IBM GENERAL PARALLEL FILE SYSTEM

Alexandre Quemy

January 12, 2016

IBM Analytics

GPFS INTRODUCTION

GPFS in a nutshell

- High-performance clustered and parallel filesystem.
- Available on AIX (1998), Linux (2001), or Windows Server (2008).
- Full POSIX filesystem semantic.
- Shared disks or shared-nothing mode.
- Share using Ethernet or Infiniband.
- HDFS compatible, so might be used by Hadoop.
- Now part of IBM Spectrum Scale (rebranded since 2015).

GPFS IN INDUSTRY

Used in some of the largest supercomputers:

- ALTAMIRA, 240 IBM nodes, 4000 Intel cores, <1ms latency between nodes, 2 PB storage.
- ASC Purple, 200 IBM nodes, 12500 cores, 50 TB memory, 2 PB storage.

GPFS IN FIGURES

Theoretical limits of GPFS:

- Maximum volume size: 8 yobibytes (2^{80} bytes).
- Maximum file size: 8 exbibytes (2^{60} bytes).
- Maximum number of files: 2^{60} per file system.

Comparison with ZFS:

- Maximum volume size: 256 zebibytes (2^{70} bytes).
- Maximum file size: 16 exbibytes (2^{60} bytes).
- Maximum number of files: 2^{40} per filesystem.

Comparison is not reason:

ZFS is not designed for clustered and parallel accesses.

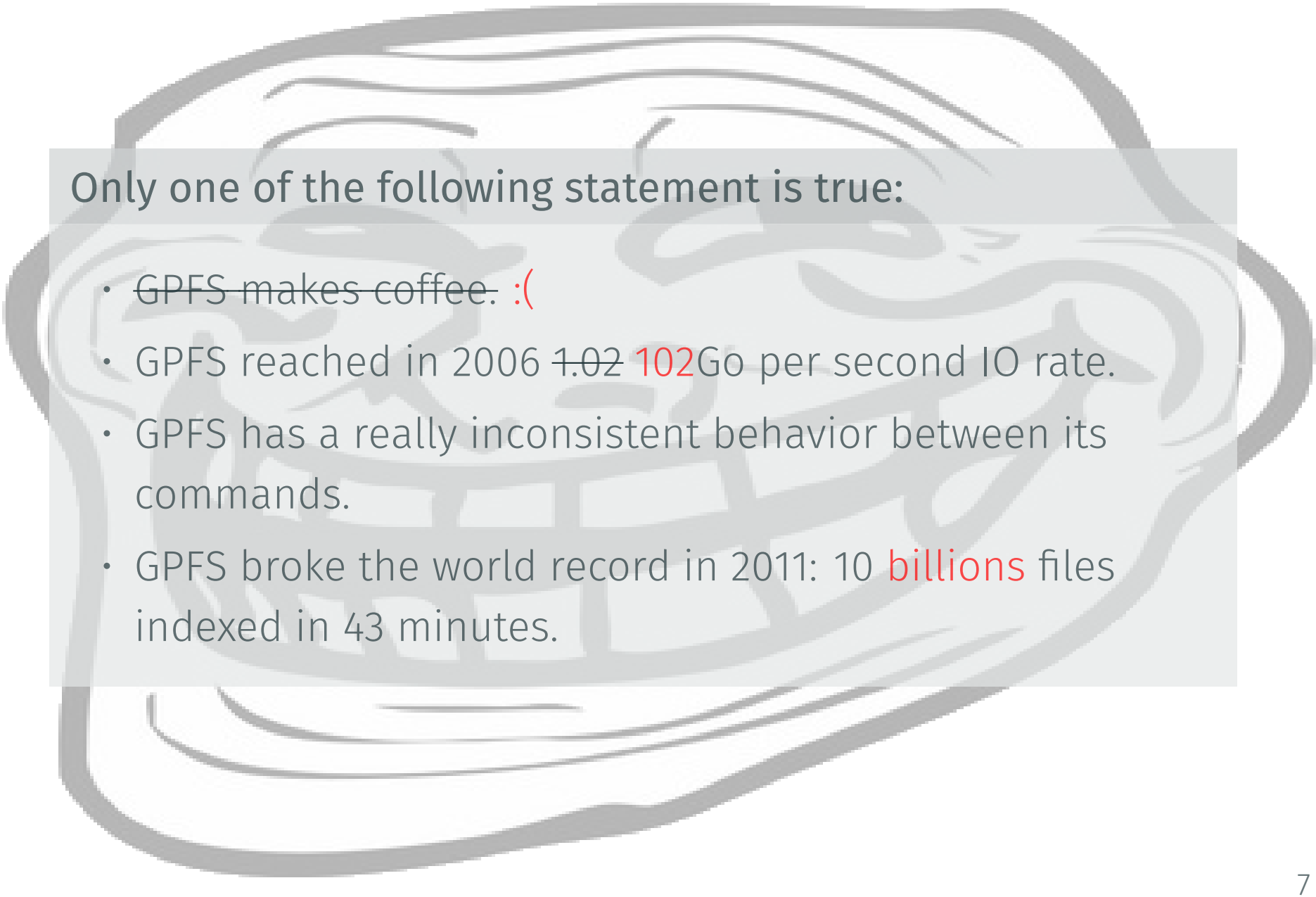
LET'S PLAY A GAME

Only one of the following statement is true:

- GPFS makes coffee.
- GPFS reached in 2006 1.02Go per second IO rate.
- GPFS has a really inconsistent behavior between its commands.
- GPFS broke the world record in 2011: 10 files indexed in 43 minutes.

Make a guess !

LET'S PLAY A GAME



Only one of the following statement is true:

- ~~GPFS makes coffee.~~ :(
- GPFS reached in 2006 ~~1.02~~ 102Go per second IO rate.
- GPFS has a really inconsistent behavior between its commands.
- GPFS broke the world record in 2011: 10 billions files indexed in 43 minutes.

HOW DOES IT WORK ?

Parallel and distributed filesystem:

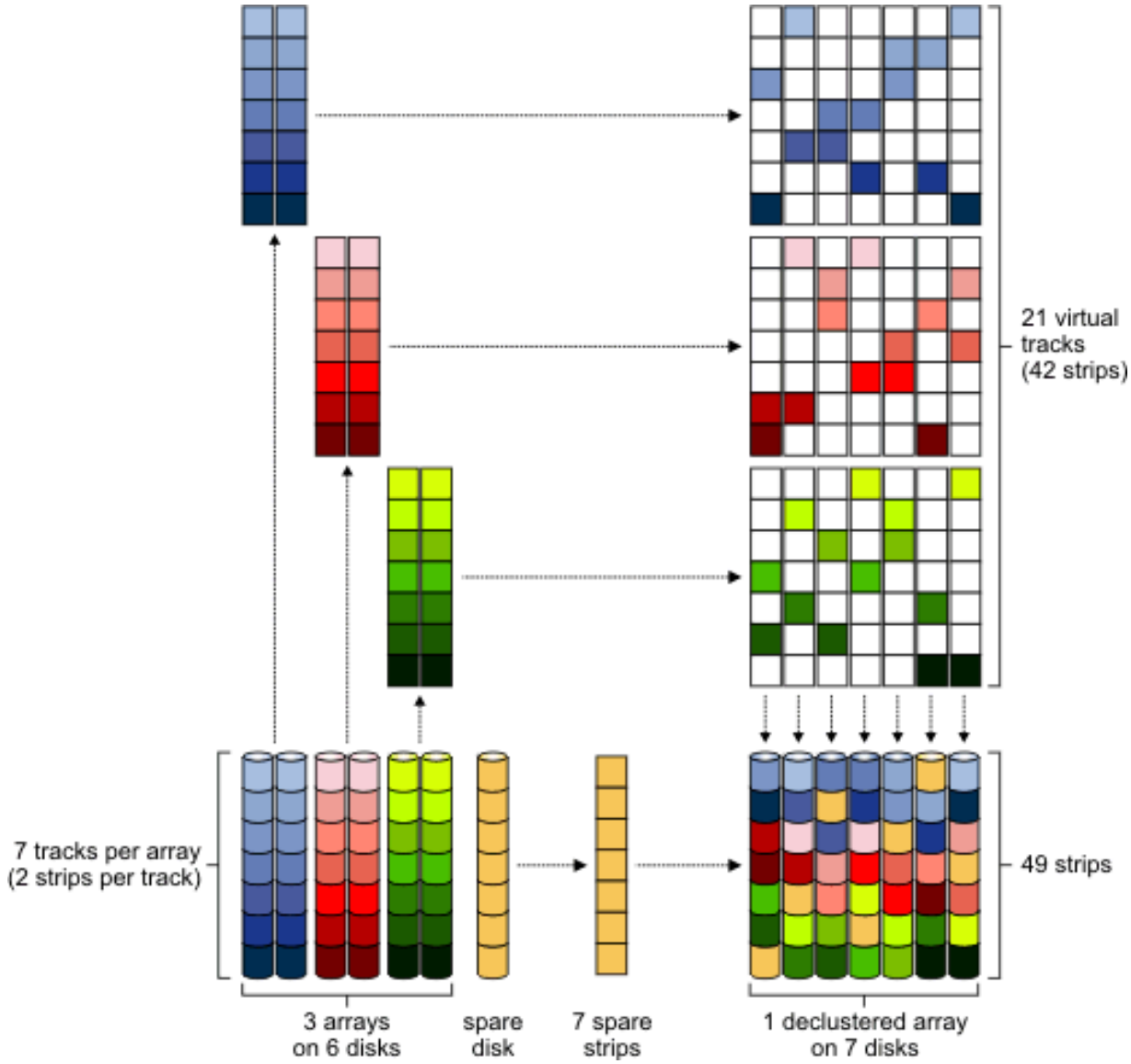
- Each file is stripped into blocks of a given size.
- Each block is replicated over multiple disks $\Longrightarrow$ parallel reading.
- Local deamon for synchronization $\Longrightarrow$ parallel writting.
- Local RAID controller for per-node fault tolerance (disk or block level).

Main advantage:

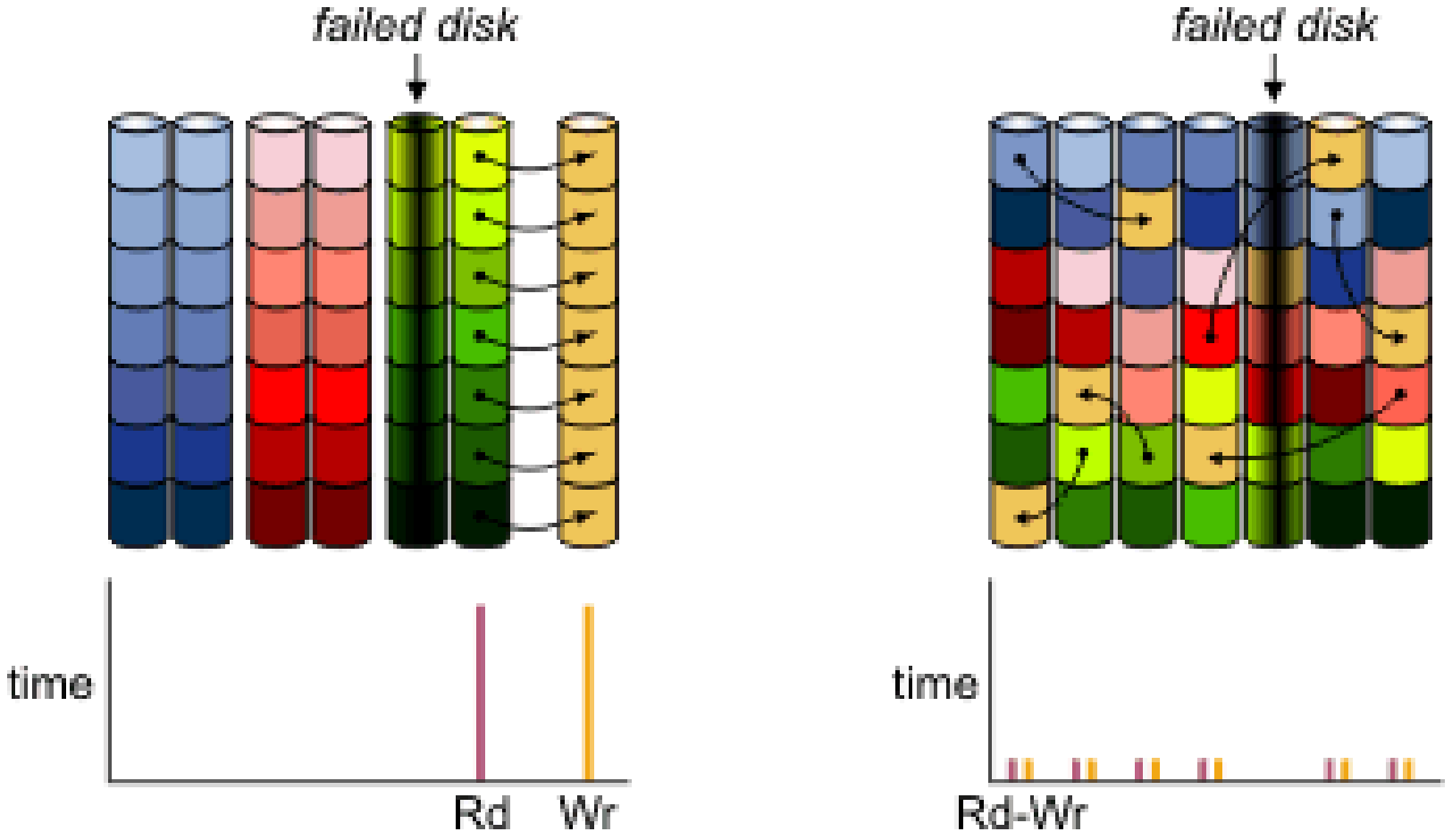
Adding disks and nodes increases the performances.

HOW DOES IT WORK ?

Block level RAID (aka RAID declustering):

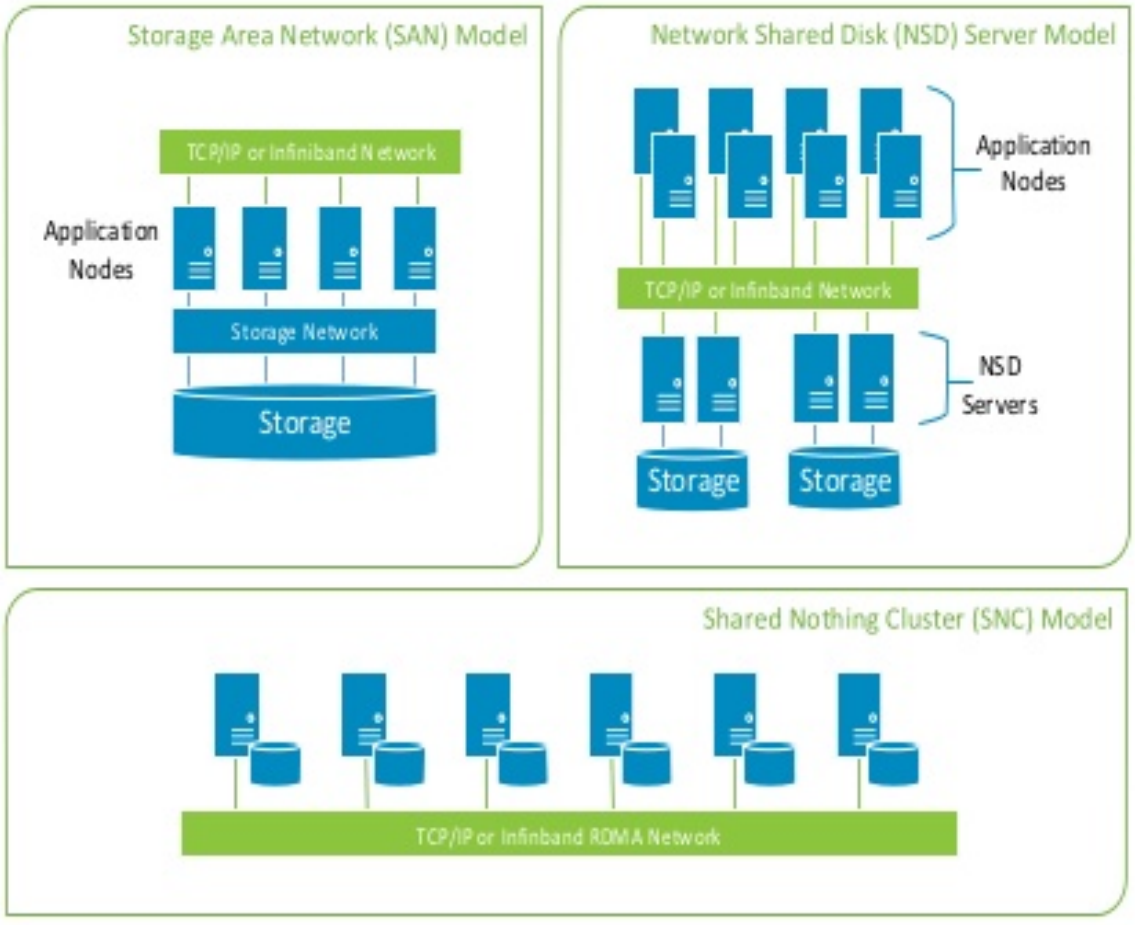


Advantage of Block level RAID:



HOW DOES IT WORK ?

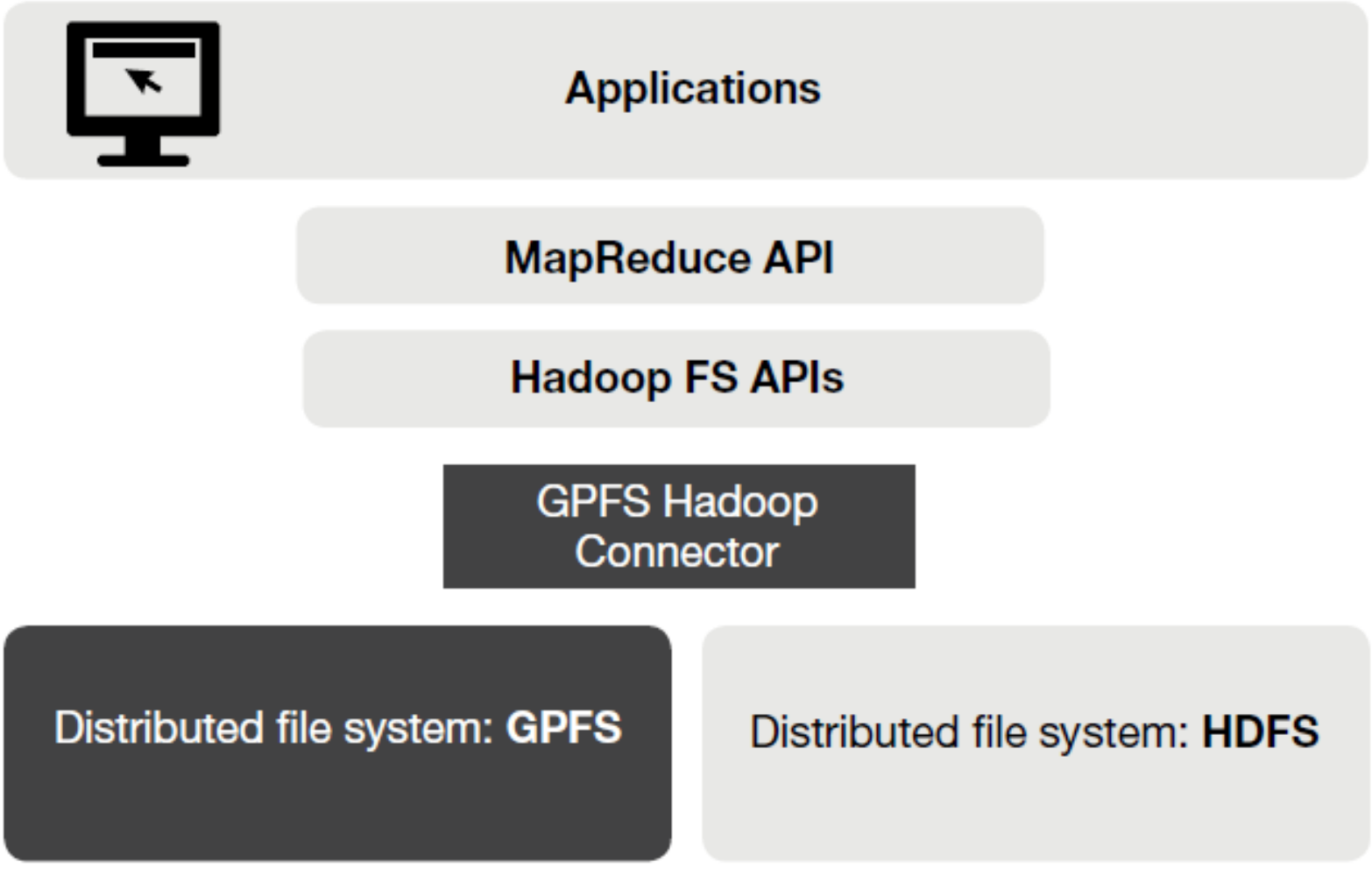
Describe Common Spectrum Scale Architectures



Software Defined Storage means

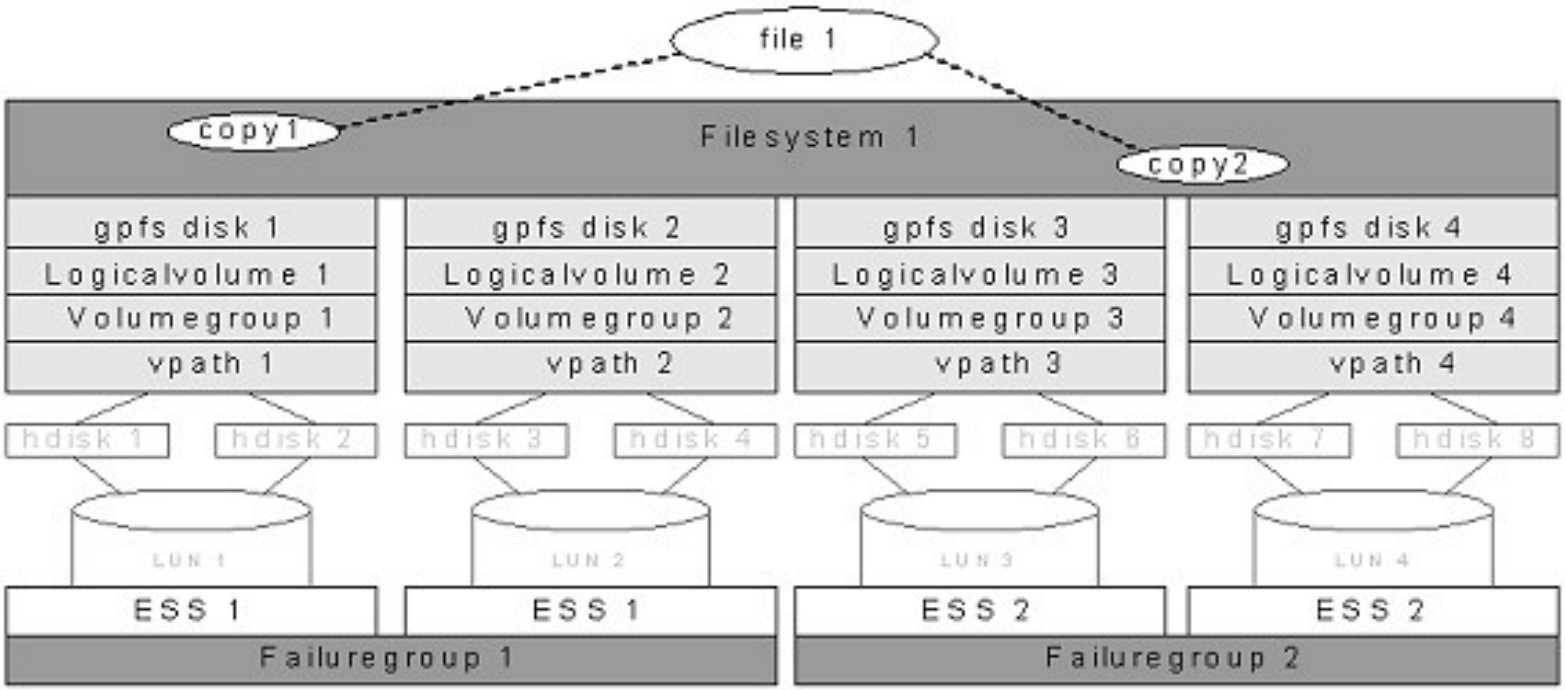
- Spectrum Scale (GPFS 4.1) is a Software product
- It provides flexibility to choose the right server & storage technology
- It allows you to mix and match storage to support a wide variety of application workloads.
- It supports virtually any Network Architecture allowing Use of Fibre Channel SAN, SAS, TCP/IP and InfiniBand RDMA to transfer data.

CAN I USE GPFS WITH HADOOP ?



THE CONCEPTS

THE BIG PICTURE



Cluster

Composed of nodes that can be heterogenous.

Communication between nodes:

- TCP/IP communication: requires consistent name resolution for forward and reverse look-up, and passwordless connection.
- Administration commands: remote shell (ssh, rsh or other) and socket based communications. By default: scp + scp.

Quorum

Maintain data consistency in case of a node failure:

- Prevents a failed note to write data in a filesystem.
- If quorum nodes are not available, GPFS unmount filesystems on the remaining nodes.

Cluster Manager

One of the quorum nodes. Performs the following tasks:

- Monitors disk leases.
- Detects failures and manages recovery from node failure within the cluster.
- Determine is there existing quorum nodes (or GPFS won't start !).
- ...

Cluster configuration repository (CCR)

Mechanism to store and maintain consistent copies of configuration over all quorum nodes.

- Alternative to standard predefined backup node.
- Can be activated on-the-fly.
- Starts a new deamon communicating over TCP/IP.

Possibility to decouple admin network from ’cluster’ network

```
Node  Daemon node name  IP address       Admin node name         Designation
-----------------------------------------------------------------------------
1     node1             8.10.10.1        node1                   quorum-manager
2     node2             8.10.10.2        node2                   quorum-manager
3     node3             8.10.10.3        node3                   quorum-manager

> mmchnode --admin-interface=node1a -N node1

Node  Daemon node name  IP address       Admin node name         Designation
-----------------------------------------------------------------------------
1     node1             8.10.10.1        node1a                  quorum-manager
2     node2             8.10.10.2        node2a                  quorum-manager
3     node3             8.10.10.3        node3a                  quorum-manager
```

Network Shared Disks (NSD)

Defined over a standard disk partition. Clients and servers communicate by GPFS RPCs over TCP/IP protocol.

NSD Threading Model

- NSD IO requests requires a memory buffer of at least the size of the IO request.
- Such buffers come out of GPFS pagepool (pinned memory to avoid page fault).
- Array of independently lockable queues to handle requests.
- Each NSD Worker is statically assigned to a pagepool buffer.

Failure Group

Set of disks with a common point of failure that could cause them all to become simultaneously unavailable. Most common example: the disks on a same node.

Filesystem

Set of NSDs sharing the same semantic: data, log,...
Seen by the end-user as a single volume to be mounted.

$\Longrightarrow$ Can be seen as a partition.

Filesystem Descriptors

- One copy on every disk of the filesystem.
- Depending on the number of failure group, more copies.
- Possibility to place descriptors on another filesytem's disk.

Filesystem Manager

- Manage FS configuration (adding disk, repair, mount,...).
- Management of disk allocation.
- Token management.
- Quota allocation.

Replication

Insure that each block of data and metadata is replicated in different failure groups. It increases the availability.

Replication factor for data and metadata: default and max (1, 2 or 3)

Usually at Filesystem level but can be done at file level.

THE CONCEPTS: STORAGE POOL, FILESET, PLACEMENT POLICY

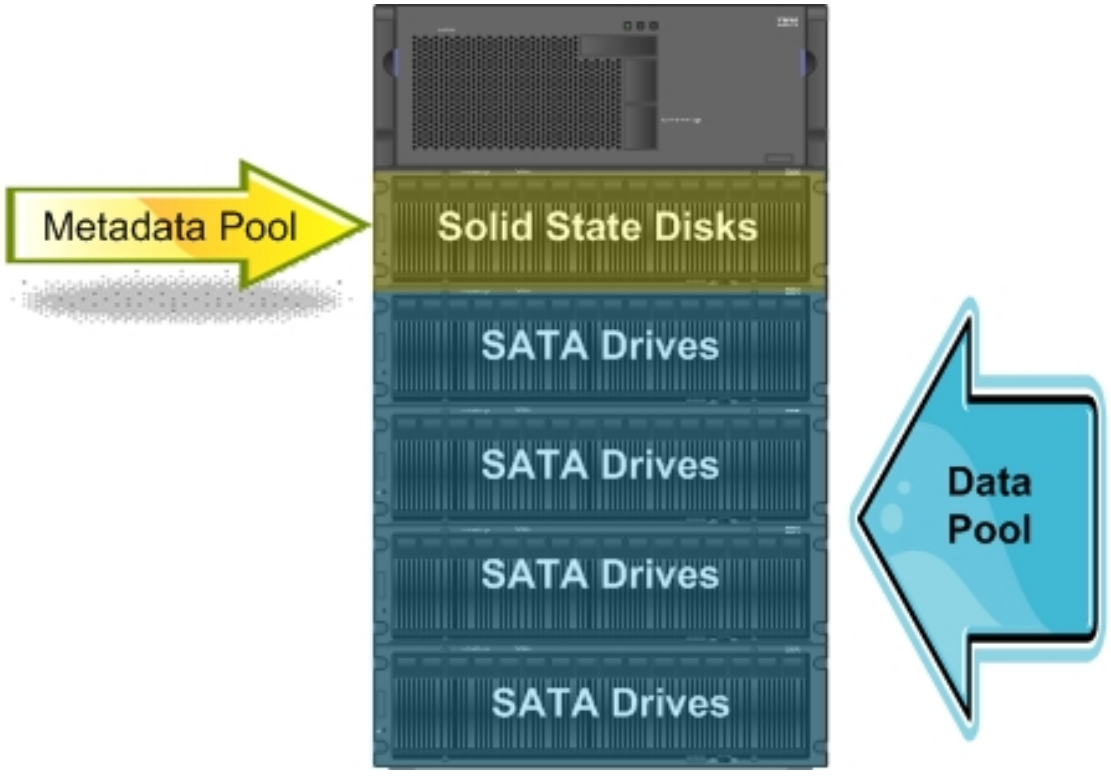
Storage Pool

Collection of physical disks with the same properties, managed as a group.

Examples of use

- Improve cost/performance.
- Improve performance:
  1. Reduce impact of slower devices.
  2. Match physical characteristics with logical characteristic.
- Improve reliability.

THE CONCEPTS: STORAGE POOL, FILESET, PLACEMENT POLICY



FileSet

- Subtree of the filesystem namespace (≅ independant filesystem).
- Allow finer grain management of filesystems (cf. Placement).
- Allow partial filesystem snapshot.
- It is a file attribute as its name.

THE CONCEPTS: STORAGE POOL, FILESET, PLACEMENT POLICY

Placement Policy

- Automatically place file in a storage pool.
- Based on file attributes (size, user/ group ID, **fileset**).
- Uses SQL-like statements.

Example

```
RULE 'datafiles' SET POOL 'data'
        WHERE UPPER(name) like '%.dat'
```

BASIC GPFS CLUSTER CREATION

TWEAKING AND OPTIMIZING GPFS

Internet communication: IPoIB or native interface ?

More like TCP/IP over InfiniBand (IPoIB) or a separate Ethernet interface ?

Non universal answer: if packets are less than 8KB, Ethernet is better than IPoIB.

In GPFS: verbsRdmaMinBytes parameter (default value: 8192B).

Client communication

IB RDMA protocol is possible:

```
mmchconfig verbsPorts="mthca0/1/1 mthca0/2/1"
mmchconfig verbsRdma=enable
```

Size does matter !

The optimal block size depends on the IO type (and thus applications):

- Large IO (Scientific computation, media): 1 to 4MB.
- Relational DB: 512kB.
- Small IO: 256kB

Benchmark it with GPFS !

```
/mmfs/sample
```

Data and MetaData

Use small and fast disks for metadata !

Pinned memory

Increase the pagepool ! (default 64M)

$$\text{FS Blocksize} = N \times 256\text{k}$$

$$\Longrightarrow \text{ pagepool} = N \times 64M$$

Data and MetaData

Default values are not optimal in general:

- maxMBpS $= 2 \times$ network bandwidth.
- maxFilesToCache to increase !

THE END

THE END

Thank you for your attention !
Questions ?