DataDirect™
NETWORKS
Information in Motion™

How to Ruin A Pefectly Good GPFS Filesystem (Yes I know its called Spectrum Scale)

SS Users Group CIUK 2015

Vic Cornell

Senior System Engineer DDN UK

Caveats

- IMHO
- YMMV
- RTFM

Using Ethernet

Using Ethernet Without a Separate Admin Network

- This one is VERY much IMHO etc etc.
- GPFS seems to be very good at . .er . . Congesting switches.
- Not many other ethernet applications are expected to run at “line speed”.
- Certain switches (enterprise) are worse than others.
- GPFS “pings” get lost.
- Expel storms follows.
- For smooth sailing without any storms add an admin network.
- Admin traffic always get through.
- Cluster stays healthy.
- . . . Or use Infiniband

Mixing Metadata and Data on the Same Disk

- So I have my lovely GPFS filesystem with 4MB blocks on NL-SAS drives, optimized for sequential reads. My clients are reading their big files and I'm getting lots of GB/s from even from 100 big slow drives.
- Then, someone goes looking for that report file that he archived in with his data files 5 years ago.
- find /gpfs –name report.pdf -print
- Suddenly all of the jobs reading from large sequential files slow down. Why?
- 100 drives is only 8000 IOPS.
- Find can blow through a significant number of these in very short order.
- Using the ILM/Policy engine can too.

Short stroking might not be as cool as it sounds

- "I want 2 filesystems from a fixed number of drives. I know that throughput is limited by number of drives – so: If I split my drives/LUNS in half I can have two filesystems with the same number of drives!"
- Not too bad with SSDs a the seek time across an SSD is uniform.
- BAD with drives – you end up "short stroking" one of your LUNS and "long stroking the other" so you will get very different performance.
- Also the two LUNS will compete for a limited (82/drive) number of IOPS.
- For GPFS "Scatter" Throughput = IOPS

Having uneven numbers of pools per NSD.

- The “dog with a wooden leg” Syndrome.
- If one NSD has more LUNS than the others then it will “probably” have poorer performance – either raw throughput or in contention for write cache etc.
- GPFS distributes data across all NSDs in a pool.
- Data transfers will run at the speed of the slowest NSD.
- Try to keep NSD loads balanced across NSD servers.

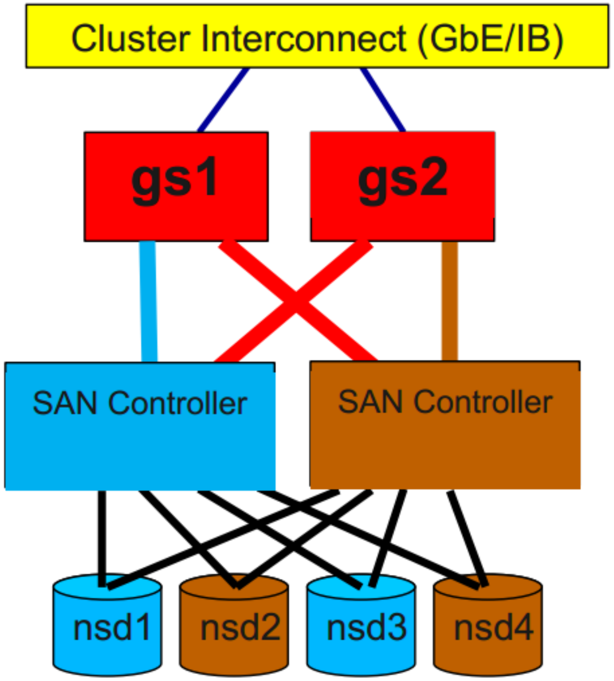
Pick The Right Path!

GRIDScaler External NSD Nodes (contd.)

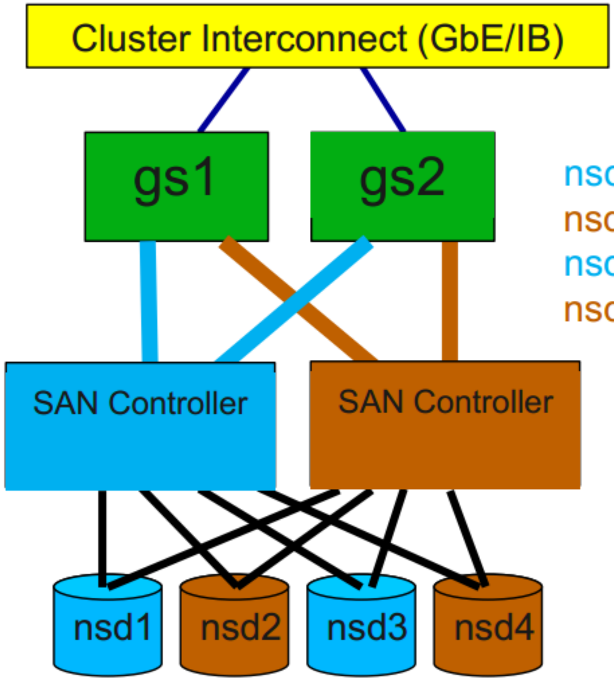
- Ensure the ServerList does not exclude the use of underlying storage device paths

WRONG

ServerList
nsd1: gs1,gs2
nsd2: gs2,gs1
nsd3: gs1,gs2
nsd4 gs2,gs1



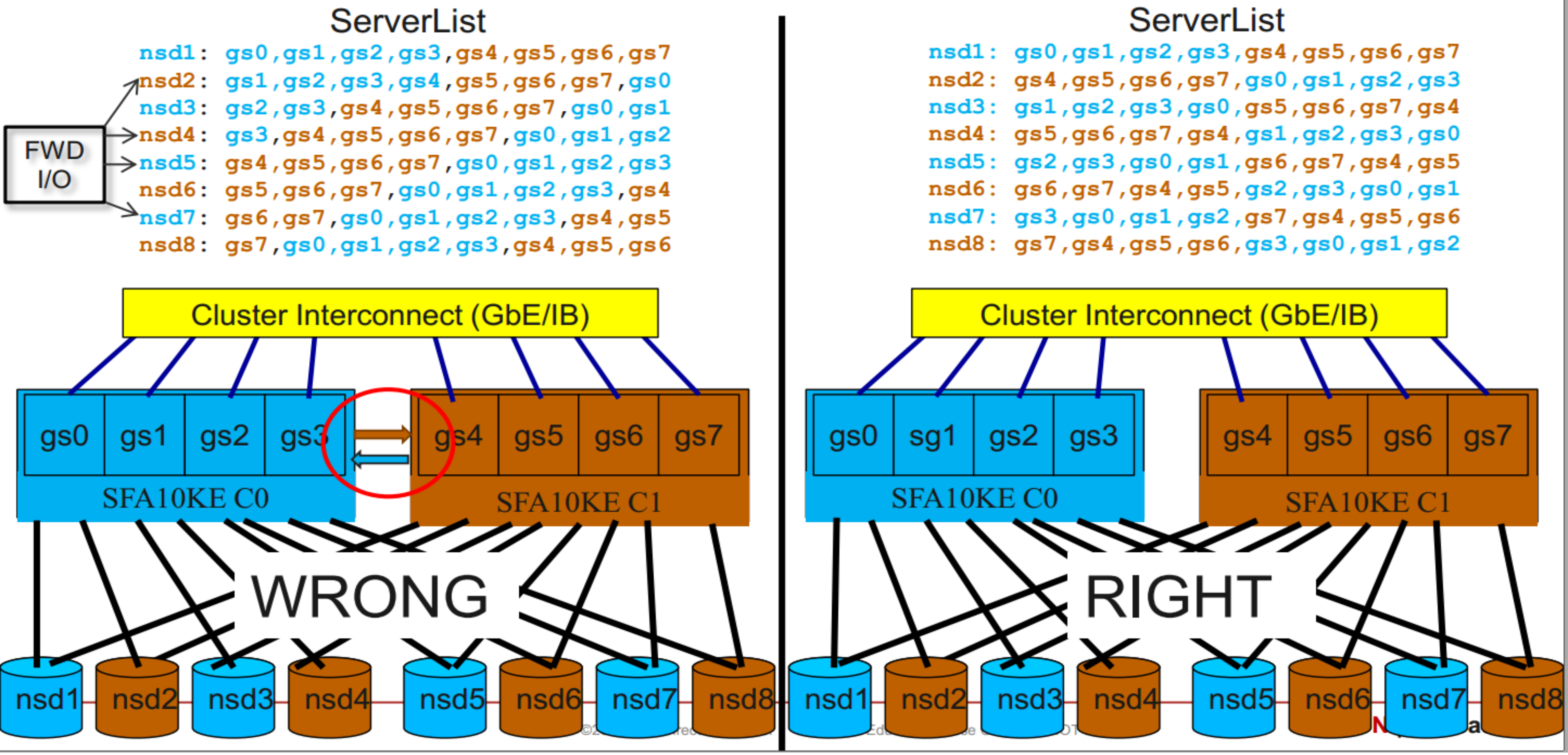
RIGHT



ServerList
nsd1: gs1,gs2
nsd2: gs1,gs2
nsd3: gs2,gs1
nsd4 gs2,gs1

Wrong Stanza File

- SFA10KE/12KE VMs
  - Primary NSD VM is running on controller that is the preferred home for the VD to avoid “Forwarded I/O” between controllers



ServerList

```
nsd1: gs0,gs1,gs2,gs3,gs4,gs5,gs6,gs7
nsd2: gs1,gs2,gs3,gs4,gs5,gs6,gs7,gs0
nsd3: gs2,gs3,gs4,gs5,gs6,gs7,gs0,gs1
nsd4: gs3,gs4,gs5,gs6,gs7,gs0,gs1,gs2
nsd5: gs4,gs5,gs6,gs7,gs0,gs1,gs2,gs3
nsd6: gs5,gs6,gs7,gs0,gs1,gs2,gs3,gs4
nsd7: gs6,gs7,gs0,gs1,gs2,gs3,gs4,gs5
nsd8: gs7,gs0,gs1,gs2,gs3,gs4,gs5,gs6
```

ServerList

```
nsd1: gs0,gs1,gs2,gs3,gs4,gs5,gs6,gs7
nsd2: gs4,gs5,gs6,gs7,gs0,gs1,gs2,gs3
nsd3: gs1,gs2,gs3,gs0,gs5,gs6,gs7,gs4
nsd4: gs5,gs6,gs7,gs4,gs1,gs2,gs3,gs0
nsd5: gs2,gs3,gs0,gs1,gs6,gs7,gs4,gs5
nsd6: gs6,gs7,gs4,gs5,gs2,gs3,gs0,gs1
nsd7: gs3,gs0,gs1,gs2,gs7,gs4,gs5,gs6
nsd8: gs7,gs4,gs5,gs6,gs3,gs0,gs1,gs2
```

Wrong Block Size

- Too Small:
  - GPFS and the storage has to work harder to move the same amount of data.
  - In Scatter mode you use up more IOPS per GB/s
  - Read-ahead will probably be smaller. (Same number of blocks?)
- Too Big
  - Less chance of a full stripe write.
  - More chance of a read/modify/write cycle with small I/O - stealing IOPS from your workload.
  - More waste as sub-block size gets bigger (1/32 of block size).

Using Cluster Mode?

- Scatter mode works well with full filesystems.
- Cluster mode is much faster.
- Are we using Scatter too much?
- Lustre uses a “Cluster-type” layout.
- Is Cluster that bad?

DMAPI?

- Brilliant but flawed.
- Only one DMAPI relationship per Filesystem.
- No support for quotas
- If your DMAPI destination breaks – it can break your GPFS filesystem or stop it from mounting.

Questions?

- ThankYou!