

Network considerations for Multi-tenancy and AI

Storage Scale User Meeting @ ISC26

Ralph Würthner
Storage Scale Development



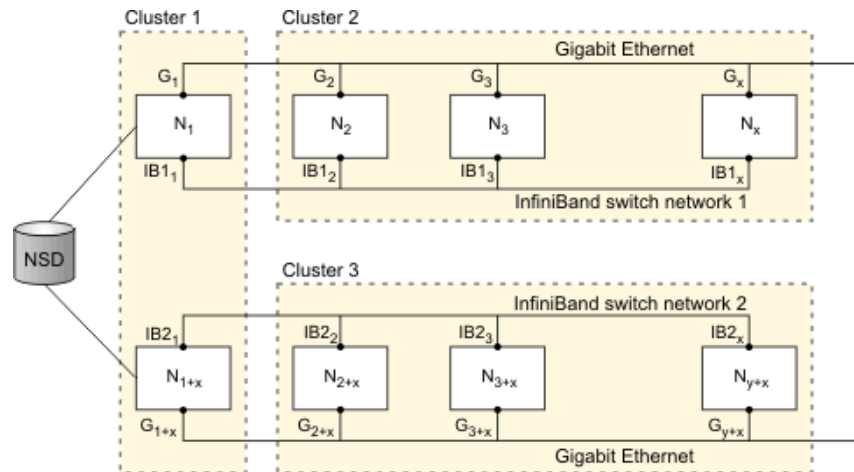
Network considerations for Multi-tenancy and AI

- Network setup for Multi-tenancy & remote cluster
- Network configuration for RoCE Deployment
- How to avoid asymmetric traffic on RDMA Ports

Network setup for Multi-tenancy & remote cluster

Important Rules

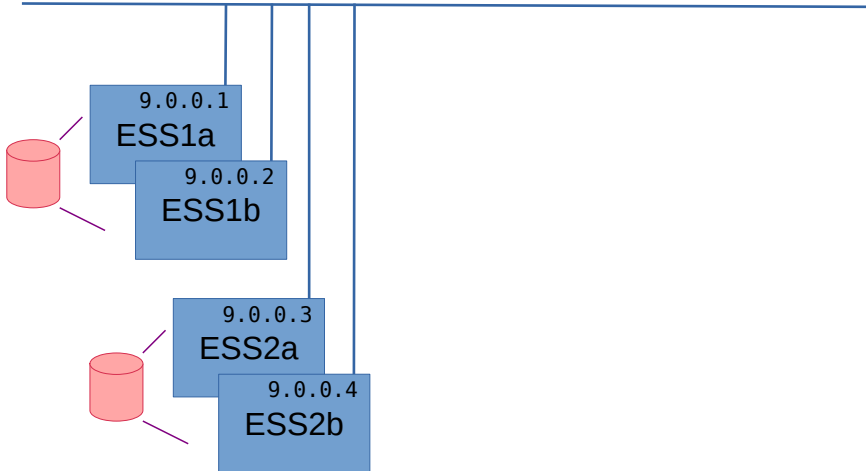
- When adding NICs, interface configuration changes, new IPs etc. to a GPFS node, you have to restart GPFS to make the GPFS daemon aware about the changes.
- GPFS on each client needs to be able to initially connect to the nodes in storage cluster on their daemon IP, which is stored in CCR (in cluster config). See administration guide, chapter Accessing a remote GPFS File system.



Storage Cluster

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@ess1a ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      storage.corp.com
GPFS cluster id:       13707083046693220691
GPFS UID domain:       storage.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:       CCR
```

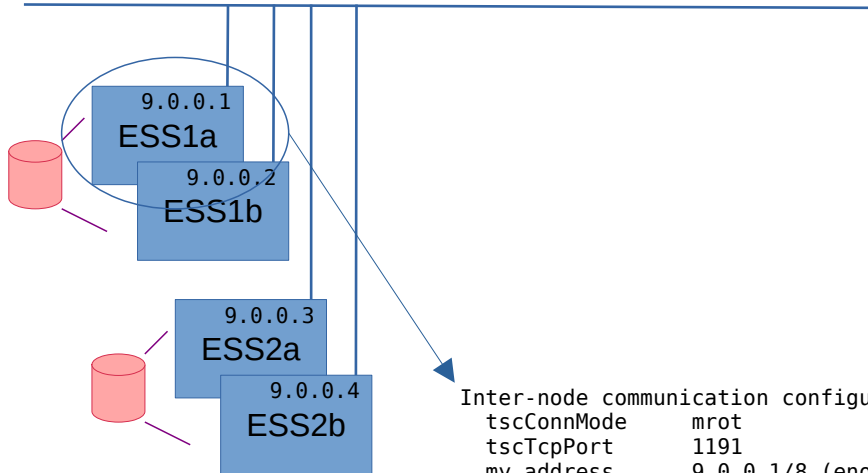
Node	Daemon node name	IP address	Admin node name	Designation
1	ess1a.corp.com	9.0.0.1	ess1a-a.corp.com	quorum-manager
2	ess1b.corp.com	9.0.0.2	ess1b-a.corp.com	manager
6	ess2a.corp.com	9.0.0.3	ess2a-a.corp.com	manager
7	ess2b.corp.com	9.0.0.4	ess2b-a.corp.com	manager

```
[root@ess1a ~]#
```

Storage Cluster

Corporate network used for daemon

Corporate network 9.0.0.0/8



Inter-node communication configuration:

```

tscConnMode    mrot
tscTcpPort     1191
my address     9.0.0.1/8 (enol) <c0n0>
my addr list   9.0.0.1/8 (enol)
my subnet list (none)
my node number 1

```

TCP Connections between nodes:

hostname	node	idx	destination	status	err	sock	...
ess1b	<c0n1>	0	9.0.0.2	connected	0	152	
ess1b	<c0n1>	1	9.0.0.2	connected	0	153	
ess2a	<c0n2>	0	9.0.0.3	connected	0	145	
ess2a	<c0n2>	1	9.0.0.3	connected	0	149	
ess2b	<c0n3>	0	9.0.0.4	connected	0	151	
ess2b	<c0n3>	1	9.0.0.4	connected	0	150	

```
[root@ess1a ~]# mmlscluster
```

GPFS cluster information

```

=====
GPFS cluster name:      storage.corp.com
GPFS cluster id:       13707083046693220691
GPFS UID domain:       storage.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:       CCR

```

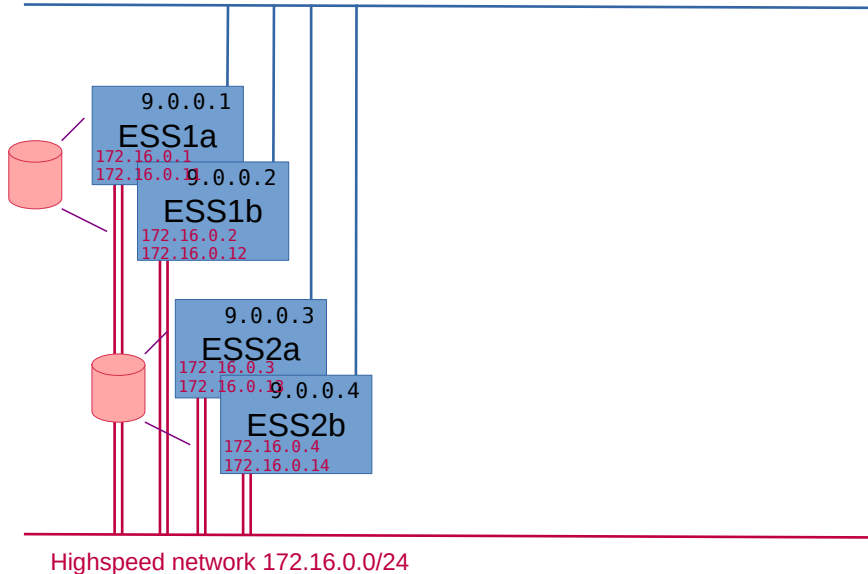
Node	Daemon node name	IP address	Admin node name	Designation
1	ess1a.corp.com	9.0.0.1	ess1a-a.corp.com	quorum-manager
2	ess1b.corp.com	9.0.0.2	ess1b-a.corp.com	manager
6	ess2a.corp.com	9.0.0.3	ess2a-a.corp.com	manager
7	ess2b.corp.com	9.0.0.4	ess2b-a.corp.com	manager

```
[root@ess1a ~]#
```

Storage Cluster

Corporate network used for daemon

Corporate network 9.0.0.0/8



mmchconfig subnets="172.16.0.0"

```
[root@ess1a ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      storage.corp.com
GPFS cluster id:       13707083046693220691
GPFS UID domain:      storage.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:      CCR
```

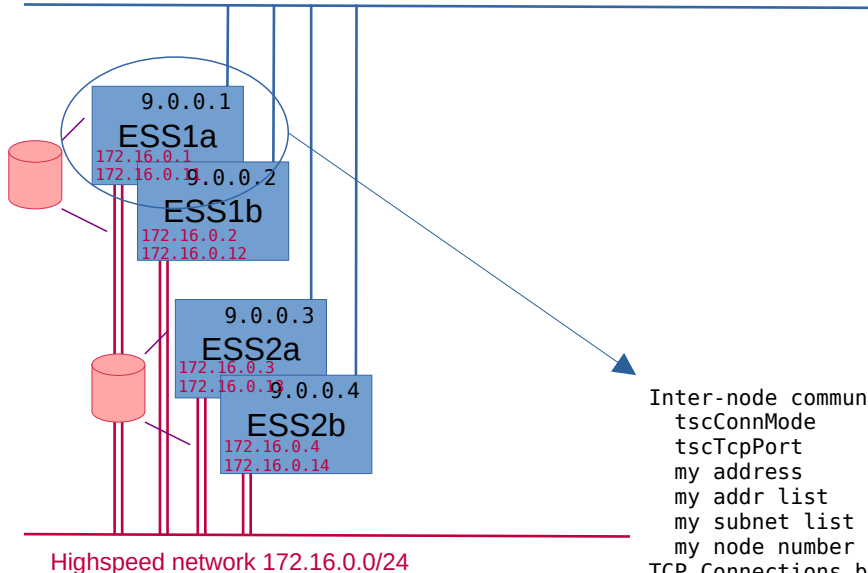
Node	Daemon node name	IP address	Admin node name	Designation
1	ess1a.corp.com	9.0.0.1	ess1a-a.corp.com	quorum-manager
2	ess1b.corp.com	9.0.0.2	ess1b-a.corp.com	manager
6	ess2a.corp.com	9.0.0.3	ess2a-a.corp.com	manager
7	ess2b.corp.com	9.0.0.4	ess2b-a.corp.com	manager

```
[root@ess1a ~]#
```

Storage Cluster

Corporate network used for daemon

Corporate network 9.0.0.0/8



mmchconfig subnets="172.16.0.0"

```
[root@ess1a ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      storage.corp.com
GPFS cluster id:        13707083046693220691
GPFS UID domain:        storage.corp.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	ess1a.corp.com	9.0.0.1	ess1a-a.corp.com	quorum-manager
2	ess1b.corp.com	9.0.0.2	ess1b-a.corp.com	manager
6	ess2a.corp.com	9.0.0.3	ess2a-a.corp.com	manager
7	ess2b.corp.com	9.0.0.4	ess2b-a.corp.com	manager

```
[root@ess1a ~]#
```

Inter-node communication configuration:

```
tscConnMode      mrot
tscTcpPort        1191
my address         9.0.0.1/8 (eno1) <c0n0>
my addr list       172.16.0.1/24 (ib0)/24  172.16.0.11/24 (ib1)/24  9.0.0.1/8 (eno1)
my subnet list     172.16.0.0/24
my node number     1
```

TCP Connections between nodes:

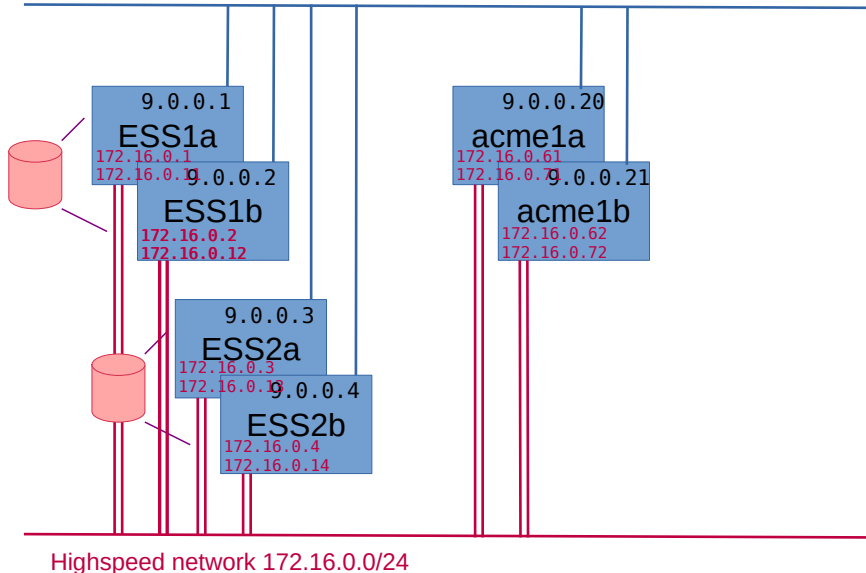
hostname	node	idx	destination	status	err	sock ...
ess1b	<c0n1>	0	172.16.0.12	connected	0	157
ess1b	<c0n1>	1	172.16.0.2	connected	0	158
ess2a	<c0n2>	0	172.16.0.3	connected	0	154
ess2a	<c0n2>	1	172.16.0.13	connected	0	155
ess2b	<c0n3>	0	172.16.0.4	connected	0	159
ess2b	<c0n3>	1	172.16.0.14	connected	0	160

Only highspeed network is used!

Adding Tenant Cluster acme1

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@acmela ~]# mmlscluster
```

GPFS cluster information

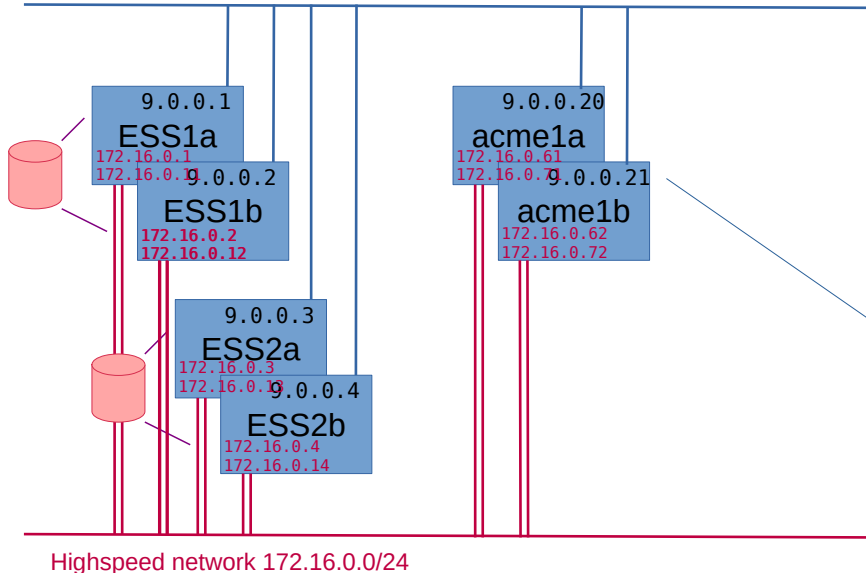
```
=====
GPFS cluster name:      acme1.corp.com
GPFS cluster id:       32700993567110668
GPFS UID domain:      acme1.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:      CCR
```

Node	Daemon node name	IP address	Admin node name	Designation
1	acmela.com	9.0.0.20	acmela-a.corp.com	quorum-manager
2	acme1b.com	9.0.0.21	acme1b-a.corp.com	manager

Updating subnets on Tenant Cluster acme1

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@acmela ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      acme1.corp.com
GPFS cluster id:       32700993567110668
GPFS UID domain:      acme1.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:      CCR
```

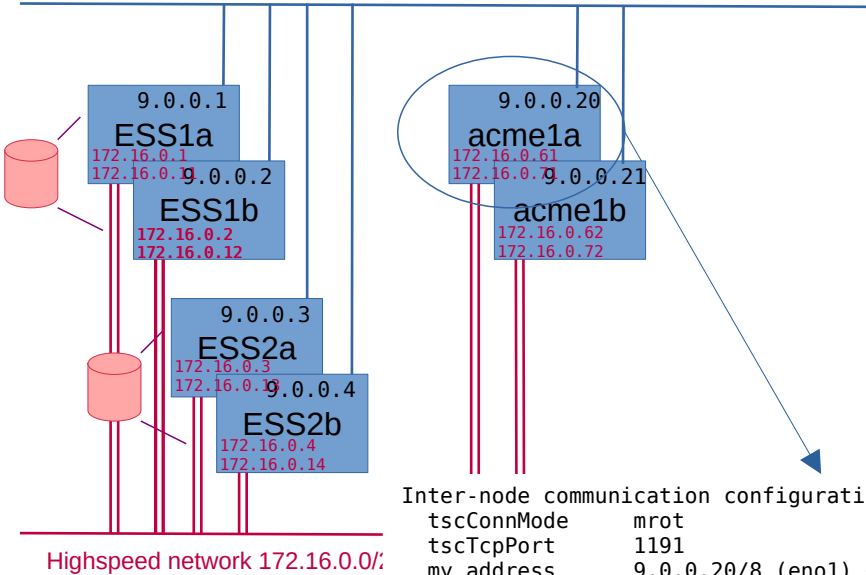
Node	Daemon node name	IP address	Admin node name	Designation
1	acmela.com	9.0.0.20	acmela-a.corp.com	quorum-manager
2	acmelb.com	9.0.0.21	acmelb-a.corp.com	manager

```
[root@acmela ~]# mmlsconfig subnets
subnets 172.16.0.0/storage.corp.com
```

Updating subnets on Tenant Cluster acme1

Corporate network used for daemon

Corporate network 9.0.0.0/8



Inter-node communication configuration:

```
tscConnMode    mrot
tscTcpPort     1191
my address     9.0.0.20/8 (eno1) <c0n0>
my addr list   172.16.0.61/24 (ib0)/storage.corp.com 172.16.0.71/24 (ib1)/storage.corp.com 9.0.0.20/8 (eno1)
my subnet list 172.16.0.0/24
my node number 1
```

TCP Connections between nodes:

hostname	node	idx	destination	status	err	sock ...
acme1a	<c0n1>	0	172.16.0.62	connected	0	108
acme1b	<c0n1>	1	172.16.0.72	connected	0	167

```
[root@acme1a ~]# mmlscluster
```

GPFS cluster information

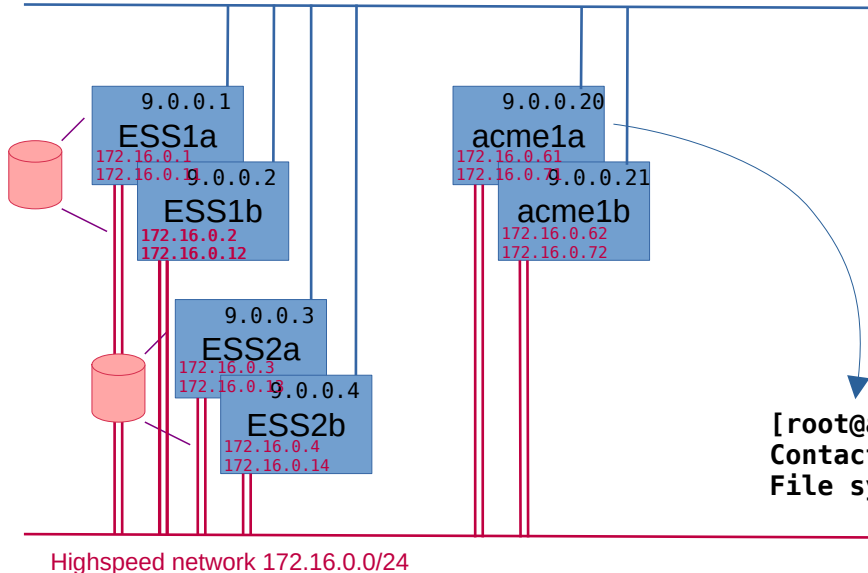
```
=====
GPFS cluster name:      acme1.corp.com
GPFS cluster id:       32700993567110668
GPFS UID domain:       acme1.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:       CCR
```

Node	Daemon node name	IP address	Admin node name	Designation
1	acme1a.com	9.0.0.20	acme1a-a.corp.com	quorum-manager
2	acme1b.com	9.0.0.21	acme1b-a.corp.com	manager

Adding Remote File System gpfs0 to acme1

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@acmela ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      acme1.corp.com
GPFS cluster id:       32700993567110668
GPFS UID domain:      acme1.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:      CCR
```

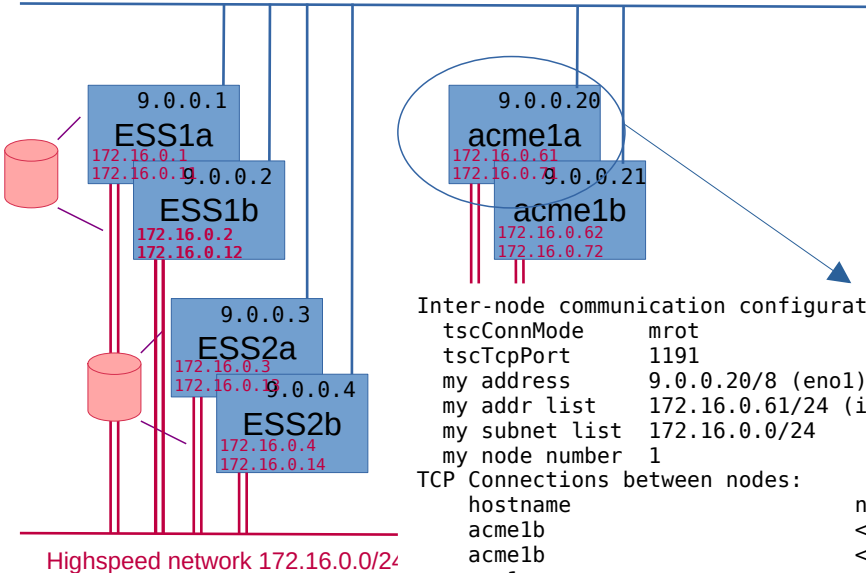
Node	Daemon node name	IP address	Admin node name	Designation
1	acmela.com	9.0.0.20	acmela-a.corp.com	quorum-manager
2	acme1b.com	9.0.0.21	acme1b-a.corp.com	manager

```
[root@acmela ~]# mmremotecluster show all | grep -e Contact -e File
Contact nodes:  ess1a,ess1b
File systems:   gpfs0 (gpfs0)
```

Mounting remote File System gpfs0 on acme1

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@acmela ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      acme1.corp.com
GPFS cluster id:        32700993567110668
GPFS UID domain:        acme1.corp.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	acmela.com	9.0.0.20	acmela-a.corp.com	quorum-manager
2	acme1b.com	9.0.0.21	acme1b-a.corp.com	manager

Inter-node communication configuration:

```
tscConnMode      mrot
tscTcpPort       1191
my address       9.0.0.20/8 (eno1) <c0n0>
my addr list     172.16.0.61/24 (ib0)/storage.corp.com 172.16.0.71/24 (ib1)/storage.corp.com 9.0.0.20/8 (eno1)
my subnet list   172.16.0.0/24
my node number   1
```

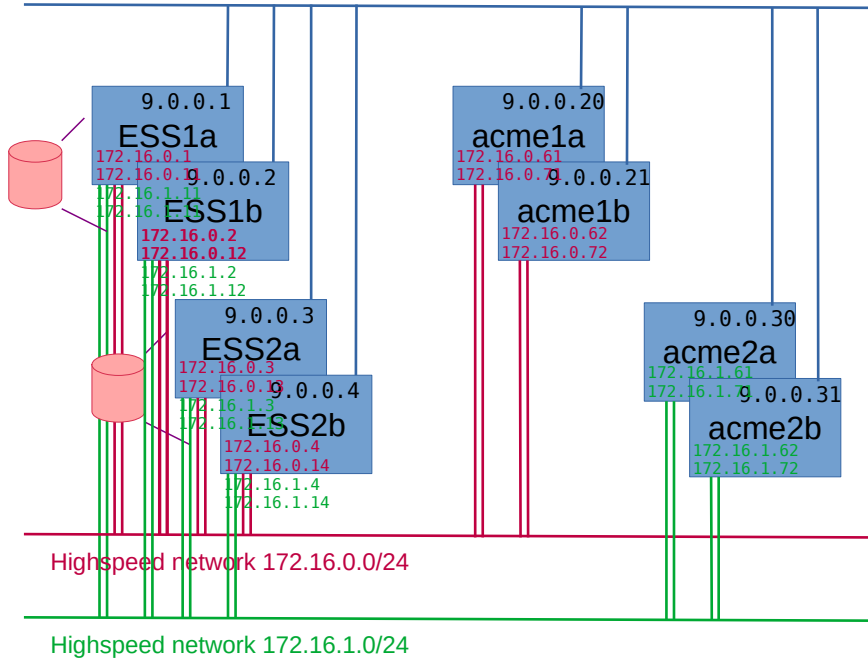
TCP Connections between nodes:

hostname	node	idx	destination	status	err	sock ...
acme1b	<c0n1>	0	172.16.0.62	connected	0	108
acme1b	<c0n1>	1	172.16.0.72	connected	0	167
ess1a	<c1n0>	0	172.16.0.1	connected	0	177
ess1a	<c1n0>	1	172.16.0.11	connected	0	178
ess1b	<c1n1>	0	172.16.0.2	connected	0	176
ess1b	<c1n2>	1	172.16.0.12	connected	0	183
ess2a	<c1n2>	0	172.16.0.3	connected	0	179
ess2a	<c1n2>	1	172.16.0.13	connected	0	180
ess2b	<c1n3>	0	172.16.0.4	connected	0	181
ess2b	<c1n3>	1	172.16.0.14	connected	0	182

Adding Tenant Cluster acme2

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@acme2a ~]# mmlscluster
```

GPFS cluster information

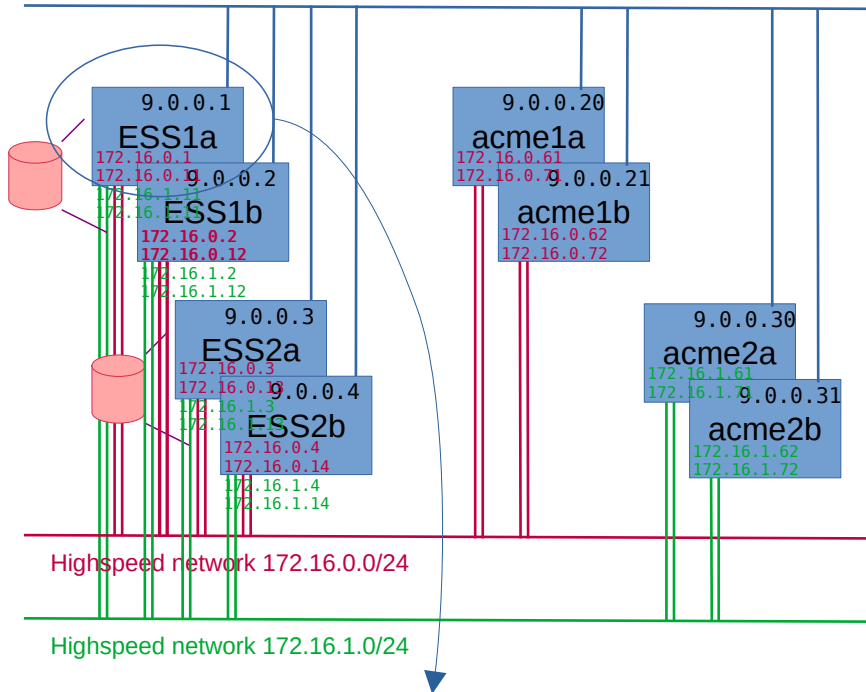
```
=====
GPFS cluster name:      acme2.corp.com
GPFS cluster id:       45241136254559417
GPFS UID domain:      acme2.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:       CCR
```

Node	Daemon node name	IP address	Admin node name	Designation
1	acme2a.com	9.0.0.30	acme2a-a.corp.com	quorum-manager
2	acme2b.com	9.0.0.31	acme2b-a.corp.com	manager

Updating subnets on Storage Cluster

Corporate network used for daemon

Corporate network 9.0.0.0/8



mmchconfig subnets="172.16.0.0 172.16.1.0"

```
[root@acme2a ~]# mmlscluster
```

GPFS cluster information

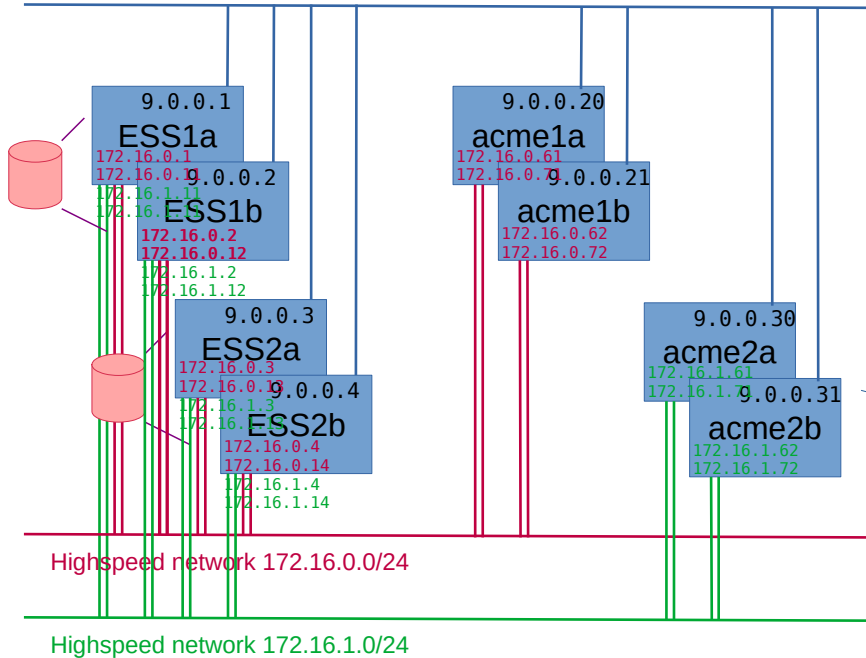
```
=====
GPFS cluster name:      acme2.corp.com
GPFS cluster id:       45241136254559417
GPFS UID domain:      acme2.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:      CCR
```

Node	Daemon node name	IP address	Admin node name	Designation
1	acme2a.com	9.0.0.30	acme2a-a.corp.com	quorum-manager
2	acme2b.com	9.0.0.31	acme2b-a.corp.com	manager

Updating subnets on Tenant Cluster acme2

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@acme2a ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      acme2.corp.com
GPFS cluster id:       45241136254559417
GPFS UID domain:      acme2.corp.com
Remote shell command:  /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:       CCR
```

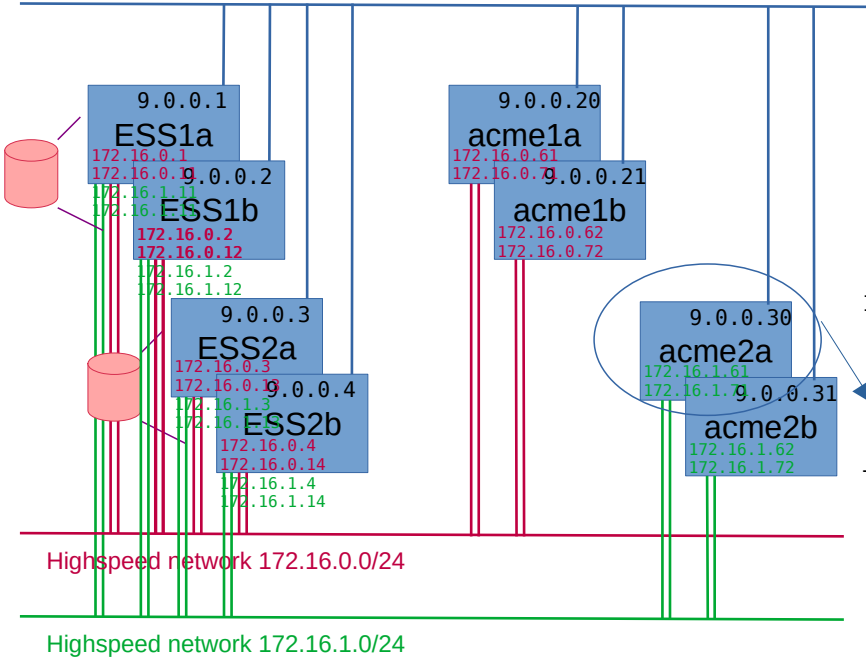
Node	Daemon node name	IP address	Admin node name	Designation
1	acme2a.com	9.0.0.30	acme2a-a.corp.com	quorum-manager
2	acme2b.com	9.0.0.31	acme2b-a.corp.com	manager

[root@acme2a ~]# mmlsconfig subnets
subnets 172.16.1.0/storage.corp.com

Mounting remote File System gpfs1 on acme1

Corporate network used for daemon

Corporate network 9.0.0.0/8



```
[root@acme2a ~]# mmlscluster
```

GPFS cluster information

```
=====
GPFS cluster name:      acme2.corp.com
GPFS cluster id:       45241136254559417
GPFS UID domain:       acme2.corp.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	acme2a.com	9.0.0.30	acme2a-a.corp.com	quorum-manager
2	acme2b.com	9.0.0.31	acme2b-a.corp.com	manager

Inter-node communication configuration:

```
tscConnMode      mrot
tscTcpPort       1191
my address       9.0.0.30/8 (eno1) <c0n0>
my addr list     172.16.1.61/24 (ib0)/storage.corp.com 172.16.1.71/24 (ib1)/st
my subnet list   172.16.1.0/24
my node number   1
```

TCP Connections between nodes:

hostname	node	idx	destination	status	err	s
acme1b	<c0n1>	0	172.16.1.62	connected	0	1
acme1b	<c0n1>	1	172.16.1.72	connected	0	1
ess1a	<c1n0>	0	172.16.1.1	connected	0	1
ess1a	<c1n0>	1	172.16.1.11	connected	0	1
ess1b	<c1n1>	0	172.16.1.2	connected	0	1
ess1b	<c1n2>	1	172.16.1.12	connected	0	1
ess2a	<c1n2>	0	172.16.1.3	connected	0	1
ess2a	<c1n2>	1	172.16.1.13	connected	0	1
ess2b	<c1n3>	0	172.16.1.4	connected	0	1
ess2b	<c1n3>	1	172.16.1.14	connected	0	1

Summary

- Different tenant clusters operate on different networks
- Common network still required for initial cluster connect
- Works for RDMA too by using different virtual fabric numbers for each tenant

Network configuration for RoCE Deployment

Network configuration for RoCE Deployment

Fabric Attributes			
<i>Fabric Configuration</i>	<i>Lossy fabric</i>	<i>Lossy fabric with QoS</i>	<i>Lossless fabric</i>
Recommended for	Ease of deployment Minimal fabric configuration	Large scale with mixed traffic (TCP/UDP and RoCE)	Uncompromised performance over large scale Storage deployment with heavy back-pressure from PCI*
Trust layer for QoS/PFC**	N/A	L3 (DSCP)	L3 (DSCP)
VLAN required	No	No	No
ECN (congestion control)	Yes	Yes	Yes
PFC (flow control)	No	No	Yes

Trust Layer: Which data is used to classify packets:

- L2: PCP bits on VLAN header
- L3: DSCP bits on IP header

IBM Redpaper:

Highly Efficient Data Access with RoCE on IBM Elastic Storage Systems and IBM Spectrum Scale
<https://www.redbooks.ibm.com/abstracts/redp5658.html>

Host configuration with DOCA OFED / Enable PFC on RoCE priority

Method 1 - Using `mlnx_qos` tool (Note: This command is nonpersistent):

```
# mlnx_qos -i <interface> --pfc 0,0,0,1,0,0,0,0
```

Method 2 - Using LLDP DCBX, and configuration in the switch (Note: This requires LLDP service to be enabled in the switch)

```
# mlxconfig -d /dev/mst/mt4115_pciconf0 -y s LLDP_NB_DCBX_P1=TRUE  
LLDP_NB_TX_MODE_P1=2 LLDP_NB_RX_MODE_P1=2 LLDP_NB_DCBX_P2=TRUE  
LLDP_NB_TX_MODE_P2=2 LLDP_NB_RX_MODE_P2=2
```

Host configuration with DOCA OFED / Set QoS parameters priority

Set DSCP (L3) as trust mode for the NIC

```
# mlx_qos -i <interface> --trust dscp
```

Set ToS to 106 (DSCP 26, AF31) for RDMA-CM traffic* (Note: This command is persistent)

```
# mmchconfig verbsRdmaRoCEToS=106
```

[Alternative] Set the RDMA-CM ToS to 106 (DSCP 26, AF31) (Note: This command is nonpersistent)

```
# cma_roce_tos -d <mlx-device> -t 106
```

[Alternative] Set ToS to 106 (DSCP 26, AF31) for ALL RoCE traffic (Note: This command is nonpersistent)

```
# echo 106 > /sys/class/infiniband/<mlx-device>/tc/1/traffic_class
```

* Requires Storage Scale 5.2.3.7 or higher, or 6.0.1.0 or higher (APAR IJ57341)

Host configuration with DOCA OFED / Set ECN

[Optional] Enable ECN for all (in- & out-going) TCP traffic (Note: This command is nonpersistent)

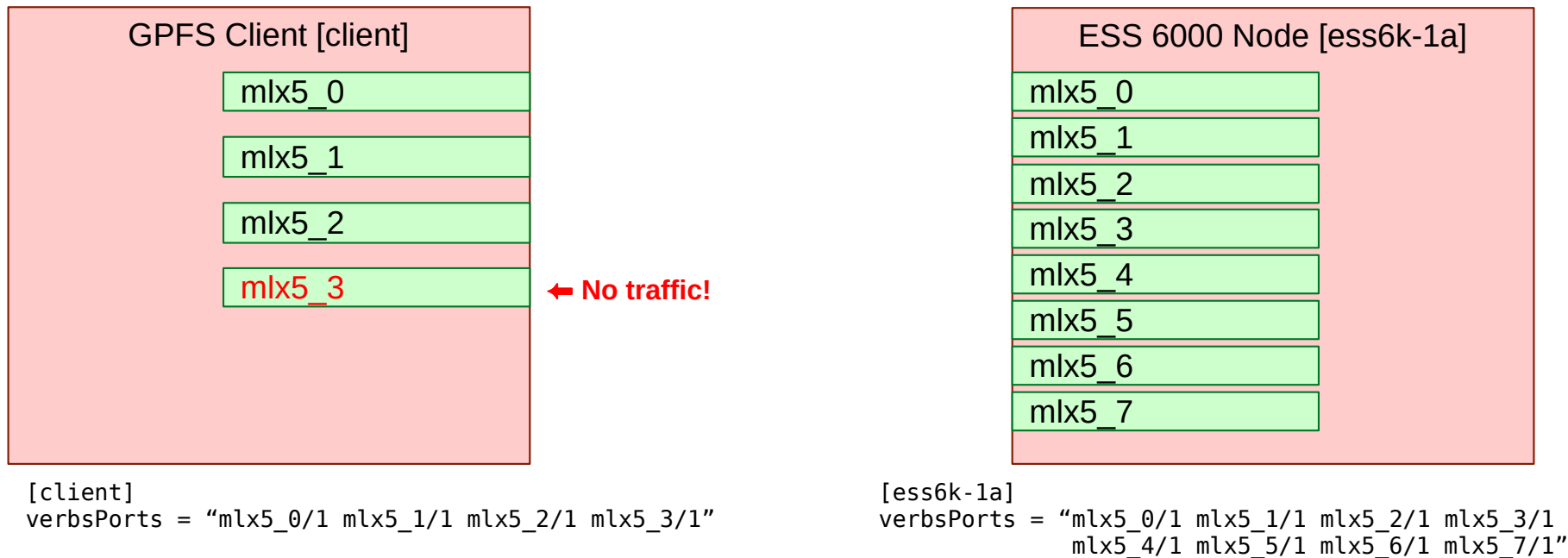
```
# sysctl -w net.ipv4.tcp_ecn=1
```

How to avoid asymmetric traffic on RDMA Ports

Best Practices: Asymmetric Traffic on RDMA Ports

Customer system with 4 RDMA ports on GPFS clients & 8 RDMA ports on ESS 6000 nodes

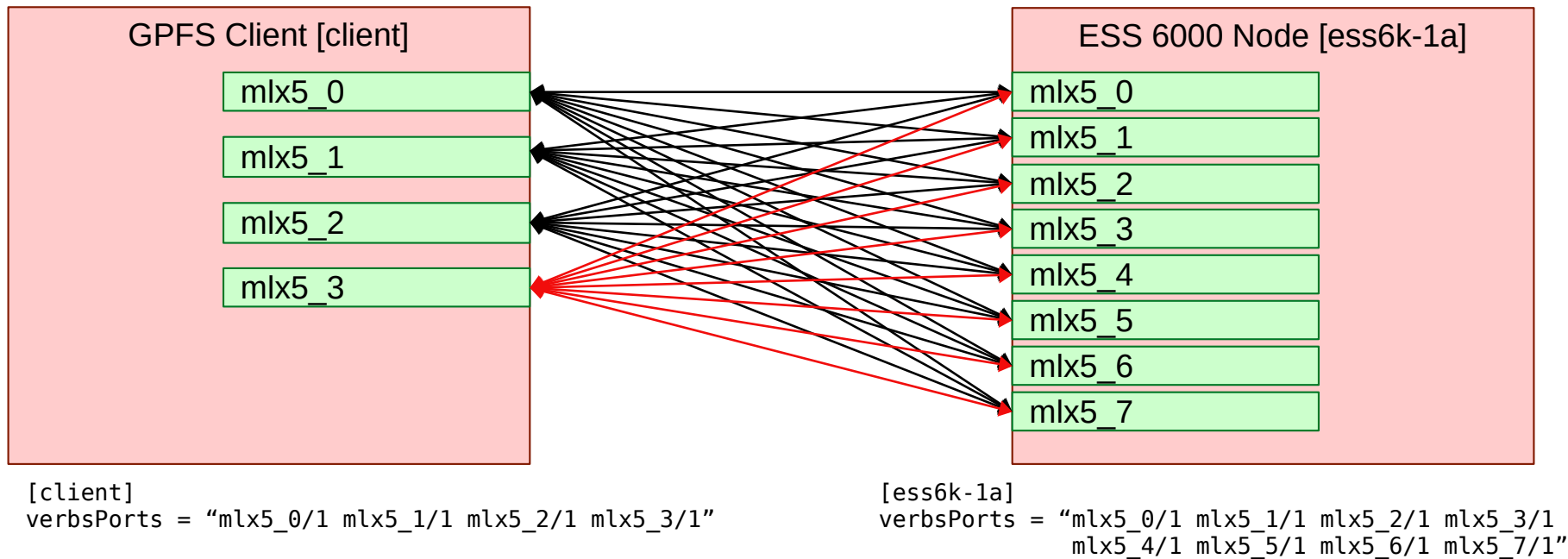
→ No Traffic on client port `mlx5_3`!



Best Practices: Asymmetric Traffic on RDMA Ports

Different number of RDMA ports on both sides: N x M connection model

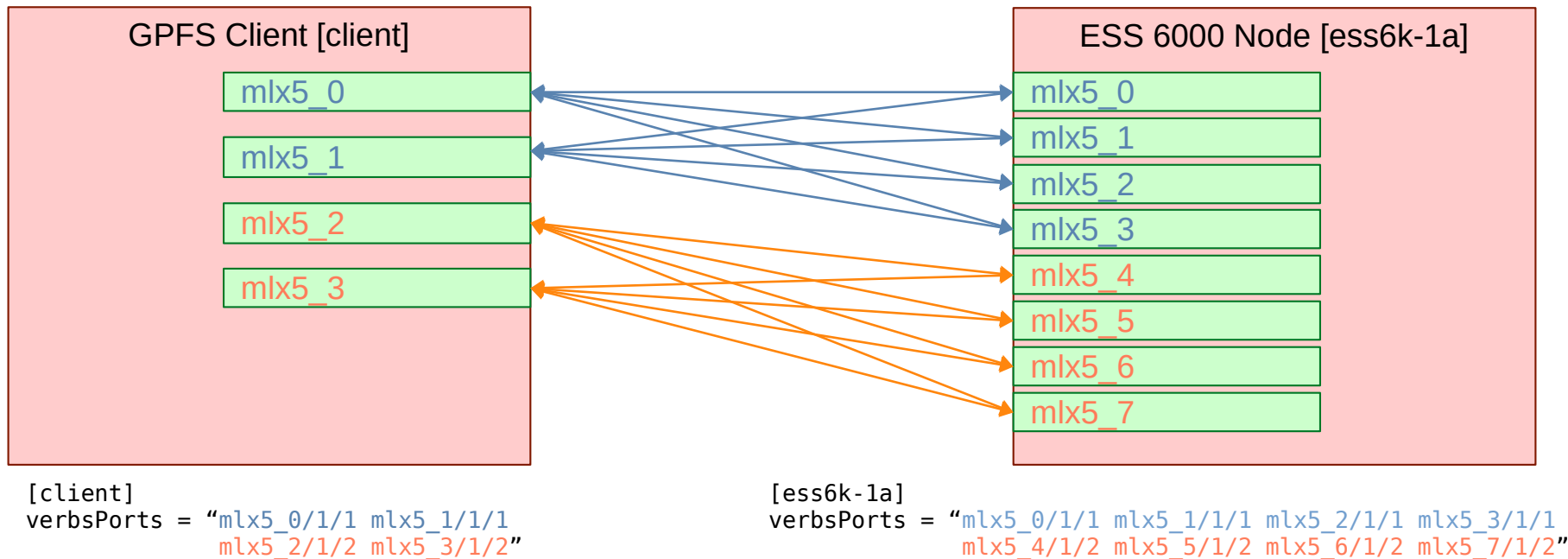
- 4 RDMA ports * 8 RDMA ports → 32 RDMA connections required
- Maximum number of RDMA connections is limited to 24! 😞



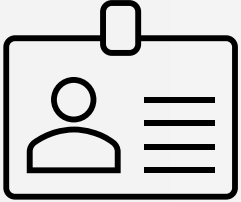
Best Practices: Asymmetric Traffic on RDMA Ports

Solution: Divide RDMA ports into two groups via virtual fabrics

- 2 times * 2 RDMA ports * 4 RDMA ports → 16 RDMA connections
- Total number is below RDMA connection limit of 24 ✓



Thank you

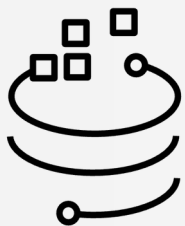


Ralph Würthner

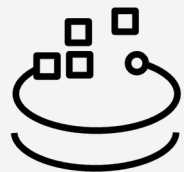
<mailto:ralph.wuerthner@de.ibm.com>



Thank you for using



Storage Scale



Storage Scale
System

