

Server and Network Specifications

This appendix provides MSDC phase 1 server testing requirements and specifications, network configurations, and buffer monitoring with configurations.

Servers

The lab testbed has forty (40) Cisco M2 Servers. Each server, 48GB RAM and 2.4Ghz, runs CentOS 6.2 64-bit OS and KVM hypervisor. There are 14 VMs configured per server and each VM is assigned 3GB RAM and allocated to one HyperThread each. The servers connect to the network via two (2) 10G NICs, capable of TSO/USO and multiple receive and transmit queues.

Figure A-1 UCS M2 Server



Server Specs

2x Xeon E5620, X58 Chipset

- Per CPU
 - 4 Cores, 2 HyperThreads/Core
 - 2.4Ghz
 - 12M L2 cache
 - 25.6GB/s memory bandwidth
 - 64-bit instructions
 - 40-bit addressing

48GB RAM

- 6x 8GB DDR3-1333-MHz RDIMM/PC3-10600/2R/1.35v

3.5TB (LVM)

- 1x 500GB Seagate Constellation ST3500514NS HD
 - 7200RPM
 - SATA 3.0Gbps
- 3x 1TB Seagate Barracuda ST31000524AS HDs
 - 7200RPM
 - SATA 6.0Gbps
- Partitions

Dual 10G Intel 82599EB NIC

- TSO/USO up to 256KB
- 128/128 (tx/rx) queues
- Intel ixgbe driver, v3.10.16

Dual 1G Intel 82576 NIC

- 16/16 (tx/rx) queues

Operating System

CentOS 6.2, 64-bit

- 2.6.32-220.2.1.el6.x86_64

Operating System Tuning

10G NIC tuning parameters (targeted at Intel 82599EB NICs):

- eth0


```
echo 1 > /proc/irq/62/smp_affinity
echo 2 > /proc/irq/63/smp_affinity
echo 4 > /proc/irq/64/smp_affinity
echo 8 > /proc/irq/65/smp_affinity
echo 10 > /proc/irq/66/smp_affinity
echo 20 > /proc/irq/67/smp_affinity
echo 40 > /proc/irq/68/smp_affinity
echo 80 > /proc/irq/69/smp_affinity
echo 1 > /proc/irq/70/smp_affinity
echo 2 > /proc/irq/71/smp_affinity
echo 4 > /proc/irq/72/smp_affinity
echo 8 > /proc/irq/73/smp_affinity
echo 10 > /proc/irq/74/smp_affinity
echo 20 > /proc/irq/75/smp_affinity
echo 40 > /proc/irq/76/smp_affinity
echo 80 > /proc/irq/77/smp_affinity
echo 40 > /proc/irq/78/smp_affinity
```
- eth1


```
echo 100 > /proc/irq/79/smp_affinity
```

```

echo 200 > /proc/irq/80/smp_affinity
echo 400 > /proc/irq/81/smp_affinity
echo 800 > /proc/irq/82/smp_affinity
echo 1000 > /proc/irq/83/smp_affinity
echo 2000 > /proc/irq/84/smp_affinity
echo 4000 > /proc/irq/85/smp_affinity
echo 8000 > /proc/irq/86/smp_affinity
echo 100 > /proc/irq/87/smp_affinity
echo 200 > /proc/irq/88/smp_affinity
echo 400 > /proc/irq/89/smp_affinity
echo 800 > /proc/irq/90/smp_affinity
echo 1000 > /proc/irq/91/smp_affinity
echo 2000 > /proc/irq/92/smp_affinity
echo 4000 > /proc/irq/93/smp_affinity
echo 8000 > /proc/irq/94/smp_affinity
echo 1000 > /proc/irq/95/smp_affinity

```

- /etc/sysctl.conf

```

fs.file-max = 65535
net.ipv4.ip_local_port_range = 1024 65000
net.ipv4.tcp_sack = 0
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_rmem = 10000000 10000000 10000000
net.ipv4.tcp_wmem = 10000000 10000000 10000000
net.ipv4.tcp_mem = 10000000 10000000 10000000
net.core.rmem_max = 524287
net.core.wmem_max = 524287
net.core.rmem_default = 524287
net.core.wmem_default = 524287
net.core.optmem_max = 524287
net.core.netdev_max_backlog = 300000

# service irqbalance stop
# service cpuspeed stop
# chkconfig irqbalance off
# chkconfig cpuspeed off

```

Virtual Machines

KVM

- libvirt-0.9.4-23.el6_2.1.x86_64
- qemu-kvm-0.12.1.2-2.209.el6_2.1.x86_64

14 VMs/server

Per VM:

- 3GB RAM
- 230.47GB 213GB HD
- Single {Hyper} Thread

Iptables Configurations

N/A.

Incast Tool Configurations

Cloudera CDH3 (Hadoop 0.20-0.20.2+923+194) with Oracle Java SE 1.7.0_01-b08

- /etc/hadoop-0.20/conf.rtp_cluster/core-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>fs.default.name</name>
  <value>hdfs://namenode.nn.voyager.cisco.com:8020/</value>
</property>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/data/tmp</value>
</property>
<property>
  <name>topology.script.file.name</name>
  <value>/etc/hadoop-0.20/conf/rackaware.pl</value>
</property>
<property>
  <name>topology.script.number.args</name>
  <value>1</value>
</property>
</configuration>
```

- /etc/hadoop-0.20/conf/rackaware.pl

```
#!/usr/bin/perl
use strict;
use Socket;

my @addrs = @ARGV;
foreach my $addr (@addrs){
  my $hostname = $addr;
  if ($addr =~ /\^(\\d{1,3}\\.(\\d{1,3}\\.(\\d{1,3}\\.(\\d{1,3})))/){
    # We have an IP.
    $hostname = gethostbyaddr(inet_aton($1), AF_INET);
  }
  get_rack_from_hostname($hostname);
}

sub get_rack_from_hostname () {
  my $hostname = shift;
  if ($hostname =~ /\^(r\\d+)/){
    print "/msdc/$1\\n";
  } else {
    print "/msdc/default\\n";
  }
}
```

- /etc/hadoop-0.20/conf.rtp_cluster1/hdfs-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>dfs.name.dir</name>
```

```

    <value>/data/namespace</value>
  </property>
  <property>
    <name>dfs.data.dir</name>
    <value>/data/data</value>
  </property>
  <property>
    <name>dfs.heartbeat.interval</name>
    <value>3</value>
    <description> DN heartbeat interval in seconds default 3 second </description>
  </property>
  <property>
    <name>heartbeat.recheck.interval</name>
    <value>80</value>
    <description> DN heartbeat interval in seconds default 5 minutes </description>
  </property>
  <property>
    <name>dfs.namenode.decommission.interval</name>
    <value>10</value>
    <description> DN heartbeat interval in seconds </description>
  </property>
</configuration>

```

- /etc/hadoop-0.20/conf.rtp_cluster1/mapred-site.xml

```

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>mapred.child.java.opts</name>
    <value>-Xmx5120m</value>
  </property>

  <property>
    <name>io.sort.mb</name>
    <value>2047</value>
  </property>

  <property>
    <name>io.sort.spill.percent</name>
    <value>1</value>
  </property>

  <property>
    <name>io.sort.factor</name>
    <value>900</value>
  </property>

  <property>
    <name>mapred.job.shuffle.input.buffer.percent</name>
    <value>1</value>
  </property>

  <property>
    <name>mapred.map.tasks.speculative.execution</name>
    <value>false</value>
  </property>

  <property>
    <name>mapred.job.reduce.input.buffer.percent</name>
    <value>1</value>
  </property>

```

```

</property>

<property>
  <name>mapred.reduce.parallel.copies</name>
  <value>200</value>
</property>

<property>
  <name>mapred.reduce.slowstart.completed.maps</name>
  <value>1</value>
</property>

<property>
  <name>mapred.local.dir</name>
  <value>/data/mapred</value>
</property>
<property>
  <name>mapred.job.tracker</name>
  <value>jobtracker.jt.voyager.cisco.com:54311</value>
</property>
<property>
  <name>mapred.system.dir</name>
  <value>/data/system</value>
</property>
<property>
  <name>mapred.task.timeout</name>
  <value>1800000</value>
</property>
</configuration>

```

Network

The following network configuration were used.

F2/Clipper References

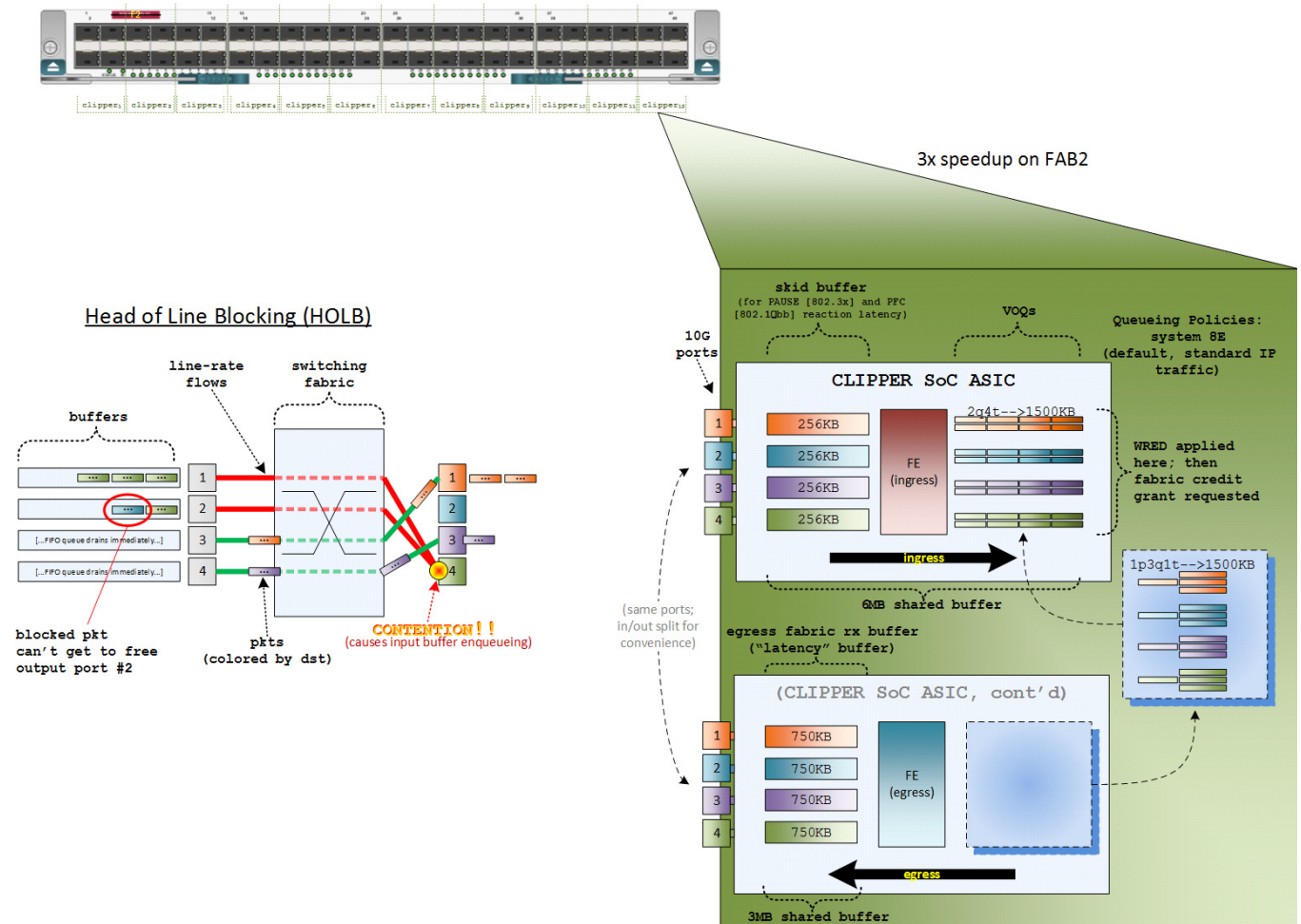
The following F2/Clipper references are available.

- Clipper ASIC Functional Specification—EDCS: 588596
- Clipper Device Driver Software Design Specification—EDCS-960101
- Packet Arbitration in Data Center Switch. Kevin Yuan

F2/Clipper VOQs and HOLB

[Figure A-2](#) shows an abstract of F2 VOQs as well as HOLB.

Figure A-2 Abstracted Diagram of F2 VOQs as well as HOLB



Python Code, Paramiko

This code implements a nested SSH session with Paramiko to attach to a linecard and issue line module specific commands:

```
def nxos_connect(nexus_host, nexus_ssh_port, nexus_user, nexus_password):
    """
    Makes an SSHv2 connection to a Nexus switch.
    """
    man = paramiko.SSHClient()
    man.set_missing_host_key_policy(paramiko.AutoAddPolicy())
    if args.verbosity > 0:
        logger.debug("nxos connect inputs are nexus_host = %s ssh port = %d User = %s passwd = %s" % (nexus_host, nexus_ssh_port, nexus_user, nexus_password))
    try:
        if args.key_file:
            man.connect(nexus_host, port=nexus_ssh_port,
                        username=nexus_user, password=nexus_password,
                        allow_agent=False, key_filename=args.key_file)
        elif args.password:
            man.connect(nexus_host, port=nexus_ssh_port,
```

```

        username=nexus_user, password=nexus_password,
        allow_agent=False, look_for_keys=False)
    else:
        man.connect(nexus_host, port=nexus_ssh_port,
                    username=nexus_user, password=nexus_password)
except paramiko.SSHException:
    return 4, man
except paramiko.BadHostKeyException:
    return 4, man
except paramiko.AuthenticationException:
    return 4, man
except socket.error:
    return 4, man

return 1, man

def attach_module(ssh, mod, check_exit_status=True, verbose=True):
    chan = ssh.invoke_shell()
    attach_cmd="attach module "+mod+"\n"

    while not chan.recv_ready():
        logger.debug("wating for channel\n")
        time.sleep(2)

    chan.send(attach_cmd)
    if args.verbosity > 0:
        logger.debug("attaching to module %s\n" % mod)
    prompt = "module-" + mod + "# "
    buff = ''
    while not buff.endswith(prompt):
        resp = chan.recv(9999)
        buff += resp

    if args.verbosity > 0:
        logger.debug("buffer output is %s" % (buff))
        logger.debug("chan is %s status is %d" % (chan, chan.recv_ready()))
    return chan

def run_lc_command(channel, mod, cmd, check_exit_status=True, verbose=True):
    processed_cmd = "term len 0 ; " + cmd
    prompt = "module-" + mod + "#"
    chan = channel
    logger.debug("prompt is %s\n" % (prompt))
    chan.send(processed_cmd)

    if args.verbosity > 0:
        logger.debug("processed command is %s" % (processed_cmd))

    buff = ''
    resp = ''

```

Spine Configuration

Forthcoming in supplemental documentation.

Leaf Configuration

Forthcoming in supplemental documentation.