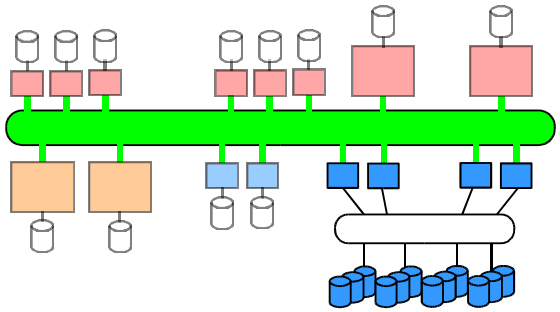
Filesystems on SSCK's HP XC6000

Roland Laifer

Computing Centre (SSCK)
University of Karlsruhe

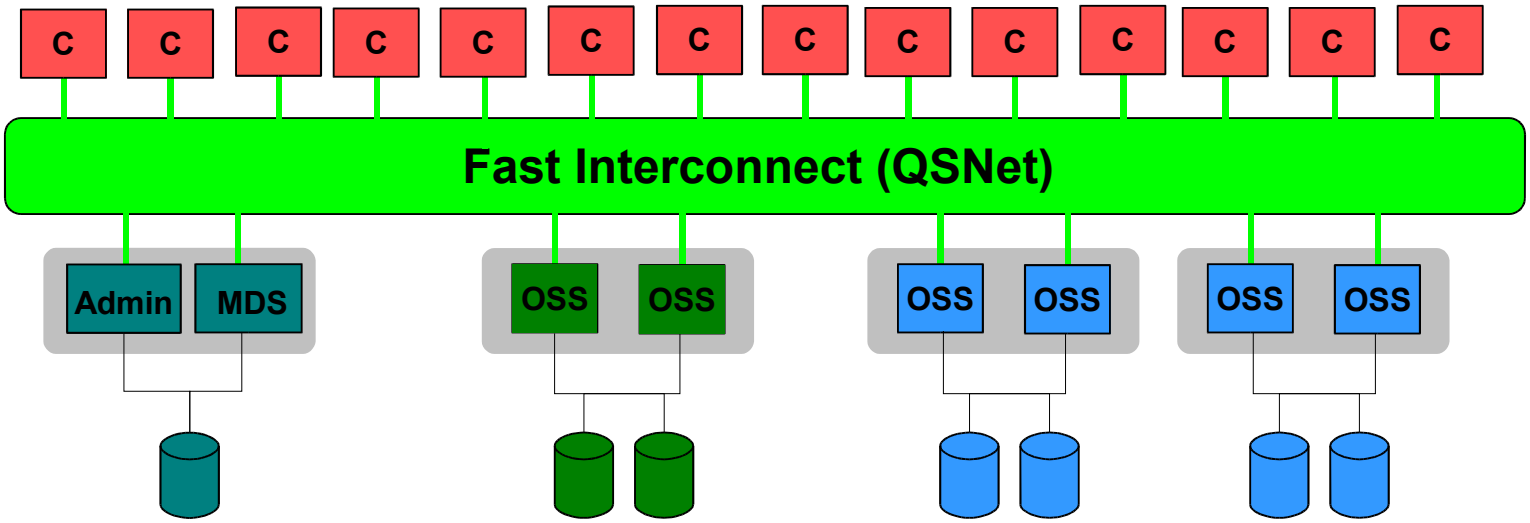
Laifer@rz.uni-karlsruhe.de

Overview



- Overview of HP SFS at SSCK
  - HP StorageWorks Scalable File Share (SFS)
  - based on Lustre from Cluster File Systems, Inc.
- Filesystems and corresponding properties
- Filesystem properties in detail
  - e.g. performance measurements
- Using HP SFS
- Background information about Lustre

HP SFS (Lustre) on SSCK's HP XC6000



MDS and Admin for $HOME and $WORK

- allows > 50 million files
- ~ 5000 file creates / sec

$HOME

- 3,8 TB storage
- ~ 380 MB/s read
- ~ 230 MB/s write

$WORK

- 7,6 TB storage
- ~ 770 MB/s read
- ~ 480 MB/s write

Legend

Admin: Administration Server
MDS: Metadata Server
OSS: Object Storage Server
C: Client

Filesystems and corresponding properties

Property	$TMP	$HOME	$WORK
Visibility	local	global	global
Lifetime	batch job	project	> 7 days
Disk space	> 10 GB	3,8 TB	7,6 TB
Quotas	no	planned	no
Backup	no	yes (default)	no
Read perf. / node	60 MB/s	260 MB/s	300 MB/s
Write perf. / node	60 MB/s	220 MB/s	410 MB/s
Total read perf.	n * 60 MB/s	380 MB/s	770 MB/s
Total write perf.	n * 60 MB/s	230 MB/s	480 MB/s

Visibility, lifetime, and usage

- **Visibility of $HOME and $WORK**
  - **all nodes show the same data**
    - **lock mechanisms guarantee that no data is lost when files are accessed from different nodes**
- **Lifetime and usage**
  - **Files on $TMP are available while the batch job is running.**
    - **$TMP should be used for temporary scratch files.**
    - **2+ MPI tasks per node, i.e. they share one $TMP file system**
  - **Files on $WORK won't be deleted for at least 7 days.**
    - **$WORK should be used for intermediary files.**
  - **Files on $HOME will be available during the project's lifetime.**
    - **$HOME should be used for permanent files.**

Disk space and quotas

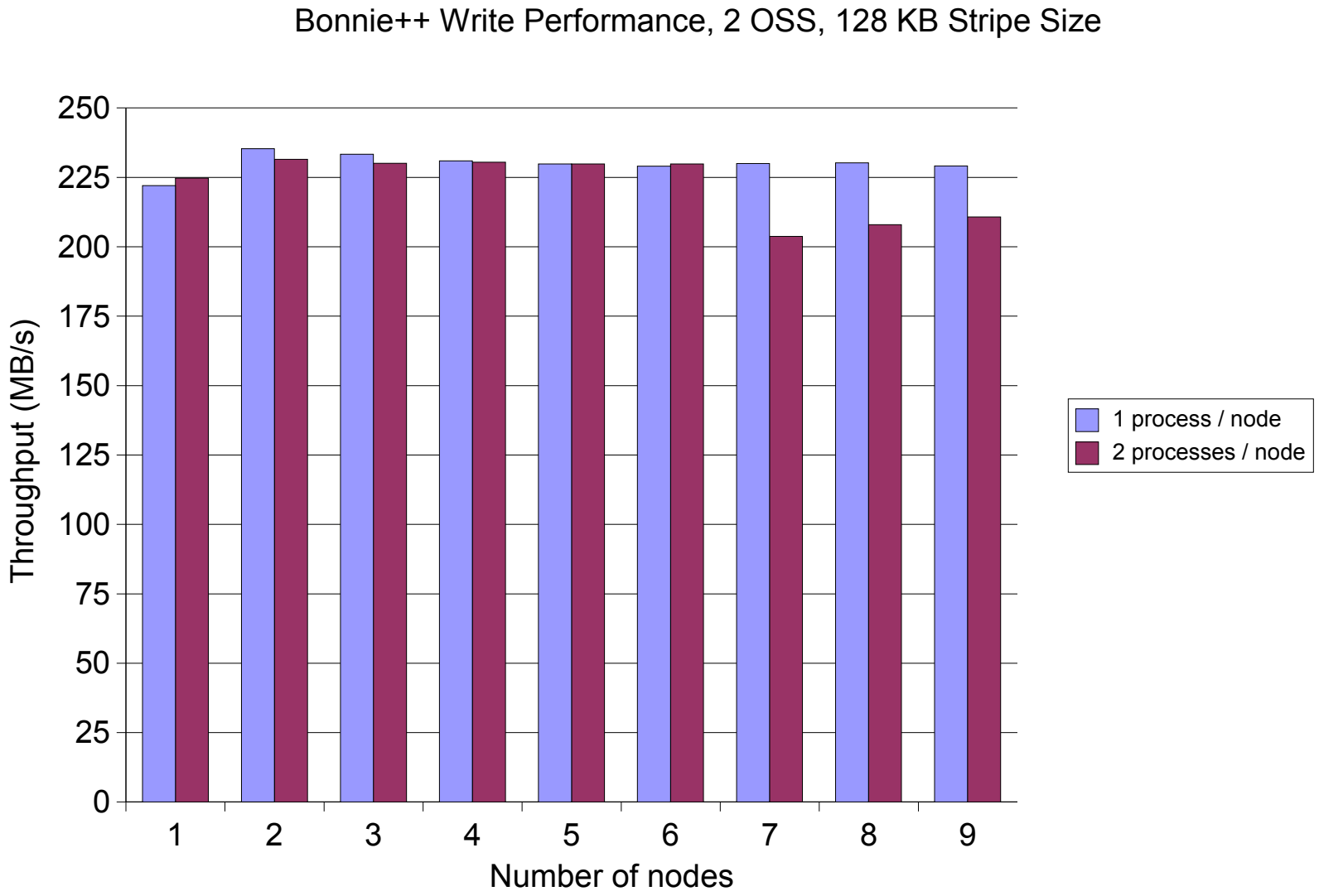
» Disk space and quotas

- **Quotas on $HOME will enforce disk usage on a project basis.**
  - **This feature will probably become available at the end of the year.**
- **The file $HOME/../diskusage reports the used disk space per user group.**
  - **This file is updated every night.**

Backup and archiving

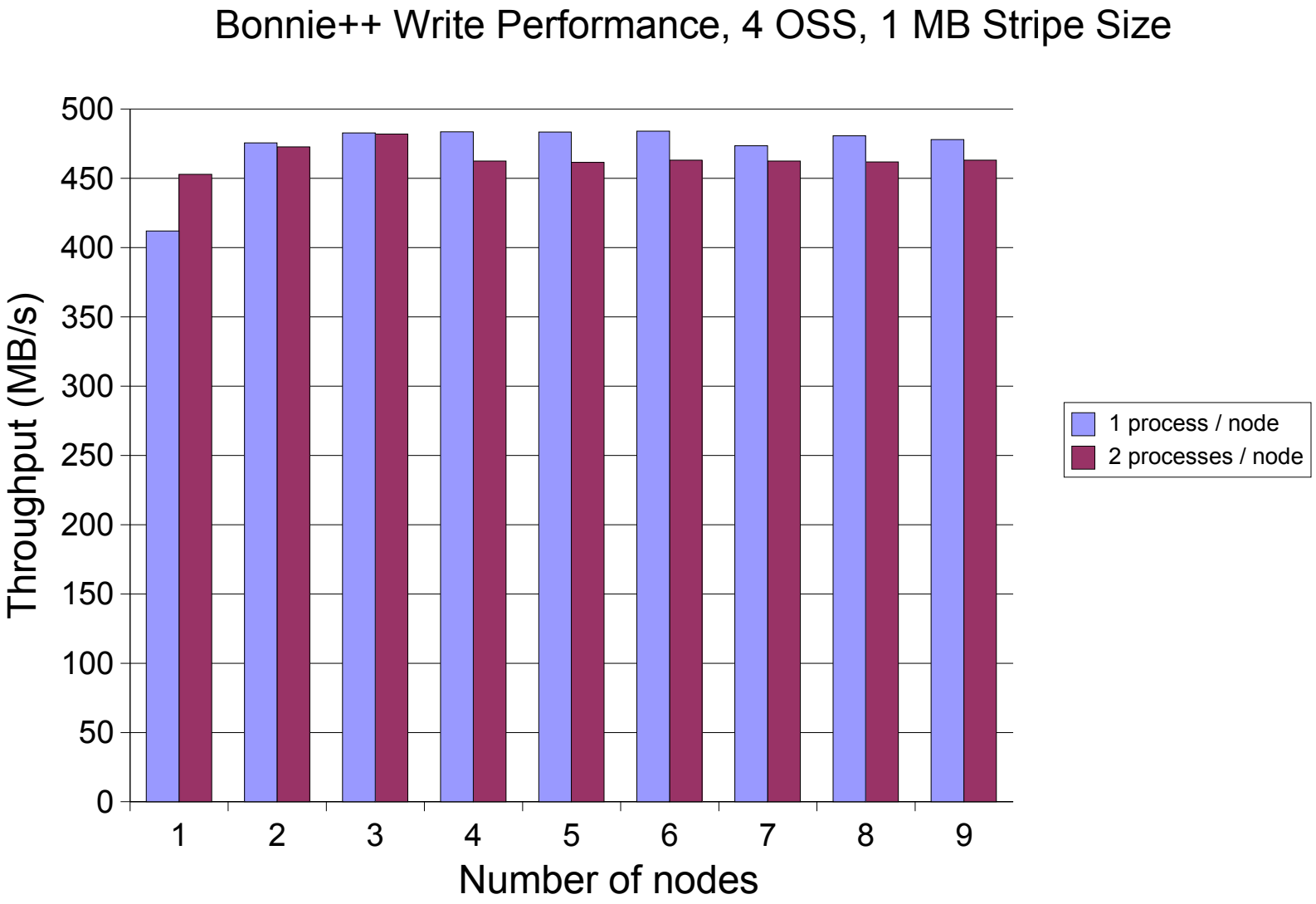
- **Backup**
  - **Files in $HOME run into backup at least twice per week.**
    - **Backup is activated depending on the project's requirement (per default for academic users).**
    - **Files without backup necessity should be stored below $WORK.**
    - **Omit renaming huge directories below $HOME.**
  - **Users can restore files by themselves.**
- **Archiving files**
  - **Can be allowed on a project basis**
    - **is per default not allowed**
  - **A special commands will be provided for archive and retrieve**
    - **can be used for files in $HOME and $WORK**
    - **is not yet fully functional**
  - **Archived files can be encrypted on a project basis**
    - **Files are stored externally to the HWW on the University's TSM servers.**

Sequential write performance of $HOME

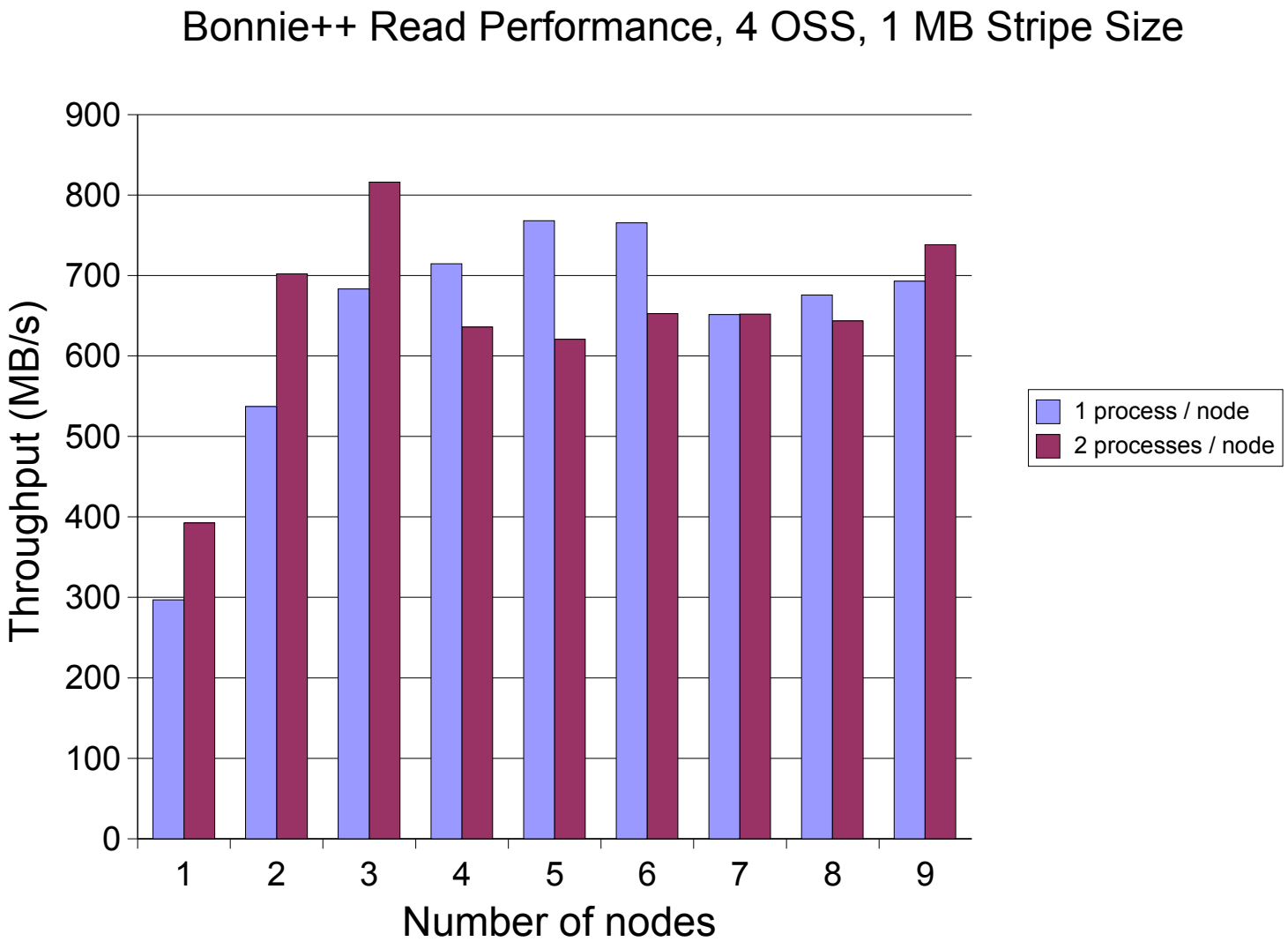


Sequential read performance of $HOME

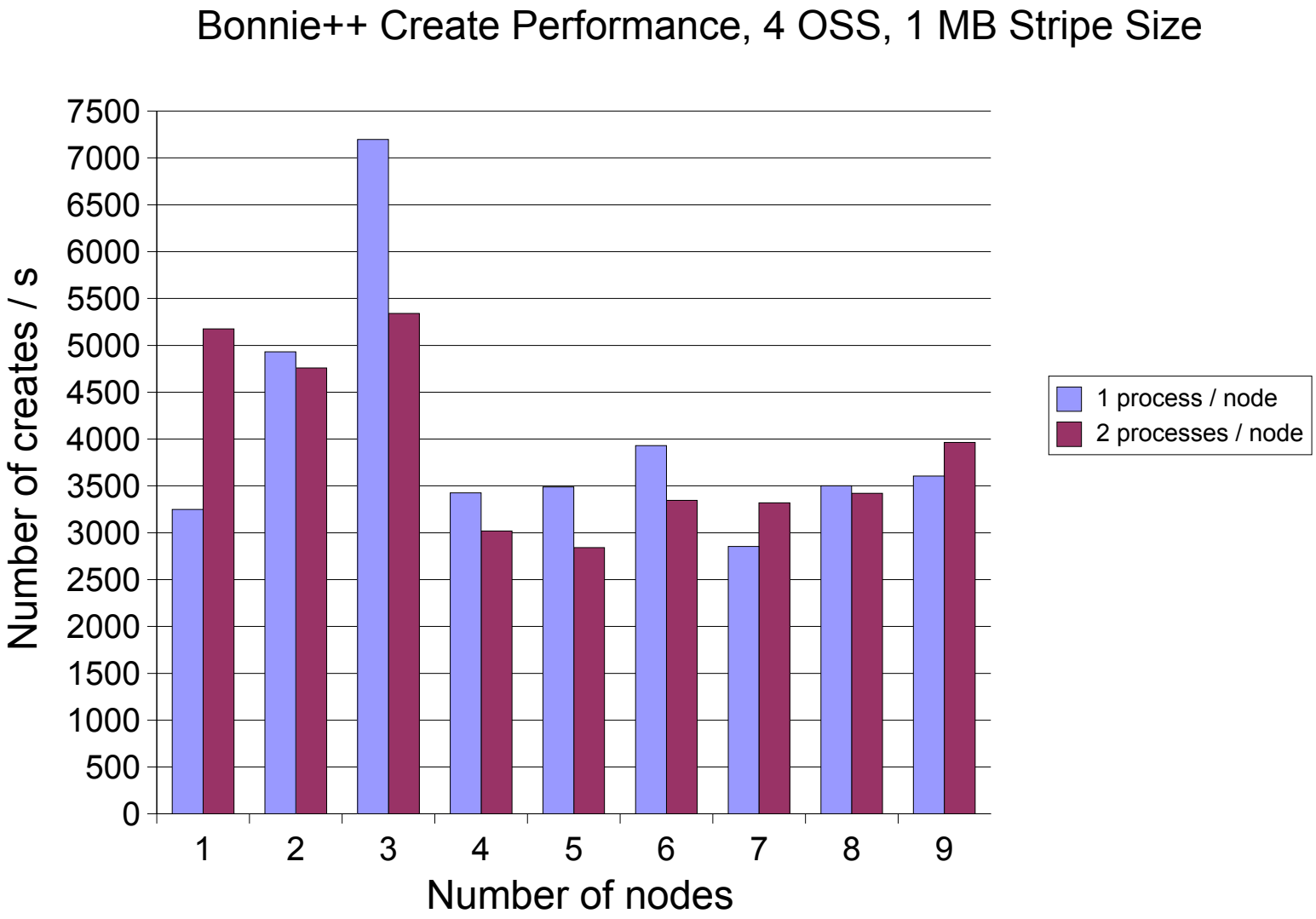
Sequential write performance of $WORK



Sequential read performance of $WORK



File creation performance



Using HP SFS

- **Just use it like a normal filesystem !**
- **Unsupported features of the POSIX standard**
  - **mmap() is not fully supported**
    - **restriction will go away in a few months**
  - **flock() is currently not supported**
    - **restriction will go away in a few months**
  - **direct IO (O_DIRECTIO option) is not supported**
    - **rarely used in applications**
  - **atime is not always accurate**
    - **does not make sense in a parallel file system**
    - ***tail -f* will not show new data at once**

Some background information about Lustre

- **Lustre is a new scalable high performance file system**
  - **"Lustre" is an amalgam of the terms "Linux" and "Clusters"**
    - **i.e. Lustre is a cluster file system for Linux**
  - **Lustre is available under the GNU General Public License**
    - **Customers with support contract get new versions in advance (up to 1 year)**
  - **http://www.lustre.org/**
- **Main development by Cluster File Systems, Inc. (CFS)**
  - **Peter J. Braam is President and Chief Technology Officer**
  - **HP is main contractor of many DOE sites**
    - **ASCI PathForward initiative pushed Lustre development by CFS, HP, and Intel**
  - **Lustre development has got serious funding from DOE**
    - **LLNL, NCSA, PNNL, Sandia, NNSA, Los Alamos**
  - **http://www.clusterfs.com/**