Spectrum Scale

& Hadoop

olaf.weiser@de.ibm.com
IBM Deutschland
SpectrumScale Support Specialist

Agenda

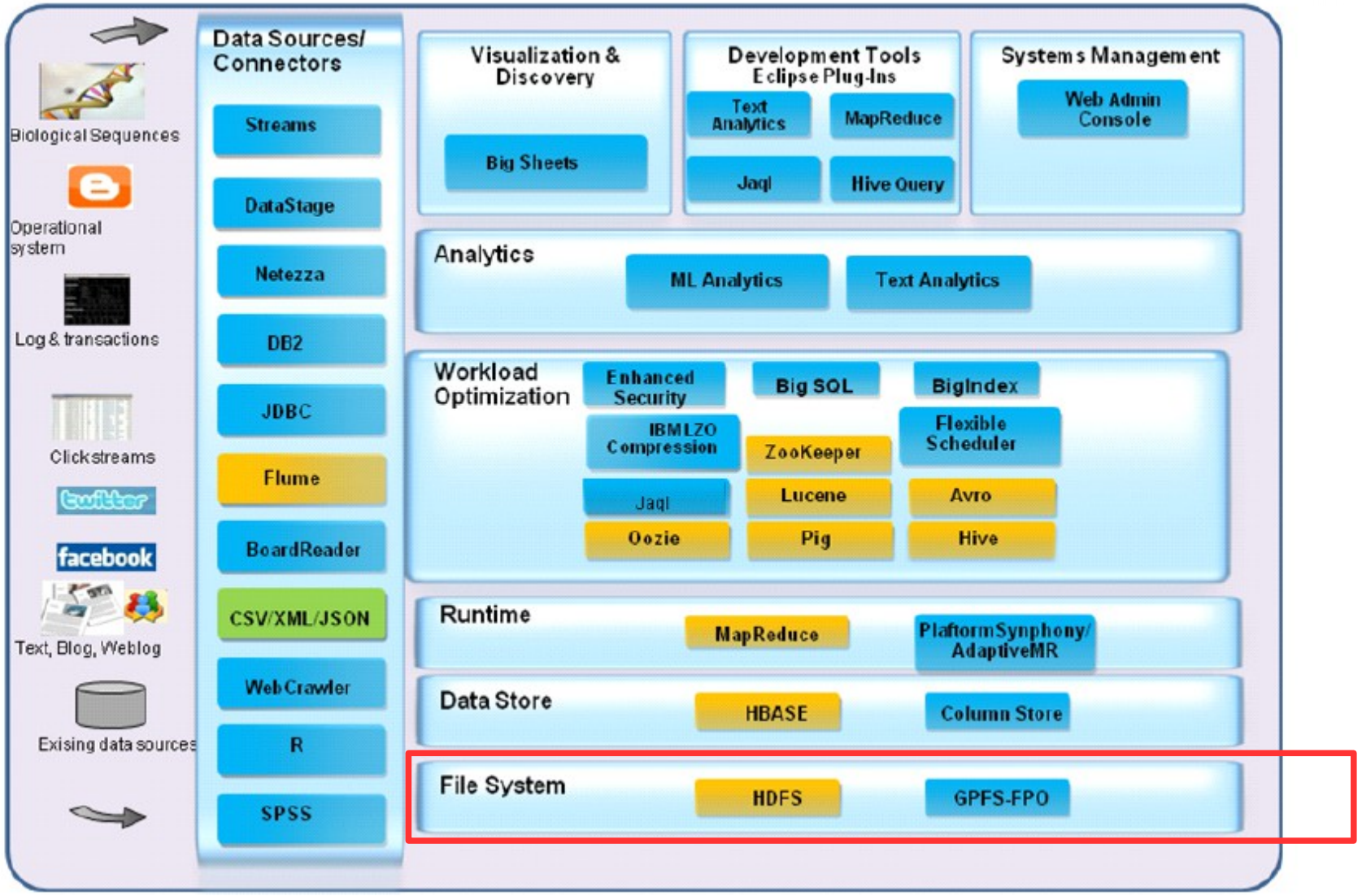
1) Hadoop, HDFS and SpectrumScale
   short - overview

2) Hadoop Connector
   a.) „old“-version
   b.) „new“-transparency

3) Configuration overview

4) Need to know - what else comes with SpectrumScale

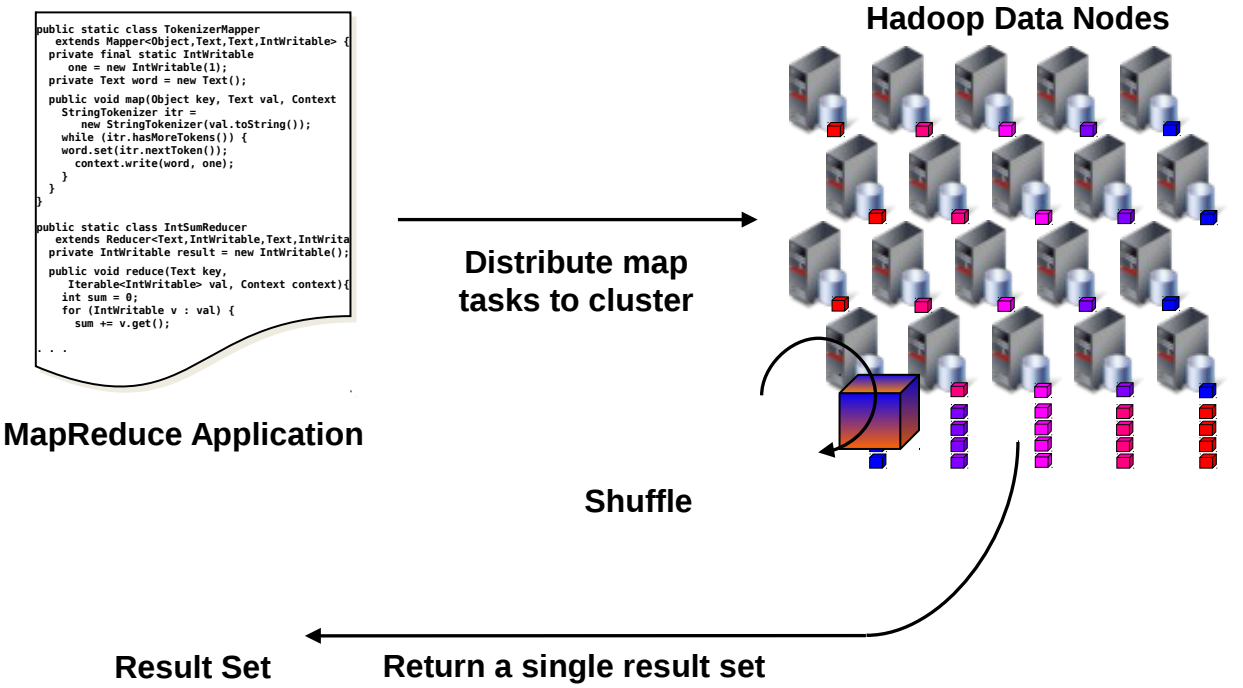
Big Data : 100 and more applications...



Basic Hadoop principles: Distributed storage and MapReduce runtime

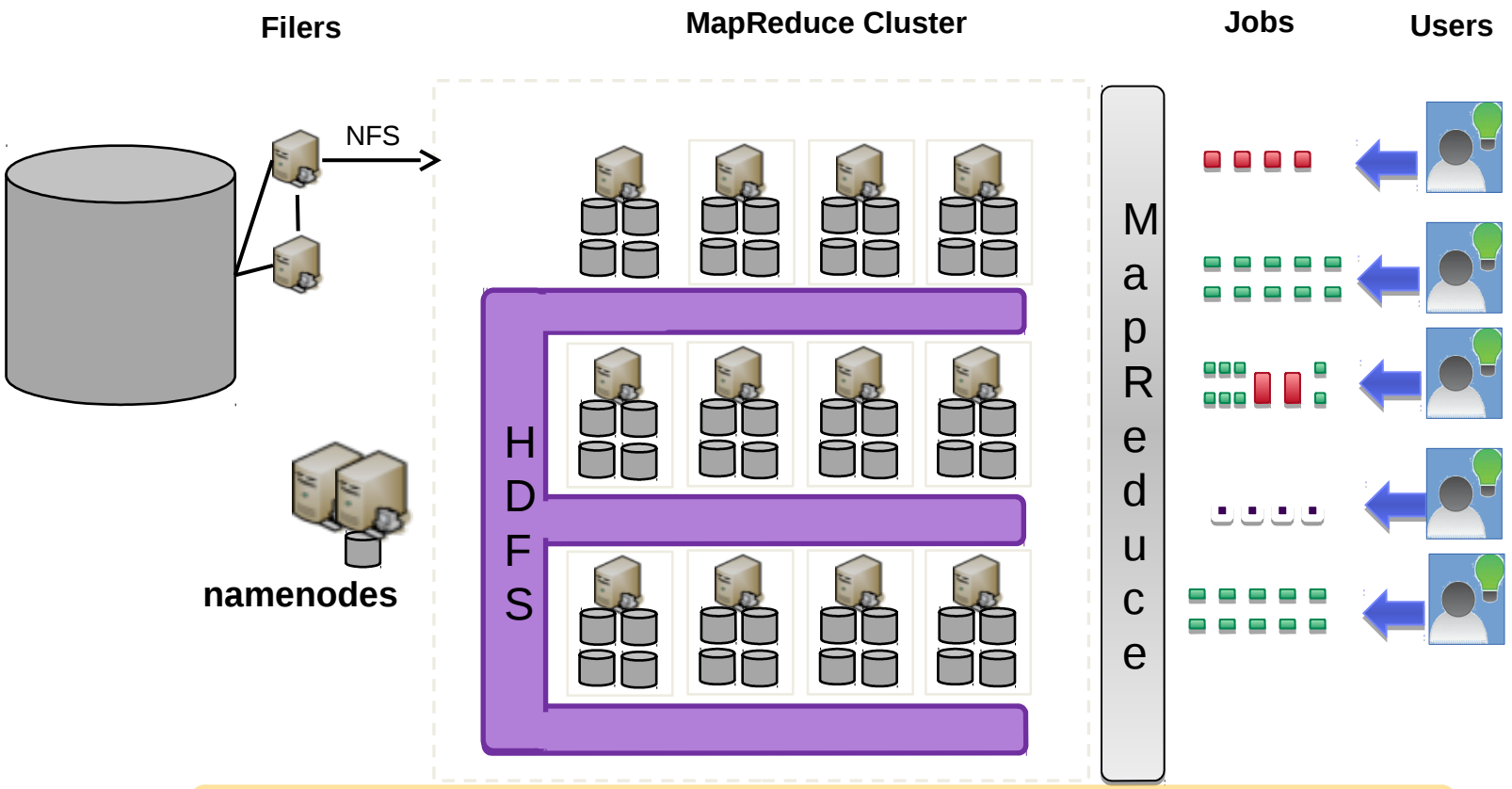
- **Hadoop Distributed File System = HDFS** : where Hadoop stores the data
  - This file system spans all the nodes in a cluster with **locality awareness**

- **Hadoop computation model = MapReduce**
  - Data stored in a distributed file system spanning many inexpensive computers
  - Bring function to the data
  - Distribute application to the compute resources where the data is stored



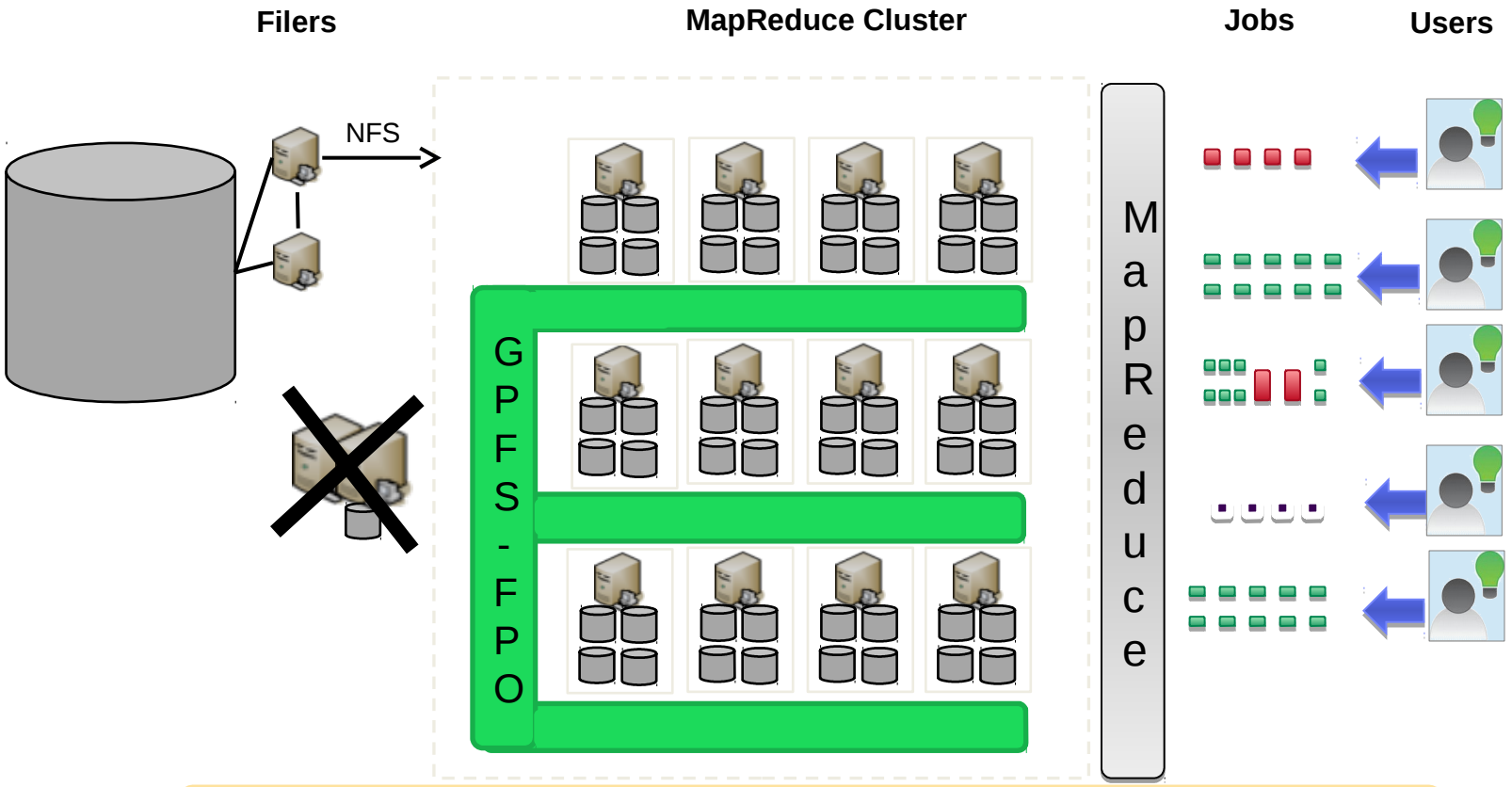
1. **Map Phase**
   (break job into small parts)
2. **Shuffle**
   (transfer interim output for final processing)
3. **Reduce Phase**
   (boil all output down to a single result set)

A Typical HDFS Environment



- Uses disk local to each server
- Aggregates the local disk space into a single, redundant shared file system
- The open source standard file systems used in partnership with Hadoop MapReduce

MapReduce Environment Using GPFS-FPO (File Placement Optimizer)



- Uses disk local to each server
- Aggregates the local disk space into a single redundant shared file system
- Designed for MapReduce workloads
- Unlike HDFS, GPFS-FPO is POSIX compliant – so data maintenance is easy
- **Intended as a drop in replacement for open source HDFS** (IBM BigInsights product may be required)

Agenda

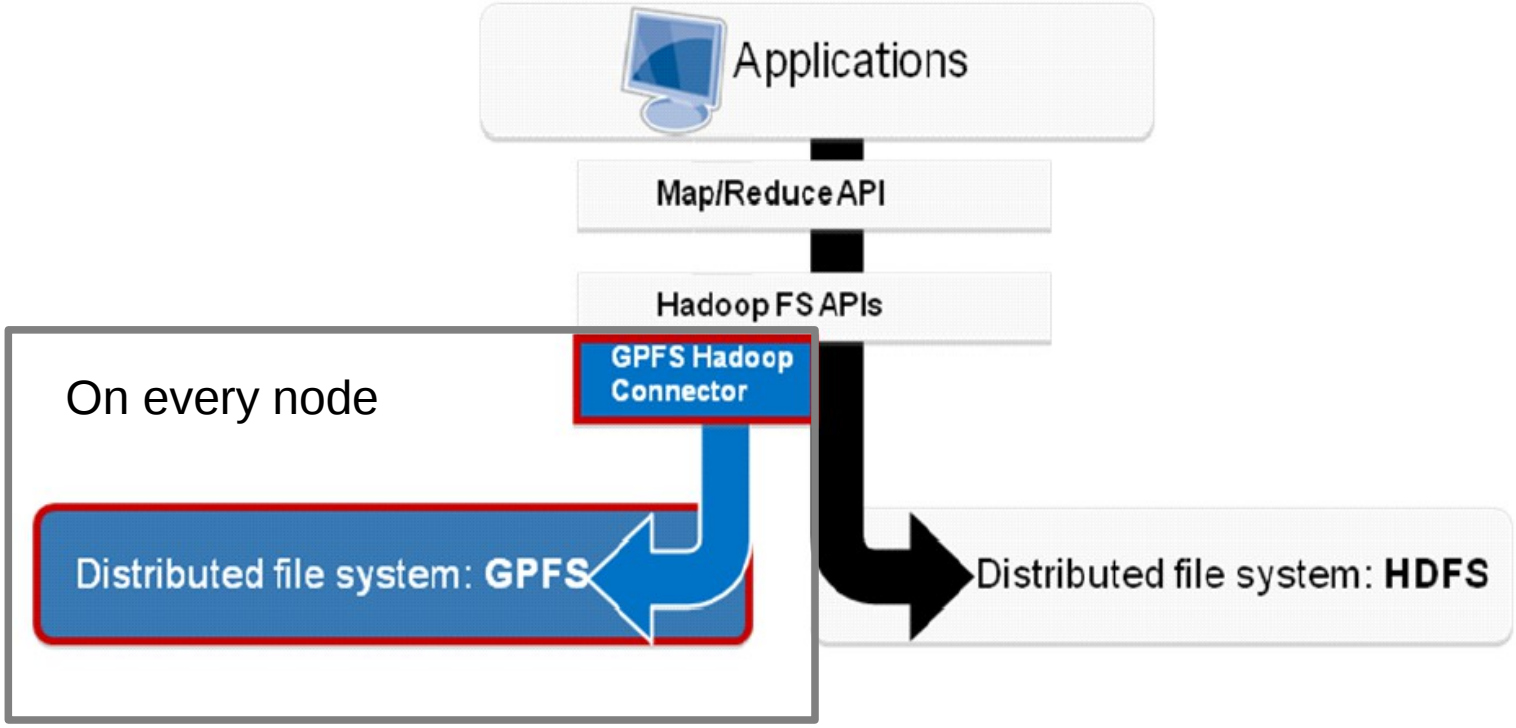
1.) Hadoop, HDFS and SpectrumScale
short - overview

2.) Hadoop Connector
a.) „old“-version
b.) „new“-transparency

3.) Configuration overview

4.) Need to know - what else comes with SpectrumScale

SpectrumScale - “old” connector

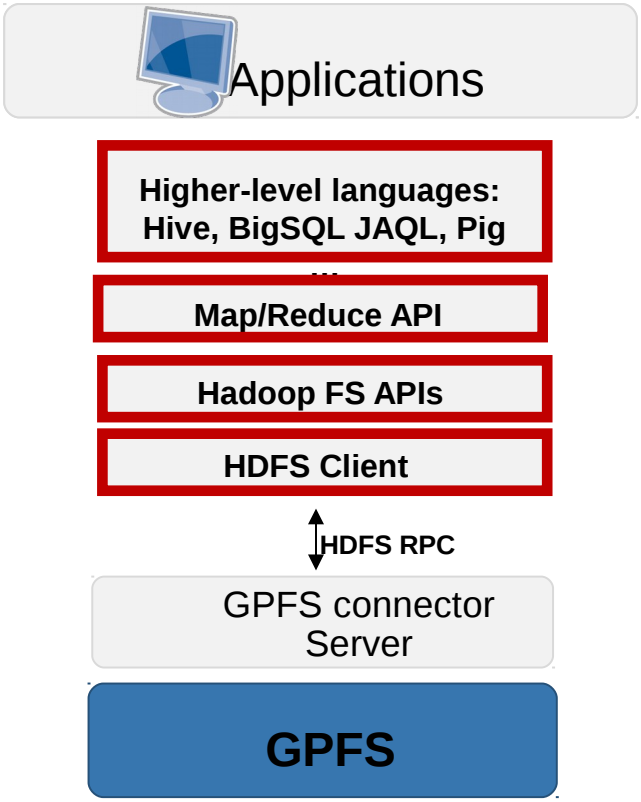


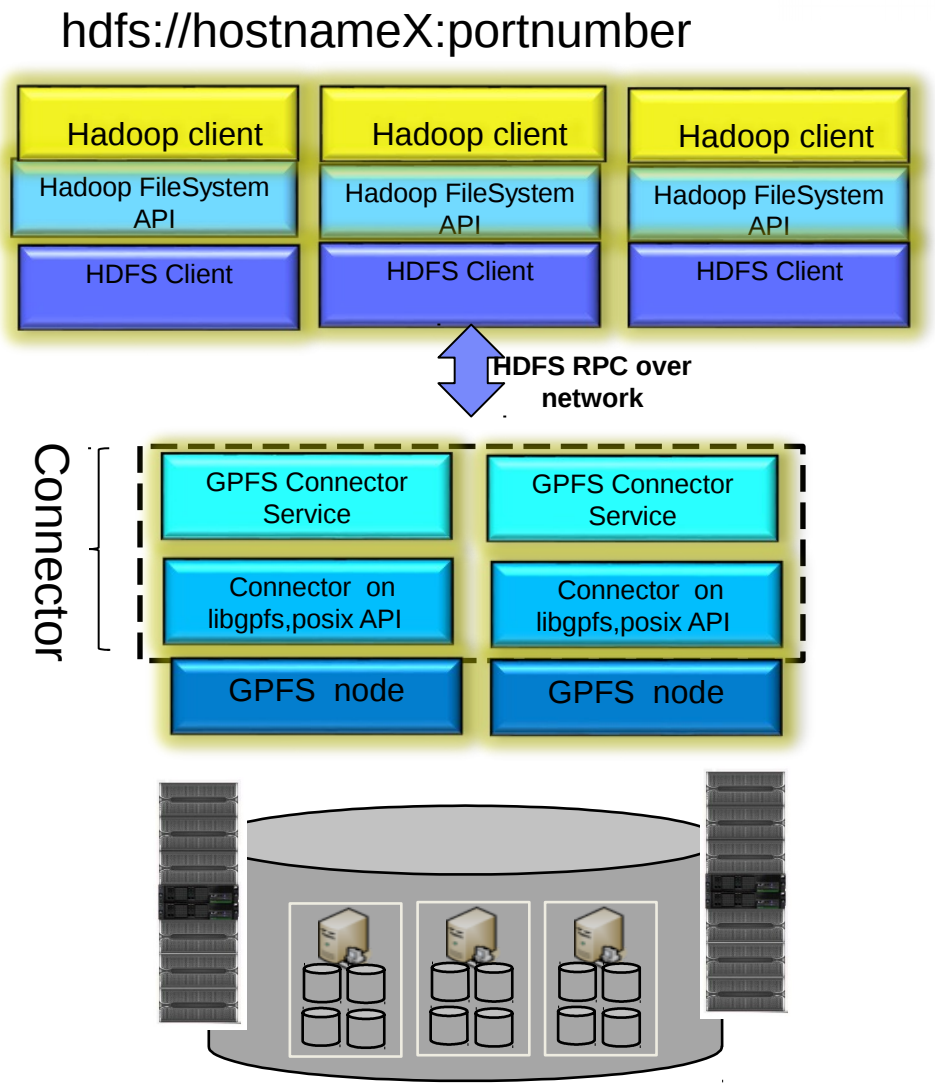
2.5.1 IBM Spectrum Scale (GPFS) Hadoop connector 2.7

gpfs.hadoop-connector-2.7.0-6	2016/2/29	x86_64/rpm x86_64/deb	ppc64/rpm	ppc64le/rpm ppc64le/deb	fixed two major issues: 1. the user root can't submit hive jobs 2. hive's ACL issues with proxy user (**Important: refer upgrade section** and **POSIX ACL**)

https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/General%20Parallel%20File%20System%20%28GPFS%29/page/Hadoop%20Connector%20Download%20%26%20Info?section=OverviewOfConnector

SpectrumScale - “new” connector - transparency

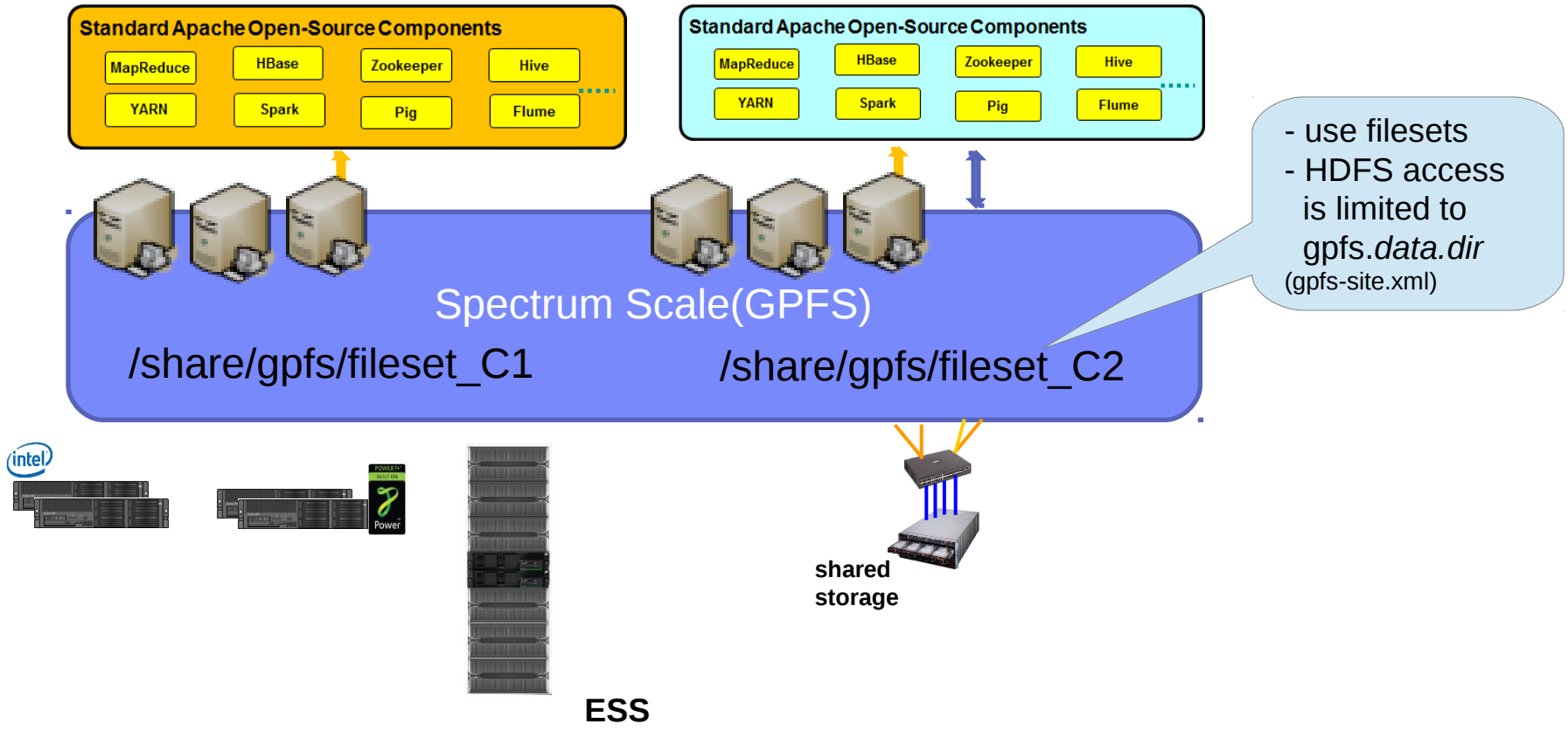




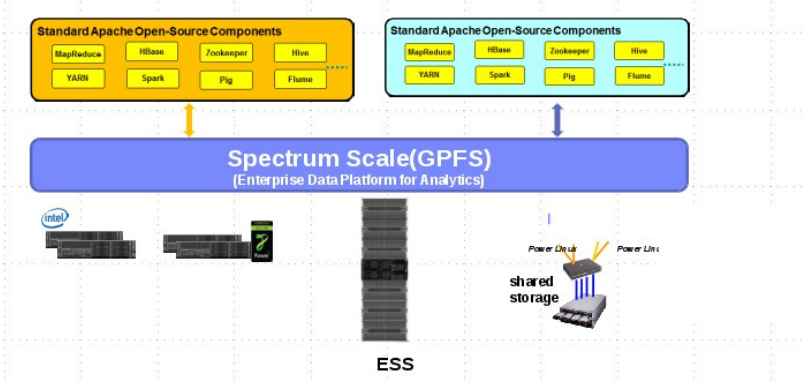
Multiple Hadoop Cluster over the same SpectrumScale FS

HDFS transparency

- **Multiple Hadoop cluster over the same GPFS file system for different application isolation(e.g. node1/2/3 for Hadoop cluster1; node4/5/6 for Hadoop cluster2)**
- **GAed in 2015/11/20**



Multiple Hadoop Cluster over the same SpectrumScale FS



- space and quota management

- storage tiers (slow, fast, flash…)

- fall back for update/upgrade scenarios of HadoopCluster
  (by re-linking fileset)

- scaling effects ..

- data ingest / extract with POSIX

- lower RAIDoverhead

Agenda

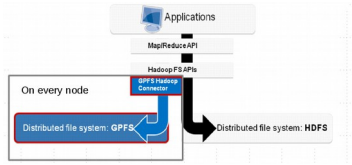
1.) Hadoop, HDFS and SpectrumScale
short - overview

2.) Hadoop Connector
a.) „old“-version
b.) „new“-transparency

3.) Configuration overview

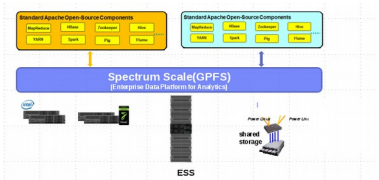
4.) Need to know - what else comes with SpectrumScale

SpectrumScale - connector versions and packaging



„old“- version

gpfs.hadoop-connector-2.7.0-6.x86_64.rpm



„new“- version / transparency

gpfs.hdfs-protocol-2.7.0-1.x86_64.rpm

- HDFS transparency is decoupled from GPFS, (not shipped in GPFS 4.2 package)
- Down load the connector from IBM developerWorks GPFS wiki
  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/General%20Parallel%20File%20System%20%28GPFS%29/page/Hadoop%20Connector%20Download%20%26%20Info?section=2.4.1GPFSHadoopConnector2.7
- GPFS 4.1/4.1.1/4.2+ can work with HDFS transparency
- Guide will be available from GPFS connector homepage in IBM developerWorks GPFS wiki

HDFS Transparency – supported environments

Table 3.2.1 GPFS HDFS Transparency Linux support on x86_64

	REDHAT		SLES		
	RHEL6	RHEL7	SLES 11 SP3+	SLES 12	Ubuntu 14
GPFS 4.1.1+	yes	yes	yes	yes	yes

Table 3.2.2 GPFS HDFS Transparency Linux support on Power(big endian)

	RHEL	
	RHEL6	RHEL7+
GPFS 4.1.1	No	Yes

ubuntu is not supporte for power big endian. RHEL6 and SLES11 are not supported because compiling the packages needs GCC4.7+

Table 3.2.3 GPFS HDFS Transparency Linux support on PowerLE

	RHEL	SLES	Ubuntu
	RHEL7+	SLES12+	Ubuntu 14+
GPFS4.1.1+	Yes	Yes	Yes

GPFS HDFS Transparency - implementation details

Supported Hadoop Version

- Hadoop 2.7.x(fully tested)
- Hadoop 2.6.x(sniff tested)
- Hadoop 2.x other than 2.6/2.7(the compatibility is ensured by community)

Key Advantages (for transparency)

- HDFS hard-coded workloads can run, e.g. impala, webHDFS etc
- Leverage HDFS client cache for better performance
- No need to install GPFS client on Hadoop computing nodes
- Fully Kerberos support in Hadoop ecosystem

Implementation Guide

https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/General%20Parallel%20File%20System%20%28GPFS%29/page/HDFS_Transparency

HDFS Transparency – configuration overview

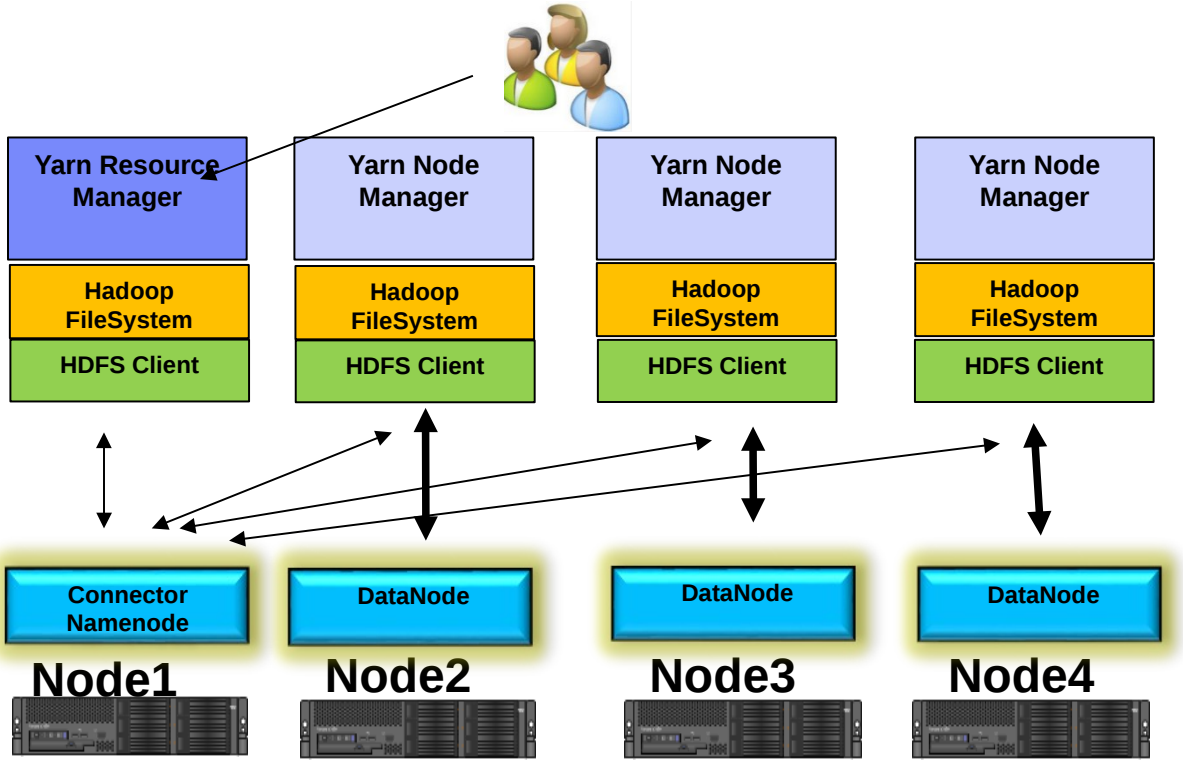
- Having a Hadoop / IOP cluster
  - stop IOP,
  - optionally : recycle hdfs
- Having an up n running GPFS cluster / filesystem

- Customize in Ambari GUI / directly
  - hdfs-site.xml
  - core-site.xml

- install **gpfs.hdfs-protocol-2.7.0-0.<arch>.rpm**
- /usr/lpp/mmfs/hadoop/sbin/mmhadoopctl connector syncconf <your hadoop config dir>
- cd /usr/lpp/mmfs/hadoop/etc/hadoop;
  cp gpfs-site.xml.template gpfs-site.xml
- Modify ***gpfs-site.xml*** according mount point and data directory
- configure slaves
- /usr/lpp/mmfs/hadoop/sbin/mmhadoopctl connector start|stop

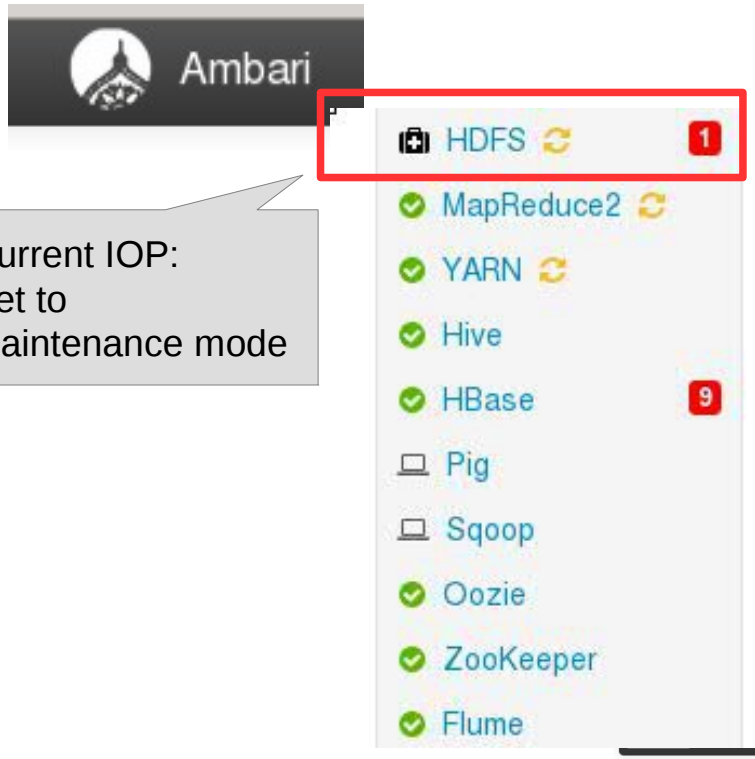
Note: connector logs are stored under /usr/lpp/mmfs/hadoop/logs/.

Hadoop Job Execution with gpfs.hdfs-protocol

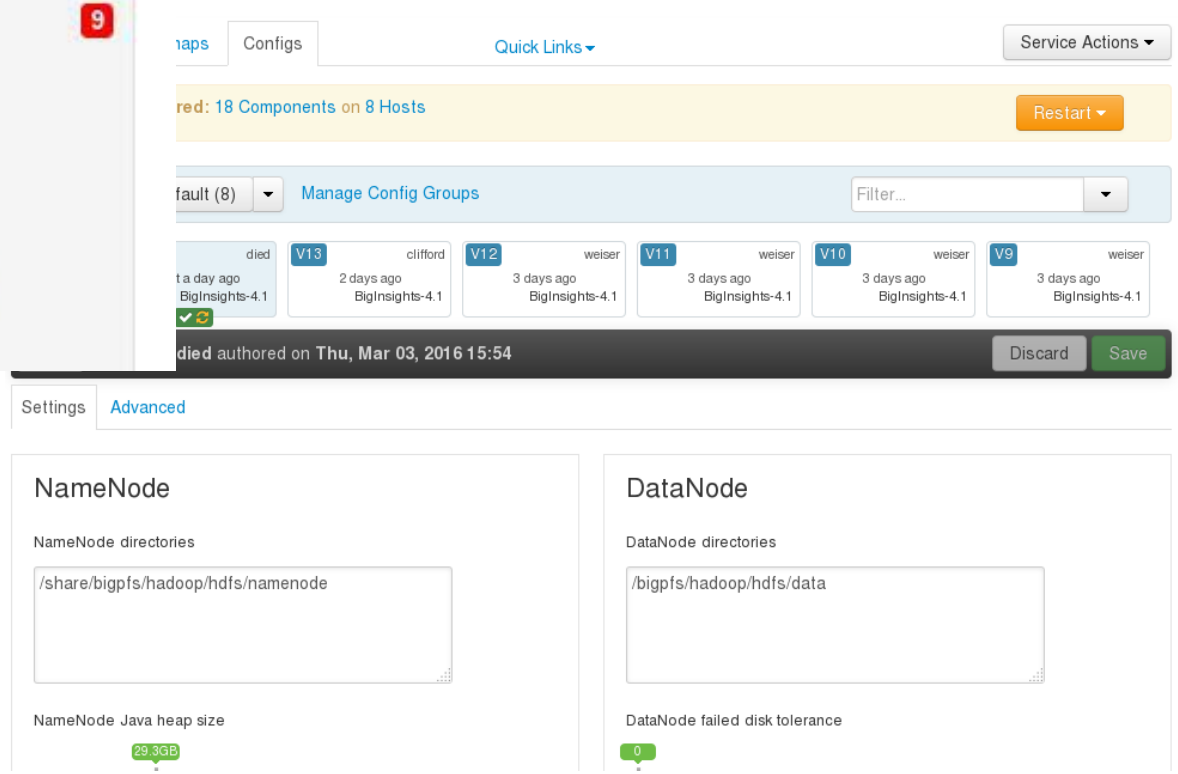


- Configure (connector) name node or nameNodeHA
- Configure data nodes (regular GPFS clients, sharing te same config gpfs-site.xml)
- Syncronize config
- Start connector
- Start applications

HDFS Transparency – configuration overview



Current IOP:
Set to
maintenance mode



Agenda

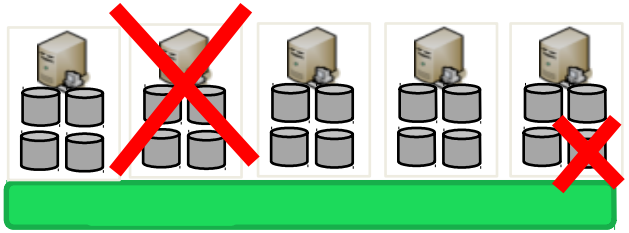
1.) Hadoop, HDFS and SpectrumScale
short - overview

2.) Hadoop Connector
a.) „old“-version
b.) „new“-transparency

3.) Configuration overview

4.) Need to know - what else comes with SpectrumScale

SpectrumScale – special configuration in FPO environments



Characteristics / considerations:

- low budget hardware
- lots of notes
- lots of drives
- MTBF / AFR of a physical disk drive

- SpectrumScale comes with a preconfigured automatism for restriping data
- (nodes) / disks, exceeding a specified counter (when down), will be emptied by restriping data with remaining drives
- Configurations change to cluster config / node based
- Control mechanism for disk local/remote/ fastest disk access

restripeOnDiskFailure

[root@n1 ~]# mmfsadm dump config | grep -i **restripeOnDiskFailure**
[root@n1 ~]#

[root@n1 ~]# mmchconfig **restripeOnDiskFailure**=yes
```
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all
  affected nodes.
```

[root@n1 ~]# mmlsconfig | grep -i restripeOnDiskFailure
restripeOnDiskFailure yes
[root@n1 ~]#
[root@n1 ~]# mmfsadm dump config | grep -i **restripeOnDiskFailure**
[root@n1 ~]#

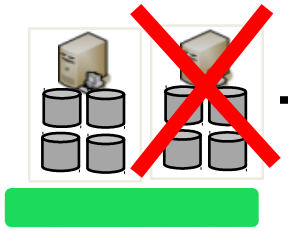
SpectrumScale – restripeOnDiskFailure

```
[root@n1 ~]# mmlscallback system | grep  Disk -A 5
gpfsRecoverFailedDisk
        command       = /usr/lpp/mmfs/bin/mmcommon
        priority      = 1
        sync          = false
        event         = diskFailure
        parms         = recoverFailedDisk %fsName %diskName

gpfsRestartDownDisks
        command       = /usr/lpp/mmfs/bin/mmcommon
        priority      = 1
        sync          = false
        event         = nodeJoin
        parms         = restartDownDisks %myNode %clusterManager %eventNode

gpfsStopFailedDisk
        command       = /usr/lpp/mmfs/bin/mmcommon
        priority      = 1
        sync          = false
        event         = nodeLeave
        parms         = stopFailedDisk %myNode %clusterManager %eventNode
```

SpectrumScale – restripeOnDiskFailure



nodeLeave,diskFailure

```
mmchconfig metadataDiskWaitTimeForRecovery=seconds   (default 2400 sec.)
mmchconfig dataDiskWaitTimeForRecovery=seconds       (default 3600 sec.)
mmchconfig minDiskWaitTimeForRecovery=seconds        (default 1800 sec.)
mmchconfig maxDownDisksForRecovery=disks          (default 16 disks)
mmchconfig maxFailedNodesForRecovery=nodes      (default 3 nodes)
```

gpfsRecoverFailedDisk

tschdisk start -a / restripefs -r

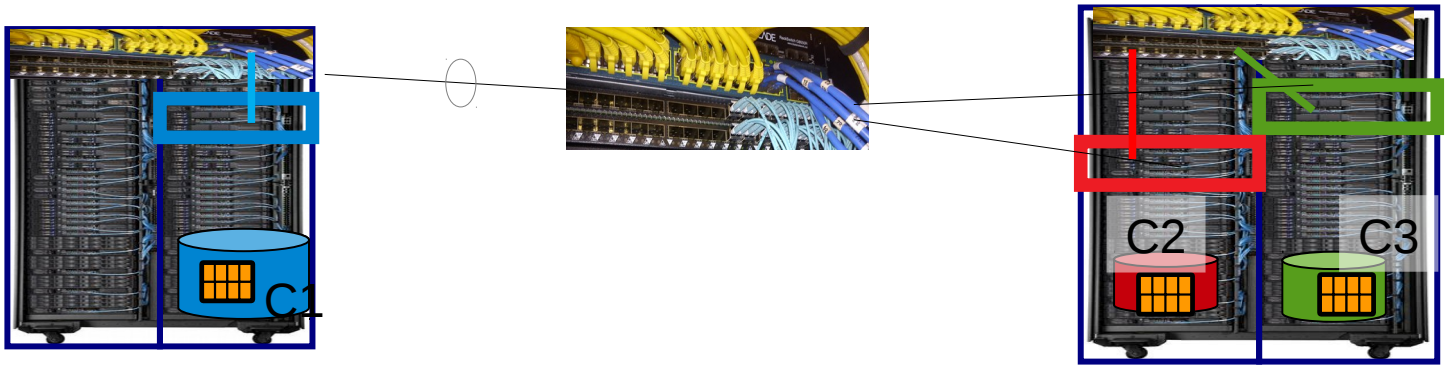
nodeJoin

gpfsRestartDownDisks

files:

/var/adm/ras/mmfs.log.latest
/var/adm/ras/restripefsOnDiskFailure.log.

SpectrumScale - RepWriteStream



Single source replication:

- throughput is bounded by outgoing bandwidth of the source/N

The more Copies to write, the higher bandwidth is needed

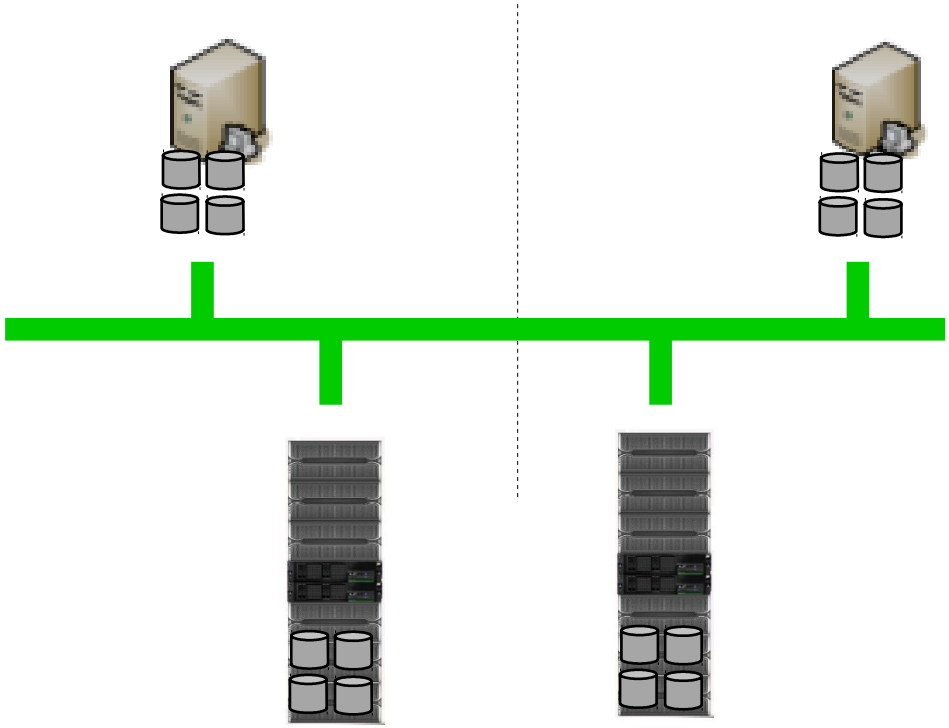
Pipelined replication

```
[root@n1 ~]# mmchconfig enableRepWriteStream=yes
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@n1 ~]#
```

SpectrumScale – avoid read from remote copy

Old: mmchconfig readReplicaPolicy=local -i -N all

New:
mmchconfig readReplicaPolicy=fastest -i -N all



read from fastest disk (1 / 2)

- In a file system with replicas > 1,
- before doing a read, we can choose, depending on some rules,

  from which disk to read the data so as to get better performance.
- termed "read replica policy".

...these rules...

In the older implementation(s), there is only one such rule called "LOCAL".

The LOCAL rule instructs gpfs to choose the replica that is closer to the node that has issued the read, where "closer" means:

1. Prefer locally attached disk over NSD servers
2. Among NSD servers, prefer the server that is on the same subnet as the node issuing the read over the one on a different subnet

read from fastest disk (2 / 2)

configure read policy and set the related parameter
- we can name the current read replica policy as "default"
- this newly designed as "fastest",
- changeable

```
mmchconfig readReplicaPolicy=fastest -i
mmchconfig readReplicaPolicy=default -i
mmlsconfig readReplicaPolicy
```

read from fastest disk - further details

For the fastest policy, we also have some configure settings to tune its behavior:

mmchconfig fastestPolicyNumReadSamples=xx –i
mmchconfig fastestPolicyCmpThreshold=xx -i
mmchconfig fastestPolicyMaxValidPeriod=xx -i
mmchconfig fastestPolicyMinDiffPercent=xx –I

fastestPolicyNumReadSamples	[3 ~ 100], the default is 5	how many latest read samples we take to evaluate the disk's recent speed
fastestPolicyCmpThreshold	range >= 3, the default is 50	if a disk's comparison count becomes greater than this value, we'll force to select this disk as the preferred disk to read so as to update its current evaluation speed
fastestPolicyMaxValidPeriod	range >= 1 and in unit of seconds, the default is 600 (i.e. 10 min)	after this period of time, the disk's current speed evaluation is considered invalid if a disk's comparison count becomes greater than this value, we'll force to select this disk as the preferred disk to read so as to update its current evaluation speed
fastestPolicyMinDiffPercent	[0, 100], the default is 50	how we judge which disk is considered fast between two disks

Agenda

1) Hadoop, HDFS and SpectrumScale
   short - overview

2) Hadoop Connector
   a.) „old“-version
   b.) „new“-transparency

3) Configuration overview

4) Need to know - what else comes with SpectrumScale

olaf.weiser@de.ibm.com
IBM Deutschland
SpectrumScale Support Specialist

Disclaimer

- The information in this document may be **IBM CONFIDENTIAL**.
- This information is provided on an "AS IS" basis without warranty of any kind, express or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. Some jurisdictions do not allow disclaimers of express or implied warranties in certain transactions; therefore, this statement may not apply to you.
- This information is provided for information purposes only as a high level overview of possible future products. PRODUCT SPECIFICATIONS, ANNOUNCE DATES, AND OTHER INOFORMATION CONTAINED HEREIN ARE SUBJECT TO CHANGE AND WITHDRAWAL WITHOUT NOTICE.
- USE OF THIS DOCUMENT IS LIMITED TO SELECT IBM PERSONNEL AND TO BUSINESS PARTNERS WHO HAVE A CURRENT SIGNED NONDISCLUSURE AGREEMENT ON FILE WITH IBM. THIS INFORMAITON CAN ALSO BE SHARED WITH CUSTOMERS WHO HAVE A CURRENT SIGNED NONDISCLOSURE AGREEMENT ON FILE WITH IBM, BUT THIS DOCUMENT SHOULD NOT BE GIVEN TO A CUSTOMER EITHER IN HARDCOPY OR ELECTRONIC FORMAT.
- Important notes:
- IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.
- IBM makes no warranties, express or implied, regarding non-IBM products and services, including but not limited to Year 2000 readiness and any implied warranties of merchantability and fitness for a particular purpose. IBM makes no representations or warranties with respect to non-IBM products. Warranty, service and support for non-IBM products is provided directly to you by the third party, not IBM.
- All part numbers referenced in this publication are product part numbers and not service part numbers. Other part numbers in addition to those listed in this document may be required to support a specific device or function.
- MHz / GHz only measures microprocessor internal clock speed; many factors may affect application performance. When referring to storage capacity, GB stands for one billion bytes; accessible capacity may be less. Maximum internal hard disk drive capacities assume the replacement of any standard hard disk drives and the population of all hard disk drive bays with the largest currently supported drives available from IBM.
- IBM Information and Trademarks
- The following terms are trademarks or registered trademarks of the IBM Corporation in the United States or other countries or both: the e-business logo, IBM, xSeries, pSeries, zSeries, iSeries.
- Intel, Pentium 4 and Xeon are trademarks or registered trademarks of Intel Corporation. Microsoft Windows is a trademark or registered trademark of Microsoft Corporation. Linux is a registered trademark of Linus Torvalds. Other company, product, and service names may be trademarks or service marks of others.