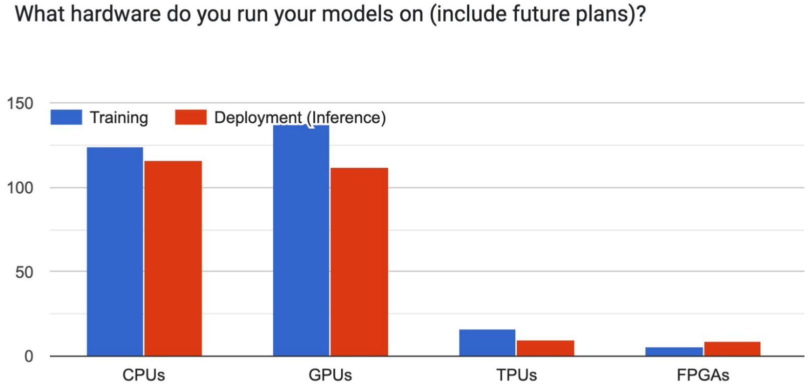


ML@NERSC 2022 Survey

Scientific AI workflows

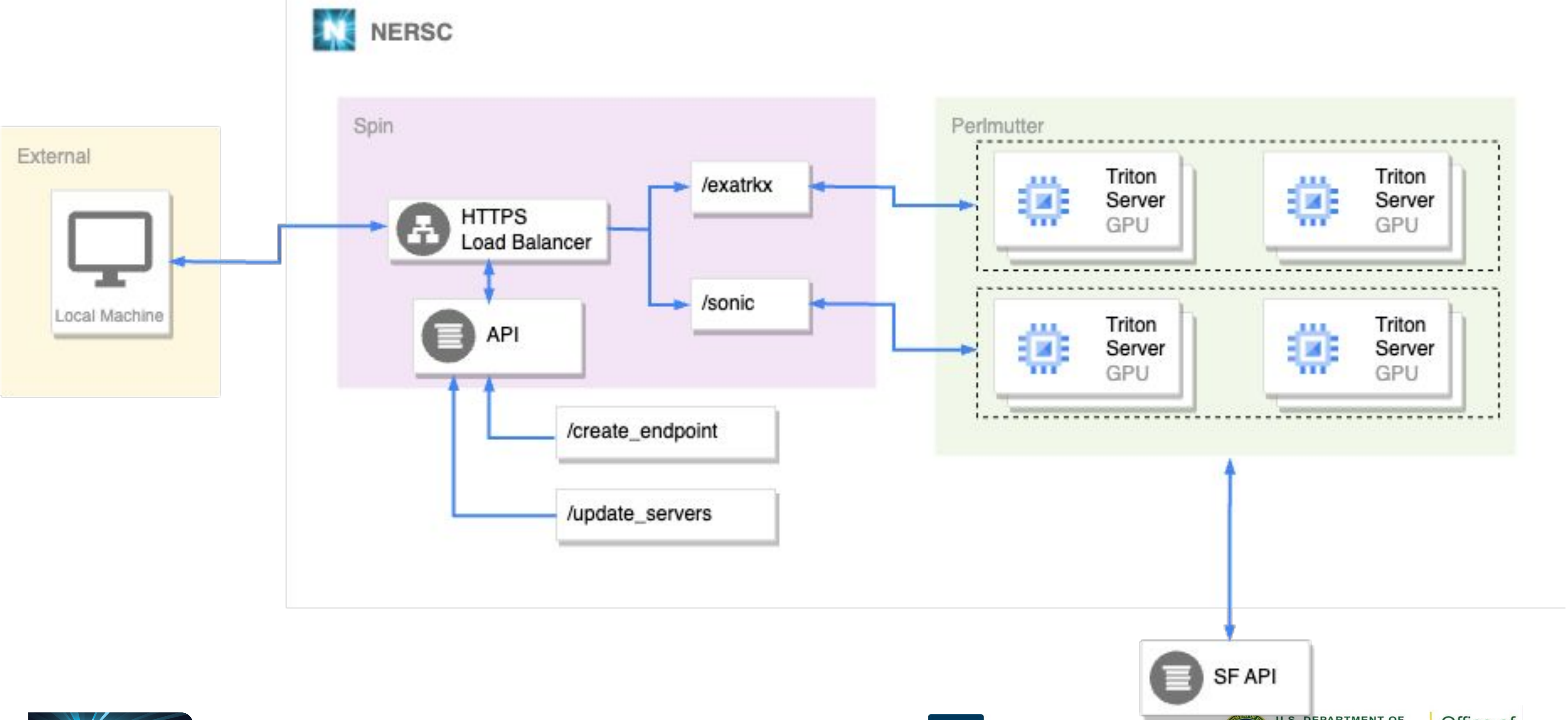






- Interesting mixture of ML applications
- Jupyter very popular for development
- CPUs still used by many
- Trends in training vs. inference

Potential Future Plans


NERSC
External
Local Machine
Spin
HTTPS Load Balancer
API
/exatrkx
/sonic
/create_endpoint
/update_servers
Perlmutter
Triton Server GPU
Triton Server GPU
Triton Server GPU
Triton Server GPU
SF API