Smarter technology for all

Eeking out performance with MROT

Simon Thompson | November 12th 2023

Lenovo

Portfolio update

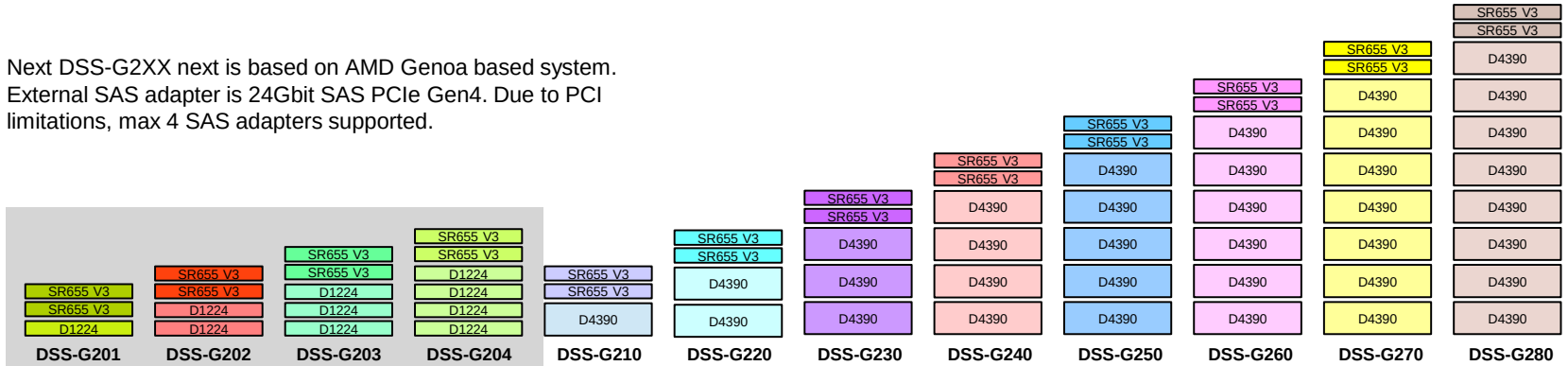
Lenovo DSS-G Gen5 Configurations (Genoa)

There is only one model number for ECE: DSS-G100.

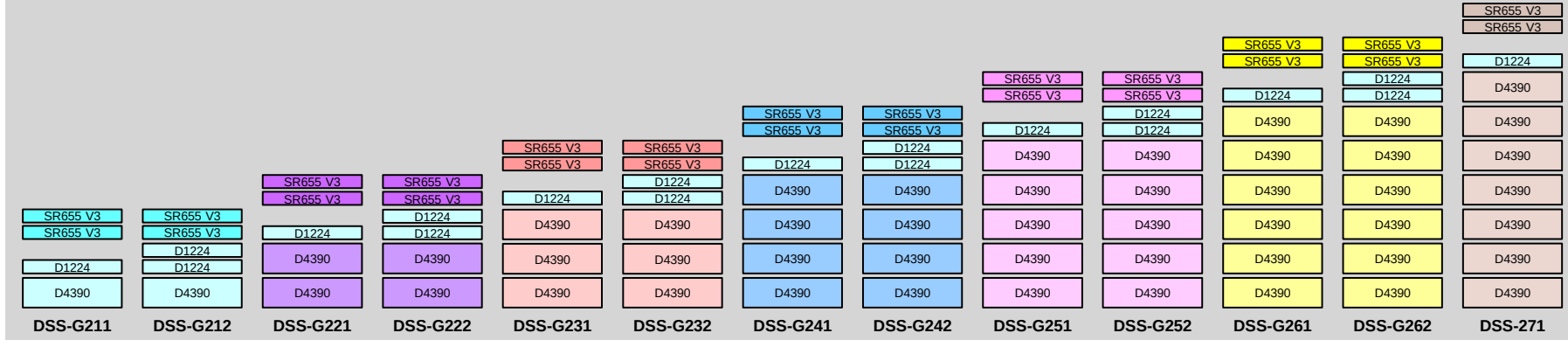
But a Lenovo ECE solution will consist of multiple ECE servers where each one will be configured with at least 6 but not more than 32 servers. Storage solutions can be a hybrid mix of DSS-G and ECE Here are 3 examples.

There are 12 homogeneous and 13 mixed model numbers for JBOD configurations:

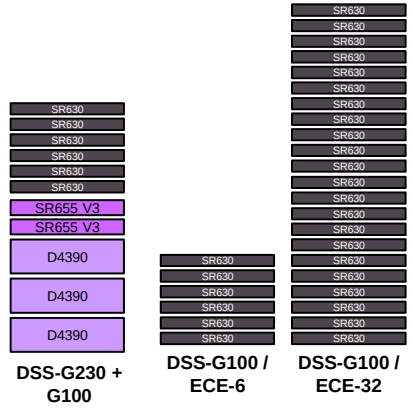
Next DSS-G2XX next is based on AMD Genoa based system. External SAS adapter is 24Gbit SAS PCIe Gen4. Due to PCI limitations, max 4 SAS adapters supported.



Support for hybrid building blocks follows

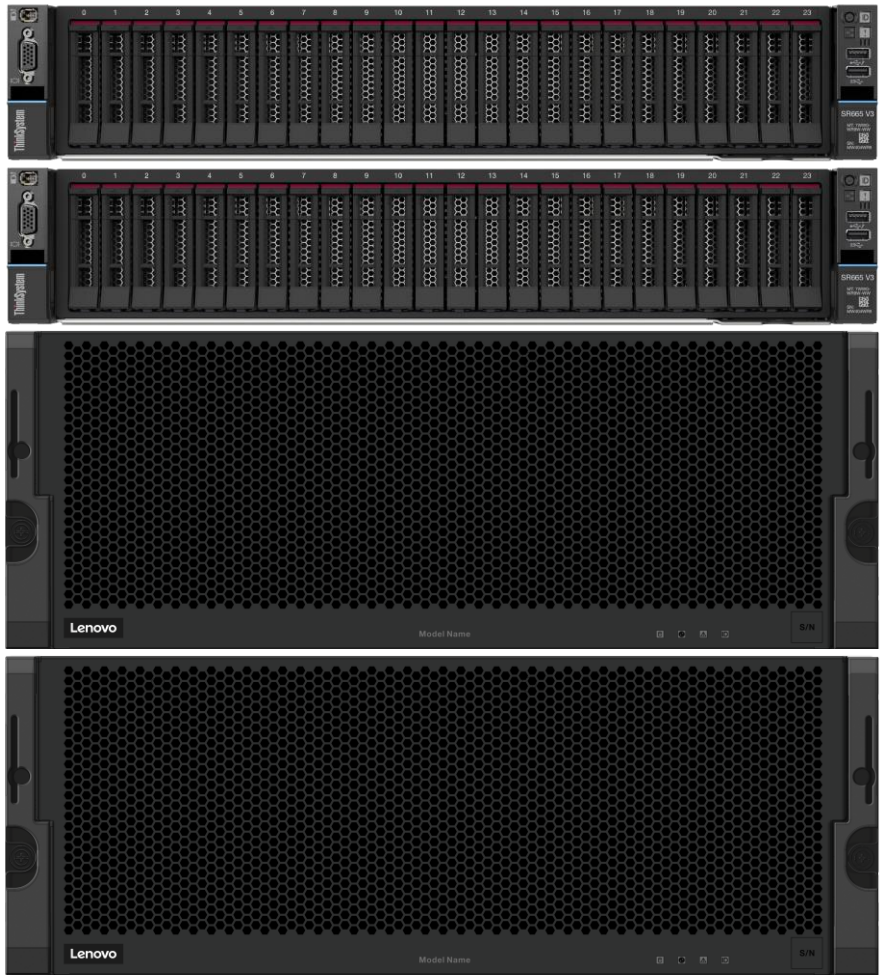


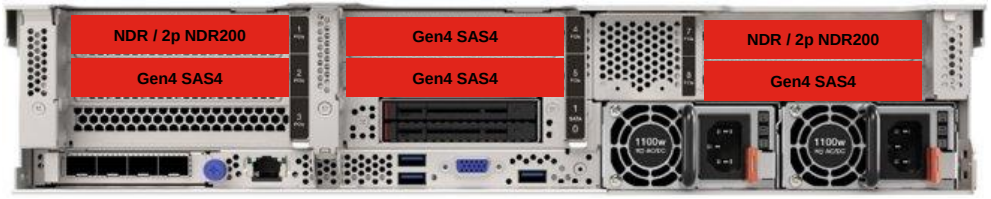
Support for hybrid building blocks follows



Footnotes:
1. Fits in 48U rack.
2. Requires 2 racks.

Gen5 DSS-G220 – SR655 V3 + D4390



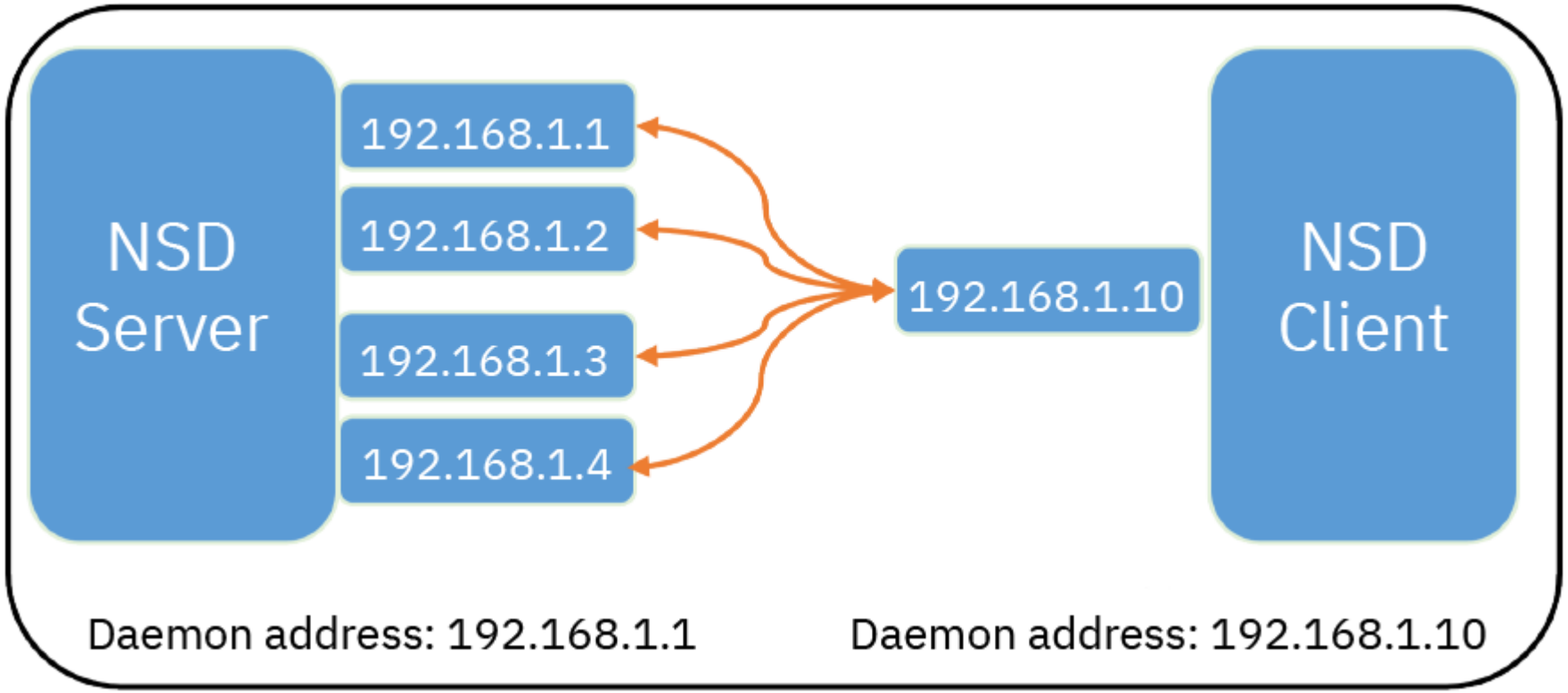


OCP Options:
2x 10G Base-T
4x 10G Base-T
2x 10/25Gb SFP28
4x 10/25Gb SFP28

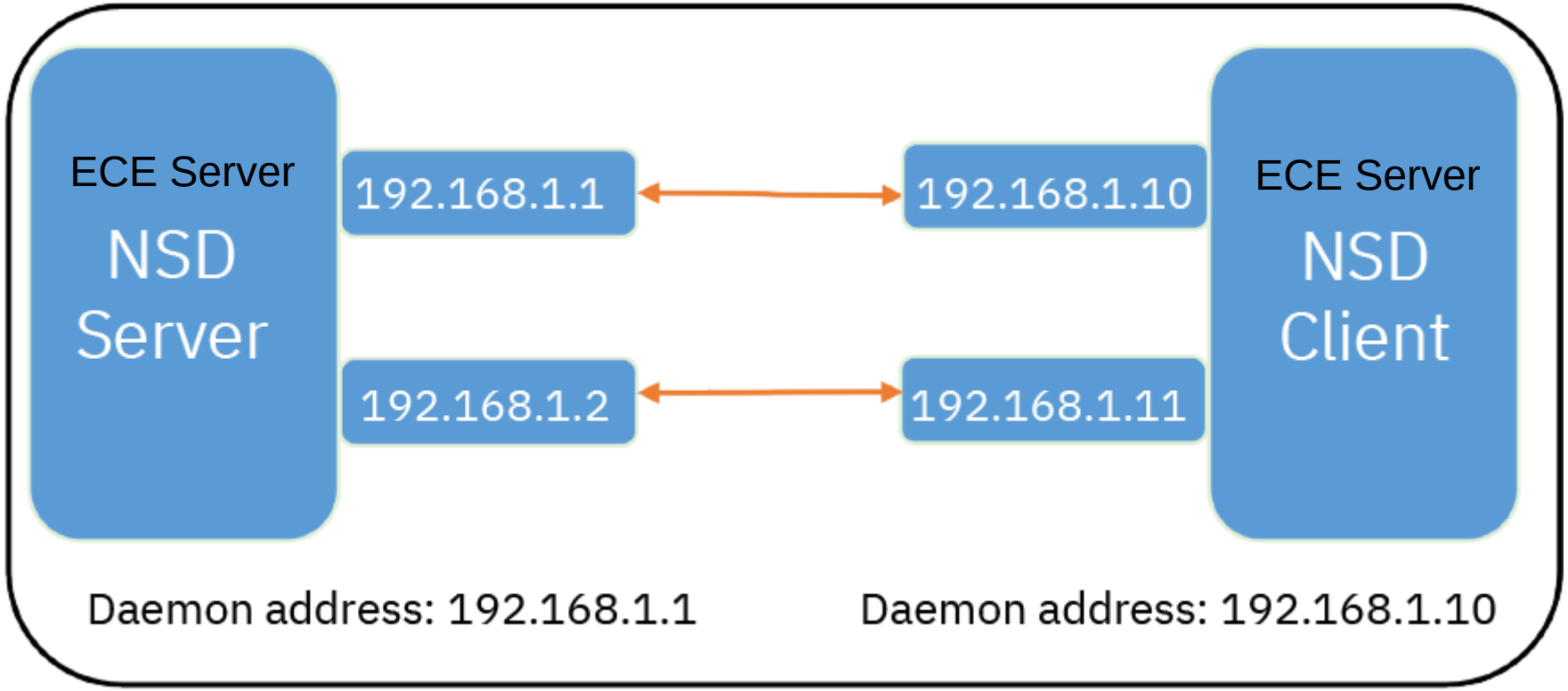
- Up to 8 D4390 HDD expansion drawers
- Up to 4 D1224 SAS SSD expansions
- Single socket AMD Genoa system
  - Single vs dual socket under evaluation
  - Min 384GB RAM
- 4x SAS Adapters (PCIe Gen4)
- Networking options (PCIe Gen5)
  - 2x NVIDIA NDR adapters
  - 2x NVIDIA NDR-200 2 port adapters

Fun and games with MROT

- For testing, we are using IPoIB, lets disable RDMA …
- `mmchconfig verbsRdma=no`
- `[I] Verbs configuration parameter "verbsRdma" is not set - skipping RDMA device discovery.`



- But with ECE between nodes …



Source policy routing

```
let rule=200+${idx}
let prior=500+${idx}
nmcli con mod ${netdev} ipv4.routing-rules "priority ${prior} iif ${netdev} table ${rule}"
nmcli con mod ${netdev} +ipv4.routing-rules "priority ${prior} from ${ipaddr} table ${rule}"
nmcli con mod ${netdev} ipv4.routes "${net_segment}/${net_cidr} table=${rule}"
# DO NOT set the route-table, otherwise the routes will not appear in the
# "default" routing table
## nmcli con mod ${netdev} ipv4.route-table ${rule}
nmcli con mod ${netdev} ipv4.route-table ""
nmcli con up ${netdev}
if [ $(sysctl --values net.ipv4.conf.${netdev}.arp_filter) != "1" ]; then
  sysctl -w net.ipv4.conf.${netdev}.arp_filter=1
fi
if [ ! -f ${SYSCTL_CNF} ]; then
  echo "net.ipv4.conf.${netdev}.arp_filter=1" > ${SYSCTL_CNF}
else
  grep -q "net.ipv4.conf.${netdev}.arp_filter=1" ${SYSCTL_CNF}
  if [ $? == 1 ]; then
    echo "net.ipv4.conf.${netdev}.arp_filter=1" >> ${SYSCTL_CNF}
  fi
fi
```

Checking the arp filter

```
[root@ece4707 ~]# sysctl net.ipv4.conf.ib0.arp_filter
net.ipv4.conf.ib0.arp_filter = 1
[root@ece4707 ~]# sysctl net.ipv4.conf.ib1.arp_filter
net.ipv4.conf.ib1.arp_filter = 1
[root@ece4707 ~]# sysctl net.ipv4.conf.ib2.arp_filter
net.ipv4.conf.ib2.arp_filter = 1
```

Checking routing

```
# ibdev2netdev
mlx5_0 port 1 ==> ib0 (Up)
mlx5_1 port 1 ==> ib1 (Up)
mlx5_2 port 1 ==> ib2 (Up)

# ip route show
default via 172.30.86.1 dev ens4f0np0 proto static metric 102
169.254.95.0/24 dev enp0s20f0u1u6 proto kernel scope link src 169.254.95.120 metric
101
172.30.80.0/20 dev ens4f0np0 proto kernel scope link src 172.30.87.67 metric 102
172.30.112.0/20 dev ib0 proto kernel scope link src 172.30.119.67 metric 150
172.30.112.0/20 dev ib2 proto kernel scope link src 172.30.119.107 metric 151
172.30.112.0/20 dev ib1 proto kernel scope link src 172.30.119.87 metric 152
```

Check the routing rules

```
# ip rule list
0:      from all lookup local
500:    from all iif ib0 lookup 200
500:    from 172.30.119.67 lookup 200
501:    from all iif ib1 lookup 201
501:    from 172.30.119.87 lookup 201
502:    from all iif ib2 lookup 202
502:    from 172.30.119.107 lookup 202
32766:  from all lookup main
32767:  from all lookup default
```

Checking Source Policy Routing

```
# ping 172.30.119.88
PING 172.30.119.88 (172.30.119.88) 56(84) bytes of data.
64 bytes from 172.30.119.88: icmp_seq=1 ttl=64 time=0.065 ms
64 bytes from 172.30.119.88: icmp_seq=2 ttl=64 time=0.065 ms
64 bytes from 172.30.119.88: icmp_seq=3 ttl=64 time=0.049 ms
64 bytes from 172.30.119.88: icmp_seq=4 ttl=64 time=0.058 ms
64 bytes from 172.30.119.88: icmp_seq=5 ttl=64 time=0.053 ms

# tcpdump -n -i ib0 icmp and host 172.30.119.67
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on ib0, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
^C
0 packets captured
308 packets received by filter
288 packets dropped by kernel

# tcpdump -n -i ib1 icmp and host 172.30.119.67
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on ib1, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
16:58:55.156108 IP 172.30.119.67 > 172.30.119.88: ICMP echo request, id 4, seq 1, length 64
16:58:55.156133 IP 172.30.119.88 > 172.30.119.67: ICMP echo reply, id 4, seq 1, length 64
16:58:56.156479 IP 172.30.119.67 > 172.30.119.88: ICMP echo request, id 4, seq 2, length 64
16:58:56.156504 IP 172.30.119.88 > 172.30.119.67: ICMP echo reply, id 4, seq 2, length 64
^C
4 packets captured
243 packets received by filter
222 packets dropped by kernel

# tcpdump -n -i ib2 icmp and host 172.30.119.67
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on ib2, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
^C
0 packets captured
464 packets received by filter
400 packets dropped by kernel
```

Traffic comes from correct interface

Subnets are important – client config

```
# mmlsconfig subnets,maxTcpConnsPerNodeConn
subnets
172.30.112.0/compute.gpfs.hpc.eu.lenovo.com;g100sr630v2.gpfs.hpc.eu.lenovo.com

maxTcpConnsPerNodeConn 4
```

mmdiag - - network is your friend …

```
Inter-node communication configuration:
  tscConnMode      mrot
  tscTcpPort       1191
 my address       172.30.87.67/20 (ens4f0np0) <c0n0>
  my addr list     172.30.119.67/20
(ib0)/compute.gpfs.hpc.eu.lenovo.com;g100sr630.gpfs.hpc.eu.lenovo.com;de6000
.gpfs.hpc.eu.lenovo.com;g204_sr650v2.gpfs.hpc.eu.lenovo.com;g2x0_sr650v2.gpf
s.hpc.eu.lenovo.com;g100_sr635v3.gpfs.hpc.eu.lenovo.com;g100sr630v2.gpfs.hpc
.eu.lenovo.com  172.30.119.87/20
(ib1)/compute.gpfs.hpc.eu.lenovo.com;g100sr630.gpfs.hpc.eu.lenovo.com;de6000
.gpfs.hpc.eu.lenovo.com;g204_sr650v2.gpfs.hpc.eu.lenovo.com;g2x0_sr650v2.gpf
s.hpc.eu.lenovo.com;g100_sr635v3.gpfs.hpc.eu.lenovo.com;g100sr630v2.gpfs.hpc
.eu.lenovo.com  172.30.119.107/20
(ib2)/compute.gpfs.hpc.eu.lenovo.com;g100sr630.gpfs.hpc.eu.lenovo.com;de6000
.gpfs.hpc.eu.lenovo.com;g204_sr650v2.gpfs.hpc.eu.lenovo.com;g2x0_sr650v2.gpf
s.hpc.eu.lenovo.com;g100_sr635v3.gpfs.hpc.eu.lenovo.com;g100sr630v2.gpfs.hpc
.eu.lenovo.com  172.30.87.67/20 (ens4f0np0)
  my subnet list  172.30.112.0/20
```

ECE server lists three addresses found on the “subnets”

From the client -> server

The client only has a single interface

```
 tscConnMode     mrot
  tscTcpPort      1191
  my address      172.30.83.5/20 (eno1np0) <c0n72>
  my addr list    172.30.115.5/20
(ibs1)/compute.gpfs.hpc.eu.lenovo.com;g100sr630.gpfs.hpc.eu.lenovo.com;de6000.gpfs.hpc.eu.lenovo.com;g204_sr
650v2.gpfs.hpc.eu.lenovo.com;g2x0_sr650v2.gpfs.hpc.eu.lenovo.com;g100_sr635v3.gpfs.hpc.eu.lenovo.com;g100sr6
30v2.gpfs.hpc.eu.lenovo.com  172.30.83.5/20 (eno1np0)
  my subnet list  172.30.112.0/20
…
 <c2n0> 172.30.87.67/2 (ece4707)
    status connected was_broken 0 err 0 reconnEnabled 1 delayedAckEnabled 1
    connMode mrot shutting 0 handlerCount 0 need_notify 0 leaseSentOn 1
    nMaxTcpConns 2 (2) nActiveCount 2 nActiveState 0x3 (1100000000000000)
    nInuseTcpConns 0 currTcpConnIndex 1 availableTcpConns (1111111111111111)
    nReservedSmallMsgTcpConns 0 currSmallMsgTcpConnIndex 0 currLargeMsgTcpConnIndex 0
    reconnectTcpConns (0000000000000000) disconnectTcpConns (0000000000000000)
    Inuse owner:
      [ 0]:0          [ 1]:0          [ 2]:0          [ 3]:0
      [ 4]:0          [ 5]:0          [ 6]:0          [ 7]:0
      [ 8]:0          [ 9]:0          [10]:0          [11]:0
      [12]:0          [13]:0          [14]:0          [15]:0

    IpPair Table (offset 2 [43592/0/3]):
      idx iface           status ping_cnt source          destination     subnet
        0 ibs1              up         0 172.30.115.5    172.30.119.67   172.30.112.0/20
        1 ibs1              up         0 172.30.115.5    172.30.119.87   172.30.112.0/20
        2 ibs1              up         0 172.30.115.5    172.30.119.107  172.30.112.0/20
```

But it knows about the three addresses on the ECE server

But we aren’t quite there yet

```
# tcpdump -n -i ib0 host ice4305-ib0
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on ib0, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
17:18:16.573357 IP 172.30.119.68.gpfs > 172.30.115.5.44309: Flags [.], seq
2042207819:2042271563, ack 3922665803, win 16384, options [nop,nop,TS val 3057381928 ecr
1012903029], length 63744
17:18:16.573395 IP 172.30.119.68.gpfs > 172.30.115.5.44309: Flags [.], seq 63744:127488, ack 1,
win 16384, options [nop,nop,TS val 3057381928 ecr 1012903029], length 63744
Same on ib1, but on ib2 …

# tcpdump -n -i ib2 host ice4305-ib0
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on ib2, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
^C
0 packets captured
1257 packets received by filter
1150 packets dropped by kernel
```

Traffic from server -> client only on 2 adapter ports
Server still has default maxTcpConnsPerNodeConn = 2

Default connection count is low …

```
# mmlsconfig maxTcpConnsPerNodeConn
maxTcpConnsPerNodeConn 2

TCP Connections between nodes:
    hostname                              node     idx destination     status     err
sock  sent(MB)  recvd(MB)  ostype
    ece4707                               <c0n0>     0 172.30.119.67   connected  0
132   1         1          Linux/L
    ece4707                               <c0n0>     1 172.30.119.87   connected  0
142   1         1          Linux/
…
 <c0n0> 172.30.87.67/0 (ece4707)
    status connected was_broken 0 err 0 reconnEnabled 1 delayedAckEnabled 1
    connMode mrot shutting 0 handlerCount 0 need_notify 0 leaseSentOn 1
    nMaxTcpConns 2 (2) nActiveCount 2 nActiveState 0x3 (110000000000000)
    nInuseTcpConns 0 currTcpConnIndex 1 availableTcpConns (111111111111111)
    nReservedSmallMsgTcpConns 0 currSmallMsgTcpConnIndex 0 currLargeMsgTcpConnIndex 0
    reconnectTcpConns (000000000000000) disconnectTcpConns (000000000000000)
```

```
# mmchconfig maxTcpCons=8

 ice4305                                 <c0n6>     0 172.30.115.5    connected
0    171   0         0          Linux/L
    ice4305                                 <c0n6>     1 172.30.115.5
connected  0    186   0         0          Linux/L
    ice4305                                 <c0n6>     2 172.30.115.5
connected  0    187   0         0          Linux/L
    ice4305                                 <c0n6>     3 172.30.115.5
connected  0    188   0         0          Linux/L

 ece4708                                 <c2n3>     0 172.30.119.68   connected
0    246   0         0          Linux/L
    ece4708                                 <c2n3>     1 172.30.119.88
connected  0    253   0         0          Linux/L
    ece4708                                 <c2n3>     2 172.30.119.108
connected  0    256   0         0          Linux/L
    ece4708                                 <c2n3>     3 172.30.119.68
connected  0    259   0         0          Linux/L
```

From the client -> NSD server

```
 <c2n3> 172.30.87.68/2 (ece4708)
    status connected was_broken 0 err 0 reconnEnabled 1 delayedAckEnabled 1
    connMode mrot shutting 0 handlerCount 0 need_notify 0 leaseSentOn -1
    nMaxTcpConns 4 (4) nActiveCount 4 nActiveState 0xf (1111000000000000)
    nInuseTcpConns 0 currTcpConnIndex 2 availableTcpConns (1111111111111111)
    nReservedSmallMsgTcpConns 0 currSmallMsgTcpConnIndex 0
currLargeMsgTcpConnIndex 0
    reconnectTcpConns (0000000000000000) disconnectTcpConns
(0000000000000000)
    Inuse owner:
```

But also between ece nodes

```
    ece4709                                 <c0n1>     0 172.30.119.109  connected  0    132   0
0           Linux/L
    ece4709                                 <c0n1>     1 172.30.119.69   connected  0    160   0
0           Linux/L
    ece4709                                 <c0n1>     2 172.30.119.89   connected  0    147   0
0           Linux/L
    ece4709                                 <c0n1>     3 172.30.119.109  connected  0    166   0
0           Linux/L
    ece4709                                 <c0n1>     4 172.30.119.69   connected  0    167   0
0           Linux/L
    ece4709                                 <c0n1>     5 172.30.119.89   connected  0    168   0
0           Linux/L
    ece4709                                 <c0n1>     6 172.30.119.109  connected  0    169   0
0           Linux/L
    ece4709                                 <c0n1>     7 172.30.119.69   connected  0    170   0
0           Linux/L

<c0n1> 172.30.87.69/0 (ece4709)
    status connected was_broken 0 err 0 reconnEnabled 1 delayedAckEnabled 1
    connMode mrot shutting 0 handlerCount 0 need_notify 0 leaseSentOn -1
    nMaxTcpConns 8 (8) nActiveCount 8 nActiveState 0xff (1111111100000000)
    nInuseTcpConns 0 currTcpConnIndex 0 availableTcpConns (1111111111111111)
    nReservedSmallMsgTcpConns 0 currSmallMsgTcpConnIndex 0 currLargeMsgTcpConnIndex 0
    reconnectTcpConns (0000000000000000) disconnectTcpConns (0000000000000000)
```

Let's check the network

```
[root@ece4708 ~]# tcpdump -n -i ib1 -c 5 host ice4305-ib0
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on ib1, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
17:27:40.360879 IP 172.30.119.88.36631 > 172.30.115.5.gpfs: Flags [P.], seq 4291401565:4291401613, ack 1186288145, win 16384, options
[nop,nop,TS val 984772677 ecr 3068480326], length 48
17:27:40.361067 IP 172.30.115.5.gpfs > 172.30.119.88.36631: Flags [P.], seq 1:63745, ack 48, win 24566, options [nop,nop,TS val 3068480368
ecr 984772677], length 63744
17:27:40.361147 IP 172.30.115.5.gpfs > 172.30.119.88.36631: Flags [P.], seq 63745:127489, ack 48, win 24566, options [nop,nop,TS val
3068480368 ecr 984772677], length 63744
17:27:40.361195 IP 172.30.115.5.gpfs > 172.30.119.88.36631: Flags [P.], seq 127489:191233, ack 48, win 24566, options [nop,nop,TS val
3068480368 ecr 984772677], length 63744
17:27:40.361243 IP 172.30.115.5.gpfs > 172.30.119.88.36631: Flags [P.], seq 191233:254977, ack 48, win 24566, options [nop,nop,TS val
3068480368 ecr 984772677], length 63744
5 packets captured
3363 packets received by filter
2992 packets dropped by kernel
[root@ece4708 ~]# tcpdump -n -i ib2 -c 5 host ice4305-ib0
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on ib2, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
17:27:44.457100 IP 172.30.115.5.gpfs > 172.30.119.108.33885: Flags [P.], seq 395082837:395082957, ack 4243168737, win 24566, options
[nop,nop,TS val 2067256736 ecr 599583632], length 120
17:27:44.468697 IP 172.30.119.108.33885 > 172.30.115.5.gpfs: Flags [P.], seq 1:49, ack 120, win 16382, options [nop,nop,TS val 599583789 ecr
2067256736], length 48
17:27:44.509335 IP 172.30.115.5.gpfs > 172.30.119.108.33885: Flags [.], ack 49, win 24566, options [nop,nop,TS val 2067256788 ecr
599583789], length 0
17:27:44.622815 IP 172.30.119.108.33885 > 172.30.115.5.gpfs: Flags [P.], seq 49:113, ack 120, win 16382, options [nop,nop,TS val 599583943
ecr 2067256788], length 64
17:27:44.622850 IP 172.30.115.5.gpfs > 172.30.119.108.33885: Flags [.], ack 113, win 24566, options [nop,nop,TS val 2067256901 ecr
599583943], length 0
5 packets captured
1014 packets received by filter
749 packets dropped by kernel
```

Are we done yet?

```
# mmlsconfig maxReceiverThreads
maxReceiverThreads 32


# mmchconfig maxReceiverThreads=128
```

- Some large clusters need to increase the value of maxReceiverThreads based on the number of TCP connections that will be needed to other nodes in both local clusters and remote clusters that are joined. The total number of TCP connections that are required is calculated by using the following formula: (maxTcpConnsPerNodeConn*(number of nodes - 1)).

- The maximum number of receiver threads that are created on any node is defined to be the minimum of the number of logical CPUs on the node and the value of the maxReceiverThreads parameter. You can specify a value in the 1-128 range for the maxReceiverThreads parameter, with the default value being 32.

ECE performance comparison

	Write	Read
3x HDR RDMA	102168.90 MB/sec	128149.30 MB/sec
3x IPoIB maxcon=2	32857.13 MB/sec	57090.77 MB/sec
3x IPoIB maxcon=8	35634.81 MB/sec	58113.13 MB/sec
3x IPoIB maxcon=8, 128	36125.54 MB/sec	59903.05 MB/sec

OK but what about DSS-G?

	Write	Read
3x HDR-RDMA	33350.24 MB/sec	79144.08 MB/sec
3x IPoIB + MROT	34055.90 MB/sec	61075.91 MB/sec
2x IPoIB + MROT	30468.80 MB/sec	56331.43 MB/sec
1x IPoIB	34312.15 MB/sec	22498.53 MB/sec

Multiple interfaces without MROT, or when using bonding could give performance of 1x IPoIB

And doing away with hooky routing in mixed fabrics …

```
5.1.8-0 was quite unstable configured like this
5.1.8-2 was significantly better
# mmlscluster

GPFS cluster information
========================
  GPFS cluster name:         g100sr630v2.gpfs.hpc.eu.lenovo.com
  GPFS cluster id:           17007464881994050140
  GPFS UID domain:           g100sr630v2.gpfs.hpc.eu.lenovo.com
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp
  Repository type:           CCR

 Node  Daemon node name  IP address    Admin node name  Designation
-------------------------------------------------------------------
   1   ece4707           172.30.87.67  ece4707          quorum-manager
   2   ece4708           172.30.87.68  ece4708          quorum-manager
   3   ece4709           172.30.87.69  ece4709          quorum-manager
   4   ece4710           172.30.87.70  ece4710          quorum-manager
   5   ece4711           172.30.87.71  ece4711          quorum-manager
   6   ece4712           172.30.87.72  ece4712          quorum-manager
```

This is the OS mgmt. interface yet Scale is using the MROT addresses internally

Smarter technology for all

Lenovo

thanks.