Acceptance Test for Jetstream Test Cluster –

Jetstream-Arizona (JA) Dell PowerEdge Test and Development Cluster

David Y. Hancock
Matthew R. Link
Craig A. Stewart
George W. Turner

Indiana University

PTI Technical Report PTI-TR15-007

25 August 2015

Citation:

Hancock, D.Y., M.R. Link, and C.A. Stewart. "Acceptance Test for Jetstream Test Cluster – Jetstream-Arizona (JA) Dell PowerEdge Test and Development Cluster," Indiana University, Bloomington, IN. PTI Technical Report PTI-TR15-007, Jan 2015. Available at http://hdl.handle.net/2022/20355

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

This material is based upon work supported in part by the National Science Foundation Award ACI-1445604: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment, Craig A. Stewart, IU, Principal Investigator. The Indiana University Pervasive Technology Institute was established with major grant funding from the Lilly Endowment, Inc. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Lilly Endowment.

Table of Contents

Table of Tables

1. Introduction

Jetstream will be a configurable, large-scale computing resource that leverages both on-demand and persistent virtual machine technology to support a much wider array of software environments and services that current NSF resources can accommodate. As a fully configurable "cloud" resource, Jetstream bridges the obvious major gap in the current ecosystem, which has machines targeted at large-scale high-performance computing, high memory, large data, high-throughput, and visualization resources. As the open cloud for science, Jetstream will:

- 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.
- Host persistent science gateways. Jetstream will support persistent science gateways, including the ability to host 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.
- Support 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 stored VM.
- Provide virtual Linux desktop services delivered from Jetstream to tablet devices. This service is designed to increase access to Jetstream for users at institutions with limited resources, including small schools, schools in EPSCoR states, and Minority Serving Institutions.

The Jetstream test cluster is one of three planned server components of the overall Jetstream cloud system. This server, called Jetstream-Arizona (JA), is a Dell PowerEdge (PE) system. It was purchased by Indiana University for eventual delivery to University of Arizona 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).

This system was ordered on 03/12/2015 via purchase order number 1615334. It arrived at the Indiana University Bloomington (IUB) Data Center on 04/27/2015. This system has the following basic characteristics:

- The hardware infrastructure is based upon Dell PowerEdge servers with a 10/40 Gb/s Fat-Tree Ethernet fabric[1].
  - Each node contains two Intel E5-2680v3 (12-core) 2.5 GHz processors and 128 GB RAM, for a total of 24 processing cores, resulting in 806.4 GFLOPS per node.
  - The overall system includes 16 nodes, 32 CPUs, 384 processor cores, 12.9 TFLOPS peak processing capability, and 2 TB RAM.
- The cloud infrastructure is based upon OpenStack with its ability to deliver virtualized compute capacity[2].

[1] http://www.dell.com/learn/us/en/04/campaigns/poweredge-13g-server

[2] https://www.openstack.org/

Following are a detailed system description, details on the performance targets, the methods used to perform the acceptance tests, and the achieved performance.

The Jetstream-Arizona Test and Development cluster was delivered to the Indiana University Bloomington Data Center. No problems were encountered installing and booting the two (2) R630 management servers, four (4) R730 storage servers, and 16 M630 compute blades.

Table 1-1. Acceptance test timeline.

Acceptance Timeline	
Purchase order	03/12/2015
System arrival	04/27/2015
System boot	05/04/2015
Functionality pass	06/01/2015
Performance pass	08/21/2015
14-day stability pass	06/05/2015

2. Role of Jetstream-Arizona test cluster as a component of the overall Jetstream system

Jetstream is a first-of-a-kind acquisition for the National Science Foundation in that it is the first cloud system to be acquired through the series of solicitations and awards of the roughly decade-old High Performance System Acquisition program. Jetstream is also the first system in recent years in which the system integrator, not a computer vendor, is acknowledged as the awardee. None of Jetstream's four major functions can be achieved by a cluster system as traditionally assembled, integrated, shipped, and delivered by a US computer system vendor.

The Jetstream-Arizona test cluster serves as the proving ground and test / development server for the overall Jetstream service, on which the Jetstream team integrates the software that will bridge the gap in functionality between the system as delivered by the vendor and the system as fully integrated. The system has three major layers of software (cloud environment, VM containers, and cloud interface system) installed above the software that is traditionally installed on a cluster as purchased by an awardee and accepted by the NSF.

The functions of the Jetstream-Arizona test cluster within the overall project varies in the two phases anticipated for the project – Construction and Operations & Management, as follows:

- Construction phase. During the Construction (current) phase of the Jetstream project, the Jetstream-Arizona test cluster:
  - Is used as the main platform supporting the software integration tasks required to build a cluster so that it functions as part of the integrated, multi-region Jetstream system.

    - Is used as a test system by early adopters and friendly users, giving us early feedback from the intended user community on whether we are successfully deploying an integrated system that meets the four major functions described for Jetstream.
- Operations and & Management phase. During the Operations & Management phases (including early operations phase per current version of the Jetstream Program Execution Plan), the Jetstream-Arizona test cluster will be used for the following purposes:
    - Regression testing for assurance of proper functionality when upgrades and/or patches become available for software components that make up part of the Jetstream cloud system software stack.
    - Support testing of possible additions to components and architectural features of Jetstream. For example, if we were to consider adding large memory nodes or GPU nodes to Jetstream, such nodes would be tested and any needed software integration would be performed on the test system.

During the first few months of the Jetstream construction phases the Jetstream-Arizona cluster will be located at IU Bloomington. At some point no later than the beginning of the formal Operations & Management phase, this system will be shipped to the University of Arizona, where it will remain for the duration of the project. We are following industry best practices in having a test system isolated from production cloud facilities so that issues, problems, and needs are encountered and addressed on the test system. A logically and physically isolated test system will allow the Jetstream team to test essential software upgrades (such as those required to maintain security or interoperability with other components of the NSF-funded XD ecosystem) and ensure that when they are applied to the production system, such changes do not adversely impact uptime, functionality, or the user experience with the production system.

Because the test system is the basis for software integration, the relevant tests are simple and show whether the test system functions as specified in the Jetstream grant proposal as accepted and funded by the NSF. The tests involve the basic hardware performance of the system. Success in those tests signifies that the system is a suitable base on which the Awardee (IU as lead institution) will perform the required tasks to integrate the hardware acquired and create the Jetstream system as described in our proposal and PEP.

In the remainder of this document we describe the system itself, the acceptance test criteria, and the acceptance test methodology and results. Our conclusion states that the system as acquired and now installed meets the criteria specified for the Jetstream test cluster in our proposal as approved by the NSF and as specified in greater detail in the Program Execution Plan for Jetstream.

3. System description

3.1. Hardware

3.1.1. System topology

The Jetstream-Arizona (JA) test cluster, as delivered to IUB, consists of two (2) PE R630 management servers, four (4) PE R730 storage servers, and 16 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 400GB 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 200GB 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. These uplink into the Dell Force10 S6000 spine switch producing 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 is populated with dual Intel Xeon E5-2680v3, 12-Core 2.5 GHz, 2133MHz bus, with 30MB L3 cache, 12x256KB L2 cache.

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

All 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 Gen3x8 slot, dedicated RAID card; Intel QPI @ 9.6 GT/s.

R730: two (2) PCIe Gen3x16 slots, four (4) PCIe Gen3x8 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 Gb/s, version 1.3.38, firmware-version: 4.25 0x8000143f 0.0.0.

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

3.1.9. *All communications hardware, including private channels*

Dual Dell Force10 MXL 10/40 Gb/s Ethernet blade switches (leaf).

Dell Force10 S6000 10/40 Gb/s Ethernet top of rack switch (spine).

Sixteen M630 blades connect at 10 Gb/s to the Force10 MXL, which uplinks to the Force10 S6000 with two bonded 40 Gb/s resulting in a two-to-one over subscription. M630 and M730 management and storage nodes respectively link into the F10 S6000 spine switch.

Dell N3048 1 Gb/s management Ethernet switch provides 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 1TB 7.2K RPM NL-SAS 6Gbps 2.5in Hot-plug system devices.

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

R730: PERC H730P integrated RAID controller, 2GB cache; dual 200GB Solid State Drive SATA Mix Use MLC 6Gbps 2.5in Flex Bay system devices; twelve 4TB 7.2K RPM NL-SAS 6Gbps 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 JAM1 management server.

3.2. Software

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

Bare metal: CentOS 7.1.1503 w/ kernel 3.10.0-229.el7.x86_64, stock.

VM: CentOS 7.1.1503 w/ kernel 3.10.0-229.el7.x86_64, stock.

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 1.3.38, firmware 4.25 0x8000143f 0.0.0.

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 0.94.2-0.el7 (Hammer) for block and object storage.

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

Intel compilers, version 15u3, Intel MPI version 5.0.3p-048.

3.2.8. All patches and bug fixes

CentOS 7.1.1503 with patches up to date as of 06/01/2015.

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

qemu-kvm-1.5.3-86.el7_1.5

libvirt-daemon-kvm-1.2.8-16.el7_1.3

OpenStack 2015.1.0-3 (Kilo)

4. Acceptance test criteria

The Project Execution Plan (PEP) between Indiana University and the National Science Foundation stipulates the acceptance criteria for Jetstream. The PEP has undergone peer review and has been recommended for acceptance to the Division of Grants and Agreements. Once that process is complete IU's cooperative agreement with the NSF will be amended. The tests from the PEP below align with the master Statement of Work (SoW) signed by Dell on July 31, 2015.

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 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 of a system implementation than a hardware implementation (as contrasted, say, to earlier systems such as Ranger, Kraken, or FutureGrid). However, it makes sense to conduct basic hardware performance tests as the first step in Jetstream acceptance testing. These criteria serve as prerequisites for other tests that verify functionality. These tests are primarily performance tests – the doing of a specific activity.

These basic hardware functionality tests are the only acceptance criteria relevant to the acceptance of the Jetstream-Arizona test cluster.

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/sec (aggregate across the node). (Measurements will be rounded to nearest 1 GB/sec.)

- 10 Gigabit Ethernet Bandwidth: the 10GigE interface on each node will achieve at least 1GB/s for large-message point-to-point transfers (Measurements will be rounded to the nearest 0.1 GB/sec.)

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.

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 in the Jetstream-Arizona test cluster.

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 was achieved. The average performance across all tested servers was 696 GFLOPS or 86% of theoretical peak performance.

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 was achieved for STREAM. The average performance across all tested servers was 90.7 GB/s for STREAM Triad.

5.1.1.3. Ethernet bandwidth

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

Iperf, with default settings, was used to measure Ethernet bandwidth and run between a management node (usually JAM1) and all M630 compute, M730 storage, and R630 management servers.

For the Ethernet Bandwidth benchmark, the performances target was achieved. The average performance across all tested servers was 1.2 GB/s.

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 were achieved for file system and storage. The performance from a single VM was 510 MB/s for writes and 849 MB/s for reads.

5.1.3. System stability and uptime performance tests

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

The Jetstream-Arizona test cluster must be up, running stably, and available for systems administrators, software developers, and performance analysts to engage in their routine activities. The cluster was up, running stably, and available for daily usage for a period exceeding 14 days.

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

5.2. Integrated cloud operations

The remaining tests described in sections 3.2, 3.3, 3.4, and 3.5 are applicable only to the production environment and were not evaluated on the Jetstream-Arizona test cluster. The environment is currently being used to prototype and test software required for production operations.

6. Conclusion

Based on the results presented in this report, we conclude that the Jetstream-Arizona Dell PowerEdge test cluster has met or exceeded the acceptance test required under the agreement between Indiana University and Dell Corporation (PO Number 1615334), 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.