GPFS Multi-Cluster Routing HOWTO *v1*

A Visual HOWTO

Prepared by:
Brian Finley, IBM

GPFS Multi-Cluster

Well, I'm making the presumption that you already know what multi-cluster is, and that you're just trying to assess how to make it work in a certain scenario.

This visual HOWTO is intended to help you to do just that.



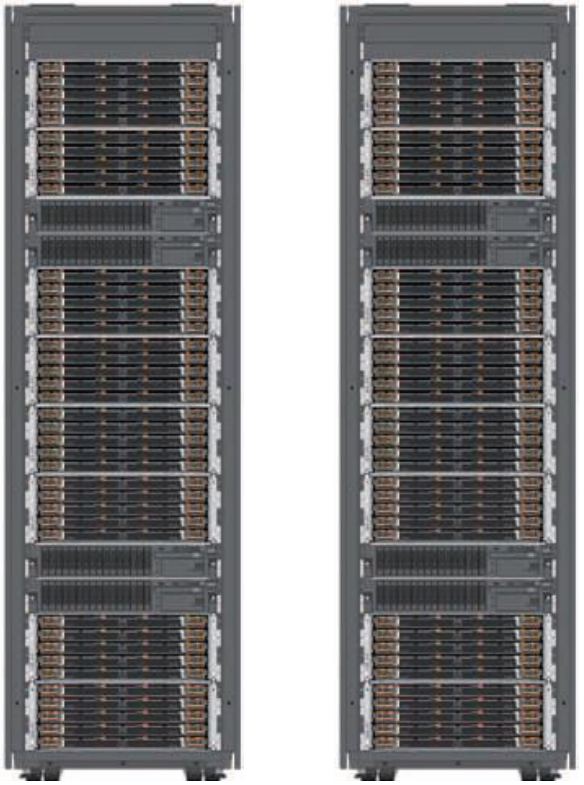
Fundamentals for Multi-Cluster

Required

- Two (or more) GPFS clusters
- Within each pair of connected clusters, you must have consistent name resolution for the GPFS daemon-interfaces for all nodes.
- Within each pair of connected clusters, all nodes must be able to route to all other nodes' GPFS daemon-interfaces via IP.
- Each GPFS cluster must maintain it's own quorum.

Optional

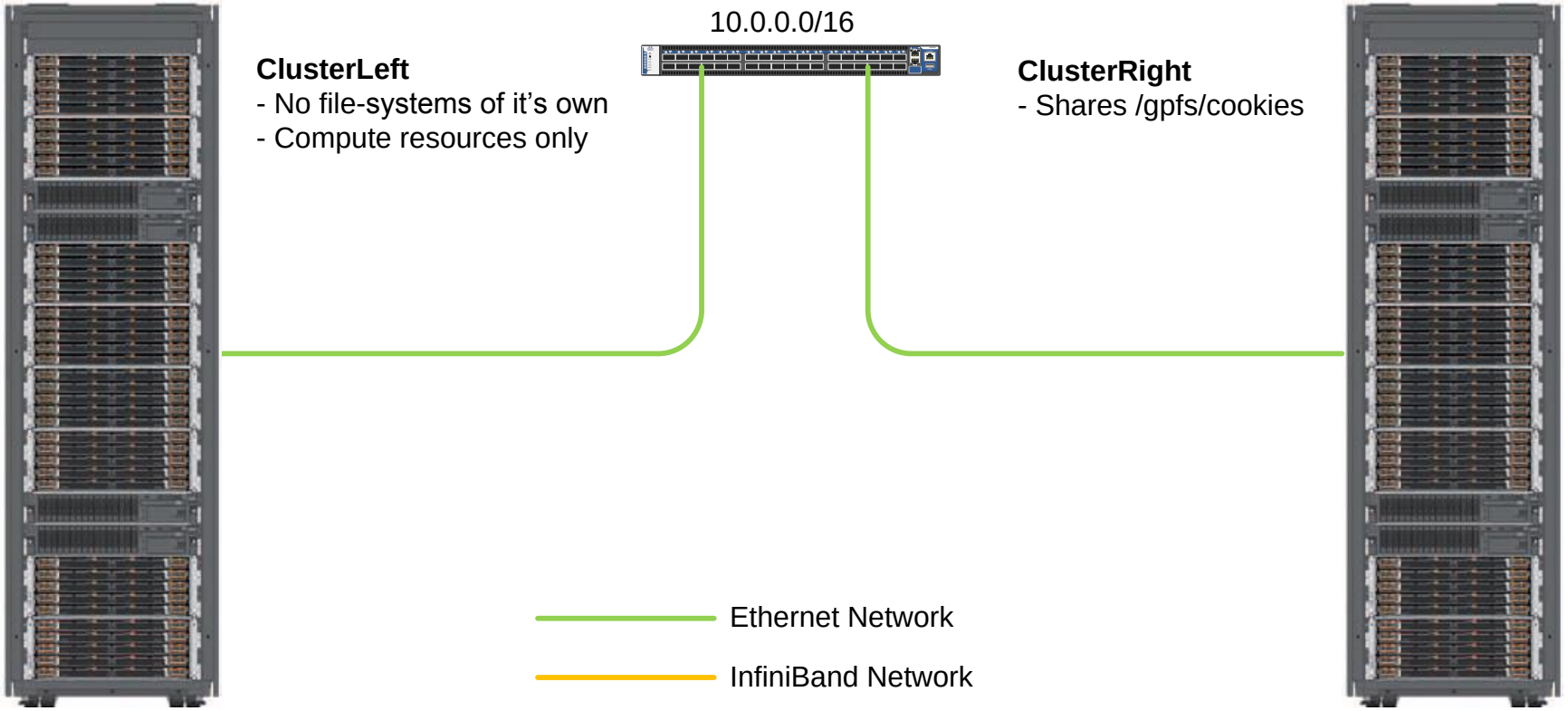
- Pairs of GPFS clusters may be on the same InfiniBand fabric, in which case RDMA can be used for ***data*** communications, even from one cluster to the other.
- It's OK if one or more of the GPFS clusters is a client only cluster, with no storage of it's own.
- IP traffic for either administrative or data communication between clusters may be passed over *any* physical layer technology, although faster is certainly better.



A Very Simple Topology

Two GPFS clusters with One IP Network over Ethernet

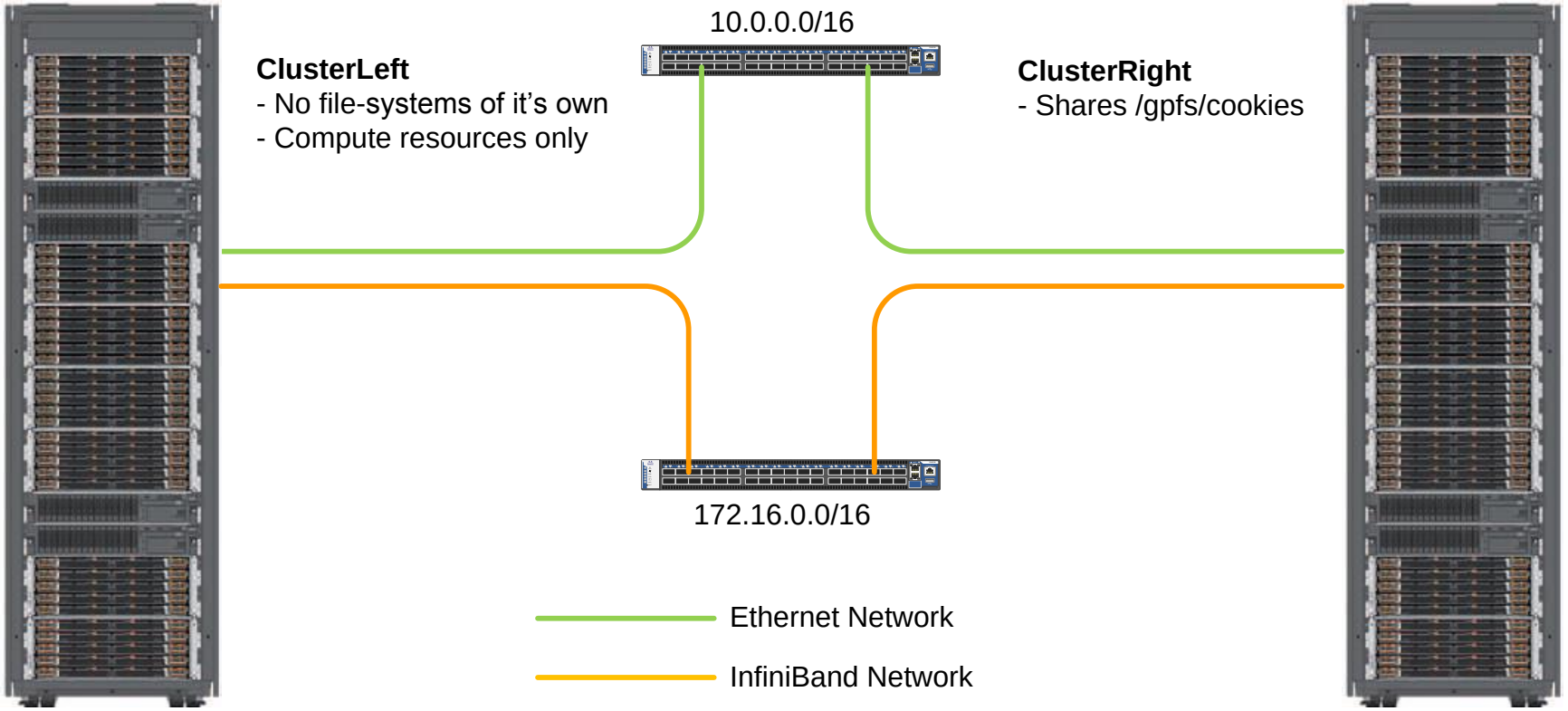
- Both clusters have their main OS and GPFS daemon-interfaces on the 10.0.0.0/16 network.
- Both clusters use IP for administrative *and* data traffic.
- ClusterLeft mounts the *cookies* file-system from ClusterRight.



A Slightly Less Simple Topology

Two GPFS clusters with One IP Network over **InfiniBand**

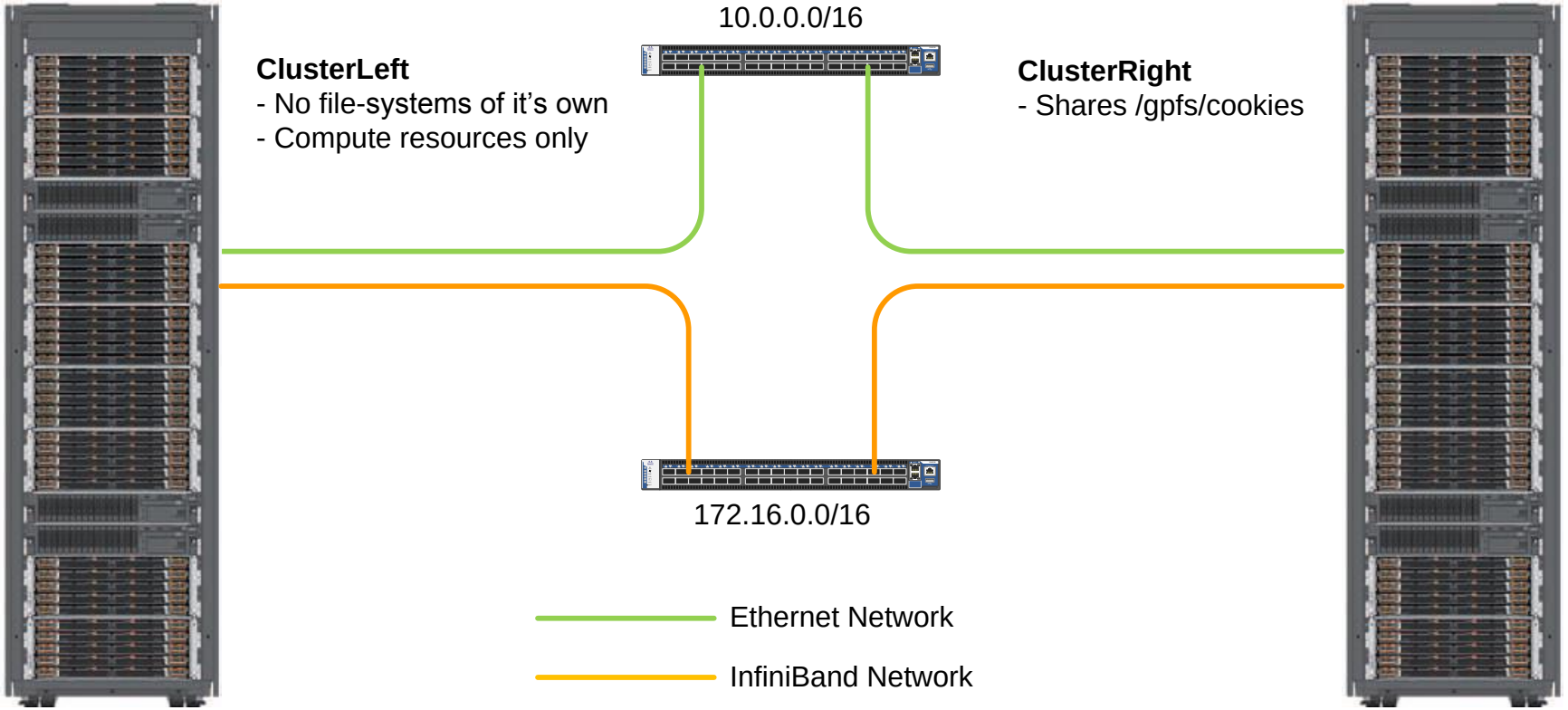
- Both clusters use 10.0.0.0/16 as their main OS network.
- **Both clusters have their GPFS daemon-interfaces on the 172.16.0.0/16 network.**
- **Both clusters use IP over 172.16.0.0/16 for administrative and data traffic.**
- ClusterLeft mounts the *cookies* file-system from ClusterRight.



Still Simple but with Better Performance

Two GPFS clusters with **Two** IP Networks (one for GPFS)

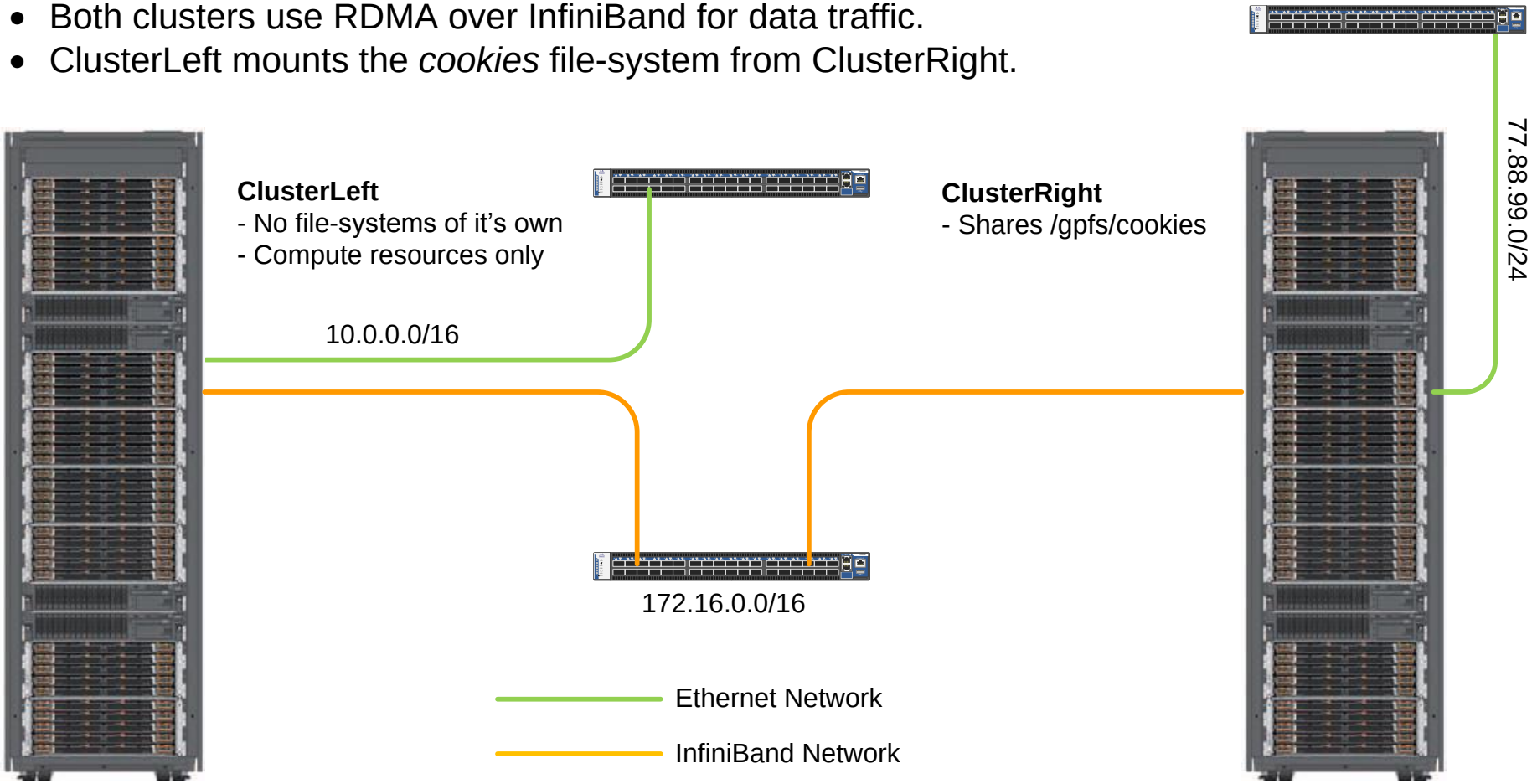
- Both clusters use 10.0.0.0/16 as their main OS network.
- Both clusters have their GPFS daemon-interfaces on the 172.16.0.0/16 network.
- Both clusters use IP over 172.16.0.0/16 for administrative traffic.
- **Both clusters use RDMA over InfiniBand for data traffic.**
- ClusterLeft mounts the *cookies* file-system from ClusterRight.



Still Simple but Three Nets

Two GPFS clusters with **Three** IP Networks (one for GPFS)

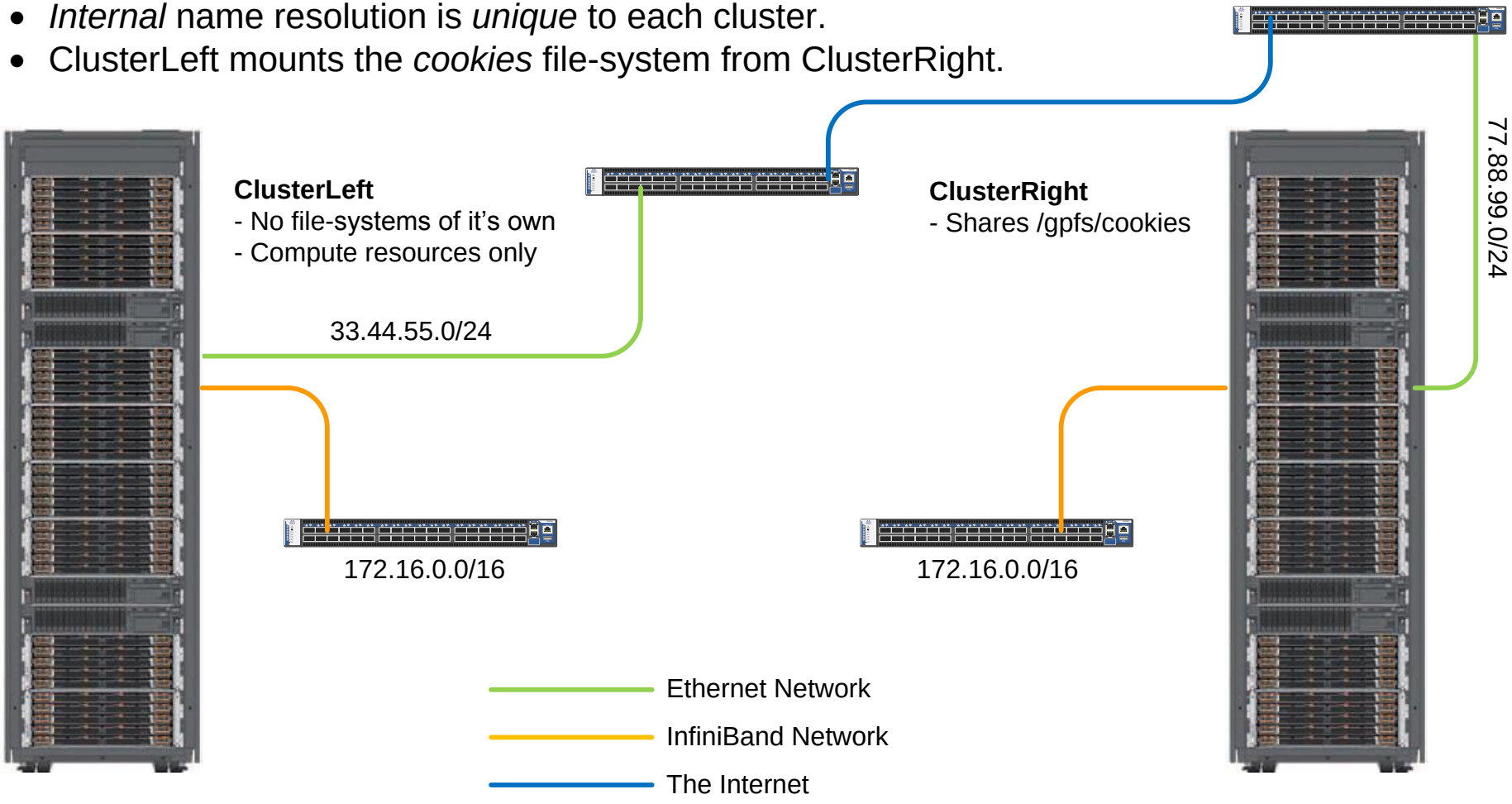
- **Still simple as both clusters have their daemon-interfaces on the 172.16.0.0/16 net.**
- **ClusterLeft uses 10.0.0.0/16 as it's main OS network.**
- **ClusterRight uses 77.88.99.0/24 as it's main OS network.**
- Both clusters use IP over 172.16.0.0/16 for administrative traffic.
- Both clusters use RDMA over InfiniBand for data traffic.
- ClusterLeft mounts the *cookies* file-system from ClusterRight.



Fairly Simple with Four Nets

Two GPFS clusters with **Four** IP Networks (**two** for GPFS)

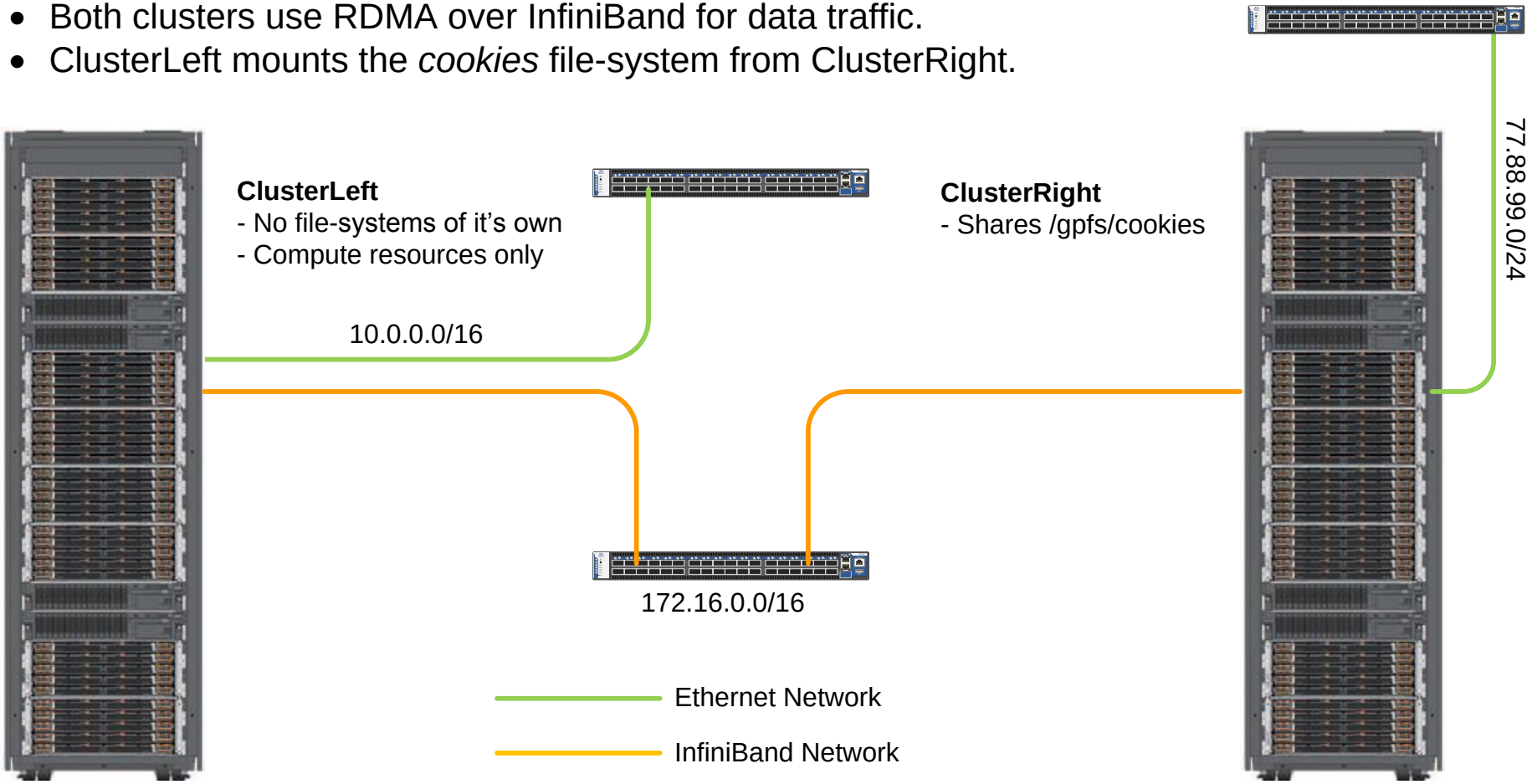
- Each cluster has it's own internal net, which both happen to use 172.16.0.0/16. This is OK because they are non-routed nets and thus do not present a conflict.
- The daemon-interfaces for each cluster are on their external nets (**green**), which are routed.
- *External* name resolution is *common and consistent* across both clusters.
- *Internal* name resolution is *unique* to each cluster.
- ClusterLeft mounts the *cookies* file-system from ClusterRight.



Now it's Gettin' Funky *(page 1)*

Two GPFS clusters with **Three** IP Networks (**two** for GPFS)

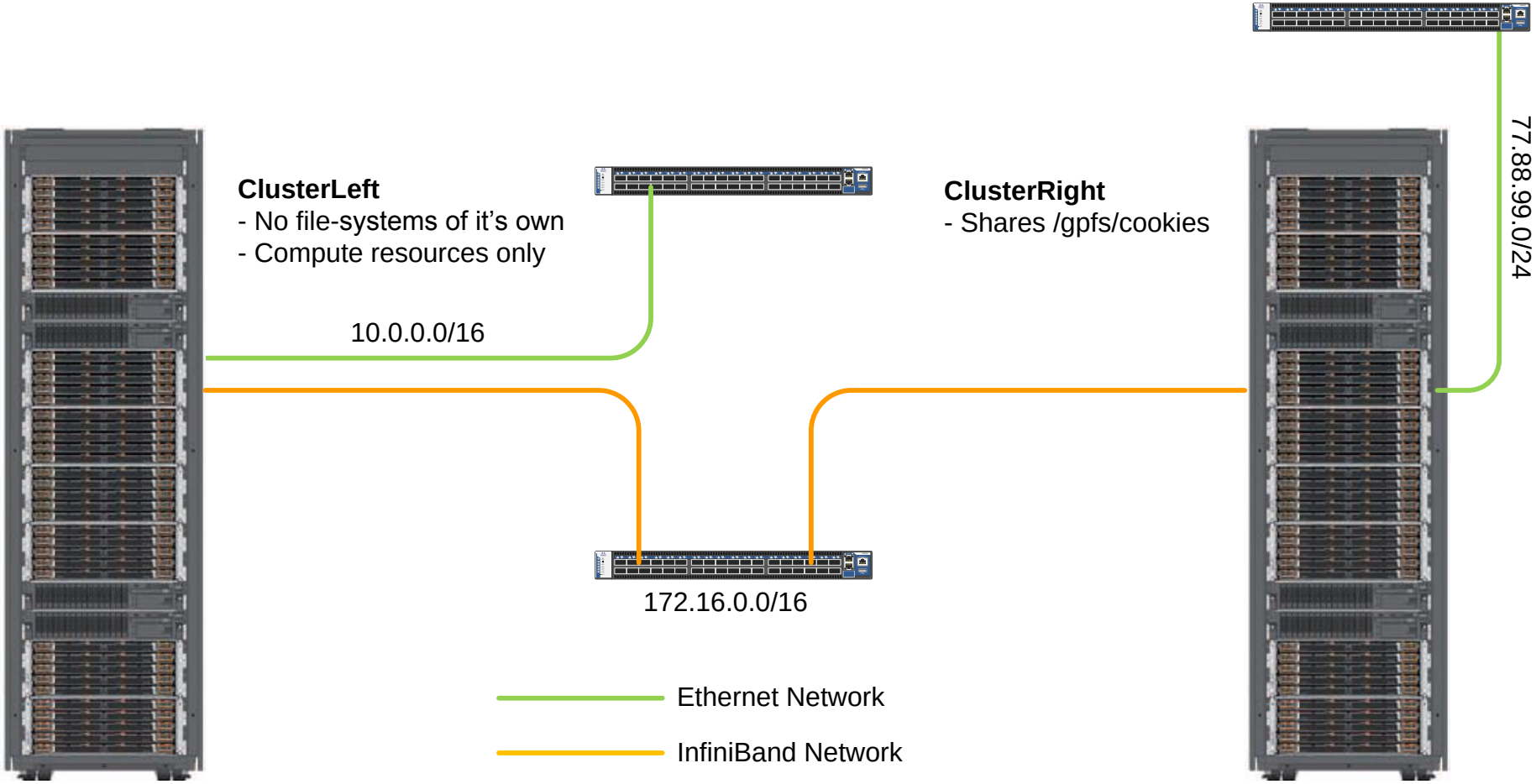
- ClusterLeft has it's daemon-interface on the 172.16.0.0/16 network.
- ClusterRight has it's daemon-interface on the 45.67.89.0/24 network.
- ClusterLeft requires static route entries to reach ClusterRight's daemon-interfaces.
- ClusterRight can already access ClusterLeft's daemon-interfaces directly.
- Both clusters use RDMA over InfiniBand for data traffic.
- ClusterLeft mounts the *cookies* file-system from ClusterRight.



Now it's Gettin' Funky *(page 2)*

Before Routing Changes

- ClusterLeft has no default gateway.
- ClusterLeft can *not* ping the daemon-interfaces of ClusterRight (77.88.99.0/24)
- ClusterLeft can ping the *non*-daemon-interfaces of ClusterRight (172.16.0.0/16)
- ClusterRight can ping the the daemon-interfaces of ClusterLeft (172.16.0.0/16)



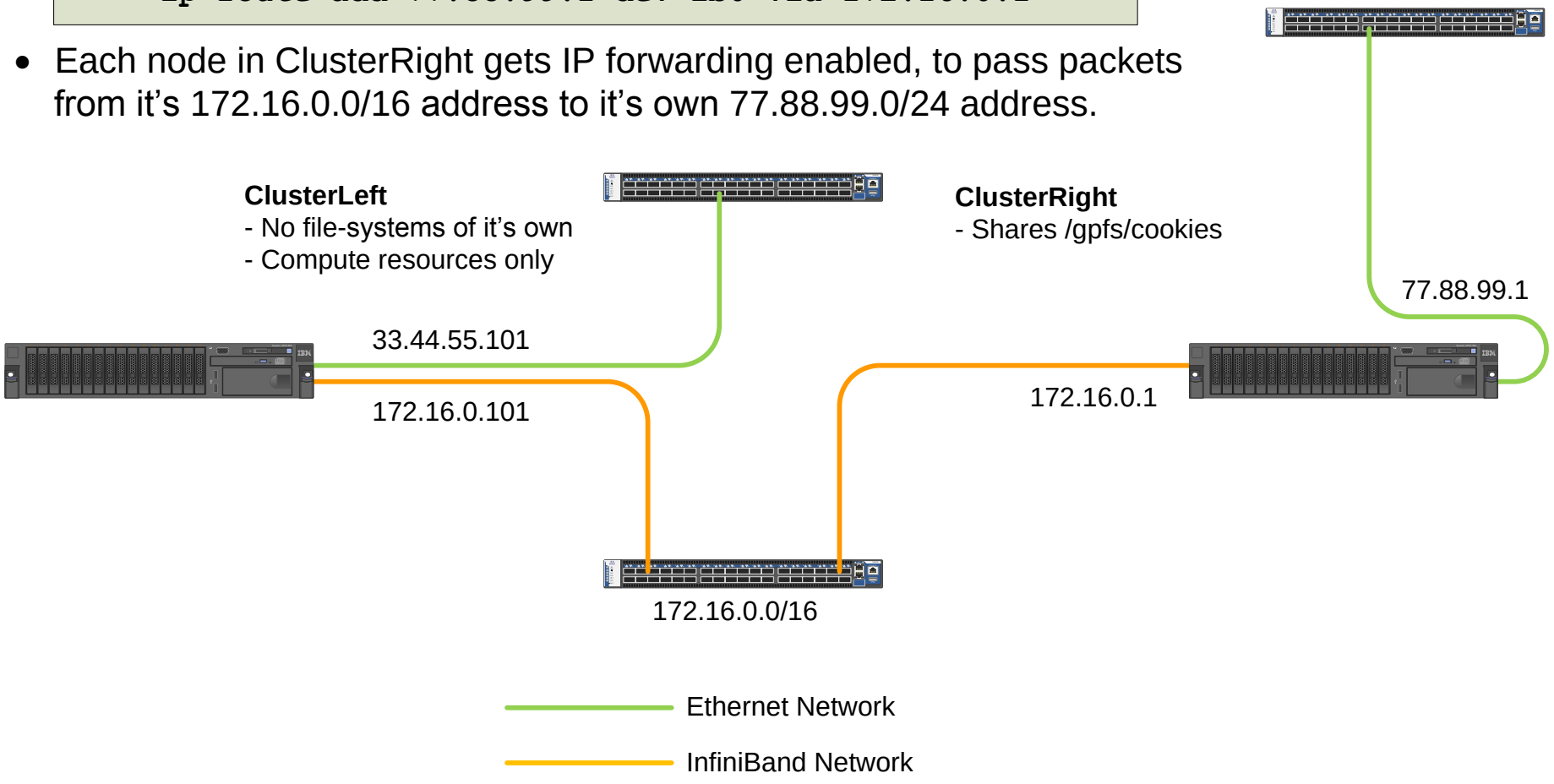
Now it's Gettin' Funky *(page 4)*

What are Those Routing Changes?

- Each node in ClusterLeft gets a shiny new static route entry for each node in ClusterRight, that makes the 172.16.0.0/16 address on each ClusterRight node the gateway to that same node's own 77.88.99.0/24 address.

```
ip route add 77.88.99.1 dev ib0 via 172.16.0.1
```

- Each node in ClusterRight gets IP forwarding enabled, to pass packets from it's 172.16.0.0/16 address to it's own 77.88.99.0/24 address.



For more information, contact:

Brian Finley
IBM, Executive IT Specialist
Scale Out Cloud and Technical Computing
Mobile: +1 469.667.2110
bfinley@us.ibm.com

Scott Denham
IBM, Consulting IT Specialist
IT Architect, Upstream Petroleum HPC
Phone: +1 713-940-1178
sdenham@us.ibm.com

Ray Paden
IBM, IT Architect - Cross Segment
STG, Software Defined Systems
Phone: +1 512-286-7055
raypaden@us.ibm.com