IBM GPFS GSS 26

GNR

Nilay K. Roy, Ph.D
• GPFS Overview
• Classical versus GSS I/O Solution
• GPFS Storage Server (GSS)
• GPFS Native RAID (GNR)
• Declustered RAID
- IBM General Parallel File System
- Shared-disk (NSD), clustered file system
- POSIX file system
- High-performance scalable parallel file system
  - Concurrent file access to applications executing on multiple nodes of the cluster
    - File data striped across all disks in file system
    - Distributed locking mechanism
    - Distributed metadata
- Fault tolerance and replication mechanism
- Extensible: adding/removing disks on-the-fly
What’s New:
- Replaces external hardware controller with software based RAID
- Modular upgrades improve TCO
- Non-intrusive disk diagnostics

Client Business Value:
- Integrated and ready to go for Big Data applications
- 3 years maintenance and support
- Improved storage affordability
- Delivers data integrity, end-to-end
- Faster rebuild and recovery times
- Reduces rebuild overhead by 3.5x

Key Features:
- Declustered RAID (8+2p, 8+3p)
- 2- and 3-fault-tolerant erasure codes
- End-to-end checksum
- Protection against lost writes
- Off-the-shelf JBODs
- Standardized in-band SES management
- SSD Acceleration Built-in
• GSS is a new IBM System-x solution based on the GPFS Native Raid (GNR) software

• GSS was officially announced at SC’12
  • Lemanicus BG/Q is the first system in the world integrating this solution

• GSS hardware consists of storage servers directly connected to enclosures with disks
  No storage controller needed

• The GSS servers perform both
  – traditional NSD service, and
  – disk controller functions (GNR software layer)
GSS Architecture

GPFS

GPFS Clients

LAN (IB)

NSD Servers

Virtual disks

GNR

SAS

JBODs

Virtual disks
GNR Features

- GPFS software layer enabling a GSS server to provide storage controller services
- Software-based error correction codes (ECCs)
  - Multiple data redundancy schemes
- Declustered RAID
  - Reduces disk rebuild overhead
  - Allows flexible virtual disk layout
- End-to-end checksums
  - Prevents silent data corruption
- Data caching
GNR Features (cont’d)

• Disk hospital
  – Asynchronous disk fault diagnostics
  – Live disk replacement
  – Background disk scrubbing

• Failover mechanisms
  – Seamlessly recovers from hardware failure

• Uniform and familiar command interface
  – Standard GPFS command syntax, including commands for maintaining and replacing failed disks
GNR – Physical disks (pdisks)

- JBOD disk enclosure
- NL-SAS 3TB disks (in BG/Q Lemanicus config)
- Dual-ported disks
  - Each pdisk is physically connected to two servers:
    - A primary GSS Server, and
    - A backup GSS server
  - GNR manages devices which map 1-to-1 to the SCSI disk devices
  - GNR provides disk failover
    - No multipathing drivers needed (no RDAC, no Device Mapper)
GNR – Recovery Groups (RGs)

- Disks of a GSS building block are divided into two non-overlapping RGs
- The RGs define the LUNs for primary and backup GSS servers
- Accesses to disks are made through the active server of the RG
- RG failover relies on inherent NSD failover capabilities of GPFS
  - when primary server fails, backup server automatically becomes the active server of the RG
Recovery Groups in GSS26 Building Block

Two RGs containing 174 pdisks each

Primary server for RG01
Backup server for RG02

Primary server for RG02
Backup server for RG01

GSS Server N01

GSS Server N02
GNR – Declustered Arrays (DAs)

- A DA is a subset of pdisks within a RG
- A DA can hold a number of virtual disks (vdisk) containing data and/or meta-data
  - Each vdisk has a specific RAID code and block size
  - Each vdisk is seen as a GPFS NSD (building block for File Systems)
- Spare space for the DA is allocated within the DA
Declustered Arrays in GSS26 Building Block

Six DAs containing 58 pdisks each (Lemanicus config)
GNR – Fault Tolerant ECCs

• 2-fault tolerant error correction codes (ECCs)
  8+2p Reed-Solomon code
  3-way replication

8 data strips = GPFS Block  2 parity strips

• 3-fault tolerant error correction codes (ECCs)
  8+3p Reed-Solomon code
  4-way replication
### Mean time to data loss 8+2 vs. 8+3

<table>
<thead>
<tr>
<th>Parity</th>
<th>50 disks</th>
<th>200 disks</th>
<th>50,000 disks</th>
</tr>
</thead>
<tbody>
<tr>
<td>8+2</td>
<td>200,000 years</td>
<td>50,000 years</td>
<td>200 years</td>
</tr>
<tr>
<td>8+3</td>
<td>250 billion years</td>
<td>60 billion years</td>
<td>230 million years</td>
</tr>
</tbody>
</table>

These figures assume uncorrelated failures and hard read errors.

**Simulation assumptions:** Disk capacity = 600-GB, MTTF = 600khrs, hard error rate = 1-in-10^{15} bits, 47-HDD declustered arrays, *uncorrelated failures*. These MTTDL figures are due to hard errors, AFR (2-FT) = 5 \times 10^{-6}, AFR (3-FT) = 4 \times 10^{-12}
GNR – Uniform and familiar interface

Command examples

• List information about GNR Recovery Groups
  – mmlsrecoverygroup

• List information for one or more GNR vdisks
  – mmlsvdisk

• Check and replace a physical disk
  – mmlspdisk all --not-ok
  – mmchcarrier gss2_L --pdisk "e1d1s04" --release
  => Replace disk
  – mmchcarrier gss2_L --replace --pdisk e1d1s04
Declustered RAID – Motivation

• Disk rebuilding becomes a critical issue in large scale storage systems
  – With 25’000 disks and an MTBF$_{\text{disk}}$ of 600 Khrs, a disk rebuild is triggered every day
  – 24-hour disk rebuild time implies permanent rebuilding activity
  – As GPFS distributes (stripes) file data over multiple LUNs to increase parallel performance, a single slower LUN impedes performance of the entire file system

• Rebuild time increases with larger disks
  – Risk of (too) many disk failures in same RAID increases
• Declustered RAID
  – Data and parity stripes are uniformly partitioned and distributed across a disk array.
  – Arbitrary number of disks per array (unconstrained to an integral number of RAID stripe widths)

• 2-fault and 3-fault tolerance
  – Reed-Solomon parity encoding
  – 2 or 3-fault-tolerant: stripes = 8 data strips + 2 or 3 parity strips
  – 3 or 4-way mirroring

• End-to-end checksum & dropped write detection
  – Disk surface to GPFS user/client
  – Detects and corrects off-track and lost/dropped disk writes

• Asynchronous error diagnosis while affected IOs continue
  – If media error: verify and restore if possible
  – If path problem: attempt alternate paths

• Supports live replacement of disks
  – IO ops continue on for tracks whose disks have been removed during carrier service
- Conventional RAID: Narrow data+parity arrays
  - Rebuild can only use the IO capacity of 4 (surviving) disks

20 disks (5 disks per 4 conventional RAID arrays)

4x4 RAID stripes (data plus parity)

Striping across all arrays, all file accesses are throttled by array 2’s rebuild overhead.

- Declustered RAID: Data+parity distributed over all disks
  - Rebuild can use the IO capacity of all 19 (surviving) disks

20 disks in 1 Declustered RAID array

16 RAID stripes (data plus parity)

Load on files accesses are reduced by 4.8x (=19/4) during array rebuild.
Declustered RAID – How it works

- Example: conventional RAID 6 (3+P+Q)
Declustered RAID – How it works

- Example: conventional RAID 6
  - 21 disks, 4 x RAID 6 (3+P+Q), 1 x hot spare

Only 5 disks participate to the rebuild operation
Declustered RAID – How it works

• How to improve (speed up) rebuild operation?

• Distribute arrays across all pdisks

=> Declustered Array
  – Data strips from any virtual \textit{vdisk} (3+P+Q) are distributed across different pdisks (example too small !)
Declustered RAID – How it works

• **Rebuild**

![Diagram of rebuild process]

- Failed disk
- All disks participate in the rebuild operation

**Rebuild time**
Rebuild Overhead Reduction Example

Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs

Rebuild activity spread across many disks, less disruption to user programs

Rebuild overhead reduced by 3.5x
Declustered RAID Example

3 1-fault-tolerant mirrored groups (RAID1)

7 stripes per group (2 strips per stripe)

3 groups
6 disks
spare disk
7 spare strips

7 disks

21 stripes (42 strips)

49 strips
Declustered RAID6 Example

14 physical disks / 3 traditional RAID6 arrays / 2 spares

14 physical disks / 1 declustered RAID6 array / 2 spares

Number of faults per stripe

Red | Green | Blue
---|---|---
0 | 2 | 0
0 | 2 | 0
0 | 2 | 0
0 | 2 | 0
0 | 2 | 0
0 | 2 | 0
0 | 2 | 0

Number of stripes with 2 faults = 7

Number of faults per stripe

Red | Green | Blue
---|---|---
1 | 0 | 1
0 | 0 | 1
0 | 1 | 1
2 | 0 | 0
0 | 1 | 1
1 | 0 | 1
0 | 1 | 0

Number of stripes with 2 faults = 1
Declustered RAID – Rebuild time

• Conventional RAID
  – Typical disk rebuild time is 24 hours
  – Once a failed disk has been replaced, data has to be copied back from the hot spare to the new disk

• GSS26 Building block
  Assuming 58 x 3 TB disks DAs with 8+2p ECC:
  – Normal rebuild time for a 3 TB disk =~ 3 hours
    • Occurs as a background task not to interfere with regular disk access workload
  – Critical rebuild time =~ 15 minutes
    • A stripe is considered critical if one more disk failure would result in data loss
Declustered RAID – Flexibility

**Conventional RAID**
- RAID groups are mapped to specific disks
- Fixed number of disks in arrays
- Performance constrained by limited number of disks
- Unused slots

**Declustered Array**
- Vdisks are uniformly distributed over all pdisks

- Vdisk0 (8+2p) - dataOnly
- Vdisk1 (8+2p) - dataOnly
- Vdisk2 (3-way replication) - metadataOnly
- Spare space
GSS Server
- Standard IBM System-x
- x3650 M4
- E-2670, 2x8 cores, 2.6 GHz
- 64 GB memory

JBOD disk enclosure
5 x drawers
58 x 3 TB NL-SAS disks

GSS26 Building Block
2 x servers (x3650 M4)
6 x JBOD enclosures
348 x 3 TB NL-SAS
- raw capacity = 972 TiB
- usable capacity ~ 726 TiB
Max streaming rate 10 GB/s

Lemanicus BG/Q config
3 x GSS26 Building Blocks
1044 x 3 TB NL-SAS
- raw capacity = 2916 TiB
- usable capacity ~ 2178 TiB
Max streaming rate 30 GB/s
# Lemanicus BG/Q – File Systems

<table>
<thead>
<tr>
<th>Mount Point</th>
<th>Size</th>
<th>Access rights</th>
<th>Backed-up</th>
<th>Quotas</th>
<th>Max Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>/home</td>
<td>217 TB</td>
<td>ro</td>
<td>Yes</td>
<td>No</td>
<td>~25 GB/s</td>
</tr>
<tr>
<td>/bgscratch</td>
<td>~ 1.9 PB</td>
<td>rw</td>
<td>No</td>
<td>No</td>
<td>~25 GB/s</td>
</tr>
</tbody>
</table>
Lemanicus BG/Q – I/O Performance

- IOR Benchmark
- Test cases
  - IOR running on 1024 nodes (on the entire BG/Q rack)
  - In each test case, the aggregate filesize is 2048 GiB
  - Transfer size is 16 MiB

<table>
<thead>
<tr>
<th>IOR test cases</th>
<th>1024 MPI processes ranks-per-node 1 blocksize = 2 GiB</th>
<th>16384 MPI processes ranks-per-node 16 blocksize = 128 MiB</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOR POSIX file-per-process</td>
<td>25037 / 28480 MiB/s</td>
<td>21263 / 25410 MiB/s</td>
</tr>
<tr>
<td>IOR POSIX single-shared-file</td>
<td>23988 / 27038 MiB/s</td>
<td>16773 / 24063 MiB/s</td>
</tr>
<tr>
<td>IOR MPIIO file-per-process</td>
<td>25166 / 28034 MiB/s</td>
<td>20996 / 23888 MiB/s</td>
</tr>
<tr>
<td>IOR MPIIO single-shared-file</td>
<td>23945 / 27740 MiB/s</td>
<td>17689 / 24281 MiB/s</td>
</tr>
</tbody>
</table>
GSS/GNR advantages

• Lower TCO
  – No external hardware controller required
  – One end-to-end technology – IBM GPFS – from the compute nodes to the disks
  – Easier maintenance
    • no training required for third-party storage system
    • GPFS standard command interface
    • standard RedHat Linux on GSS servers
  – Easier end-to-end I/O tuning
    • no Device-Mapper Multipath
    • nothing to tune at the back-end storage level
GSS/GNR advantages (cont’d)

- Flexible vdisk and file system configuration
  - Any number of vdisks of different sizes, ECCs and block sizes can be defined, and integrated in FS

- Improved data integrity
  - 2- and 3-fault tolerant ECCs
  - End-to-end checksum
  - Faster rebuild times using declustered RAIDs

GSS/GNR drawback

- All file systems share disk bandwidth
GPFS Native Raid Advantages

- **Lower Cost!**
  - Software RAID – No *hardware* storage controller
    - 10-30% lower cost with higher performance
  - Off-the-shelf SBODs
    - Generic low-cost disk enclosures
    - Standardized in-band SES management
  - *Standard* Linux or AIX
  - Generic high-volume servers
  - Component of GPFS

- **Extreme Data Integrity**
  - 2- and 3-fault-tolerant erasure codes
    - 80% and 73% storage efficiency
  - End-to-end checksum
  - Protection against *lost writes*
  - **Fastest** Rebuild times using Declustered RAID

- **Industry Leading Performance**
  - Declustered RAID – *Reduced* app load during rebuilds
    - Up to 3x lower overhead to applications
  - Aligned full-stripe writes – disk limited
  - Small writes – backup-node NVRAM-log-write limited
  - Faster than alternatives today – and tomorrow!
Complete Storage Solution
Data Servers, Disk (NL-SAS and SSD), Software, InfiniBand and Ethernet

Model 24:
Light and Fast
4 Enclosures, 20U
232 NL-SAS, 6 SSD
10 GB/Sec

Model 26:
HPC Workhorse!
6 Enclosures, 28U
348 NL-SAS, 6 SSD
12 GB/sec

High-Density HPC Option
18 Enclosures
2 - 42U Standard Racks
1044 NL-SAS 18 SSD
36 GB/sec

Performance based on IOR BM
Thank You