# Expanse: Computing without Boundaries

## Architecture, Deployment, and Early Operations Experiences of a Supercomputer Designed for the Rapid Evolution in Science and Engineering

Ilkay Altintas, Haisong Cai, Trevor Cooper, Christopher Irving, Thomas Hutton, Marty Kandes, Amitava Majumdar, Dmitry Mishin, Ismael Perez, Wayne Pfeiffer, Manu Shantharam, Robert S. Sinkovits, Subhashini Sivagnanam, Shawn Strande, Mahidhar Tatineni, Mary Thomas, Nicole Wolter, Michael Norman
University of California, San Diego, La Jolla, CA

## ABSTRACT

We describe the design motivation, architecture, deployment, and early operations of Expanse, a 5 Petaflop, heterogenous HPC system that entered production as an NSF-funded resource in December 2020 and will be operated on behalf of the national community for five years. Expanse will serve a broad range of computational science and engineering through a combination of standard batch-oriented services, and by extending the system to the broader CI ecosystem through science gateways, public cloud integration, support for high throughput computing, and composable systems. Expanse was procured, deployed, and put into production entirely during the COVID-19 pandemic, adhering to stringent public health guidelines throughout. Nevertheless, the planned production date of October 1, 2020 slipped by only two months, thanks to thorough planning, a dedicated team of technical and administrative experts, collaborative vendor partnerships, and a commitment to getting an important national computing resource to the community at a time of great need.

## CCS CONCEPTS

• **Computer Systems Organization -> Architectures -> Distributed Architectures**;

## KEYWORDS

High performance computing, high throughput computing, science gateways, scientific applications, user support

**ACM Reference Format:**
Ilkay Altintas, Haisong Cai, Trevor Cooper, Christopher Irving, Thomas Hutton, Marty Kandes, Amitava Majumdar, Dmitry Mishin, Ismael Perez, Wayne Pfeiffer, Manu Shantharam, Robert S. Sinkovits, Subhashini Sivagnanam, Shawn Strande, Mahidhar Tatineni, Mary Thomas, Nicole Wolter, Michael Norman. 2021. Expanse: Computing without Boundaries: Architecture, Deployment, and Early Operations Experiences of a Supercomputer Designed for the Rapid Evolution in Science and Engineering. In *Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465588



*PEARC '21, July 18–22, 2021, Boston, MA, USA*

ACM ISBN 978-1-4503-8292-2/21/07.
https://doi.org/10.1145/3437359.3465588

## 1 SYSTEM ARCHITECTURE

High-performance computing system architectures, software, and expertise are evolving to support the growing diversification in computing that is being driven by rapidly evolving science and engineering research. Now more than ever, supercomputers must be part of a more integrated, national cyberinfrastructure that comprises distributed computing and data resources, scientific instruments, R&E networks, and expertise. Accordingly, systems and application software, and user support must also evolve to support this expanding ecosystem [1, 2].

Developed in response to NSF Solicitation 19-534, *Expanse* is an evolutionary system, designed in large part from lessons learned from the operation of *Comet* [3, 4], a supercomputer that has been operated by SDSC for the last six years. Like *Comet*, *Expanse* was designed to support the "longtail" of computing, which we define as the broad spectrum of computational science and engineering research that is carried out at modest scale, but with increasingly diverse system and application software. A summary of the major subsystems is given in Table 1

Notable features of this design include: first large-scale NSF system to feature AMD EPYC processors; 13 identical Scalable Units, each with 56 compute and 4 GPU nodes; rich storage environment; full-bisection, low-latency Mellanox HDR-100 interconnect at the rack level, accessing 7,168 EPYC cores, and 16 V100 GPUs; support for Slurm and Kubernetes; integration with the Open Science Grid; scheduler-based integration with public cloud; and support for composable systems. The system includes a Lustre file system for compute and a Ceph file system, built from repurposed hardware from the *Comet* system that will be retired in July 2021, and will support the composable systems and cloud integration capability, and limited second copy data.

## 2 ACQUISTION AND DEPLOYMENT

### 2.1 Project Changes

Following a thorough assessment of the original design in light of vendor developments following the award and the potential impact on the planned deployment, a decision was made to execute a Project Change Request (PCR) for the standard compute node processor (the *Expanse* Risk Mitigation Plan had a provision for

**Table 1: Expanse System Specifications**

| System Component | Configuration |
|---|---|
| *AMD EPYC (Rome) 7742 Compute Nodes* | |
| Node count: procs/node: cores/proc; total cores | 728:2:64:93,184 |
| Total DRAM: Total node NVMe | 182 TB: 728 TB |
| *NVIDIA V100 GPU Nodes* | |
| Node count: total GPUs | 52: 208 |
| Total GPU Memory | 6.6 TB |
| Total NVMe Storage | 83 TB |
| *Large Memory Nodes (EPYC 7742)* | |
| Number of nodes | 4 |
| Memory per node | 2 TB |
| *Storage* | |
| Lustre file system | 12 PB |
| Ceph file system | 7 PB (coming July 2021) |
| Home File system (NFS) | 1 PB |

this change.) Whereas the proposed system was based on an Intel processor, the deployed system is based on AMD's EPYC processor. This change resulted in doubling of the original core count from 46,592 to 93,184. Extensive benchmarking was performed in advance of the PCR to demonstrate that the change would result in a system that delivered as good or better performance than the proposed system. The results showed that the change would deliver both improved per-core performance over *Comet* on real applications, and a substantial increase in overall system throughput. The other notable PCR was the addition of direct-to-chip cooling for the standard compute nodes. Both changes were accommodated within the original acquisition budget.

Acquisition, deployment, and transition to production operations were carried out under COVID-19 restrictions. We estimate an approximate 2-month delay due to COVID-19 related delays, including data center access, and supply chain.

## 2.2 Installation and Acceptance Testing

System testing comprised several distinct phases, including: early access for benchmarking and software; early delivery of a single compute node in advance of the full system for system and application stack development; system burn-in and testing at vendor facilities; onsite vendor regression testing following initial deployment; SDSC verification of bills of materials; unit level testing; full systems benchmarking using micro benchmarks and applications; reliability testing; and 30-day Early User Program. Application testing was defined by SDSC and approved as part of the Cooperative Agreement with NSF.

SDSC staff received remote access to Dell and AMD clusters with EPYC processors and used the time to develop application build recipes with various compilers (GNU, Intel, and AOCC) and find optimal run configurations. We used the NPS4 configuration - which is 4 NUMA domains per socket. This is the AMD-recommended configuration for HPC workloads which are typically NUMA aware. NPS1 is suggested for codes that are not NUMA aware or cannot handle NUMA complexity, but we did not explore this.

Synthetic benchmarks included HPL, STREAM, OSU Micro Benchmarks (OMB), IOR, mdtest, and Tartan. Using HPL v2.3, an average floating-point performance of 3,711 GFLOPS (80% of theoretical peak) was measured for *Expanse's* standard AMD compute nodes. An average memory bandwidth for STREAM's copy function of 346,072 MB/s was measured for *Expanse's* standard AMD compute nodes (85% of theoretical peak). OMB results show the average point-to-point in-rack latency between the AMD compute nodes was 1.184 $\mu$s and the point-to-point in-rack bandwidth was 12,335 MB/s. IOR results showed an average read bandwidth of 2,694 MiB/s on the node-local NVMe drives. The Lustre filesystem performed at an average bandwidth of 143,098 MiB/s on IOR tests and achieved a maximum of 211,369 IOPs on file reads with the mdtest benchmark.

The CPU applications benchmarks included GROMACS, NAMD, NEURON, OpenFOAM, Quantum Espresso, WRF, and RAxML. The RAxML runs were on single nodes with core counts ranging from 10 to 40. WRF, NEURON, OpenFOAM, and Quantum Espresso were run with core counts ranging from 96 to 768. NAMD, GROMACS were run with core counts ranging from 96 to 1,536 cores. Speedups relative to *Comet* were calculated on a per-core basis and ranged from 0.97X to 1.9X with an average speed up of 1.41X combining all cases.

The GPU application benchmarks included: AMBER - run with a single GPU and in throughput mode using all 4 GPUs; BEAST, GROMACS, NAMD, MXNet, PyTorch, and TensorFlow - all run with 1, 2, and 4 GPUs on a single node. Speedups relative to the *Comet* P100 GPUs ranged from 1.29X to 1.89X with an average speed up of 1.63X combining all cases.

Compute node reliability was tested by running five of the benchmark application performance tests (HPL, RAxML, NEURON, Quantum Espresso, and WRF) described above over a period of three days. A total of 47,319 jobs were run in 3 days and the overall success rate for all of the benchmarks jobs run was 99.84%. The reliability of *Expanse's* GPU nodes was tested by running the AMBER and GROMACS benchmark application performance tests over a period of three days. A total of 7,214 jobs were run with only one job failure detected during the 3-day period. A few of the GROMACS run were slightly slower than the norm and two GPU nodes were fixed to address the issue. Full-scale HPL was run to further identify hardware issues.

## 2.3 Early User Program

Following the activities described above, which addressed issues of performance and reliability, and prior to full transition to production operations, *Expanse* underwent a 30-day Early User Program (EUP) to ensure that the system was highly usable and stable under a real world mix of users and applications. Forty-one projects were allocated time on the system, thirty two of which made use of their allocations during the program.

During the EUP, the users ran 87k CPU and 5.7k GPU jobs, consuming a total of 6M CPU core-hours and 20k GPU hours, with an overall job completion rate of 99%. Users were surveyed as part of the program to assess their experience and determine if the system was ready for transition. Overwhelmingly, users expressed high satisfaction with the system, with many noting that it provided

significant performance improvements over either *Comet* or other systems they were using. A few notable user comments: ***"Simple** benchmark tests indicate that running our model code on Expanse provides a speedup of about 1.6 over Comet"; **"These** rules call for RAxML analyses of the most common, small data sets to run on 40 cores of Expanse where they are 1.6x to 2.0x faster than on 48 cores of Comet"; **"We** were able to run the GPUs based simulations smoothly without any issues. The user document was especially helpful. We also appreciate the quick response of the EXPANSE support team."*

The EUP was considered very successful with the system performing superbly and without any major system or reliability issues that would keep it from entering production operations.

## 3 PREPARING THE USER COMMUNITY

As the first NSF system to feature the AMD EPYC processor significant resources went into easing user transition to the new architecture. Given the high core count and unique design of the EPYC processor, and our initial benchmarking and performance analysis, we expected that some applications might encounter performance bottlenecks. Accordingly, we developed targeted training to guide users in application builds and proper job placement, and to facilitate transition from *Comet* to *Expanse*.

Since Spring, 2020, we have hosted 15 *Expanse* related training activities, reaching over 2,100 users, including: 9 webinars; 4 user forum meetings; a 12-week HPC User Training class customized to train the Student Cluster Competition team; and an annual Summer institute. Future training plans include: a second annual SDSC NVIDIA GPU Hackathon; the first annual Cyberinfrastructure-Enabled Machine Learning (CIML) Summer Institute [5]; the annual SDSC High Performance Computing and Data Science Summer Institute, and the 4th annual HPC User Training class. More details about our future training activities can be found on the *Expanse* web site [6].

We also worked closely with AMD to conceive and launch the AMD HPC User Forum. The Forum focuses on connecting HPC administration and support communities with AMD experts. This provides members with the unique opportunity to provide feedback to AMD and share experiences, expertise, code, and knowledge among the user community. The Forum membership is approaching 100 with representatives from over 30 national, international, and commercial organizations.

## 4 INNOVATIVE FEATURES

### 4.1 Public Cloud Integration

Under funding via the Internet2 E-CAS Phase 1 project [7] SDSC developed tools enabling direct scheduler integration of the CIPRES Gateway with public cloud resources from *Comet* (Figure 1). This foundation was migrated to *Expanse* as part of early system acceptance and updated to support a second project, the Neuroscience Gateway, during the first quarter of production operation. As part of this early effort the SDSC tooling and resource provisioning process was updated to allow cross-account sharing of critical network resources managed by SDSC simplifying the configuration of cloud-resources and reducing cost for the end user. Through funding in the *Expanse* award, additional projects will be supported that will

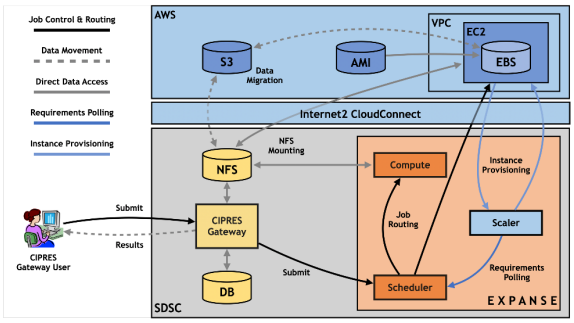


**Figure 1: Expanse Cloud Integration Architecture**

expand the public cloud resource integration options, with specific effort made to utilize Bright Cluster Manager [8] hybrid cloud capabilities. Work is underway to integrate *Expanse* with the NSF-funded CloudBank Project [9]. Specifically, *Expanse* will leverage CloudBank's cloud accounting and monitoring capabilities.

In summary, this integration includes dedicated high-speed networking to the public cloud, infrastructure-as-code to deploy public cloud resources, custom software monitoring on-premises batch scheduler resource demand and gateway modifications to support submission targeted at cloud resources all of which are described in detail in the E-CAS final report [10].

### 4.2 Composable Systems

A key innovation of *Expanse* is its support for composable systems, allowing for the creation of a virtual cluster of resources for a specific project. Composable systems workloads use Kubernetes through the Bright Cluster Manager (Figure 2) Bright Cluster Manager includes capabilities for configuring, managing and deploying Slurm and Kubernetes-based clusters from a single interface and includes features to dynamically re-provision nodes from the same hardware pool to run either under Slurm or Kubernetes based on resource needs. Bright Cluster Manager automates many of the steps that would otherwise require manual administrative intervention using Kubernetes commands. The cm-scale tool within Bright Cluster Manager is a meta-scheduler that can monitor workloads and dynamically repurpose a set of compute nodes for Kubernetes. The task of provisioning a new cluster, or potentially joining an existing one is simplified, creating additional possibilities for users to automate their workflows and leverage the existing containerized software repositories. The "EXPANSE Cluster" box in Figure 2 further illustrates the details of the cm-scale resizing of nodes and the Ceph based storage cloud integration. A multi-user JupyterLab [11] hub instance running as an interactive application development interface enables deployment of applications on top of the Kubernetes microservice architecture.

We have done an initial implementation of this for a wildfire modeling and prediction application. As part of the *Expanse* Operations funding, staff resources will be allocated to a small number of projects selected through NSF allocations processes.

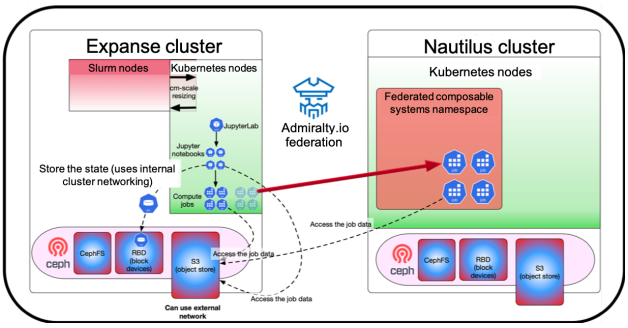


**Figure 2: Expanse Composable Systems Architecture**

## 5 PRODUCTION OPERATIONS

Following a successful NSF review in November, *Expanse* entered production operations on December 7, 2020, approximately 2 months later than originally planned. Given the COVID-19 constraints, we considered this an excellent result. *Expanse* was available in the August 2020 XSEDE allocations cycle for projects starting on October 1 and has been allocated in two allocation cycles since. Demand is already high, with requests and recommended awards exceeding available resources. Thankfully, there are new NSF-funded systems coming that will go far to addressing user demand.

*Expanse* has several important support and operational features that are geared toward supporting as many research projects as possible, including: shared node computing for CPU and GPU nodes; maximum job size of 4,096 cores; and per project allocation limits for a single PI, with exceptions for science gateways; direct integration with the Open Science Grid; and use of the NSF-funded Open OnDemand User portal [12–14].

Since the start of production operations, *Expanse* has delivered over 143M core-hours, and 262k GPU hours on behalf of 346 allocated projects, including several science gateways. Compute and GPU utilization is increasing rapidly, and we expect full utilization in the coming months as *Expanse* goes through a full year of allocation cycles. The system continues to be very reliable, with availability well above the required 95%. Hardware failure has been minimal.

## 6 CONCLUSIONS AND FUTURE WORK

*Expanse* was designed to support the growing diversity in computational and data-driven research and engineering which requires a mix of compute and data elements as well as distributed resources. Following a rigorous deployment and testing process, *Expanse* entered production as a resource for the national community on December 7, 2020, where it will serve for 5 years. Over the coming years, we expect to support important research and discoveries by pushing the boundaries of *Expanse* through integration with public cloud, remote instruments, R&E networks, innovative software, and a technical team of experts who will support the community through the full complement of support and training needed to advance their research.

## ACKNOWLEDGMENTS

This work was supported by the NSF under Award #1928224. Additional support was provided by NSF Award #1548562. The authors are grateful for the support of their partners at Dell, Aeon, AMD, NVIDIA, Intel and Mellanox.

## REFERENCES

[1] *Rethinking NSF's Computational Ecosystem for 21st Century Science and Engineering. NSF Workshop, May 30-31, 2018.*; Available from: https://uiowa.edu/nsfcyberinfrastructure/sites/uiowa.edu.nsfcyberinfrastructure/files/wysiwyg_uploads/report.pdf.

[2] Transforming Science Through Cyberinfrastructure. NSF's Blueprint for National Cyberinfrastructure Ecosystem in for Science and Engineering in the 21st Century. April 2019. Available from: https://www.nsf.gov/cise/oac/vision/blueprint-2019/Overview-Computational.pdf

[3] Strande, S.M., H. Cai, T. Cooper, K. Flammer, C. Irving, G. von Laszewski, A. Majumdar, D. Mishin, P. Papadopoulos, W. Pfeiffer, R.S. Sinkovits, M. Tatineni, R. Wagner, F. Wang, N. Wilkins-Diehr, N. Wolter, and M.L. Norman. *Comet: Tales from the Long Tail: Two Years In and 10,000 Users Later.* in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact.* 2017. ACM.

[4] Moore, R.L., C. Baru, D. Baxter, G.C. Fox, A. Majumdar, P. Papadopoulos, W. Pfeiffer, R.S. Sinkovits, S. Strande, and M. Tatineni. *Gateways to Discovery: Cyberinfrastructure for the Long Tail of Science.* in *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment.* 2014. ACM.

[5] NSF Award OAC 2017767 CyberTraining: Implementation: Small: Developing a Best Practices Training Program in Cyberinfrastructure-Enabled Machine Learning Research

[6] Expanse website. Available at: http://expanse.sdsc.edu

[7] Exploring Clouds for Acceleration of Science (E-CAS). Available at: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1904444

[8] Bright Cluster Manager. Available at: https://www.brightcomputing.com/brightclustermanager

[9] CloudBank: Managed Services to Simplify Cloud Access for Computer Science Research and Education. Available at: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1925001

[10] Phase 1 Final Report: Cloud Bursting to AWS from the CIPRES Science Gateway. Available at: https://drive.google.com/file/d/1_BM4LowbeRhVU6dslAcxrl5fC2aZn7H0/view

[11] https://github.com/sdsc-hpc-training-org/reverse-proxy

[12] Expanse User Portal. Available at: https://portal.expanse.sdsc.edu.

[13] Hudak *et al.*, (2018). Open OnDemand: A web-based client portal for HPC centers. Journal of Open Source Software, 3(25), 622, https://doi.org/10.21105/joss.00622

[14] NSF award # 1534949. SI2-SSE: Open OnDemand: Transforming Computational Science through Omnidisciplinary Software Cyberinfrastructure