Data Ecosystem Overview

New User Training
June 16, 2020

Wahid Bhimji
Data And Analytics Services Group

Afternoon Schedule

1:00 pm	Data Ecosystem Overview	Wahid Bhimji
1:10 pm	Workflows at NERSC	Bill Arndt
1:30 pm	File Systems and Burst Buffer	Wahid Bhimji
1:50 pm	Data Transfer Best Practices	Lisa Gerhardt
2:10 pm	I/O Best Practices	Quincey Koziol
2:30 pm	Break	
2:50 pm	Python at NERSC	Laurie Stephey
3:10 pm	Jupyter at NERSC	Rollin Thomas
3:30 pm	Shifter	Shane Canon
3:50 pm	Deep Learning	Mustafa Mustafa

Production Data Stack

Capabilities	Technologies				
Data Transfer + Access	globus online	GridFTP	jupyter	!M	python django	newt
Workflows	Parsl	GNUparallel	papermill	FireWorks	TaskFarmer	
Data Management	HDF	netCDF	ROOT Data Analysis Framework	mongoDB	MySQL	PostgreSQL
Data Analytics	python	R	Spark	julia	TensorFlow	PyTorch
Data Visualization		VisIt	ParaView			

Cori's Data Friendly Features

12 32-core Haswell
500 GB Login Nodes

Jupyter
Notebook
Node

768 GB "bigmem"
Haswell Nodes

Pipeline/Workflow
Management Nodes

Serial/Shared-node
Queue
Transfer Queue

Containerized
Environments

SHIFTER

Real-Time Queues
for Co-Scheduling
w/Experiments

slurm
workload manager

Streaming Data to
Compute Nodes

Cray DataWarp:
Burst Buffer for
I/O acceleration

CRAY
DATAWARP

Interactive Queue:
64 Nodes x 4 Hours

External Network
Access to/from
Compute Nodes

NERSC

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science

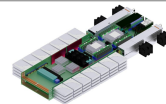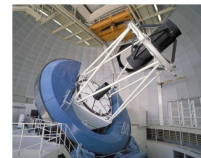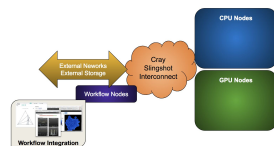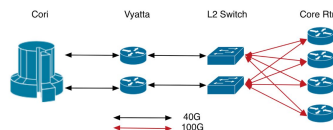| Data Features	Cori experience	Perlmutter
I/O and Storage	Burst Buffer	All-flash file system: performance with ease of data management
Analytics - **Production stacks** - **Analytics libraries** - **Machine learning**	User defined images with Shifter NESAP for data New analytics and ML libraries	Production analytics workflow benchmark. Data apps in NESAP at start Optimised analytics libraries and deep learning application benchmark
Workflow integration	Real-time and interactive queues	SLURM co-scheduling Workflow nodes integrated
Data transfer and streaming	SDN	Slingshot ethernet-based converged fabric
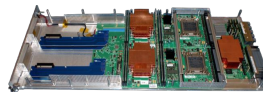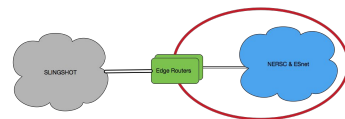
Perlmutter: Workflow, Analytics and Data
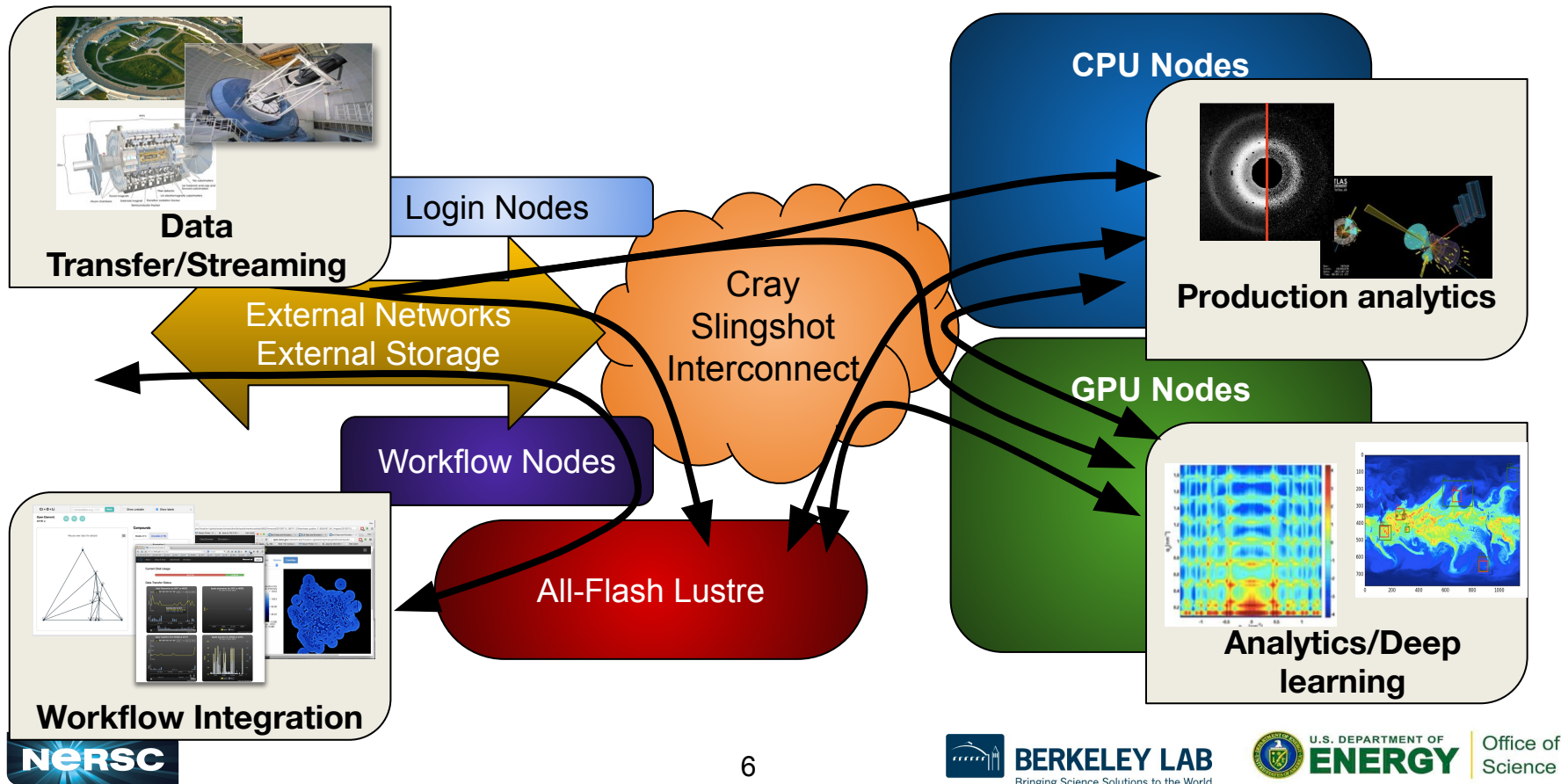
NERSC Data Crew Are Here to Help

- **Feel free to reach out to us**

    E.g. on NERSC users slack

    https://www.nersc.gov/users/NUG/nersc-users-slack/

    Provide feedback, critique
    Tell us about interesting science problems!

- **Questions?**

Thank You and Welcome to NERSC!