Nautilus Distributed Hypercluster

Mahidhar Tatineni
Director, User Services, San Diego Supercomputer Center

4th GLOBAL RESEARCH PLATFORM WORKSHOP
Oct 9-10, 2023

Acknowledgement for presentation:
Dmitry Mishin, UCSD/SDSC

Nautilus: Past and the future!

- Nautilus has been the K8S infrastructure of PRP for the last 5+ years and incubator for a lot of the ideas that led to a **successful** Category II system proposal to the NSF.
- Category II system has operations funded for 5 years and longer with the possibility of a renewal. Provides the support for Nautilus going forward.
- Introduces several new hardware options - composable hardware, FPGA hardware, and a significant addition of FP32 GPUs => Nautilus continues to be the K8S infrastructure of NRP at present and will continue in the future!

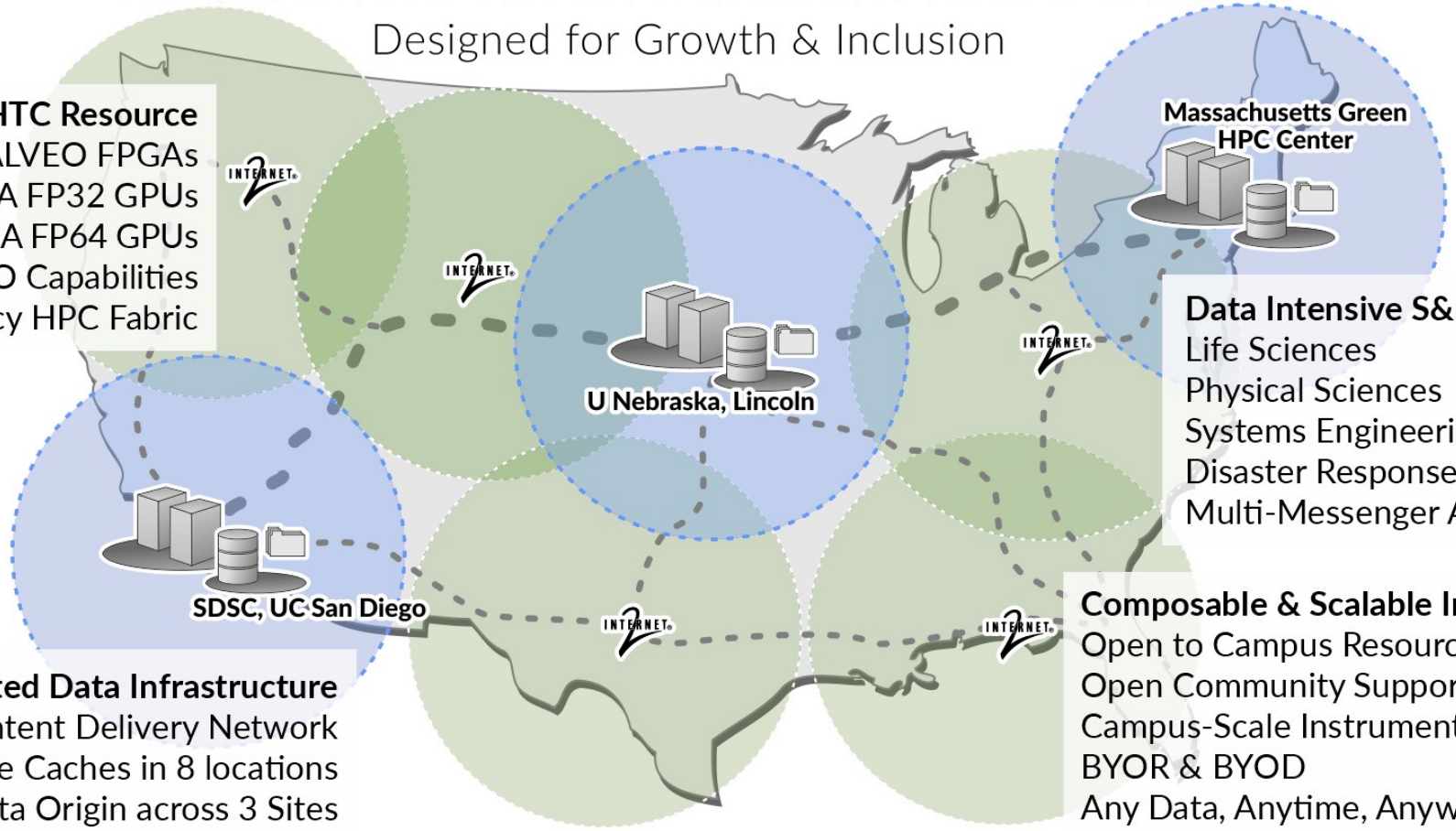
The NSF Cat-II Program

- NSF supports via the Cat-II program novel systems ideas.
  - 3 year “testbed” phase
    - The PI owns the resource and has (some) freedom regarding who uses it.
      - No requirements for making it available via any specific allocation mechanism.
    - It is expected that not all features work on day 1.
      - 3 years of experimentation & development of features
  - 2 year “allocation” phase
    - The resource is made available via an NSF supported allocation mechanism.
  - The solicitation mentions the possibility of an additional 5-year renewal

- **Ideal program to target addition of innovative hardware to Nautilus and use the testbed phase to build a user base for the new hardware (like FPGAs). Also, funds Ops for duration of project**

NATIONAL RESEARCH PLATFORM

Designed for Growth & Inclusion



HPC/HTC Resource
32 ALVEO FPGAs
A10 288 NVIDIA FP32 GPUs
80GB A100 64 NVIDIA FP64 GPUs
Tbps WAN IO Capabilities
GigaIO's Low Latency HPC Fabric

Data Intensive S&E
Life Sciences
Physical Sciences
Systems Engineering
Disaster Response
Multi-Messenger Astrophysics

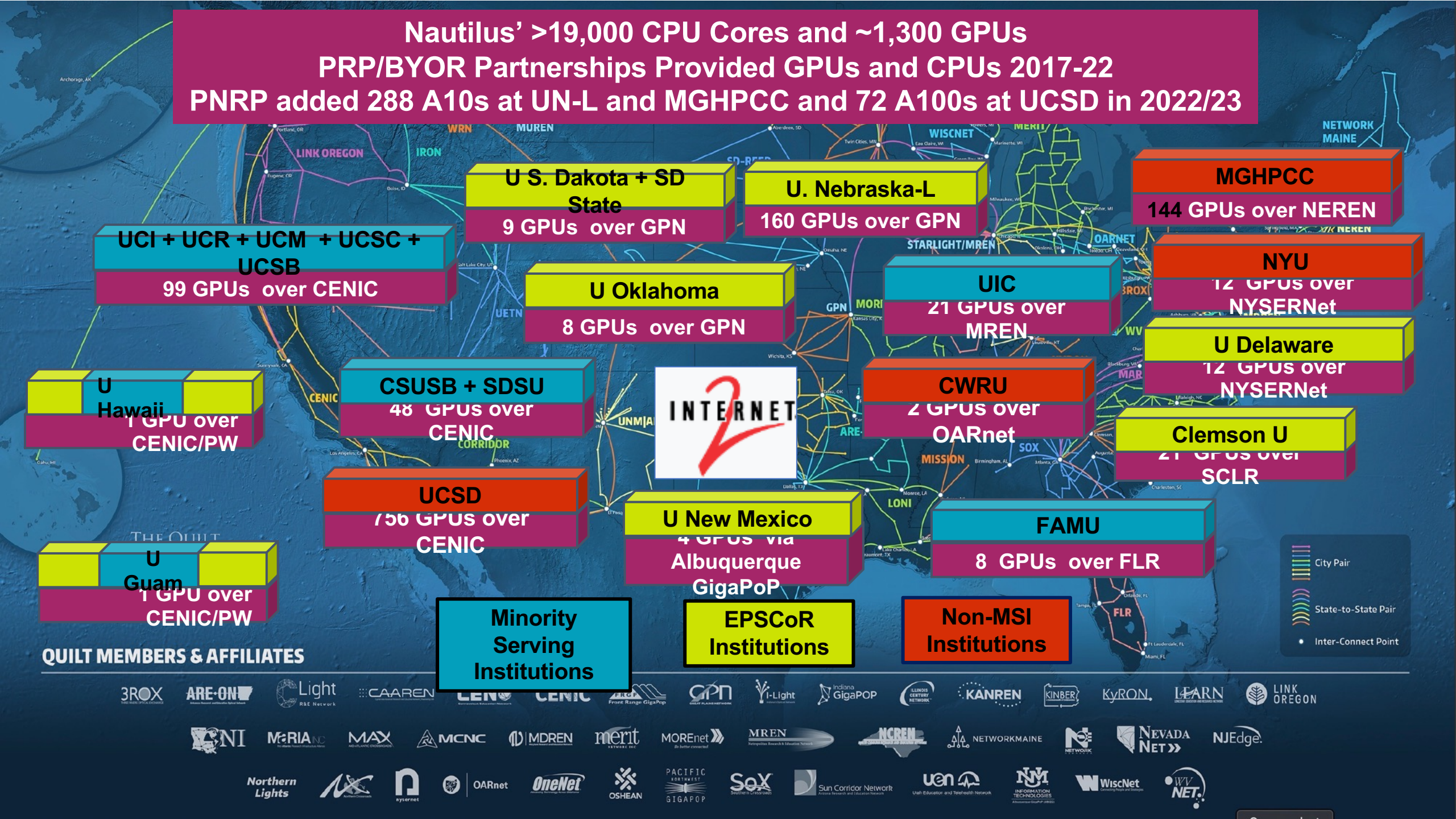
Distributed Data Infrastructure
National Scale Content Delivery Network
50TB 100Gbps NVMe Caches in 8 locations
4.5PB Distributed Data Origin across 3 Sites

Composable & Scalable Innovation
Open to Campus Resource Integration
Open Community Support Model
Campus-Scale Instrument integration
BYOR & BYOD
Any Data, Anytime, Anywhere

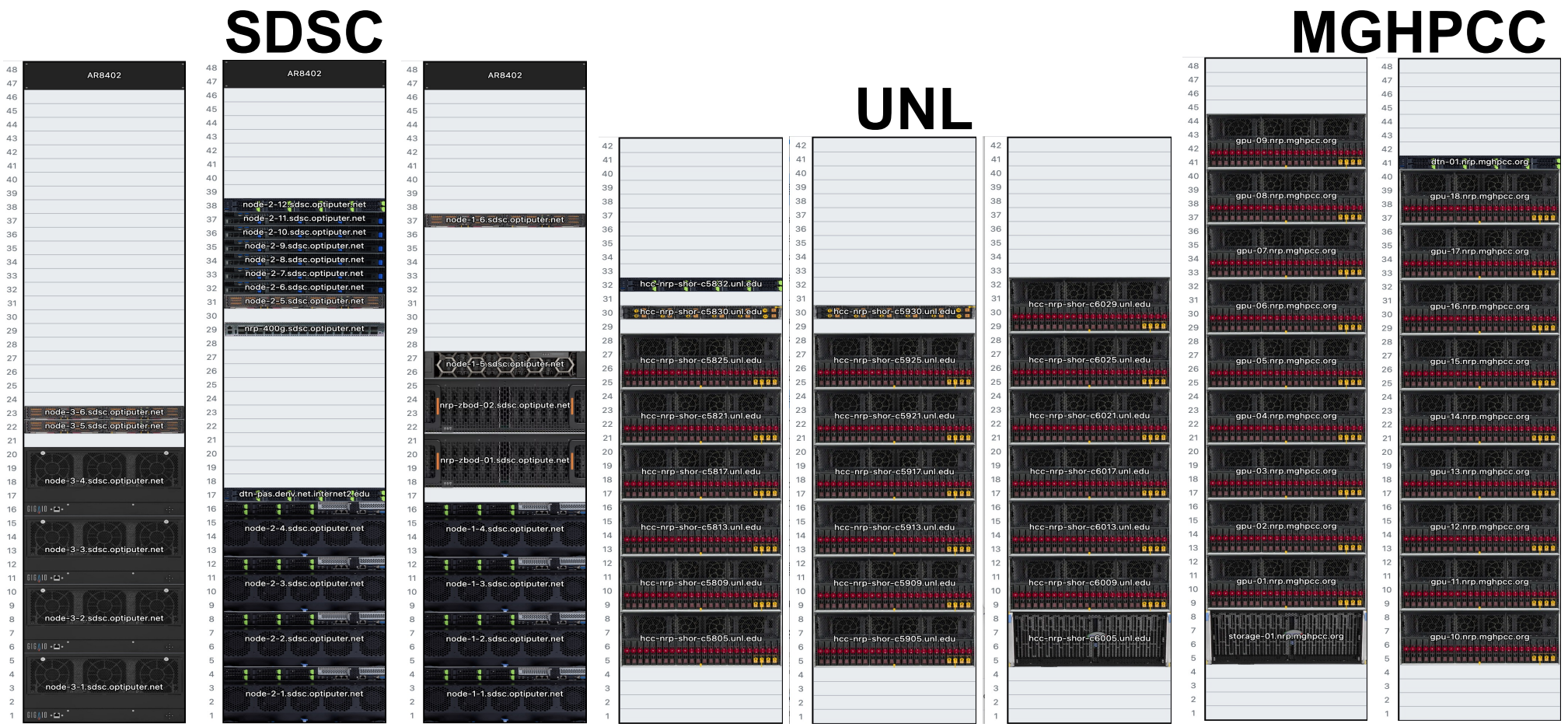
5-year project: $5M for Acquisition and Deployment; $7.25 for Operations and Maintenance

PI = Wuerthwein; Co-PIs: DeFanti, Rosing, Tatineni, Weitzel

Funded as NSF 2112167


Nautilus' >19,000 CPU Cores and ~1,300 GPUs
PRP/BYOR Partnerships Provided GPUs and CPUs 2017-22
PNRP added 288 A10s at UN-L and MGHPCC and 72 A100s at UCSD in 2022/23
U S. Dakota + SD State
9 GPUs over GPN
U. Nebraska-L
160 GPUs over GPN
MGHPCC
144 GPUs over NEREN
UCI + UCR + UCM + UCSC + UCSB
99 GPUs over CENIC
U Oklahoma
8 GPUs over GPN
UIC
21 GPUs over MREN
NYU
12 GPUs over NYSERNet
U Delaware
12 GPUs over NYSERNet
U Hawaii
1 GPU over CENIC/PW
CSUSB + SDSU
48 GPUs over CENIC
CWRU
2 GPUs over OARnet
Clemson U
21 GPUs over SCLR
UCSD
756 GPUs over CENIC
U New Mexico
4 GPUs via Albuquerque GigaPoP
FAMU
8 GPUs over FLR
U Guam
1 GPU over CENIC/PW
Minority Serving Institutions
EPSCoR Institutions
Non-MSI Institutions
City Pair
State-to-State Pair
Inter-Connect Point
QUILT MEMBERS & AFFILIATES

PNRP contribution to Nautilus



PNRP contribution to Nautilus

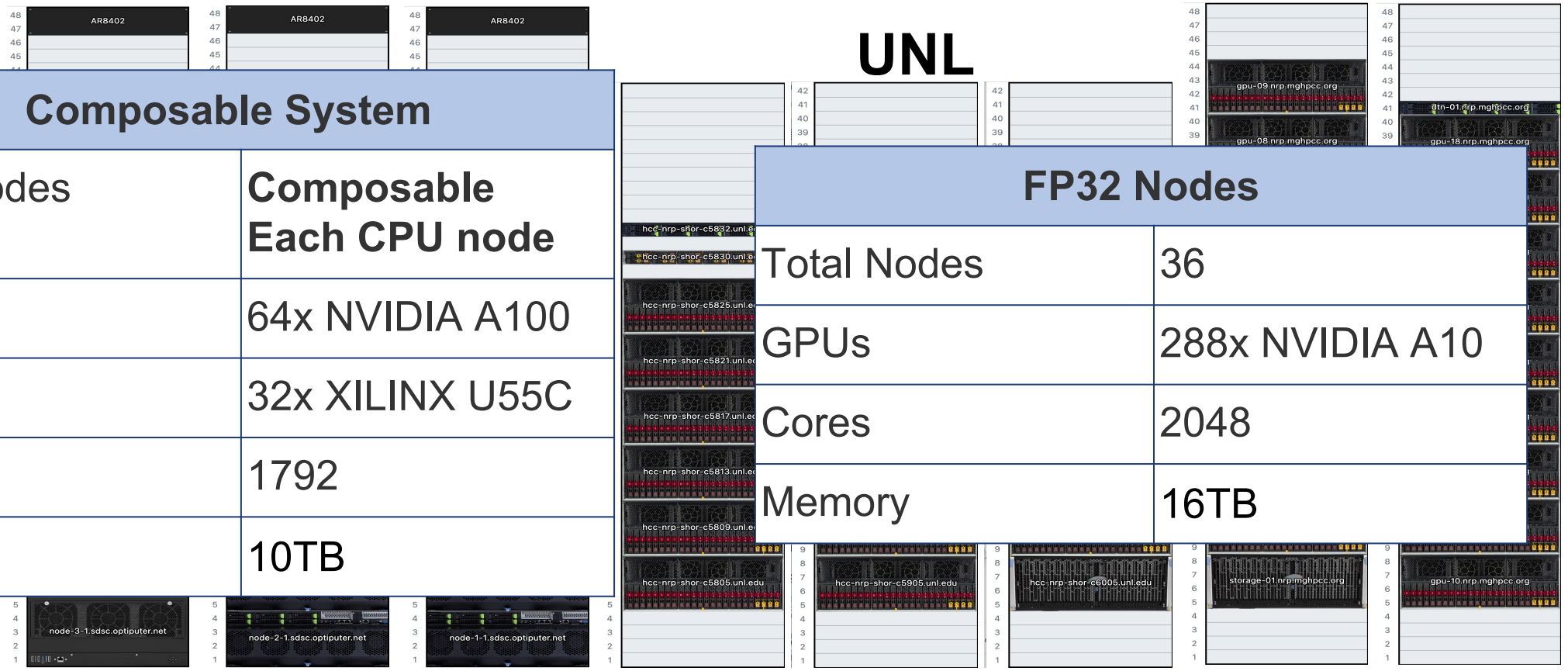
SDSC

Composable System	
Total Nodes	**Composable Each CPU node**
GPUs	64x NVIDIA A100
FPGAs	32x XILINX U55C
Cores	1792
Memory	10TB

UNL

FP32 Nodes	
Total Nodes	36
GPUs	288x NVIDIA A10
Cores	2048
Memory	16TB

MGHPCC



How to compose nodes

- Composition is done by hand by admins
- Composition of FPGA and GPUs are done by GigaIO User Interface
- Exporting NVMe uses JSON and a tool on the NVMe node

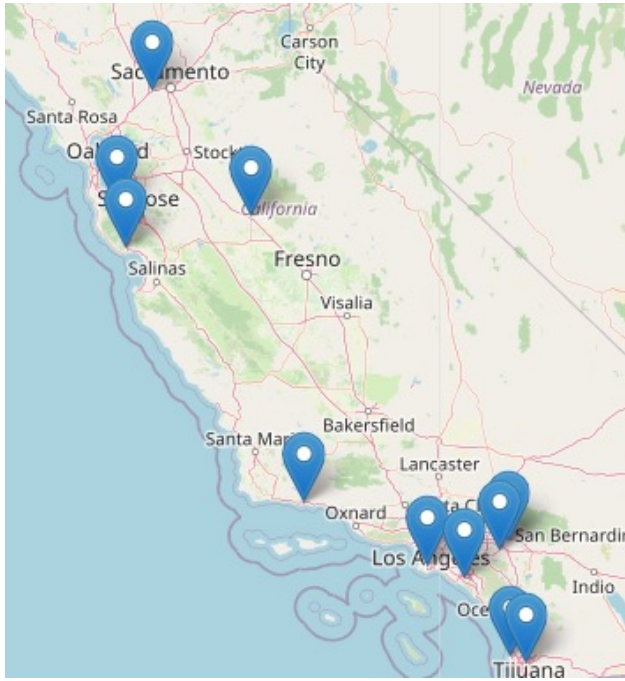
Nautilus

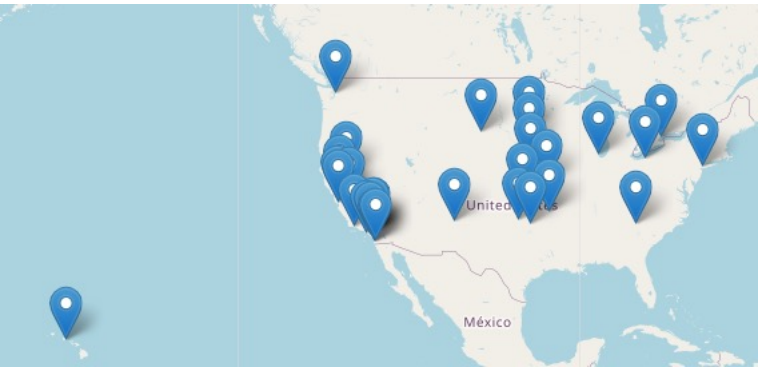
- System managed with  kubernetes
- Logins managed by  CILogon
- Truly federated resource!

Bring Your Own Resources

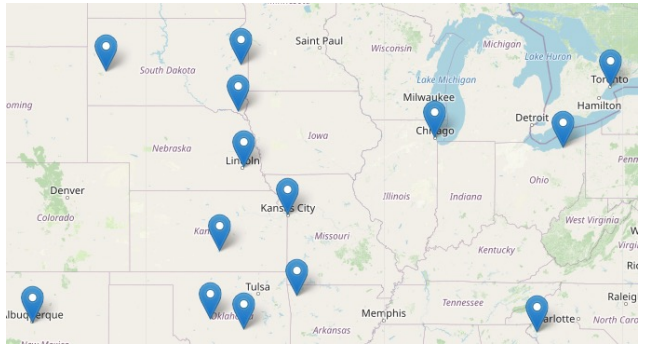
- Researchers and providers can add their own resources
- Researchers provide hardware maintenance, power, cooling, networking
- NRP admins support the OS and above
- 5 minutes and a single ansible command from bare Ubuntu OS to Nautilus node
- Storage drives management and volumes creation in kubernetes
- Ansible : https://gitlab.nrp-nautilus.io/prp/nautilus-ansible

BYOR Status: Hardware on NRP is globally distributed



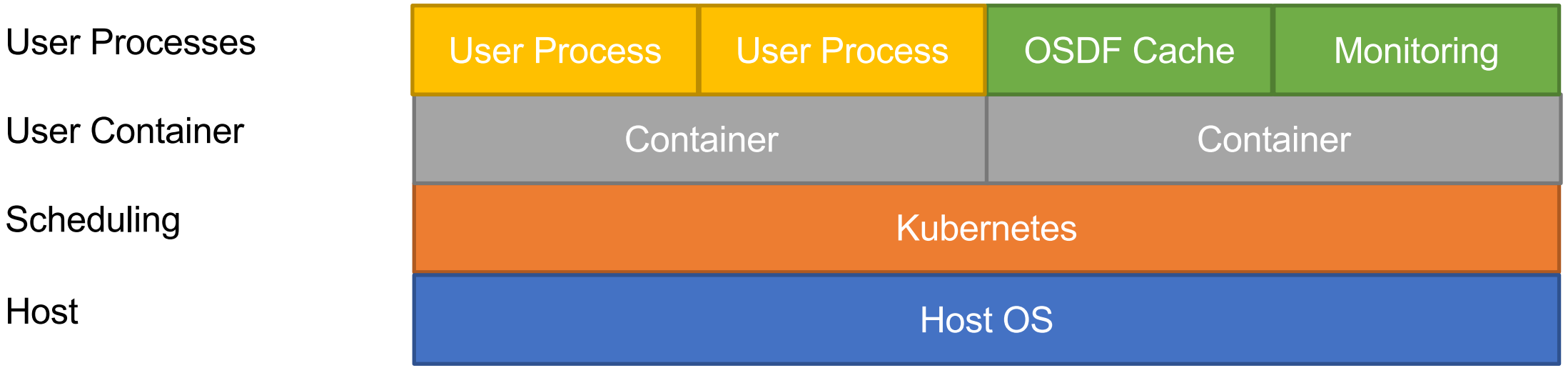


Hardware in many unusual locations



System Software

- Services are also hosted in Kubernetes, such as the OSDF caches



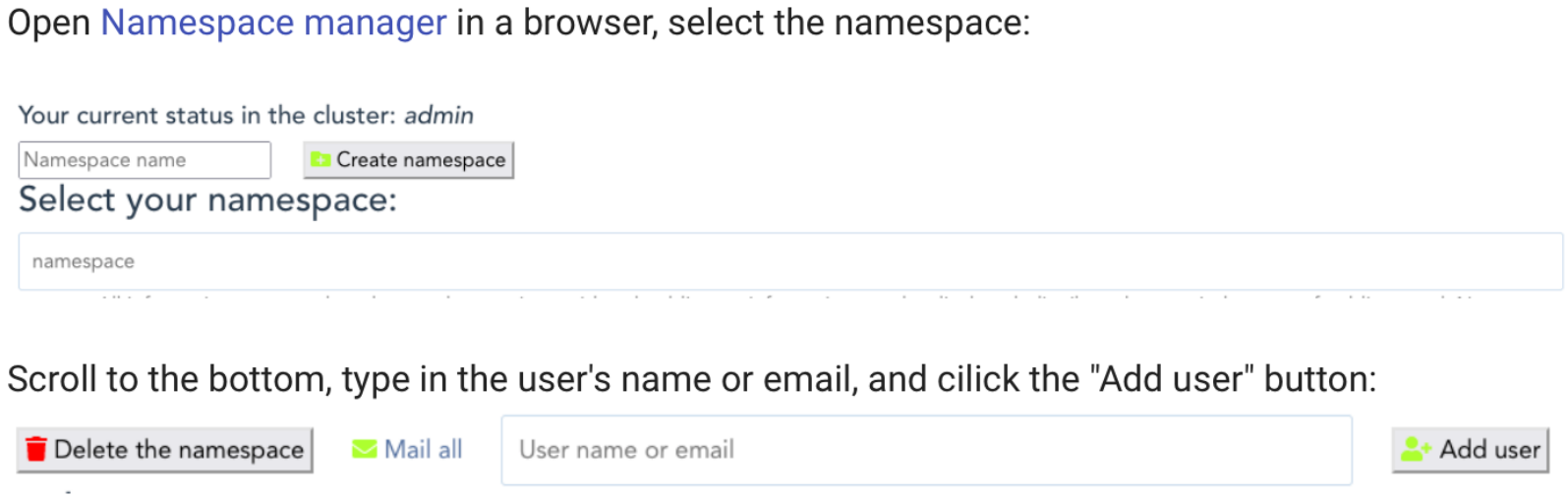
Credit: Derek Weitzel, UNL

User Access

- All access to PNRP is through CILogon federated login
- Institutions enforce their own rules, but most use two factor login
- No passwords are ever transferred to PNRP, only OAuth tokens from CILogon

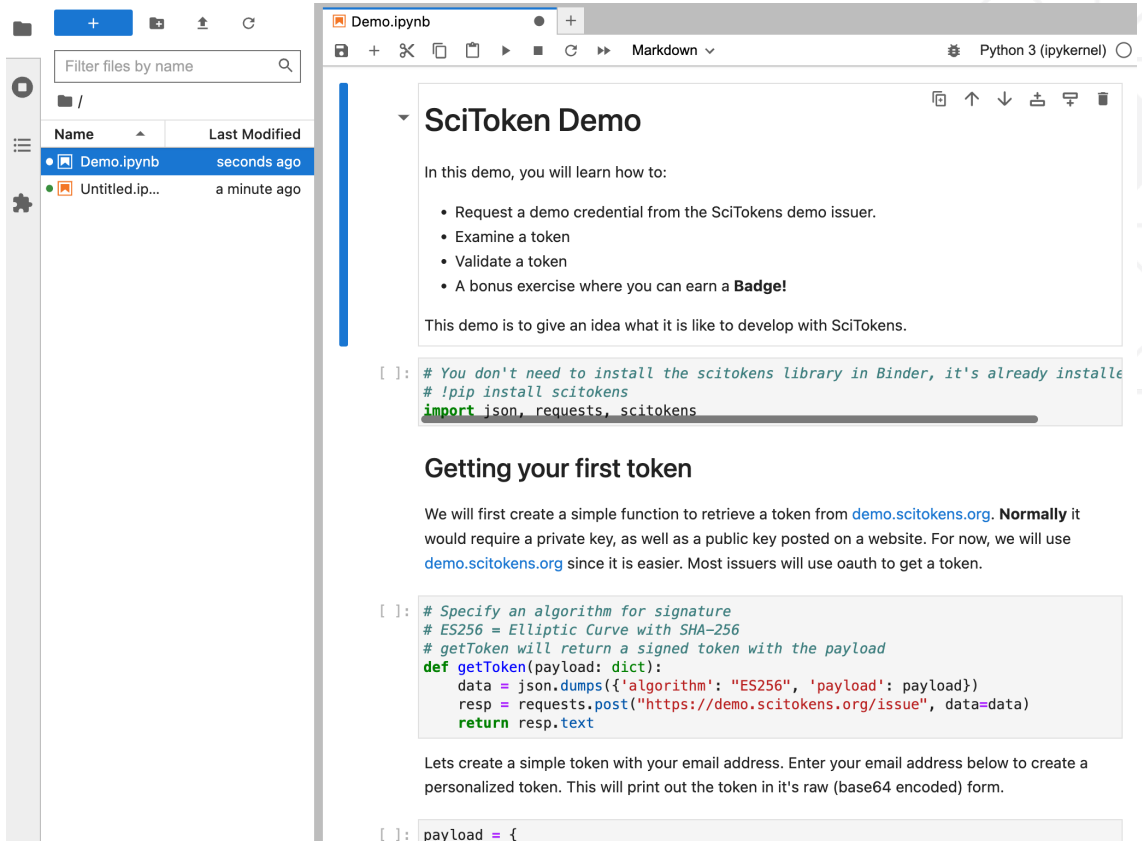
Nautilus User Management

- Initial user login with CILogon
- Cluster admins can promote users to namespace admins
- Delegate user addition to specific namespaces to their admins. So XYLab namespace admin can add all their project members to their namespace (without Cluster admin intervention)

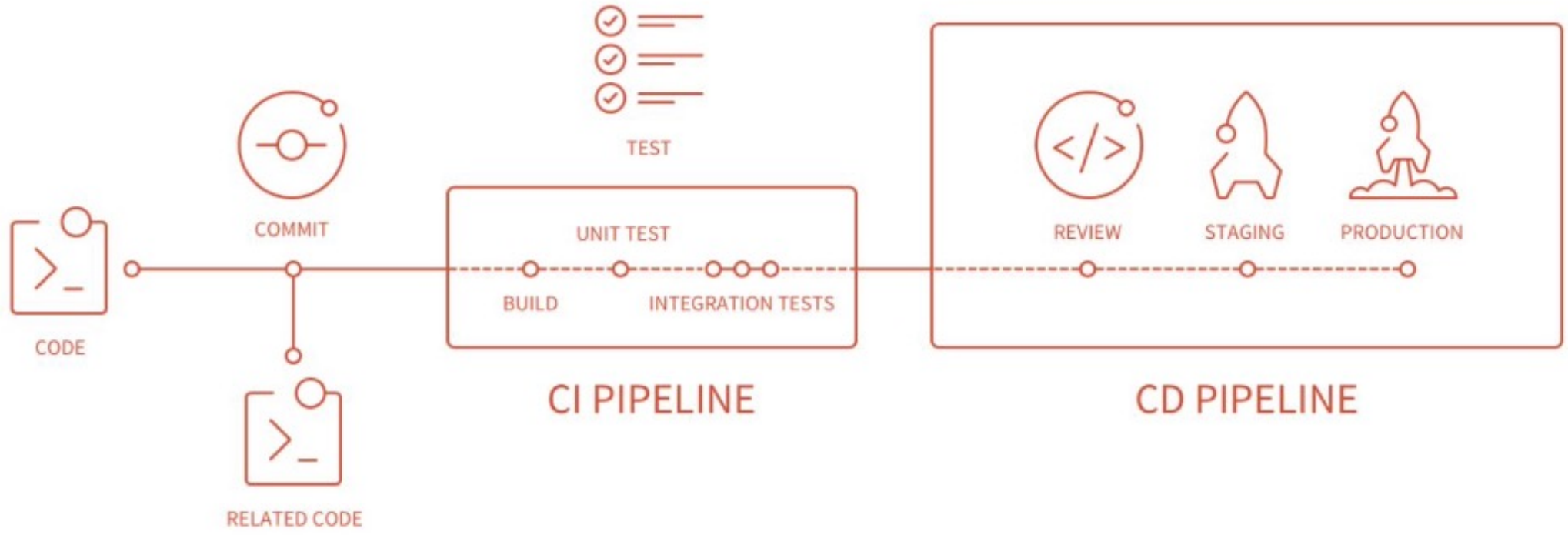


Nautilus Jupyter Optons

- JupyterHub service, use institutional credentials via CILogon
- Jupyter pods using, GUI w/ port forwarding
- Custom Jupyter deployment w/ config in Gitlab



Automating applications deployment: Gitlab CI/CD



Credit: Dmitry Mishin, SDSC

Utilization monitoring

Prevent resource wastage

A single user can't submit more than 256 pods violating the following limits:

GPU: > 0.4

CPU: 0.2 - 2

Memory: 0.2 - 1.5

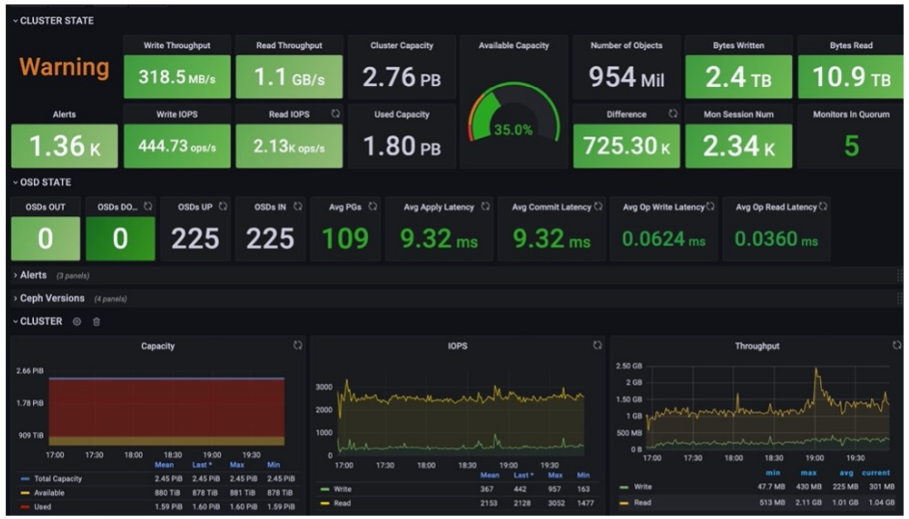
http://cilogon.org/serverA/users/10844

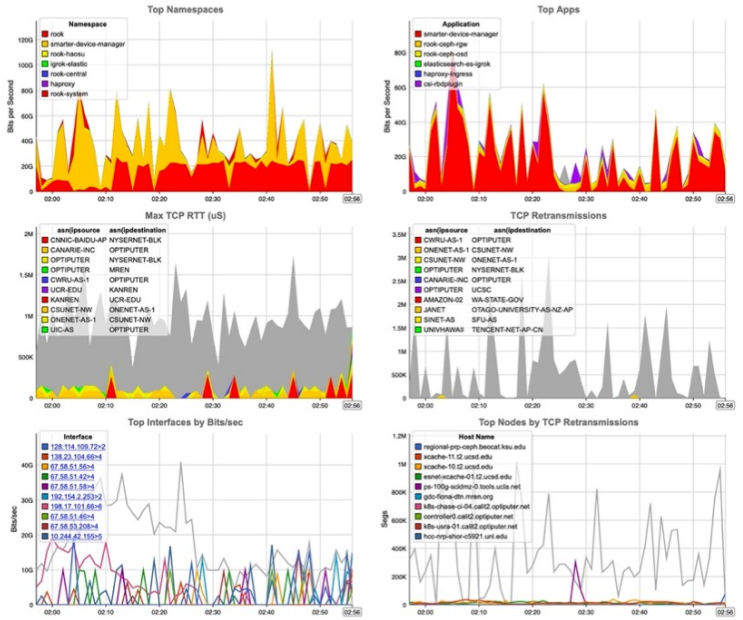
Pod Name	% GPU utilization (1 is 100%)	% CPU utilization (1 is 100%)	% Memory utilization (1 is 100%)	GPUs requested	CPUs requested	Memory request
atlas/v4cvmfs-starlight-bbdb6b78c-sj667		0.00	0.00	0	4.00	35 GB

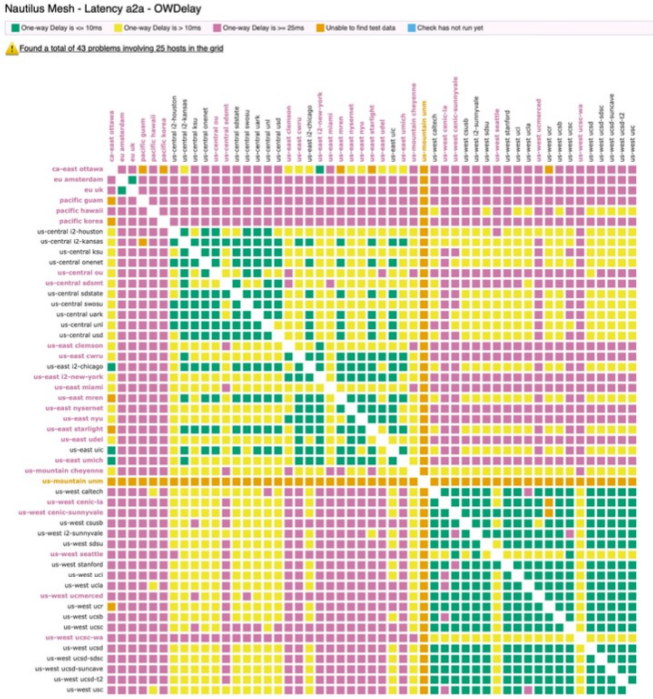
http://cilogon.org/serverA/users/11152391

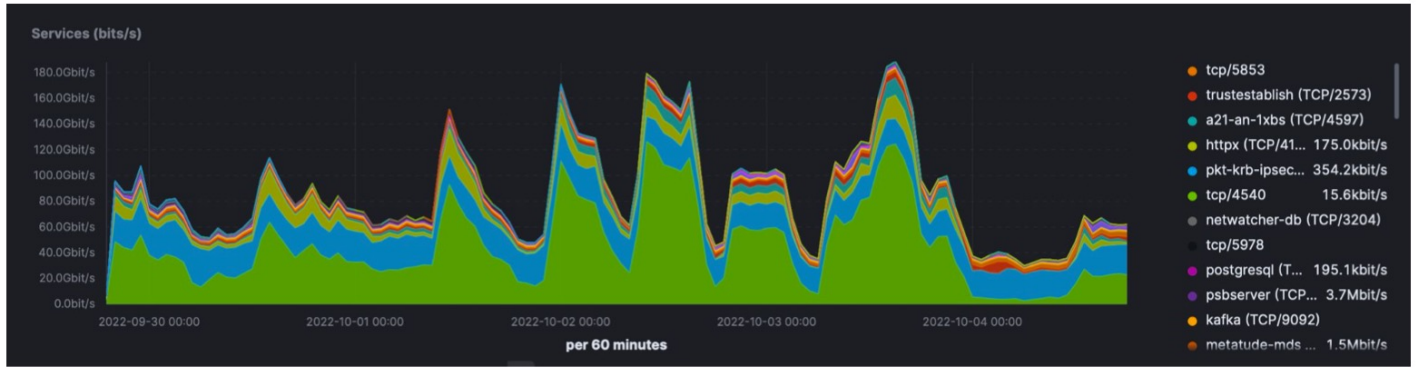
Pod Name	% GPU utilization (1 is 100%)	% CPU utilization (1 is 100%)	% Memory utilization (1 is 100%)	GPUs requested	CPUs requested	Memory request
ecewcsng/ece-wcsng-ant-array-6c44d6888f-87hwp		0.00	0.00	0	32.00	137 GB

Monitoring: Prometheus, Inmon TS, Perfsonar, Elasticsearch









Summary

- The PNRP builds upon the successful PRP and NRP communities
- Nautilus system substantially eased the deployment of the new PNRP resources
- Provides innovative hardware and software for exploring wide range of S&E research
- BYOR allows users to add their hardware to the system
- Provides opportunities for system administrators to learn about emerging systems software
- Users have access at various levels - from easy Jupyter notebook access to full deployments going from development to production (w/ Gitlab).

Acknowledgements

- US National Science Foundation (NSF) awards to UCSD
  - CNS-1456638, CNS-1730158, CNS-2100237, CNS-2120019
  - ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167
- DOD DURIP awards to UCSD
- UC Office of the President, Calit2 and Calit2's UCSD Qualcomm Institute
- San Diego Supercomputer Center and UCSD's Research IT and Instructional IT
- Partner Campuses: UCB, UCSC, UCI, UCR, UCLA, USC, UCD, UCSB, SDSU, Caltech, NU, Uwash,
- UChicago, UIC, UHM, CSUSB, UMo, FAMU, MSU, NYU, UNeb, UNM, UNC, FIU, UDel, UDak,
- SDakSU, Stanford, UArk, UOk, UoGuam, UKansas, CWRU, Clemson, MGHPCC, KISTI, UVA, AIST
- CENIC, Pacific Wave/PNWGP, StarLight/MREN, The Quilt, Great Plains Network, NYSERNet, Open
- Science Grid, Internet2, DOE ESnet, NCAR/UCAR & Wyoming Supercomputing Center,
- AWS, Google, Microsoft, Cisco, Juniper, Arista