BeeGFS

Introduction

HPCwire
2022
READERS' CHOICE AWARDS
Best HPC Storage Product or Technology
BeeGFS
Awards Winner for 5 Years

BeeGFS.io

2022

Agenda

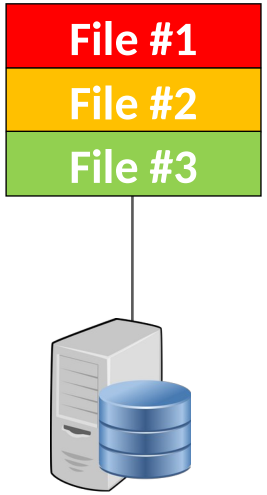
- **About Us**
  - Fraunhofer Center for HPC
  - ThinkParQ
- **History**
  - How it all started
  - Main motivation
- **Basic Concepts**
  - Key aspects
  - Main characteristics
  - Architecture

Why the bee?

Why the bee?

Introduction

- **Local File System**
  - Widely used, easy to use, simple
  - Examples: XFS, ext4, ZFS
  - Files stored on a local storage devices
  - Limited storage capacity
  - Limited performance
  - IO operations processed by a single machine



File #1

File #2

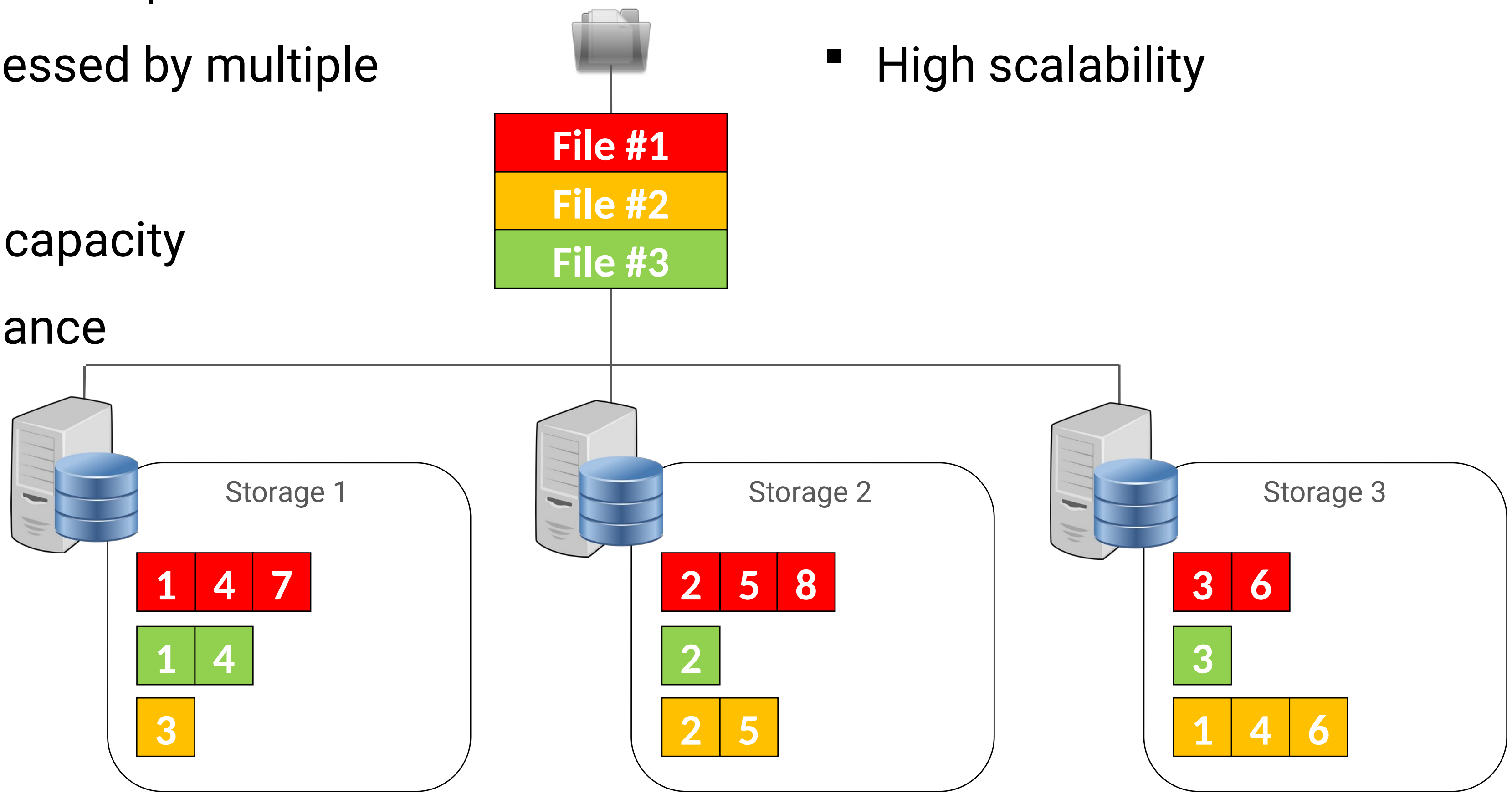
File #3

  - Limited fault tolerance
  - Limited scalability

Introduction

Parallel File System

- Data striped across multiple servers
- IO operations processed by multiple servers
- Increased storage capacity
- Increased performance
- Increased fault tolerance
- High scalability



File #1
File #2
File #3

Storage 1
1 4 7
1 4
3

Storage 2
2 5 8
2
2 5

Storage 3
3 6
3
1 4 6

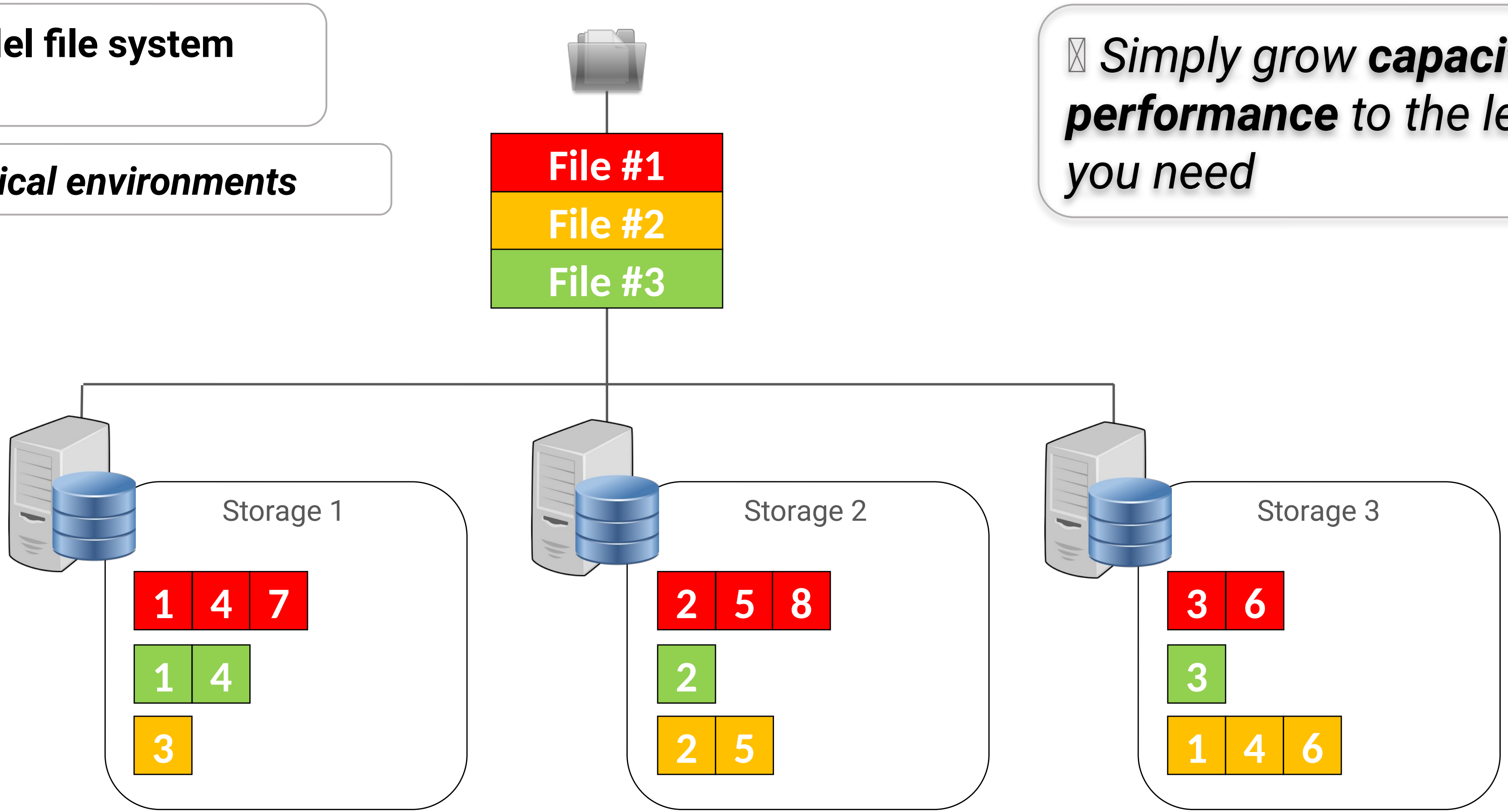
Introduction

BeeGFS is...

...a *hardware-independent* parallel file system (aka Software-defined Storage)

...designed for *performance-critical environments*

▯ *Simply grow **capacity** and **performance** to the level that you need*



File #1

File #2

File #3

Storage 1

1 4 7

1 4

3

Storage 2

2 5 8

2

2 5

Storage 3

3 6

3

1 4 6

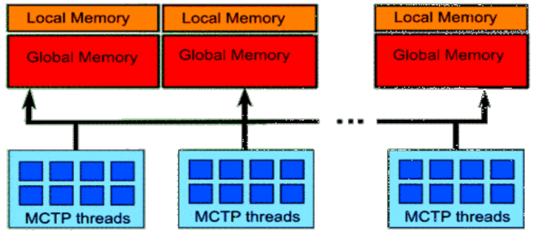
About Us

- BeeGFS was originally developed at the Fraunhofer Center for HPC

- The Fraunhofer Gesellschaft (FhG)
  - Largest organization for applied research in Europe
  - Special base funding by German government
  - Institutes, research units and offices around the globe
  - Staff: ~24000 employees

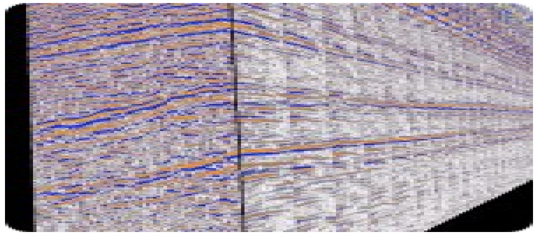
Fraunhofer

The Fraunhofer Center for HPC



Parallel Programming Models & Tools

Photo Realistic Real Time Ray Tracing



Interactive Seismic Imaging

Parallel File System

Big Data

Smart Energy / Green by IT

About Us

- ThinkParQ
  - A Fraunhofer spin-off
  - Based in Kaiserslautern (right next to Fraunhofer HPC Center)
  - Founded in 2014 specifically for BeeGFS
  - Consulting, professional services & support for BeeGFS
  - Cooperative development together with Fraunhofer
  - First point of contact for BeeGFS

© BIC

Business Model

- BeeGFS is free to use for end users
  - Ready-to-use binary packages
  - Complete source code also available (but: BeeGFS is intentionally not a community project)
  - BeeGFS is not open source under the GPL license, except the client module

- System integrators/partners for turn-key solutions
  - System setup and tuning
  - 1st- and 2nd-level support
  - Partners make back2back contract with ThinkParQ for 3rd-level support

Business Model

- Professional 3rd-level support contract
  - Allows to use the enterprise edition features
  - Allows to open tickets at support@thinkparq.com
  - Pricing based on number of servers and timeframe (e.g. 3 or 5 years)

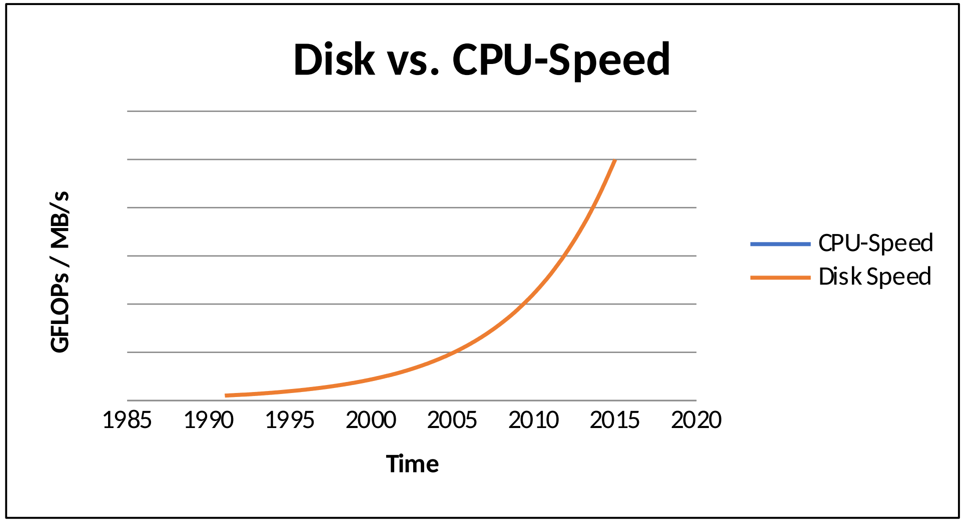
□ Support contracts are also the financial basis for development of great new features

History

- Development started in 2005 (old name: FhGFS, aka “Fraunhofer File System”)

- Why?

> *“A supercomputer is a device for turning compute-bound problems into I/O-bound problems.“*
> *- Ken Batcher*



- So we evaluated existing solutions, but…

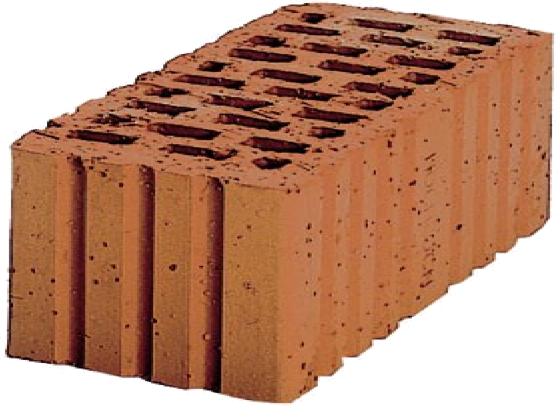
Existing solutions seemed like this…

History

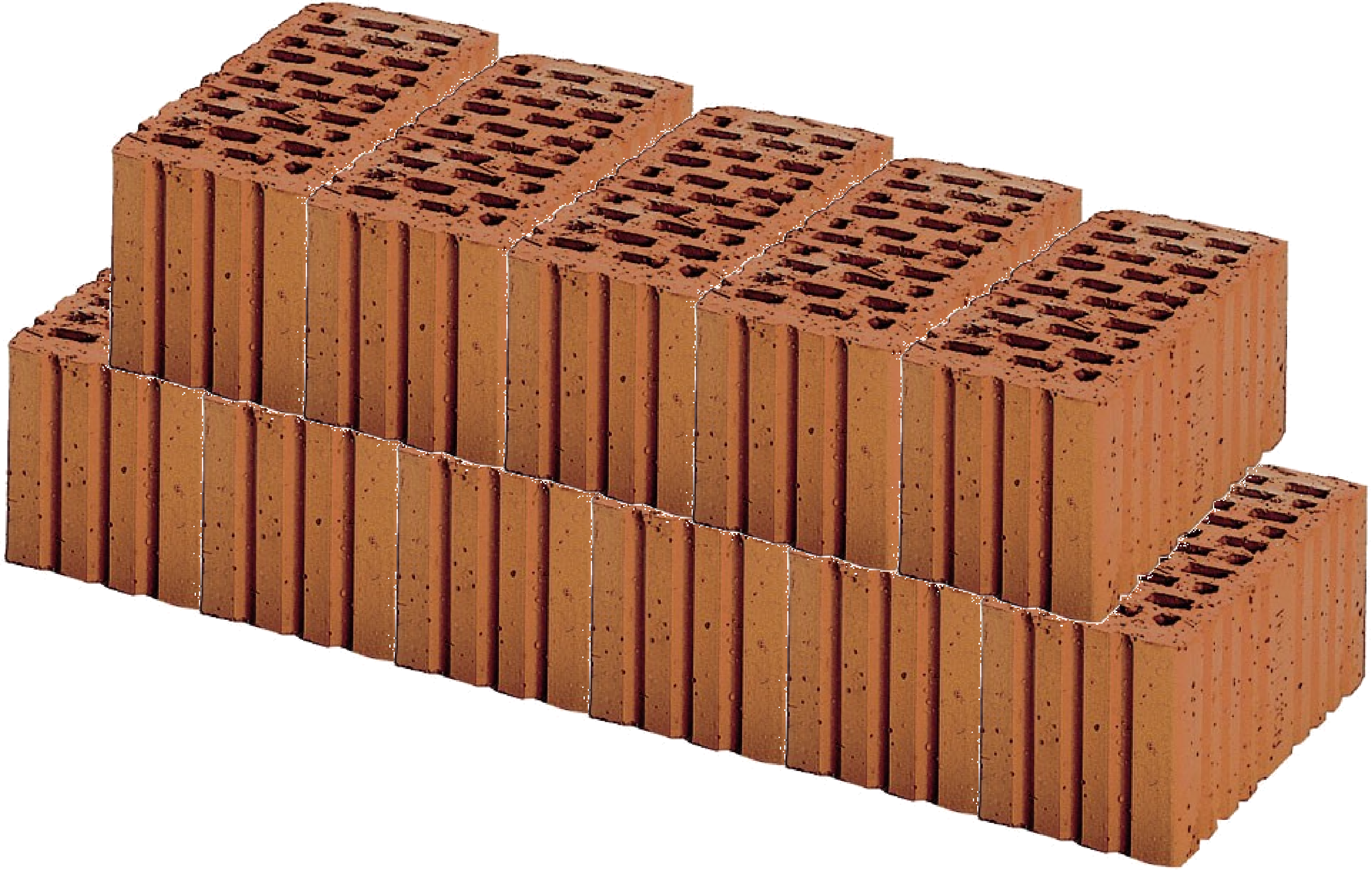
- Evaluated existing solutions, not happy with what we found:
  - Very complex and inflexible
  - Required dedicated staff for continuous maintenance
  - Expensive
  - Scalability and performance problems for metadata access, shared file writes, single-stream I/O, …

- We‘re a HPC center, so a lot of knowledge and users in-house

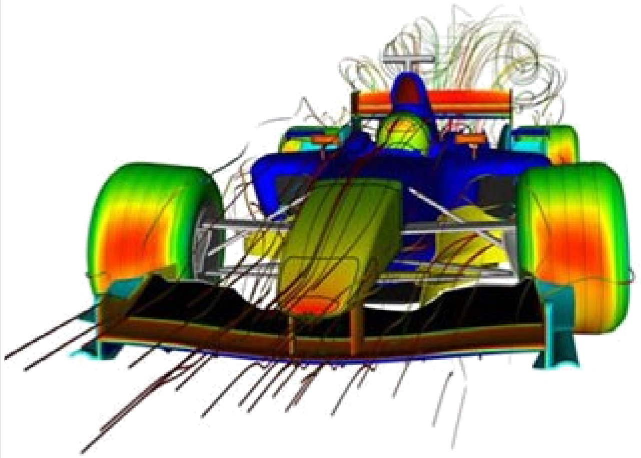
This is how we want it to be…



This is the flexibility that we want…



Key Aspects



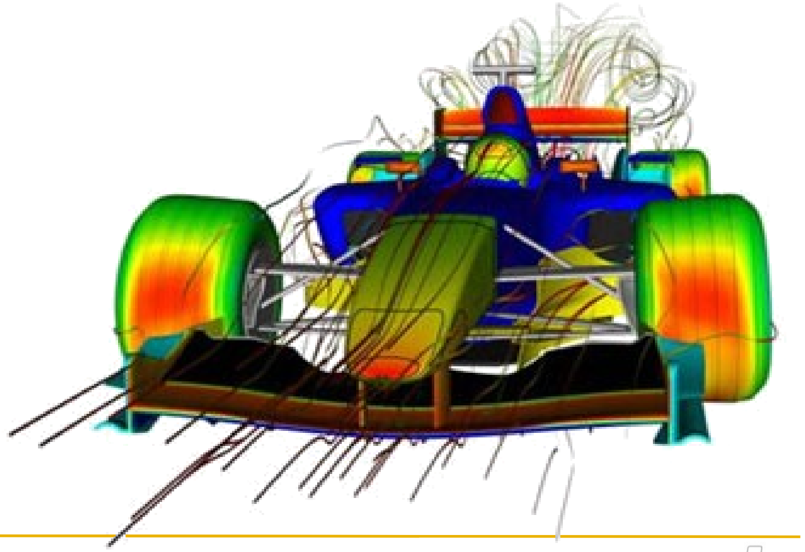
Maximum Performance & Scalability

High Flexibility

Robustness & Easy to use

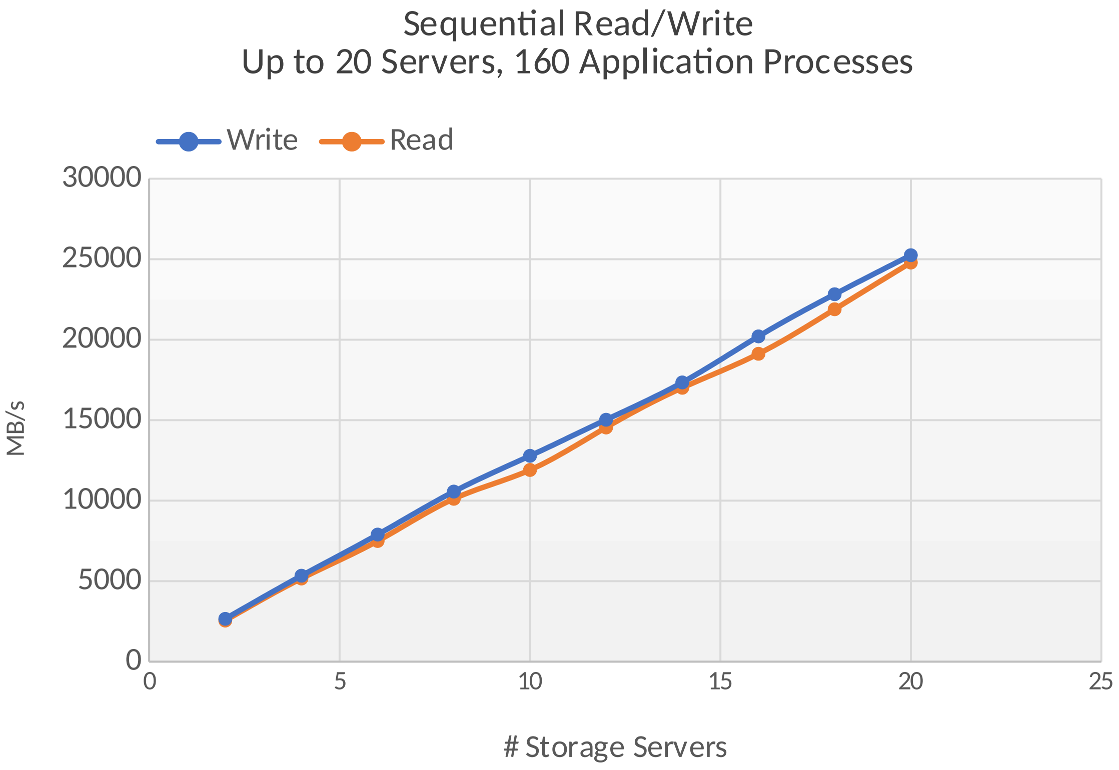
Key Aspects

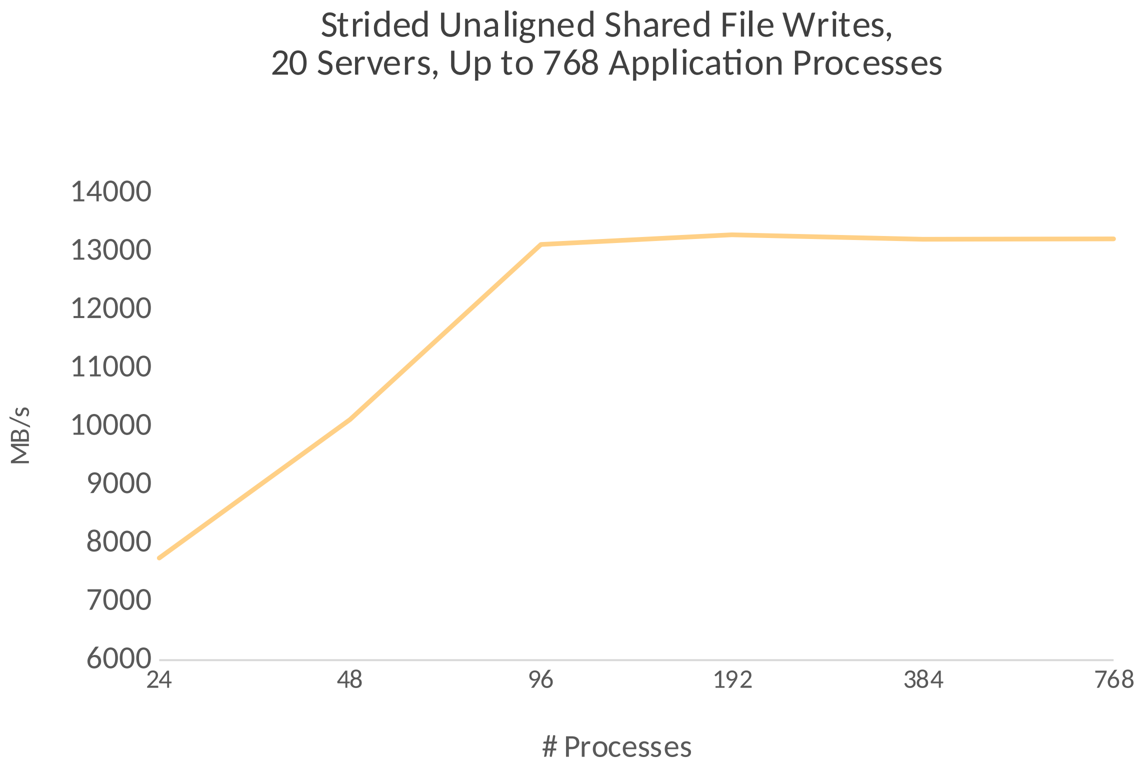
- Performance & Scalability
  - Initially optimized for performance-critical workloads
  - Efficiently multi-threaded and light-weight design
    - “Not even breaking a sweat: BeeGFS at 10GB/s on single node all-flash unit over 100Gbit network” -ScalableInformatics
  - Supports RDMA/RoCE and TCP (InfiniBand, Omni-Path, 100/40/10/1GbE, …)
  - Aggregated IOPS and throughput of multiple servers
  - Distributed file contents & distributed metadata
  - High single stream performance
    - 9 GB/s single-stream throughput with Mellanox EDR
      (Few file streams completely saturate a 100 Gbit link.)



Key Aspects

- Performance & Scalability

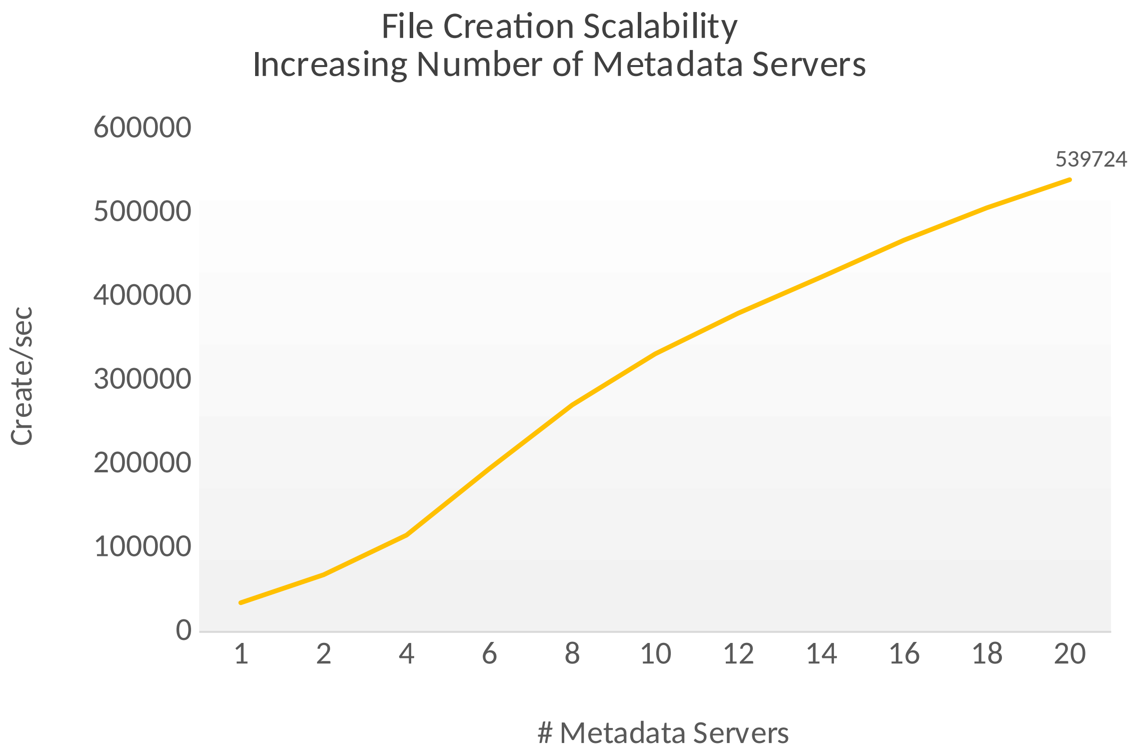


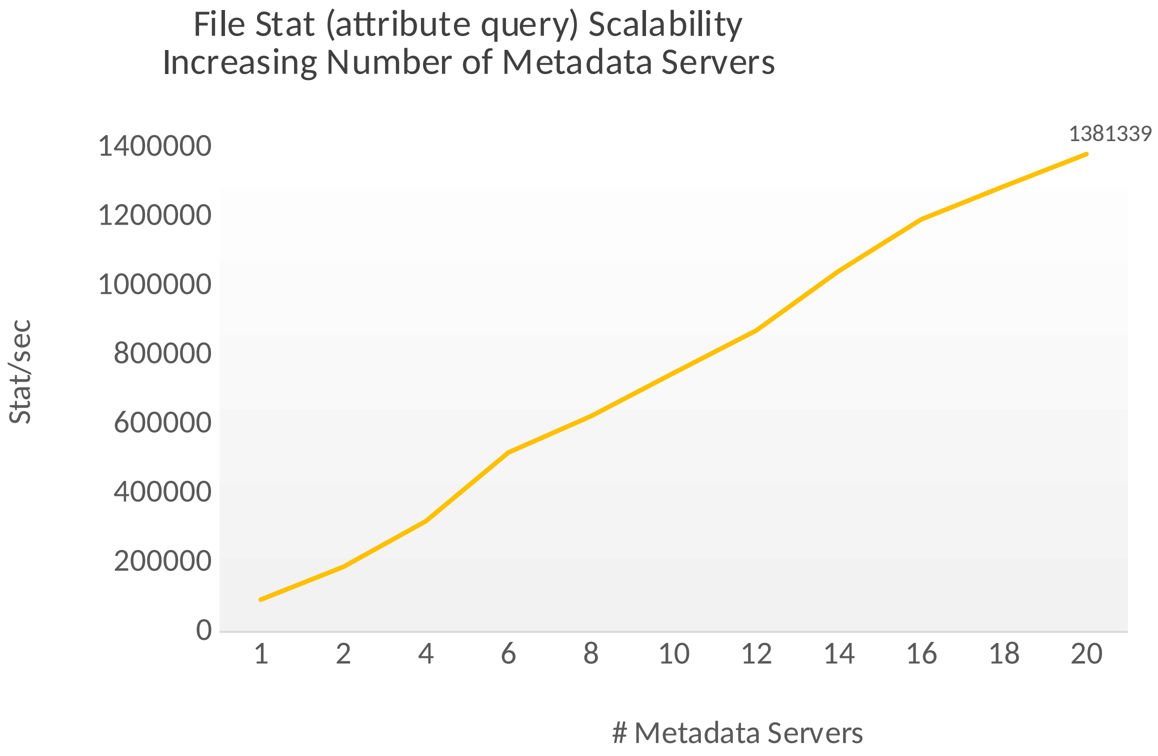


Note: Absolute values in these cases depend on per-server hardware performance, of course.

Key Aspects

Performance & Scalability





Note: Absolute values in these cases depend on per-server hardware performance, of course.

Key Aspects

BeeGFS®

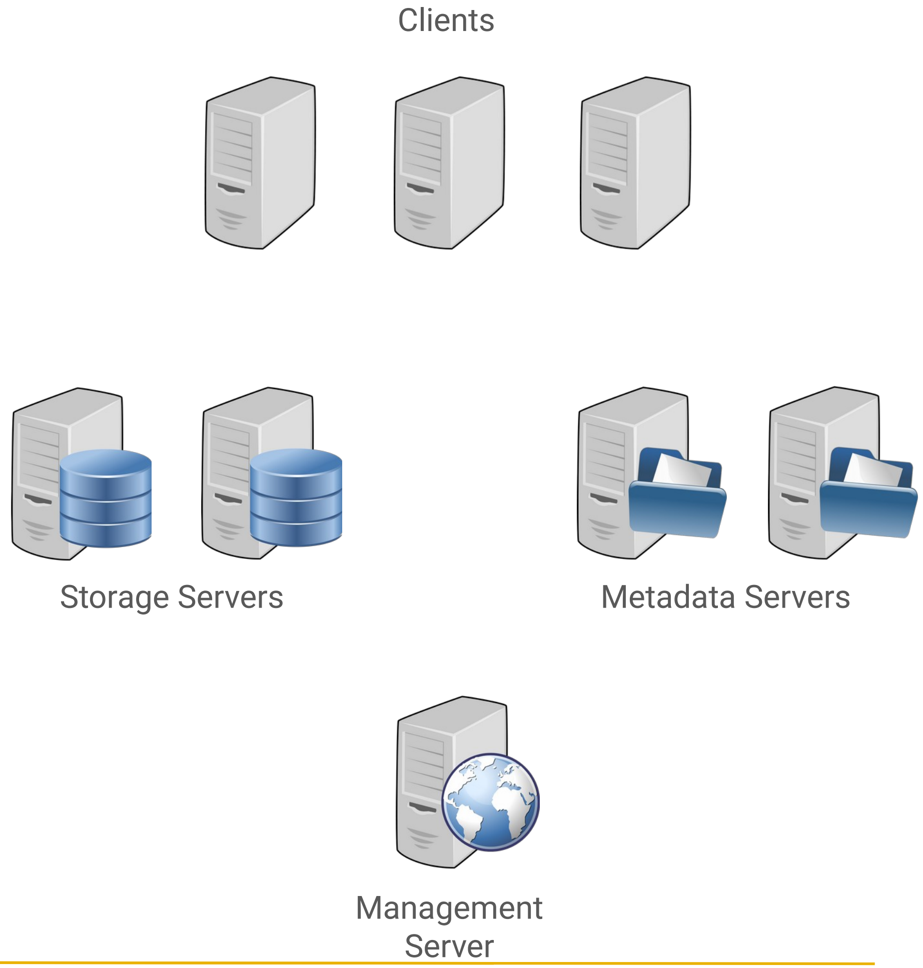
- Flexibility
  - Runs on different architectures, e.g.: intel ARM OpenPOWER™ Sunway
  - No special hardware requirements
  - Packages for several Linux distributions and kernels: debian ubuntu SUSE redhat LINUX
  - Multiple BeeGFS services (any combination) can run together on the same machine
  - NFS & Samba re-export possible
  - Flexible data striping per-file / per-directory
  - Add servers or storage devices at runtime
  - Installation & updates without even rebooting

Key Aspects

- Robust & Easy to use
  - Very intensive suite of release stress tests, in-house production use before public release
    - The move from a 256 nodes system to a 1000 nodes system did not result in a single hitch, similar for the move to a 2000 nodes system.
  - No kernel patches
    - Updates of system packages, kernel and BeeGFS are trivially simple
  - Servers run on top of standard local file systems (ext4, XFS, ZFS, …)
  - Graphical tools
  - Comprehensive documentation (online, built-in)

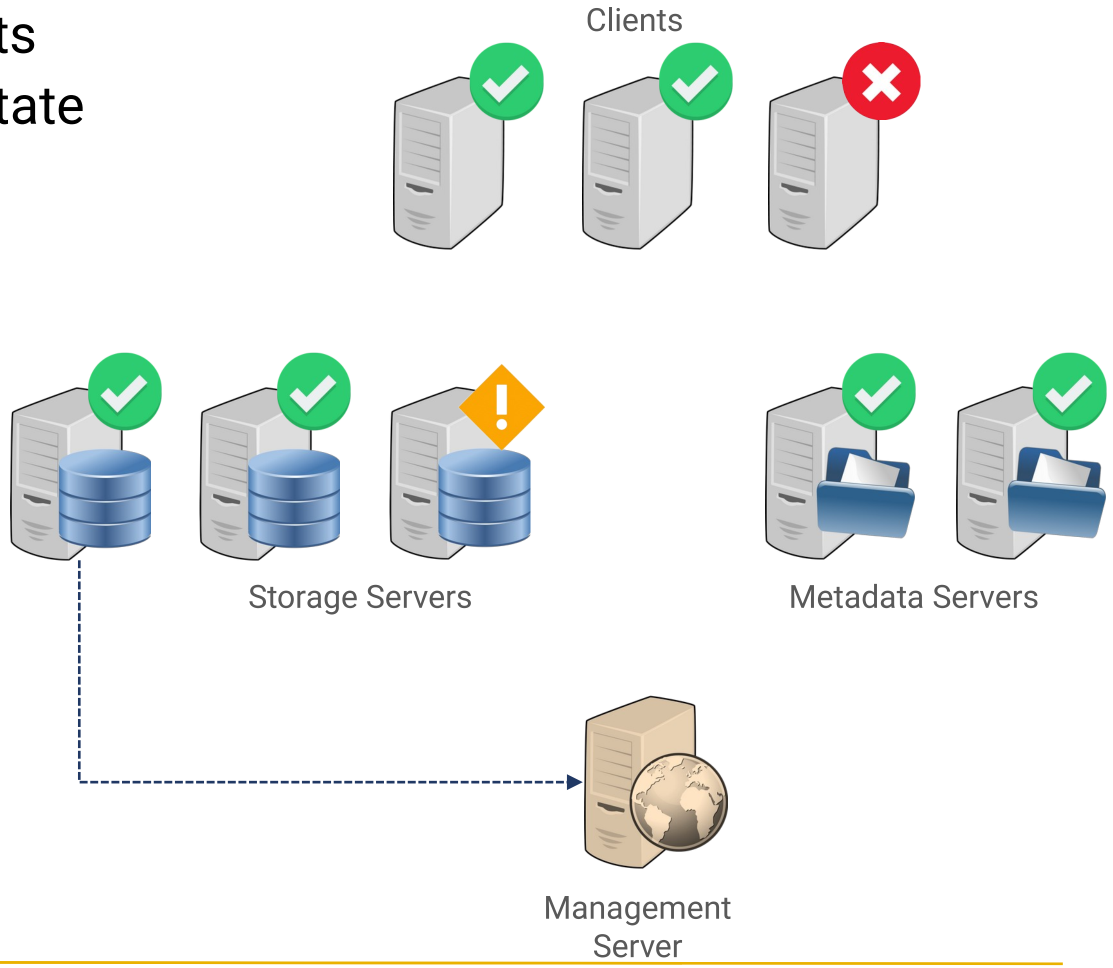
BeeGFS Architecture

- Main Services
  - Management service
  - Storage service
  - Metadata service
  - Client service



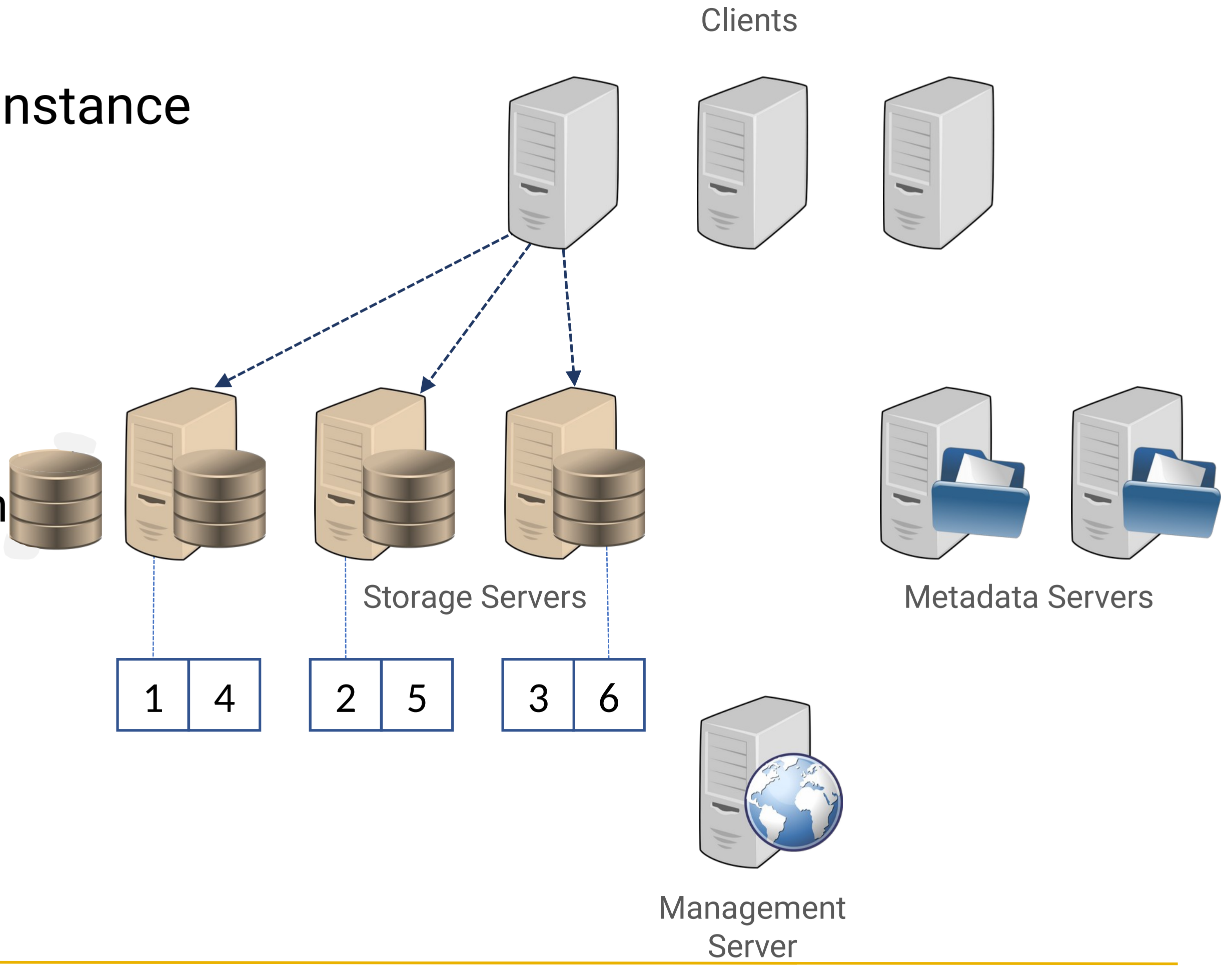
BeeGFS Architecture

- Management Service
  - Rendevouz point for (new) servers and (new) clients
  - Watches registered components and check their state
  - Not performance-critical, stores no user data



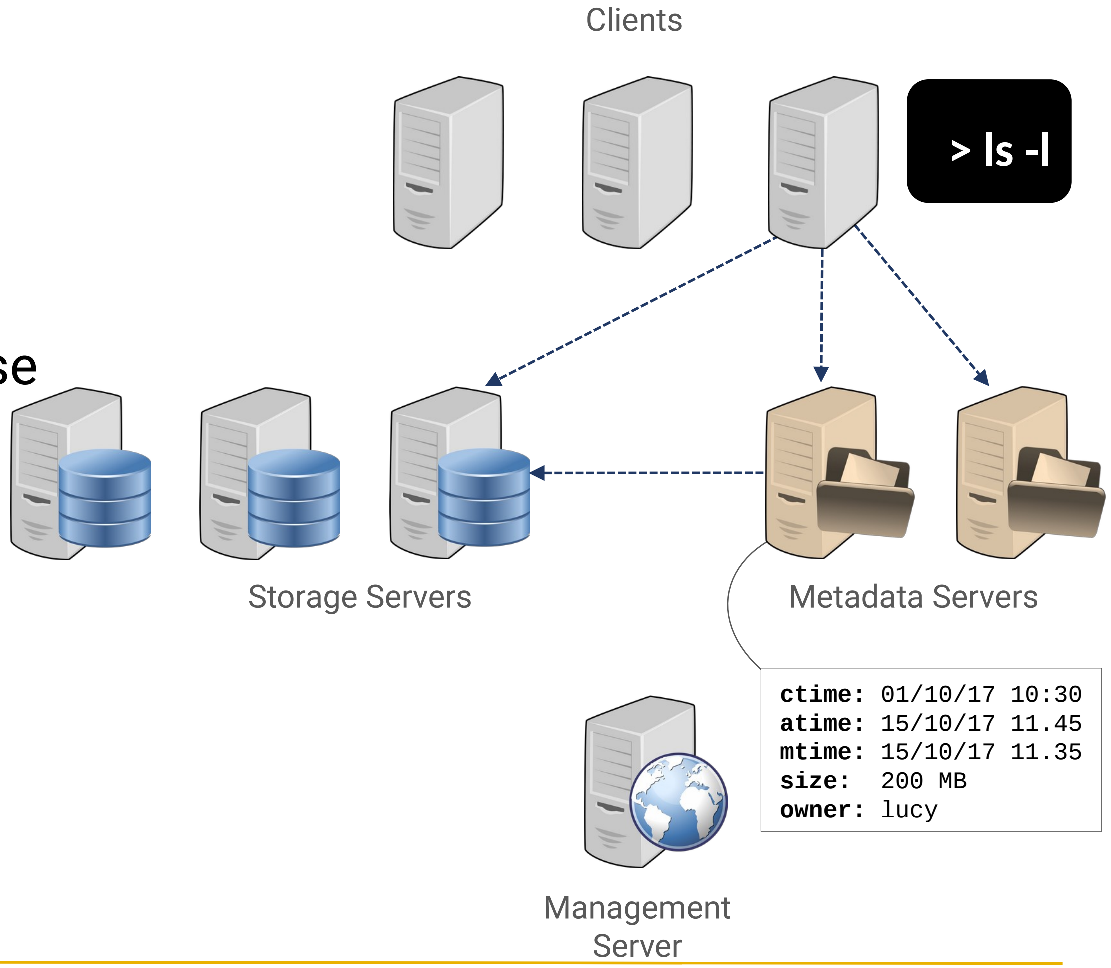
BeeGFS Architecture

- Storage Service
  - Stores user file contents (data chunk files)
  - One or multiple storage services per BeeGFS instance
  - Manages one or more storage devices
    - Typically a RAID volume
    - Internally or externally attached
    - It can also be a single HHD, NVMe, or SSD
    - Called storage targets
    - In general, any directory on an local file system



BeeGFS Architecture

- Metadata Service
  - Stores information about the data
    - Directory entries
    - File and directory ownership
    - File size, update time, creation time, etc
    - Location of user data files on storage targets
  - Not involved in data access between file open/close
  - Manages one metadata target
    - Typically an SSD or NVMe device
    - In general, any directory on local file system

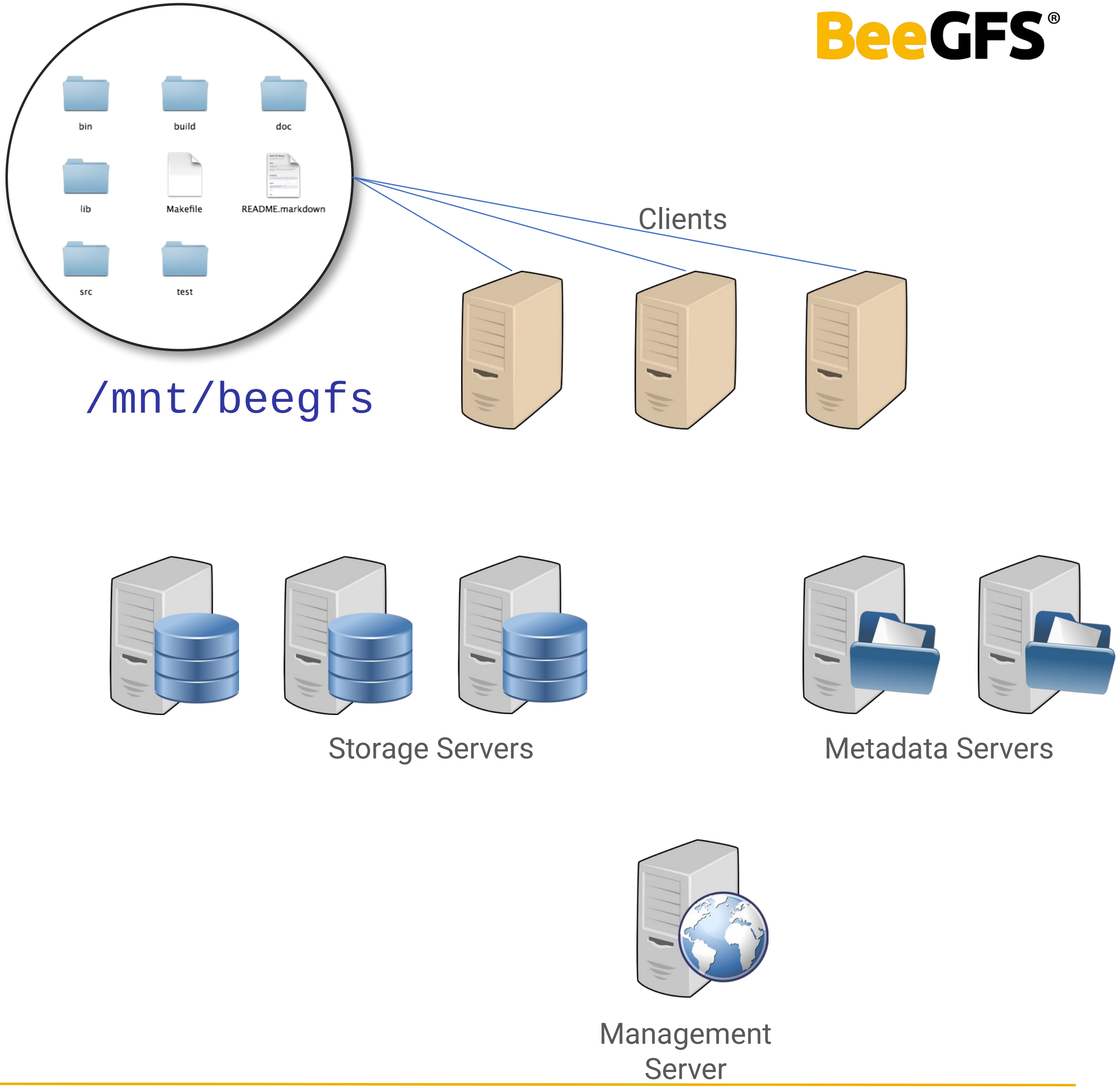


```
> ls -l
```

```
ctime: 01/10/17 10:30
atime: 15/10/17 11.45
mtime: 15/10/17 11.35
size:  200 MB
owner: lucy
```

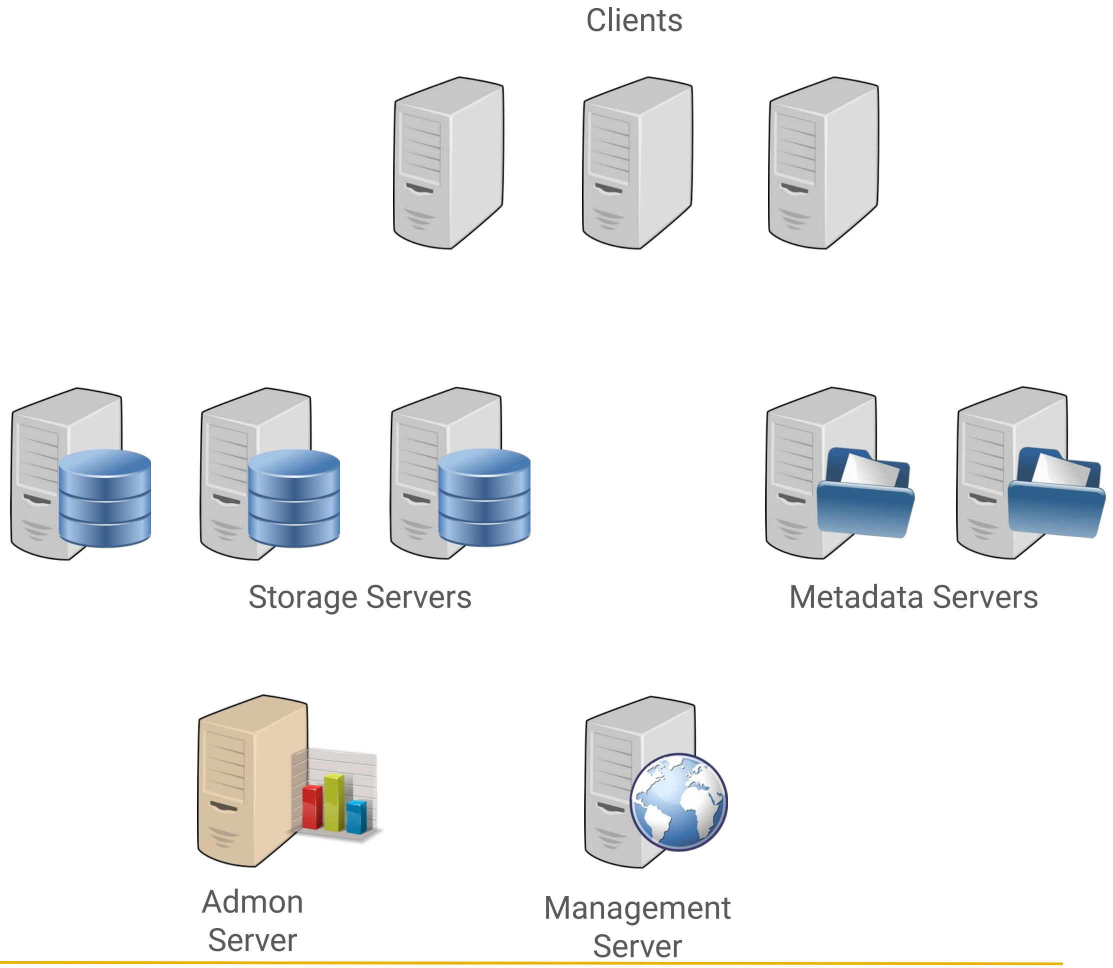
BeeGFS Architecture

- Client Service
  - Native Linux module
  - Mount the file system



BeeGFS Architecture

- Optional Service
  - Graphical Monitoring System
    - System information monitoring



What next?

✓

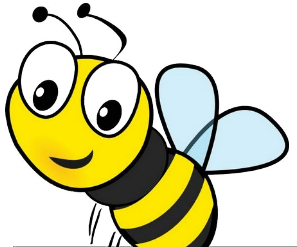
BeeGFS Training for System Engineers

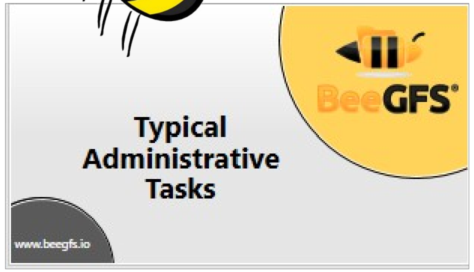
www.beegfs.io

✓

Introduction

www.beegfs.io





Typical Administrative Tasks

www.beegfs.io

BeeGFS®

Thank You

Follow BeeGFS: