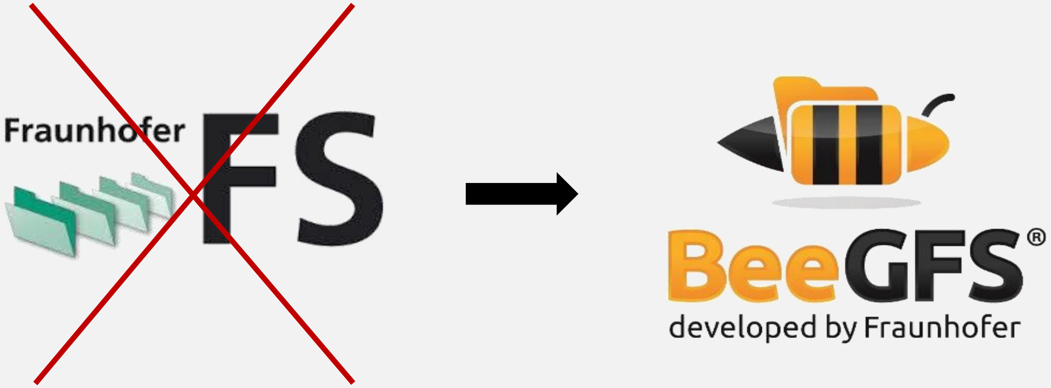
BeeGFS – not only for HPC

www.beegfs.com

24/09/2015 | Christian Mohrbacher

What is BeeGFS?

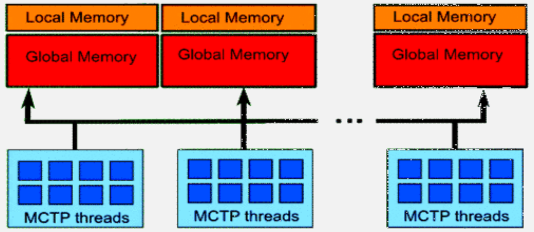
- A hardware independent parallel filesystem
- Designed for high performance and high throughput environments
- Developed at Fraunhofer ITWM (original name: FhGFS)
- Productive installations since 2007
- First commercial installation in 2009
- Renamed to BeeGFS in 2014
- Free to use
- Commercial support available



Fraunhofer ITWM

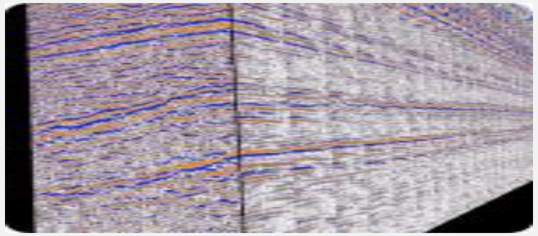
- Institute for Industrial Mathematics
- Located in Kaiserslautern, Germany
- Staff: ~ 260 employees + ~ 60 PhD students in 8 departments

Fraunhofer ITWM – Competence Center HPC



Parallel Programming models & tools

Photo realistic real time ray tracing



Interactive seismic imaging

Parallel File Systems

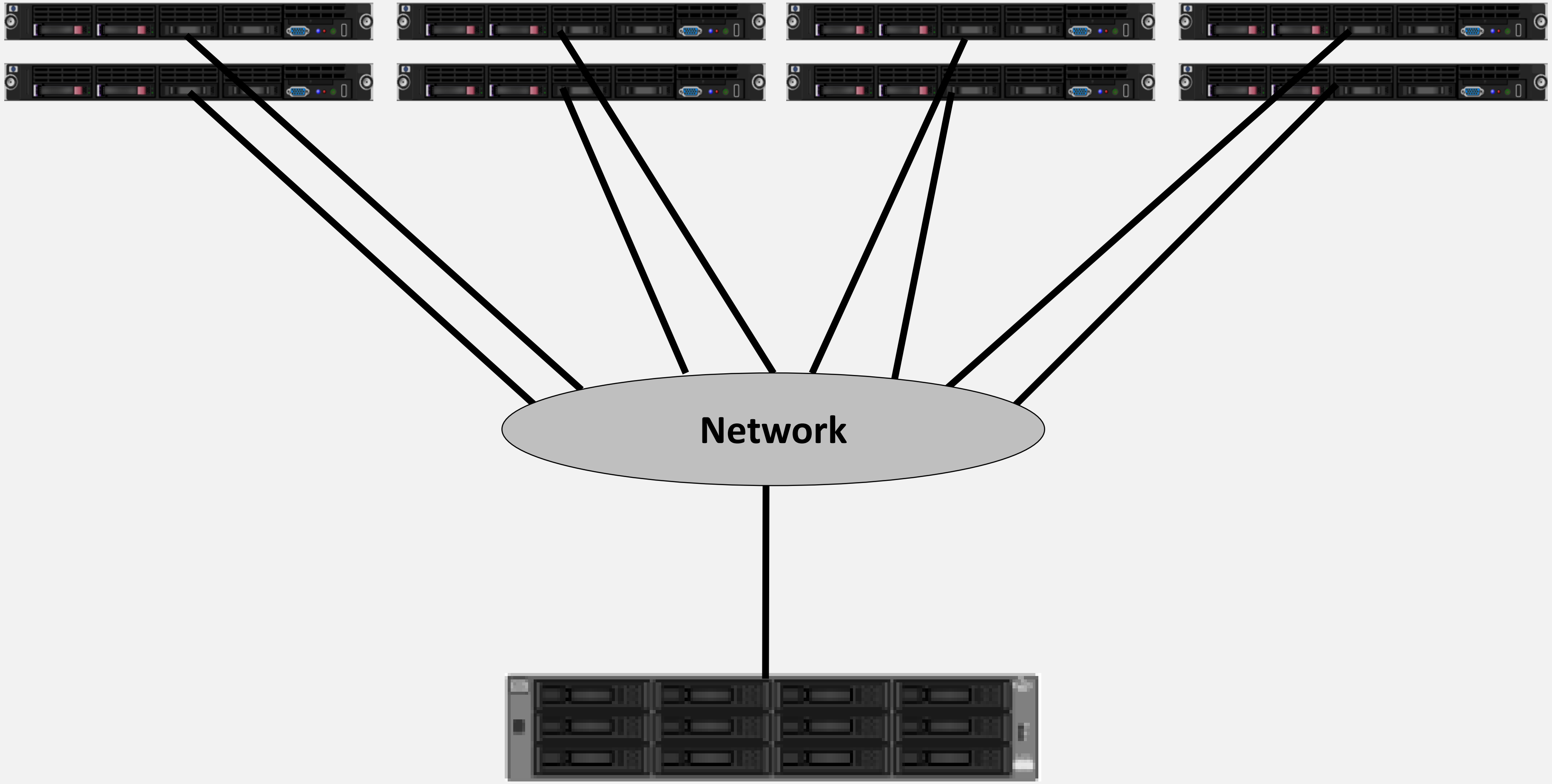
Big Data

Smart Energy / Green by IT

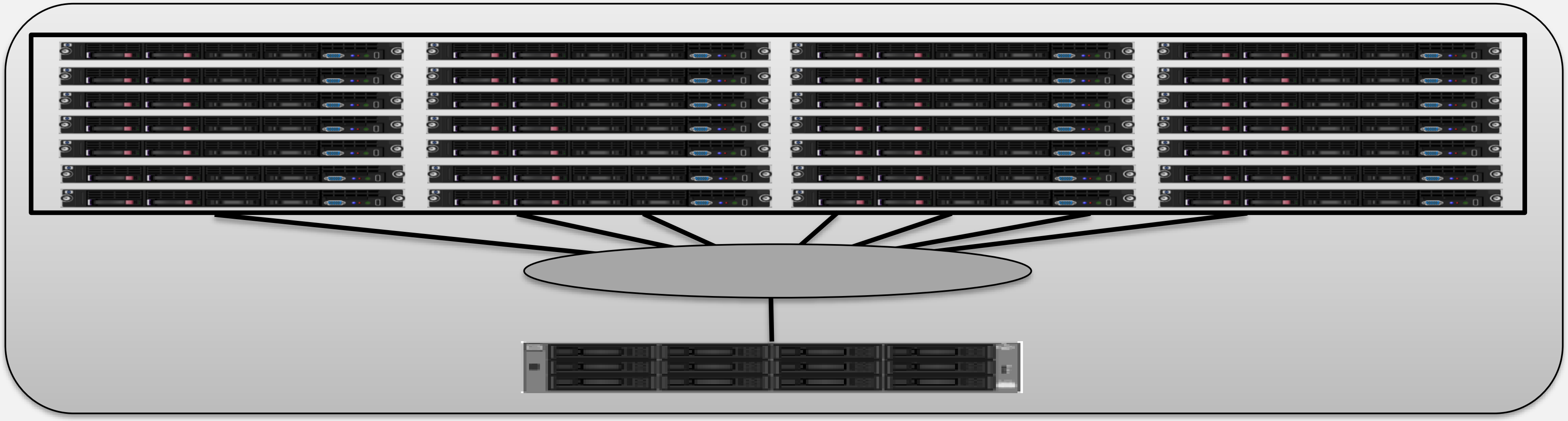
ThinkParQ?

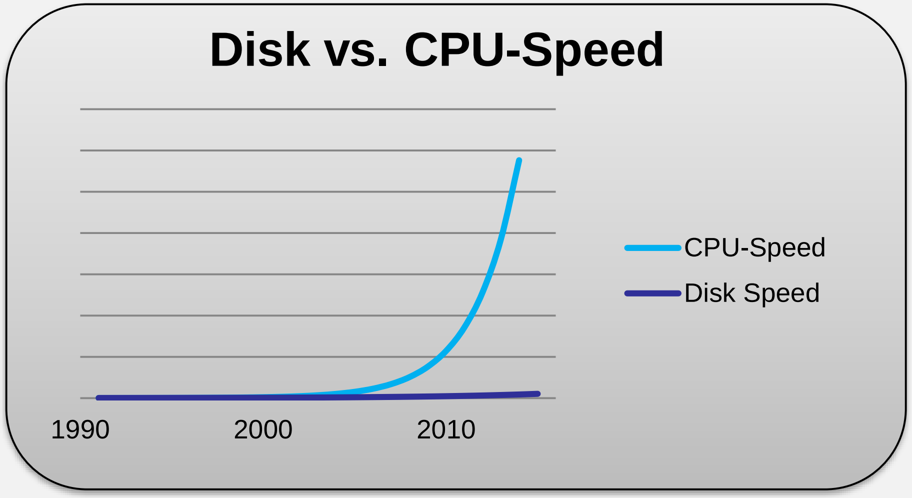
- ThinkParQ…
  - … is a spin-off from Fraunhofer ITWM to bring BeeGFS to the market
  - … does consulting, services and support around BeeGFS
  - … manages partner relationships
  - … will develop own add-ons for BeeGFS

Most storage systems still look like this...



But why bother?

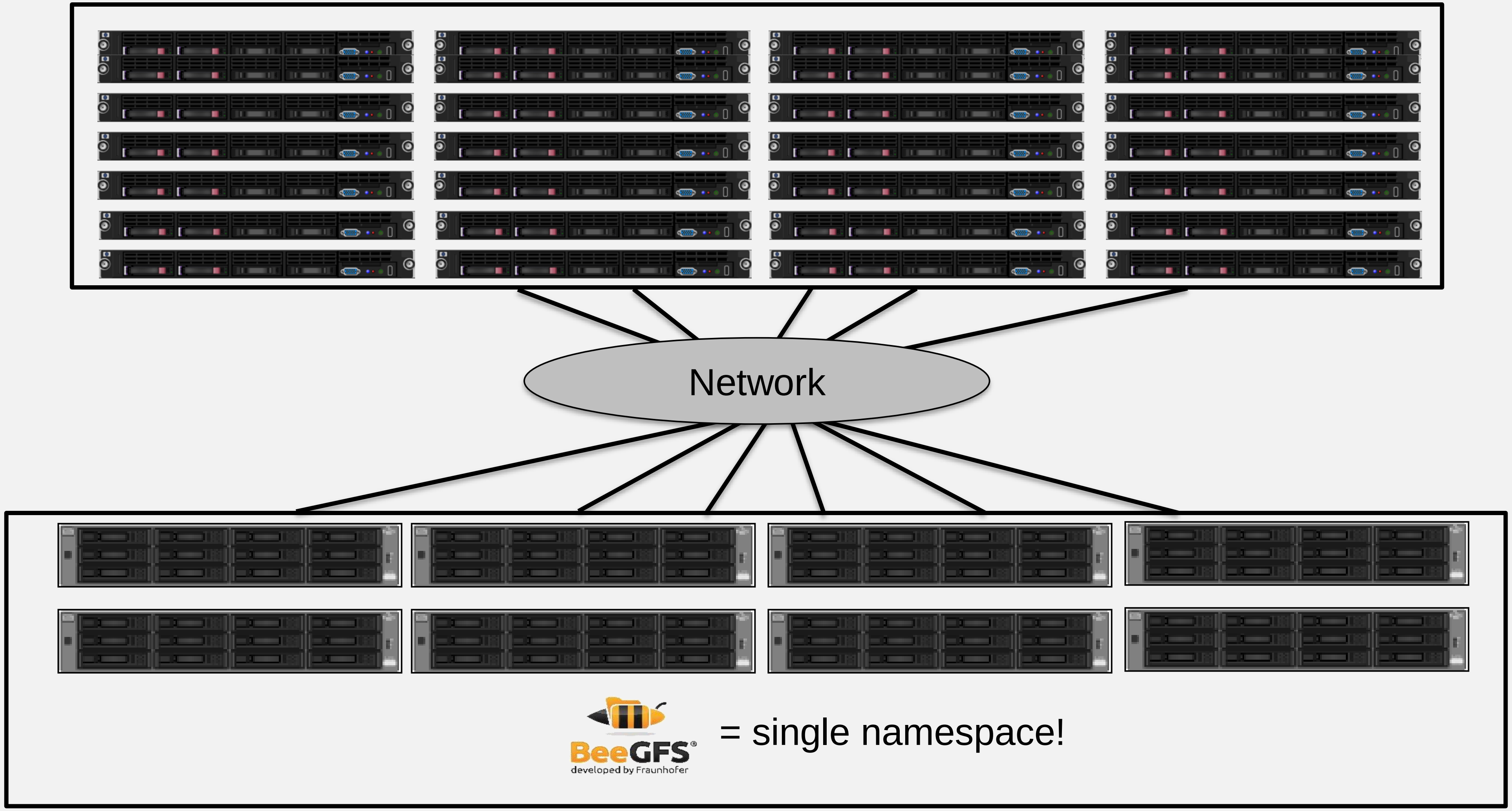




„A supercomputer is a device for turning compute-bound problems into I/O-bound problems."

- Ken Batcher

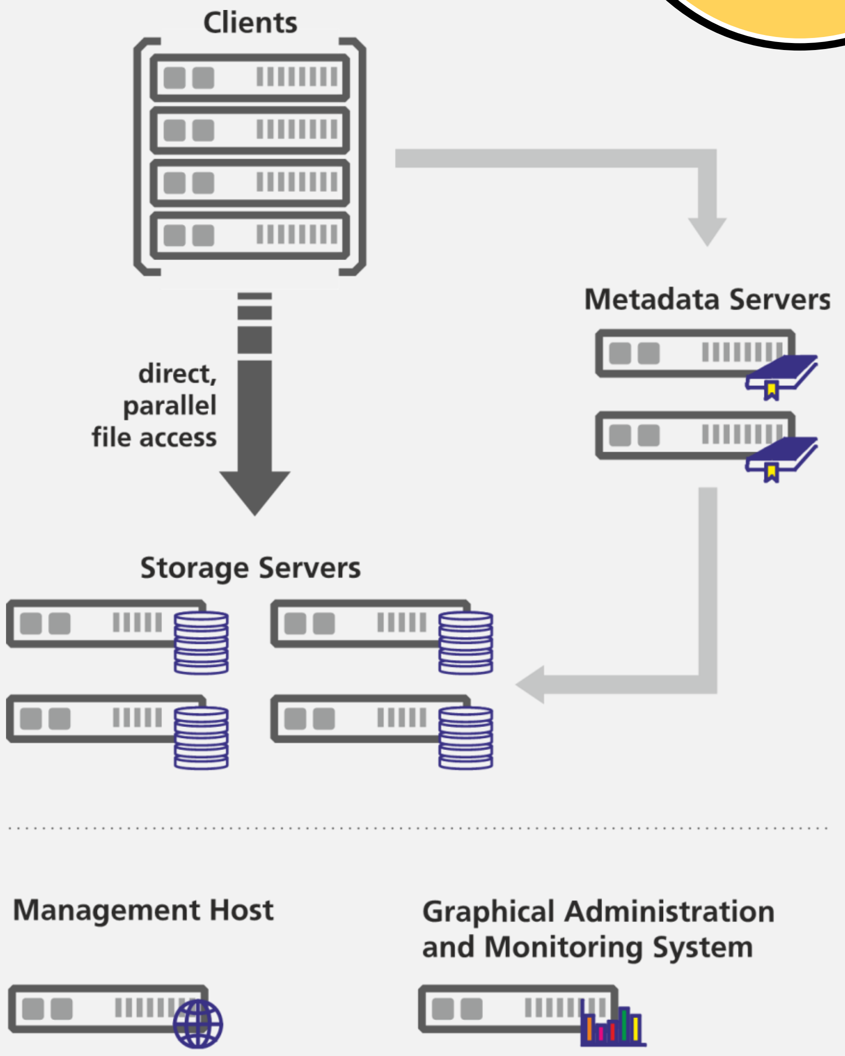
Scale out storage



= single namespace!

BeeGFS Architecture

- Management Host
  - Maintains a list of all components in the system
  - Provides all neccessary information to new components
- Storage Servers
  - Store the (distributed) file contents
- Metadata servers
  - Manage the metadata of file system entries
  - Maintain striping information for files
  - Not involved in data access
- Client
  - Native client module to mount the file system
- Graphical Administration and Monitoring System
  - GUI to perform administrative tasks and monitor system information



Key Concepts

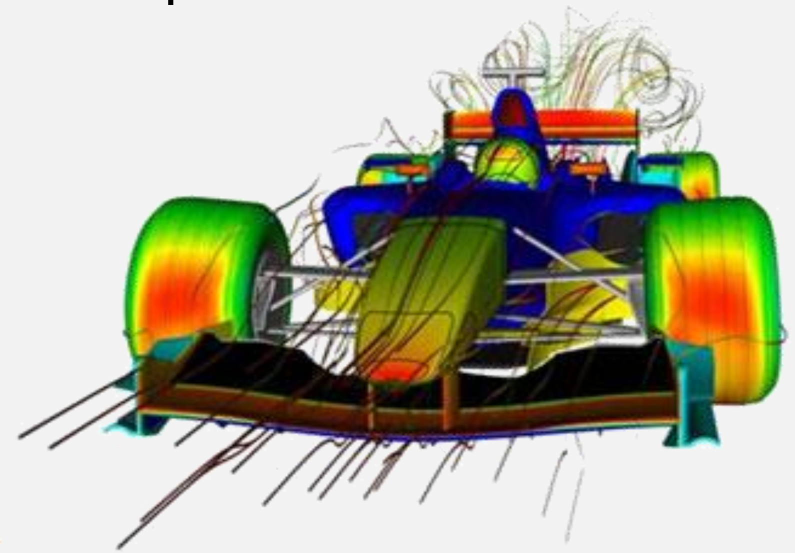
Performance Scalability

Flexibility

Easy to use

Key Features

- Performance & scalability
  - Initially optimized for HPC
  - Completely multithreaded – lightweight design
  - Supports GigE/10GE/40GE (TCP/RoCE) and InfiniBand (TCP/RDMA)
  - Distributed file contents: aggregated throughput of multiple servers
  - Distributed metadata across multiple servers
  - Excellent single stream performance



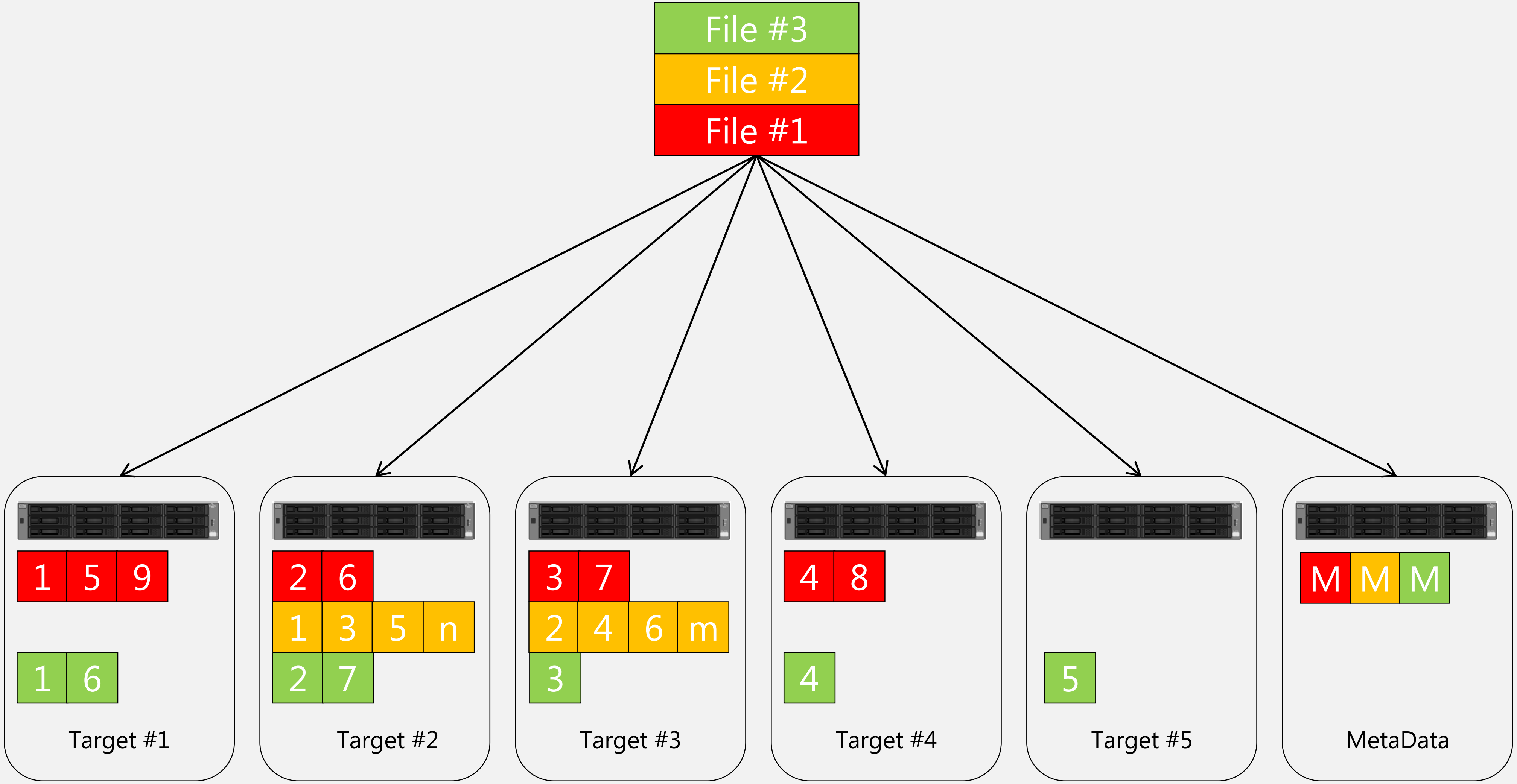
Key Features

- Performance & scalability
- Flexibility
    - Multiple daemons (any combination) can run on the same machine
    - Flexible striping per file/per directory
    - Add servers without downtime
    - On demand filesystem „per job“ possible
    - Client runs on any kernel >2.6.18
    - Client runs on Xeon PHI
    - ARM port available
    - NFS & SMB/CIFS re-export possible

Key Features

- Performance & scalability
- Flexibility
- Easy to use
  - Servers run in user space
  - No kernel patches
  - Servers use existing local filesystems (ext4, xfs, zfs, ...)
  - Packages for RHEL/SL/CentOS/SLES/Debian/Ubuntu
  - Hardware independent
  - Graphical monitoring tool

Striping



File #3

File #2

File #1

1 5 9

1 6

Target #1

2 6

1 3 5 n

2 7

Target #2

3 7

2 4 6 m

3

Target #3

4 8

4

Target #4

5

Target #5

M M M

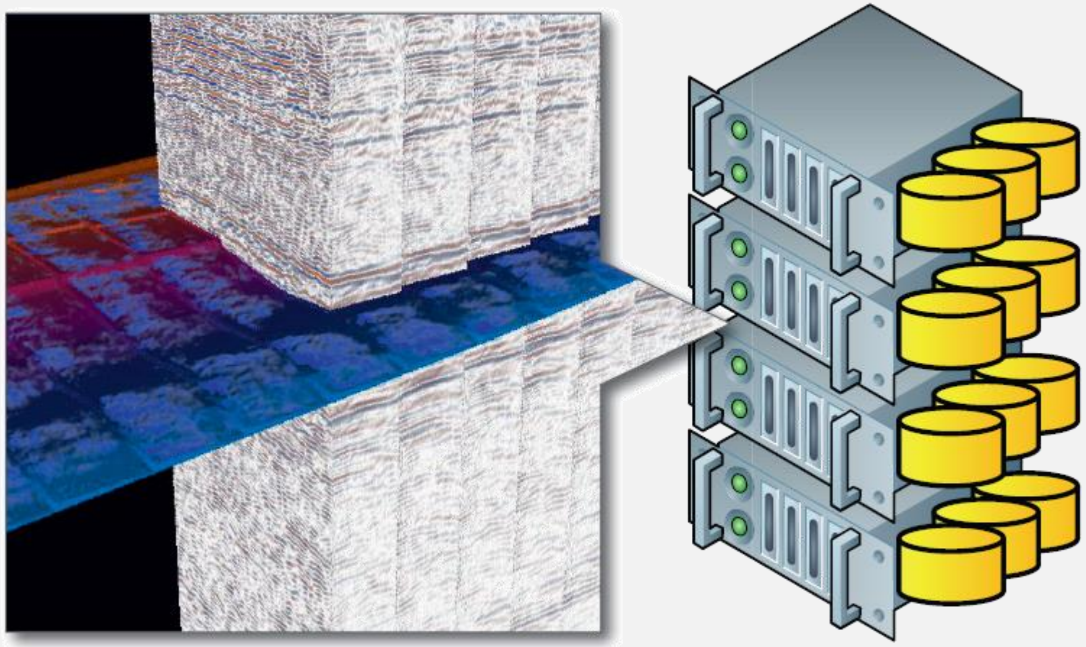
MetaData

BeeOND – BeeGFS On Demand

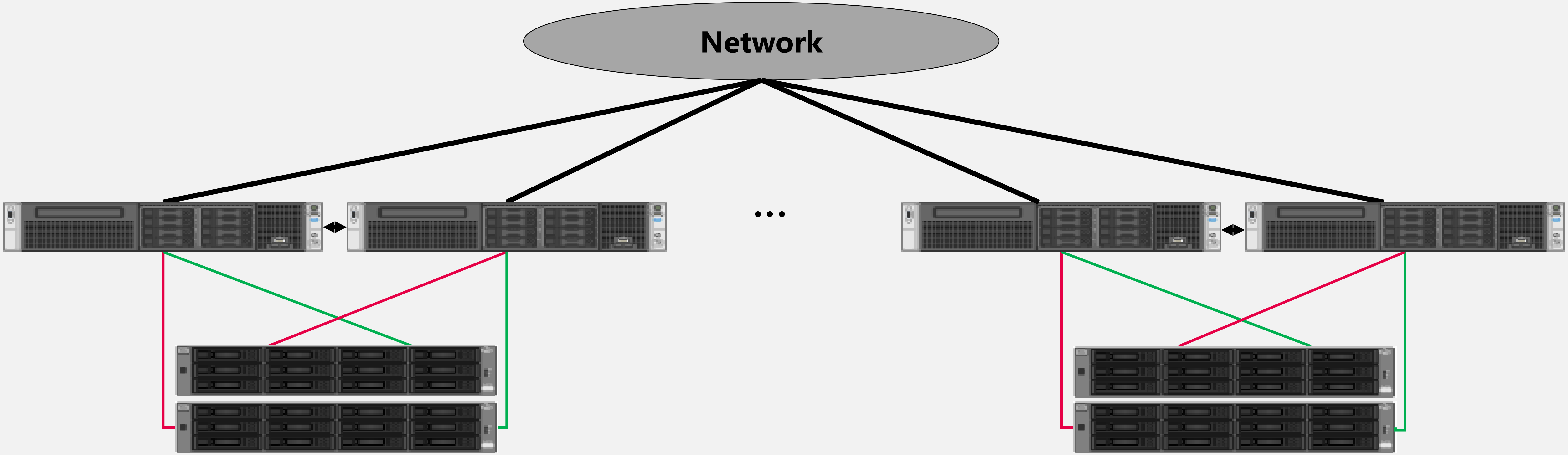
- Create a parallel file system instance with one command
- Use cases: cloud computing, test systems, cluster compute nodes, …..
- Can be integrated in cluster batch system (e.g. PBS)
  - Suitable for a private „per-job parallel file system“
- Used in Fraunhofer Seislab
  - Take load from global storage
  - Speed-up certain I/O patterns

BeeOND – Use in Fraunhofer Seislab

- Fraunhofer Seislab
    - In-house cluster of CC-HPC at Fraunhofer ITWM
    - 92 compute nodes with 1 TB of SSDs each
    - Global BeeGFS storage on 3,5" SATA drives
- Create BeeOND on SSDs on job startup
- Stage-in input data, work on BeeOND, stage-out results



High Availability - Shared Storage



High Availability - Shared Storage

➕

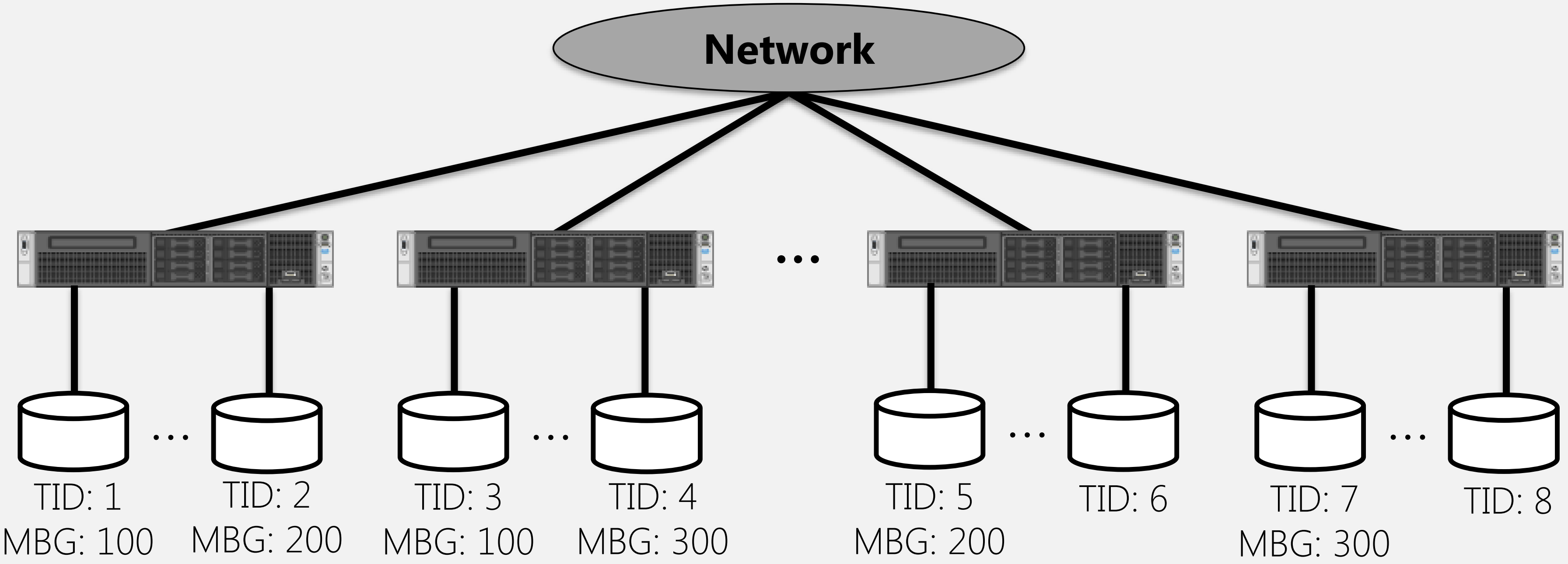
- No system downtime in case of server failure
- No additional storage capacity needed

➖

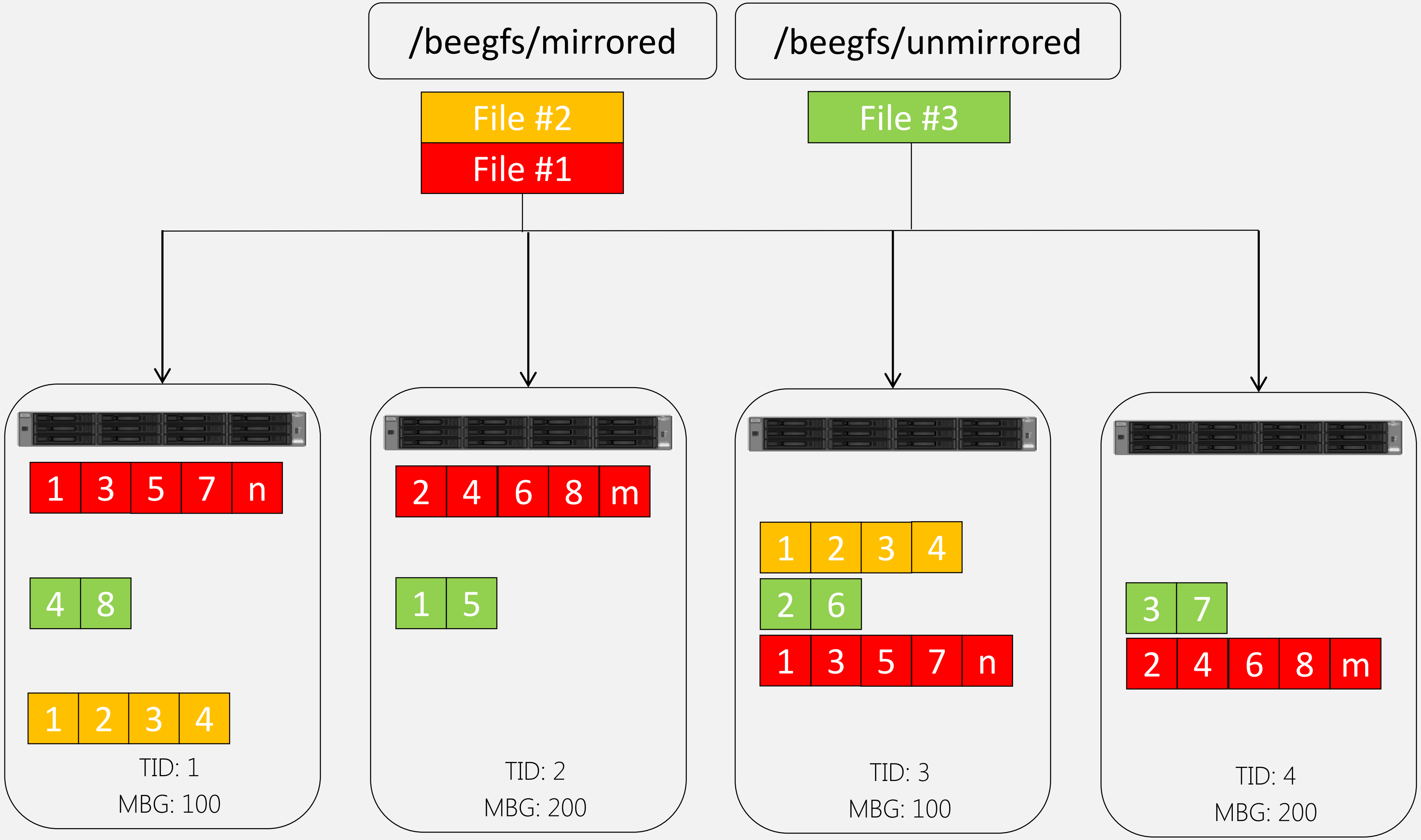
- Expensive storage components needed
- 3rd party software components needed
- Complex to set up and maintain
- Failover Risk
- No increased data safety

High Availability – Built-in Replication

- Assign targets to „mirror buddy groups“
- MBGs replicate chunks (but can also store non-replicated data)
- Internal HA/failover mechanisms



Built-in Replication - Striping



High Availability – Built-in Replication

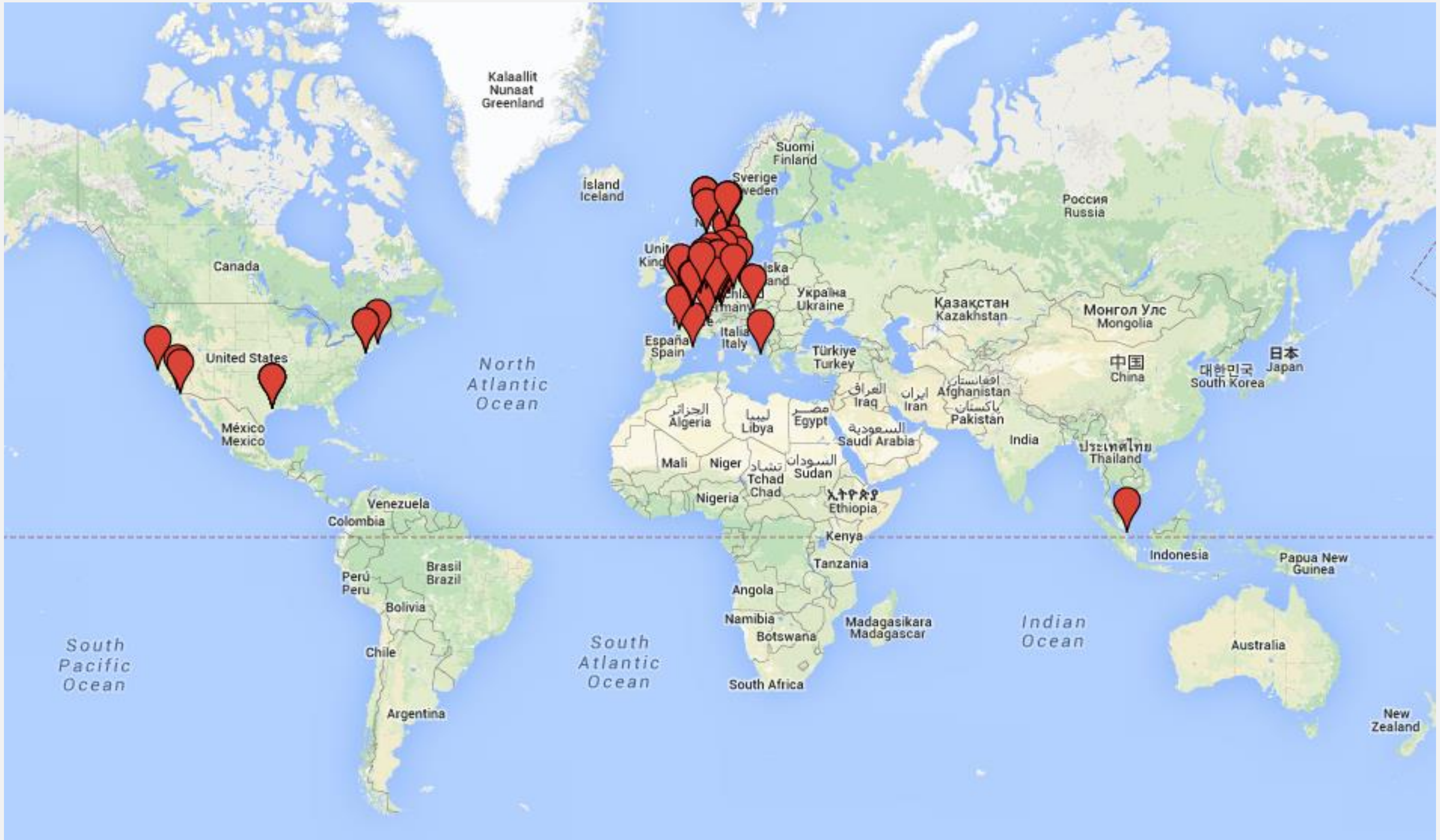
➕

- Flexible (replication configurable per-directory)
- Easy to scale/extend
- No 3rd party tools for monitoring and failover functionality
- Any storage backend can be used
- Additional data safety

➖

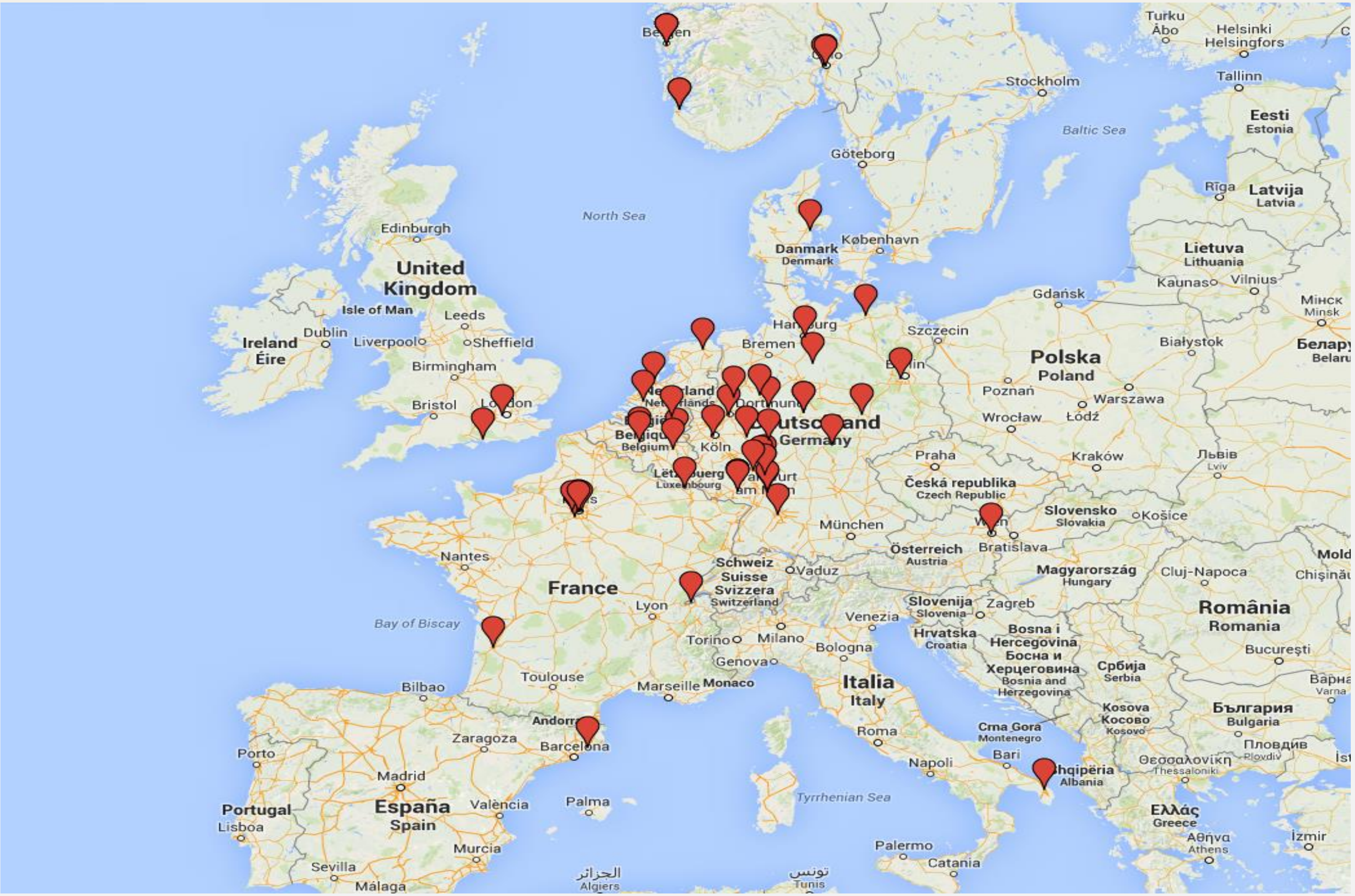
- Overhead in storage capacity
- Write penalty for replicated data

More than 100 happy customers



www.beegfs.com

More than 100 happy customers



Customer Examples

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

UNIVERSITAS OSLOENSIS MDCCCXI

l'Observatoire de Paris

Fraunhofer EMI

GL Germanischer Lloyd

TU WIEN

Heidelberger Institut für Theoretische Studien | HITS

DET NORSKE

DNV

DTU

iit istituto italiano di tecnologia

university of groningen

UNIVERSITÉ BORDEAUX 1 Sciences Technologies

EMBL

MAX-PLANCK-GESELLSCHAFT

SeismicCity

BNP PARIBAS

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

GWDG

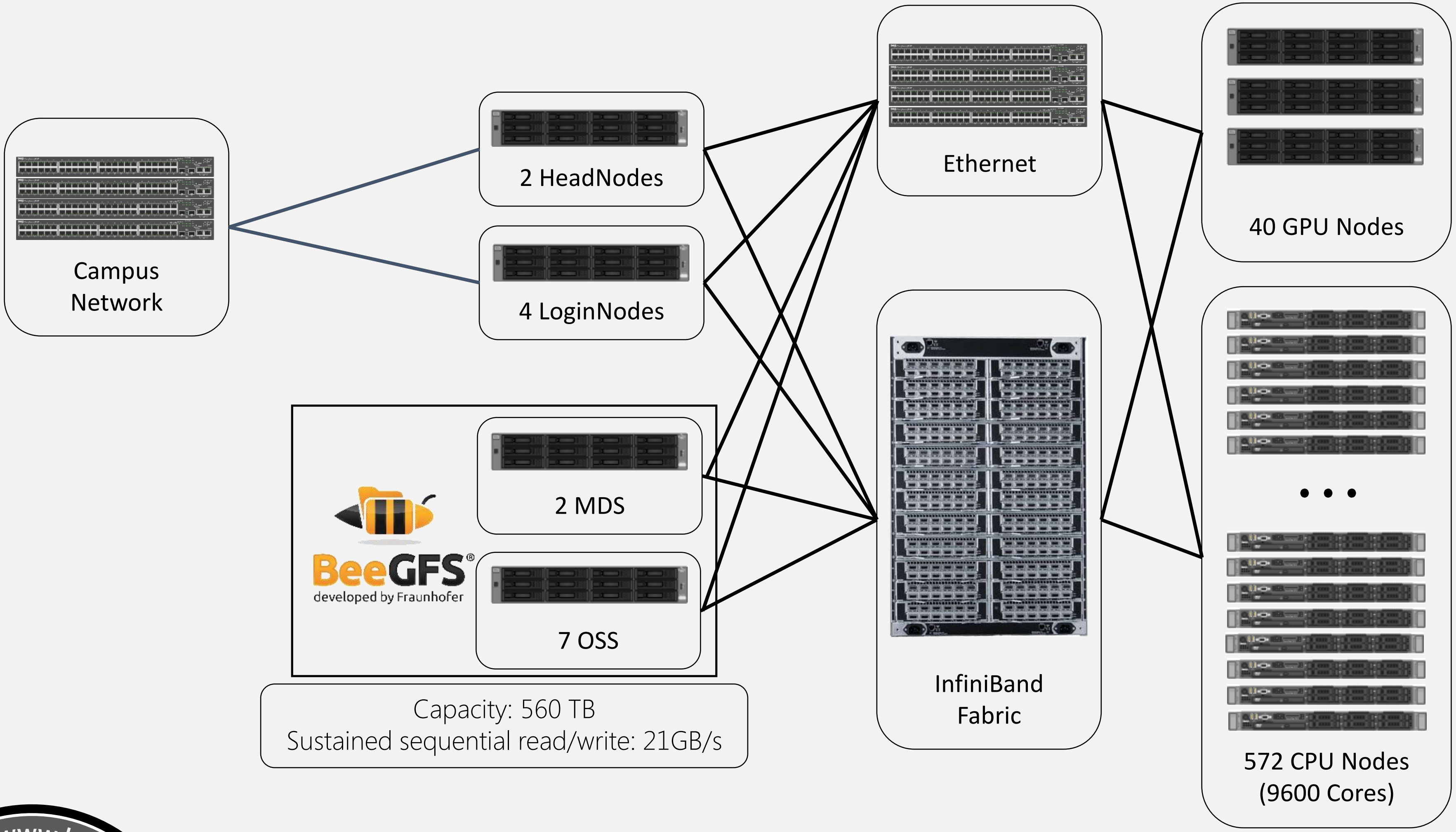
BROAD INSTITUTE

TECHNISCHE UNIVERSITÄT ILMENAU

BiRC Bioinformatics Research Center Aarhus

MIT Dept of Chemical Engineering

OcULUS – A Typical HPC Installation

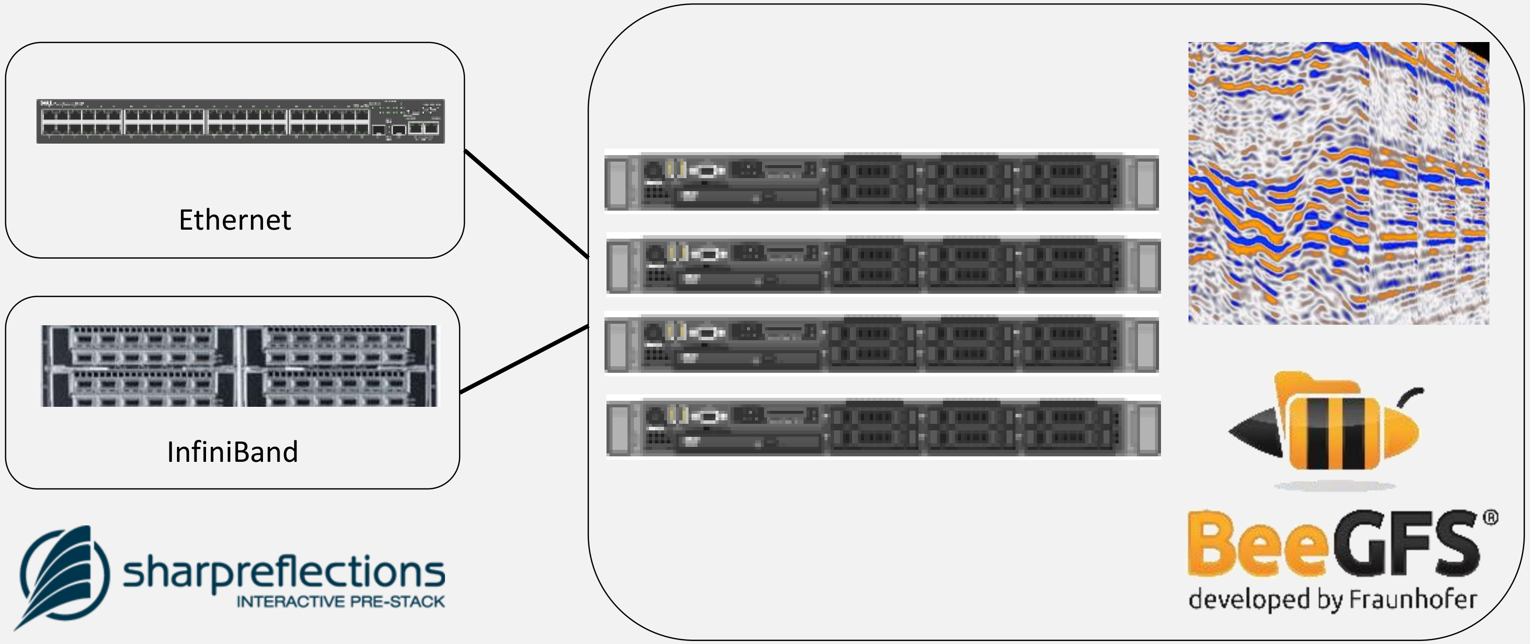


Capacity: 560 TB
Sustained sequential read/write: 21GB/s

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

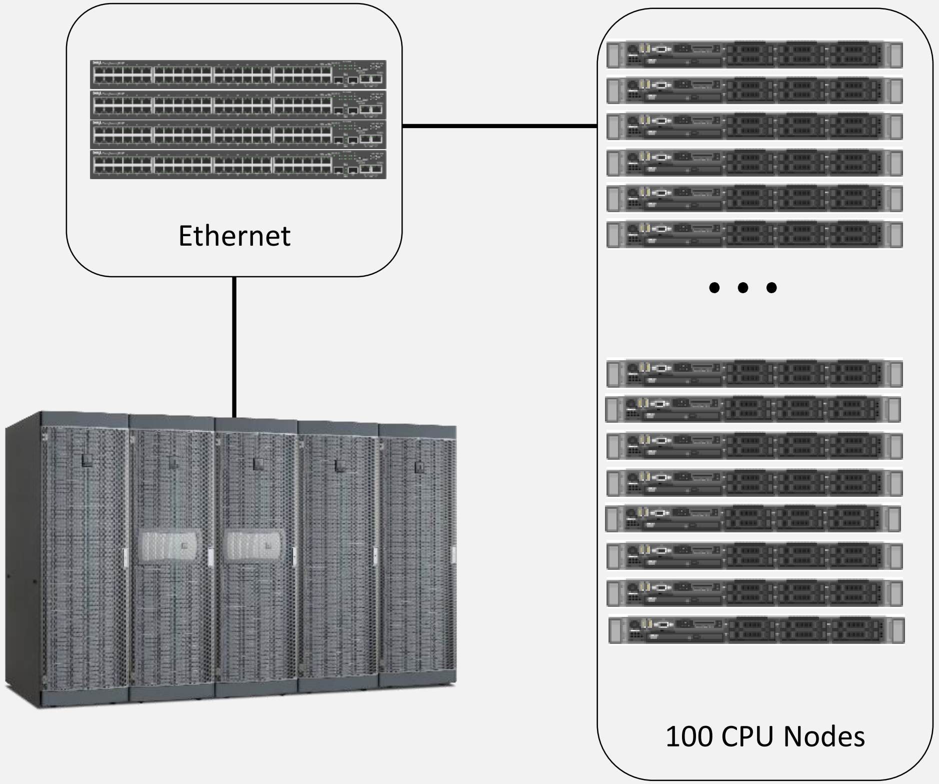
Sharp Reflections – A Lightweight Solution

- Four compute nodes for seismic data interpretation
- 12 3.5" SATA drives per compute node
- BeeGFS running on compute nodes



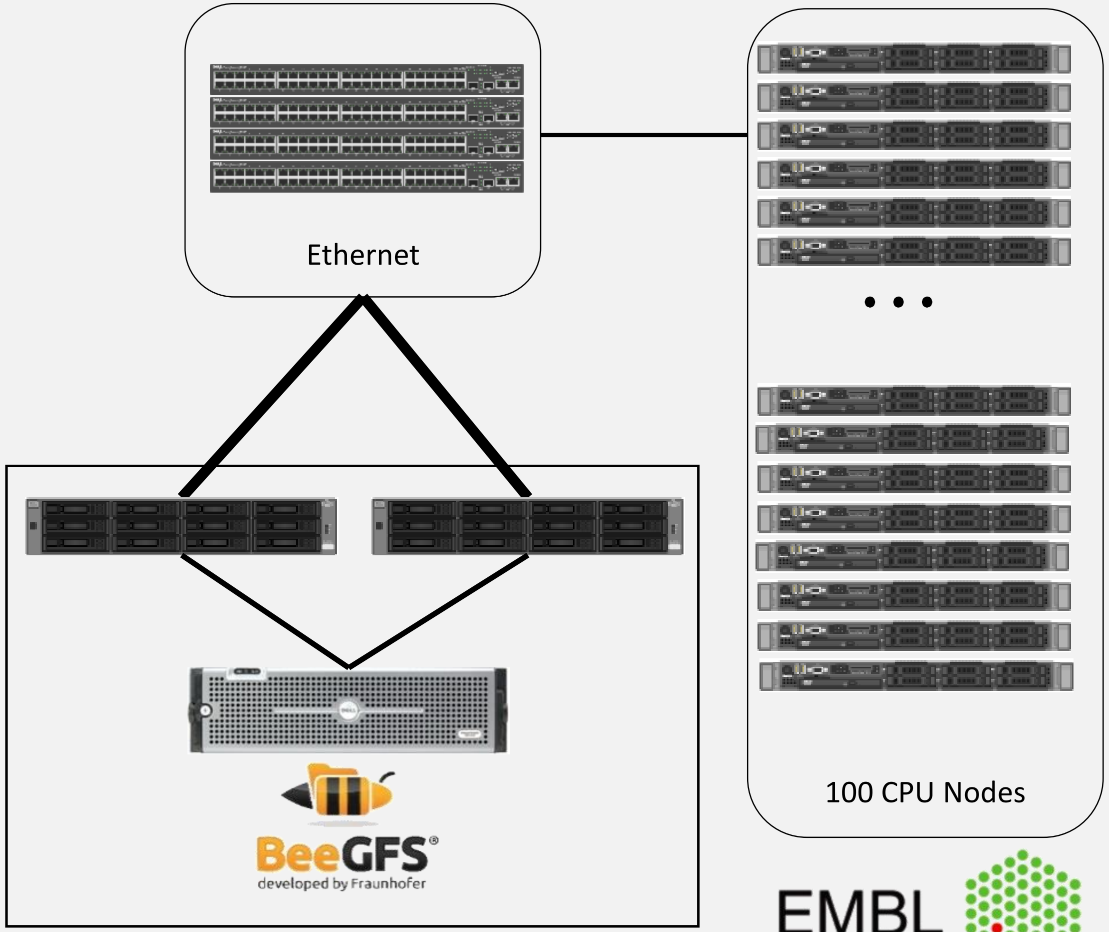
EMBL – A Workload Optimized Solution

- Situation
  - 100 compute nodes (connected with GigE)
  - „Traditional“ NAS storage
  - Only single core compute jobs (life science)
  - Jobs working on the same data
  - Random data access
  - 100k+ IOPS needed
  - ~ 40 jobs saturate NAS



EMBL – A Workload Optimized Solution

- BeeGFS solution
  - 2 storage servers
  - 30 disk drives each
  - 0.5 TB RAM each
  - 40GbE uplinks for storage
  - Up to 600 jobs at peak performance



Questions?

• Wiki	wiki.beegfs.com
• Twitter	www.twitter.com/BeeGFS
• MailingList	fhgfs-user@googlegroups.com
• NewsList	beegfs-news@googlegroups.com
• Mail	sales@thinkparq.com
	support@beegfs.com