

# Updated Acceptance Test Results for the Jetstream Production Environment

*David Y. Hancock  
George Turner  
John Michael Lowe  
Michael Packard  
Craig A. Stewart*

Indiana University

PTI Technical Report PTI-TR16-005

August 20, 2016

## Citation:

Hancock, D.Y., Turner, G., Lowe, J.M., Packard, M., Stewart, C.A. (2016). Updated Acceptance Test Results for the Jetstream Production Environment (PTI Technical Report PTI-TR16-005). Bloomington, IN: Indiana University. Retrieved from <http://hdl.handle.net/2022/20958>



**PERVASIVE TECHNOLOGY  
INSTITUTE**

INDIANA UNIVERSITY



**RESEARCH  
TECHNOLOGIES**

INDIANA UNIVERSITY

University Information Technology Services  
Pervasive Technology Institute

## Table of Contents

1. Executive Summary .....	1
2. Introduction.....	2
3. System Description.....	3
3.1. Hardware.....	3
3.2. Software .....	5
4. Acceptance Test Criteria .....	6
4.1. Basic Hardware performance.....	6
4.2. Provide "self-serve" academic cloud services.....	7
4.3. Host Persistent Science Gateways (production system only).....	7
4.4. Data movement, storage and dissemination (production system only).....	8
4.5. Provide virtual Linux desktop services to tablet devices .....	8
5. Acceptance Test Methodology and Results .....	8
5.1. Basic Hardware Performance .....	8
5.2. Integrated Cloud Operations .....	14
6. Conclusion.....	15

# 1. Executive Summary

The Jetstream-IU and Jetstream-TACC Dell PowerEdge (PE) systems were purchased by Indiana University for eventual delivery to the Indiana University – Bloomington (IUB) and Texas Advanced Computing Center (TACC) as part of the Jetstream project (National Science Foundation Award ACI-1445604: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment, Craig A. Stewart, IU, Principal Investigator).

The Jetstream project underwent a formal National Science Foundation (NSF) site visit and review on May 3-4, 2016; subsequently, on May 27, 2016, the cooperative agreement between IU and the NSF was amended to increase spending authority for the project and enter into the management and operations phase. The technical data contained in this report represent the detailed information presented to the NSF review panel in May 2016 and also serve to describe the additional testing discussed and recommended by the review panel. A formal response was requested on the following technical evaluations identified in the course of the site visit and review:

- previously performed technical evaluations applied to storage components contributing to the TACC hosted subsystem be re-applied with a fully configured Ceph environment as to enable more comprehensive characterization of Jetstream's capabilities, including specifically as maybe advanced through further network performance verification applying a tool such as nuttcp or other equivalent tool;
- additional technical evaluations applying IOR in single-node mode or other equivalent technical evaluation (for example such as sysbench or bonnie++ executed in parallel) be performed at a higher scale to stress the underlying infrastructure for example as previously performed to assess IOPS on Amazon Web Service; and
- additional technical evaluations of applying a random read/write benchmark code, as well as further load testing be performed applying diverse evaluative approaches including for example to apply file system and storage evaluative criteria.

The Jetstream team has performed extensive additional testing both in preparation for a paper submission and presentation at XSEDE16 as well as during normal user operations to further tune and demonstrate the capabilities of the Jetstream system. Section 5.1.1.4 **Error! Reference source not found.** details the additional network testing performed since May 2016. Section 5.1.2.1 details the additional storage testing with the versions of Ceph listed in section 3.2.6. Changes were made to Jetstream-TACC to improve network performance as the results of these tests. Ceph updates have been also been applied making Jetstream as a whole more consistent. We believe these changes and tests address the technical points identified above. Additional tuning and topology changes are still possible for the Jetstream environment and we will provide an update on related activity during our next review. It is notable that the additional effort put into the XSEDE16 paper as a result of discussions with the NSF was productive; the paper "Jetstream early implementation and experiences" won the best technical paper award at XSEDE16. It is clear from the NSF review, and supported by the data within this report, that the system as proposed by Indiana University, and awarded by the NSF, is the system that is available to researchers throughout the United States.

## 2. Introduction

This Jetstream system was ordered on 07/29/2015 via purchase order numbers 1681608 and 1681609 respectively. The respective clusters arrived at TACC's data center on 10/16/2015 and the IUB Data Center on 10/19/2015. Each of these production clusters have the following basic characteristics:

- The hardware infrastructure is based upon Dell PowerEdge servers with a 10/40 Gbps Fat-Tree Ethernet fabric<sup>1</sup>.
  - Each node contains two Intel E5-2680v3 (12-core) 2.5 GHz processors for a total of 24 processing cores resulting in a peak performance of 806.4 GFLOPS and 64 GB RAM for the management and storage servers and 128 GB RAM for the compute servers.
  - The compute portion system includes 320 Dell M630 blades with a total of 640 CPUs, 15,360 processor cores, 258 TFLOPS peak processing capability, and 40 TB RAM.
  - The management portion consists of 7 Dell R630 servers, with a total of 14 CPUs, 168 processing cores, 448 GB RAM, 5.6TB local storage, and a peak processing capability of 5.6TF
  - The storage portion consists of 20 Dell R730 servers, with a total of 40 CPUs, 960 processing cores, 1.2TB RAM, 16 TB local storage, 960 TB of block/object storage and a peak processing capability of 16.1 TFLOPS.
- The cloud infrastructure is based upon OpenStack with its ability to deliver virtualized compute capacity<sup>2</sup>.

The acceptance timeline is highlighted in Table 1 below with additional information regarding the timeline immediately following. A detailed system description follows in Section 3, after which are details on the performance targets, the methods used to perform the acceptance tests, and the achieved performance.

**Table 1. Acceptance timeline for Jetstream**

Acceptance Task	Jetstream-IU	Jetstream-TACC
Purchase Order	07/29/2015	07/29/2015
System arrival	10/19/2015, 10/23/2015	10/16/2015
System boot	11/11/2015	11/03/2015
Functionality pass	01/14/2016	11/05/2015
Performance pass	11/30/2015	02/24/2016
14-day stability pass	04/21/2016	04/21/2016
NSF Review	05/03-04/2016	
Amended Agreement	05/27/2016	

The majority of the Jetstream-IU Production cluster was delivered to the Indiana University – Bloomington's Data Center on 10/19/2015. One compute rack was held at the Dell Merge Center due to a single top of rack S6000 switch that failed to pass Dell's acceptance criteria as well as

---

<sup>1</sup> <http://www.dell.com/learn/us/en/04/campaigns/poweredge-13g-server>

<sup>2</sup> <https://www.openstack.org/>

four blades with non-functioning Ethernet network interface controllers (NICs). Additional issues were encountered plumbing the water cooling doors resulting in the inability to power on the Jetstream-IU system in its entirety until 01/14/2016. Otherwise, there were no problems encountered installing and booting the five (5) R630 management servers, 20 R730 storage servers, and the 320 M630 compute blades. Two (2) additional R630 management servers were removed from the Jetstream-AZ cluster and installed in the Jetstream-IU cluster before the Jetstream-AZ was packed up for shipment to Arizona where Jetstream-AZ will serve as the test and development resource for Jetstream.

The Jetstream-TACC Production cluster was delivered to the TACC data center. No problems were encountered installing and booting the seven (7) R630 management servers, 20 R730 storage servers, and the 320 M630 compute blades.

The Jetstream system as an integrated environment was evaluated by the NSF on May 3-4, 2016 and the project's cooperative agreement with the NSF was amended as of May 27, 2016 indicating formal acceptance of the system as proposed and awarded.

## 3. System Description

### 3.1. Hardware

#### 3.1.1. System topology

The Jetstream-IU and Jetstream-TACC production clusters consists of 7 PE R630 management servers, 20 (4) PE R730 storage servers and 320 PE M630 compute blades. The PE R630 management servers are configured with dual Intel 2.5 GHz, 120W, Xeon E5-2680v3 "Haswell" chips, 64 GB RAM, dual 400 GB SSD system devices and are wired directly into the Dell Force10 (F10) S6000 spine network switch.

The PE R730XD storage servers are configured with dual Intel E5-2680v3 "Haswell" chips, 64 GB DDR4 RAM, dual 200 GB SSD system devices, and 12 – 4 TB Near-Line Serial Attached SCSI (NL-SAS) storage disks and are wired directly into the Dell Force10 S6000 spine network switch.

The PE M630 blade servers are installed in a PE M1000 blade enclosure and are configured with dual Intel E5-2680v3 "Haswell" chips, 128 GB RAM, and dual 1 TB NL-SAS disk drives and are wired into the Dell PE MXL1000 chassis switches which then uplink into the Dell Force10 S6000 spine switch for a two-to-one oversubscribed Fat-Tree topology.

#### 3.1.2. Memory boards, sections, and/or banks

Each M630, R630, and R730 has 24 DDR4 DIMM slots running at 2133MT/s.

#### 3.1.3. Memory size

Each M630 compute blade has eight (8) 16 GB RDIMM running at 2133MT/s for a total of 128 GB.

Each R630, R730 management, storage server respectively has eight (8) 8 GB RDIMM running at 2133MT/s for a total of 64 GB.

#### 3.1.4. CPU manufacturer, model, and speed

Each M630, R630, and R730 are populated with dual Intel Xeon E5-2680v3, 12-Core 2.5 GHz, 2133MHz bus with 30 MB L3 cache, 12 x 256KB L2 cache.

#### 3.1.5. Speed of the memory and memory bus (if applicable)

Each of the M630, R630, and R730 utilize DDR4 memory running at 2133MT/s.

**3.1.6. I/O boards and bus interfaces**

M630: internal RAID controller, Intel QPI @ 9.6 GT/s.

R630: two (2) PCIe Gen3x16 slots, one (1) PCIe Gen3 x8 slot, dedicated RAID card; Intel QPI @ 9.6 GT/s.

R730: two (2) PCIe Gen3x16 slots, four (4) PCIe Gen3 x8 slot, dedicated RAID card; Intel QPI @ 9.6 GT/s.

**3.1.7. HBAs, Network Interface Cards and TCO Offload Engine (TOE) cards including firmware**

None.

**3.1.8. Network adapters, including firmware**

M630: Intel X710 dual port, 10 Gbps, version 1.3.38, firmware-version: 4.25 0x8000143f 0.0.0.

R630, R730: Intel X710 quad port, 10 Gbps, version 1.3.38, firmware-version: 4.25 0x8000143f 0.0.0.

**3.1.9. All communications hardware, including private channels**

Dual Dell Networking MXL 10/40 Gbps Ethernet blade switches (leaf).

Dell Force10 S6000 10/40 Gbps Ethernet top of rack switch and spine.

Sixteen M630 blades connect via bonded 2 x 10 Gbps links to the two Dell MXL switches in each blade chassis (with virtual link trunking enabled) which each uplink to the Top-of-Rack (ToR) Dell Force10 (F10) S6000 switches via two bonded 40 Gbps links resulting in 2:1 over-subscription to the blades. Each ToR S6000 then connects via 2 x 40 Gbps links into each of the two Spine S6000 switches. The two spine F10 S6000 are cross linked at 3x40 Gbps for 120 Gbps aggregate. Each F10 S6000 spine switch is then uplinked to the data center's Science DMZ via 2 x 40 Gbps uplinks for a total of 4 x 40 Gbps. The M630 management and M730 storage nodes link into the F10 S6000 spine switch via dual bonded 10 Gbps links.

Dell N3048 1 Gbps management Ethernet switch provide out-of-band management control of the overall system.

**3.1.10. RAID hardware including disks, cache, firmware, channels, GBICS and interfaces**

M630: PERC H330 RAID controller; dual 1 TB 7.2K RPM NL-SAS 6 Gbps 2.5in Hot-plug system devices.

R630: PERC H330 integrated RAID controller; dual 400 GB Solid State Drive (SSD) SATA Mix Use MLC 6 Gbps 2.5in Hot-plug system devices.

R730: PERC H730P integrated RAID controller, 2 GB cache; dual 200 GB Solid State Drive SATA Mix Use MLC 6 Gbps 2.5in Flex Bay system devices; twelve 4 TB 7.2K RPM NL-SAS 6 Gbps 3.5in Hot-plug Hard storage devices.

**3.1.11. Fibre Channel switches, if used**

None

**3.1.12. Any other hardware used as part of the benchmark configuration**

Benchmarks were run from an NFS mounted file system exported from the respective cluster's management server.

## 3.2. Software

### 3.2.1. Operating system, including all tunable parameters and their values

M630: CentOS 7.1.1503 w/ kernel 3.10.0-229.el7.x86\_64, stock.

VM: CentOS 7.2.1511 w/ kernel 3.10.0-229.el7.x86\_64, stock.

MXL: 9.9.9.0.0

S6000-ON: 9.8.0.0p9

### 3.2.2. BIOS tunable parameters and their values

Firmware

M630: 2.10.10.10, build 49, 04/06/2015 09:05:28.

R630: 2.02.01.01, build 92, 09/15/2014 09:45:31.

R730: 2.02.01.01, build 92, 09/15/2014 09:45:31.

BIOS

M630: 1.1.10, default performance values.

R630: 1.0.4, default performance values.

R730: 1.2.10, default performance values.

### 3.2.3. Network drivers

Intel i40e, version 1.3.47, firmware 4.41 0x8000186a 16.5.20.

### 3.2.4. Network stacks, including TOEs

Standard Linux network stack.

### 3.2.5. I/O drivers

N/A

### 3.2.6. File system software and/or volume manager

XFS for local file systems.

Ceph 0.94.5 (Hammer) for initial block and object storage testing at IU.

Ceph 9.21 (Infernalis) for initial block and object storage testing at TACC.

Ceph 10.2.2 (Jewel) for subsequent testing at both IU and TACC.

### 3.2.7. Compiler and libraries, including I/O and MPI libraries

Intel compilers, version 15 update 3, Intel MPI version 5.0.3p-048

### 3.2.8. All patches and bug fixes

CentOS 7.2.1511 with patches up to date as of the Performance Pass date identified in tables above.

### 3.2.9. Any additional software used as part of the benchmark configuration

qemu-kvm-1.5.3-105.el7\_2.1

libvirt-daemon-kvm-1.2.17-13.el7\_2.2

OpenStack 2015.2.0 (Liberty)

## 4. Acceptance Test Criteria

The Project Execution Plan (PEP) between Indiana University and the National Science Foundation stipulate the acceptance criteria for Jetstream.

The purpose of the acceptance testing is to ensure that the system as implemented is the system described in the original proposal as modified by a scope of work change document. The following acceptance criteria demonstrate the functionality of Jetstream based exclusively on the terms of NSF Request for Proposals, the original proposal by IU and its partners, and the scope of work change document submitted to the NSF as a supplement to the original proposal. If completed successfully these tests will comprehensively demonstrate that the computational resource satisfies the capabilities of the Jetstream system that Indiana University and its subcontractors have been contracted to integrate and deliver.

IU and NSF retain the right by mutual agreement to change these tests should one or more prove not informative, or if the software underlying the tests proves itself to be faulty in terms of demonstrating the capabilities of Jetstream.

### 4.1. Basic Hardware performance

Jetstream is a first-of-a-kind acquisition and implementation for the NSF and for the NSF-funded national cyberinfrastructure. It is more a system implementation than a hardware implementation (as contrasted, say, to earlier systems such as Ranger, Kraken, or FutureGrid). However, it makes sense to have some basic hardware performance tests as the first step in the acceptance testing of Jetstream. These criteria are, in a sense, prerequisites for other tests that verify the functionality of the system. These tests are primarily performance tests – the doing of a specific activity.

#### 4.1.1. Single Node Performance

- High-Performance Linpack (HPL): Single node Linpack performance will achieve 80% of the peak floating-point performance for a problem size that uses at least half of the on-node memory. (Measurements will be rounded to nearest %).
- STREAM: Single node OpenMP threaded STREAM performance will be at least 65 GB/s (aggregate across the node). (Measurements will be rounded to nearest 1 GB/s).
- 10 Gigabit Ethernet Bandwidth: the 10GigE interface on each node will achieve at least 1 GB/s for large-message point-to-point transfers (Measurements will be rounded to the nearest 0.1 GB/s).

#### 4.1.2. File System and Storage Benchmarks

- The system will achieve a minimum of 200 MB/s data transfer rate for data reads and a minimum of 100 MB/s writes from within a virtual machine from/to the block storage. (Measurements will be rounded to the nearest MB/s).

#### 4.1.3. System Reliability Tests

System reliability will be tested by operating the system during the friendly user mode with uptime of at least 95% for a period of 14 days. Appendix 1 of the PEP describes the rationale for a 14-day reliability test.

Neither the solicitation nor our proposal included any terms regarding Mean Time Between Failures (MTBF), so MTBF is not included as part of the acceptance criteria. However, we can place a lower bound on MTBF from the system reliability metrics. 95% uptime implies that the system won't be down more than 36 hours per month.

## 4.2. Provide "self-serve" academic cloud services

The full text of the capability described in Section 2 of the PEP is *'Provide "self-serve" academic cloud services, enabling researchers or students to select a VM image from a published library, or alternatively to create or customize their own virtual environment for discipline- or task-specific personalized research computing. Authentication to this "self-serve" environment will be via Globus.'* Implicit in the sense of the words 'cloud services' is that the two production components of Jetstream function as parts of an integrated whole. There are both capability and capacity issues to providing a cloud environment.

Much of the description of Jetstream as a cloud resource describes capabilities so the tests of these aspects are 'capability' tests, and a first of a kind system the test will consist simply of demonstrating the following functions:

- An authorized and knowledgeable user will be able to authenticate to the Jetstream user interface (which uses Globus as the mechanism for verification of credentials).
- After so doing, an authorized and knowledgeable user will be able to launch a virtual machine from a menu of pre-packaged VMs on the production hardware located in Indiana or Texas.
- After so doing, an authorized and knowledgeable user will be able to quiesce a VM image running on production hardware in Indiana or Texas, move it from one production system to another, and reactivate said VM.
- An authorized and knowledgeable user can create and access persistent cloud storage on the Indiana or Texas production hardware
- An authorized and knowledgeable user can modify a preexisting VM image and manually store that VM image to one of the production locations within Jetstream.

There is a capacity (load) goal that can be derived from the proposal as well. Working backwards from the final budget and configuration and the statements in the original proposal limiting the amount of oversubscription that would be permitted on Jetstream, and VM configurations, we can create a metric of the minimum number of active VMs that Jetstream should support: 640. (This is based on 640 nodes in the system, with the largest VMs to be supported on Jetstream taking a full node) This leads to the following system capacity test:

- Jetstream will support a minimum of 640 VMs operating simultaneously

## 4.3. Host Persistent Science Gateways (production system only)

The full text of the capability described in Section 2 of the PEP is *'Host persistent Science Gateways. Jetstream will support persistent science gateways, including the capability of hosting persistent science gateways within a VM when the nature of the gateway is consistent with operation within a VM. Galaxy will be one of the initial science gateways supported.'*

This is a 'capability' and functionality test, and a first of a kind system the test will consist of:

- The Galaxy bioinformatics gateway is installed and will operate a demonstration workflow providing correct results, based on comparison with output results from a known reference installation. The job will no more than 25% slower than the time required to complete an analysis running on an equivalent system.
- One other exemplar science gateway that is known to function properly in other XSEDE-supported gateway hosting environments will function and remain reliable to within 2% of the overall system availability achieved during system reliability tests during a 14-day test period. (E.g. if the system turns out to be available with an uptime of 96%, the gateway used to test this criterion will be available 96% - 2% or 94%). The test period may be contemporaneous with the overall system test period or done at some other time. The critical metric here is that Gateway Availability track overall availability within a delta of 2% of total potential system uptime.

#### **4.4. Data movement, storage and dissemination (production system only)**

The full text of the capability described in Section 2 of the PEP is *'Data movement, storage and dissemination.*

- *'Jetstream will support data transfer with Globus Connect.*
- *Users will be able to store VMs in the Indiana University persistent digital repository, IUScholarWorks (scholarworks.iu.edu) and obtain a Digital Object Identifier (DOI) that is associated with the VM stored.'*

The performance characteristics of the storage system are verified through item 4.1.2. Globus Connect is a service offered by a partner organization that contains a set of performance characteristics that are well understood, and not affected by this solicitation. The first item above becomes a functionality test:

- An authorized and knowledgeable user can select a file to which they have rights on a system outside Jetstream, and move that file and save it on storage on Jetstream (with the condition that the file size is within the storage quota set for their use on Jetstream).
- An authorized and knowledgeable user can select a file to which they have rights on Jetstream, and move that file and save it on storage to a system on which that user has rights and which is accessible from open public networks (with the condition that the file size is within the storage quote set for their use on Jetstream).

The second feature described above is again a capability test, satisfied by the following:

- An authorized and knowledgeable user can successfully save a VM previously stored to disk storage on Jetstream into a format supported by DSspace, upload that file to IU Scholarworks.iu.edu, and using the existing online forms submit that document for publication via IUScholarworks. Subsequent to that, provided the relevant and required information has been provided by the user, the VM will appear in IUScholarworks and the user will receive a DOI identifier for that object. Note: This is a "human in the loop" process and may take days from upload and submission to publication and receipt of DOI. Email transactions may be required beyond the initial submission.

#### **4.5. Provide virtual Linux desktop services to tablet devices**

The full text of the capability described in Section 2 of the PEP is *'Provide virtual Linux desktop services delivered from Jetstream to tablet devices. This service is aimed to increase access to Jetstream for users at institutions with limited resources including small schools, schools in EPSCoR states, and Minority Serving Institutions.'*

This test is a functionality test, with some time constraints. This feature will be satisfied by the following:

- An authorized and knowledgeable user can access Jetstream from a tablet device, and load a virtual Linux desktop configured in a way that allows the user to access Jetstream services.

### **5. Acceptance Test Methodology and Results**

#### **5.1. Basic Hardware Performance**

##### **5.1.1. Single Node Performance Tests**

Single node performance benchmarks were run on all M630 compute servers on both the Jetstream-Indiana and Jetstream-TACC clusters. The Jetstream-Arizona system was previously accepted.

#### 5.1.1.1. *HPL*

The theoretical peak performance for the Dell M630 server is 806.4 GFLOPS. For a node to pass acceptance, it must achieve 80% of this value or 645.1GFLOPS on the HPL. Measurements will be rounded to the nearest 1%.

HPL was run as part of the HPCC benchmark suite. No modifications to the source code were made. It was compiled with the Intel compiler version 15.0.3 with options “-O3 -DRA\_SANDIA\_OPT2 -mP2OPT\_hlo\_loop\_intrinsic=F.”

The performance target as outlined in the PEP was achieved. The average performance across all tested servers was 697 GFLOPS (86% of theoretical peak performance) for Jetstream-IU and 701 (87%) GFLOPS for Jetstream-TACC.

#### 5.1.1.2. *STREAM*

The memory performance target for an M630 node is 65 GB/s rounded to the nearest 1 GB/s.

STREAM was run as a separate benchmark with no modifications to the source code. It was compiled with the Intel compiler version 15.0.3 with options “-O3 -xCORE-AVX2 -openmp.”

The performance target as outlined in the PEP was achieved for STREAM. The average STREAM Triad performance across all tested servers was 100.5 GB/s for Jetstream-IU and 113.1 for Jetstream-TACC.

#### 5.1.1.3. *Ethernet bandwidth*

Each node will need to demonstrate 1 GB/s rounded to the nearest 0.1 GB/sec across its 10 Gbps interfaces.

Iperf<sup>3</sup>, with default settings, was used to measure Ethernet bandwidth and was executed between the management node and all M630 compute, M730 storage, and R630 management servers.

For the Ethernet Bandwidth benchmark, the performance target as outlined in the PEP was achieved. The average performance across all tested servers was 1 GB/s for Jetstream-IU and 1.2 GB/s for Jetstream-TACC.

#### 5.1.1.4. *Ethernet bandwidth, subsequent testing as recommended by the review panel*

Further network testing relating to 10 Gbps performance, as described on ESnet’s Network Performance Testing web page<sup>4</sup> was recommended by the review panel.

We were able to replicate the initial TCP testing done with the iperf3 tool and were able to achieve 1.2 GB/s performance on both the Jetstream-IU and Jetstream-TACC clusters.

UDP bandwidth testing showed 1.2 GB/s on Jetstream-IU upon initial testing using nuttcp, matching what was seen with iperf3. At TACC initial nuttcp bandwidth performance was in the range of 0.9 to 1.2 GB/s. Similarly, the UDP burst test looking for packet loss showed zero packet loss across the IU topology for both small and large bursts; whereas, the tests on Jetstream-TACC showed packet loss of 15-20% for the small burst and approaching 60% on the large burst tests. Network kernel parameters were adjusted on Jetstream-TACC to provide sufficient memory buffers for the desired 10 Gbps with an MTU size of 9000. The performance after these adjustments indicated full 1.2 GB/s UDP bandwidth and zero packet loss for both the small and large UDP burst tests.

As a result of these tests, we have demonstrated that all components within the network fabric are capable of generating, transmitting, and receiving 10 Gbps bandwidth at the 9000 MTU packet size and that the intermediary buffers are sufficiently sized to handle this workload.

---

<sup>3</sup> <https://fasterdata.es.net/performance-testing/network-troubleshooting-tools/iperf-and-iperf3/>

<sup>4</sup> <http://fasterdata.es.net/performance-testing/network-troubleshooting-tools/nuttcp/>

### 5.1.2. File System & Storage Performance Tests

A minimum read performance of 200 MB/s and a minimum write performance of 100 MB/s from within a virtual machine to block storage. Measurements will be rounded to the nearest MB/s.

Read/write I/O performance was measured via the dd Linux utility to/from an OpenStack Cinder block device mounted within a running VM instance. A file size of 2 GB was used with a block size of 1 MB.

The performance targets as outlined in the PEP were achieved for file system and storage. The write performance for a single VM was 359 MB/s for Jetstream-IU and 108 MB/s for Jetstream-TACC. The read performance for a single VM was 244 MB/s for Jetstream-IU and 210 MB/s for Jetstream-TACC.

#### 5.1.2.1. *File system & Storage Performance Tests, subsequent testing as recommended by the review panel*

Two further performance tests of the network and storage systems were recommended by the review panel. The first recommendation was to utilize a random read/write I/O pattern in order to determine more realistic I/O characteristics, one that users might experience in the real world. The second recommendation involved utilizing a parallel distributed I/O benchmark to determine the maximum capacity that the network and storage systems could sustain.

To explore the random read/write performance, fio<sup>5</sup> was used to simulate random seek+reads and seek+writes IO patterns that might be generated by a single user running on a single instance. As might be anticipated, and documented in Table 1 below, performance was on order 1/5 that of single instance serial reads and on order 1/10 that of single instance serial writes.

To examine the upper boundaries of the network and storage performance capabilities, parallel IOR<sup>6</sup> benchmarks were ran from XXL-size instances. The XXL size allocates one whole compute blade to one running virtual machine instance. Parameter sweeps ranging from one to 166 instances were executed to determine the basic performance characteristics of the system. Write performance at IU and TACC plateaued easily within this range. Read performance on Jetstream-TACC at this point appeared to plateau within this range and Jetstream-IU's read performance saturation point exceeded 166 tasks. It is known problem that a kernel bug on the host operating system limits the number of instances on the same virtual network to 166.

The Jetstream-IU blades are configured with dual, bonded 10 Gbps NICs. After some experimentation, it was determined that up to three MPI ranks could be executed per instance with little to no noticeable performance degradation. This allowed the realm from 167 up to 498 tasks to be explored. Read performance on Jetstream-IU was found to plateau in the range of 350-400 tasks and saturate around 420 tasks.

---

<sup>5</sup> <http://freecode.com/projects/fio>

<sup>6</sup> [http://www.cs.sandia.gov/Scalable\\_IO/ior.html](http://www.cs.sandia.gov/Scalable_IO/ior.html)

Write performance for both IU and TACC clouds tends to plateau around 32 tasks. Since tuning efforts are continuing on Jetstream-TACC and because the TACC blades are not bonded, the search for the upper read performance boundary has not yet been pursued.

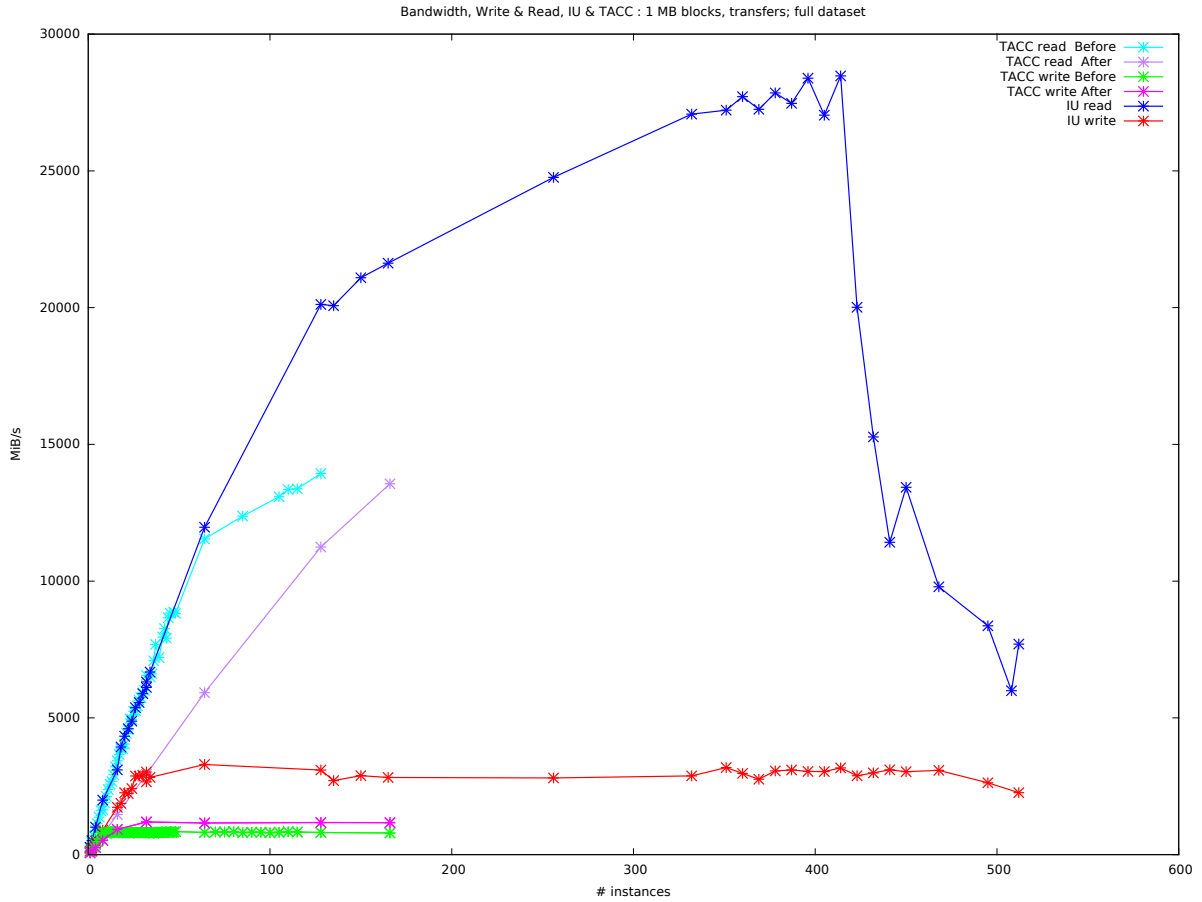
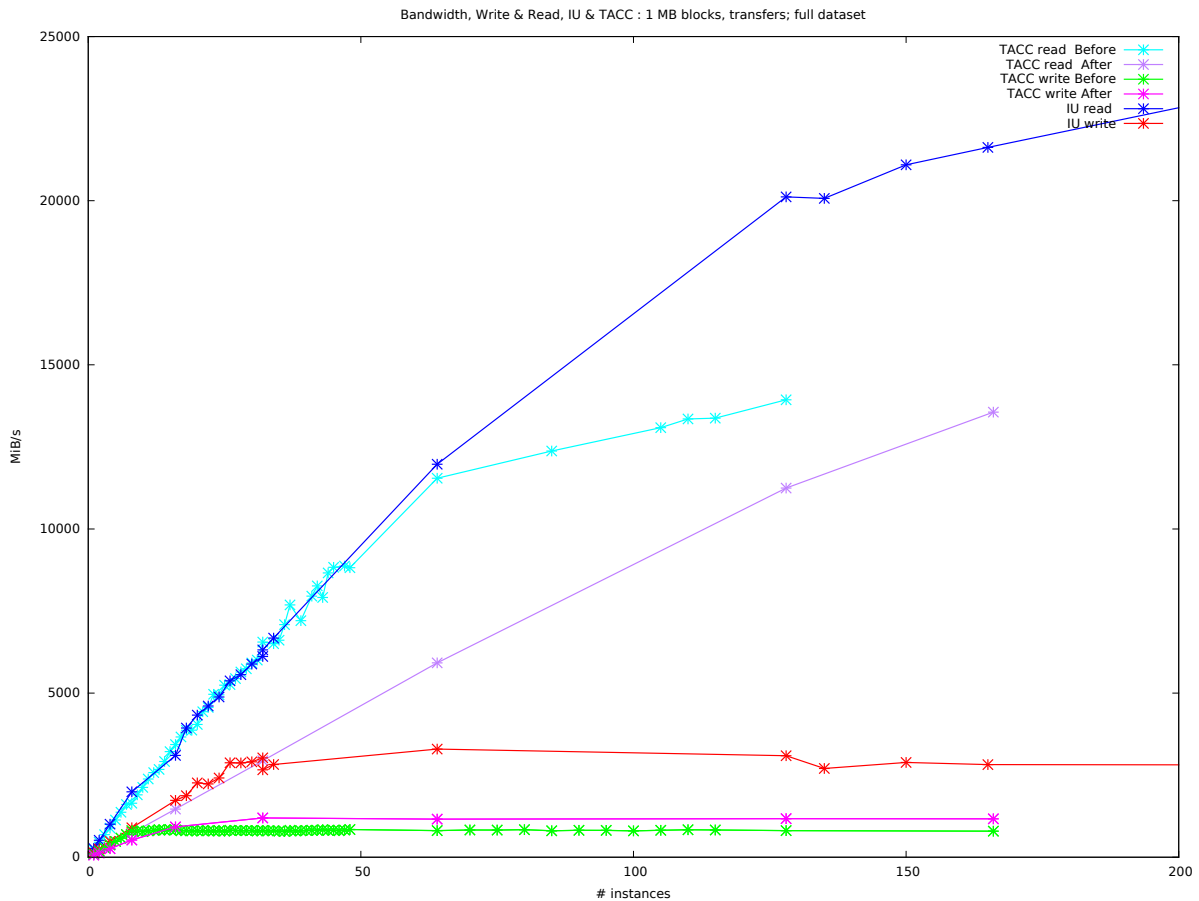


Figure 1: Network performance (MiB/s) for reads and writes at Jetstream-IU and Jetstream-TACC using a 1 MB block size with IOR up to 512 tasks (one file per task). Before and After indicate results before or after tuning of Jetstream-TACC detailed in Section 5.1.1.4.



**Figure 2: Network performance highlighting the lower bandwidth domain on Jetstream-TACC up to 166 tasks (one task per instance, one file per task) of Figure 1 above.**

It is our conclusion that the initial poor performance on Jetstream-TACC was due to the newer “Infernalis” version of Ceph being installed there as compared to the “Hammer” version installed on Jetstream-IU. An aggressive write lock algorithm was implemented in the “Infernalis” version and further refined with the “Jewel” version. When Jetstream-IU was upgraded to the “Jewel” version, it too experienced the marked decrease in write performance. Jetstream-TACC experienced a smaller decrease in performance when upgrading from “Infernalis” to “Jewel”.

The parallel IOR benchmarks have also illuminated the marked decrease in maximum performance with Jetstream-TACC relative to Jetstream-IU. As is documented in Table 2 and described above, some increase in maximum performance has been attained on Jetstream-TACC via blade-level kernel tuning and NIC configuration changes. Further cluster-wide tuning and configuration changes continue to be explored and we will discuss the results of those efforts at our next review. Jetstream-IU and Jetstream-TACC also vary in low-level RAID configuration which has an impact on performance, which should be less noticeable. The RAID configuration is unlikely to change in future tuning efforts. Jetstream-IU storage servers have disks configured in RAID 0 (striped) pairs for all drives whereas Jetstream-TACC is not leveraging RAID groups. Note that Ceph replication provides redundancy making traditional RAID5/6 configuration unnecessary.

It is also worth noting when viewing the data in Table 2 that tests prior to, and including, June 14<sup>th</sup> were the average of multiple runs. The systems were much less loaded in the early operations phase and standard deviation of each run was small. As the systems have become more utilized there’s a higher variance between each test. The Jetstream team did not feel it was in the best interest of the research community or the NSF to quiesce the systems in order to perform further storage or network parameter studies in isolation.

**Table 2: I/O benchmark summary**

Benchmark	Reads (MB/s)	Writes (MB/s)	Program	Dates	Notes
Serial IU	244	359	dd	04/2016	Acceptance report (average)
Serial TACC	210	107	dd	04/2016	Acceptance report (average)
Serial IU	258	316	dd	05/04/2016	Review panel rerun (average)
Serial TACC	198	90	dd	05/04/2016	Review panel rerun (average)
Serial IU	273	173	dd	06/14/2016	Average R/W
Serial IU	279	165	IOR	06/14/2016	XSEDE16 paper (average)
Serial TACC	247	251	dd	06/14/2016	Average R/W
Serial TACC	241	255	IOR	06/14/2016	XSEDE16 paper (average)
Random IU	51	17	fio	06/14/2016	Average R/W
Random TACC	59	20	fio	06/14/2016	Average R/W
MPI IU, single	281	144	IOR	08/06-07/2016	Max R/W
MPI TACC, single	239	96	IOR	08/04-05/2016	Before tuning (max)
MPI TACC, single	95	68	IOR	08/04-05/2016	After tuning (max)
MPI IU, parallel	29766	3325	IOR	08/06-07/2016	max write at 414 tasks write plateau at 32 tasks max read at 394 tasks, read plateau at 396 tasks
MPI TACC, parallel	14611	885	IOR	08/04-05/2016	Before tuning, max write at 48 write plateau at 32 tasks max read at 128 read plateau at 128
MPI TACC parallel	14217	1253	IOR	08/12-14/2016	After tuning, max write at 32 tasks write plateau at 32 tasks max read at 166< tasks read plateau at 166< tasks

**5.1.3. System Stability and Uptime Performance Tests**

The system must maintain a continuous 95% availability for a period of 14 days.

For the Jetstream system, i.e. Jetstream-IU and Jetstream-TACC running as an integrated entity, the system must be up, running stably, and available for users to engage in their routine research activities. The Jetstream system was up, running stably, and available for daily usage for a period exceeding 14 days starting 6-Apr-2016 and maintained stability beyond the 20-Apr-2016 testing period.

MTBF requirements are not applicable to the test environment but the system operated continuously for over 28 days.

## **5.2. Integrated Cloud Operations**

The inherent value of Jetstream is not in its hardware components; but rather, in the integration of the various software and hardware parts into its whole. Metrics designed to demonstrate the achievement of this are listed below.

### **5.2.1. Provide “Self-service” Academic Cloud Services**

On a routine and daily basis, users of the Integrated Jetstream system are:

- Authenticating via GlobusAuth to the Jetstream Atmosphere user interface. As of 21-Apr-2016 159 users from 68 distinct projects have accessed Jetstream.
- Launching virtual machine instances from the menu of pre-packaged VM images installed in Jetstream’s libraries.
- Suspending a running instance to a disk image.
- Creating and accessing persistent cloud storage (volumes).
- Modifying a running instance, creating a snapshot of that instance to disk and storing that image. New community contributed images such as R OpenSci, MAKER, ASTRAL, NeurDenian, Newbler, and Galaxy are available as examples.

Jetstream has also demonstrated the ability to instantiate and sustain more than 640 VMs operating simultaneously. The Jetstream system has instantiated and supported 993 VMs running simultaneously on 15-Apr-2016.

### **5.2.2. Host Persistent Science Gateways**

*Jetstream has shown the ability to support persistent science gateways*

- The Galaxy bioinformatics gateway has been installed and became operational on 15-Apr-2016. A known Galaxy workflow was executed on 20-Apr-2016 and provided the correct results within the specified performance parameters.
- The SEAGrid gateway was installed and became operational on 15-Apr-2016 and operated until 29-Apr-2016 with performance within the specified criteria.

### **5.2.3. Data Movement, Storage, and Dissemination**

- Authorized users routinely transfer files into and out of Jetstream utilizing Globus Connect Personal from within a running instance. In addition, users can leverage their own desired transfer tools such as SFTP, iRODS, and HTTP protocols.
- Users have been able to suspend a running instance, save it to disk, upload it to IU scholarworks.iu.edu, and using the existing online forms, submit that document for publication via IUScholarworks. E.g. CentOS 7 (7.2) Development; Fischer, Jeremy; Stewart, Craig; DOI: doi:10.5967/P93W2M; URI: <http://hdl.handle.net/2022/20772>

### **5.2.4. Provide Virtual Linux Desktop Services to Tablet Devices**

- Authorized users have accessed Jetstream from a tablet device and loaded a virtual Linux desktop configured in a manner that allowed them to access Jetstream services. This functionality was tested using the RealVNC viewer app on an iOS device.

## 6. Conclusion

It is evident from the rapid, dynamic nature of cloud software development that although Jetstream is a hardware acquisition functioning as both a pilot and production system the software components will have a large impact on the performance and features of the Jetstream system throughout the operations and management phase of the project. The Jetstream team will continue to optimize and harden the resources and software based on community best practices. Knowing the results of the Jetstream NSF review, and supported by the data presented in this report, we conclude that the integrated Jetstream system based on Dell PowerEdge clusters has met or exceeded the acceptance tests required under the agreement between Indiana University and Dell, Corporation (PO Numbers 1681608 and 1681609), pursuant to IU's Cooperative Service Agreement with the National Science Foundation for award # ACI-1445604: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment). The PI, Dr. Craig A. Stewart, has approved this report and endorsed this conclusion.