Computing without boundaries on Expanse

8th Annual MVAPICH User Group (MUG) Meeting
August 26, 2020

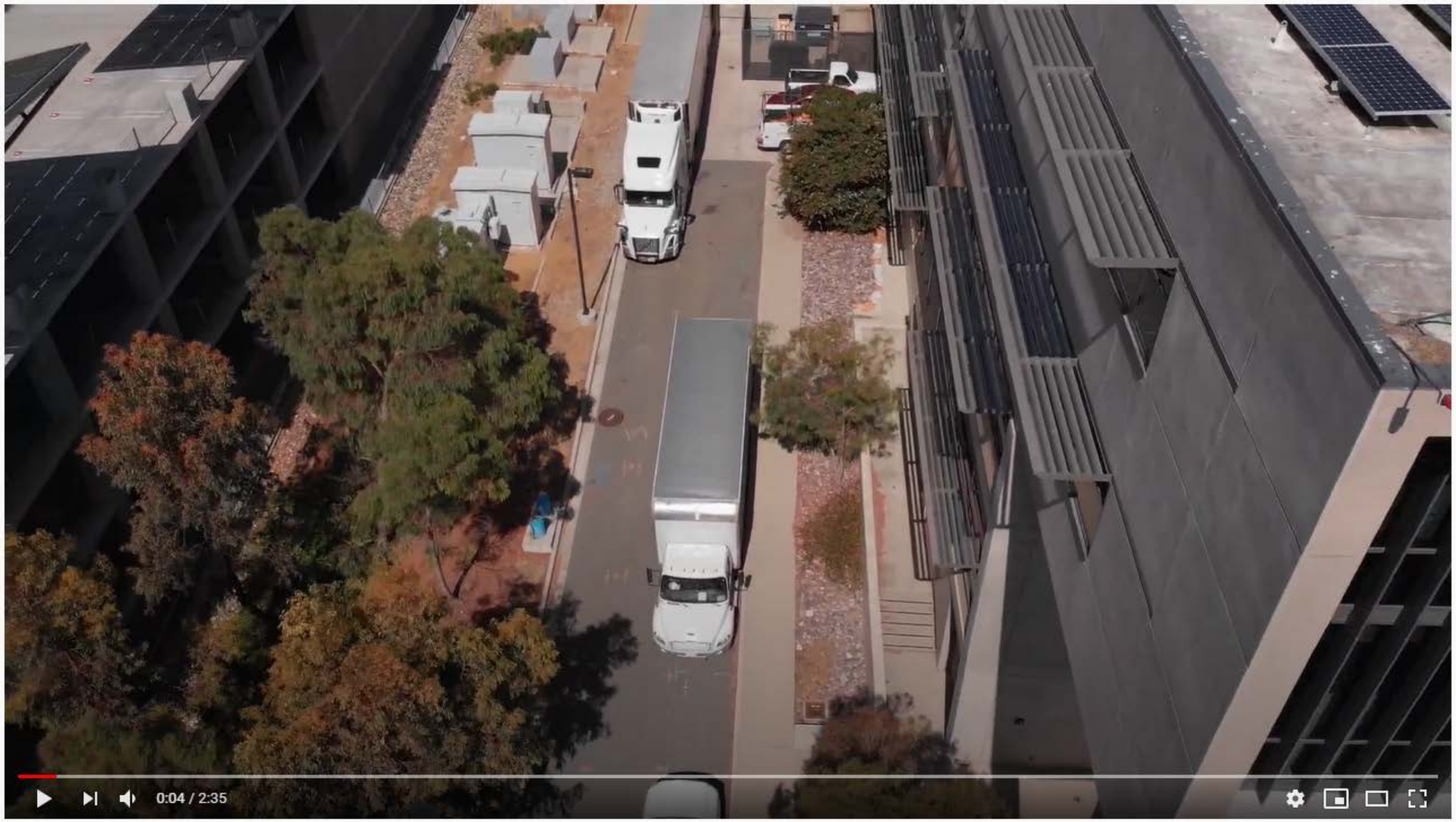
Michael L Norman
Director, SDSC

EXPANSE
COMPUTING WITHOUT BOUNDARIES

San Diego Supercomputer Center

NSF Award 1928224

Time-lapse construction video



Expanse installation is underway now!

…Dell handed the cluster over to us a week ago for SW install

Photos courtesy of Jeff Filliez. Taken July 30, 2020

Computing Without Boundaries: Cyberinfrastructure for the Long Tail of Science

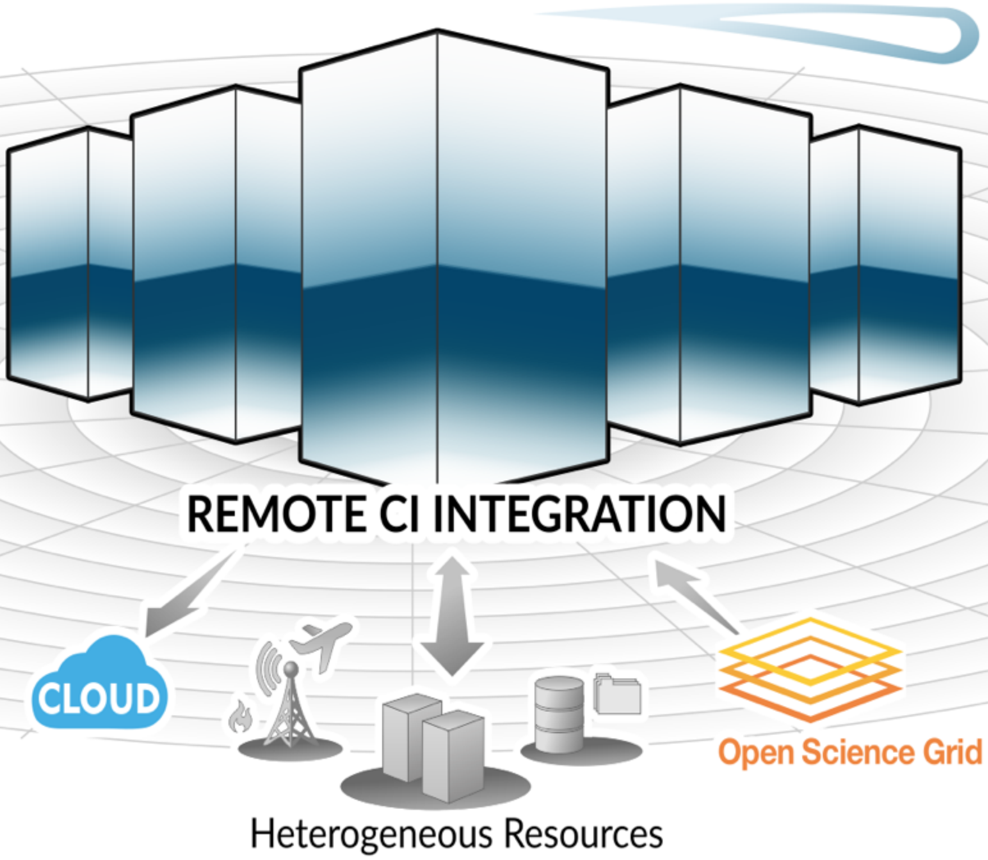
- NSF Solicitation 19-534: Advanced Computing Systems & Services: Adapting to the Rapid Evolution of Science and Engineering Research
- Category 1: Capacity System, NSF Award # 1928224
- NSF Program Officer: Robert Chadduck
- PIs: Mike Norman (PI), Ilkay Altintas, Amit Majumdar, Mahidhar Tatineni, Shawn Strande
- $10M Acquisition; Operations and Maintenance funding est. $2.5M/year
- Primary Vendors: Dell (HPC system); Aeon Computing (storage)
- Compute, interconnect, NVMe: AMD, Intel, NVIDIA, Mellanox
- Liquid cooling: CoolIT

EXPANSE

COMPUTING WITHOUT BOUNDARIES
5 PETAFLOP/S HPC and DATA RESOURCE

HPC RESOURCE
13 Scalable Compute Units
728 Standard Compute Nodes
52 GPU Nodes: 208 GPUs
4 Large Memory Nodes

DATA CENTRIC ARCHITECTURE
12PB Perf. Storage: 140GB/s, 200k IOPS
Fast I/O Node-Local NVMe Storage
7PB Ceph Object Storage
High-Performance R&E Networking

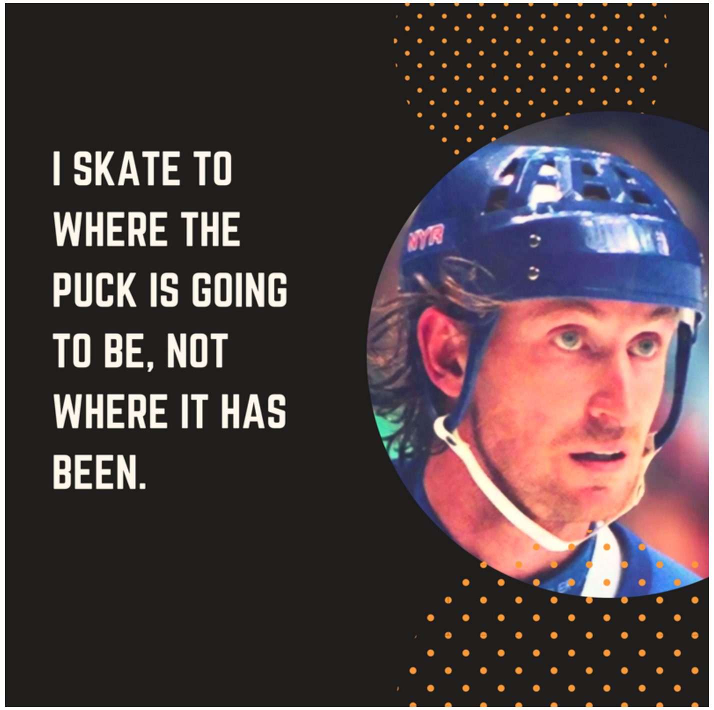


LONG-TAIL SCIENCE
Multi-Messenger Astronomy
Genomics
Earth Science
Social Science

INNOVATIVE OPERATIONS
Composable Systems
High-Throughput Computing
Science Gateways
Interactive Computing
Containerized Computing
Cloud Bursting

Expanse is the latest incarnation of SDSC's evolving HPC strategy

- How did we get here?
- Brief history of SDSC HPC systems 2009-present
  - Adapting to user needs
  - Innovating on HW, SW and OPS
  - Taking some risks
  - "Skating to where the puck is going to be"
    Wayne Gretsky
- *This is my contribution as SDSC Director and PI of 3 NSF funded HPC systems*



“Flash” Gordon: data-intensive HPC system

NSF production: 2011-2017

Innovations

- 300TB IB-connected flash SSD storage system
- Node local flash SSD
- vSMP supernodes (256 core)
- Dual rail 3D torus
  - Rail 1: MPI traffic
  - Rail 2: Lustre IO
- Early OSG integration pilot for massive LHC data analysis

Gordon: lessons learned

- Computational chemists love flash SSD for “scratch IO”
  - Flocked to SDSC
- First in TeraGrid/XSEDE storage allocations a big hit
- Surprisingly low demand for vSMP supernodes
  - Virtual shared memory across nodes
  - 16-way supernode aggregates 256 cores and 2 TB RAM
  - Ease of use for threaded big data analysis SW
- → Most NSF researchers did not have Big Data

Trestles: long-tail capacity system pilot

NSF production: 2011-2013

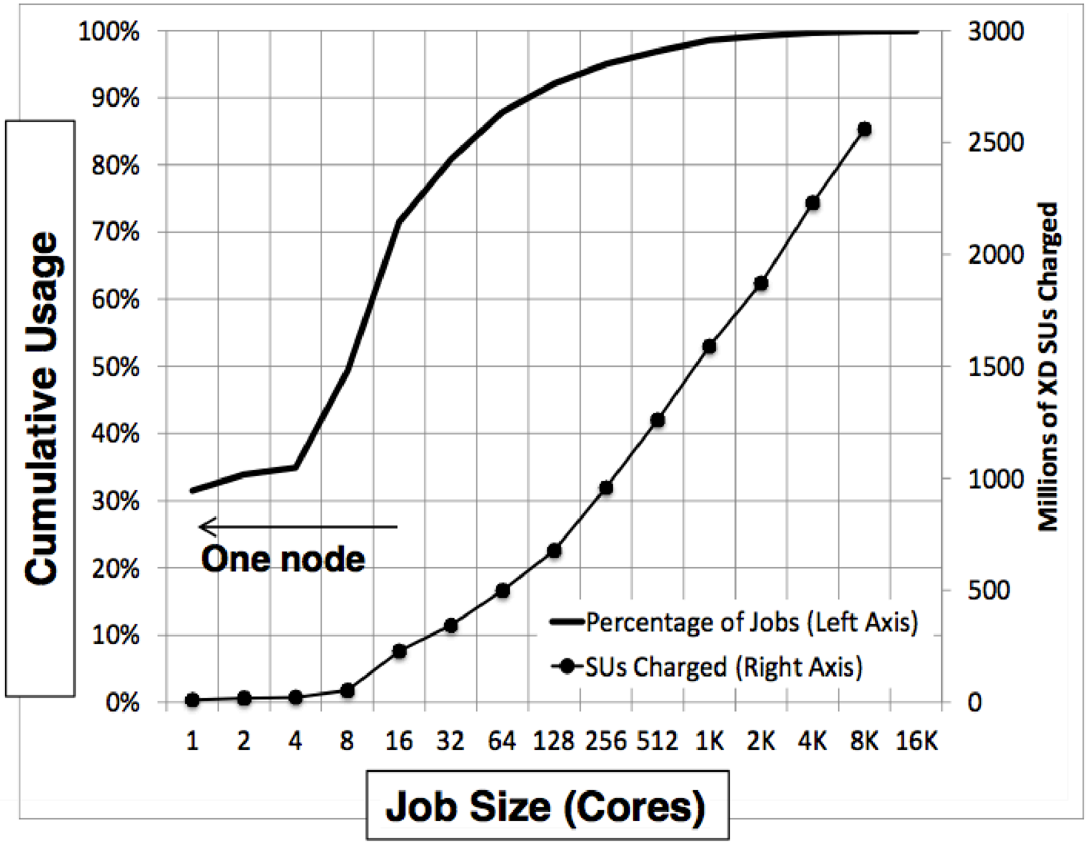
Innovations

- Architected to support modest scale HPC and Gateway users
- Node-local SSD like Gordon
- Under-allocate to ensure good turnaround
- Limit maximum job size (cores)
- Scheduling polices favored long-tail jobs

Trestles: lessons learned

- Long-tail users are really out there!
  - Demand for capacity systems
- Users loved the throughput
- Gateway access becoming more important access mechanism for non-traditional HPC users
- Potential for many more users and jobs to be supported

2011 TG workload analysis



Comet: virtualized, hybrid, long-tail capacity system

NSF production: 2015-2021

Innovations

- Hybrid CPU/GPU cluster
- Add’l HW to host Gateways
- Virtual Cluster capability at native IB speeds (SRIOV)
    - Total OSG integration
- Trestles operating policies
- Shared node scheduling
- Pioneered 24-hr trial account
- First XSEDE deployment of Singularity

Comet: lessons learned

- Unforeseen higher demand for GPUs
  - Doubled capacity in 2018
- Users love Comet, esp. chem/mat'l/bio
- Gateways are dominant usage mode
- Initial long-tail goal:
  - 10,000 unique users in 5 years
  - Achieved that in first year
- New long-tail goal:
  - 50,000 unique users in 5 years
  - >80,000 currently

SC '18 HPCWire Awards presentation

MVAPICH @ SDSC

- SDSC (Majumdar, Tatineni) has long collaborated with the MVAPICH team led by DK Panda on NSF-funded projects
- Gordon, Trestles, and Comet all deployed MVAPICH2
- Expanse will deploy MVAPICH2, MVAPICH2-X, and MVAPICH2-GDR

EXPANSE

COMPUTING WITHOUT BOUNDARIES

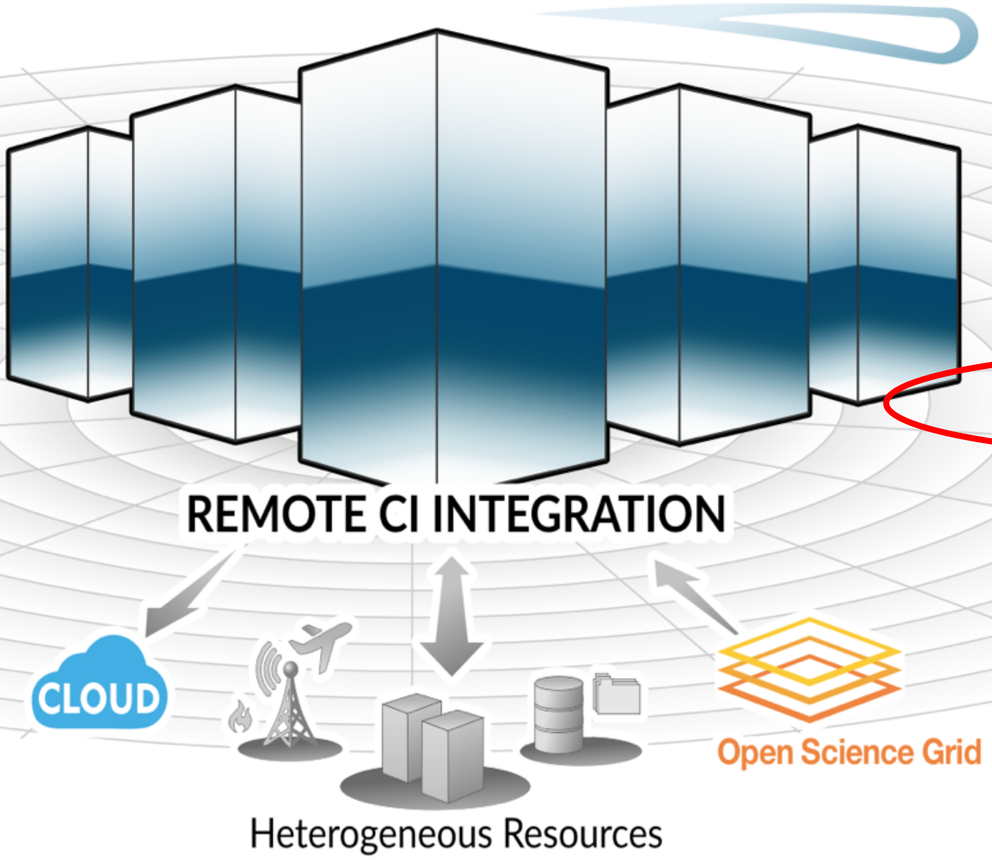
5 PETAFLOP/S HPC and DATA RESOURCE

HPC RESOURCE

13 Scalable Compute Units
728 Standard Compute Nodes
52 GPU Nodes: 208 GPUs
4 Large Memory Nodes

DATA CENTRIC ARCHITECTURE

12PB Perf. Storage: 140GB/s, 200k IOPS
Fast I/O Node-Local NVMe Storage
7PB Ceph Object Storage
High-Performance R&E Networking



LONG-TAIL SCIENCE

Multi-Messenger Astronomy
Genomics
Earth Science
Social Science

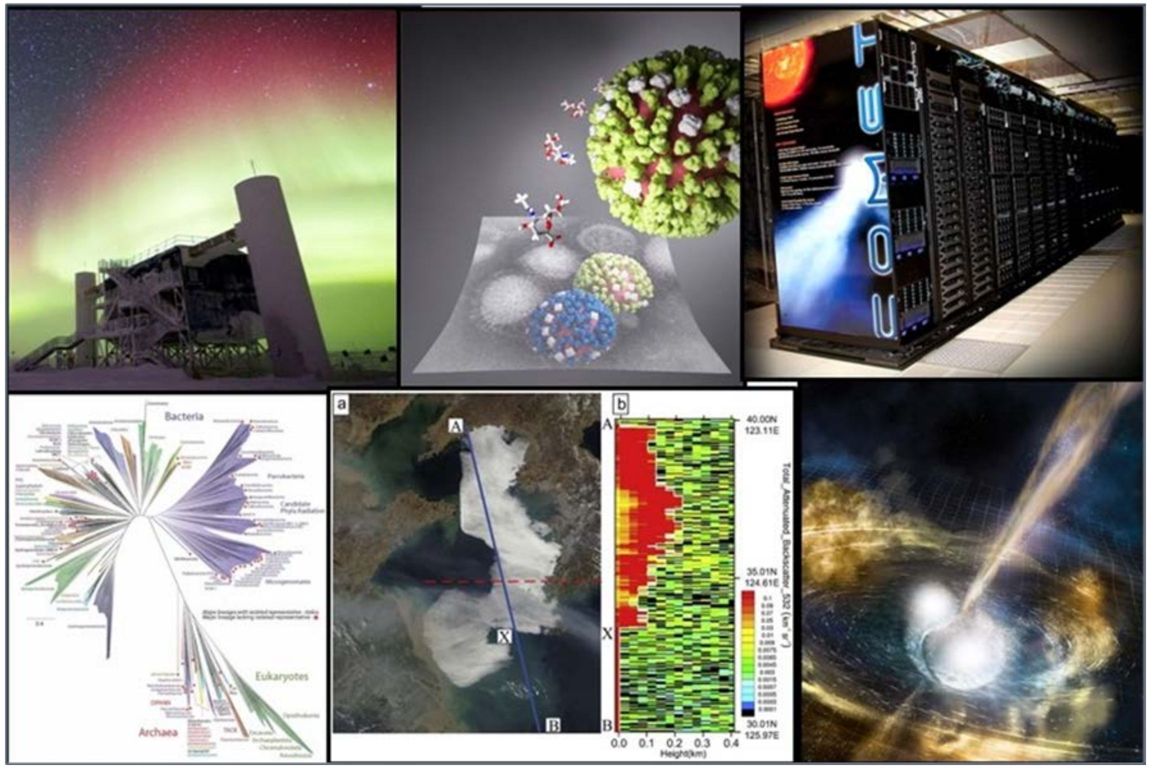
INNOVATIVE OPERATIONS

Composable Systems
High-Throughput Computing
Science Gateways
Interactive Computing
Containerized Computing
Cloud Bursting

Overview

- 728, 2-socket AMD-based compute nodes (2.25 GHz EPYC; 64-core/socket)
- 93,184 compute cores (that's 2x the cores in Comet in about ½ as many racks!)
- 52 4-way GPU nodes based on V100 w/NVLINK
- Based on benchmarks we've run, we expect > 2x throughput over Comet; and a 1-1.8 per-core improvement over Comet's Haswell processors.
- **Expect a smooth transition from Intel to AMD**
- **SDSC team has compiled and run many of the common software packages on AMD Rome based test clusters**
- Available in the Sep 15 – Oct 15 XSEDE Allocations Review for Jan 1 2021 start.
- **October 1, 2020: Operations for 5-years; 5-year follow-on system anticipated**

Like *Comet*, which concludes operations in March 2021, Expanse will advance science and engineering discovery

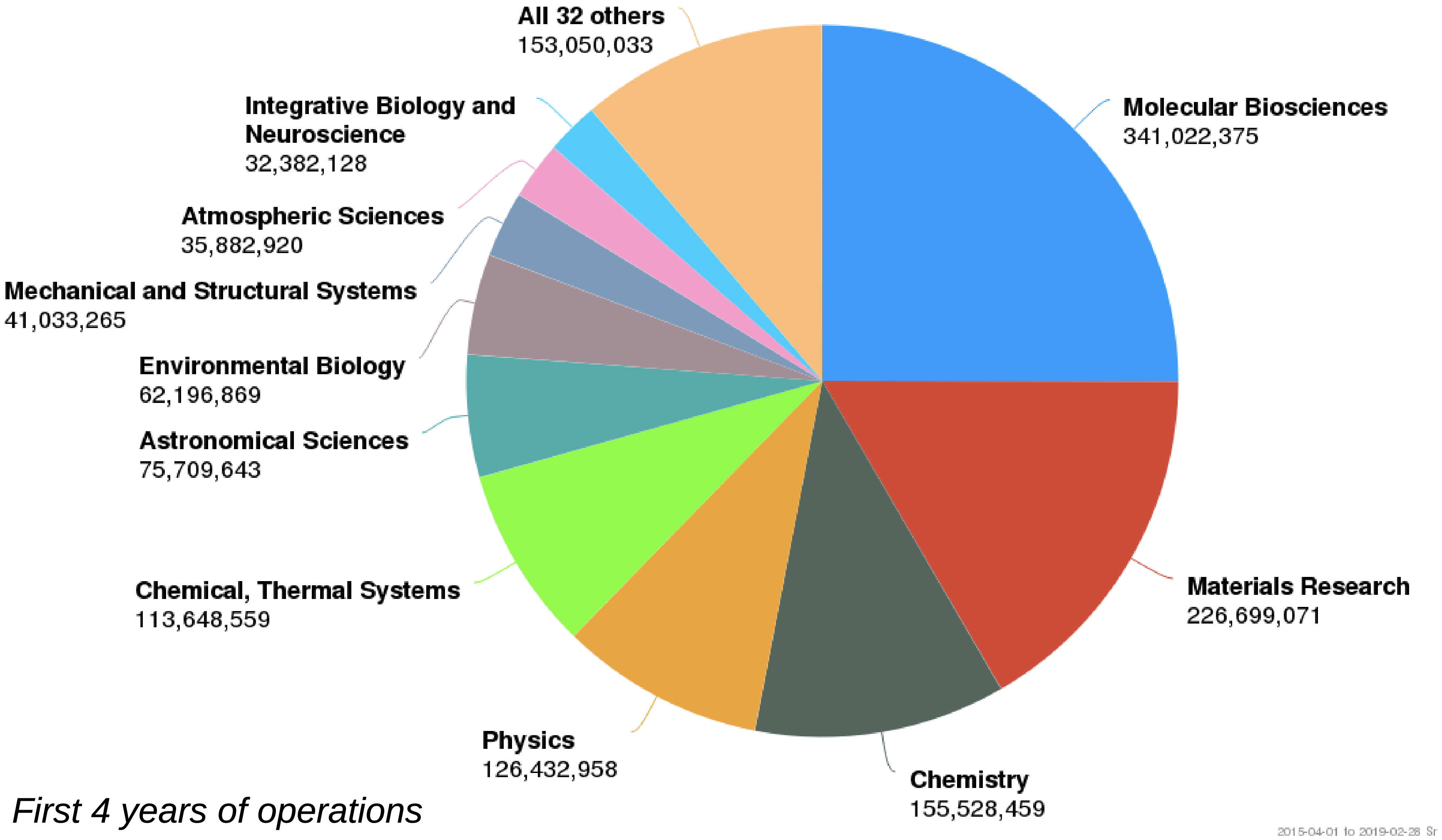


Clockwise from upper left: IceCube Neutrino Detection; Battling Influenza; Comet Surpasses 40,000 Users; Detecting Gravitational Waves; Predicting Sea Fog; Defining a New Tree of Life

In just over 4 years of Comet:

- 40,000+ Unique Users
- 1,200+ Publications
- ~2,000 Research, education and startup allocations
- 400+ Institutions
- Scientific discoveries and breakthroughs
- Overlap of 6 months for Comet and Expanse operations will provide ample transition time for users.

Comet historical usage is a good indicator of the science we expect to see on Expanse

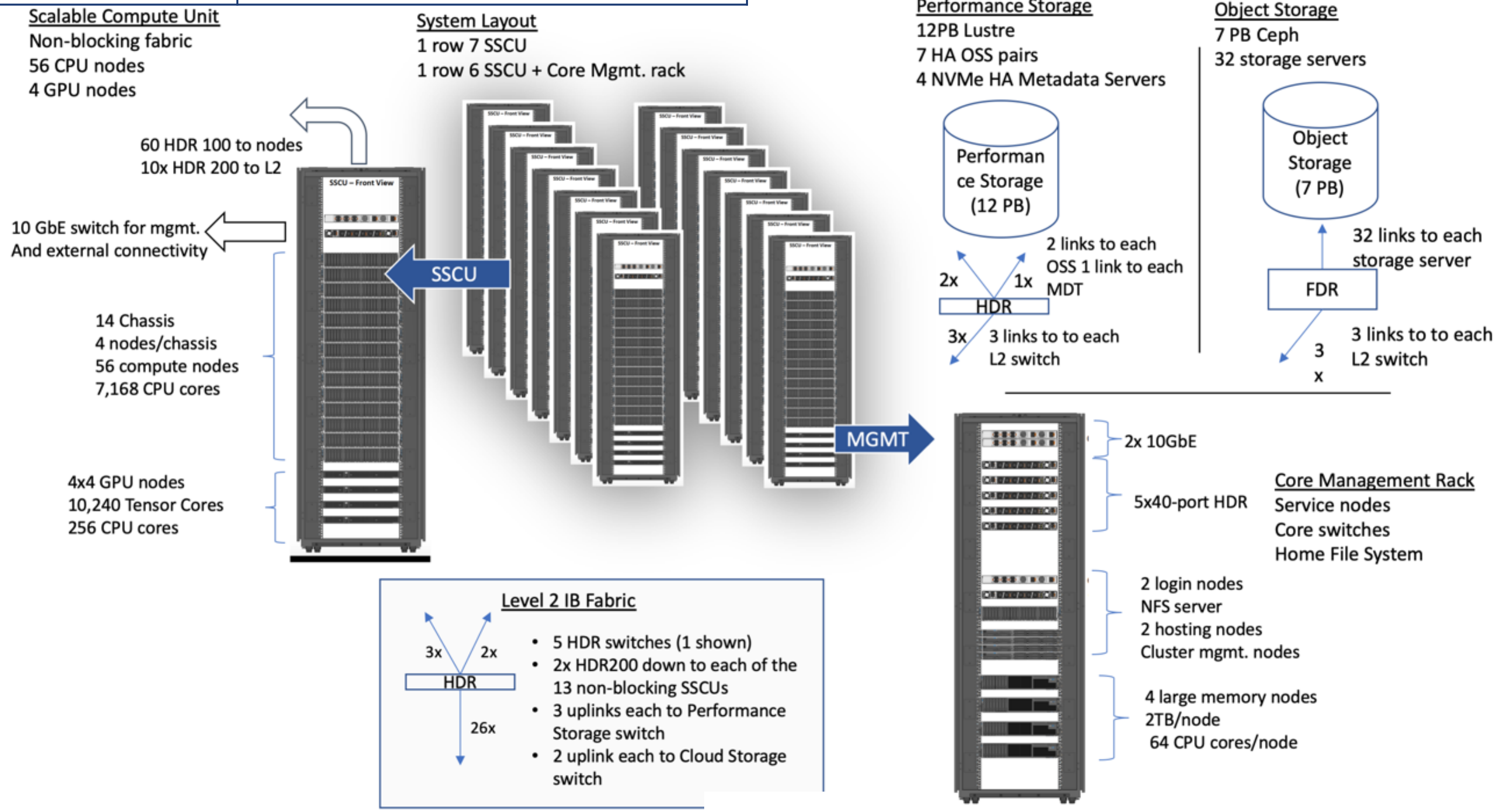


First 4 years of operations

Expanse System Summary

System Component	Configuration
AMD EPYC (Rome) 7742 Compute Nodes	
Node count	728
Clock speed	2.25 GHz
Cores/node	128
Total # cores	93,184
DRAM/node	256 GB
NVMe/node	1 TB
NVIDIA V100 GPU Nodes	
Node count	52
Total # GPUs	208
GPUs/node	4
GPU Type	V100 SMX2
Memory/GPU	32 GB
CPU cores; DRAM; clock (per node)	40; 384 GB; 2.5 GHz;
CPU	6248 Xeon
NVMe/node	1.6TB
Large Memory Nodes	
Number of nodes	4
Memory per node	2 TB
CPUs	2x AMD 7742/node;

Storage	
Lustre file system	12 PB (split between scratch & allocable projects)
Ceph file system	7 PB (coming April 2021)
Home File system	1 PB



The SSCU is Designed for the Long Tail Job Mix, Maximum Performance, Efficient Systems Support, and Efficient Power and Cooling

Standard Compute Nodes

- *2x AMD EPYC 7742 @2.25 GHz*
- *128 Zen2 CPU cores*
- *PCIe Gen4*
- *256 GB DDR4*
- *1 TB NVME*

GPU Nodes

- *4x NVIDIA V100 w/NVLINK*
- *10,240 Tensor Cores*
- *32 GB GDDR*
- *1.6 TB NVMe*
- *Intel CPUs*

SSCU Components

- *56x CPU nodes*
- *7,168 Compute Cores*
- *4x GPU nodes*
- *1x HDR Switch*
- *1x 10GbE Switch*
- *HDR 100 non-blocking fabric*
- *Wide rack for serviceability*
- *Direct Liquid Cooling to CPU nodes*

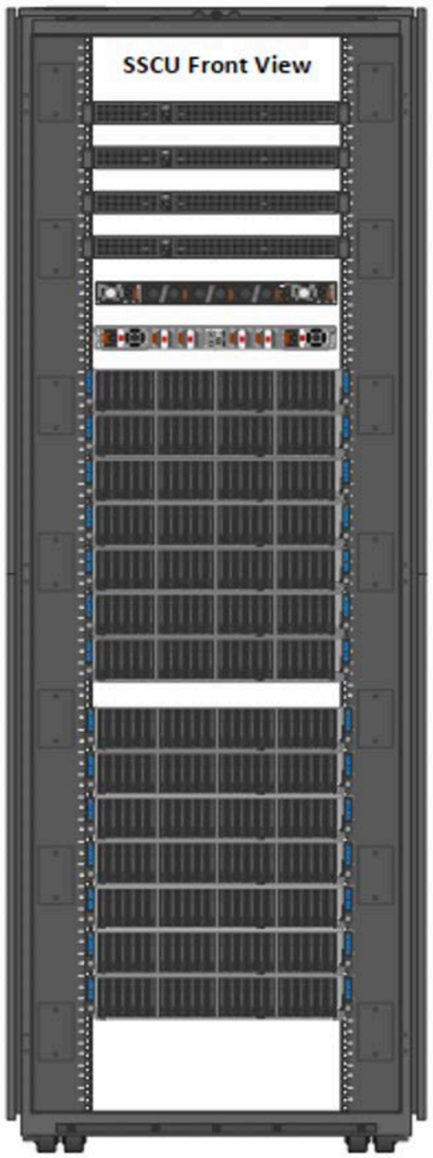
Non-blocking Interconnect

1 HDR Switch/SSCU

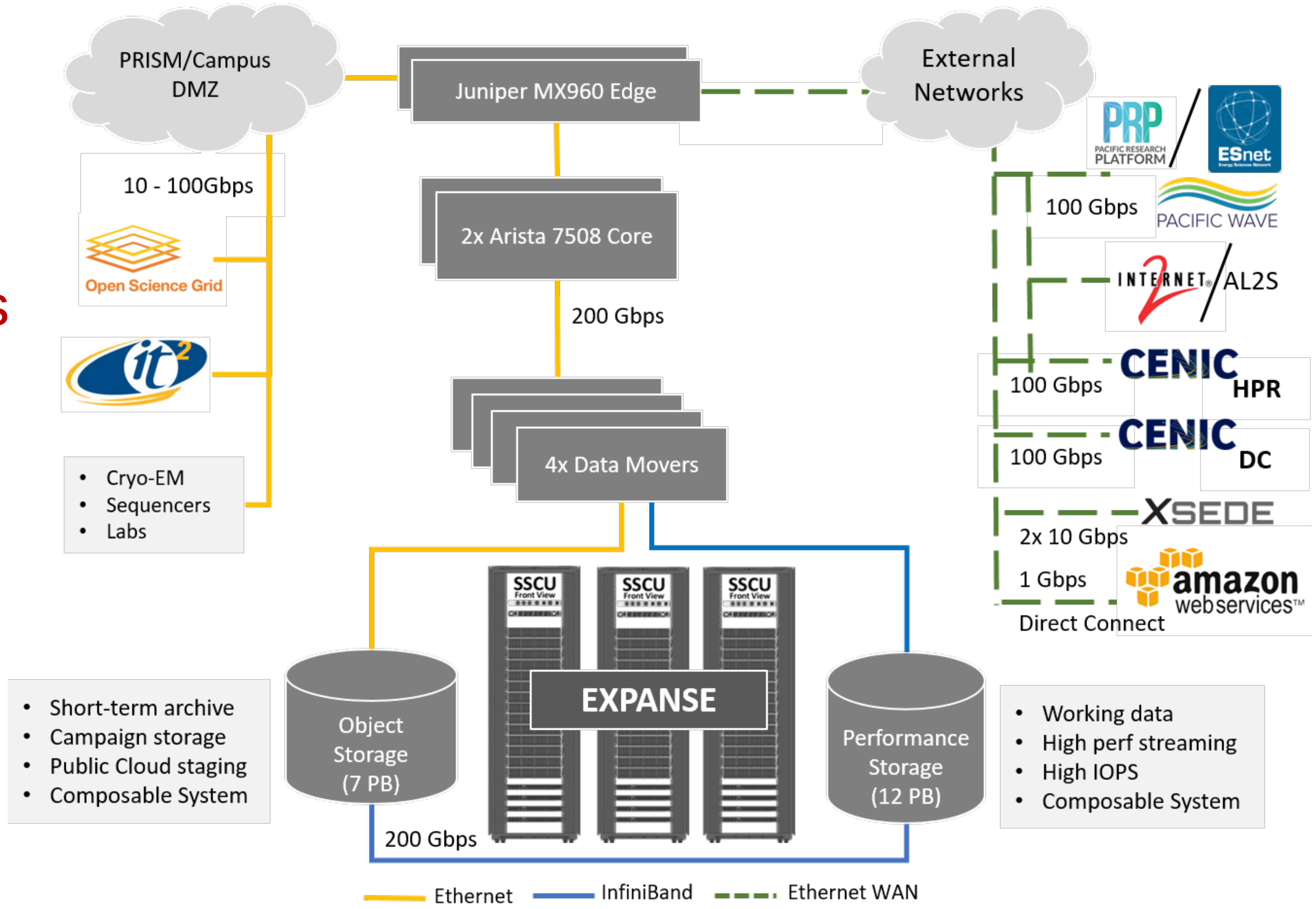
10x (200 Gbps)
HDR
56x *4x*
Compute Nodes *GPU Nodes*

5 Level 2 switches

Performance Storage *Cloud Storage*
3x *2x*
HDR
26x (200 Gbps)



Connectivity to R&E Networks Facilitates Compute and Data Workflows



Initial Benchmarks of Applications on AMD Rome Hardware

- Benchmarked CPU Applications: GROMACS, NAMD, NEURON, OpenFOAM, Quantum Espresso, RAxML, WRF, and ASTRAL.
- MPI, Hybrid MPI/OpenMP, and Hybrid MPI/Pthreads cases. Compilers used included AOCC, gnu, and Intel.
- Early results on test clusters shows per-core performance of 1-1.8X faster than Comet’s Haswell cores
- Overall throughput is expected to be easily more than 2X of Comet.
- As Expanse hardware comes online at SDSC, more benchmarks will be performed.

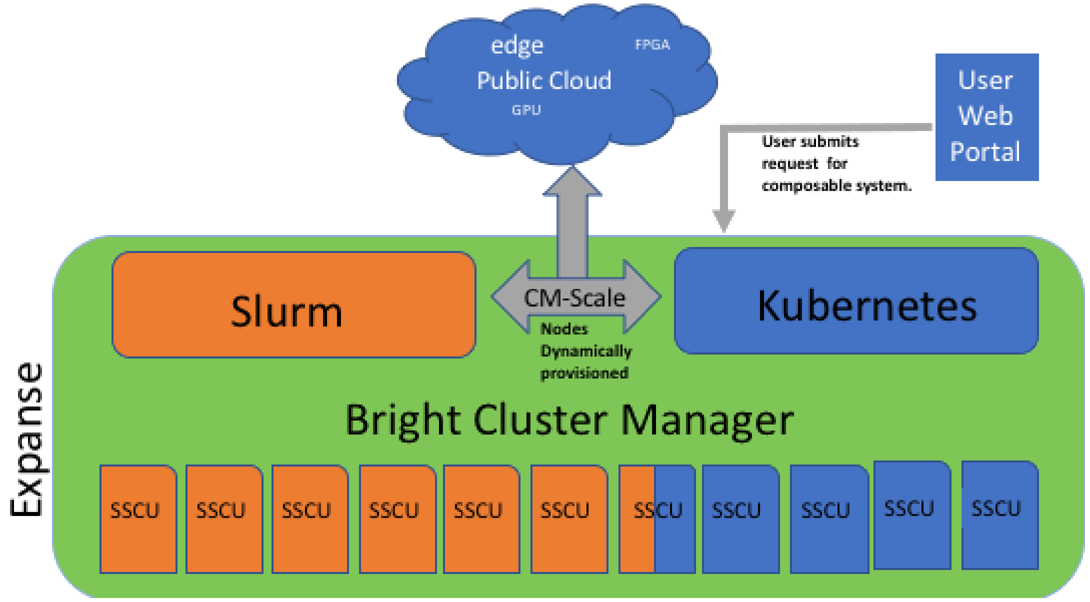
Integration with public cloud supports projects that share data, need access to novel technologies, and integrate cloud resources into workflows

- Slurm + in-house developed software + Terraform (Hashicorp)
- Early work funded internally and via NSF E-CAS/Internet2 project for CIPRES (Exploring Cloud for the Acceleration of Science, Award #1904444).
- Approach is cloud-agnostic and will support the major cloud providers
- Users submit directly via the Slurm, or as part of a composed system
- Options for data movement: data in the cloud; remote mounting of file systems; cached filesystems (e.g., StashCache), and data transfer during the job.

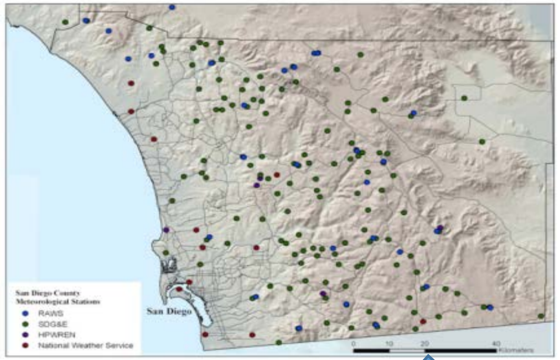
* Funding for user cloud resources is not part of the Expanse award. Researcher must have access to these via other NSF awards and funding.

Composable Systems will support complex, distributed, workflows – making Expanse part of a larger CI ecosystem

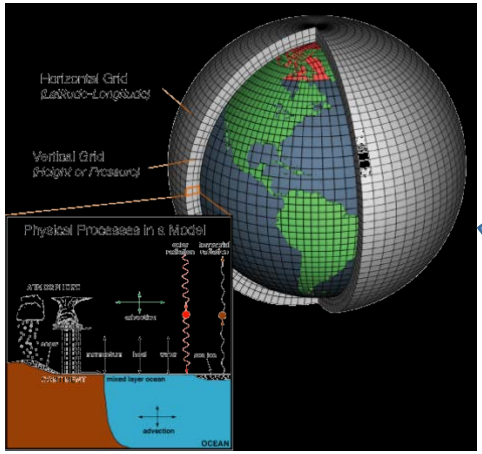
- Bright Cluster Manager + Kubernetes
- Core components developed via NSF-funded CHASE-CI (NSF Award # 1730158), and the Pacific Research Platform (NSF Award # 1541349)
- Requests for a composable system will be part of an XRAC request
- Advanced User Support resources available to assist with projects - **this is part of our operations funding.**



Fire Weather Monitoring and Prediction in WIFIRE

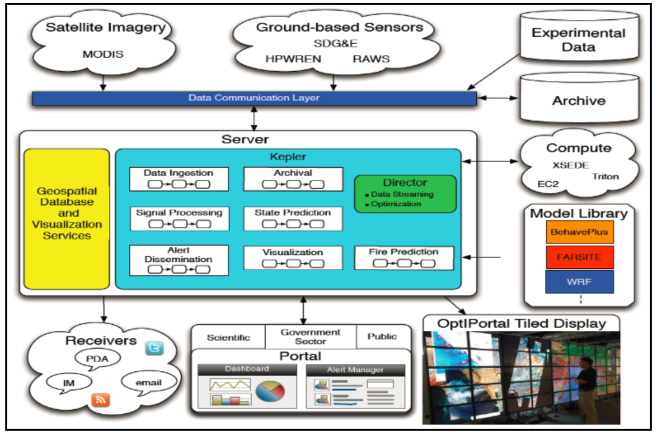


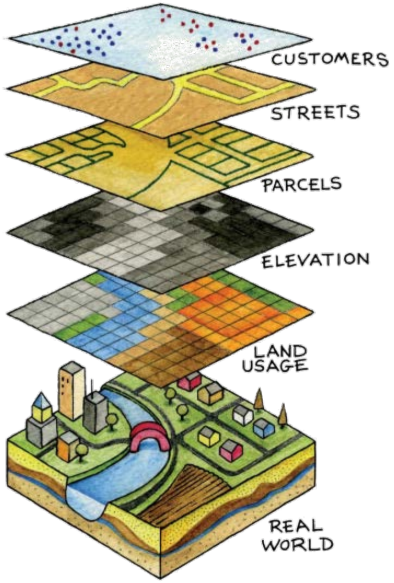
Real-time sensors



Weather forecast

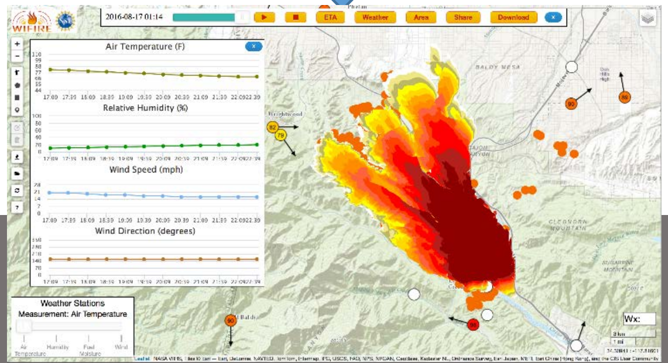
Fire perimeter





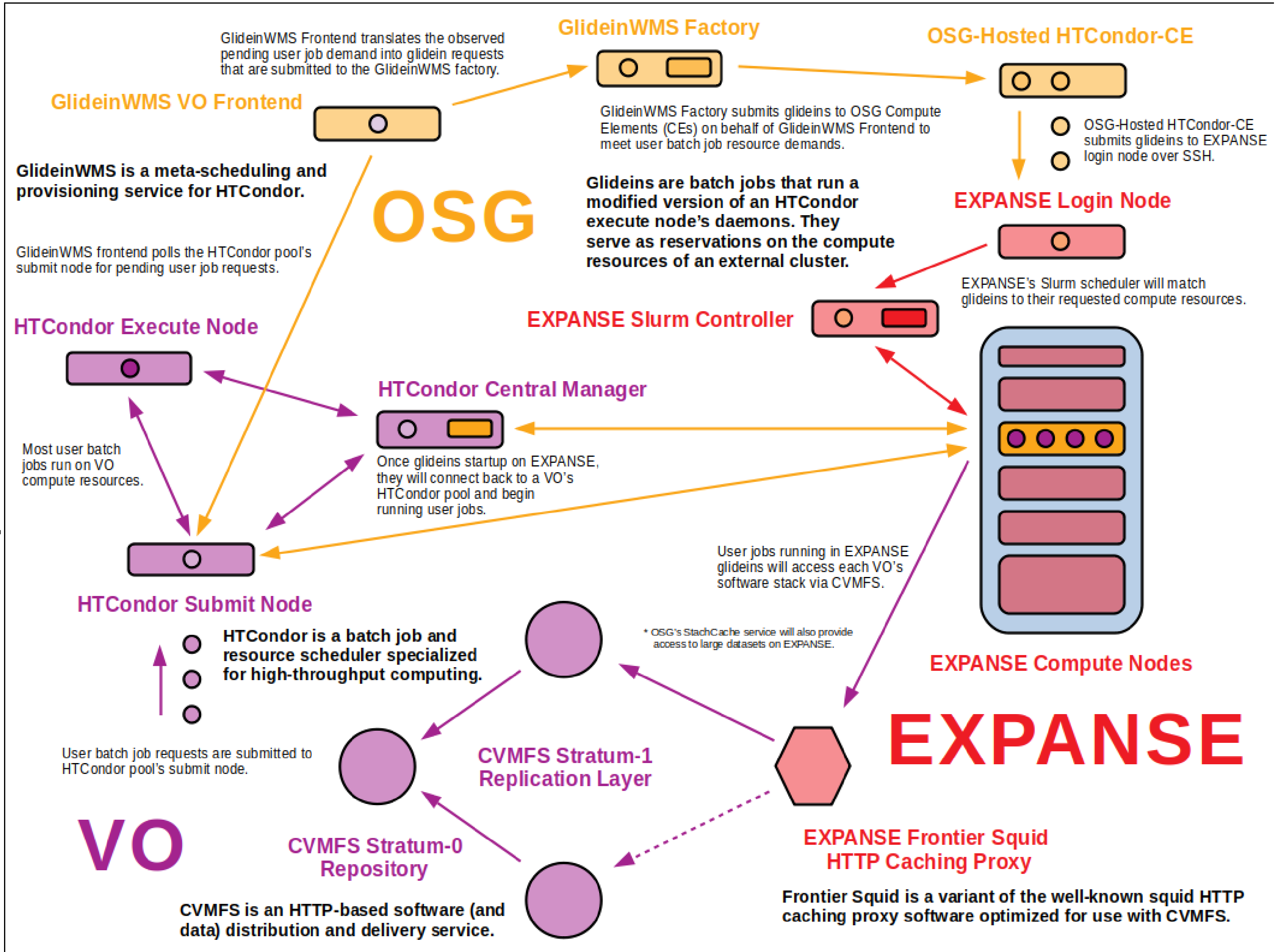
Landscape data

Monitoring & fire mapping



Expanse will integrate with the Open Science Grid

- HTCondor-CE per VO
- Allocations made directly to XSEDE at a project level ->> on behalf of a Virtual Organization (VO)
- CVMFS and StashCache for efficient software and data distribution
- Preemptable queue will run at a reduced rate
- Slurm TRES for fine-grained node partitioning



Source: Marty Kandes, SDSC

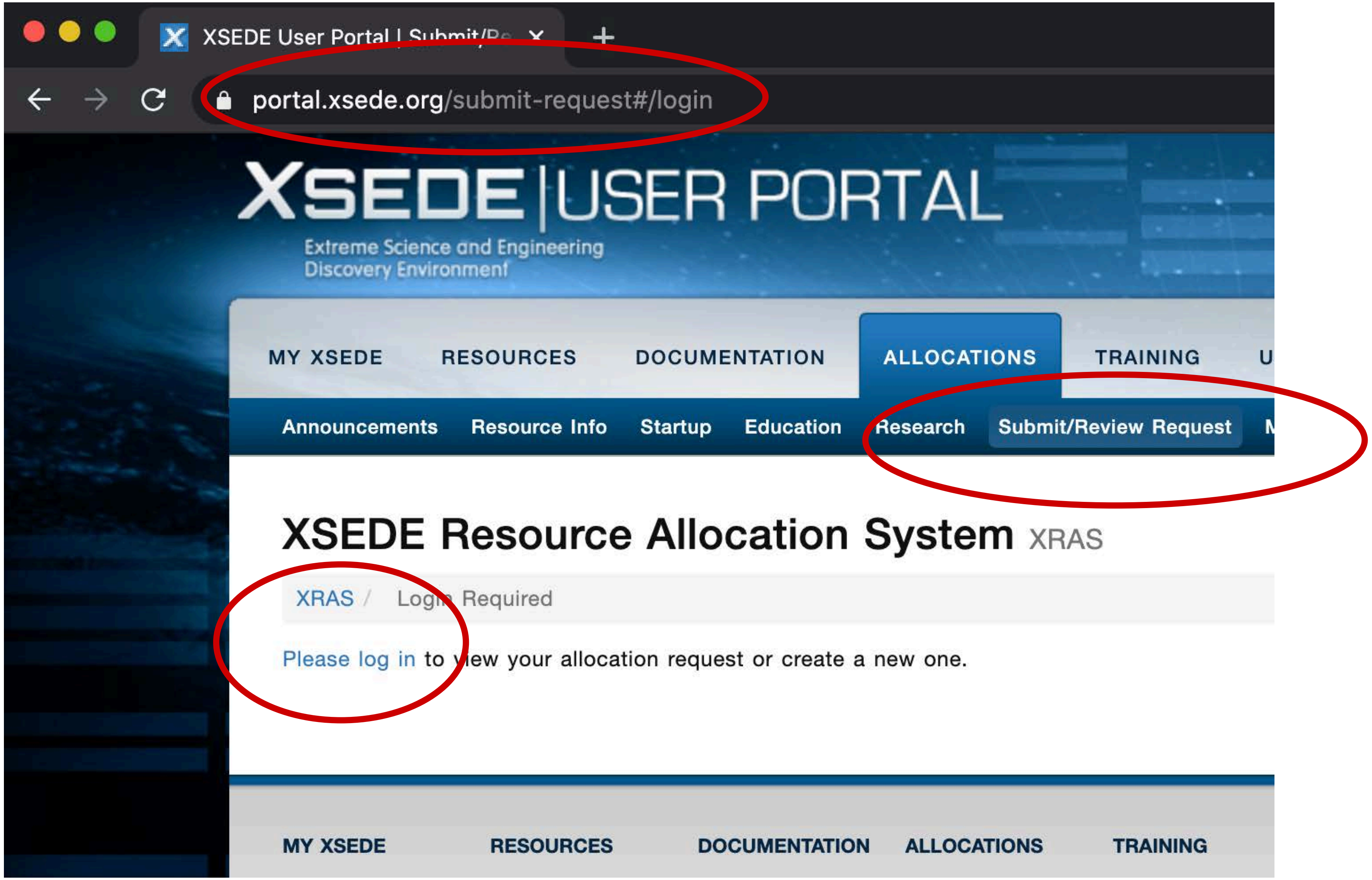
User support, training, outreach, and education will help users make the most of Expanse's traditional and innovative features

- Fully integrated as an XSEDE Level 1 Resource
- Overlap of 6 months in Comet and Expanse operations. Training for users transitioning from Comet to Expanse.
- A new program, HPC@MSI targeted at Minority Serving Institutions will make use of Directors Discretionary time that can be awarded via a rapid review process
- Advanced Support available from SDSC staff for cloud integration and composable systems projects.

I want to use Expanse, what do I do?

- Learn about XSEDE
  - https://www.xsede.org
- Learn about XSEDE Allocations
  - https://portal.xsede.org/allocations/policies#30
- Determine your eligibility:
  - *Researcher or educator at a U.S. academic or non-profit research institution; Post-doctoral researcher; NSF Graduate Student Fellows and Honorable Mention recipients; Qualified advisor e.g., a high school teacher or faculty member on behalf of high school students or undergraduate and graduate students*
  - https://portal.xsede.org/allocations/policies#22
- Determine what kind of allocation is right for you: Trial account? Startup? Educational? Research?
  - https://portal.xsede.org/allocations/policies#30
- Start with a small allocation and work your way up
- Use XSEDE and SDSC resources to help you develop your allocation request
- See if your campus has a Campus Champion (and allocation)
  - https://www.xsede.org/community-engagement/campus-champions

Allocations (XSEDE Portal)



Expanse Allocations

- Expanse resources can be requested in the upcoming XRAC submission period (September 15 - October 15) for allocations starting January 1, 2020.
  - https://portal.xsede.org/submit-request
- Startup and Trial allocations will be available at production launch and can be requested at any time
- Three resources related to Expanse:
  - **Expanse**: For allocations on compute (AMD Rome) part of the system.
  - **Expanse GPU**: For allocations on the GPU (V100) part of the system.
  - **SDSC Expanse Projects Storage**: Allocations on Expanse projects storage space* (will be mounted on both compute and GPU part of system).
- **Ceph** storage option coming next year

*Total space available will be 5PB (The 12 PB Lustre based filesystem will be split between projects and scratch areas)

Important Dates

- **Hardware delivery,** installation, application stack development, and initial testing. Now!!
- **Expanse Early Access Period:** Sept 1-30, 2020
- **Training for Comet to Expanse transition:** September 2020
- **6-month overlap with Comet.** Existing users with allocations will be transferred
- **Expanse 101: Accessing and running jobs:** Late September 2020
- **Production operations begin:** October 1, 2020
- **Next XRAC Allocation submission period:** Sep 15 – Oct 15, 2020. Review of these submissions will be in December for allocations that start January 1, 2021.

Thank you!!

We look forward to seeing you on Expanse!!

Follow all things Expanse at https://expanse.sdsc.edu

Thank you to our collaborators, partners, users, and the SDSC team!

XSEDE
Extreme Science and Engineering Discovery Environment

Ilkay Altintas	Mike Norman
Haisong Cai	Wayne Pfeiffer
Amit Chourasia	Scott Sakai
Trevor Cooper	Fernando Silva
Jerry Greenberg	Bob Sinkovits
Eva Hocks	Subha Sivagnanam
Tom Hutton	Michele Strong
Christopher Irving	Shawn Strande
Marty Kandes	Mahidhar Tatineni
Amit Majumdar	Mary Thomas
Dima Mishin	Nicole Wolter
Sonia Nayak	Frank Wuerthwein

EXPANSE
COMPUTING WITHOUT BOUNDARIES

In Production October 2020