DANCES Phase 1 Pilot Report

25 June 2015

Version 0.2

Table of Contents

A. Document History

Entire Document	0.1	04/15/2015	Created Document	Peter Enstrom
Entire Document	0.2	6/25/2015		V. Hazlewood

B. Document Scope

This document is the evaluation report for the pilot evaluation of DANCES Phase 1. The results of the pilot and the collected information are intended for the benefit of the XSEDE project and the general eScience community.

C. Executive Summary

A pilot for the Developing Applications with Networking Capabilities via End-to-End SDN (DANCES) Phase 1 [1] was performed by the Technology Investigation Service (TIS) evaluation team led by Bryan Webb. The other members of the pilot team were George Butler, Victor Hazlewood and Jesse Hanley. The pilot also had access to Eric Boyer at NCSA as a subject matter expert in networking.

The major project goal of DANCES project is to add network bandwidth scheduling capability via software-defined networking (SDN) programmability to selected cyberinfrastructure applications which is of strong interest to XSEDE Operations networking and XSEDE service providers. The TIS DANCES pilot scope aligned with DANCES Phase 1 which was to evaluate one or more vendor hardware devices and the corresponding SDN software in a test environment to see if it meets the OpenFlow 1.3 requirements need by the DANCES project in order to attempt to add network scheduling capability in a supercomputer environment. DANCES Phase 1 consisted of work to transition a DANCES test environment (called Phase 0) from the Phase 0 virtual environment to a physical installation in the NICS machine room. DANCES Phase I attempted to test Juniper, Brocade and Dell network equipment, however, only Juniper equipment was obtained and tested during the DANCES Phase I performance period. In the NICS network laboratory two Juniper MX80-48T switches were obtained and installed. Vendors described these were supposed to support OpenFlow 1.3 capabilities. Several servers were configured as VM hosts to simulate a "local side" and a "remote side" for the application servers and a Ryu OpenFlow controller was configured. DANCES Phase I was able to test the OpenFlow controller environment with existing OpenFlow 1.3 commands that were available on the Juniper MX80-48Ts using Juniper beta code obtained from PSC that was described as OpenFlow 1.3 compliant to determine its applicability for use as a component in a network bandwidth scheduling capability. It was determined that the Juniper software version 14.2R1.9 had some of the OpenFlow 1.3 features and capabilities, but the two important capabilities of queuing and slicing for bandwidth reservation (also called quality of service) specified in the OpenFlow 1.3 specification were not available in the Juniper JUNOS beta code. Therefore, DANCES Phase I and this TIS pilot project specifically cannot recommend Juniper MX80-48T devices with the latest software, available in beta by March 2015, to have the OpenFlow 1.3 capabilities needed to perform network bandwidth reservation capabilities as described by the DANCES project.

D. Introduction

TIS and XSEDE, in general, is interested in software defined networking capabilities that can improve the performance and efficiency of scientific workflows.   TIS was made aware of the Developing Applications with Networking Capabilities via End-to-End SDN (DANCES) project that was being developed by staff from some XSEDE service providers.

The DANCES acronym stands for Developing Applications with Networking Capabilities via End-to-End SDN. The major project goal of DANCES is to add network bandwidth scheduling capability via software-defined networking (SDN) programmability to selected cyberinfrastructure applications. The DANCES bandwidth control and scheduling capability is designed to provide bandwidth reservation capabilities and mitigate congestion-induced throughput problems on end-site networks. The selected cyberinfrastructure applications for DANCES include GridFTP and wide-area distributed file systems implemented, where possible, using resource management and job scheduling on supercomputers. DANCES work includes extensions to the TORQUE/Moab scheduling and management currently in use at the XSEDE supercomputing sites to support networking as a requested resource. The wide area distributed file systems selected for bandwidth scheduling integration are the XSEDE-wide File System, implemented using IBM's General Parallel File System (GPFS) and SLASH2 a file system which was developed at Pittsburgh Supercomputing Center.

The overall purpose of the DANCES project is to investigate and select software and hardware that meet the needs to engineer and integrate end-to-end software defined networking (SDN) into supercomputing infrastructure with the goal of improving the stability, predictability, and performance of the data flows across the wide area network infrastructure. DANCES Phase I is one step towards this goal to evaluate vendor hardware and software and other associated SDN software that would provide the OpenFlow 1.3 capabilities needed to implement the goal. DANCES Phase 1 identified and evaluated prospective hardware and software to support the project’s advanced network control requirements. In Phase 0 virtual environment testing, staff determined that Ryu OpenFlow controller software provided the necessary OpenFlow 1.3 capabilities needed by DANCES so Ryu was used in DANCES Phase I.   Contact was made with Juniper, Brocade and Dell to identify network equipment that may be OpenFlow 1.3 feature capable and attempts were made to obtain two network devices for Phase I. Only Juniper provided two devices by the DANCES Phase I deadline of February 2015.

The overall architecture of DANCES Phase I is described in Figure 1. Phase I consisted of two Juniper MX80-48Ts each connected to virtual hosts that were in the same machine room and represent NICS hosts (left side) and PSC hosts (right side) and a Ryu OpenFlow controller (top). The regular Juniper JUNOS software was not advertised as having OpenFlow 1.3 capabilities.

OpenFlow 1.3 capabilities were described as available in the JUNOS beta program as of February 2015. PSC participated in the JUNOS beta program and provided the JUNOS 14.2R1.9 beta code to NICS staff for use in DANCES Phase I. Note that JUNOS 14.2R3 was not released until June 9, 2015.

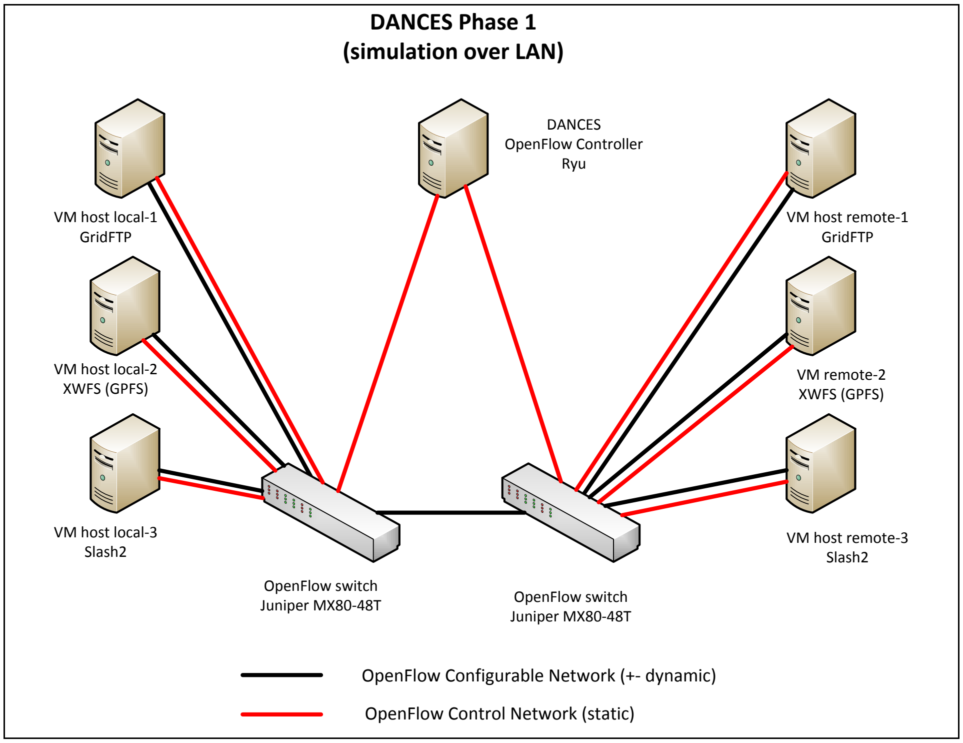

Figure 1: DANCES Phase I Architecture Overview

E. Prerequisites

Prerequisites for DANCES Phase I include network vendor hardware and software that is OpenFlow 1.3 compliant including specifically OpenFlow queueing and slicing capabilities to support minimum guaranteed bandwidth capabilities (also called quality of service) and OpenFlow 1.3 compliant controller software.   This project will focus on open source packages for the network controller software. The OpenFlow 1.3 spec allows vendors to claim support once some portion of the specified features are implemented. This can cause difficulties for the pilot.

F. Pilot Details

The requirements for this evaluation were:

- Install and configure networking hardware at multiple sites that meets the following guidelines:
  - Supports OpenFlow 1.3
  - must support queueing and slicing for bandwidth reservation (quality of service)
  - Has deep port buffers to support bulk data flows
- be able to schedule and reserve bandwidth for particular hosts, groups, etc.
- be able to sustain reserved bandwidth for given amounts of time.
- be able to measure performance tests of SDN implementations versus non-SDN baselines.
-

The open source Ryu package was selected for use as the OpenFlow controller in the DANCES Phase I. See http://osrg.github.io/ryu/

The overall architecture of DANCES Phase I is described in Figure 1. Phase I consisted of two Juniper MX80-48Ts each connected to virtual hosts that were in the same machine room and represent NICS hosts (left side) and PSC hosts (right side) and a Ryu OpenFlow controller (top). The regular Juniper JUNOS software was not advertised as having OpenFlow 1.3 capabilities. OpenFlow 1.3 capabilities were described as available in the JUNOS beta program as of February 2015. PSC participated in the JUNOS beta program and provided the JUNOS 14.2R1.9 beta code to NICS staff for use in DANCES Phase I.

The Juniper MX80-48T edge routers were configured with the following software packages:

- JUNOS Base OS Software Suite [14.2R1.9]
- JUNOS Crypto Software Suite [14.2R1.9]
- JUNOS Online Documentation [14.2R1.9]
- JUNOS Kernel Software Suite [14.2R1.9]
- JUNOS Packet Forwarding Engine Support (MX80) [14.2R1.9]
- JUNOS Routing Software Suite [14.2R1.9]
- JUNOS SDN Software Suite [14.2R1.9]
- JUNOS Services Application Level Gateways [14.2R1.9]
- JUNOS Services Crypto [14.2R1.9]
- JUNOS Services IPSec [14.2R1.9]
- JUNOS Services Jflow Container package [14.2R1.9]
- JUNOS Services NAT [14.2R1.9]
- JUNOS Services RPM [14.2R1.9]

- JUNOS Services Stateful Firewall [14.2R1.9]
- JUNOS Services SSL [14.2R1.9]
- JUNOS Base OS boot [14.2R1.9]

The OpenFlow features support configured on the MX80's are as follows:

Openflowd platform feature support

- Flow statistics: Yes
- Table statistics: Yes
- Port statistics: Yes
- Group statistics: Yes
- 802.1d spanning tree: No
- Reassemble IP fragments: No
- Queue statistics: Yes
- Match IP addresses in ARP pkts: No

Openflowd platform match condition support

- Switch input port: Yes
- VLAN vid: Yes
- Ethernet source address: Yes
- Ethernet destination address: Yes
- Ethernet frame type: Yes
- IP protocol: Yes
- TCP/UDP source port: Yes
- TCP/UDP destination port: Yes
- IPv4 source address: Yes
- IPv4 destination address: Yes
- IPv6 source address: Yes
- IPv6 destination address: Yes
- VLAN priority: Yes
- IP ToS (DSCP field): Yes

Openflowd platform action support

- Output to switch port: Yes
- Set the 802.1q VLAN id Yes
- Set the 802.1q priority: No
- Strip the 802.1q header: Yes
- Ethernet source address: No
- Ethernet destination address: No
- IP source address: No
- IP destination address: No

- IP ToS (DSCP): No
- TCP/UDP source port: No
- TCP/UDP destination port: No
- Output to queue: No
- Execute Group: Yes

The two virtual machine nodes each had an interface dedicated to the OpenFlow network. The OpenFlow network covered the 10.1.50.0/24 IP space. Both nodes ran Centos 6.5 with kernel 2.6.32-431.

Results

Issues Encountered

The Juniper routers appear to provide the wrong version info when providing switch information to a controller.

```
connected socket:<eventlet.greenio.GreenSocket object at 0x1b82050>
address:('10.10.10.1', 49803)
hello ev <ryu.controller.ofp_event.EventOFPHello object at 0x1b82590>
unsupported version 0x4. If possible, set the switch to use one of the
versions [5]
error msg ev version: 0x4 msg_type 0x1 xid 0x0
OFPErrorMsg(code=0,data='',type=1) type 0x1 code 0x0
```

Another issue encountered was that the OpenFlow controller did not have to be present for traffic to proceed. We believe this was due to the hosts being on the same router and the underlying network interface taking control.

The biggest issue encountered was the lack of queuing and slicing support in the MX80 [2]. Though queue statistics could be gathered, there was no way to enqueue flows on these routers. **This prevents further work on the pilot since the basic functionality needed is not available.** An example output of the queuing stats follows:

```
admin@ofrtr01> show openflow statistics queue
Openflow queue statistics information:
Switch Name   Port No Queue Id TX bytes TX packets Tx errors
OFswitch1    34483  0      436638  5180      0
OFswitch1    34483  1      0       0         0
OFswitch1    34483  2      0       0         0
OFswitch1    34483  3      0       0         0
OFswitch1    34483  4      0       0         0
OFswitch1    34483  5      0       0         0
OFswitch1    34483  6      0       0         0
OFswitch1    34483  7      0       0         0
OFswitch1    44383  0      0       0         0
OFswitch1    44383  1      0       0         0
```

```
OFswitch1    44383   2     0     0     0
OFswitch1    44383   3     0     0     0
OFswitch1    44383   4     0     0     0
OFswitch1    44383   5     0     0     0
OFswitch1    44383   6     0     0     0
OFswitch1    44383   7     0     0     0
```

Due to the 1 GigE limitation of the Openflow network, we were unable to test if the Juniper can sustain high (~10 GigE) throughput. Iperf tests were performed with the results below. Gigabit performance was experienced.

```
OpenFlow Network Iperf performance
[5] local 10.10.50.101 port 50570 connected with 10.10.50.102 port 5001
[ ID] Interval Transfer Bandwidth
[ 5] 0.0-5.0 sec 562 MBytes 943 Mbits/sec
[ 5] 5.0-10.0 sec 561 MBytes 942 Mbits/sec
[ 5] 10.0-15.0 sec 561 MBytes 942 Mbits/sec
[ 5] 15.0-20.0 sec 561 MBytes 941 Mbits/sec
[ 5] 20.0-25.0 sec 561 MBytes 941 Mbits/sec
[ 5] 25.0-30.0 sec 561 MBytes 941 Mbits/sec
[ 5] 0.0-30.0 sec 3.29 GBytes 942 Mbits/sec
[ 4] local 10.10.50.101 port 5001 connected with 10.10.50.102 port 39481
[ 4] 0.0-5.0 sec 561 MBytes 941 Mbits/sec
[ 4] 5.0-10.0 sec 561 MBytes 941 Mbits/sec
[ 4] 10.0-15.0 sec 561 MBytes 941 Mbits/sec
[ 4] 15.0-20.0 sec 561 MBytes 941 Mbits/sec
[ 4] 20.0-25.0 sec 561 MBytes 941 Mbits/sec
[ 4] 25.0-30.0 sec 561 MBytes 941 Mbits/sec
[ 4] 0.0-30.0 sec 3.29 GBytes 941 Mbits/sec
```

The OpenFlow network shows no clear signs of performance degradation at the 1 GigE level. Because queues are at only partially implemented in the JUNOS SDN software stack, these routers are not useful for DANCES Phase I but may eventually become useful. Without this ability to create, modify, and remove queue capabilities, the Juniper MX80-48T does not benefit DANCES Phase I which has the goal of network bandwidth reservation (minimum guarantee).

G. Installation Information

Hardware

The hardware obtained was two Juniper MX80-48T and three virtual machines (VM) were setup. Two hosts configured as the application servers and one configured as the Ryu openflow controller system.

Software

The Juniper MX80-48T JUNOS software initially loaded on the system was not OpenFlow compliant. JUNOS 14.2R1.9 was obtained via a Juniper JUNOS beta program from PSC and installed on the MX80-48T network devices. The VM hosts were configured as CentOS 6.5 kernel version 2.6.32-431. Ryu from the February 2015 build was used for the OpenFlow controller. See http://osrg.github.io/ryu/ for Ryu information

H. Usage Information

Since the Juniper MX80’s did not support the queuing and slicing features of OpenFlow 1.3 the DANCES Phase I testing was not able to continue past the configuration of the Juniper devices, the Ryu openflow controller and initial setup of the host virtual machines.

I. Pilot Result

The Juniper MX80-48T hardware with JUNOS 14.2R1.9 does not provide queue and slicing capabilities of OpenFlow 1.3.0 and therefore is not recommended for use for projects that need OpenFlow 1.3 software defined networking capabilities.

The Ryu OpenFlow controller software has capabilities to fully support OpenFlow 1.0, 1.2, 1.3, 1.4 specifications, however, those capabilities were not thoroughly tested due to the limitations of the Juniper MX80-48T hardware and JUNOS 14.2R1.9 software.