Expanse Overview

Robert Sinkovits, PhD

Director of Education

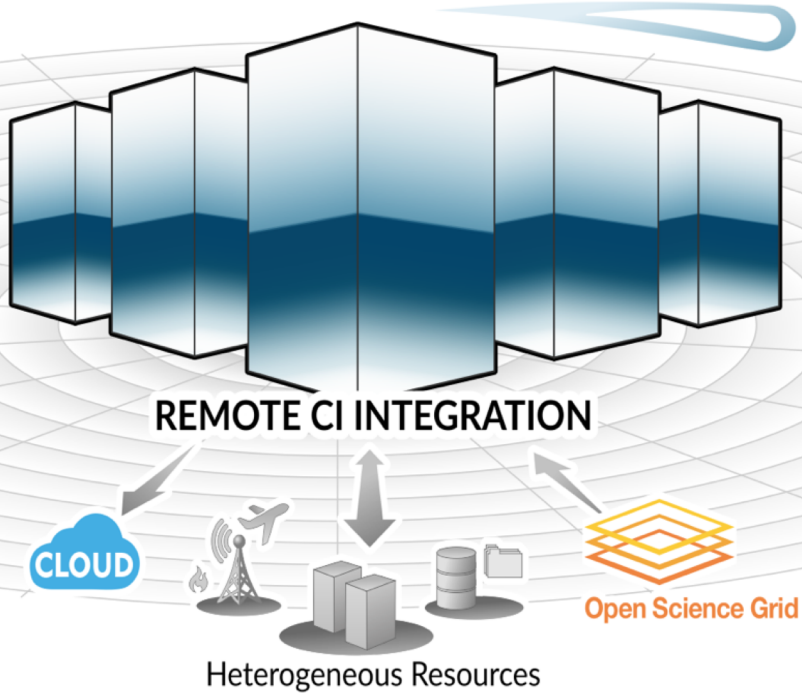
San Diego Supercomputer Center

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 dedicated to the commitment and hard work of SDSC staff, who carried out and supported this work in the context of COVID-19

Ilkay Altintas
Haisong Cai*
Amit Chourasia
Trevor Cooper*
Csilla Csori
Mike Dwyer
Jeff Filliez*
Keith Green*
Jerry Greenberg
Eva Hocks*
Tom Hutton*
Christopher Irving*
Marty Kandes
Sophorn Khem*
Amit Majumdar
Jeff Makey
Tim McNew*
Dima Mishin
Sonia Nayak
Mike Norman
Nicolas Patience*
Ismael Perez
Wayne Pfeiffer
Susan Rathbun
Scott Sakai*
Jeff Sale
Manu Shantharam
Fernando Silva*
Bob Sinkovits
Subha Sivagnanam
Fred Spinny*
Anthony Steinell*
Michele Strong
Shawn Strande
Mahidhar Tatineni
Mary Thomas
Ben Tolo
Nicole Wolter
Cindy Wong
Frank Wuerthwein

*A special thanks to the HPC Systems Group and the Data Center staff who have been performing work onsite under very difficult constraints

Expanse System Summary

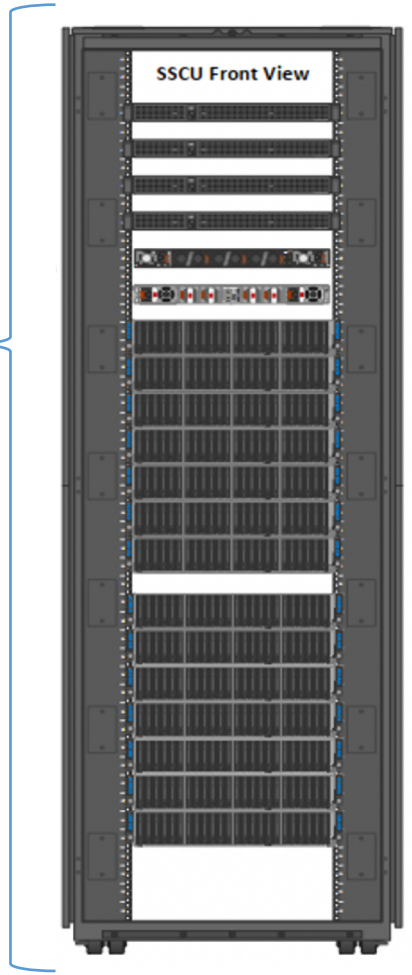
System Component	Configuration
AMD EPYC (Rome) 7742 Compute Nodes	
Node count	728
Total # cores	93,184
Total DRAM	182 TB
Total NVMe Stroage	728 TB
NVIDIA V100 GPU Nodes	
Node count	52
Total # GPUs	208
Total GPU Memory	1.6 TB
Total # cores;	2080
Total NVMe	1.6TB
Large Memory Nodes	
Number of nodes	4
Memory 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

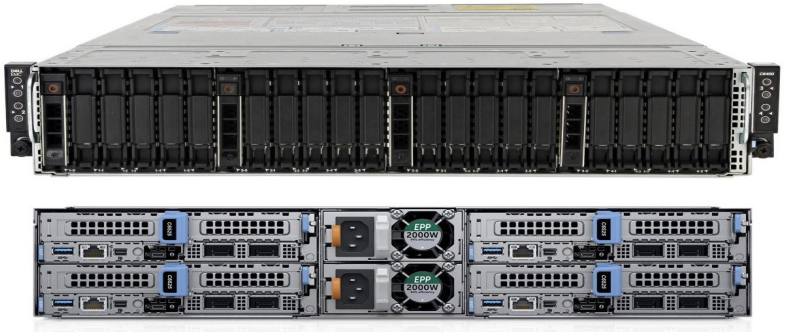
The SSCU is designed for maximum performance, efficient systems support, and efficient power and cooling

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*



Compute nodes are next generation and familiar to SDSC



- **Dell PowerEdge C6525**
- Dual Socket AMD Epyc 7742 with Direct Liquid Cooling
- 64 Cores per socket, 128 per node
- 16 Dual Ranked 16GB DDR4 DIMMs per nodes.
- 1 DIMM per memory channel
- 1 TB NVMe drive for local scratch space.
- 100 Gb InfiniBand HDR Data I/O Network
- 25 Gb Ethernet for management and External access.

Dell PowerEdge C6525	
AMD EPYC (Rome) 7742 Compute Nodes	
CPU count	2 X AMD EPYC 7742 (DLC)
Clock speed	2.25 GHz
Cores/node	128
DRAM/node	256 GB. (16 X 16GB DDR4 3200)
NVMe/node	1 TB
SSD/node	1 240 GB
Infiniband	100 Gbps HDR
Ethernet	25 Gbps Ethernet

GPU Nodes are also a generational improvement on *Comet* GPU nodes

- **Dell PowerEdge C4140**
- 2 Intel
- 24 Dual Ranked 16GB DIMMs per node
- 4 Nvida V100 SMX2
- NVLink GPU Interconnect
- 32 GB RAM per GPU
- 1.6 TB NVMe Drive for local Scratch
- 100 Gb InfiniBand HDR
- 10 Gb Ethernet

Dell PowerEdge C4140	
CPUs	2 X Intel(R) Xeon(R) Gold 6248
Clock speed	2.5 GHz
Cores/node	40
DRAM/node	384 GB. (24 X 16GB DDR4 2933)
GPUs	4
GPU Type	V100 SMX2
GPU Mem	32 GB / GPU
NVMe/node	1.6 TB
SSD/node	1 240 GB
Infiniband	100 Gbps HDR
Ethernet	10 Gbps Ethernet

Large Memory Nodes

- **Dell PowerEdge R6525**
- Dual Socket AMD Epyc 7742
- 64 Cores per socket, 128 per node
- 2 TB of RAM per node.
- 32 Dual Ranked 16GB DDR4 DIMMs per nodes.
- Two 1.6 TB NVMe drives for local scratch space.
- 100 Gb InfiniBand HDR Data I/O Network
- 25 Gb Ethernet for management and External access.

Dell PowerEdge R6525	
CPU count	2 X AMD EPYC 7742
Clock speed	3.3 GHz
Cores/node	128
DRAM/node	2 TB. (32 X 64 GB DDR4 2933)
NVMe/node	2 X 1.6TB
SSD/node	1 240 GB
Infiniband	100 Gbps HDR
Ethernet	25 Gbps Ethernet

Collaboration with Dell and AMD led to optimized system EPYC settings for maximum performance

NUMA per Socket (NPS) set to 4

- Separates the CPU into four NUMA domains
- 8 Cores share two interleaved memory channels
- Provides best overall memory performance

Preferred IO Devices settings

- These impact I/0 traffic prioritization and are crucial for insuring best IB performance
- PCIe Preferred IO Bus set to enabled
- PCIe Preferred IO Bus Val set the PCIe Bus address of the IB device
- Enhanced Preferred IO is enabled

Other Settings

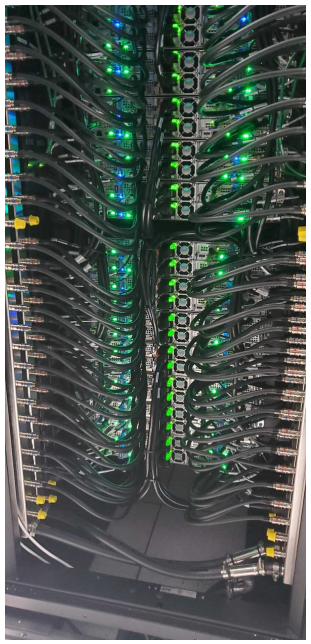
- CCX as NUMA is disabled
- MADT Core Enumeration set to linear
- Logical Processors is disabled (for now)

https://www.dell.com/support/article/en-bb/sln319015/amd-rome-is-it-for-real-architecture-and-initial-hpc-performance?lang=en

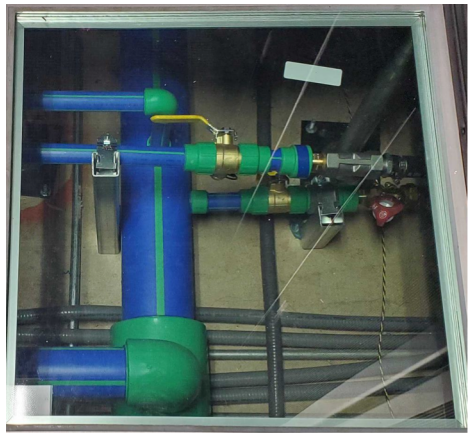
DLC =

Typical C6525 sled with DLC

+



+



+

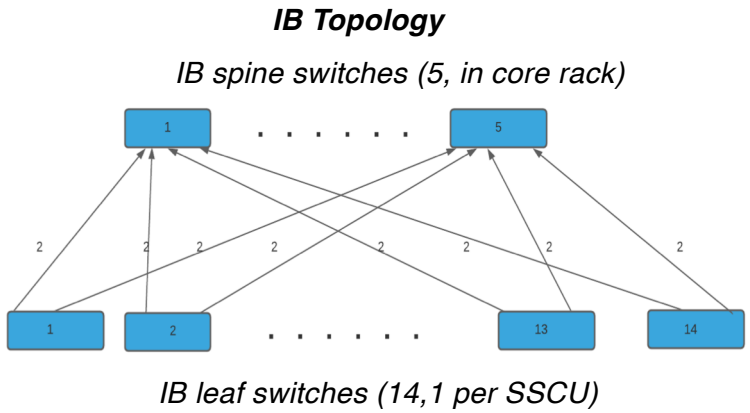
HDR non-blocking topology in each SSCU, and 100GbE core provide outstanding performance and manageability

InfiniBand Fabric is a two-layer HDR Fat Tree

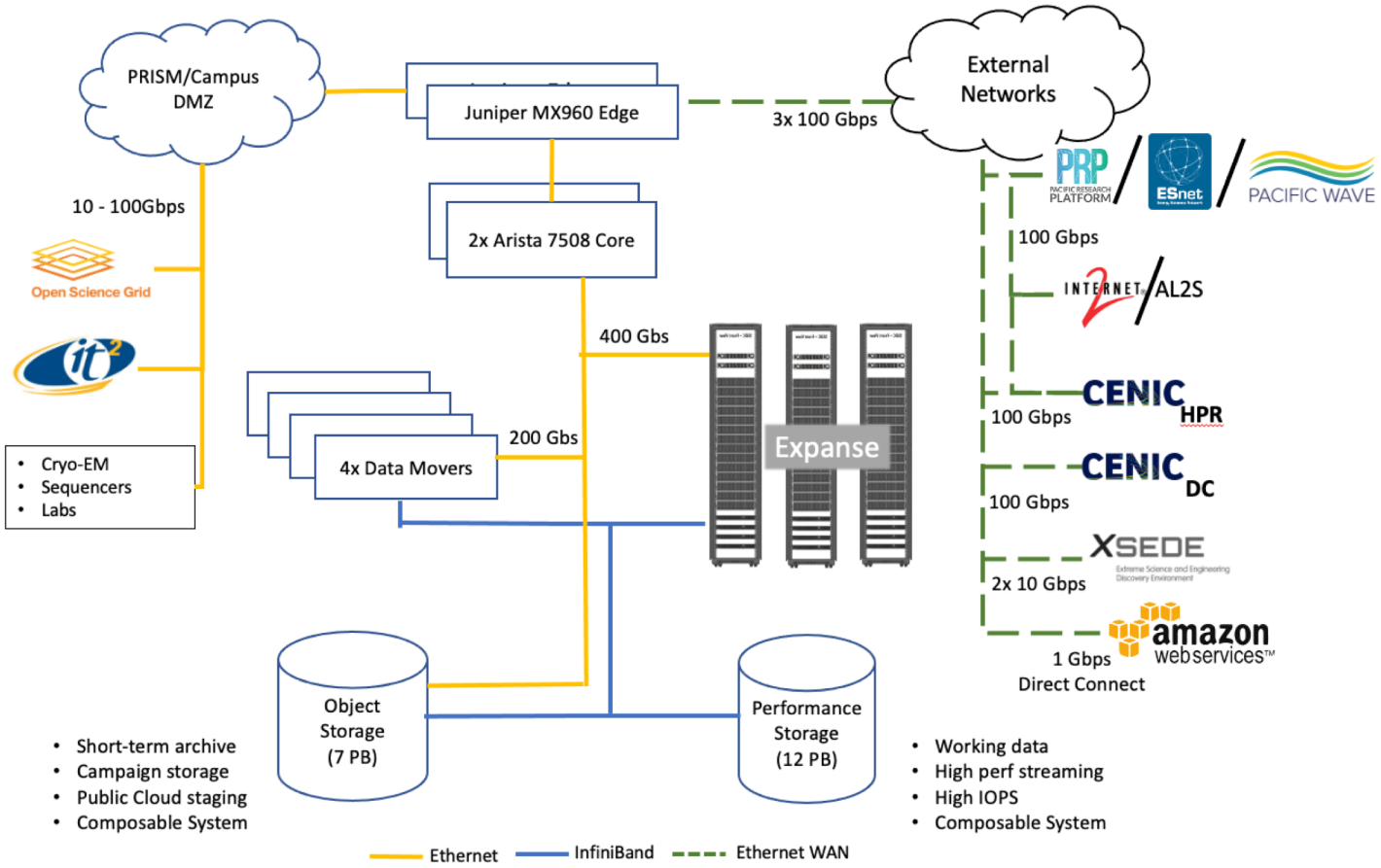
- Five spine switches and fourteen leaf switches
- One leaf switch per SSCU
- 60x HDR 100 downlinks to the compute nodes
- 10x HDR 200 uplinks from each SSCU to the spine switches (two per spine switch)
- 3:1 Oversubscription to the core.
- Within each SSCU there is full bisection

Ethernet

- Two 100GbE Core switches
- One Ethernet switch per SSCU with a 100GbE uplink to each core switch
- 25GbE down to the compute nodes, 10GbE to the GPUs
- Four 100 GbE Connections From Expanse to the Datacenter Core



Connectivity to R&E Networks Facilitates Compute and Data Workflows



Tiered storage leads to improved job performance and overall better system utilization and throughput

Node local NVMe Scratch drives provide excellent performance for workloads that don't need to share data files between nodes.

Performance Parallel Distributed Lustre filesystem for I/O workloads that require high-bandwidth and large capacity shared storage.

Ceph Object Storage for Sort-term archival storage and staging data transfers to cloud-base storage.

High-Availablity Network Files System (NFS) Cluster for user home directory storage.

Expanse continues our practice from *Gordon* and *Comet* of having user accessible local SSD in every compute and GPU node

- SDSC's Trestles and Gordon systems were the first large HPC system to feature Solid State Drives (SSDs)
- This has proven to be an excellent approach to improving user application and system performance
- Expanse continues with NVMe local scratch on every node

Benefits of local NVMe

1. Low latency Read/Write operations
2. High IOPS rate
3. Reduce load on network file systems
4. Improved user application performance
5. Easily managed and serviced

Node	Size	#
Compute	1 TB	1
GPU	1.6 TB	1
Large Mem	1.B TB	2

FIO Test on local NVMe: 100,000 640K files, sequential write with 4k blocksize.

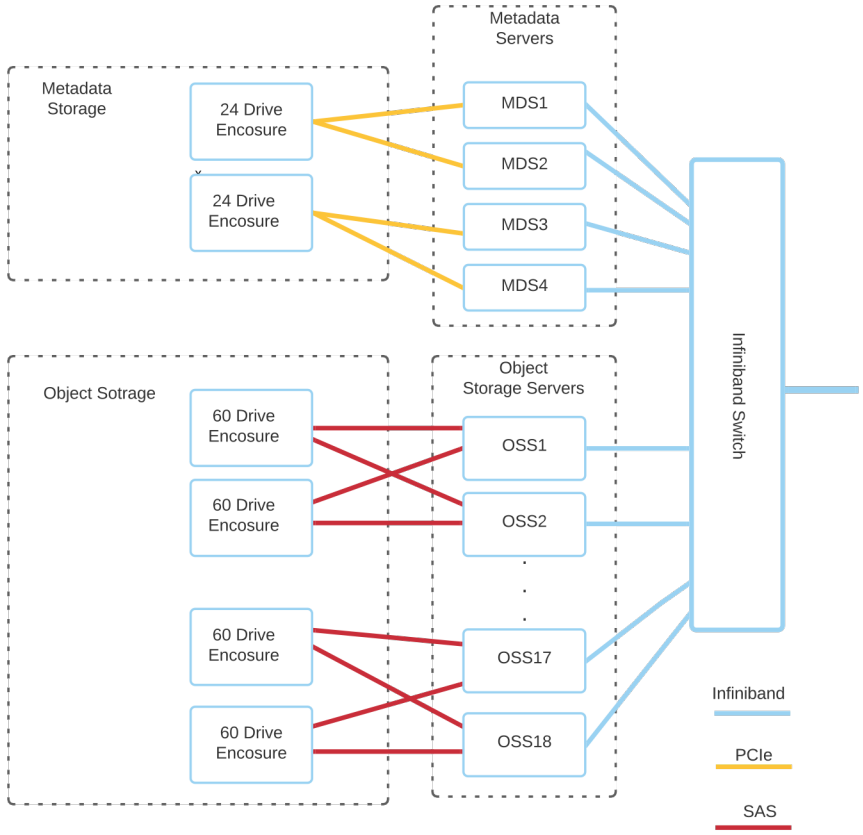
FIO Test	min	max	avg	stdev
slat (usec)	2	6861	4.31	5.10
clat (usec):	7	123512	185.58	266.65
Latency (usec)	10	123516	189.96	266.72
write: IOPS=167k		BW=654MiB/s (686MB/s)		

High-performance, Lustre filesystem architecture

- Hardware Provided by Aeon Computing
- Native IB throughout
- 12 Peta Bytes of RAW capacity, approx. 11 PB formatted
- File Capacity of approx. 3 billion files.
- 140 GB/s Filesystem Bandwidth
- 200K IOPS
- DNE Phase II, Striped across the four MDTs
- Data on MDT (DoM) for small file performance

4 Lustre MDS	
Processor	2 X AMD Epyc 7302 (16 Cores)
Memory	512 GB (16 X 32GB DDR4 3200)
MDT Drives	24 X 3.8 TB NVMe per pair
Interconnect	InfiniBand HDR 200
System Drives	2 X 240 GB Intel SSDs

18 Lustre OSS	
Processor	1 AMD Epyc 7402 (24 Cores)
Memory	512 GB (16 X 32 GB DDR4 3200)
JBODS	2 Cross Connected 60 Bay JBODS
OSS Drives	120 X 14 TB 7200 SAS Drives
Interconnect	InfiniBand HDR 200
System Drives	2 X 240 GB Intel SSDs



Maintaining the flexibility and adaptability of proven open-source solutions while benefiting from the advantages of commercial software

Function	Name	Version
Cluster Management	Bright Cluster Manager (CM)	9.0
Primary OS	CentOS	8.1
HPC Scheduler	Slurm	20.2
Non HPC Scheduler	Kubernetes	1.18
IB Subnet Manager	Mellanox OpenSM	5.5.1
OFED	Mellanox OFED	4.7

Flexible batch scheduling policies for high system utilization and reduced charging for users

- Full node allocation for jobs that need whole node resources
- Compute nodes have 128 cores, which is more than many users require for a job
- Shared node partitions allow for flexible and efficient use of system resource
- Users pay only for the portion of the system they need
- Users only changed for the maximum portion of trackable resources they used on the system.
  - Previously (on Comet and Gordon) the only trackable resource was CPU usage and resources were allocated based on proportional CPU usage
  - Expanse has multiple resources that can be tracked to determine job cost based on maximum billing weight (e.g., cores, memory, wallclock, charge factor)

Expanse partitions support a wide range of job types

Partition Name	*Comments*
compute	Used for **exclusive access** to regular compute nodes (Max 32 nodes/job)
shared	Single-node jobs using **fewer than 128 cores** (Maximum 4096 jobs)
gpu	Used for **exclusive access** to the GPU nodes
gpu-shared	Single-node job using fewer than 4 GPUs
large-shared	Single-node jobs using large memory **up to 2 TB (minimum memory required 256G)**
debug	**Priority access** to compute nodes set aside for testing of jobs with **short walltime and limited resources**
gpu-debug	**Priority access** to **GPU** nodes set aside for testing of jobs with **short walltime and limited resources**
preempt and gpu-preempt	Discounted (.8) jobs to run on free nodes that can be pre-empted by jobs submitted to any other queue