Fraunhofer FS

A parallel file system – made in Germany

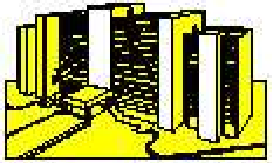
March 7th 2012
Dr. Franz-Josef Pfreundt   Competence Center for HPC
Sven Breuner

Fraunhofer
ITWM

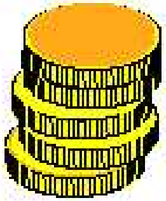
Fraunhofer

Non Profit Applied Research

18,000 employees



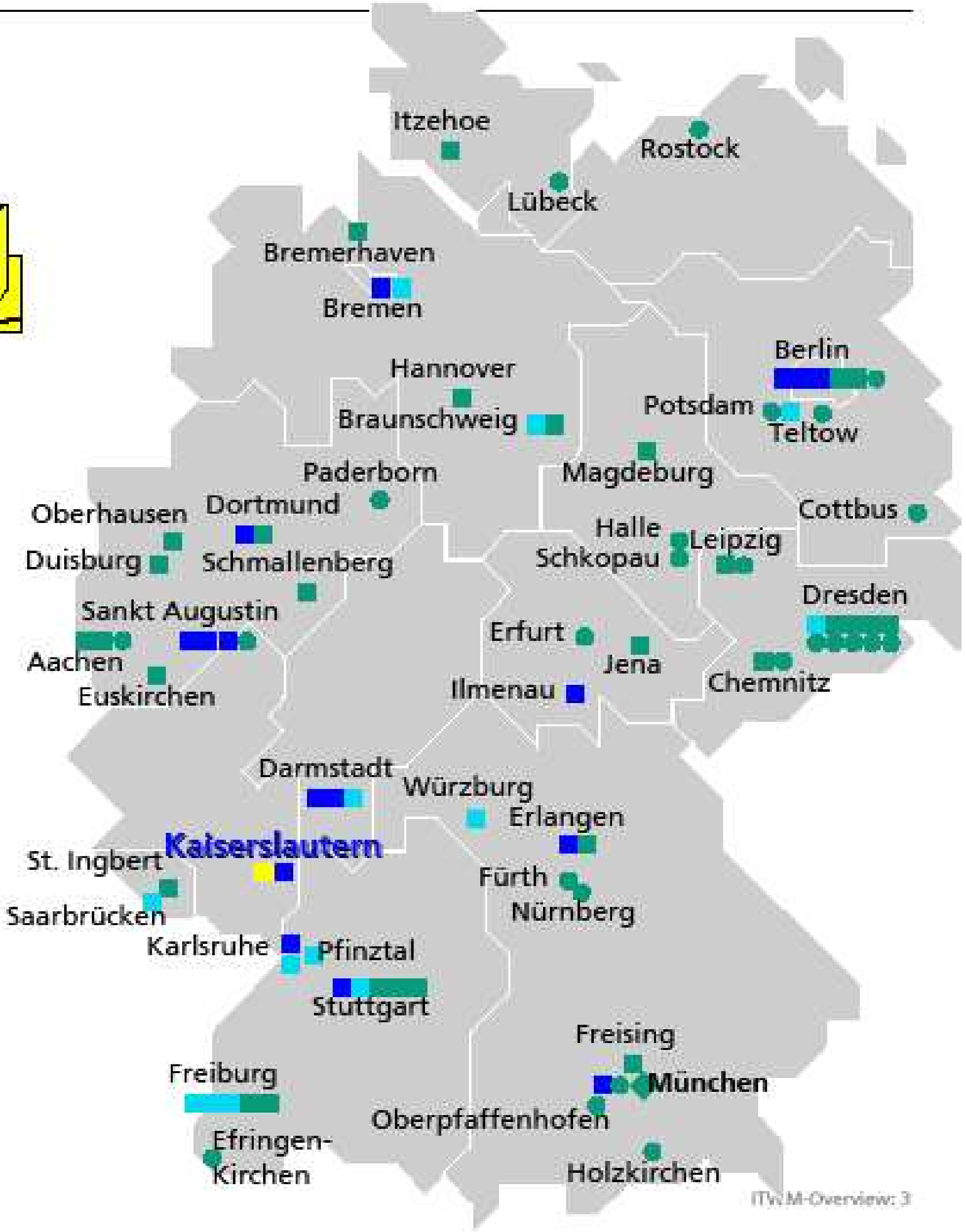
62 institutes



1.8 billion € budget

7 alliances

- Microelectronics
- Production
- **Information and Communication**
- Materials and Components
- Life Sciences
- Surface Technology and Photonics
- Defense and Security Research



Fraunhofer Institut for Industrial Mathematics

2

Mathematical models

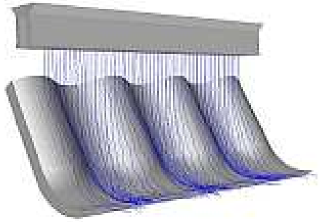
Algorithms

Simulations

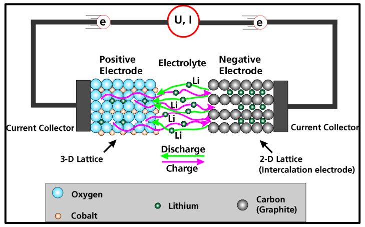
Software

Visualization

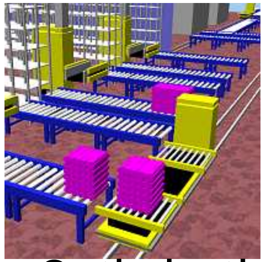
Data mining



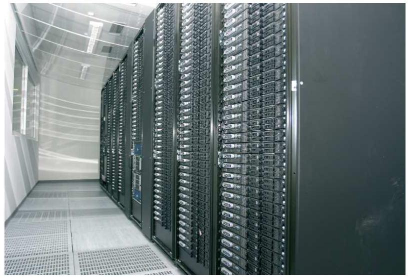
Fluid Dynamics



LI-ION Battery Sim



Optimization



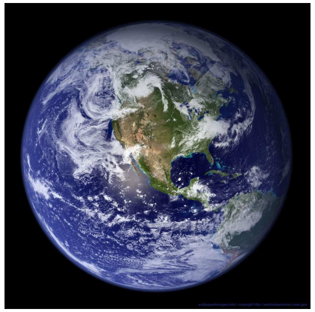
HPC

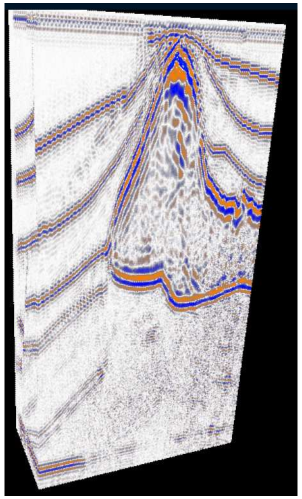
Fraunhofer Competence Center for HPC

Business Fields

Fraunhofer FS

GPI-Space





HPC –Tools **Visualization** **Green IT** **HPC Apps Seismic**

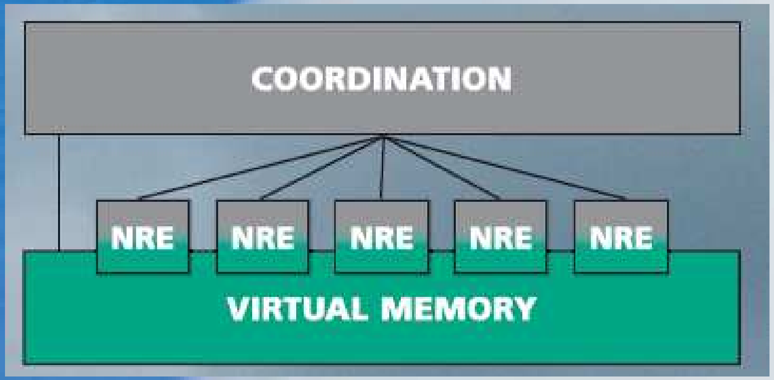
Research

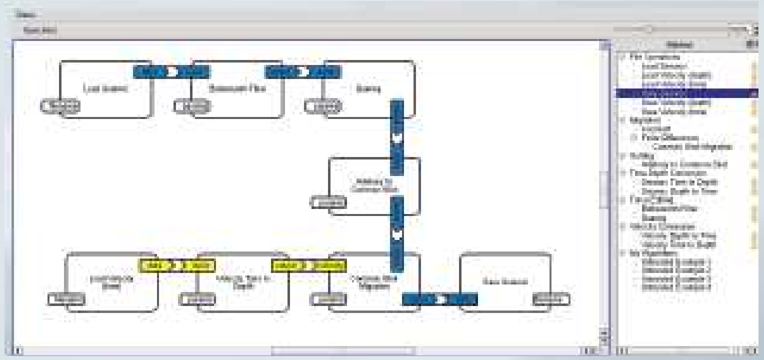
- parallel programming models
- distributed computing
- parallel algorithms
- parallel file systems
- ray tracing in visualization
- seismic imaging
- distributed energy managment

Staff : about 40 people

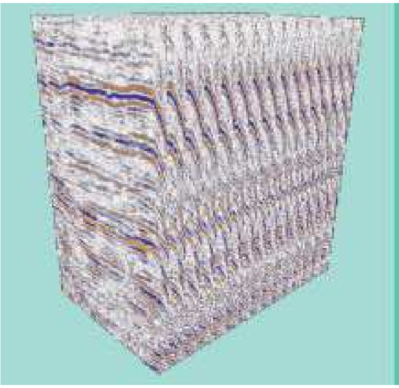
GPI -SPACE

Productive Development and Execution of Cluster&Cloud Applications





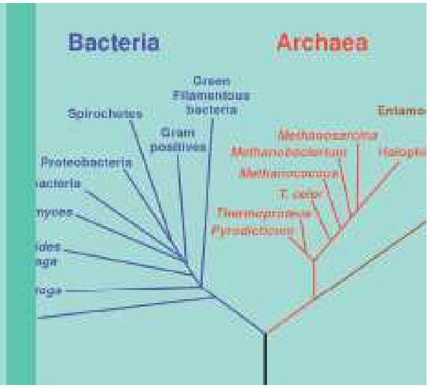
GPI Space



Seismic

Finance

Engineering



Life Sciences

- ➢ One large distributed virtual memory space
- ➢ Optimal throughput - dynamic load balancing
- ➢ Failure tolerant execution
- ➢ Autoparallelization of complex workflows

FraunhoferFS How it started

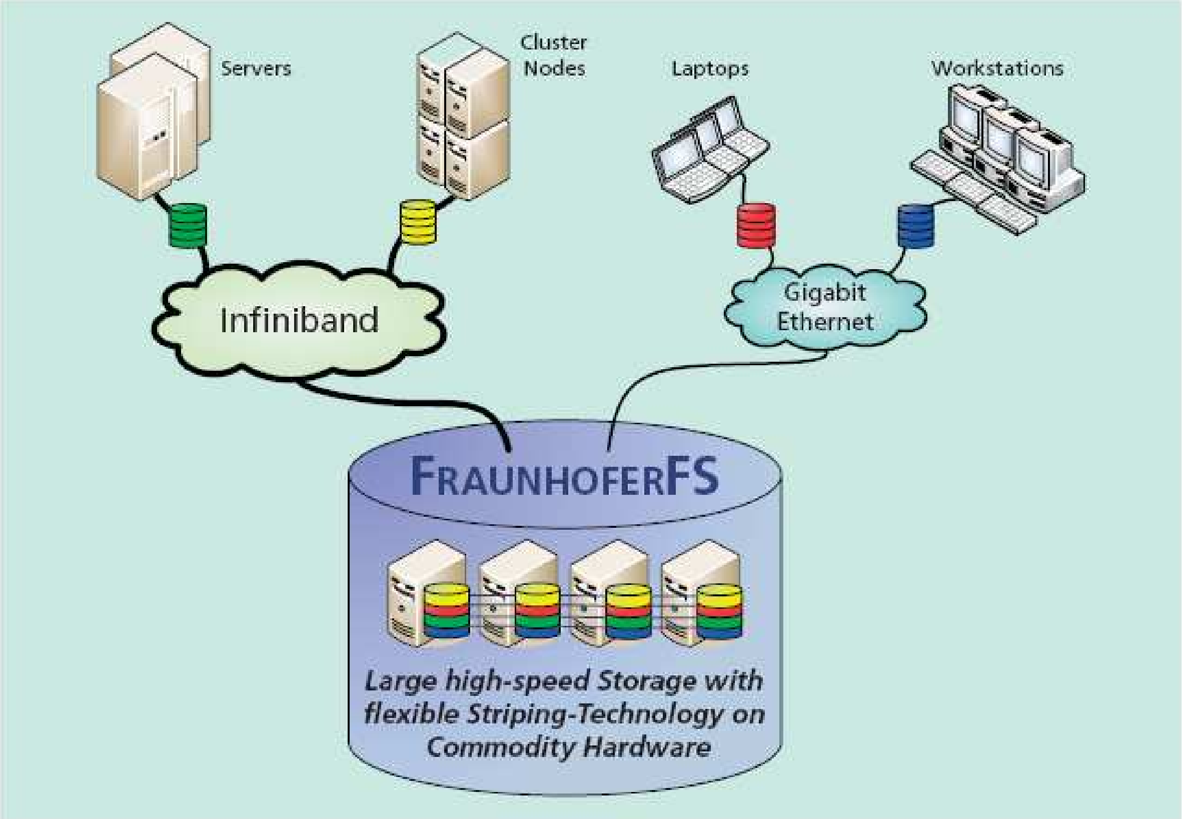
As part of a cooperation with Linux NetworX

2003 First Lustre Installation at Fraunhofer
2004 Port of the Blue Order Media Server System
on top of Lustre

2004 Decision to develop the Fraunhofer FS

Requirements: Distributed Metadata
No Kernel patches, zero config clients
Scalable multithreaded architecture
Native IB and Ethernet
Easy to install and maintain
Use P2P technology

Anouncement after 3 years of development (ISC Dresden)

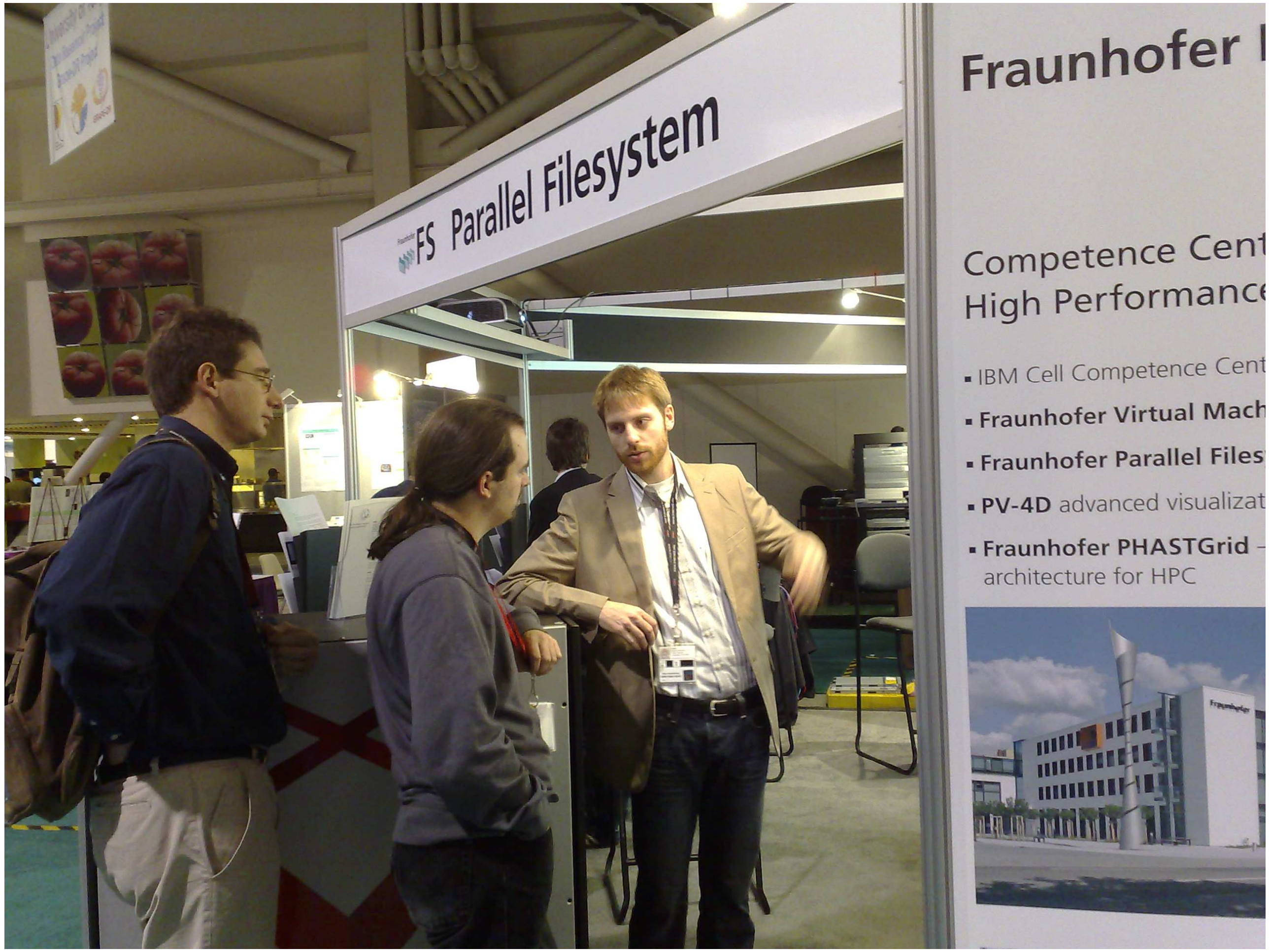


Large high-speed Storage with flexible Striping-Technology on Commodity Hardware

Fraunhofer Parallel Filesystem

Available Q4/2007

SC 2007 Reno Introduction of the FhGFS



SC 2010 New Orleans

Fraunhofer ITWM
GLOBAL ADDRESS SPACE PROGRAMMING INTERFACE
GPI
A PARALLEL FILE SYSTEM DEDICATED TO THE HPC COMMUNITY
Fraunhofer FS
www.fhgfs.com
Gate 18
GPI – GLOBAL ADDRESS SPACE PROGRAMMING INTERFACE
Efficient and scalable multicore programming
Superior to MPI in Perfomance and Scalability
FHGFS – FRAUNHOFER PARALLEL FILE SYSTEM
LexisNexis

Customer Base

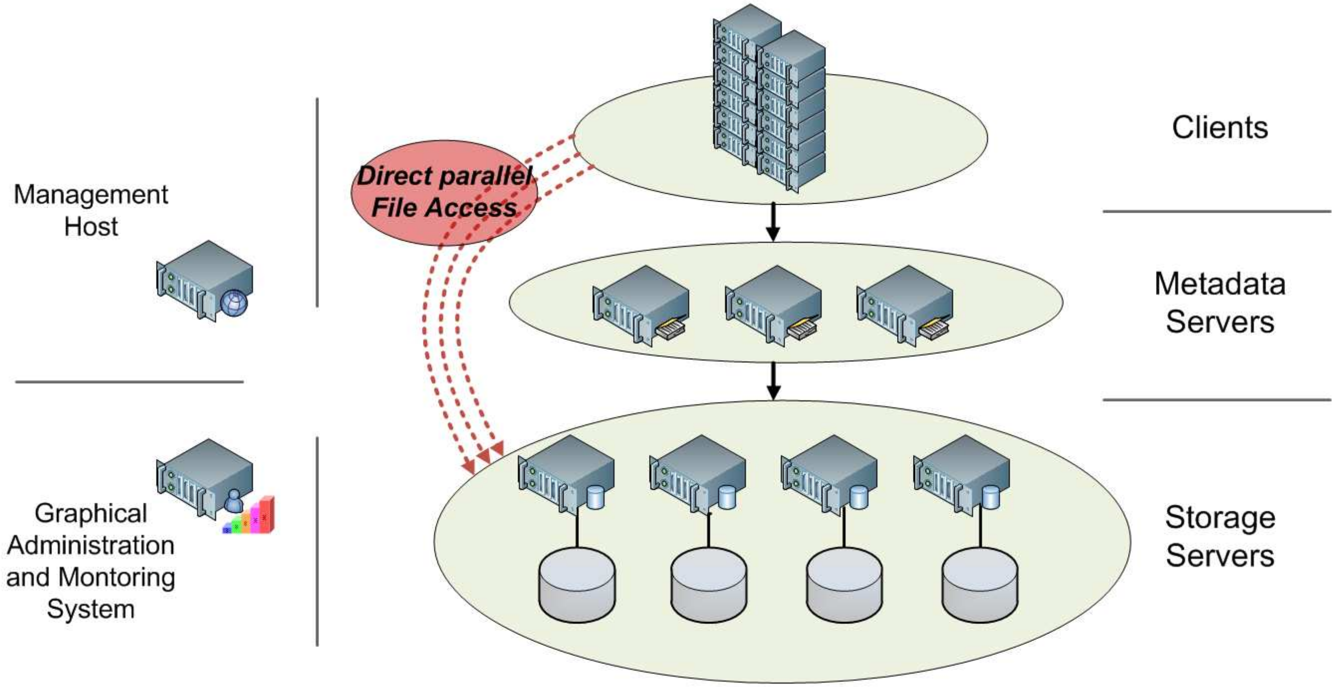
Oil&Gas
Universities

~ 30 supported customers

FhGFS Key Features

- ❑ **Maximum Scalability**
  - Distributed File Contents & Metadata
  - Low server load – efficient multithreading

→ Installed system
More than 300 servers

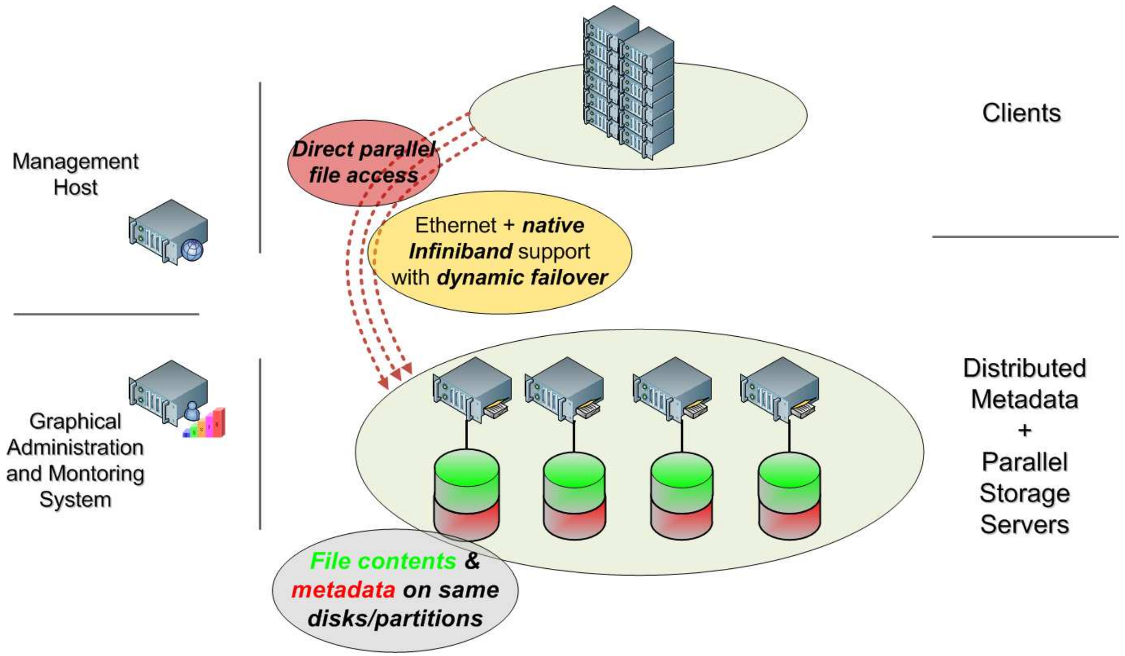

Management Host
Graphical Administration and Montoring System
Direct parallel File Access
Clients
Metadata Servers
Storage Servers

Object storage , servers use a local file system (XFS, EXT,…)

FhGFS Key features

- ❑ **Flexibility**
  - ❍ Add Clients and Servers without Downtime
  - ❍ Client and Servers can run on same Machine
  - ❍ On-the-fly storage init (mkfs)
  - ❍ Multiple Networks with dynamic Failover

Storage Cluster
Compute + Storage


Management Host
Direct parallel file access
Clients
Ethernet + native Infiniband support with dynamic failover
Graphical Administration and Montoring System
Distributed Metadata + Parallel Storage Servers
File contents & metadata on same disks/partitions

- o Flexible Striping: individual Settings on a per-File /per-Directory Basis

Fraunhofer Seislab - interactive seismic imaging

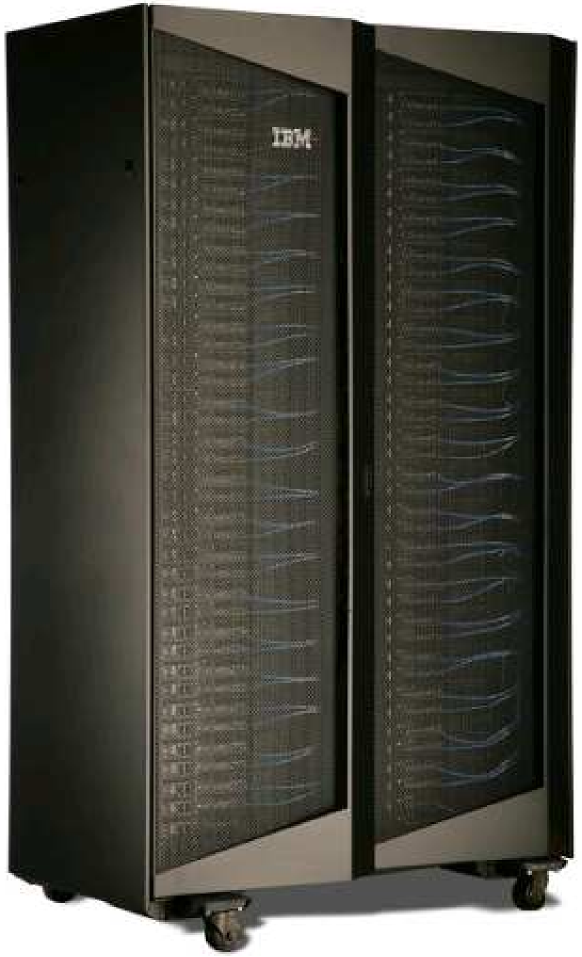
Compute & Storage

20 Compute Nodes
48 -96 GB RAM
4 x 256 GB SSD striped
QDR Infiniband

5 Compute&Storage Nodes
20 TB SATA , RAID5 (Archive)
QDR Infiniband

On demand SSD based FhGFS
per job up to 20 TB

Read: 30 GB/sec Write: 20GB/sec



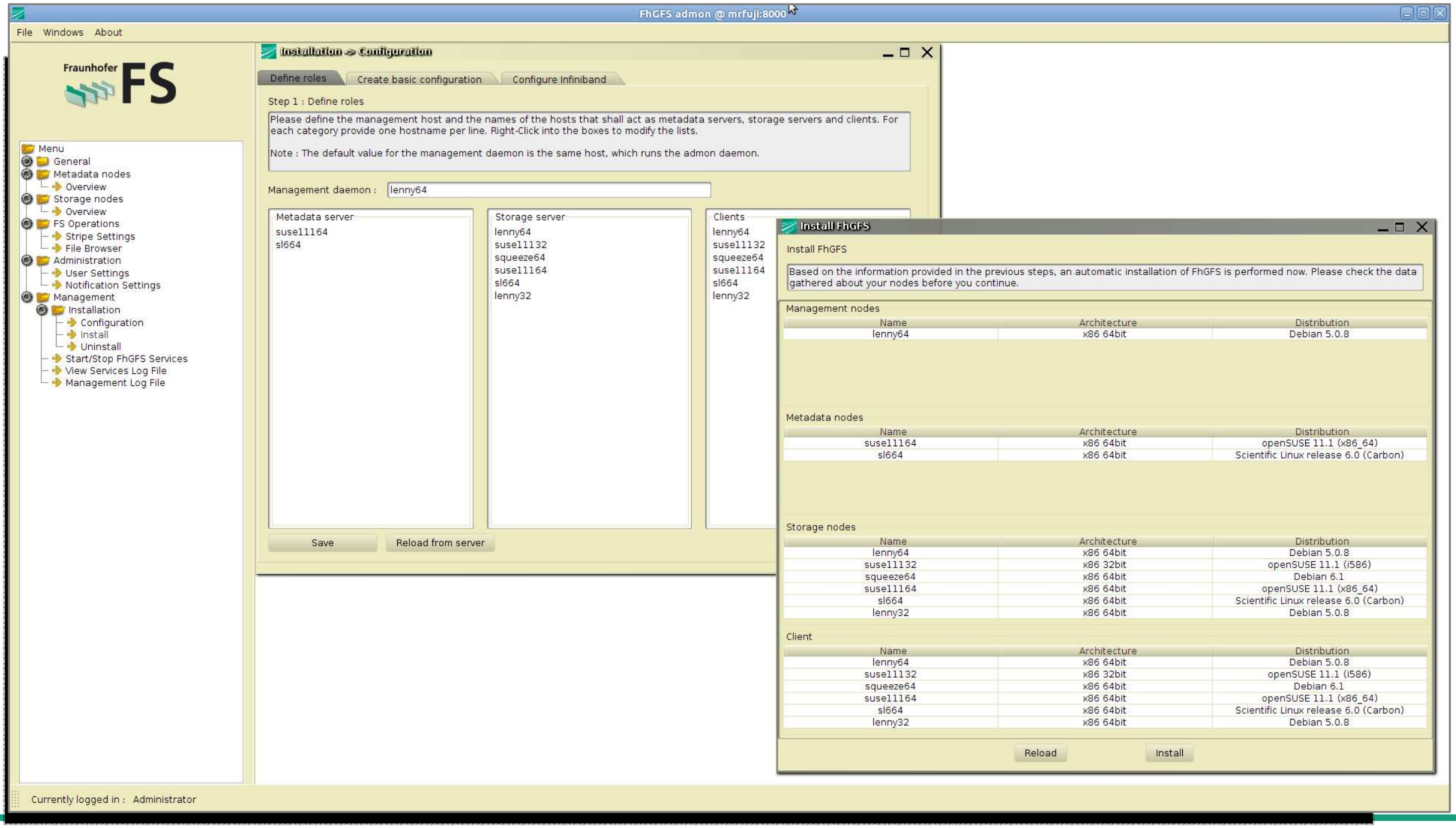
Network bisection BW ~ I/O performance

FhGFS Key Features

Easy to use

- Automated Cluster Installation
- Kernel Module
- Graphical System Administration & Monitoring

- No specific Linux Distribution,
- no special Hardware required



Step 1 : Define roles

Please define the management host and the names of the hosts that shall act as metadata servers, storage servers and clients. For each category provide one hostname per line. Right-Click into the boxes to modify the lists.

Note : The default value for the management daemon is the same host, which runs the admon daemon.

Management daemon : lenny64

Metadata server
suse11164
sl664

Storage server
lenny64
suse11132
squeeze64
suse11164
sl664
lenny32

Clients
lenny64
suse11132
squeeze64
suse11164
sl664
lenny32

Currently logged in : Administrator

Install FhGFS

Based on the information provided in the previous steps, an automatic installation of FhGFS is performed now. Please check the data gathered about your nodes before you continue.

Management nodes

Name	Architecture	Distribution
lenny64	x86 64bit	Debian 5.0.8

Metadata nodes

Name	Architecture	Distribution
suse11164	x86 64bit	openSUSE 11.1 (x86_64)
sl664	x86 64bit	Scientific Linux release 6.0 (Carbon)

Storage nodes

Name	Architecture	Distribution
lenny64	x86 64bit	Debian 5.0.8
suse11132	x86 32bit	openSUSE 11.1 (i586)
squeeze64	x86 64bit	Debian 6.1
suse11164	x86 64bit	openSUSE 11.1 (x86_64)
sl664	x86 64bit	Scientific Linux release 6.0 (Carbon)
lenny32	x86 64bit	Debian 5.0.8

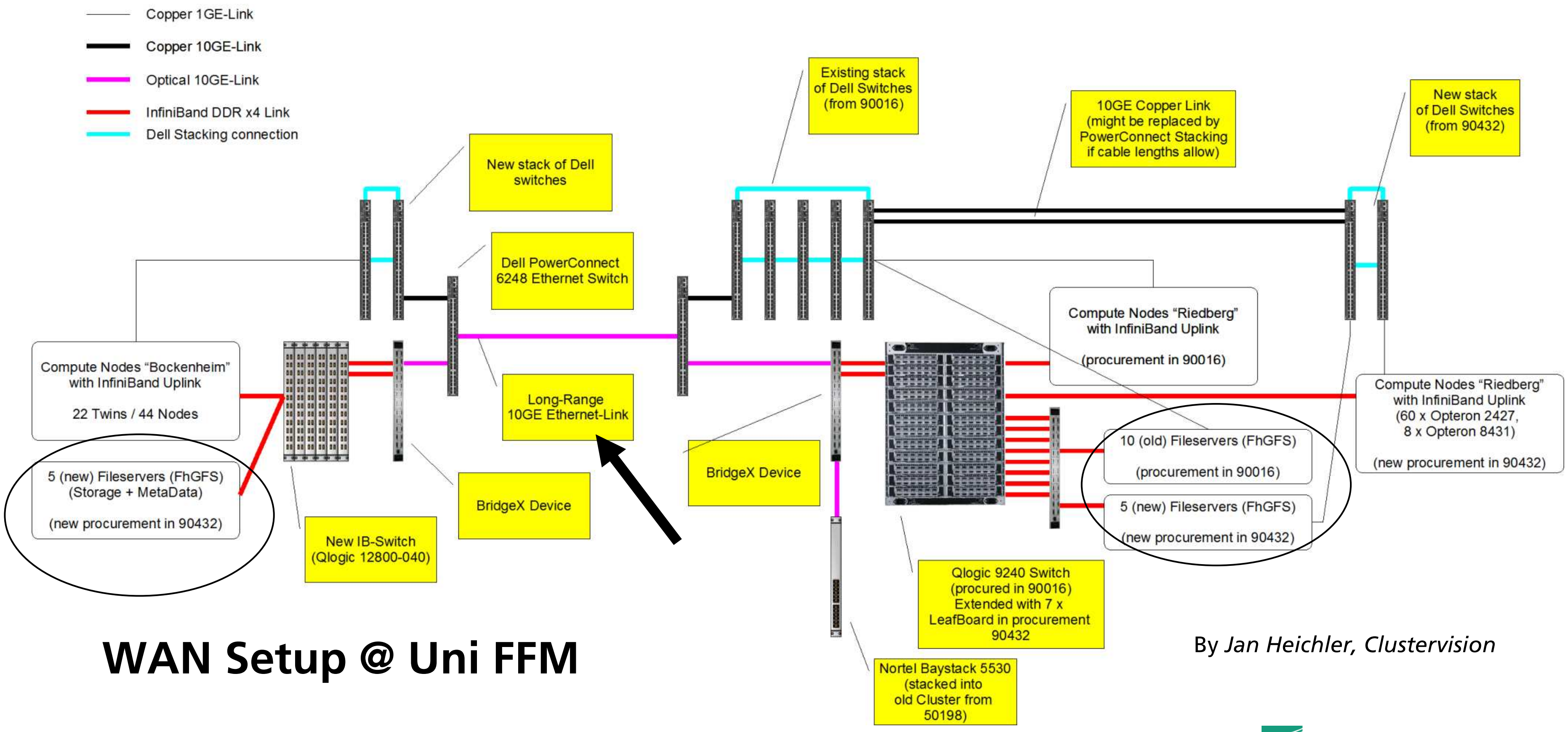
Client

Name	Architecture	Distribution
lenny64	x86 64bit	Debian 5.0.8
suse11132	x86 32bit	openSUSE 11.1 (i586)
squeeze64	x86 64bit	Debian 6.1
suse11164	x86 64bit	openSUSE 11.1 (x86_64)
sl664	x86 64bit	Scientific Linux release 6.0 (Carbon)
lenny32	x86 64bit	Debian 5.0.8

FhGFS Key Features

- Light-weight Client Kernel Module
  - ❑ High Single-Stream Throughput (>2.7GB/s on QDR IB)

- Server Preference
  - ❑ Clients can prefer a Subset of Servers => Support for multiple Data Centers


Copper 1GE-Link
Copper 10GE-Link
Optical 10GE-Link
InfiniBand DDR x4 Link
Dell Stacking connection
New stack of Dell switches
Dell PowerConnect 6248 Ethernet Switch
Existing stack of Dell Switches (from 90016)
10GE Copper Link (might be replaced by PowerConnect Stacking if cable lengths allow)
New stack of Dell Switches (from 90432)
Compute Nodes "Bockenheim" with InfiniBand Uplink
22 Twins / 44 Nodes
5 (new) Fileservers (FhGFS) (Storage + MetaData)
(new procurement in 90432)
New IB-Switch (Qlogic 12800-040)
BridgeX Device
Long-Range 10GE Ethernet-Link
BridgeX Device
Compute Nodes "Riedberg" with InfiniBand Uplink
(procurement in 90016)
Compute Nodes "Riedberg" with InfiniBand Uplink (60 x Opteron 2427, 8 x Opteron 8431)
(new procurement in 90432)
10 (old) Fileservers (FhGFS)
(procurement in 90016)
5 (new) Fileservers (FhGFS)
(new procurement in 90432)
Qlogic 9240 Switch (procured in 90016) Extended with 7 x LeafBoard in procurement 90432
Nortel Baystack 5530 (stacked into old Cluster from 50198)

WAN Setup @ Uni FFM

By *Jan Heichler, Clustervision*

Typical Size and performance of current installations

Frankfurt University : 12 servers ,1 PByte , 20GB/sec , 900 clients
measured single stream performance : 2,0 GB/sec

TU Vienna : 12 servers, 300 TByte, 6GB/sec, 1200 clients
12 metadata server(SSD), x00 000 I/O Ops/sec

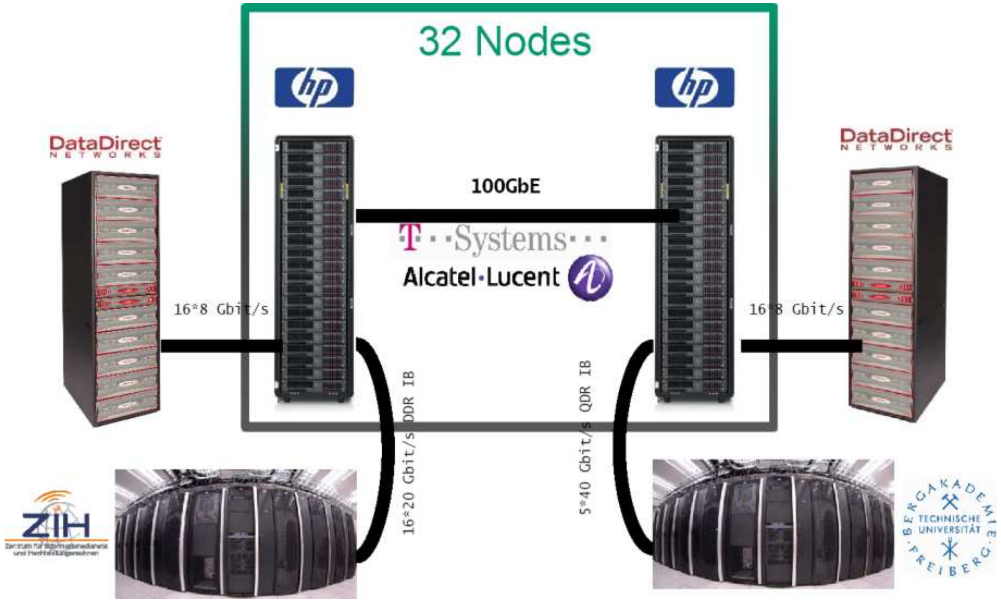
RSI (Houston) : 12 server, 300 Tbyte, 6 GB/sec, 28 clients
client and server on same machine

Fraunhofer Seislab : 20 servers, 20 TB SSD,120TB SATA, 30 GB/sec
server and clients

DTU Kopenhagen : 5 servers, 200 TByte, 5GB/sec, 100 clients
port to BSD UNIX

100GBit Testbed (Dresden <-> Freiberg)

Dresden
Freiberg
100 Gbit/s



By *Michael Kluge, TU Dresden*

Uni-directional

- GPFS - 10,1 GB/s (60 km)
- Lustre - 11,8 GB/s (60 km)
- FhGFS - 12,4 GB/s (400 km)

Bi-directional

- GPFS - *n/a*
- Lustre - 21,9 GB/s (60 km)
- FhGFS - 22,5 GB/s (400 km)

Last major release August 2011

- Client operation counters
- All file attributes stored on metadata server
- **Distributed POSIX file locking**
- Simplified automatic updates via software repositories
- Multiple storage targets per server
- Re-designed metadata request handling to scale to high numbers of CPU cores
- Parallel online file system check/repair

Faster,

more flexible,

easier to use

Business Model

No license fees

Pay for support and maintenance

Open Source - on a individual basis

So far not a community request

Our supported customers (~ 50)

HPC Centers

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

universität wien

University Oslo

Universität Stuttgart

DESY

University of Zagreb

UNIVERSITÉ BORDEAUX 1 Sciences Technologies

university of groningen

Oil&Gas

MARATHON

Statoil

NORECO

RSI

REPSOL

DETNORSKE

... And more

Cloud Computing

Social Media

No system Halt for Software resons

Happy users

About the FhGFS Roadmap

Some FhGFS roadmap pillars are fixed, e.g.:

- HA
- HSM

We leave some room to implement interesting user ideas, e.g.:

- Server affinity
- Client operation counters

We learned that we need to leave some room to improve Linux kernel / tools, e.g.:

- tail, ls -l, Linux RDMA

And we have enough people in the institute that develop HPC appplications with disruptive new ideas

- Fraunhofer Seislab
  - 20 compute nodes with SSDs
  - Runs FhGFS on-demand
  - Jobs store temporary data on SSDs and move it to dedicated servers afterwards
  - 20GB/s write, 25GB/s read sustained

Next major release 2012

- Data/metadata mirroring over multiple FhGFS servers
- Configurable on a per-file (per-directory) basis
- Server groups for remote mirroring
- Quota/ACL support
- MAC support (Q2 2012)

→ Next major release Q2 2012

Hierarchical Storage Management

GrauData provides Grau ArchiveManager (GAM) as a solid single-server HSM solution

Fraunhofer and Grau teamed up to integrate GAM and FhGFS

The combined solution will support…

- Parallel data migration (e.g. recall all file chunks at once)
- Collocation IDs
- Asynchronous recalls

First prototype will run at HLRS Q3 2012

Questions ?

http://www.fhgfs.com

Franz-Josef Pfreundt , Sven Breuner
pfreundt@itwm.fhg.de ,breuner@itwm.fhg.de