Expanse: System Overview
October 29, 2020

Mahidhar Tatineni
SDSC

EXPANSE
COMPUTING WITHOUT BOUNDARIES

SAN DIEGO SUPERCOMPUTER CENTER

NSF Award 1928224

Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse innovative features
- Allocations
- Summary

Computing Without Boundaries: Cyberinfrastructure for the Long Tail of Science

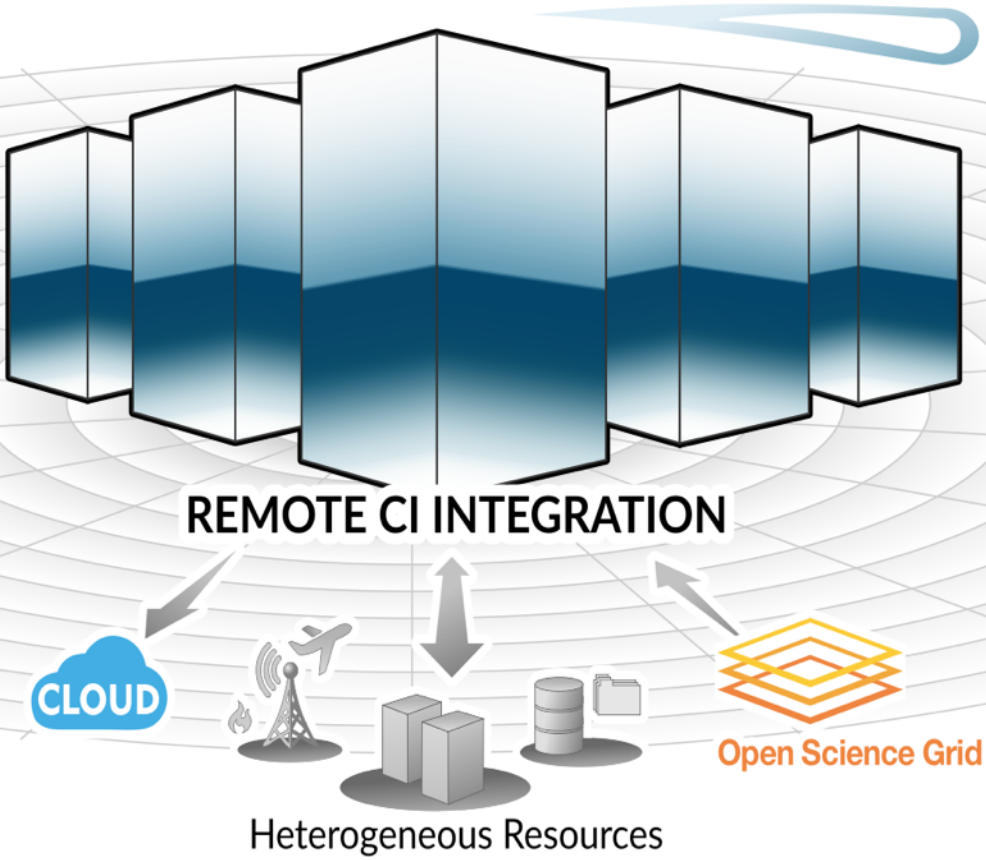
- NSF Solicitation 19-534: Advanced Computing Systems & Services: Adapting to the Rapid Evolution of Science and Engineering Research
- Category 1: Capacity System, NSF Award # 1928224
- NSF Program Officer: Robert Chadduck
- PIs: Mike Norman (PI), Ilkay Altintas, Amit Majumdar, Mahidhar Tatineni, Shawn 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

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
- 52 4-way GPU nodes based on V100 w/NVLINK
- Based on benchmarks we've run, we expect > 2x throughput over Comet (per-core improvement over Haswell, and 2x the core counts)
- Expect a smooth transition from Intel to AMD processors. SDSC team has compiled and run many of the common software packages on Expanse.
- At present, system is in early user period.
- **Production:** *November 2020 with operations for 5-years*

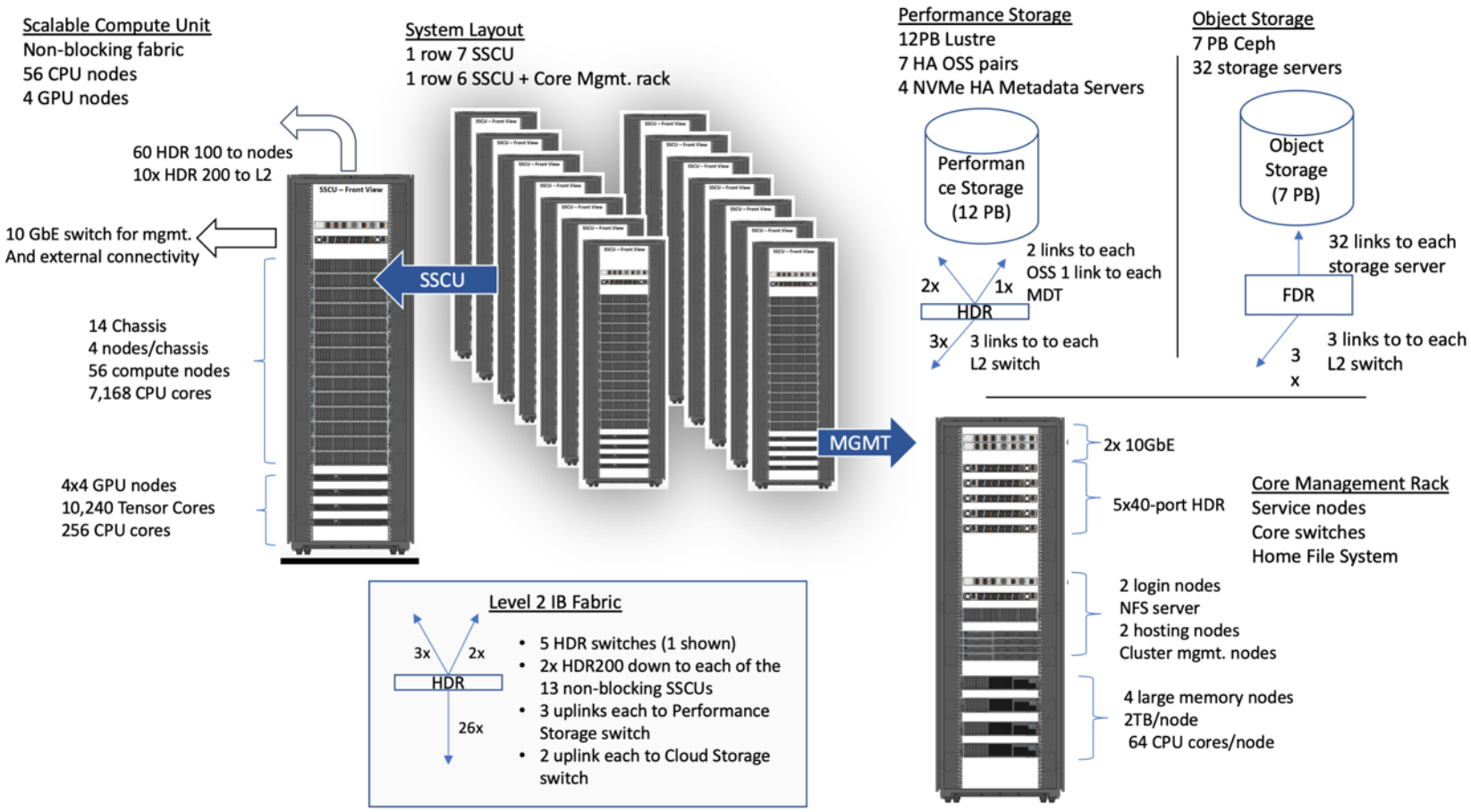
Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse innovative features
- Allocations
- Summary

Expanse is a heterogeneous architecture designed for high performance, reliability, flexibility, and productivity

System Summary

- *13 SDSC Scalable Compute Units (SSCU)*
- *728 x 2s Standard Compute Nodes*
- *93,184 Compute Cores*
- *200 TB DDR4 Memory*
- *52x 4-way GPU Nodes w/NVLINK*
- *208 V100s*
- *4x 2TB Large Memory Nodes*
- *HDR 100 non-blocking Fabric*
- *12 PB Lustre High Performance Storage*
- *7 PB Ceph Object Storage*
- *1.2 PB on-node NVMe*
- *Dell EMC PowerEdge*
- *Direct Liquid Cooled*



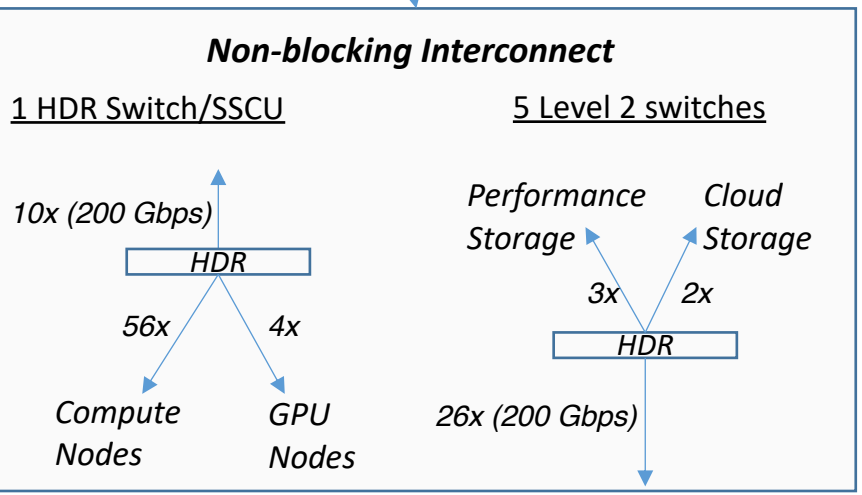
The SSCU is Designed for the Long Tail Job Mix, Maximum Performance, Efficient Systems Support, and Efficient Power and Cooling

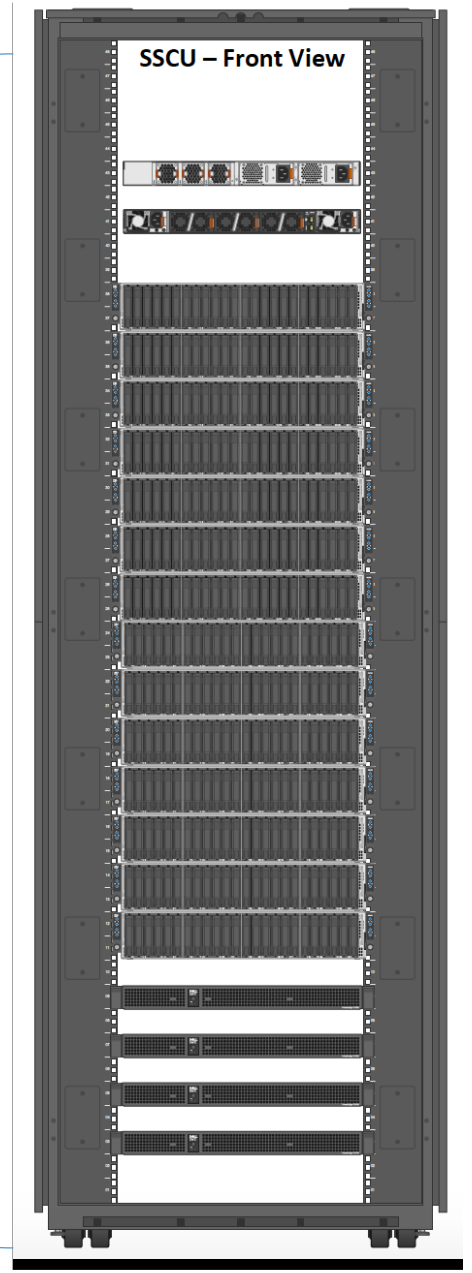
Standard Compute Nodes
- 2x AMD EPYC 7742 @2.25 GHz
- 128 Zen2 CPU cores
- PCIe Gen4
- 256 GB DDR4
- 1 TB NVME

GPU Nodes
- 4x NVIDIA V100/follow-on
- 10,240 Tensor Cores
- 32 GB GDDR
- 1.6 TB NVMe
- Intel CPUs

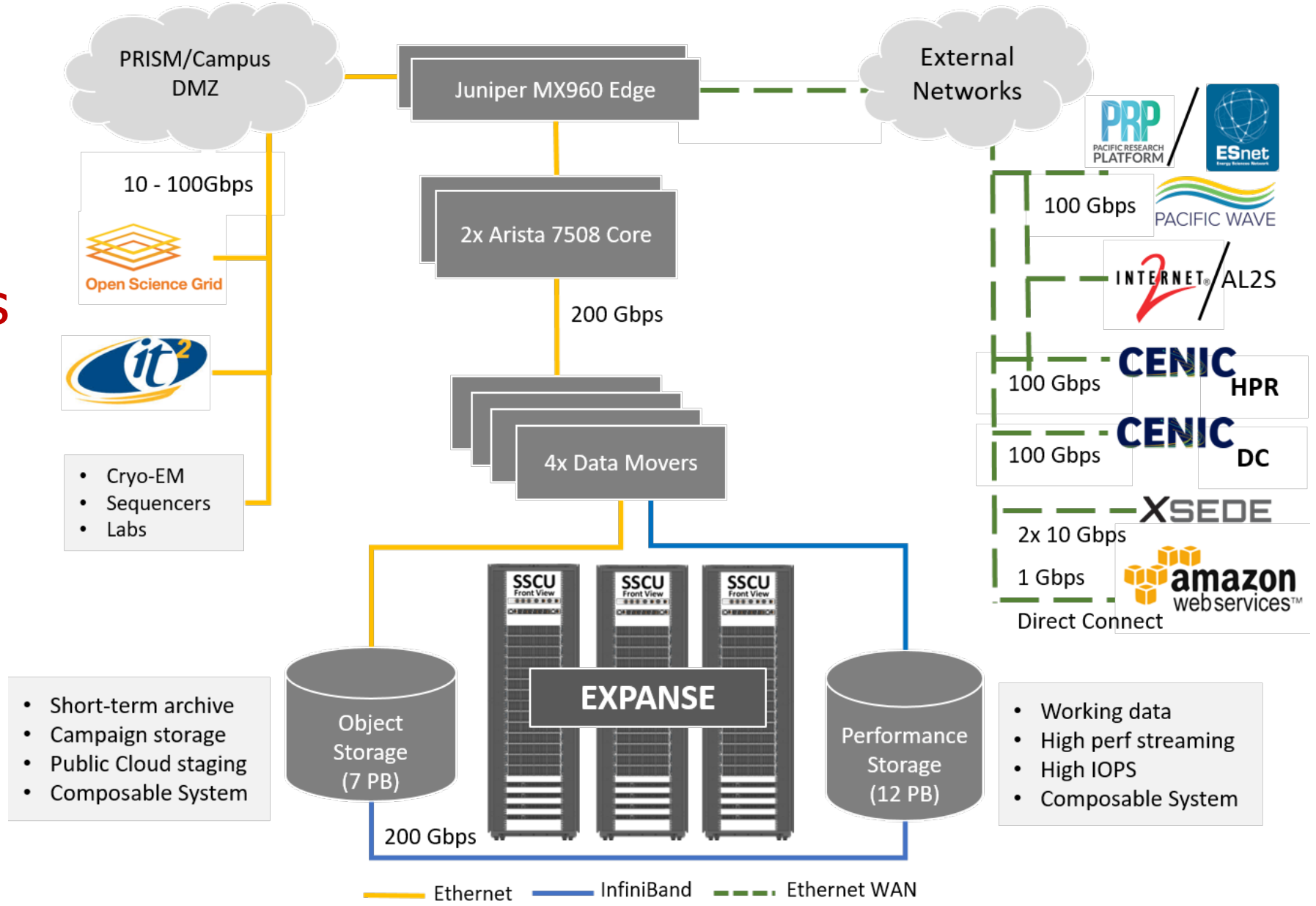
SSCU Components
- 56x CPU nodes
- 7,168 Compute Cores
- 4x GPU nodes
- 1x HDR Switch
- 1x 10GbE Switch
- HDR 100 non-blocking fabric
- Wide rack for serviceability
- Direct Liquid Cooling to CPU nodes

Non-blocking Interconnect





Connectivity to R&E Networks Facilitates Compute and Data Workflows

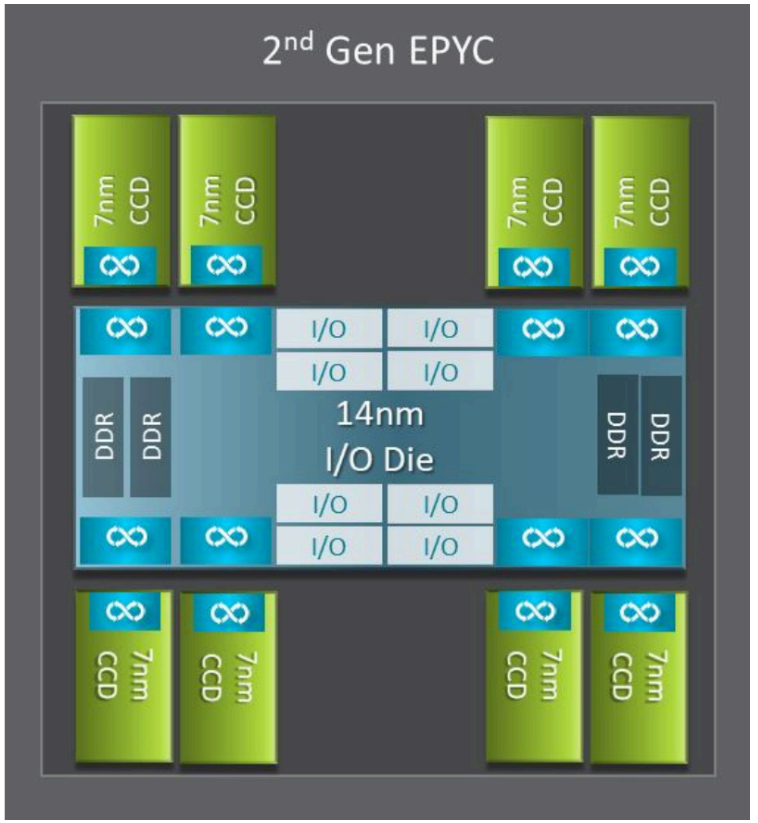


Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse innovative features
- Allocations
- Summary

AMD EPYC 7742 Processor Architecture

- 8 Core Complex Dies (CCDs).
- CCDs connect to memory, I/O, and each other through the I/O Die.
- 8 memory channels per socket.
- DDR4 memory at 3200MHz.
- PCI Gen4, up to 128 lanes of high speed I/O.
- Memory and I/O can be abstracted into separate quadrants each with 2 DIMM channels and 32 I/O lanes.



Reference: https://developer.amd.com/wp-content/resources/56827-1-0.pdf

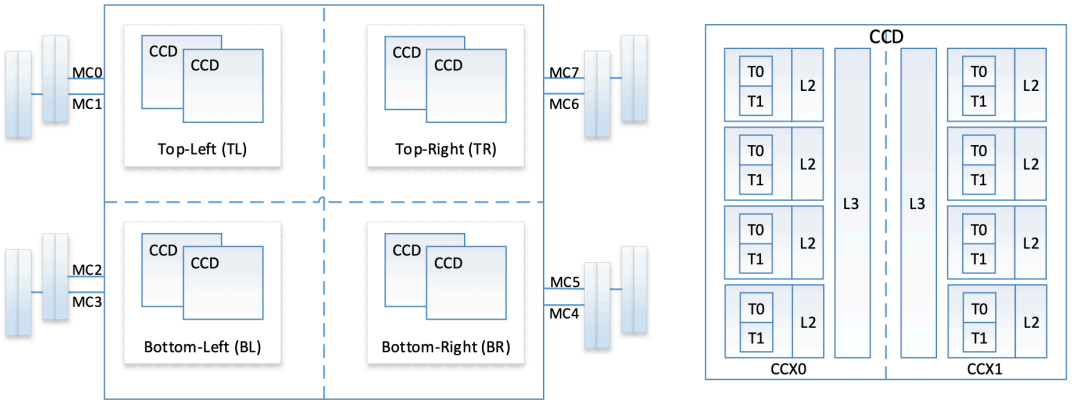
AMD EPYC 7742 Processor: Core Complex Die (CCD)

- 2 Core Complexes (CCXs) per CCD
- 4 Zen2 cores in each CCX shared a 16M L3 cache. Total of 16 x 16 = 256MB L3 cache.
- Each core includes a private 512KB L2 cache.

Reference: https://developer.amd.com/wp-content/resources/56827-1-0.pdf

AMD EPYC 7742 Processor : NUMA Nodes Per Socket

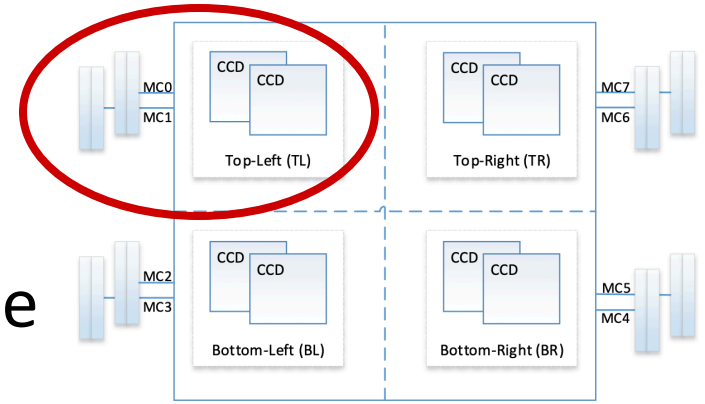
- The four logical quadrants allow the processor to be partitioned into different NUMA domains. Options set in BIOS.
- Domains are designated as NUMA per socket (NPS).
- **NPS4:** Four NUMA domains per socket is the typical HPC configuration.

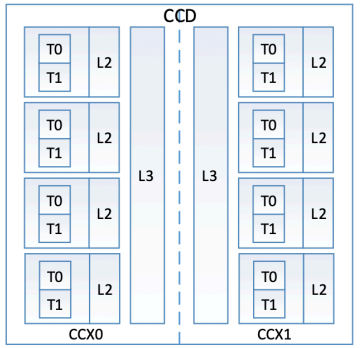


https://developer.amd.com/wp-content/resources/56338_1.00_pub.pdf

NPS4 Configuration

- The processor is partitioned into four NUMA domains.
- Each logical quadrant is a NUMA domain.
- Memory is interleaved across the two memory channels
- PCIe devices will be local to one of four NUMA domains (the IO die that has the PCIe root for the device)
- ***This is the typical HPC configuration*** as workload is NUMA aware, ranks and memory can be pinned to cores and NUMA nodes.





https://developer.amd.com/wp-content/resources/56338_1.00_pub.pdf

Compile and run time considerations

- Tested with AOCC, gnu, and Intel compilers. MPI versions include MVAPICH2, OpenMPI, and Intel MPI.

- Specific optimization flags:
  - AOCC, gcc/10.2.0: -march=znver2
  - Intel, gcc/9.2.0 : -march=core-avx2

- Runtime considerations:
  - MPI: Use binding options such as --map-by core (OpenMPI); I_MPI_PIN, I_MPI_PIN_DOMAIN (Intel MPI)
  - Open MP: Use affinity options like GOMP_AFFINITY, KMP_AFFINITY
  - Hybrid MPI/OpenMP, MPI/Pthreads: Keep threads on same NUMA domain (or CCX) as parent MPI task using affinity flags or wrapped with taskset (in case of MPI/Pthreads; used in RAxML runs for example)

Benchmarks of Applications on Expanse

- 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.
  - Results on Expanse show performance ranges from matching on a per core basis to 1.8X faster on a per core basis compared to Comet.
  - Overall throughput is expected to be easily more than 2X of Comet.

- Benchmarked GPU Applications: NAMD, AMBER, TensorFlow, PyTorch, MXNET, GROMACS, and BEAST
  - Results on Expanse show >1.5X per GPU improvement over the Comet P100 nodes.

GPU Node Architecture

- 4 V100 32GB SMX2 GPUs
- 384 GB RAM, 1.6 TB PCIe NVMe
- 2 Intel Xeon 6248 CPUs
- Topology:

```
        GPU0    GPU1    GPU2    GPU3    mlx5_0  CPU Affinity
GPU0     X      NV2     NV2     NV2     SYS     0-0,4-4,8-8,12-12,16-16,20-20,24-24,28-28,32-32,36-36
GPU1    NV2      X      NV2     NV2     SYS     0-0,4-4,8-8,12-12,16-16,20-20,24-24,28-28,32-32,36-36
GPU2    NV2     NV2      X      NV2     SYS     1-1,5-5,9-9,13-13,17-17,21-21,25-25,29-29,33-33,37-37
GPU3    NV2     NV2     NV2      X      SYS     1-1,5-5,9-9,13-13,17-17,21-21,25-25,29-29,33-33,37-37
mlx5_0  SYS     SYS     SYS     SYS      X

Legend:

  X    = Self
  SYS  = Connection traversing PCIe as well as the SMP interconnect between NUMA nodes (e.g., QPI/UPI)
  NODE = Connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA node
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe bridges (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing at most a single PCIe bridge
  NV#  = Connection traversing a bonded set of # NVLinks
```

Software Stack

- Expanse will support a broad application base with installs and modules for commonly used packages in bioinformatics, molecular dynamics, machine learning, quantum chemistry, structural mechanics, and visualization.

- The current application stack on Comet will be replicated on Expanse. Today's talks are aimed at informing Comet users about transition to the Expanse application environment.

- Primarily Spack based installs. Continued support for Singularity based containerization on Expanse.

Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse innovative features
- Allocations
- Summary

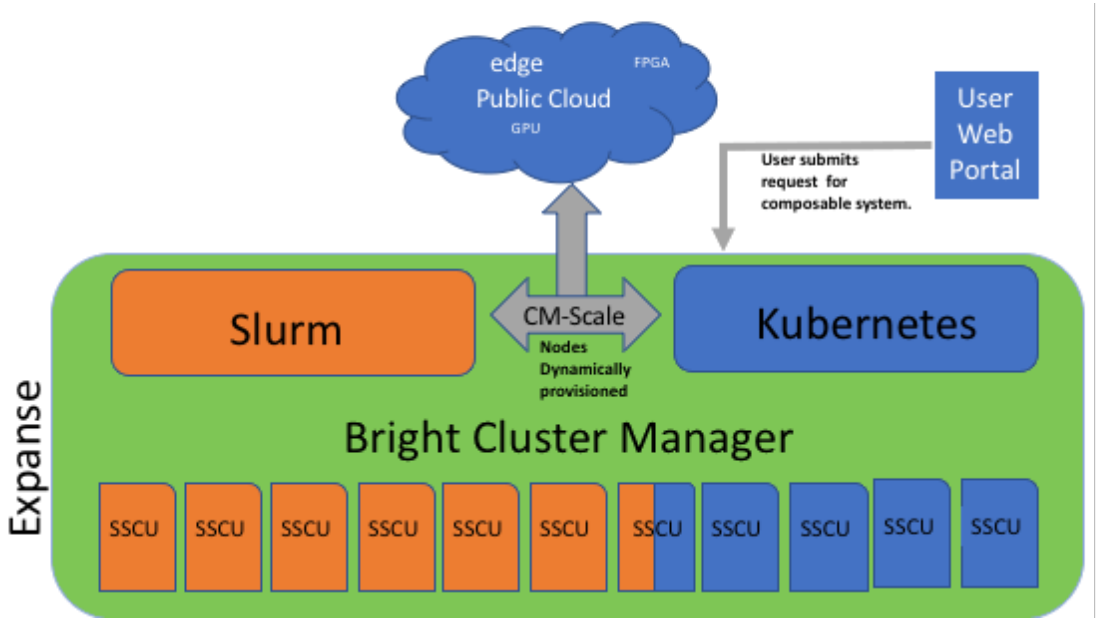
Integration with public cloud supports projects that share data, need access to novel technologies, and integrate cloud resources into workflows

- 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.**



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.
- ***Today's workshop: 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.
- We will be working with XSEDE/XRAC to develop review criteria for the innovative elements of Expanse

Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse innovative features
- Allocations
- Summary

Allocations

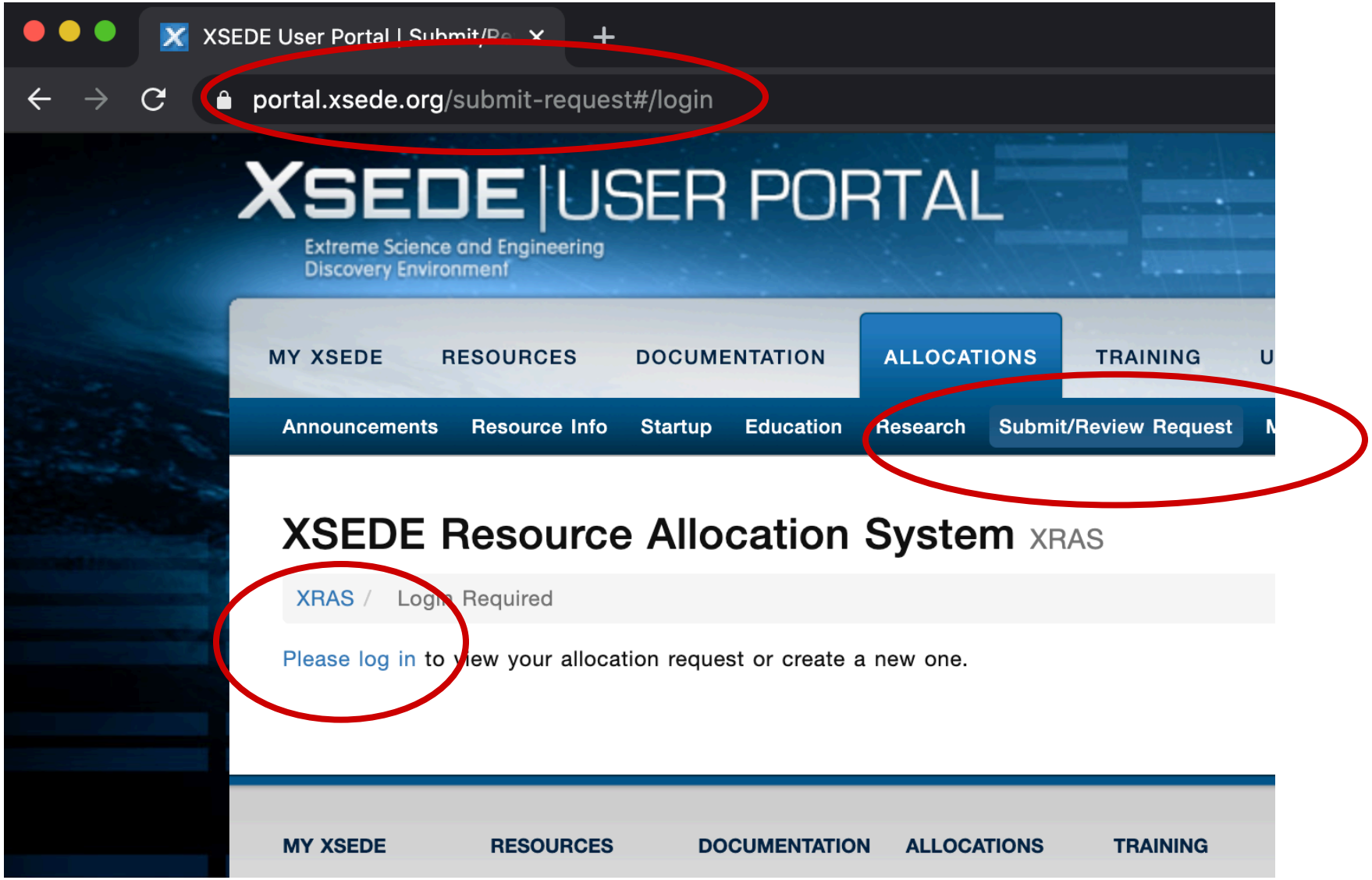
- Expanse resources can be requested in XSEDE XRAC windows.

  https://portal.xsede.org/submit-request

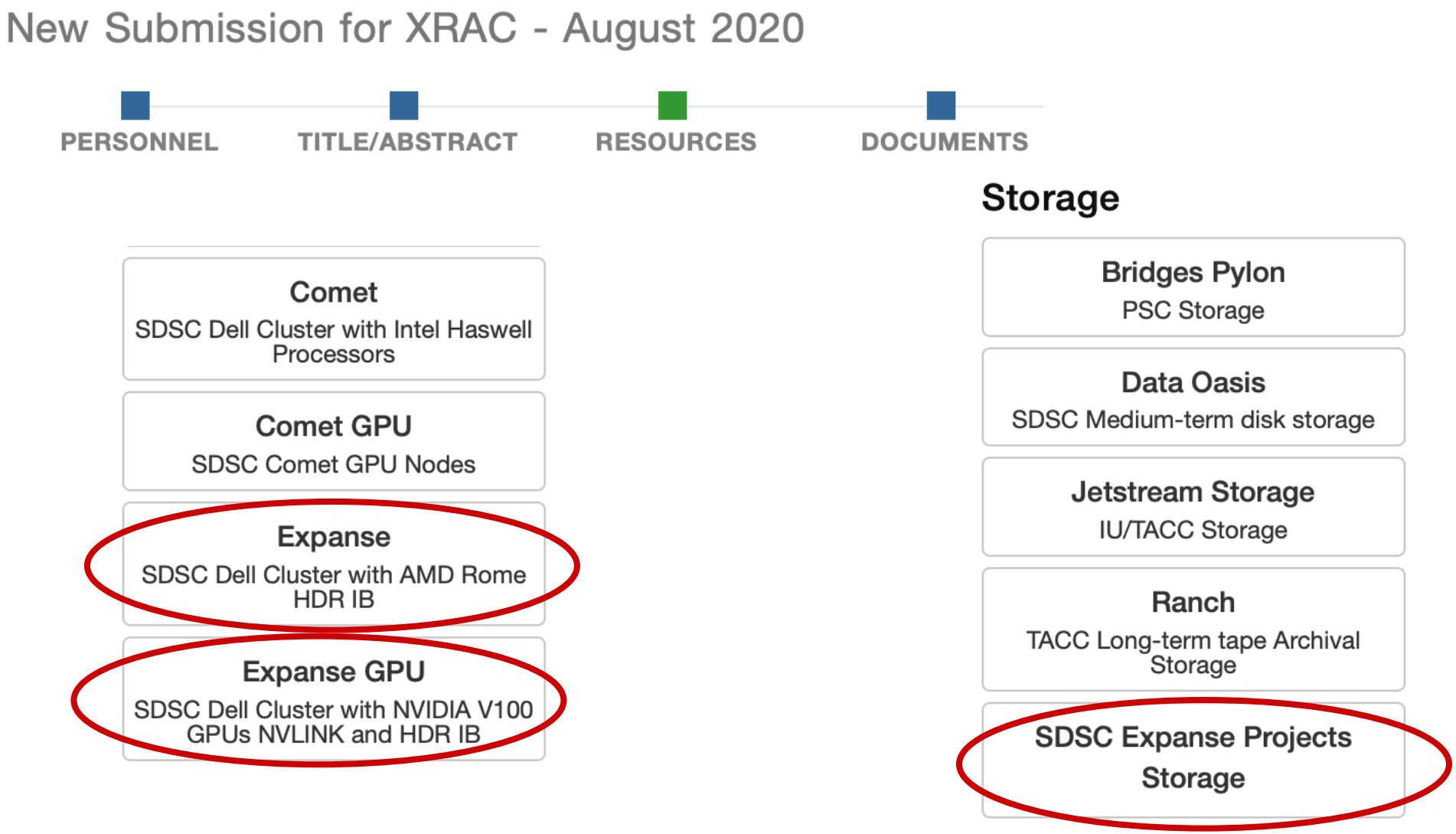
- Three resources related to Expanse:
  - **Expanse**: For allocations on compute (AMD Rome) part of the system.
  - **Expanse GPU**: For allocations on the GPU (V100) part of the system.
  - **SDSC Expanse Projects Storage**: Allocations on Expanse projects storage space* (will be mounted on both compute and GPU part of system).

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

Allocations (XSEDE Portal)



Allocations Request (Resources Tab)



Summary

- Expanse will provide a substantial increase in the performance and throughput compared to the highly successful, NSF-funded Comet supercomputer.
- Expanse is an evolution of the Comet design with innovations in cloud integration and composable systems and continued support for science gateways and distributed computing via the Open Science Grid.
- 728, 2-socket AMD-based compute nodes (2.25 GHz EPYC; 64-cores/socket) and 52 4-way GPU nodes based on V100 w/NVLINK.
- HDR InfiniBand interconnect – HDR100 to the nodes and HDR200 switches.
- **Early access period ongoing**. *Production November, 2020*.
- Follow all things Expanse at https://expanse.sdsc.edu !

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

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

EXPANSE
COMPUTING WITHOUT BOUNDARIES

SAN DIEGO SUPERCOMPUTER CENTER

IN PRODUCTION OCTOBER 2020