# BeeGFS at DESY

## -A short introduction

## -Our experience


Yves Kemp
on behalf of Sven Sternberger
HEPiX Spring 2015
25.3.2015 Oxford

HELMHOLTZ ASSOCIATION

DESY

# BeeGFS Basics

- High performance parallel filesystem developed 2007 from Competence Center for High-Performance Computing, Fraunhofer ITWM
  - Aim: Replace GPFS and Lustre by something easy to deploy, config and administer
- Originally named FhGFS it was renamed in 2014 to allow a commercial spin off
- Development is driven by Fraunhofer, the company ThinkparQ offers support.
- The Software is free of charge
- License and costs
  - The client kernel module is under the GPL
  - Storage and Management Daemons, currently closed source, but guaranteed in the context of the DEEP-ER project to become open-source
  - Commercial support offered: Annual license per storage target

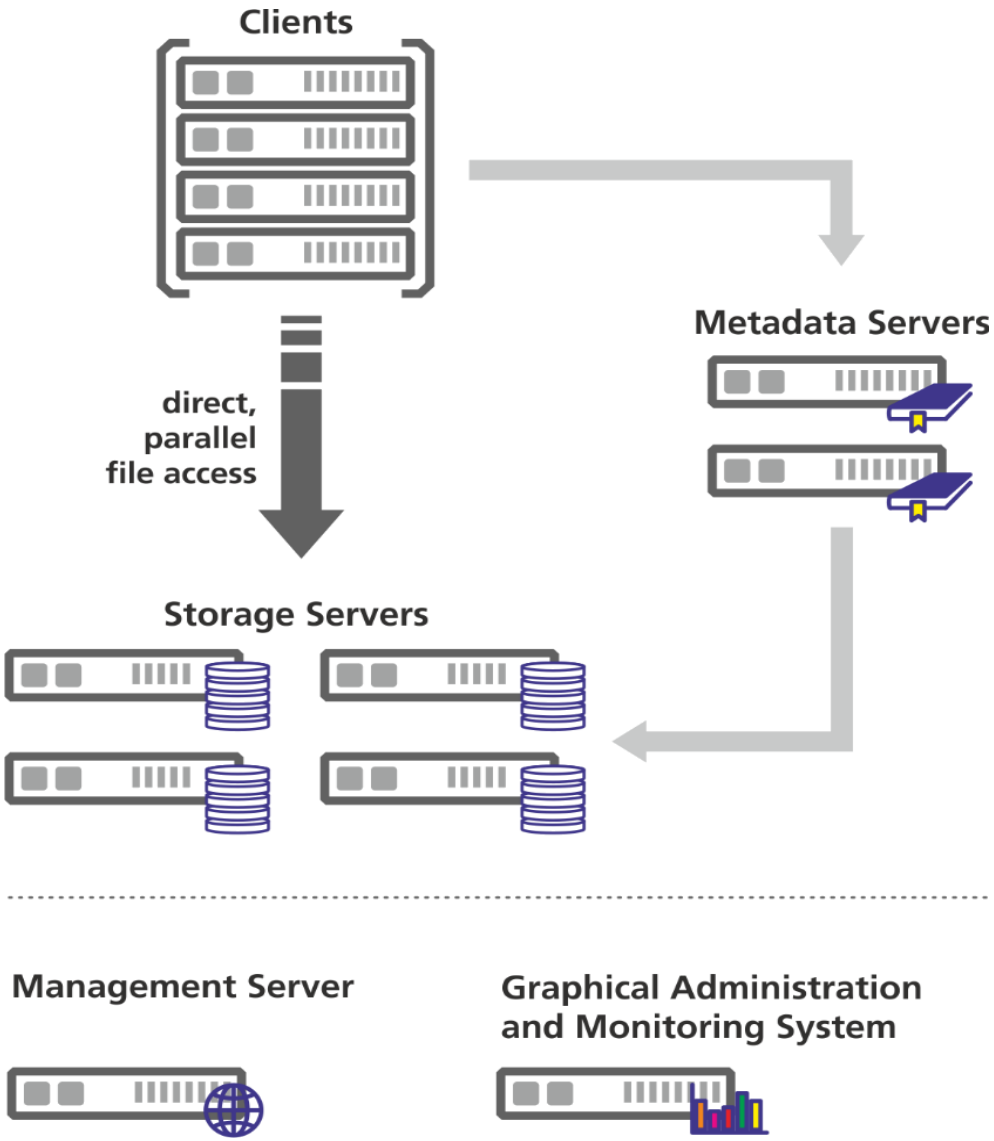
# BeeGFS Technical Facts

- Distributed Object and Metadata
  - Aggregated throughput for objectdata (using striping)
  - Loadbalancing for Metadata
- Linux (only) based
  - Packed for for RHEL, Debian, Suse
  - Support x86_64 and XeonPHI (proof of concept for ARM)
- Server runs in userspace, and use supported filesystem of the OS
  - Object Store tested with xfs, ext4 and zfs
  - Metadata Stored on ext4 filessystem (use extended Attributes)
- Clients are kernel modules
  - Support all kernels from 2.6.16 to latest vanilla, no Kernel Patch
  - Automatic rebuild after kernel update
- Support native Infiniband/1GE/10GE/40GE
- BeeGFS is improving fast
  - Already there: Raid Level, HSM integration
  - Upcoming: HA, Data integrity, erasure coding



DESY

# BeeGFS components

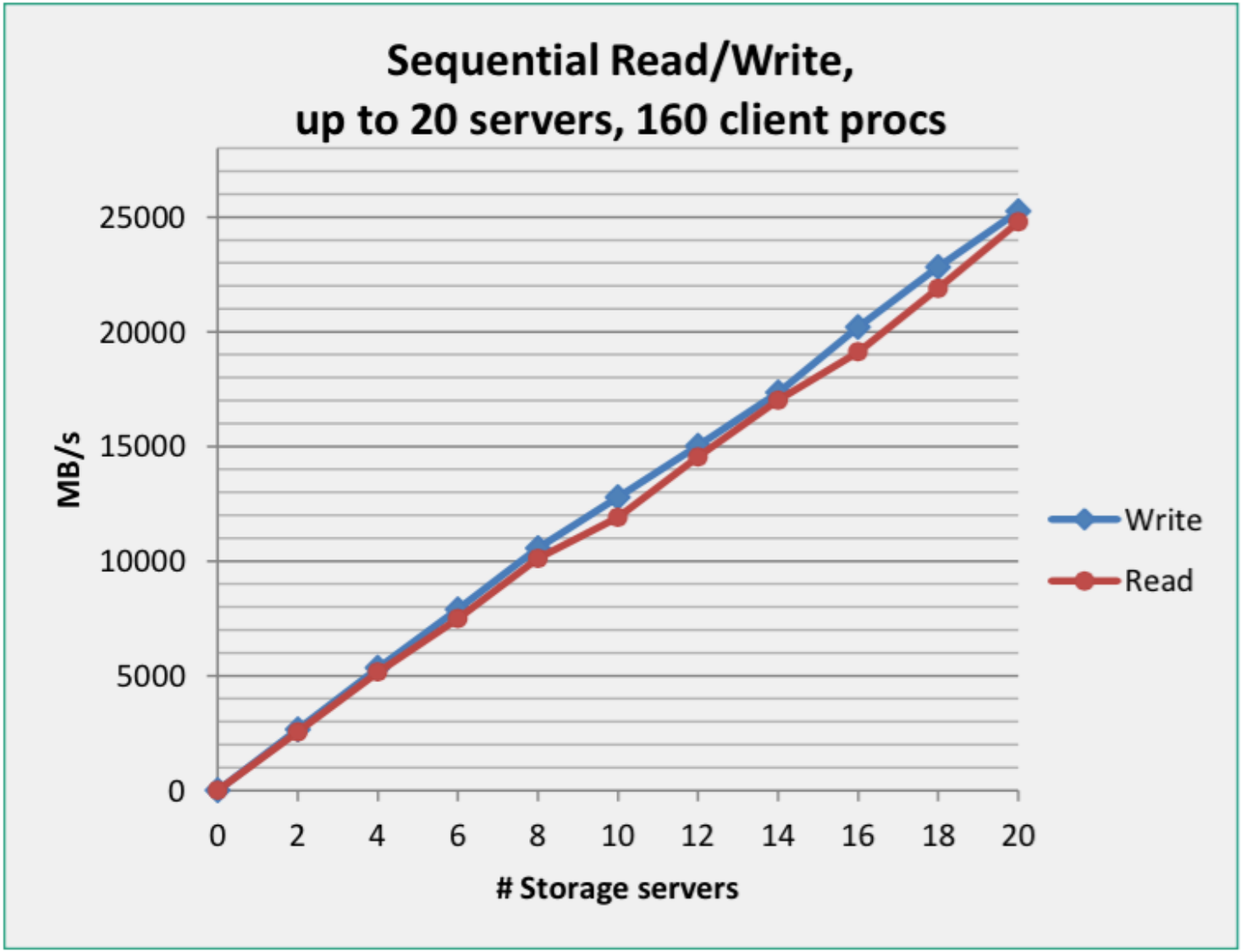



- Clients and server services can run in any combination, e.g.:
- Dedicated objectstore, Metadata and clients
  - Current DESY setup
- Combined Object and Metadatastore, with dedicated client
  - For extreme metadata performance
- Combined Objectstore and client with dedicated Metadata
  - Initial DESY setup

# Benchmarks by the Fraunhofer people

- *"I've found that when you want to know the truth about someone that someone is probably the last person you should ask."*
  Dr. Gregory House
- … nevertheless, we chose to show some benchmarks made by Fraunhofer and presented at SC12
  - http://www.fhgfs.com/docs/SC12_FHGFS_Presentation.pdf
  - Please apply some grain of salt
- Many variables influencing the benchmarks
  - e.g. firmware & kernel versions
  - socket binding and so on
- No systematic benchmarks performed by DESY

# Streaming Throughput




Slide by Christian Mohrbacher



Fraunhofer ITWM

# Streaming Throughput (2)

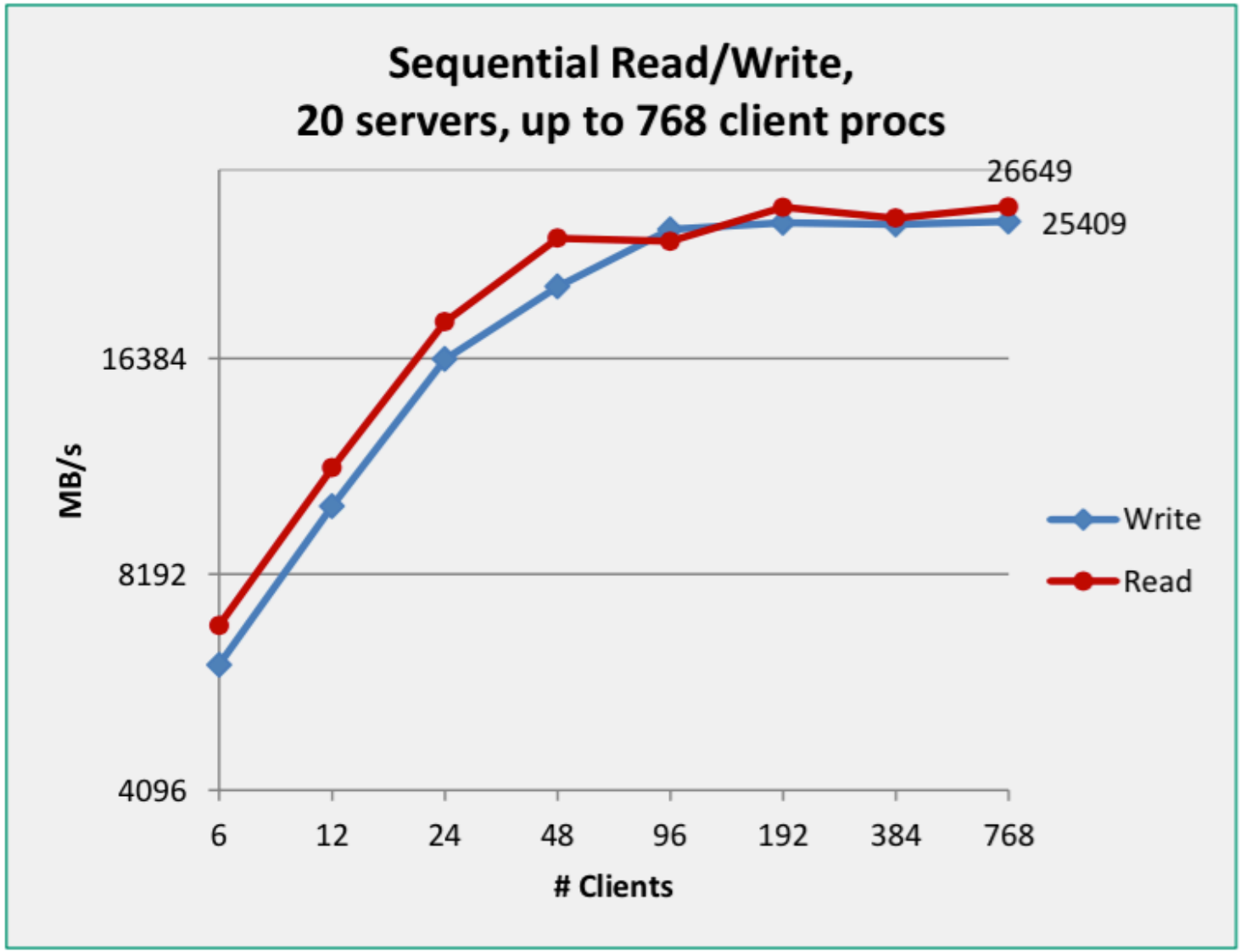
- **Single node local performance**
  - Write: 1332 MB/s
  - Read: 1317 MB/s
- **20 nodes (theoretical)**
  - Write: 26640 MB/s
  - Read: 26340 MB/s
- **FhGFS**
  - Write: 26247 MB/s (98,5%)
  - Read: 24789 MB/s (94,1%)


Slide by Christian Mohrbacher



Fraunhofer ITWM

# Streaming Throughput (3)

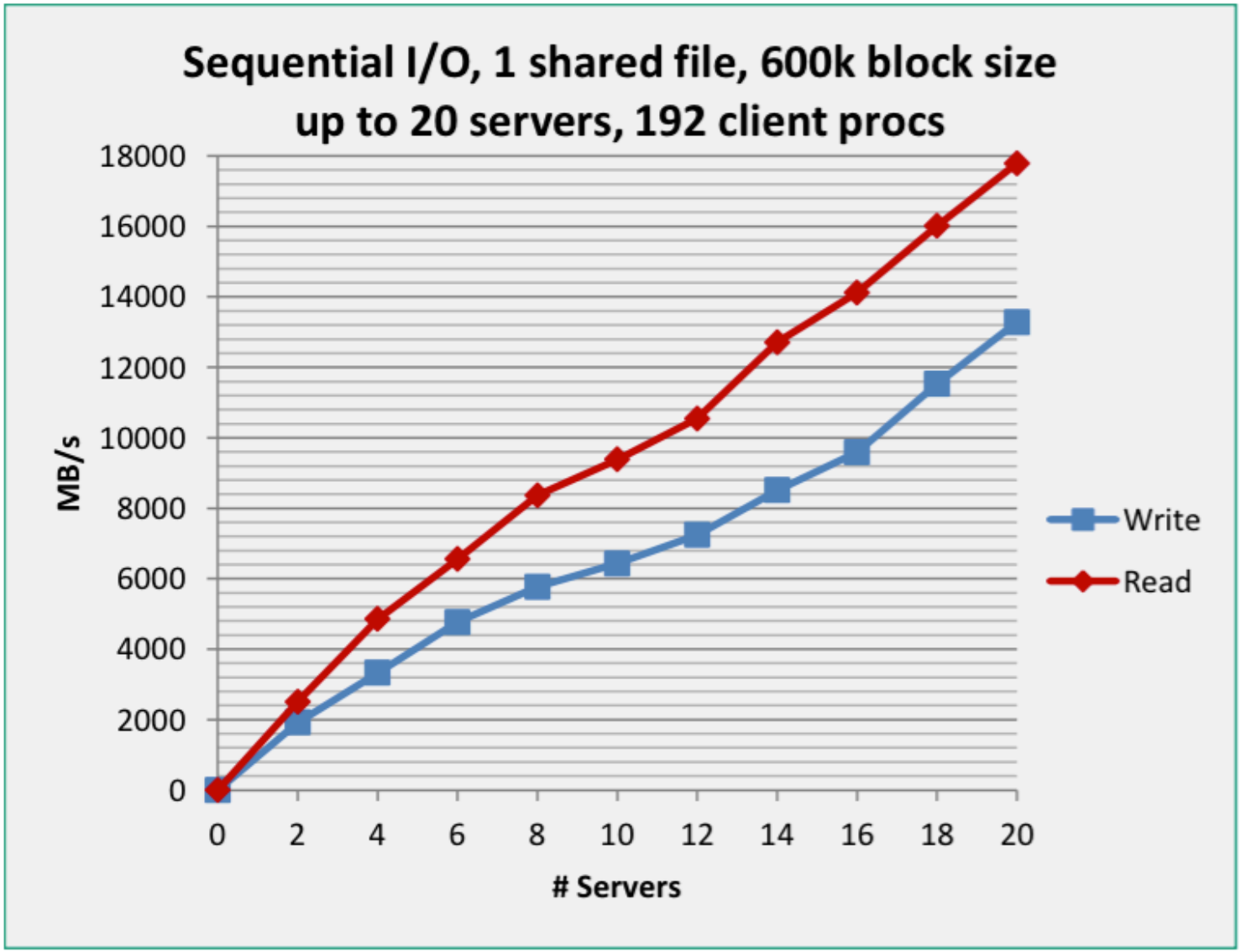



Slide by Christian Mohrbacher



Fraunhofer ITWM

# Shared file access (1)

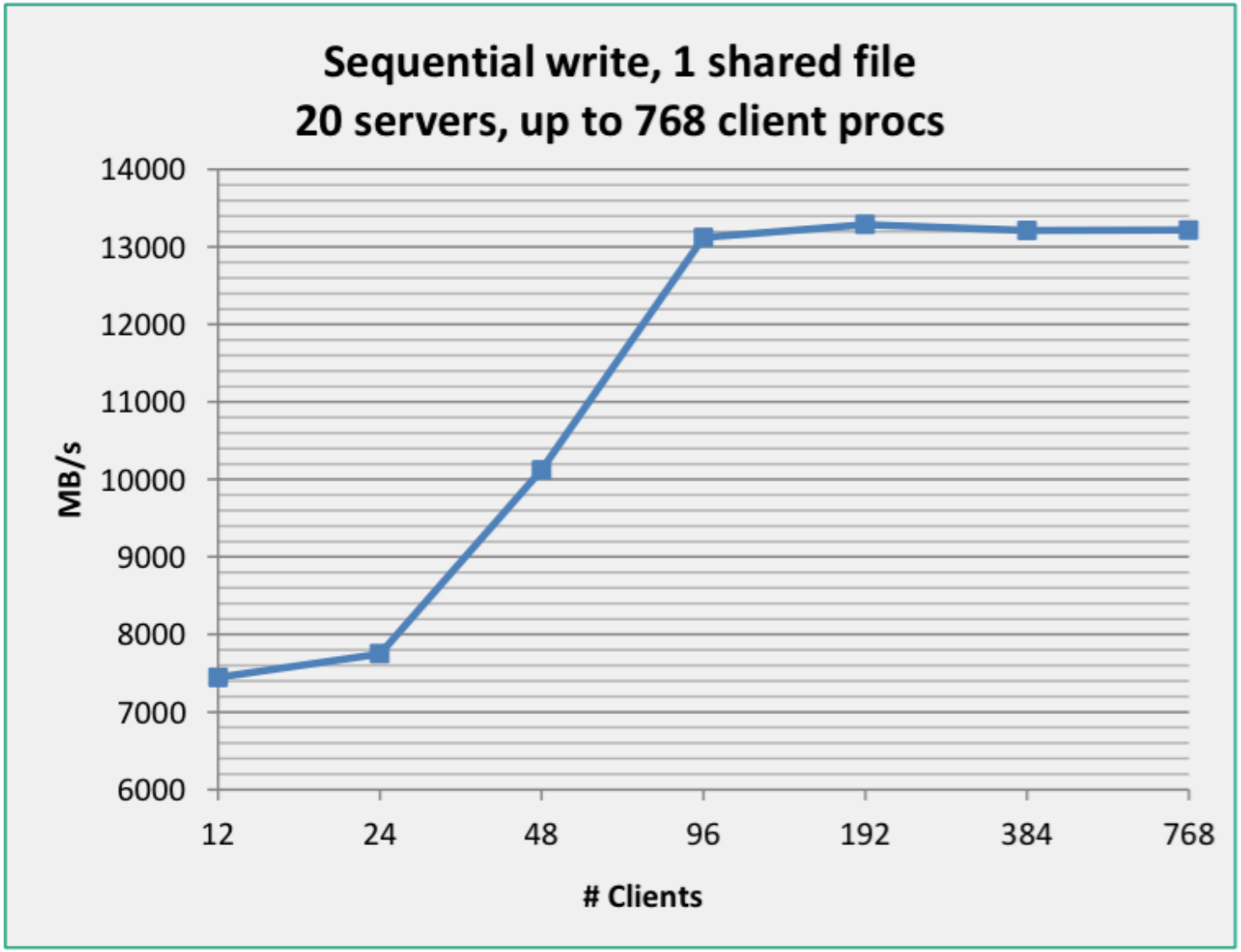



Slide by Christian Mohrbacher



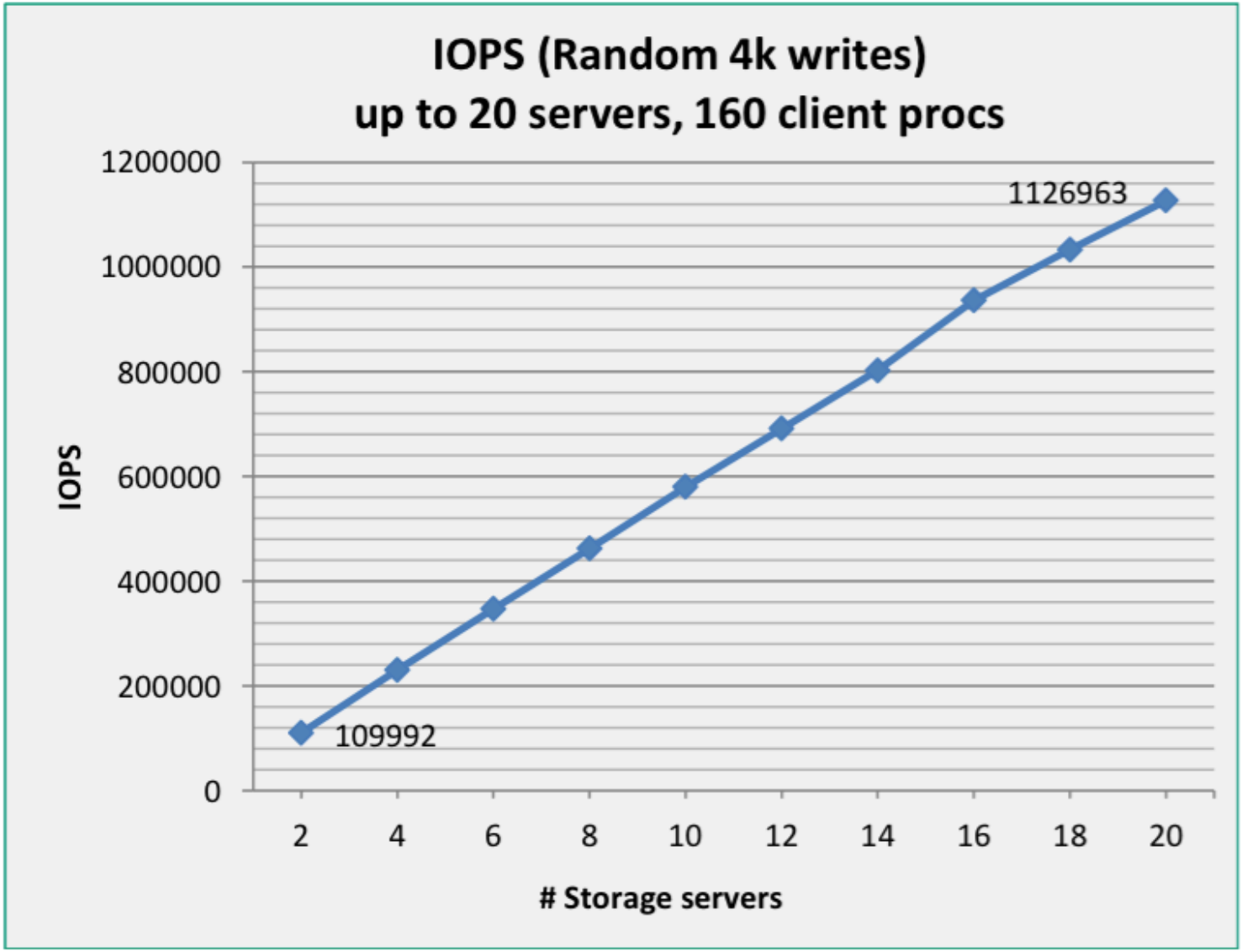
Fraunhofer ITWM

# Shared file access (2)




Slide by Christian Mohrbacher

# IOPS




Slide by Christian Mohrbacher



Fraunhofer ITWM

# Metadata performance


Slide by Christian Mohrbacher



Fraunhofer ITWM

# The first HPC cluster & FHGSF setup: 2012 – mid 2013

- Using 64 decommissioned HP nodes (originally purchased in 2007)
  - 64 x 2-Socket, each 4 Cores and 8 GB RAM
  - DDR InfiniBand
  - 2x150 GB SAS Disks in RAID-0
- Wanted a simple and fast cluster file system for *scratch*
- FhGFS setup:
  - Each WN was both Client and Objectstore: 64 nodes
  - Dedicated Metadataserver
  - Pros: Very good performance
  - Cons: When a cluster node crashed it had an impact on other nodes

# The second HPC cluster and second setup: ~mid 2013

- Worker nodes: 18x AMD 4-way systems (initial setup)
  - 18 x 4 x 16 cores, 256 GB per system
  - QDR IB based
  - Some local disk for $TMP

- FhGFS setup:
  - Four (decommissioned) R510 (=12disks @ 2 TB) with 8 GB RAM each
  - Bad performance: Small number of nodes, too little RAM, slow controller
  - Changed rapidly to newer system

# Third FhGFS setup

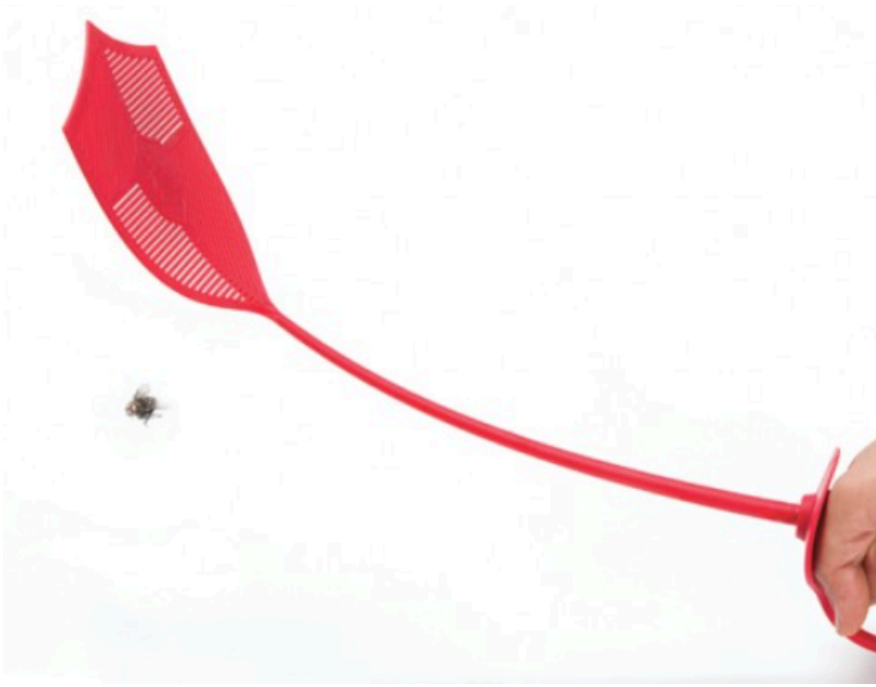
- Mid/End 2013:
  - Dedicated Objectdata nodes: 10x R320 with each 8x1.2 TB disk RAID 10, 48 GB RAM
  - 50 TB net over 10 nodes
  - FDR IB
  - Dedicated metadata nodes: 2x R620 with SSD
  - Pros: Very good performance
  - Cons: Little space, rather high price/space
- End 2014: Added four larger server:
  - 4 * Dell R720xd/Raid 60 + Spare ~ 20 TB per server
  - Still good performance, more TB/EUR
  - Now ~130 TB net capacity
- (HPC: currently at ~50 fat nodes, will add more nodes, but Intel)

# Experience at DESY

- Really reliable high quality software, low administration effort
- frequently updated packages (with yum) and Kernel updates without any glitches
- Good commandline tools for inspecting and monitoring
  - As it works without any problems not used often
- Helpful mailinglist
- Excellent performance (Metadata and Data)

# The Accident

- Missed a disk failure on one of the small storage target. Second disk failed on the same node crashed the RAID10 volume
- About 40% of all files were corrupted afterwards
  - At this time 10 storage targets, files striped over 4 storage nodes
  - Fsck to analyze the problem lasted 2 days
  - Repair lasted 7 days
- The open source documentation for these type of critical situation, is not sufficient.
- Needed help from the mailing list and some reverse engineering to fix the situation



# What we learned

- BeeGFS has its roots in the HPC world.
  - To loose data means to loose time
  - Performance over data safety
- If you can't afford the commercial license, you should have a concept how to deal with accidents
  - Inform the user to check and save data if possible
  - Re-Initialize the Objectstore
- Plain BeeGFS is as reliable as your hardware
- We advertised the storage as volatile/scratch storage and we will continue

# … wait, what about GPFS and the previous talk?

- Beginning 2014, when taking decision for Petra-III online storage:
  - No commercial support for (then) FhGFS, ThinkparQ not yet founded
  - IBM/GPFS more features than just a cluster file system (helping data migration e.g.)
- We took the decision to continue with FhGFS/BeeGFS for now as scratch in HPC *and* build up GPFS expertise for datataking
  - Will reevaluate the situation once major change needed within BeeGFS (e.g. HW exchange, different setup, performance problems)