# “Utilizing Nautilus and the National Research Platform for Big Data Research and Teaching”

**Panel Presentation**

**Larry Smarr and Grant Scott**

**MOREnet 2022 Annual Conference**

**October 19, 2022**

**Dr. Larry Smarr**

**Founding Director Emeritus, California Institute for Telecommunications and Information Technology;**

**Distinguished Professor Emeritus, Dept. of Computer Science and Engineering**

**Jacobs School of Engineering, UCSD**

**http://lsmarr.calit2.net**

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# Abstract

Thanks to a grant from the National Science Foundation, several organizations are investing in hardware that supports a distributed compute and storage cluster termed the National Research Platform which leverages shared resources between about 40 universities across the US. As a higher education member of the MOREnet Consortium, you can utilize these resources for data analysis or machine learning on large datasets. Learn from key enablers Larry Smarr (University of California, San Diego) and Grant Scott (University of Missouri – Columbia) about how Google's Kubernetes system can orchestrate the movement of your containerized application software across the distributed NRP cyberinfrastructure from your institution. Whether you are interested in hearing more about easily deploying Jupyter Notebooks for teaching STEM subjects with programming, developing workbench research codes, workflows, or data analysis – this session is for you. Your access to terabytes of data and significant compute resources through this collaboration can also be considered a match for grants.

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# 1948-1970: I Was Born and Raised in Central Missouri

My Family Earned
16 MU Degrees
Over 70 Years



Grandfather, Father, Me
At My Mizzou Graduation 1970

Nautilus

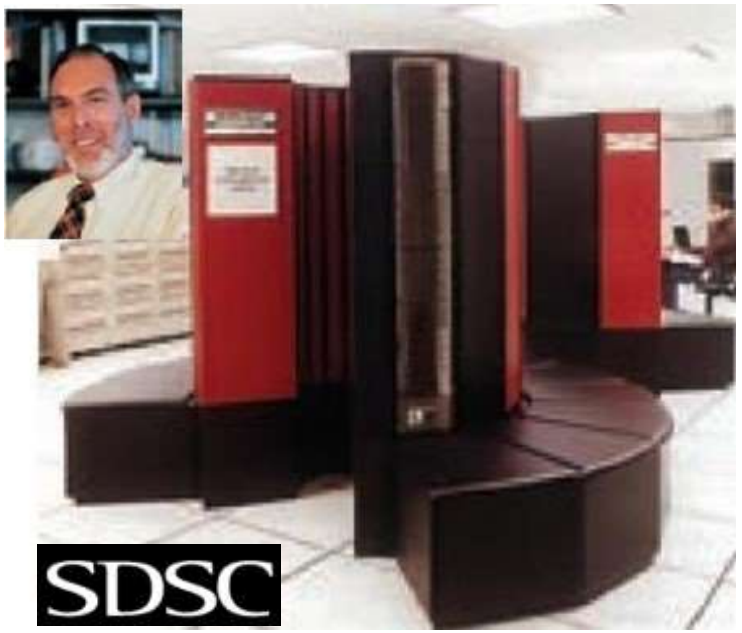
NRP NATIONAL RESEARCH PLATFORM

# 1985:
# NSF Adopted a DOE High-Performance Computing Model

**NSFNET 56 Kb/s Backbone (1986-8) Adopted TCP/IP**




NCSA Was Modeled on LLNL



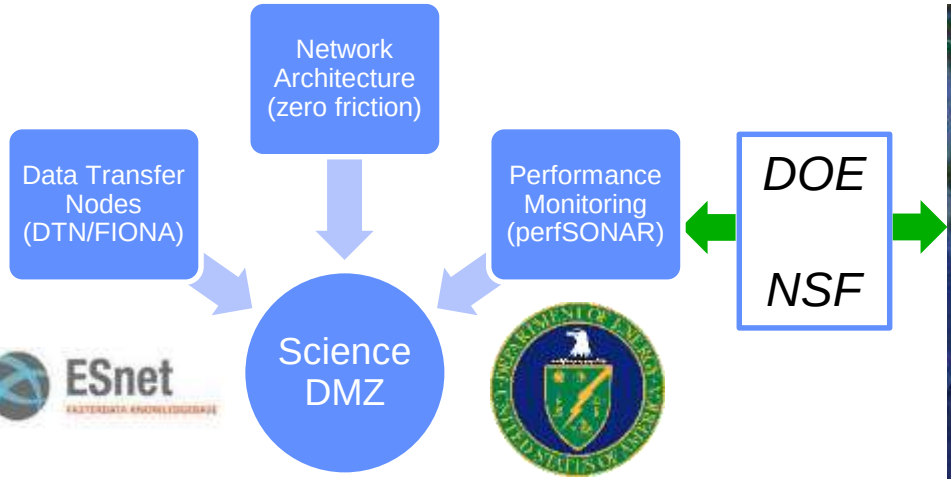

SDSC Was Modeled on MFEnet

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# 2010-2022:
# NSF Adopted a DOE High-Performance Networking Model

ScienceDMZ Coined in 2010 by ESnet




http://fasterdata.es.net/science-dmz/

Slide Adapted From Inder Monga, ESnet

NSF Campus Cyberinfrastructure Program
2012-2022



Has Made Over 340 Awards:
Across 50 States and Territories

Slide Adapted From Kevin Thompson, NSF

Internet Backbone 100Gbps
= 2 Million x NSFnet in 1986

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# 2015 Vision: The Pacific Research Platform Will Connect Science DMZs Creating a Regional End-to-End Science-Driven Community Cyberinfrastructure




**NSF CC*DNI Grant**
**$7.3M 10/2015-10/2022**




Source: John Hess, CENIC

# 2015-2022: UCSD Designs PRP Data Transfer Nodes (DTNs) -- Flash I/O Network Appliances (FIONAs)

**FIONAs Solved the Disk-to-Disk Data Transfer Problem at Near Full Speed on Best-Effort 10G, 40G and 100G**

**FIONA**

Software Specifications

Installation Instructions

Hardware Specifications

2021 – 2020 – 2019 – 2018 – 2017 – 2016 – 2015

**https://pacificresearchplatform.org/fiona/**



Add Up to 8 Nvidia GPUs Per 2U FIONA
To Add Machine Learning Capability
Up to 240TB Storage

SDSC

Qualcomm Institute

Nautilus

NSF

FIONAs Designed by UCSD's Phil Papadopoulos, John Graham, Joe Keefe, and Tom DeFanti

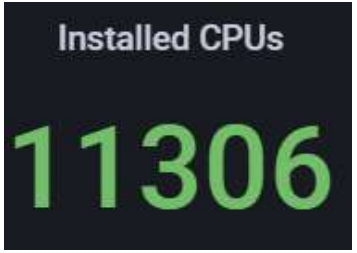
NRP NATIONAL RESEARCH PLATFORM

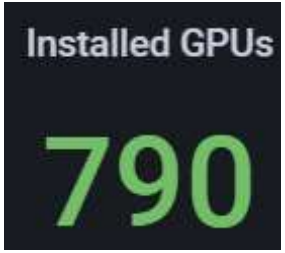
# PRP's Nautilus is a Multi-Institution Hypercluster Connected by Optical Networks

As of October 15, 2022

160 GPU & Storage FIONAs on 27 Partner Campuses Networked Together at 10-100Gbps















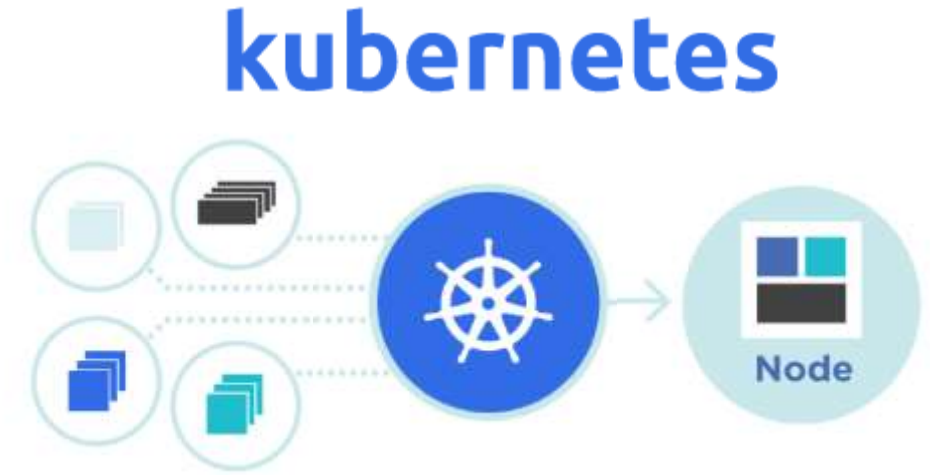






# 2018/2019: PRP Game Changer!
# Using Google's Kubernetes to Orchestrate Containers Across the PRP

User Applications




Containers

PRP PACIFIC RESEARCH PLATFORM

Clouds

Nautilus

CLOUD NATIVE COMPUTING FOUNDATION

NRP NATIONAL RESEARCH PLATFORM

# PRP's Nautilus Hypercluster Adopted Open-Source Kubernetes and Rook to Orchestrate Software Containers and Manage Distributed Storage







"Kubernetes with Rook/Ceph Allows Us to Manage Petabytes of Distributed Storage and GPUs for Data Science, While We Measure and Monitor Network Use."
--John Graham, UC San Diego

# The PRP Web Site Provides Widely-Used Open-Source Services Supporting Joining, Application Research, Development, and Collaboration




**Nautilus documentation**

Get Access

Docker and containers

Basic kubernetes

**[matrix] Chat for Nautilus Users**

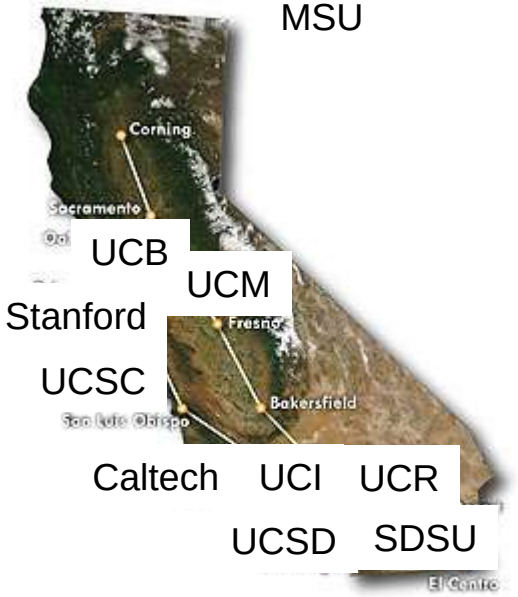
**JupyterLab for Nautilus**

GitLab Continuous Integration & Delivery

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# 2017-2020: NSF CHASE-CI Grant Adds a Machine Learning Layer Built on Top of the Pacific Research Platform




NSF Grant for High Speed "Cloud" of 256 GPUs
For 30 ML Faculty & Their Students at 10 Campuses
for Training AI Algorithms on Big Data

**CI-New: Cognitive Hardware and Software Ecosystem Community Infrastructure (CHASE-CI)**

**For the Period September 1, 2017 – August 31, 2020**

**SUBMITTED – January 18, 2017**

PI: Larry Smarr, Calit2, UCSD
Co-PIs:

- Tajana Rosing, CSE, UCSD
- Ken Kreutz-Delgado, ECE, UCSD
- Ilkay Altintas, SDSC, UCSD
- Tom DeFanti, QI, UCSD

NSF Has Funded Two Extensions:
CHASE-CI ABR-Smarr PI &
CHASE-CI ENS-DeFanti PI
$2.8M

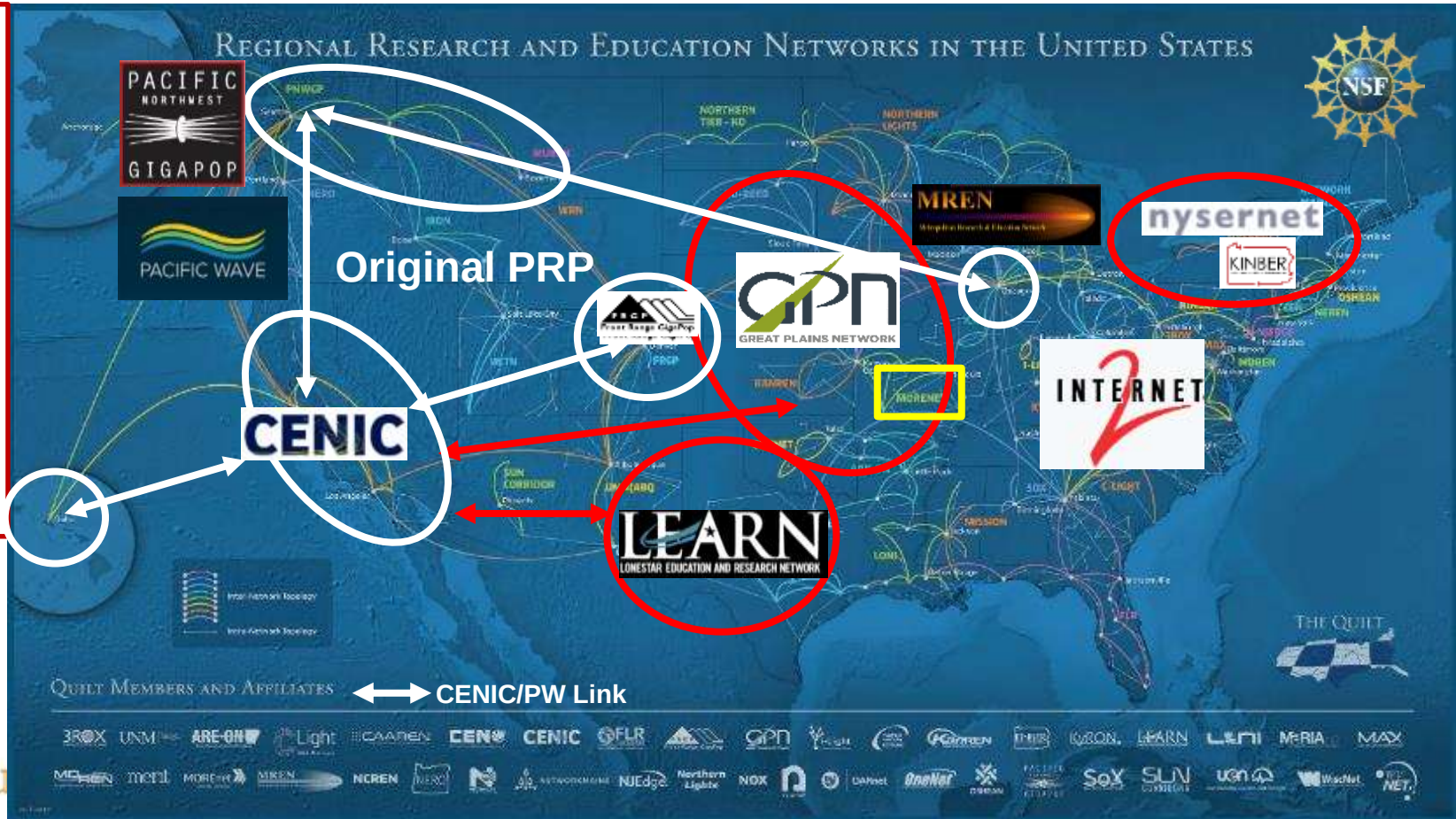
Nautilus

NRP NATIONAL RESEARCH PLATFORM

# 2018-2022: Toward the National Research Platform (TNRP) - Using CENIC & Internet2 to Connect Quilt Regional R&E Networks

**"Towards The NRP"**
**3-Year Grant Funded by NSF**
**$2.5M**
**October 2018**
**Award #1826967**

**PI Smarr**
**Co-PIs Altintas**
**Papadopoulos**
**Wuerthwein**
**Rosing**
**DeFanti**



# The Pacific Research Platform Video Highlights 3 Different Applications Out of 700 Nautilus Namespace Projects

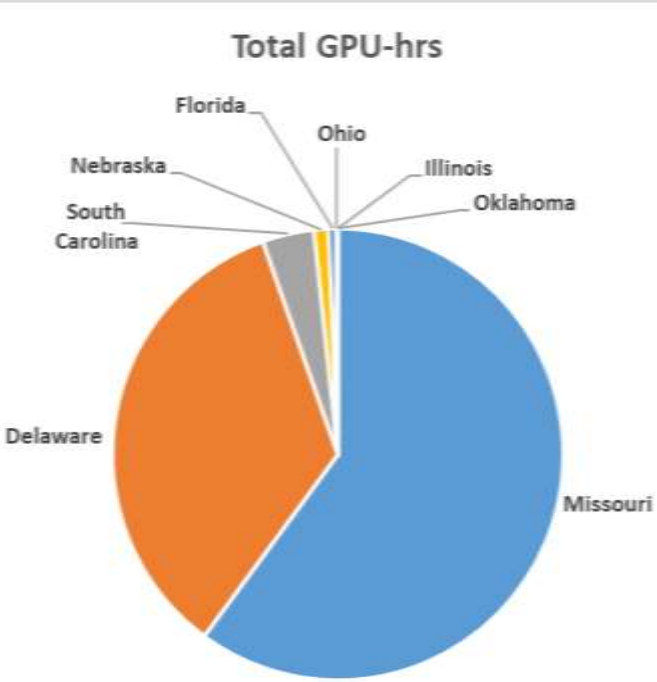



Pacific Research Platform Video:
https://nationalresearchplatform.org/media/pacific-research-platform-video/

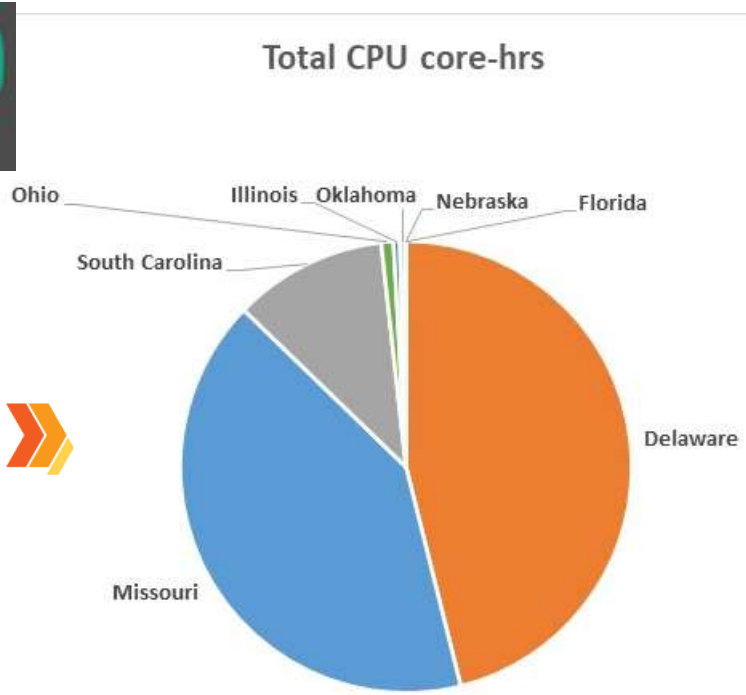
# The PRP Has Emphasized Expanding Diversity and Inclusion

- **When the PRP Grant Was Funded in 2015, It Started With:**
  - **6 States ➡ Now 40 States**
  - **19 Campuses ➡ Now 95 Campuses**
  - **9 Minority Serving Institutions ➡ Now 20 MSIs**
  - **2 NSF EPCoR States ➡ Now 19 EPSCoR States, 2 Territories, and Wash DC**

# Non-California Nautilus PI Namespace 2021 Usage by State: "Big MO!"




17,217 GPU-hrs

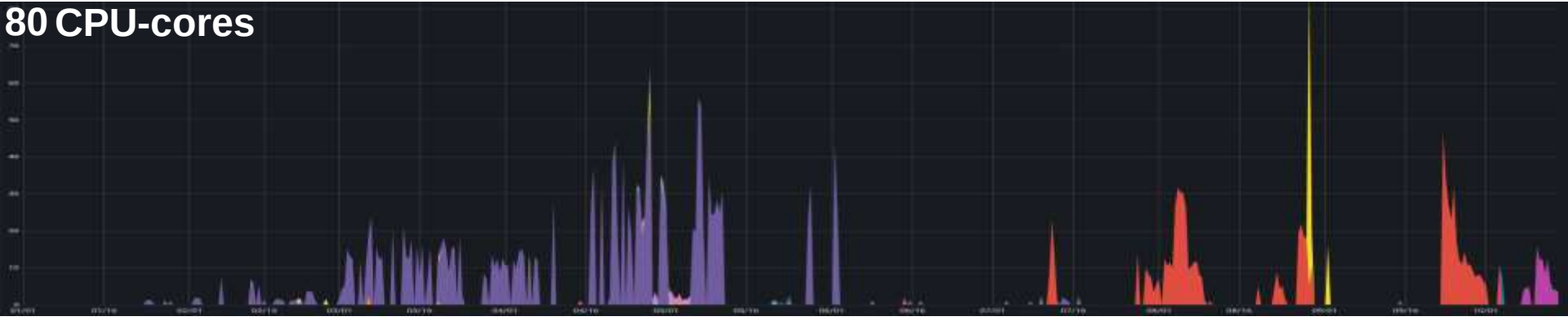



28,088 CPU core-hrs

Grant Scott, UMC
Helped Organize the UMC PRP Usage

# Missouri Campus PRP Namespace Usage
# Calendar 2022



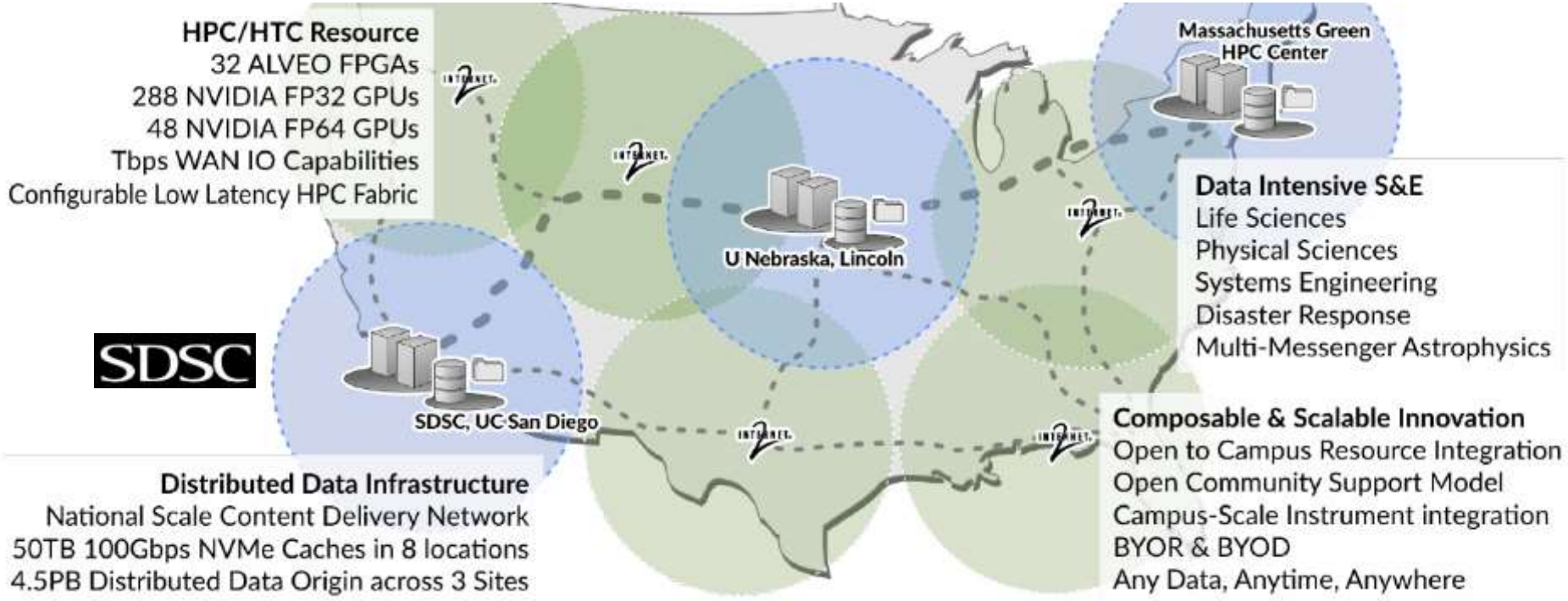




Nautilus

Campuses UMC & WUSTL

NRP NATIONAL RESEARCH PLATFORM

# 2022-2026 NRP Future: PRP Federates with NSF-Funded *Prototype National Research Platform*

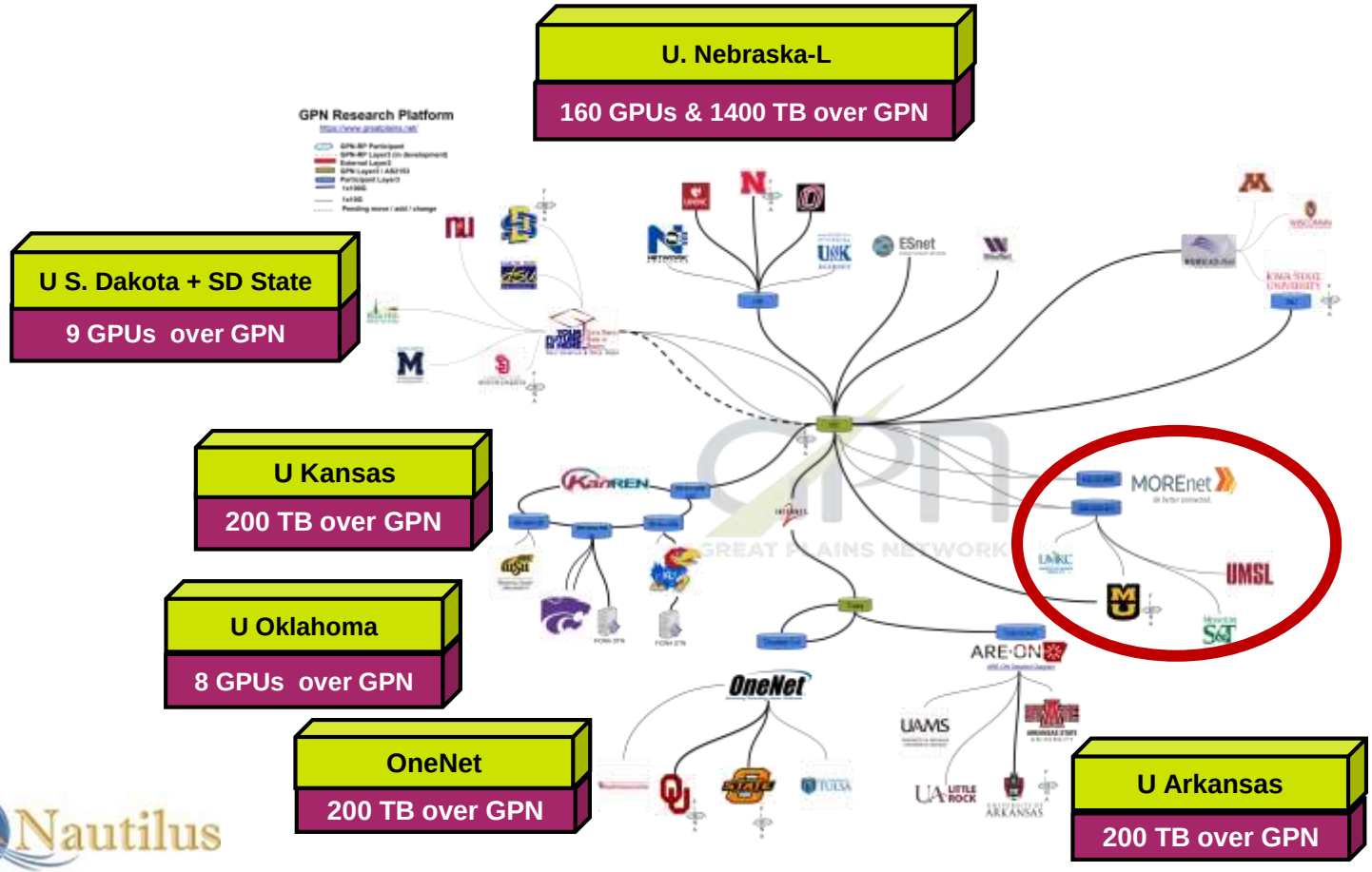


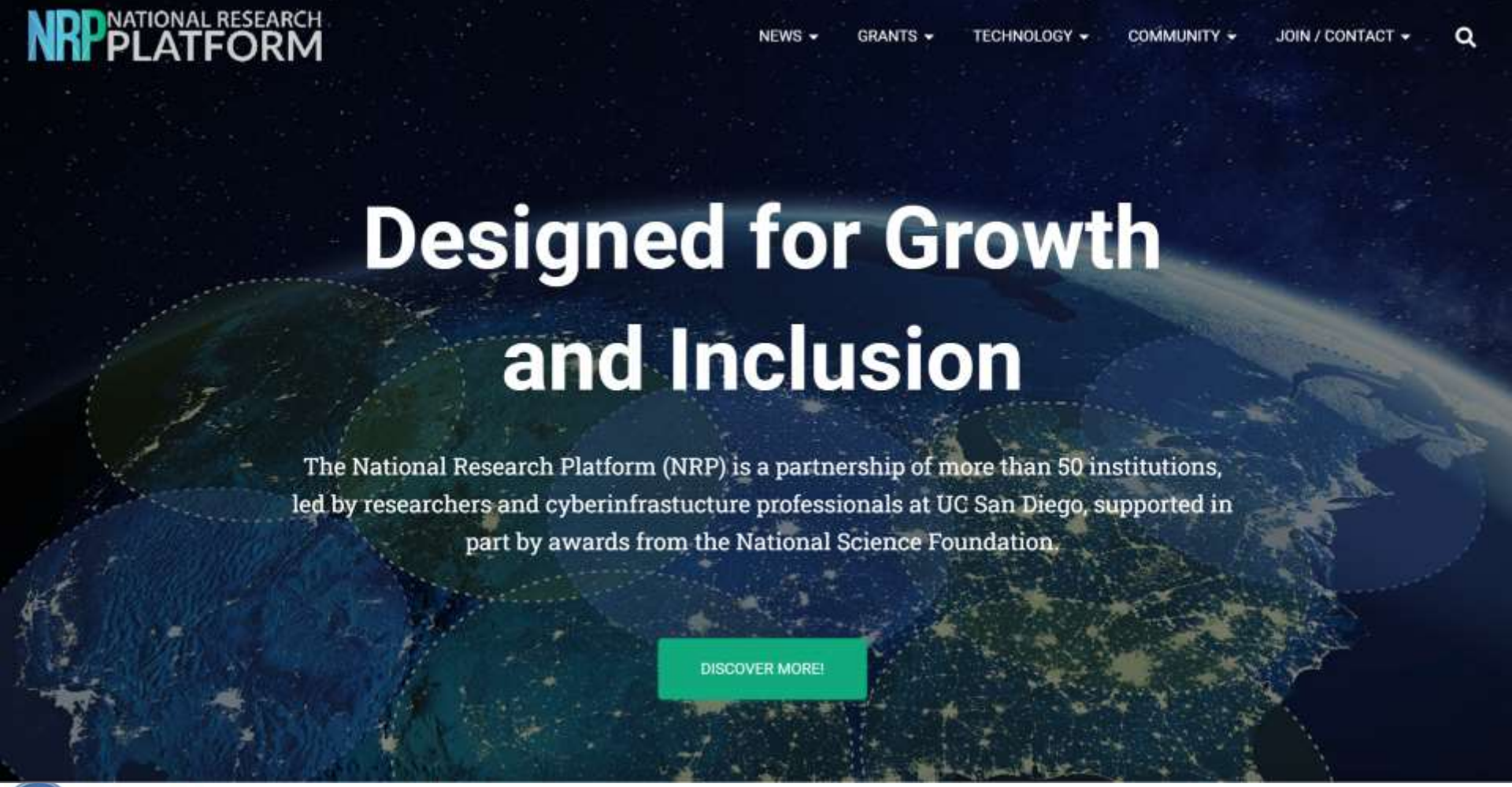

**NSF Award OAC #2112167 (June 2021) [$5M Over 5 Years]**
**PI Frank Wuerthwein (UCSD, SDSC)**
**Co-PIs Tajana Rosing (UCSD), Thomas DeFanti (UCSD), Mahidhar Tatineni (SDSC), Derek Weitzel (UNL)**

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# NRP Brings More Regional Computational and Storage Assets to MOREnet via GPN in 2022




NRP NATIONAL RESEARCH PLATFORM
NEWS
GRANTS
TECHNOLOGY
COMMUNITY
JOIN / CONTACT
Designed for Growth and Inclusion
The National Research Platform (NRP) is a partnership of more than 50 institutions, led by researchers and cyberinfrastucture professionals at UC San Diego, supported in part by awards from the National Science Foundation.
DISCOVER MORE!

# Using Nautilus For Teaching @ MU

**Grant Scott**

https://scottgs.mufaculty.umsystem.edu/

**Assistant Professor**

- Computer Science, Computer Engineering
- College of Engineering

**Director,**

- Data Science and Analytics, MS Program
- Institute for Data Science and Informatics

**Provided Data Science and Computer Science Learning Outreach using MU & NRP Jupyter**

- US Government Intelligence Agency
- USDA ARS Long-Term Agroecosystem Research (LTAR) Data Managers
- Tutorials at Regional and International conferences

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# Nautilus Supports Jupyter Hub & Jupyter Lab

- Rich environments for STEM education
- Programming
- Scientific Computing
- Centralized Administration
- Powerful Computing Resources

- Use institutional *CILogon*

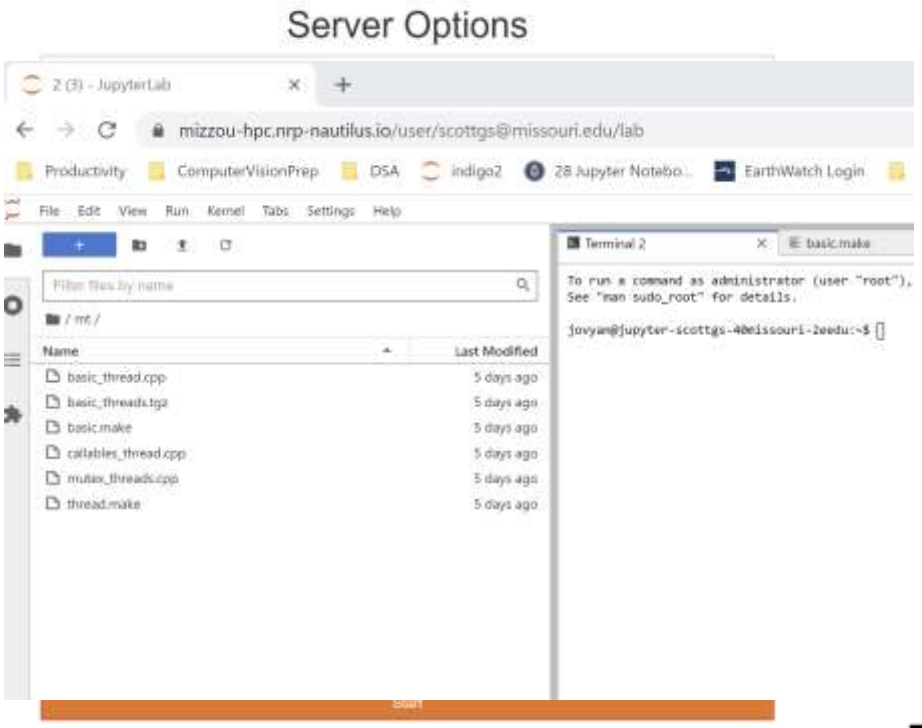
## Nautilus for STEM Teaching
## mizzou*.nrp-nautilus.io




Nautilus

NRP NATIONAL RESEARCH PLATFORM

# Nautilus Supports Jupyter Hub & Jupyter Lab




## For Computer Science mizzou-hpc.nrp-nautilus.io

- Offers rich analytics focused software stacks
- Offers specialized Jupyter Lab with scientific programming
- Students clone courseware and submit work with Gitlab

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# Nautilus Supports Jupyter Hub & Jupyter Lab

DEPARTMENT OF HIGHER EDUCATION & WORKFORCE DEVELOPMENT

## Basic Programming Course




**GEER Excels**

- **Basic Python Programming**
- **Used Nautilus Jupyter Lab**

- **Trainees authenticated with Github accounts**

Nautilus

NRP NATIONAL RESEARCH PLATFORM

# Using Nautilus For Research @ MU

**Grant Scott with mentees: Alex Hurt, Anes Ouadou**

### Case Study in Deep Learning for Computer Vision

- Journal Publication Results in Weeks
- Using State-of-the-Art Deep Learning Computer Vision Models for Satellite Imagery Data Sets

### Tutorials

- Great Plains Network Annual Meeting 2022
  - Getting Started on Nautilus and Kubernetes
- Great Plains Network Annual Meeting 2023
  - Getting Started on Nautilus and Kubernetes
  - Scaling Deep Learning with Nautilus

Nautilus

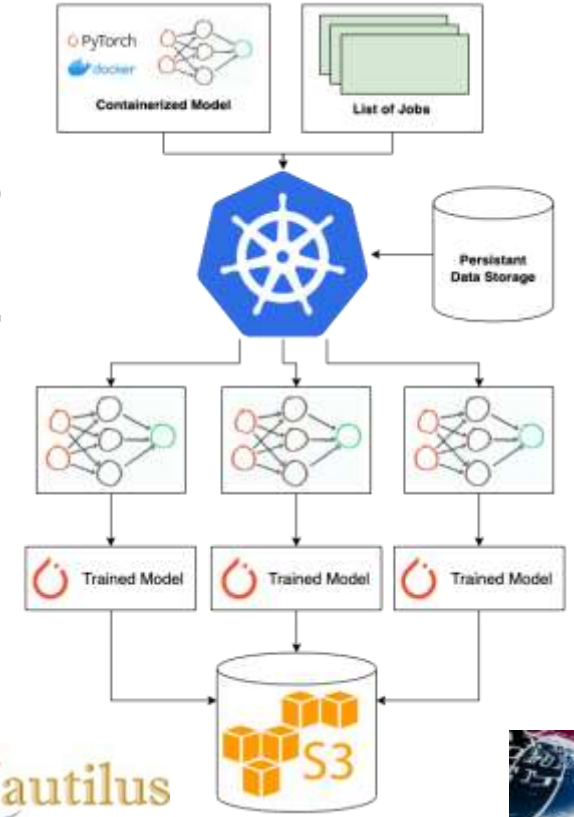



NRP NATIONAL RESEARCH PLATFORM

# Scaling Deep Learning with Kubernetes on Nautilus

Dr. Alex Hurt

Nautilus for Accelerated Research Computing




- Using containerized model definition and list of jobs
- Mounted persistent data storage to each pod
- Each GPU job produces an associated trained model
- Automation currently performed via environment variables and bash, but more sophisticated methods in development
- Models are sync'd to Nautilus S3 bucket for later use in evaluation or other ML applications

Nautilus

High-Performance Data-Intensive Computing Systems Laboratory

NRP NATIONAL RESEARCH PLATFORM

# Deep Learning on Nautilus: By the Numbers

Dr. Alex Hurt

## Compute Intensive

- Containerized deep neural architectures: 9
- Datasets trained on: 3
- PyTorch Models Trained: 27
- Training Epochs Completed: 8,100
- Iterations of Training Completed: 30,088,125
- Number of Images Processed: 240,705,000
- Trainable Parameters Optimized: 1,730,368,875

## Data Intensive

- Data loading: 415.8GB
- Neural Model Loading: 124,740 GB

- Wall-Clock: ~77 days
- Human Effort: <3 hours

# CC* Team: Great Plains Regional CyberTeam




- PI NSF Award OAC #1925681

- Helping the Great Plains region better leverage collective cyberinfrastructure resources
- GPN Contributions to Nautilus and the NRP