

---

# Performance Measurements and Experiences with the HP SFS EA Release

Roland Laifer

Computing Centre (SSCK)  
University of Karlsruhe

[Laifer@rz.uni-karlsruhe.de](mailto:Laifer@rz.uni-karlsruhe.de)



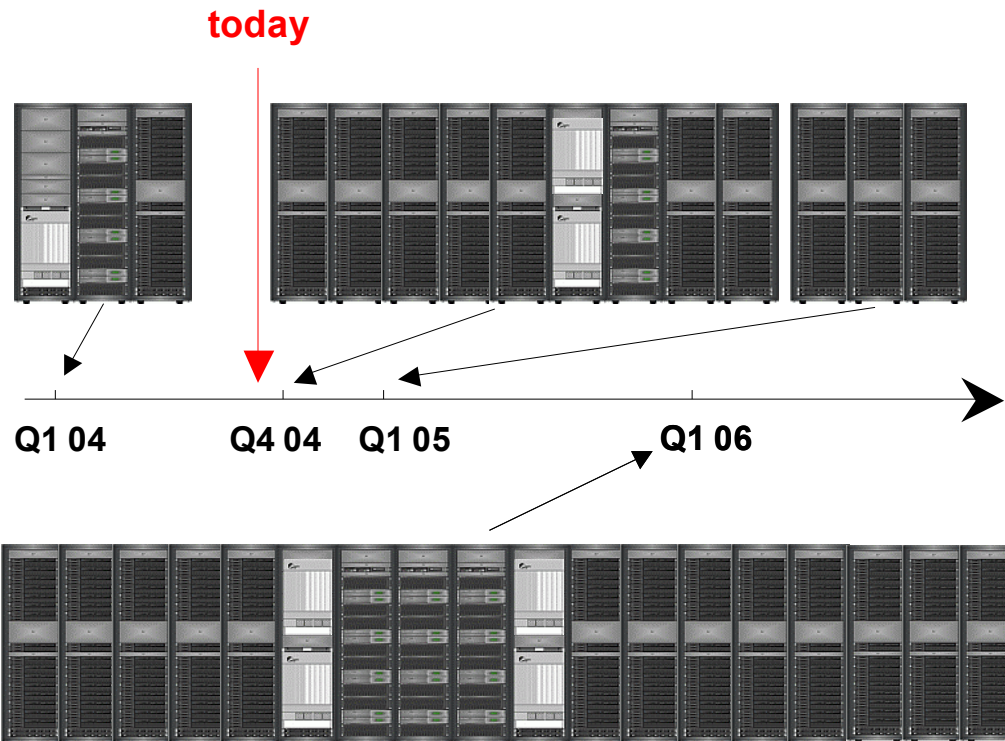
# Overview

---

- » **Introduction**
- » **Performance measurements**
- » **Experiences with the HP SFS EA release**
- » **Necessary enhancements**
- » **Configuration of our first production system**



# HP XC 6000 Cluster installation schedule at SSK



## Phase 0 (Q1 2004)

- » 16 node test system
  - Madison
  - Single rail Quadrics interconnect
- » 2 TB storage system

## Phase 1 (Q4 2004)

- » 116 two-way nodes
  - Madison 9M
  - Single rail Quadrics interconnect
- » 10 TB storage system

## Phase 1 (Q1 2005)

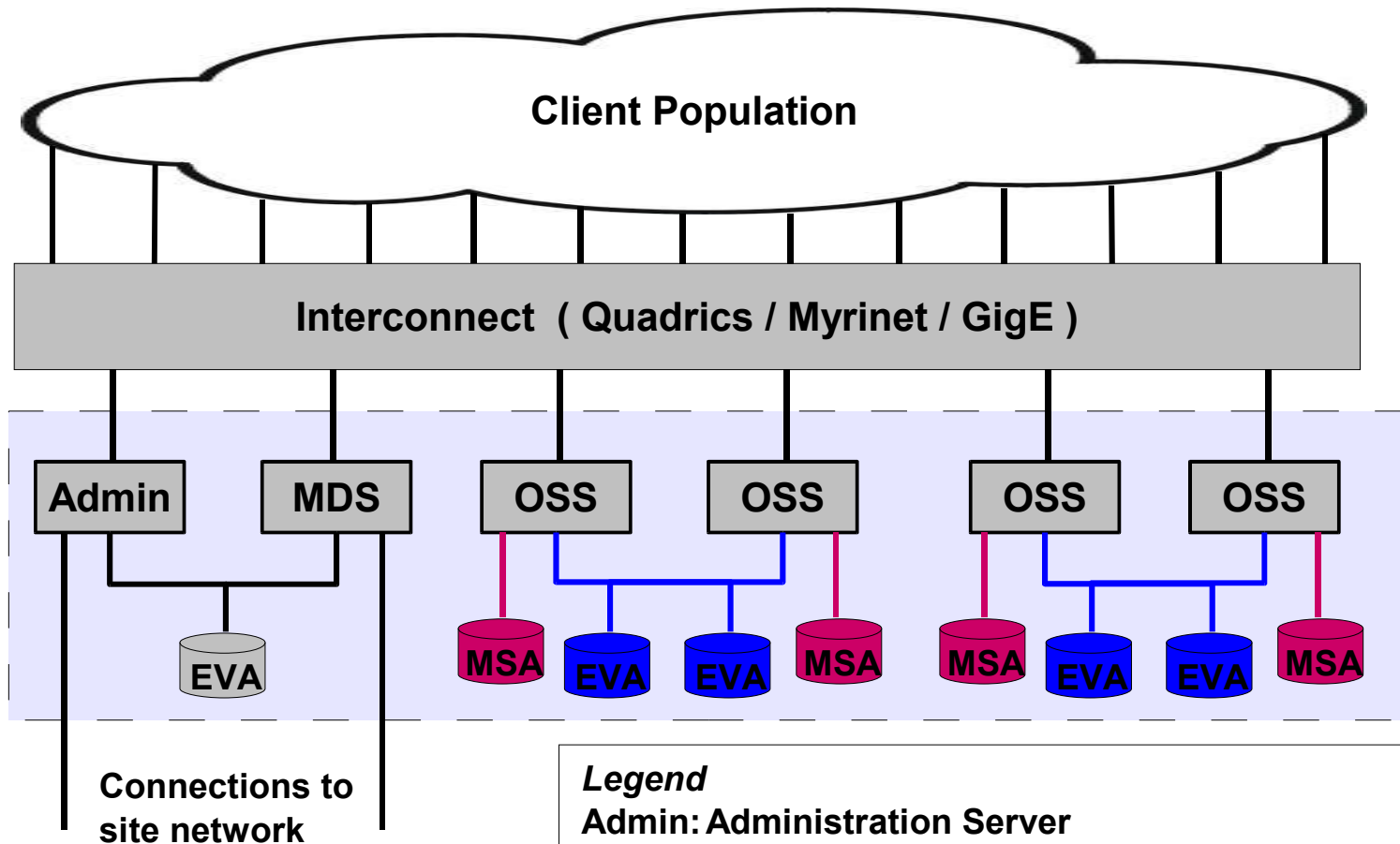
- » 6 16-way nodes
  - Madison 9M
  - Single rail Quadrics interconnect

## Phase 2 (Q1 2006)

- » 218 four-way nodes
  - Two sockets
  - Dual core Montecito
  - Single or dual rail Quadrics interconnect
- » 30 TB storage system



# HP SFS system architecture



## Legend

**Admin:** Administration Server

**MDS:** Metadata Server

**OSS:** Object Storage Server

**EVA:** Enterprise Virtual Array 3000 storage alternative

**MSA:** Modular Smart Array 20 storage alternative



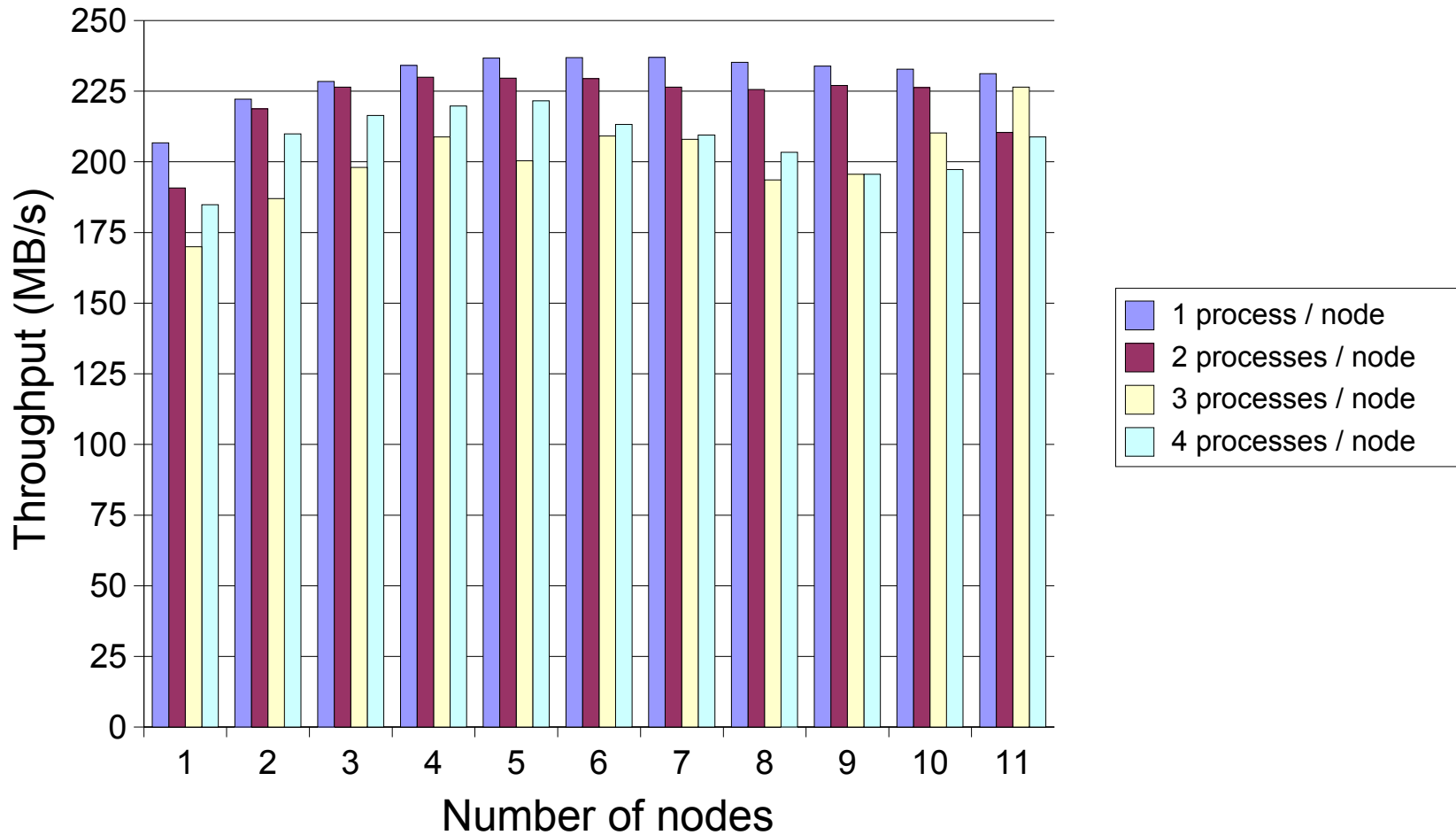
# Performance measurement environment

---

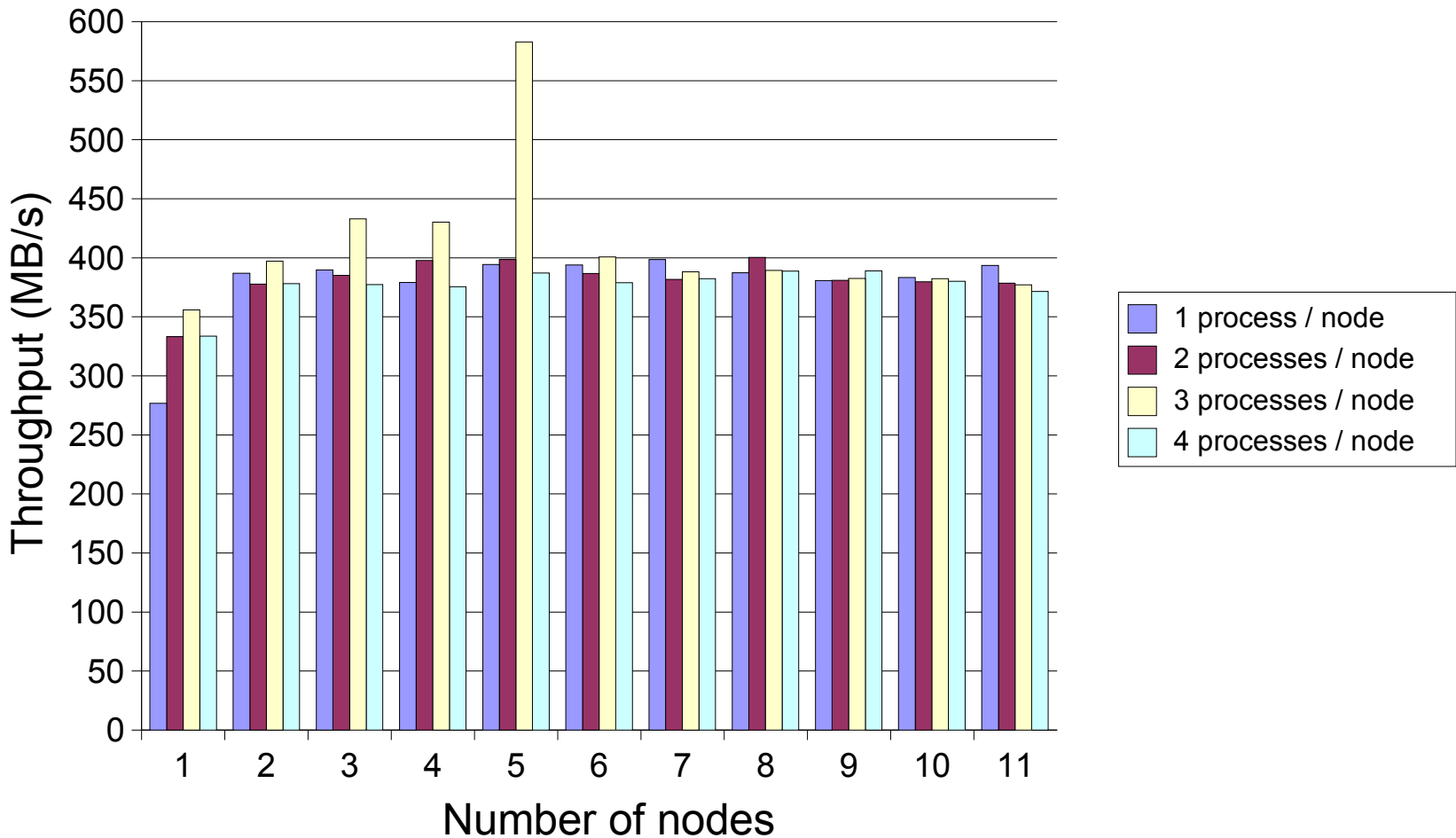
- » **Used HP SFS software version was Early Access (EA)**
  - Is based on CFS Lustre version 1.2.3
  
- » **Underlying HW**
  - Clients are IA64 systems (rx2600, 1.3 GHz, 2 CPUs, 4 GB memory)
  - Servers are IA32 systems (DL360, 3.2 Ghz, 2 CPUs, 4/2 GB memory)
    - We had 2 OSS with 2 attached EVAs (2 controllers per EVA)
  - Quadrics QNet-2 (Elan4) interconnect
  - EVA5000 (not EVA3000) storage systems
    - We would expect similar results for both EVA types
  
- » **Performance measurement strategy**
  - First test the performance of the underlying raw device
  - Saturate clients with multiple processes per client
  - Saturate servers by using multiple clients
  
- » **Benchmarking software was bonnie++**



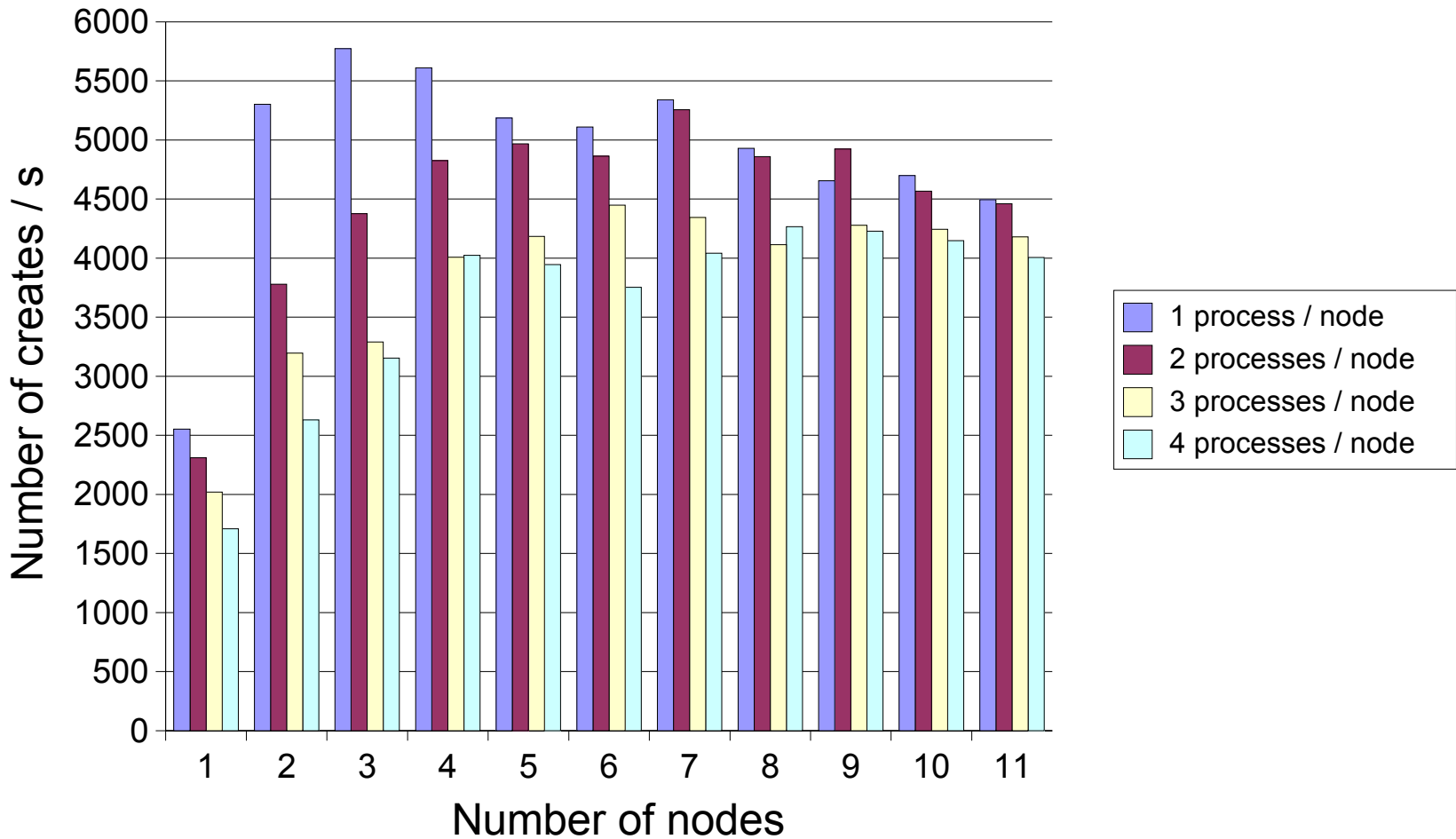
# Sequential write performance with 2 OSS



# Sequential read performance with 2 OSS

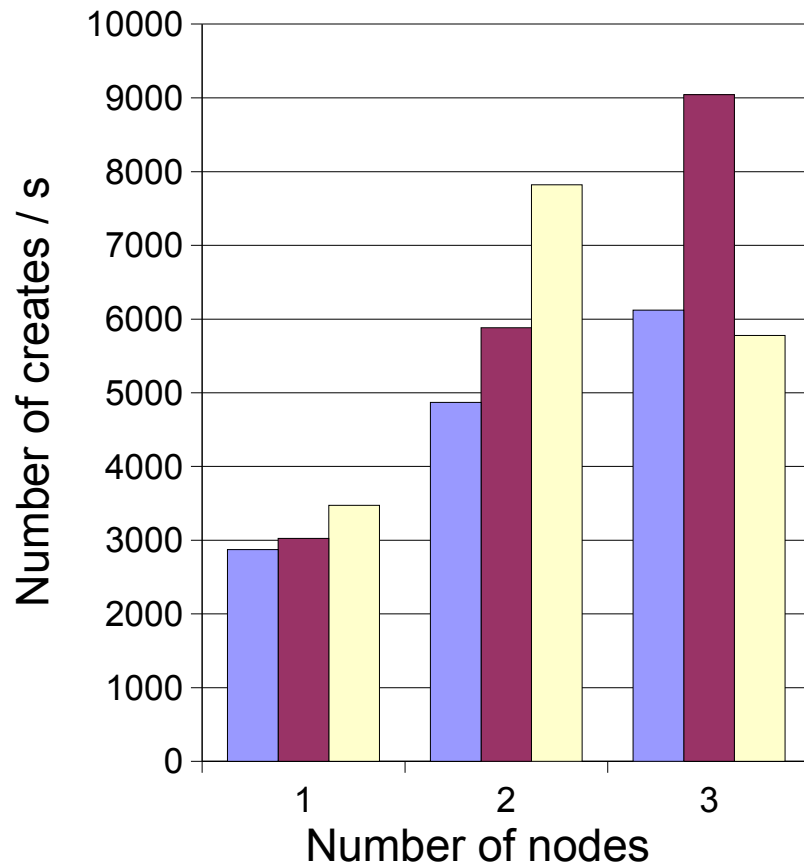


# File creation performance with 1 MB stripe size

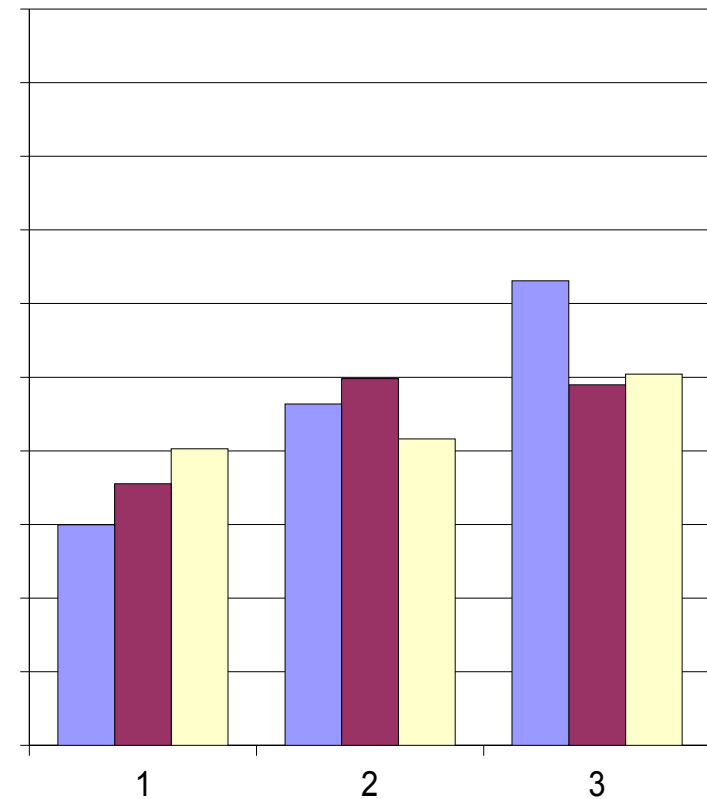


# File creation performance with different stripe sizes

1 OSS, 128 KB Stripe Size



1 OSS, 1 MB Stripe Size



# Performance measurement results

---

## » EVA seems to be bottleneck

- Raw device write performance (8 disks) is about 60 MB/s per controller
  - Bottleneck seems to be mirrored cache between the 2 controllers
  - Read performance is much better
- HP internal white paper mentions about 75 MB/s (with more disks)

## » Main benchmarking results (8 disks per disk group)

- Write performance is about 120 MB/s per OSS
  - 207 MB/s with 1 process
- Read performance can reach 300 MB/s per OSS (normally only 200 MB/s)
  - 277 MB/s with 1 process
- File creation performance can reach 9000 creates/s
  - 3000 creates/s with 1 process
- Smaller stripe sizes
  - Show better file creation performance
  - Show worse read performance



# Experiences with the HP SFS EA release

---

- » Reasonable performance results
  - Could be enhanced with better storage subsystem
- » Failover works with the following exceptions
  - LDAP IP address does automatic fallback but admin service does not
  - MDS failover does not work with multiple file systems
- » Bugs which should be fixed in the GA release
  - mmap() is not fully supported
    - Starting huge executables may not work
  - ASSERTION() failure causes crash during MDS recovery
    - CFS bug id 3440
  - Communication problem causes reduced stripe count
    - CFS bug id 4787
- » Problem diagnostic is difficult
  - Support is really needed



# Necessary enhancements

---

## » Necessary new features

- **Quota support**
- **Support for snapshots (point in time copies)**
  - Multiple snapshots would be even better
  - Will greatly reduce the need to restore unintentionally deleted files
- **Efficient backup support**
  - e.g. NDMP or fileset level backup
- **Additional tools**
  - For performance monitoring
  - For problem notification

## » More information is needed

- **About the roadmap**
- **About possible configurations**
  - e.g. with gateways, or multiple networks, or multiple Lustre engines
- **About the system internals**



# HP SFS configuration of our phase 1 system

---

## » Lustre server engine with 8 servers

- 2 MDS / Admin with disk space for 50 million files
  - RAID 1 on MDS luns
- 2 OSS for home directories and software with ~ 3 TB storage
- 4 OSS for work (scratch) file system with ~ 6 TB storage
- 7 EVA5000 2C2D storage systems

## » File system properties for home directories and software

- Used for permanent data (source code, executables, input and output files)
- File level backup with IBM Tivoli Storage Manager (TSM)
- Planned stripe size is 128 KB

## » File system properties for work directories

- Used for temporary data (checkpoints, restart files, and output files)
  - Old files will be deleted on a regular basis
- Planned stripe size is 1 MB



# Summary

---

- » **HP SFS has the most important features of a parallel file system**
  - **Performance (throughput and metadata), high availability, and scalability**
    - **Also allows ease of administration**
  - **Additional features are still needed**
  
- » **We expect a hard time to reach a highly reliable and stable system**
  - **This is something usual with a new file system**
  - **The problem tracking system works**
  - **We still need a very good support**

