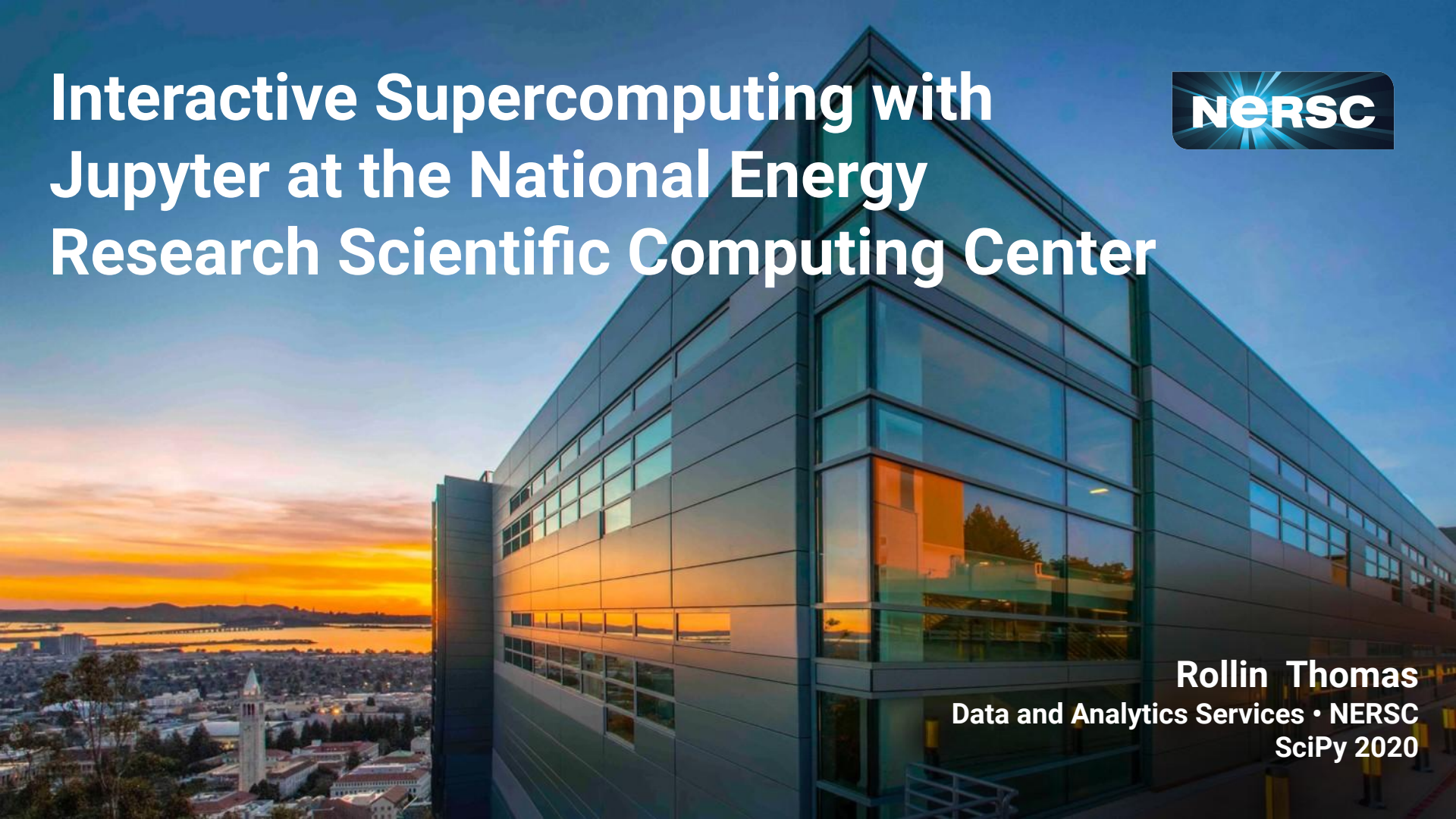
Interactive Supercomputing with Jupyter at the National Energy Research Scientific Computing Center



Rollin Thomas
Data and Analytics Services • NERSC
SciPy 2020

Jupyter at NERSC Collaborators

Rollin Thomas
Data and Analytics Services, NERSC

Shane Canon
Data and Analytics Services, NERSC

Kelly Rowland
Data Science Engagement Group, NERSC

Shreyas Cholia
Usable Software Systems Group, CRD

Matt Henderson
Usable Software Systems Group, CRD

Jon Hays
⇒ SalesForce

William Krinsman
⇒ National Microbiome Data Collective

Trevor Slaton
⇒ Apple

Labanya Mukhopadhyay
(Current, UC Berkeley)

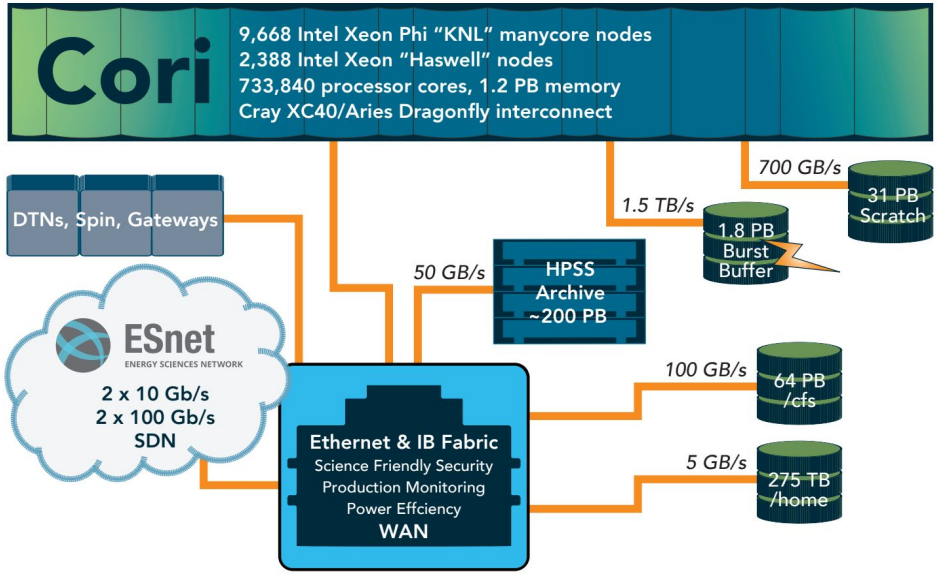
What is NERSC?

Cori (2015-today)

Gerty Cori: Biochemist, first American woman to win a Nobel Prize in science

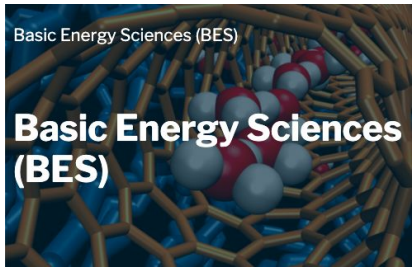
First NERSC supercomputer to support both simulation *and data analysis* workloads

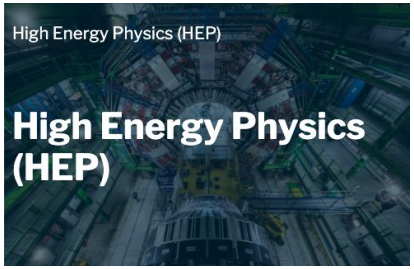
… the primary scientific computing facility for the US Department of Energy (DOE) Office of Science

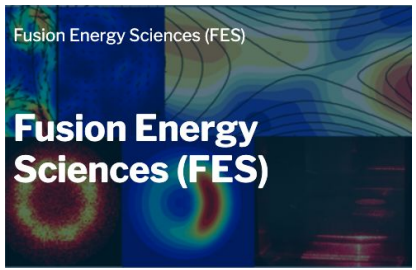


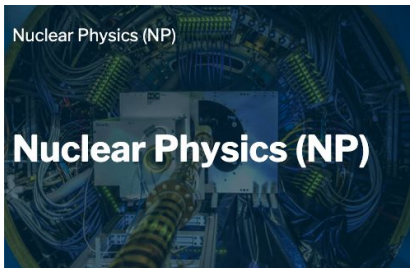
High-performance computing, networking, and storage
+ infrastructure systems, services, software and support

What Research Does NERSC Support?



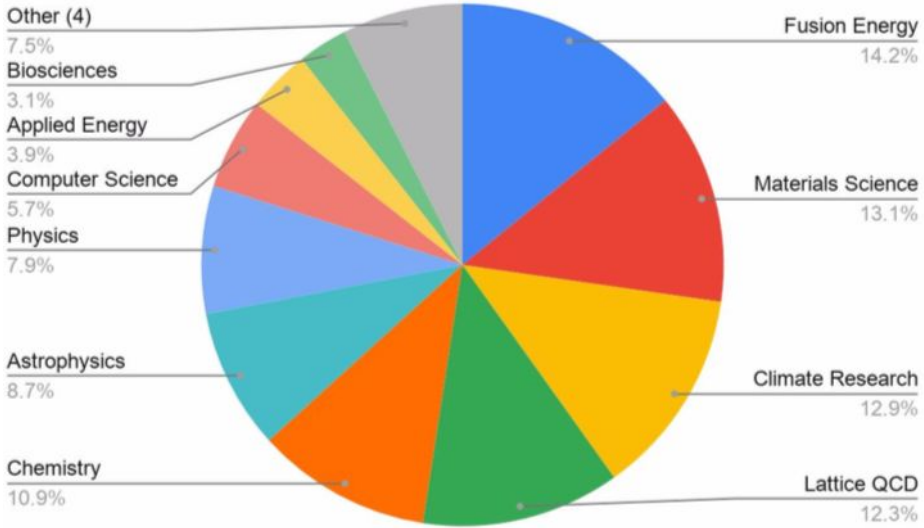






Research funded through Office of Science programs can get access NERSC compute, storage, and services.

Currently: ~7000 users, ~850 projects



Python at NERSC

FORTRAN

Python? Oh you mean Perl?

Python is OK...

Yes, we'll help you use Python!

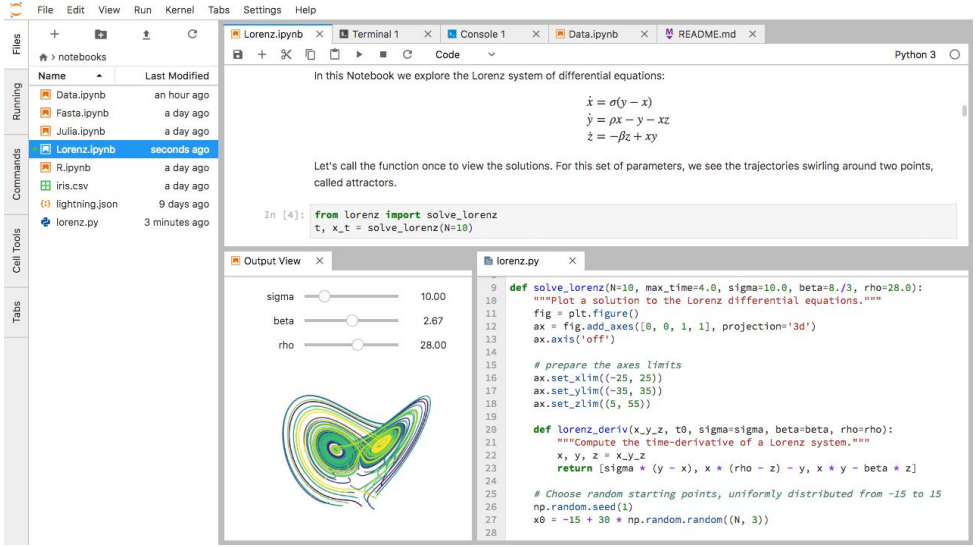
Open-source Scientific Python Libraries • Frameworks • Tools	Users begin migrating toward Python & open-source for data
Bindings for MPI through mpi4py	Multi-node parallelism on interconnects that ~only speak MPI
Support for HDF5 through h5py	Analysis of big simulation data from Scientific Python code
Optimized distros (Anaconda, Intel) Improved Packaging • Environments	Familiarity • Vendors • User-centered locus of software control
Containers on HPC	Build → Ship → Run • Mitigate slow launch at launch on PFS

What are Jupyter, JupyterLab, JupyterHub?

Interactive open-source web application

Allows you to create and share documents, "notebooks," containing:

Live code
Equations
Visualizations
Narrative text
Interactive widgets



You can use Jupyter notebooks for:

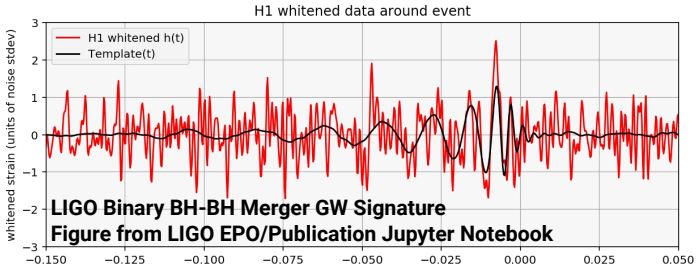
Data cleaning and data transformation
Numerical simulation
Statistical modeling
Data visualization
Machine learning
Workflows and analytics frameworks

NERSC Has Reasons to Care about Jupyter

Data 8: Foundations of Data Science, Fall 2018, Zellerbach Hall

2017 ACM Software System Award: jupyter + acm

"… ***a de facto standard for data analysis in research, education, journalism and industry.*** Jupyter has broad impact across domains and use cases. Today more than ***2,000,000 Jupyter notebooks are on GitHub***, each a distinct instance of a Jupyter application—covering a range of uses from technical documentation to course materials, books and academic publications."



Integral part of experimental and observational data science

LSST-DESC, DESI, ALS, LCLS, Materials Project, NCEM, LUX, LZ, KBase

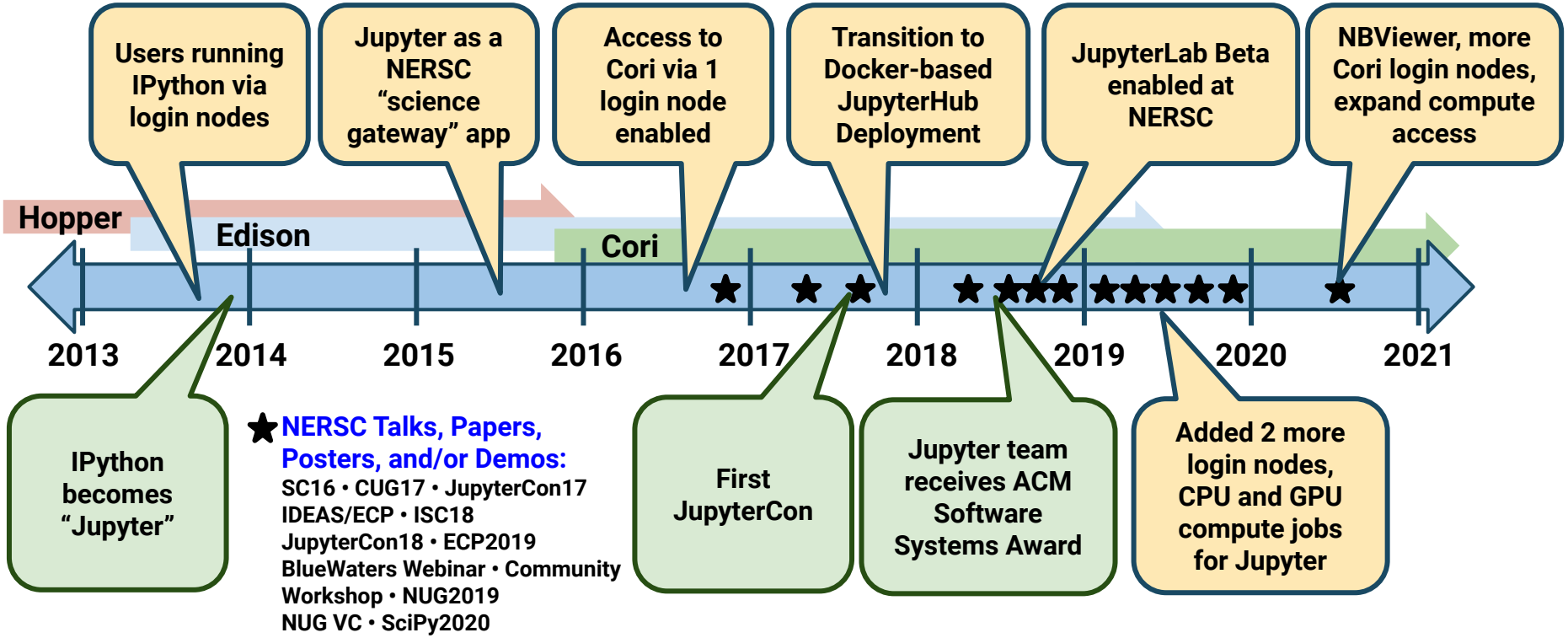
Generational shift in data science:

UCB's Data 8 course, entirely in Jupyter
"I'll send you a copy of my notebook"
Training events adopting notebooks (DL)

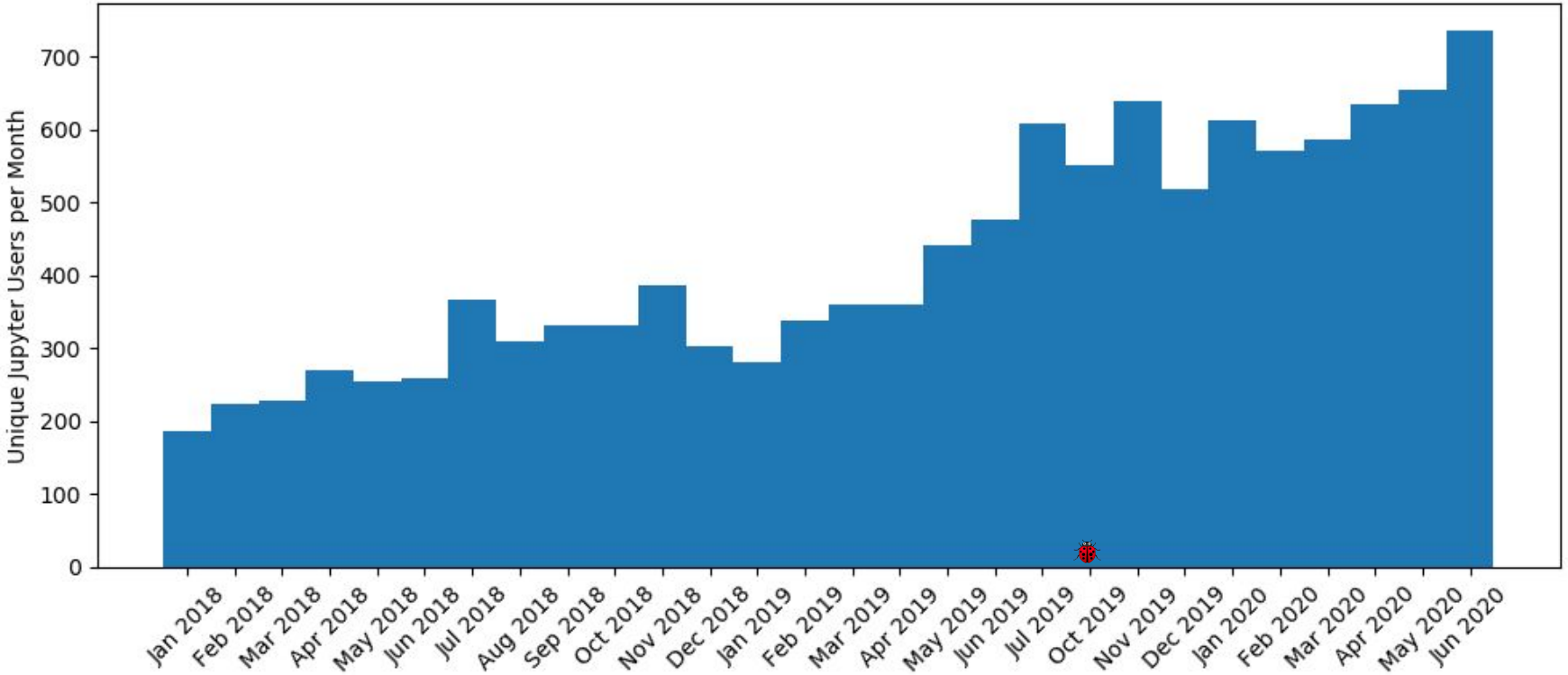
Reproducibility and science outreach:

Open source code and open science
Jupyter notebooks alongside publications

Jupyter at NERSC Timeline



Number of Jupyter Users per Month



: Bug in monitoring, data missing Aug, Sep 2019.

High Level: Jupyter Activities at NERSC

JupyterHub	How do we configure and manage Jupyter at NERSC? What resources can we expose for users and how? What delivery process best serves Jupyter users at NERSC? What changes to JupyterHub need to happen to help our users?
JupyterLab/Notebooks	What JupyterLab features would make HPC easier? What Jupyter tools help people shift workflows to Jupyter? Can we develop those features and share them?
Engagement	How do we help users, individually and at group/project level? What policies (resource access) help Jupyter users at NERSC? How do we get help, not reinvent the wheel, and share?

Support from Center management to spend time on software development is essential.

JupyterHub at NERSC

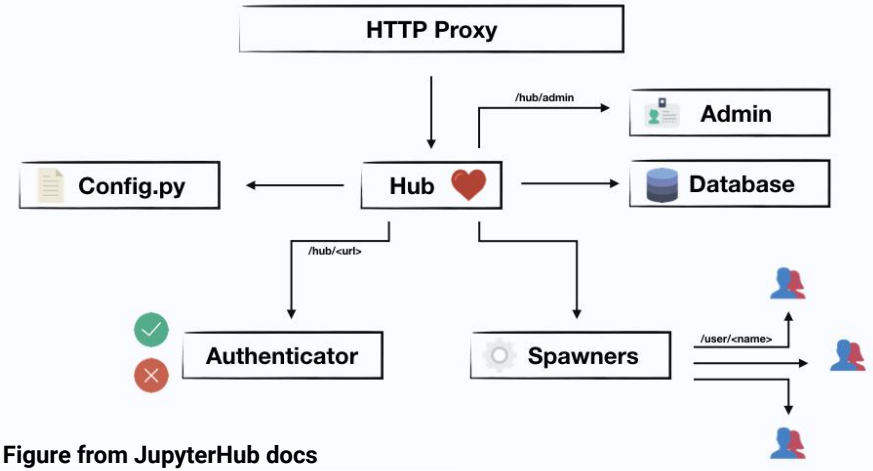


Figure from JupyterHub docs

NERSC custom Authenticator supporting multifactor auth	wrapspawner batchspawner ssh-based spawner
Identity-based access control What systems, queues for this user?	pre-spawn hook: are you over file quota?

What JupyterHub Does:

- Hub handles user login and spawns single-user servers (notebooks) on demand
- Hub launches a proxy that forwards
  - All requests to the Hub by default and
  - Notebook URL requests to running notebooks
- Can manage and interact with services

REST API for administration of Hub, users, and services.

Deployment:

Container-based but not zero-to-jupyterhub (yet?)

Custom classes and configuration

Subclass wrapspawner

Subclass batchspawner a lot

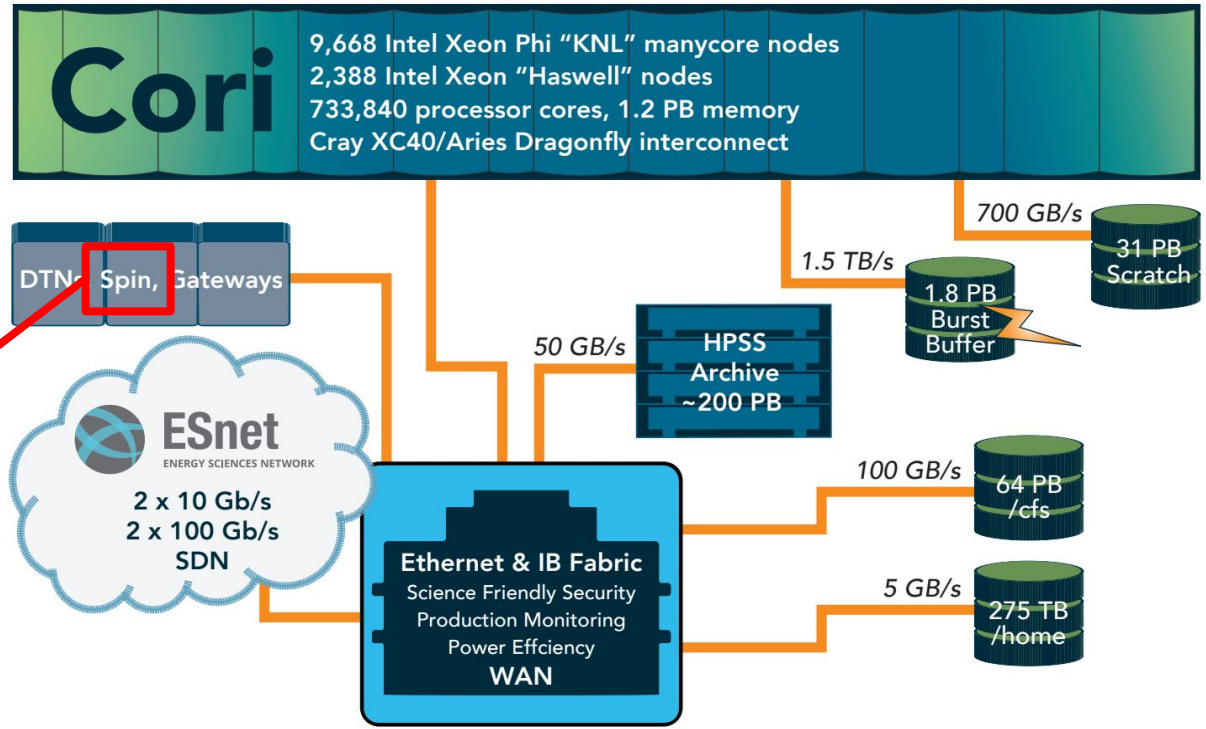
JupyterHub at NERSC Leverages Spin

Spin

RANCHER

Spin:
Not part of Cori
Containers-as-a-service
Based on Rancher
Rancher 2: k8s

User-manage services
User-facing services
Staff/infrastructure
… JupyterHub!



Nuts and Bolts: JupyterHub at NERSC

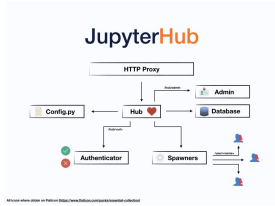
Spin

RANCHER

reverse-proxy

web-offline

jupyterhub:configurable-http-proxy
web-proxy



app-monitoring

web-jupyterhub

web-announcement

postgres:10-alpine
db-jupyterhub

app-notebooks

NGF

web-nbviewer

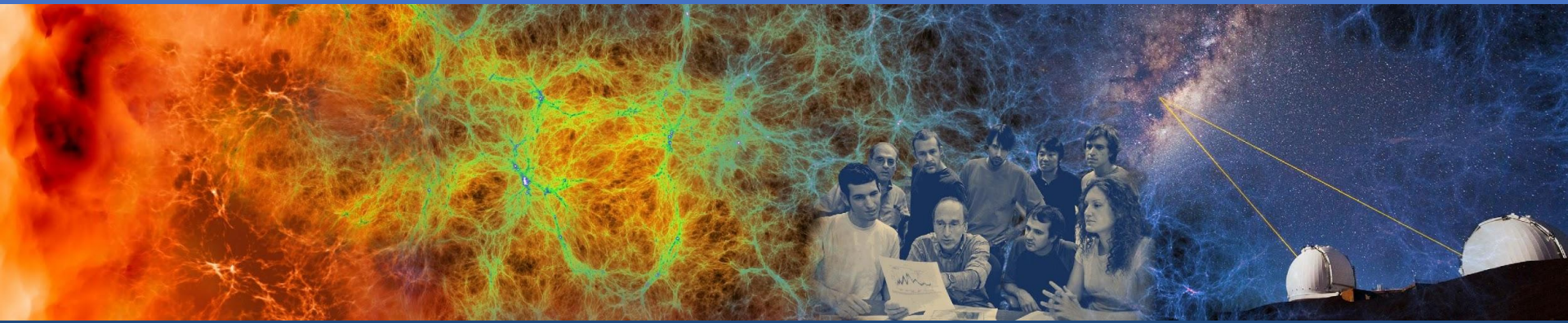
HPC

CRAY
a Hewlett Packard Enterprise company

Notebook options:

Non-batch nodes (4)
CPU compute nodes
GPU cluster compute nodes

A Guided Tour of the Jupyter Deployment at NERSC



High Level: Jupyter Activities at NERSC

JupyterHub
(jupyter.nersc.gov)

Deployment using Rancher

Configuration management

Software development
Authenticator
Spawners
Services
Pre-spawn hooks
Named servers
Custom console template
NERSC user info service calls

JupyterLab/Notebooks

Central multi-user install*

Preset kernel definitions

Notebook server configuration
e.g. jupyter-server-proxy

JupyterLab extensions
Most commonly needed

Software development
JupyterLab extensions
Contributions to JupyterLab
"Helper scripts"
Adapt workflows to notebooks

Engagement

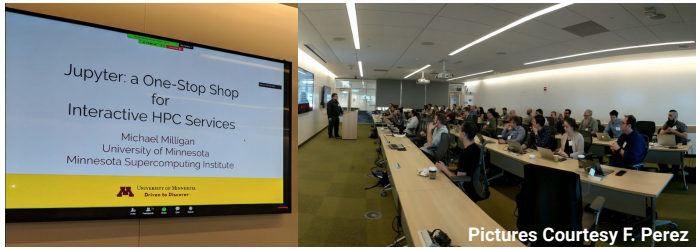
NERSC users
Directly (e.g. tickets)
Training events, mass emails
Through the hub itself

Experimental user facilities
Superfacility initiative

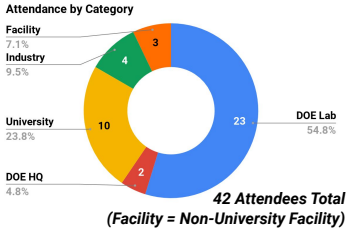
Internal stakeholders
Systems, security, networking
Staff as users
Training event support

External stakeholders
Jupyter Developers
HPC community and vendors

Jupyter Community Workshop in 2019



Pictures Courtesy F. Perez



42 Attendees Total
(Facility = Non-University Facility)

Joint Workshop w/BIDS: June 2019 at NERSC & BIDS

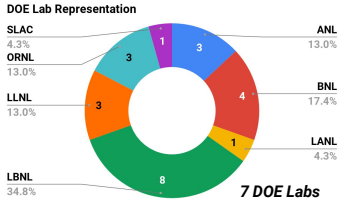
Committee: Rollin Thomas • Shane Canon • Shreyas Cholia • Kelly Rowland
Debbie Bard • Dan Allan (BNL) • Chris Holdgraf (BIDS)

Part of "Jupyter Community Workshop" Series

Funds from Bloomberg, managed by NumFOCUS and Project Jupyter

User Facilities, HPC & Data Centers Represented

NSLS-II • LSST • APS • SLAC • JGI • ARM • European XFEL
NERSC • ALCF • TACC • MSI (@UMN) • Compute Canada • ESA



7 DOE Labs

Content

Talks: **Deployment • Infrastructure • Extending Jupyter for HPC • Use Cases**
Breakouts: **Organizing Collaboration • Securing Jupyter • Sharing Notebooks**
Reproducibility • Best Practices • Future Plans • Tutorials
Roundtable Meeting with Core Jupyter Developers

Engagement

NERSC users
Directly (e.g. tickets)
Training events, mass emails
Through the hub itself

Experimental user facilities
Superfacility initiative

Internal stakeholders
Systems, security, networking
Staff as users
Training event support

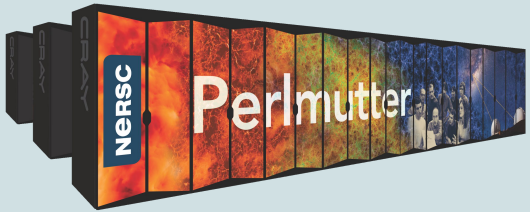
External stakeholders
Jupyter Developers
HPC community and vendors

Future Jupyter Activities at NERSC

JupyterHub
(jupyter.nersc.gov)

Migration to Rancher 2 (k8s)

Software development
SSO/Federated login
Expanded access to computes
Environment/image service
Integration with Perlmutter P1



P1: NVIDIA A100 GPUs
Arriving this year

JupyterLab/Notebooks

Per-user/collab JLab installs
User control of sw stack
Per-user/group extensions
Binder for HPC
Facilitate sharing notebooks
Ease-of-use for Dask etc.

Issues
Educating users on how to take charge of JupyterLab

Innovating on Jupyter when fancy JavaScript toolkits are increasingly how it's done

Engagement

NERSC users
Directly (e.g. tickets)
Training events, mass emails
Through the hub itself

Experimental user facilities
Superfacility initiative

Internal stakeholders
Systems, security, networking
Staff as users
Training event support

External stakeholders
Jupyter Developers
HPC community and vendors

Conclusion: What Have We Learned?

Things that work well for HPC folks working on Jupyter:

Manage JupyterHub service through containers
Leverage container management and orchestration frameworks for Jupyter

Keep a delivery cycle that impels you to deliver reliability and innovation
Develop productive relationships with internal stakeholders
Work flexibly with your colleagues on alternatives when necessary

Identify and engage with stakeholders with focus on what matters to your users
Share code with the Jupyter community, and try to *retire* code more than you write!

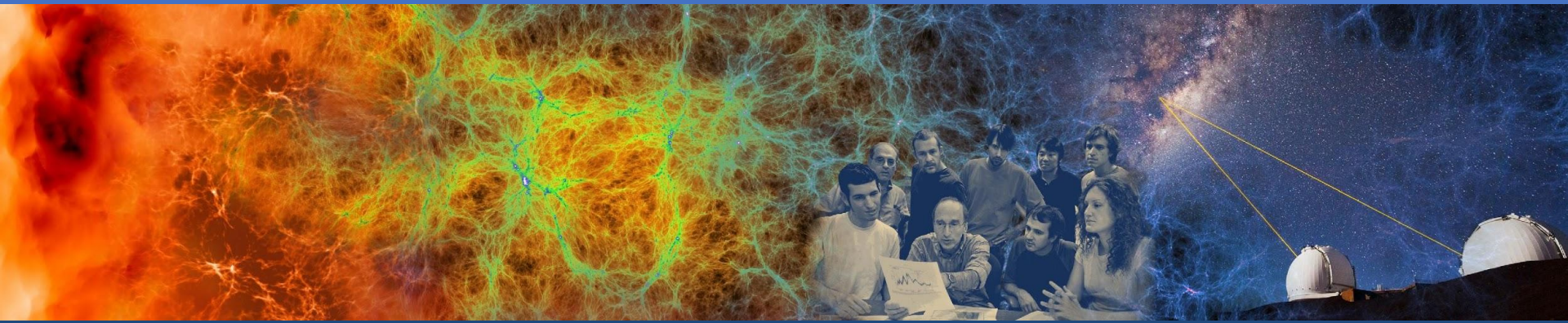
Challenging things for HPC staff working on Jupyter:

Successful HPC Management commits to innovation through software development
HPC devs can write Python, but JavaScript, React, TypeScript…?
If you are bad at UI design, *and you are,* get a UI/UX expert to help you
Resources for Jupyter can be hard to acquire sometimes within the center
Technical problems can be challenging, but they are most always temporary

Useful Links

NERSC:	https://www.nersc.gov/
Jupyter@NERSC:	https://jupyter.nersc.gov/ **(requires NERSC account)**
Spin:	https://www.nersc.gov/systems/spin/
Rancher:	https://rancher.com/
Jupyter Community Workshop 2019:	https://jupyter-workshop-2019.lbl.gov/agenda [YouTube]
Jupyter @NERSC Internships!!!:	https://www.nersc.gov/research-and-development/internships/ **or please contact rcthomas@lbl.gov !!!**
batchspawner:	https://github.com/jupyterhub/batchspawner
wrapspawner:	https://github.com/jupyterhub/wrapspawner
jupyterlab-favorites:	https://github.com/NERSC/jupyterlab-favorites
jupyterlab-recents:	https://github.com/NERSC/jupyterlab-recents
jupyterhub-announcement:	https://github.com/rcthomas/jupyterhub-announcement
Project Jupyter:	https://jupyter.org/
JupyterHub:	https://jupyterhub.readthedocs.io/en/stable/
JupyterLab:	https://jupyterlab.readthedocs.io/en/stable/

Thank You!



NERSC Operates Supercomputers

Cray-2

Cray T3E

Seaborg

Franklin

Carver

PDSF & Genepool

Hopper

Edison

Jupyter Matters to our Users

Users appreciate Jupyter @ NERSC...

"Great interactive workflow (e.g. for postprocessing) via JupyterHub"

"New jupyter notebooks are awesome!"

"I really like the jupyter interface."

"... the ability to access data from the scratch directories through the Jupyter hub is very important to my workflow. The Jupyter hub has been running more and more consistently, but it still seems to lag or stall sometimes. I guess ***my only thought on how to improve (currently)*** would be to improve the stability of the Jupyter hub."

"… jupyter notebooks are very important for me: ***The 3 most important things in life: food, shelter and jupyter… everything else is optional.***"

"I absolutely love the fact that I can use the Jupyter hub to access the Cori scratch directory. This allows me to analyze data through the browser ... or to quickly check that simulation runs are going as expected without having to transfer data to a different location. ***I actually also have access to other supercomputer clusters, but this is one of the biggest reasons I mainly use Cori and Edison for debugging and production runs.***"

...but need increased stability and to scale up.

"I would really appreciate it if jupyter.nersc.gov wouldn't go down as much as it does."

"MPI cannot be used in jupyter notebook as well, where the jupyter hubs run on login nodes (unless when using the compute nodes through SLURM.)"

NERSC Annual User Survey Comments (2018, for CY17) & User Comments

NERSC and COVID-19

NERSC is still up and running:

SF Bay Area SIP began in mid-March
System utilization went up a bit
NERSC staff mostly still working from home

NERSC supporting COVID-19 research:

National COVID-19 HPC Consortium Member
ExaLearn Exascale Computing Project

Resources for COVID-19 researchers:

Hours from Director's Discretionary Reserve
Trying to minimize impact on existing workload but accommodate urgent needs
Expedited access to staff, services, hardware