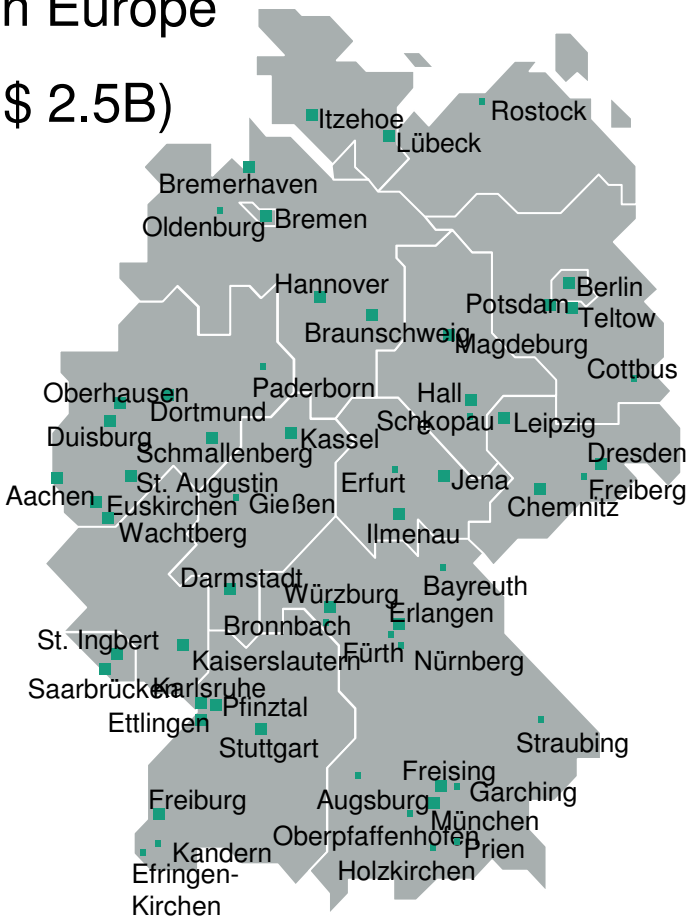
FhGFS - A Flexible Parallel File System for Performance Critical Applications

Christian Mohrbacher
christian.mohrbacher@itwm.fraunhofer.de

Fraunhofer ITWM

The Fraunhofer Gesellschaft (FhG)

- Fraunhofer is based in Germany
- Largest organization for applied research in Europe
- Annual research volume of 1.9 billion € (~ $ 2.5B)
- 22,000 employees
- > 60 Fraunhofer institutes with different business fields



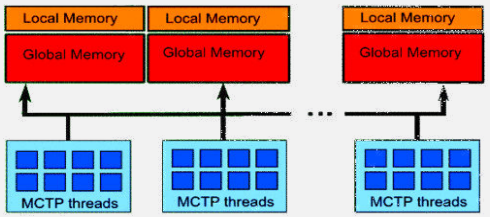
Fraunhofer
ITWM

The Fraunhofer ITWM

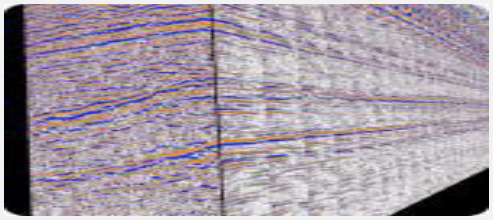
- Institute for Industrial Mathematics
- Located in Kaiserslautern, Germany
- Staff: ~ 200 employees + ~ 50 PhD students in 8 departments

Fraunhofer ITWM

ITWM's Competence Center HPC



Programming models / tools



Interactive seismic imaging

Fraunhofer FS

Filesystem Development

Photorealistic RT rendering

Green IT Smart Grids

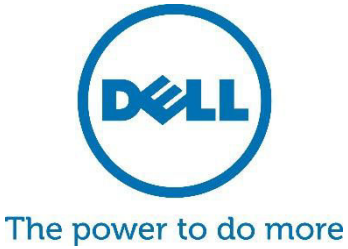
Fraunhofer ITWM

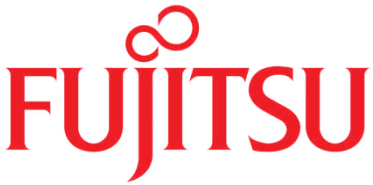
FhGFS – Some Quick Facts

- Development started in 2005
- First public beta in 2007
- First release in 2008

- Free to use
- Fraunhofer offers professional support
- Supported installations all around the world

Partners / Vendors





Fraunhofer ITWM

Users (Examples)

Heidelberger Institut für Theoretische Studien
HITS
sharpreflections INTERACTIVE PRE-STACK
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
UNIVERSITAS OSLOENSIS MDCCCXI
University of Zagreb
MARATHON
SeismicCity
UNIVERSITÉ BORDEAUX 1 Sciences Technologies
l'Observatoire de Paris
NORECO
DNV
RSI
DETNORSKE
university of groningen
CERN
DESY
DTU
GL Germanischer Lloyd
universität wien
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN
TECHNISCHE UNIVERSITÄT ILMENAU
iit istituto italiano di tecnologia
BNP PARIBAS
Freie Universität Berlin
UNIVERSITÄT PADERBORN Die Universität der Informationsgesellschaft

Fraunhofer ITWM

FhGFS - Overview

Maximum Scalability

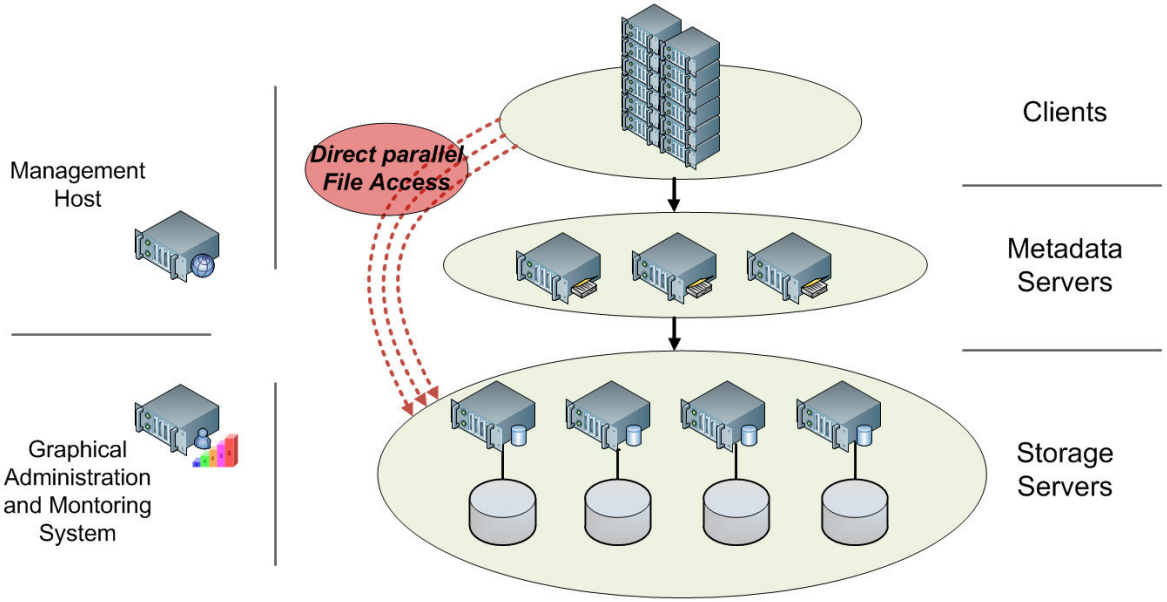
Flexibility

Easy to use

Fraunhofer FS

Maximum Scalability

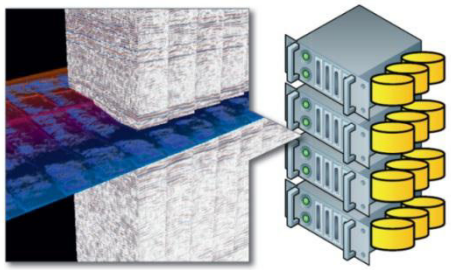
- Distributed file contents
  - Flexible striping across storage servers
- Distributed metadata
- Initially optimized especially for HPC
- Native Infiniband / RDMA



Fraunhofer ITWM

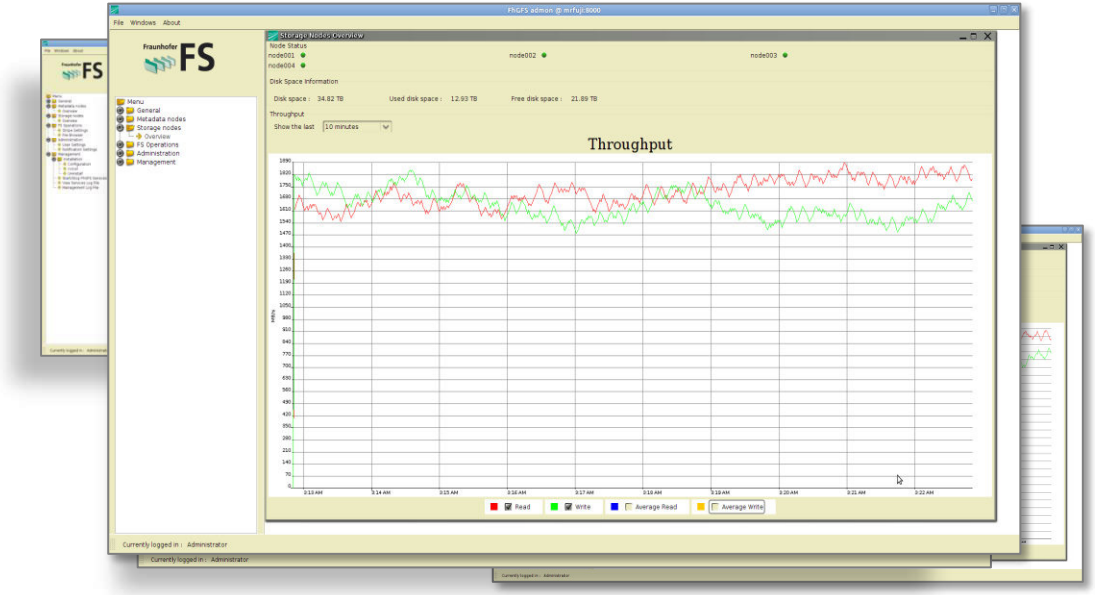
Flexibility

- Wide range of Linux distributions (RHEL/Fedora, SLES/OpenSuse, Debian/Ubuntu)
- Wide range of Linux kernels (from 2.6.16 up to latest vanilla)
- Storage servers run on top of a local filesystem
- Add clients and servers without downtime
- Multiple networks with dynamic failover
- Multiple FhGFS services on the same machine
  - No dedicated servers needed
  - Any combination of client and servers can run on the same machine
  - Computation and storage on same machines possible



Easy to use

- Servers: Userspace daemons
- Client: Kernel module without patches to the kernel
- Simple Setup/updates and startup mechanism (rpm/deb packages, init scripts)
- No special hardware requirements
- Graphical administration and monitoring system



Fraunhofer ITWM

Users (Examples)

12 Servers
900 Clients
800TB

5 Servers
100 Clients
1PB

2 Servers
2 Clients
8 TB

6 Servers
1200 Clients

7 Servers
600 Clients

5 Servers
100 Clients
200 TB

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Heidelberger Institut für Theoretische Studien

HITS

sharpreflections INTERACTIVE PRE-STACK

University of Zagreb

MARATHON

SeismicCity

UNIVERSITÉ BORDEAUX 1 Sciences Technologies

UNIVERSITAS OSLOENSIS MDCCCXI

l'Observatoire de Paris

NORECO

DNV

CERN

university of groningen

DETNORSKE

RSI

DESY

DTU

GL Germanischer Lloyd

universität wien

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

TECHNISCHE UNIVERSITÄT ILMENAU

istituto italiano di tecnologia

BNP PARIBAS

Freie Universität Berlin

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Fraunhofer ITWM

Current Release

- Last update: 2012.10-r9 (relased 2013/10)
- Completely re-designed, faster metadata layout
- Metadata mirroring
- Flexible file contents mirroring
- Online file system checking
- Built-in benchmarking tools
- More useful utilities
  (e.g. fhgfs-ondemand)
- ……

```
= FhGFS Changelog (2012.10 Release Series) =

== Changes in 2012.10-r9 ==
* general: Added support for hardlinked files in same directory.
* client: Updated to be compatible with linux-2.6.16 up to linux-3.12.
* client: New hash algorithm to generate inode numbers.
* fhgfs-ondemand: Added option to prefer local storage servers on clients.
* admon: Fixed GUI-based installation of packages on SLES11.
  [Thanks to Medizinische Hochschule Hannover for reporting.]
* fsck: Improved runtime of some checks.
* client: Automatically generate 32bit inode numbers for 32bit programs using
   32bit readdir() on 64bit systems.
  [Thanks to University of Oklahoma, Research Campus Computing Center for
  reporting.]
* client: Fixed setgid bit handling of directories.
  [Thanks to Cambridge MRC Laboratory of Molecular Biology for reporting.]
* fsck: Report file paths (not only IDs) in log messages when available.
* servers: Fixed potential problem with cleanup of syslog logger on shutdown.
* client: Package is now of architecture type "noarch" to allow installation on
  non-x86/x64 architectures.
* meta: Fixed wrong modification event flusher error log message on shutdown.
* client: Fixed missing file attributes revalidation after rename in certain
  cases.
* meta: Fixed missing file size attribute update on file close in certain cases.
* fsck: Fixed potentially wrong termination in certain error cases.
* meta: Use random targets chooser if preferred targets are given by client
  (relevant for fhgfs-ctl --migrate).


== Changes in 2012.10-r8 ==

* admon: Added download support for Debian 7 repository.
* client: Temporarily disable file create intent optimization due to problems
   with dangling symlinks.
   [Thanks to University of Iowa for reporting.]
* general: Fixed potential problems during syslog logger initialization.


== Changes in 2012.10-r7 ==

* fsck: Re-introducing online check (i.e. checking while users are accessing the
   file system), as known from the former 2011.04 release series. This mode is
   no longer experimental now.
* meta/storage: New option to use per-user message queues for improved fairness
   in multi-user environments (experimental, see option tuneUsePerUserMsgQueues
```

fhgfs-ondemand

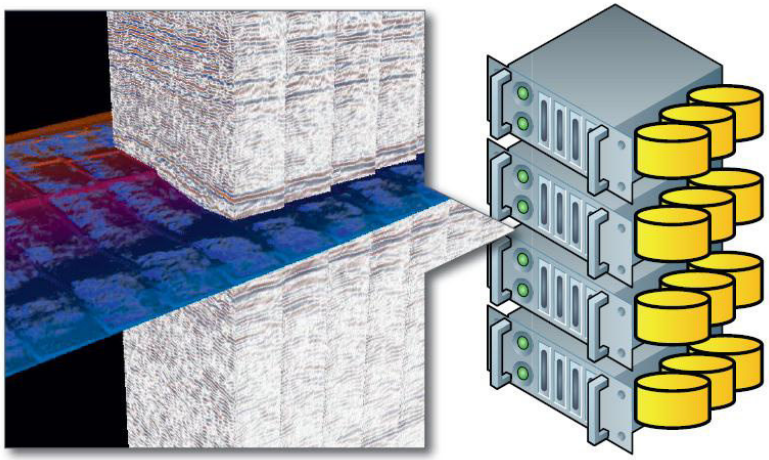
- Helper script in fhgfs-utils
- On demand creation of file system instance possible
- Can be started / stopped with a single command (e.g. integrated in cluster batch system)

```
USAGE: fhgfs-ondemand start -n <nodefile> \
          -d <storagedir> -c <clientmount>


$ fhgfs-ondemand start -n $NODEFILE \
          -d /local_disk/fhgfs -c /my_scratch

Starting FhGFS Services…
Mounting FhGFS at /my_scratch…
Done.
```

fhgfs-ondemand

- Some possible use cases
  - Dedicated file system for cluster job
  - Fast and easy setup of temporary file system for tests
  - Cloud computing
- How Fraunhofer ITWM uses fhgfs-ondemand: Fraunhofer Seislab
  - in-house cluster for development of seismic codes
  - 32 compute nodes, 4x 256GB SSD local storage per node
  - FhGFS storage system with spinning disks
  - Use built-in functionality of batch system (Torque) to run script at the beginning of each job

Fraunhofer Seislab (2)



- 2 “Storage Tiers”
  - Main storage
    - ~ 100 TB on SATA-HDDs
  - Local node Storage
    - 1 TB SSD per node
- Creating an “on-demand-FhGFS” on job start
- Each job: dedicated FhGFS; compute nodes as servers/clients
- Calculations with temporary data can use “local FhGFS”
- Only results need to be written to “slower” main storage

Exascale Plans

- Fraunhofer participates in DEEP-ER project (booth #3741)
- The gap between compute speed an I/O is one major challenge
- Plan to keep POSIX interface
- Introduce API extensions
  - Give users more control over storage
- Exascale-ready local file systems?
  - ZFS?
  - BTRFS?

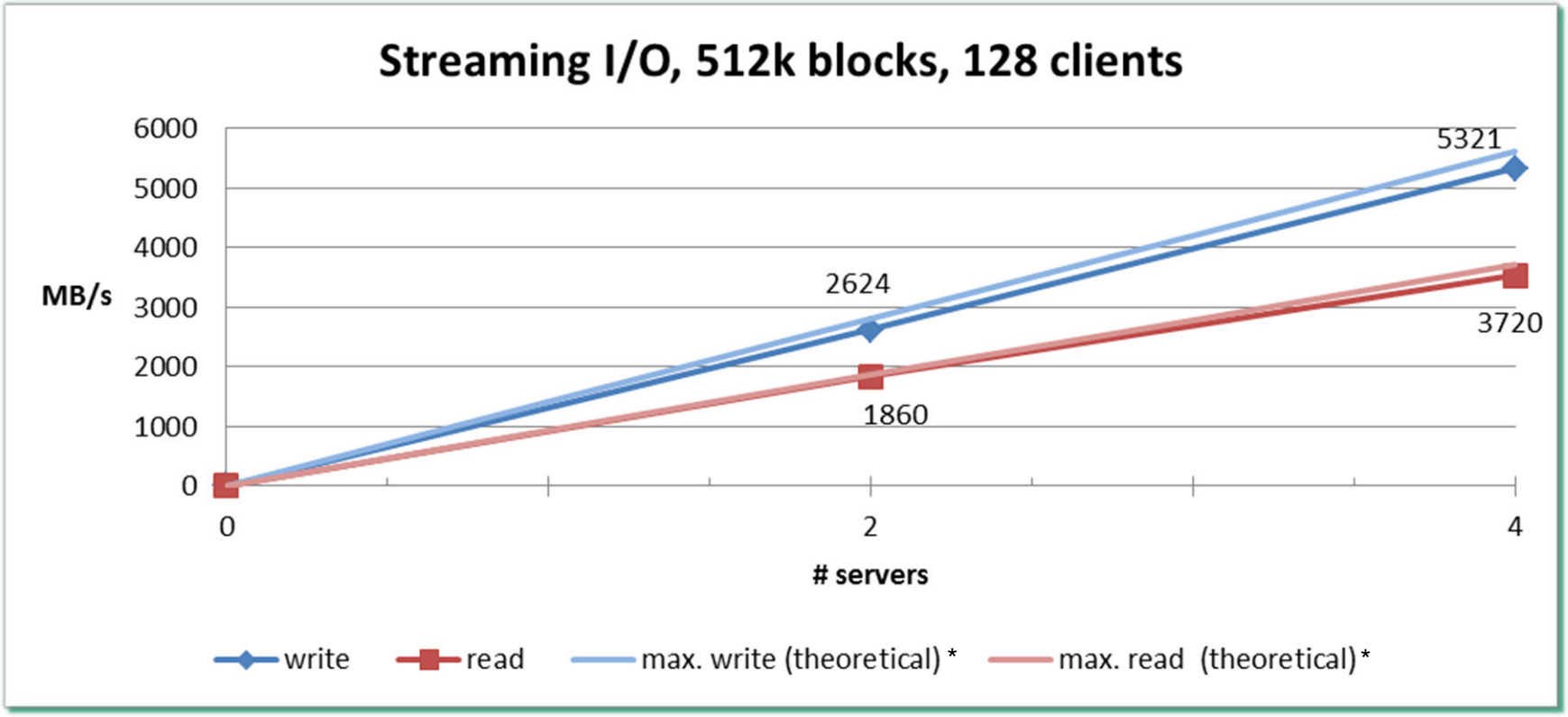
DEEP-ER

http://www.deep-er.eu

Playing /w BTRFS

- 4 servers as FhGFS metadata and storage servers
  - Intel Xeon E5-2640 @ 2.5GHz
  - 64 GB RAM
  - FDR Infiniband
  - 12x 2TB SAS HDD, 7200rpm
    - FhGFS storage targets
    - 2 BTRFS Raid 5 Volumes (5+1 each)
  - 2x 512GB SSD
    - FhGFS metadata
    - Ext4 Software Raid 1

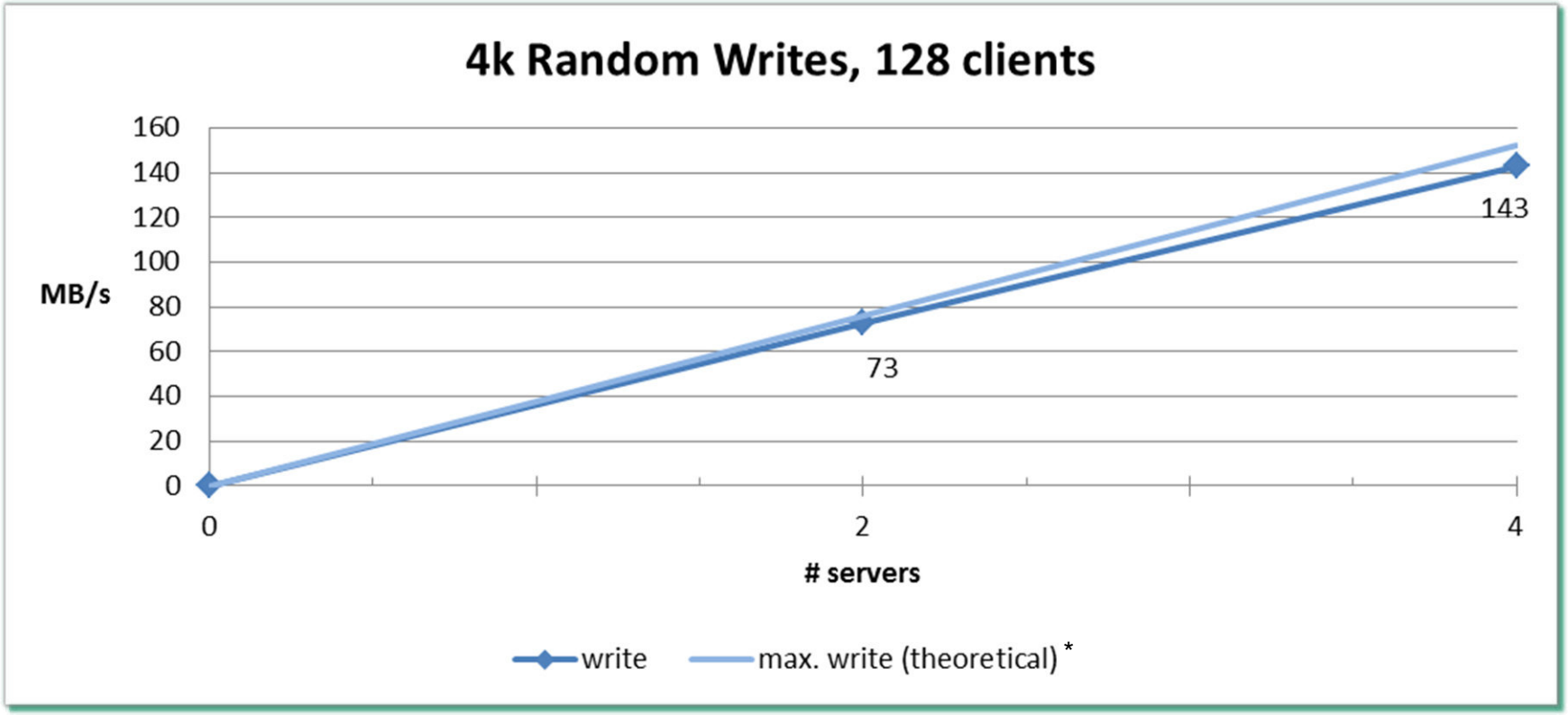
Benchmarks



* 1 local BTRFS Raid 5 set: 700MB/s write; 465MB/s read

Fraunhofer ITWM

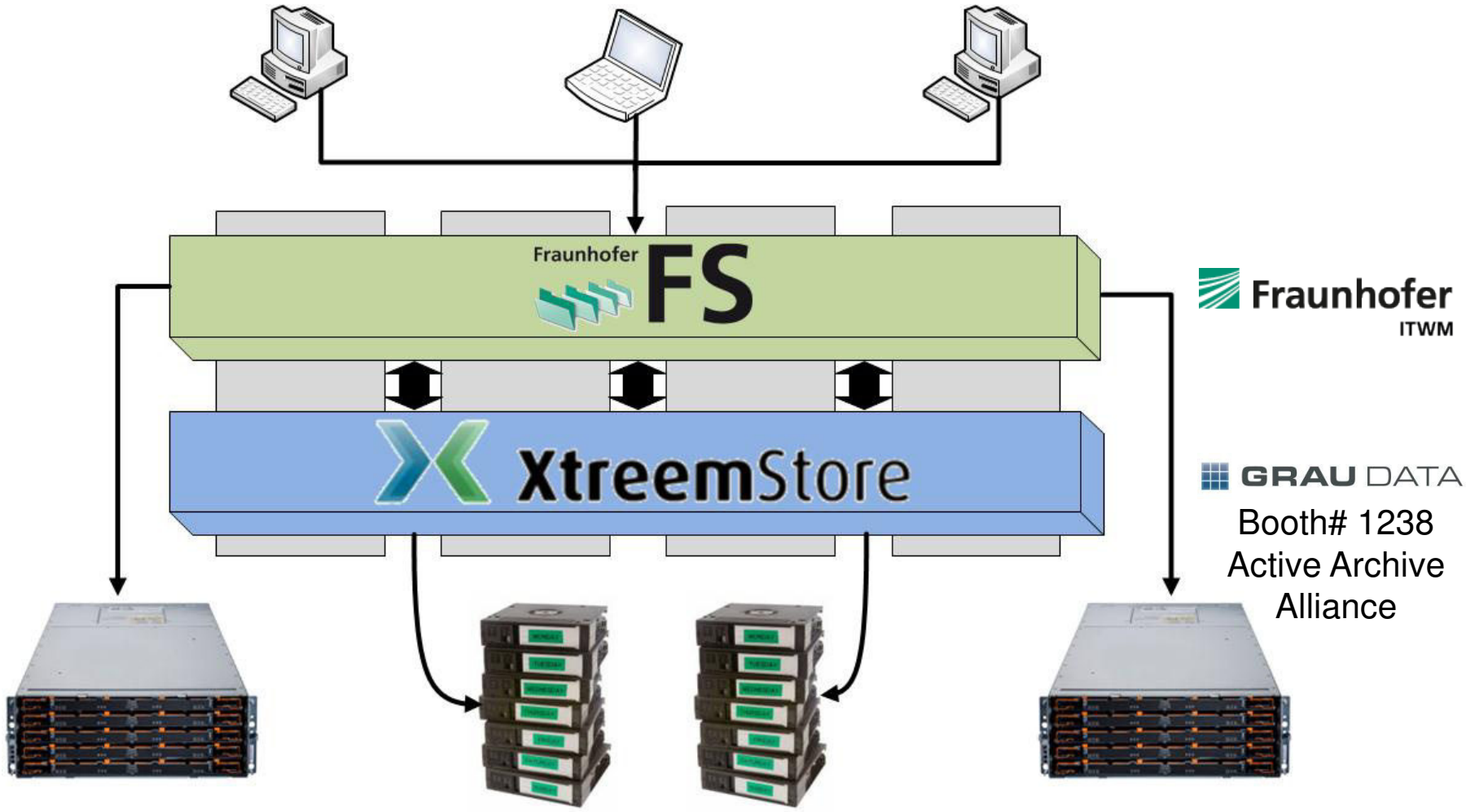
Benchmarks



* 1 local BTRFS Raid 5 set: 19 MB/s

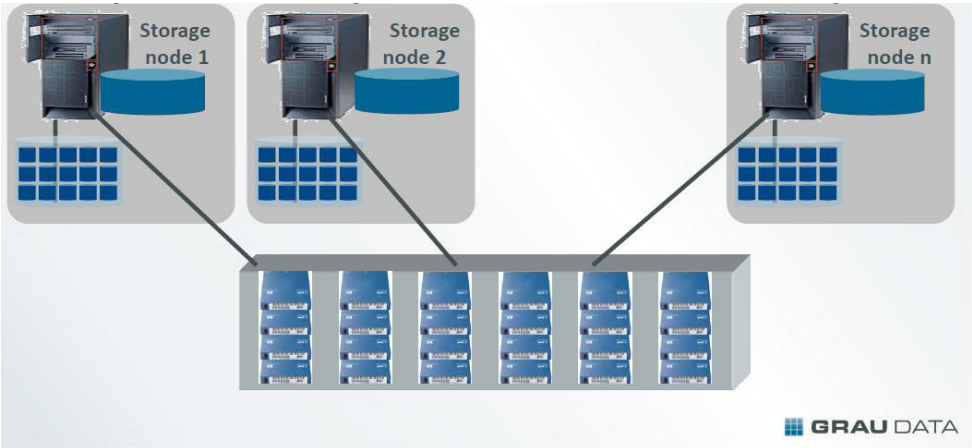
Fraunhofer ITWM

HSM Integration



FhGFS & Xtreemstore

- Fast Filesystem and fast HSM/Archiving
- Prototype system currently running at Fraunhofer
  - 4 FhGFS servers
    - currently ~100TB raw storage
  - 8 Tape drives
    - max. capacity: ~ 240TB / ~585TB (compressed)
  - Extension to 10 servers and 20 tape drives in progress



Fraunhofer ITWM

A New Name For FhGFS

BEEGFS

developed by Fraunhofer

...might remind you of these guys

...but actually means this

Questions?

http://www.fhgfs.com

http://wiki.fhgfs.com

support@fhgfs.com

Fraunhofer Booth

 # 1941

Fraunhofer
ITWM