The Parallel File System HP SFS/Lustre on xc2

Roland Laifer

Computing Centre (SSCK)
University of Karlsruhe
Germany

Laifer@rz.uni-karlsruhe.de

Outline

- What is Lustre?
- What is HP SFS?
- Overview of HP SFS on xc2
- Properties of the different file systems
- Restrictions for using HP SFS
- Performance diagrams
- IO performance monitoring
- Backup and archiving
- Quota
- How does striping work?
- Possible optimization with striping parameters

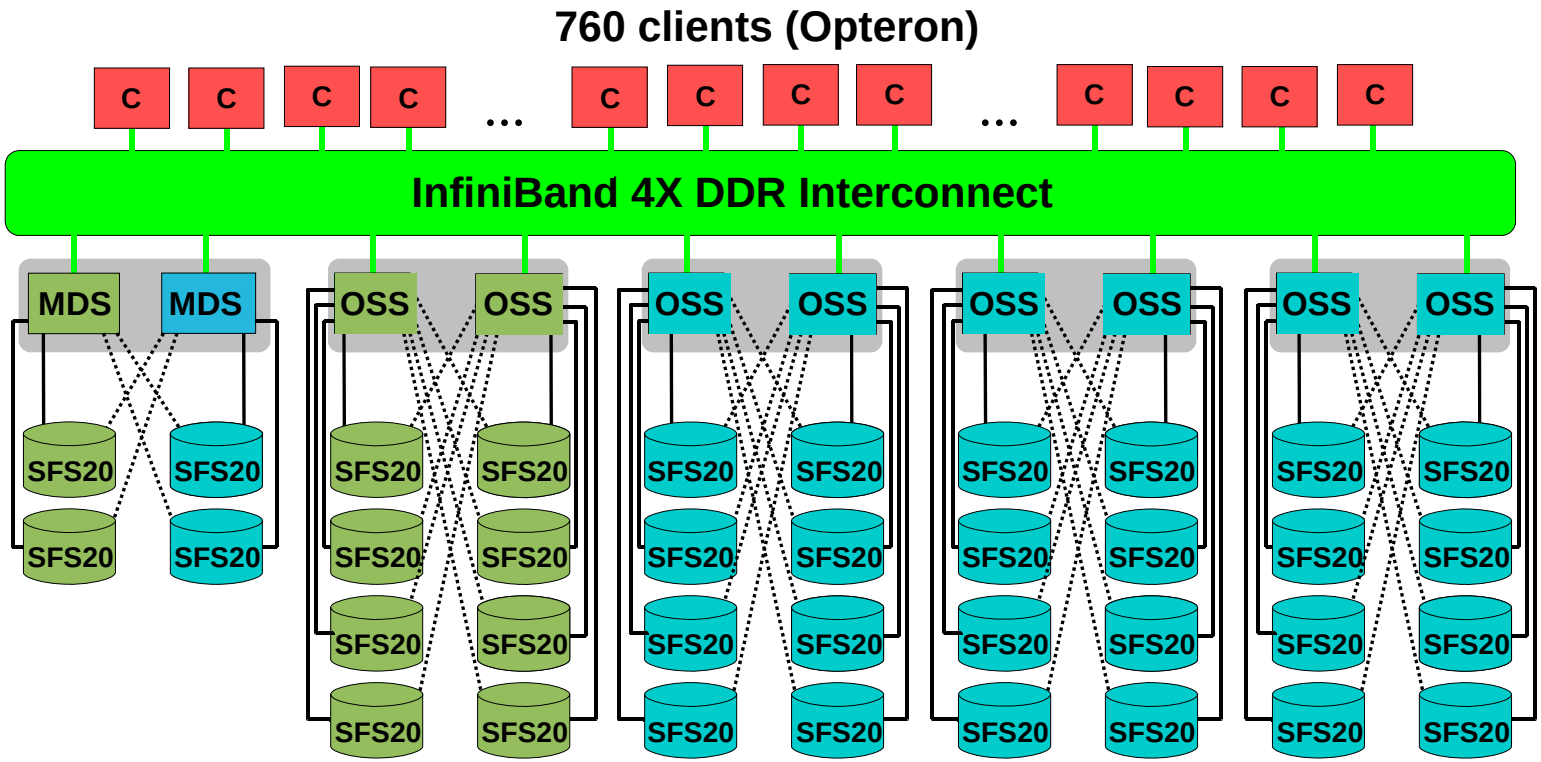
What is Lustre?

- "Lustre" is an amalgam of the terms "Linux" and "Clusters "
- Lustre is a scalable high performance file system for Linux
- Main development by Cluster File Systems, Inc. (CFS)
  - Roadmap, FAQs and source code at http://www.clusterfs.com/
    - Lustre products are available from many vendors
- Pros and Cons
  - \+ Runs very stable
    - User base is rapidly growing
    - Scalable up to 10000's of clients
    - Allows failover support on servers
  - \+ Excellent throughput and metadata performance
    - High throughput with multiple network protocols
  - \+ POSIX file system semantics
  - – Administration is not easy
  - – Currently supports only Linux clients

What is HP SFS?

- **HP SFS means HP StorageWorks Scalable File Share**
  - **Our experiences: http://www.rz.uni-karlsruhe.de/dienste/lustretalks**
- **HP SFS is a Lustre appliance from HP**
  - **Only dedicated hardware is supported:**
    - **Servers are Xeon based Proliant systems from HP**
    - **Storage arrays are SFS20 with SATA disks or EVA3000 with FC disks**
  - **Includes a special software package**
- **Advantages of HP SFS software**
  - **HP supplies a hardened Lustre version**
  - **Includes additional software**
    - **for failover and management**
    - **for sending problem alerts and to verify the system's health**
    - **for performance monitoring**
    - **client build kits and client rpm packages**
  - **Easy installation, configuration and upgrade**
  - **Good support**

Overview of HP SFS on xc2



	$HOME	$WORK
Capacity	8 TB	48 TB
Total write / read perf.	360 / 800 MB/s	2100 / 3200 MB/s
Single client write / read	360 / 320 MB/s	400 / 320 MB/s

Note:

- **$HOME file system is mirrored**

Properties of the different file systems

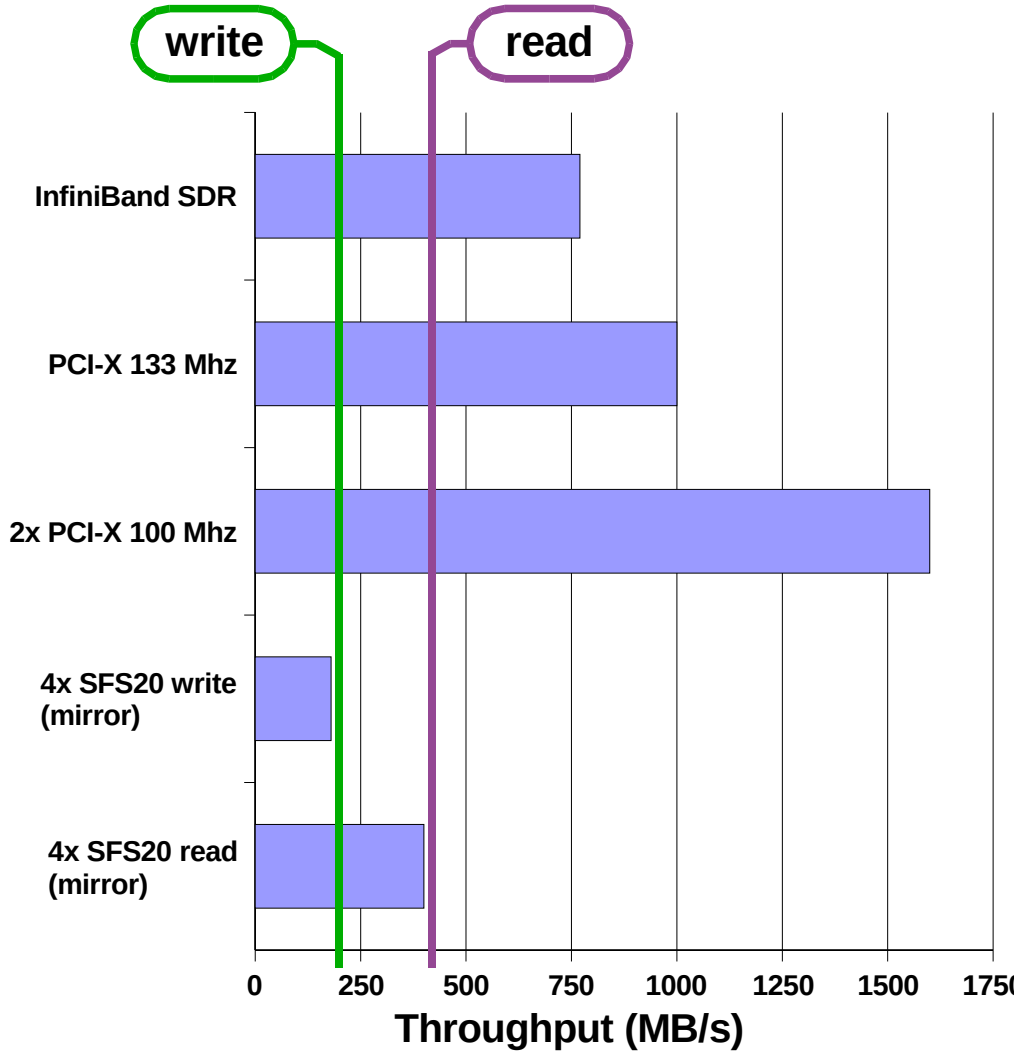
Property	$TMP	$HOME	$WORK
Visibility	local	global	global
Lifetime	batch job	project	> 7 days
Capacity	70 GB	8 TB	48 TB
Quotas	no	planned	no
Backup	no	yes (default)	no
Read perf. / node	60 MB/s	320 MB/s	320 MB/s
Write perf. / node	60 MB/s	360 MB/s	400 MB/s
Total read perf.	n*60 MB/s	800 MB/s	3200 MB/s
Total write perf.	n*60 MB/s	360 MB/s	2100 MB/s

Restrictions for using HP SFS

- **Nearly everything works like on a local file system !**
- **Unsupported features**
  - flock(2) **does not lock files over Lustre**
    - **Identical to behaviour of NFS**
  - **Direct IO (**O_DIRECTIO **option) is not supported**
    - **Rarely used in applications**
  - ***atime* is not always accurate**
    - **Similar to other parallel file systems due to performance reasons**
    - tail –f **and** ls –l **does not always show latest data**
  - **Total metadata performance is limited to several 1000 operations / sec**
    - **Operations are open, close, create, delete, stat**
- **Good practise is to**
  - **omit lots of metadata operations**
  - **write or read data in large blocks**
  - **change striping parameters if lots of clients use one very large file**

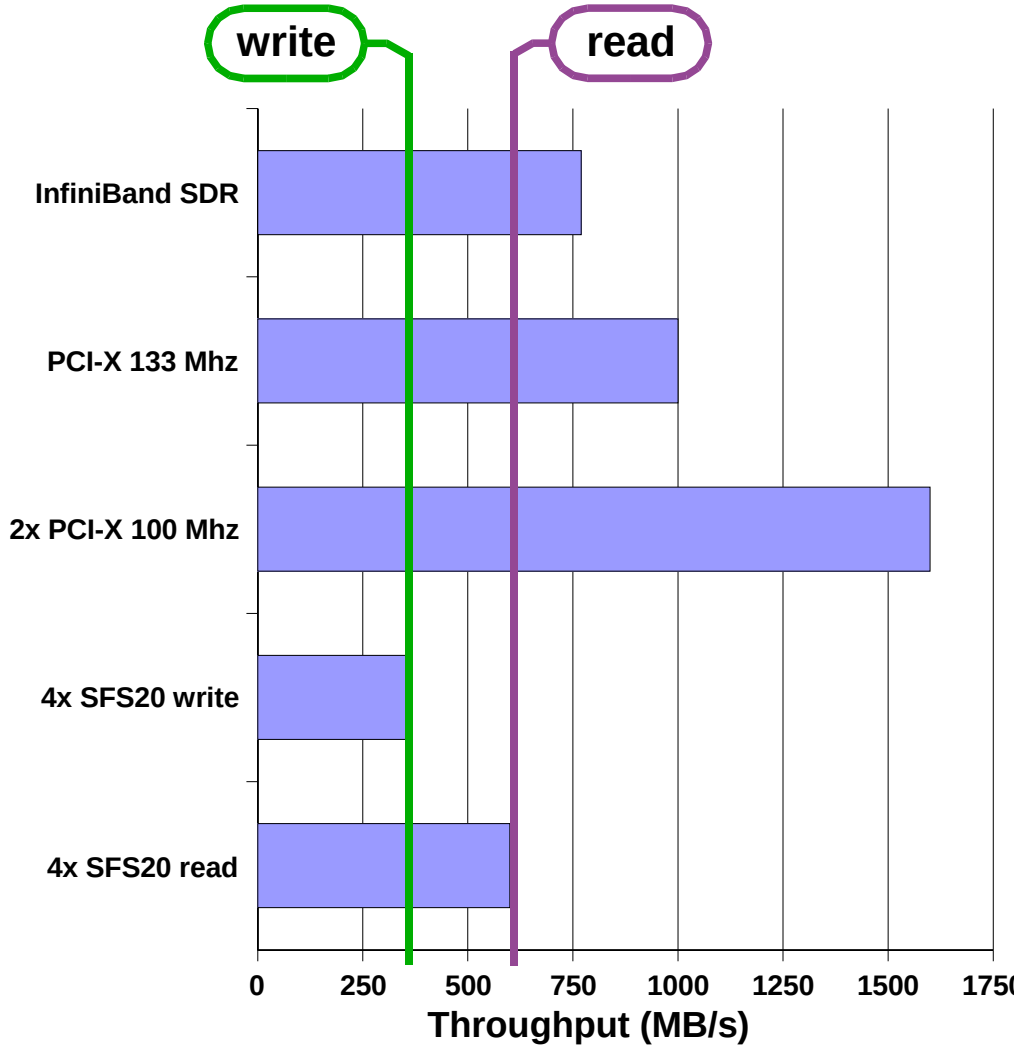
Performance of one OSS for $HOME

- InfiniBand SDR
  - 770 MB/s measured
- 133 MHz PCI-X bus for IB
  - About 1000 MB/s
- 2x 100 MHz PCI-X bus for SCSI adapters
  - About 1600 MB/s
- 4x SFS20 storage array
  - About 180 MB/s for writes
    - Mirrored, RAID6
  - About 400 MB/s for reads
    - Mirrored, RAID6

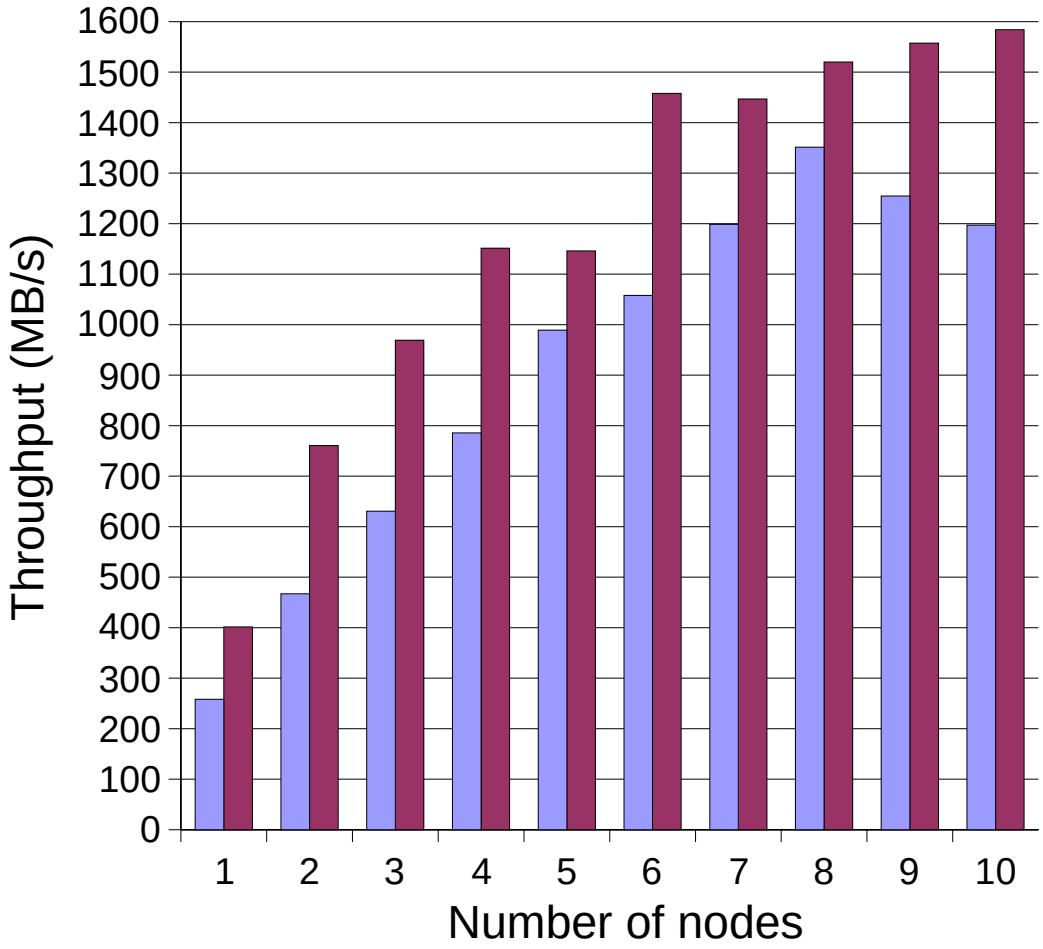


Performance of one OSS for $WORK

- InfiniBand SDR
  - 770 MB/s measured
- 133 MHz PCI-X bus for IB
  - About 1000 MB/s
- 2x 100 MHz PCI-X bus for SCSI adapters
  - About 1600 MB/s
- 4x SFS20 storage array
  - About 360 MB/s for writes
    - RAID6
  - About 600 MB/s for reads
    - RAID6

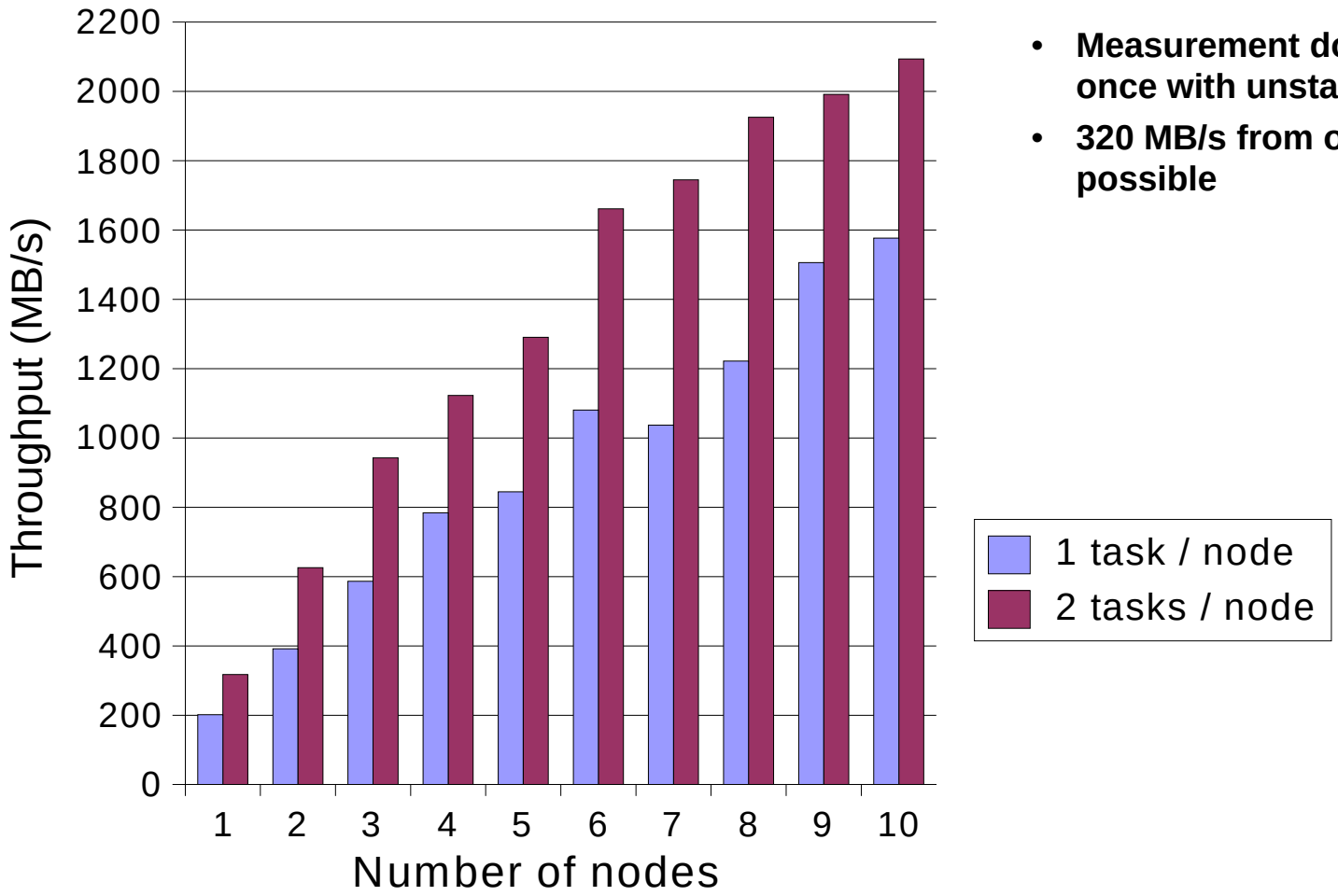


Write performance of $WORK



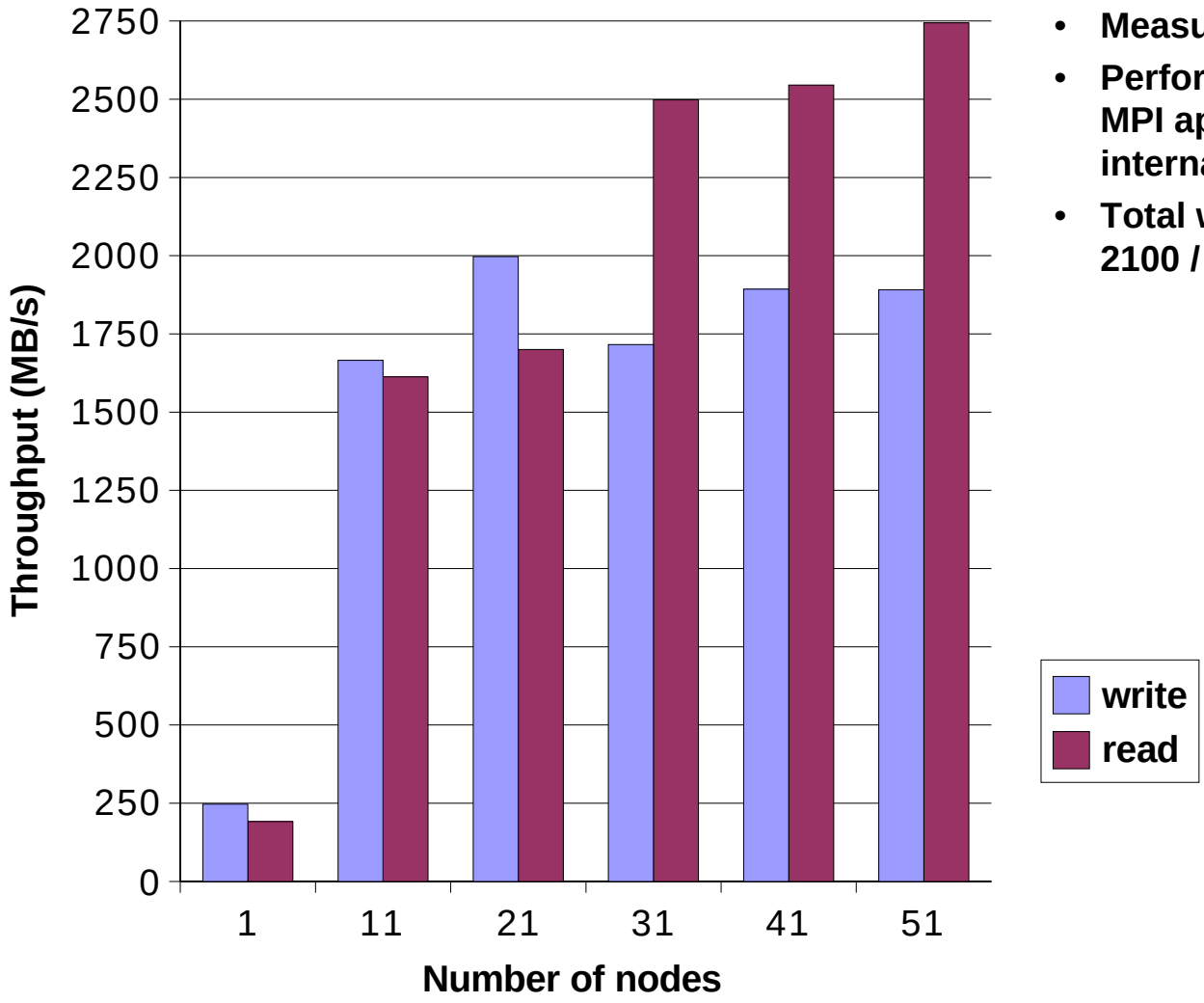
- **Measurement done only once with unstable switch**
- **360 MB/s from one task is possible**

Read performance of $WORK



- **Measurement done only once with unstable switch**
- **320 MB/s from one task is possible**

Write / read scalability of $WORK



- **Measurement done only once**
- **Performance impact by other MPI applications due to internal IB link saturation**
- **Total write / read performance 2100 / 3200 MB/s is possible**

IO performance monitoring

- **Unfortunately no easy way for users to do performance monitoring**
  - **Hopefully in later XC version a tool will become available**
    - **Problem is mainly restricted access to batch nodes**
- **Possibly add IO performance measurement to your application**
  - **Measure time for large IO operations in order to get bandwidth**
    - **This is an easy way for portable performance monitoring**
- **Contact SFS admins if you assume IO performance problems**
  - **We have tools to do fine grained performance monitoring**
  - **Possible performance impact of RAID rebuild or other applications**

Backup and Archiving

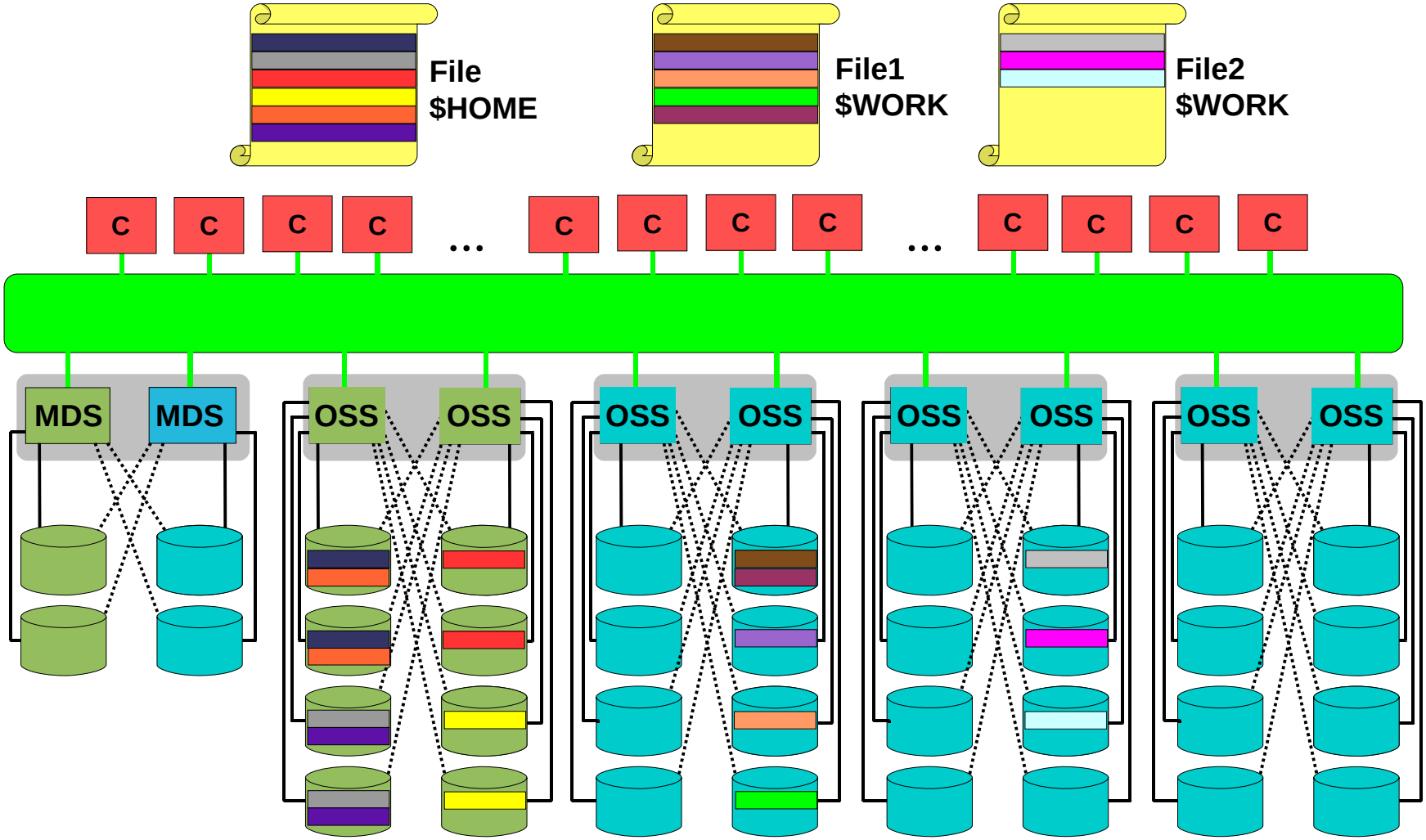
- **Commands to show and restore data of the backup:**
  - tsm_q_backup **shows one, multiple or all files stored in the backup**
  - tsm_restore **restores saved files**
    - **Use option –h to get help**
    - **Only $HOME data runs into the backup**
- **Commands to show, archive and retrieve data:**
  - tsm_q_archive **shows files in the archive**
  - tsm_archiv **archives files**
    - **Files of $HOME or $WORK can be archived**
  - tsm_retrieve **retrieves archived files**
  - tsm_d_archiv **deletes files from the archive**

Quota

- **Show reserved disk space:**
  - kontingent_get **shows reserved amount of disk space of your project**
    - **Initially set depending on your project proposal**
    - **Ask XC hotline if you need more disk space**
    - **Disk space limit (or quota) is not yet enforced**
- **Show quota in future SFS version:**
  - **Syntax:** lfs quota -g <group name> /lustre/data
  - **Example:** lfs quota -g wkrz00 /lustre/data
    Disk quotas for group wkrz00 (gid 40997):

```
     Filesystem  blocks   quota   limit   grace   files   quota   limit   grace
  /lustre/data 117342008       0 1000000000          279957       0       0
xc3-ls-mds1_UUID
                   59856       0  204800          279957       0       0
xc3-ls-ost1_UUID
                  117282152       0 117350400
```

How does striping work?



Possible optimization with striping parameters

- **Change striping parameters:**
  - lfs setstripe <dirname> <stripe size> <stripe start> <stripe count>
    - **Always use -1 as stripe start !**
    - **Has only effect on new files on this directory !**
    - **Changed parameters are not saved in the backup !**
    - **Create own documentation of changes**
- **Small file performance could be improved with stripe count 1**
  - **Example to stripe count 1:** lfs setstripe . 4194304 -1 1
- **Large IO on $WORK slightly improved with stripe count 8**
  - **Example to stripe count 8:** lfs setstripe . 4194304 -1 8
- **If many clients use a > 100 MB file on $WORK use stripe count -1 !**
  - **Example to stripe over all OSTs:** lfs setstripe . 4194304 -1 -1

Possible optimization with striping parameters (cont.)

- **Change stripe size if your application uses a special chunk size**
  - **For MPI-IO best performance was observed with stripe size 16 MB**
  - **Example to change stripe size to 16 MB:** lfs setstripe . 16777216 -1 -1
- **If application uses many files do not change striping parameters**
  - **Default stripe count is 4 and files are automatically distributed**
  - **Check stripe count:** lfs getstripe <filename>

```
OBDS:
   0: xc2-ls-ost5_UUID
…
23: xc2-ls-ost28_UUID
./my_striped_file
        obdidx       objid       objid       group
            12     3430634    0x3458ea           0
            13     3430885    0x3459e5           0
            14     3430874    0x3459da           0
            15     3431045    0x345a85           0
```

  - **Number of lines below** obdidx **shows stripe count**

Questions ?