

David Ramírez Alvarez

HPC INTEGRATOR MANAGER

WWW.SIE.ES

dramirez@sie.es

ADMINTECH 2016

BeeGFS

*Solid, fast and
made in Europe*

www.beegfs.com



Thanks to Sven for info!!!

February 2016 | Sven Breuner, CEO, ThinkParQ

What is BeeGFS?



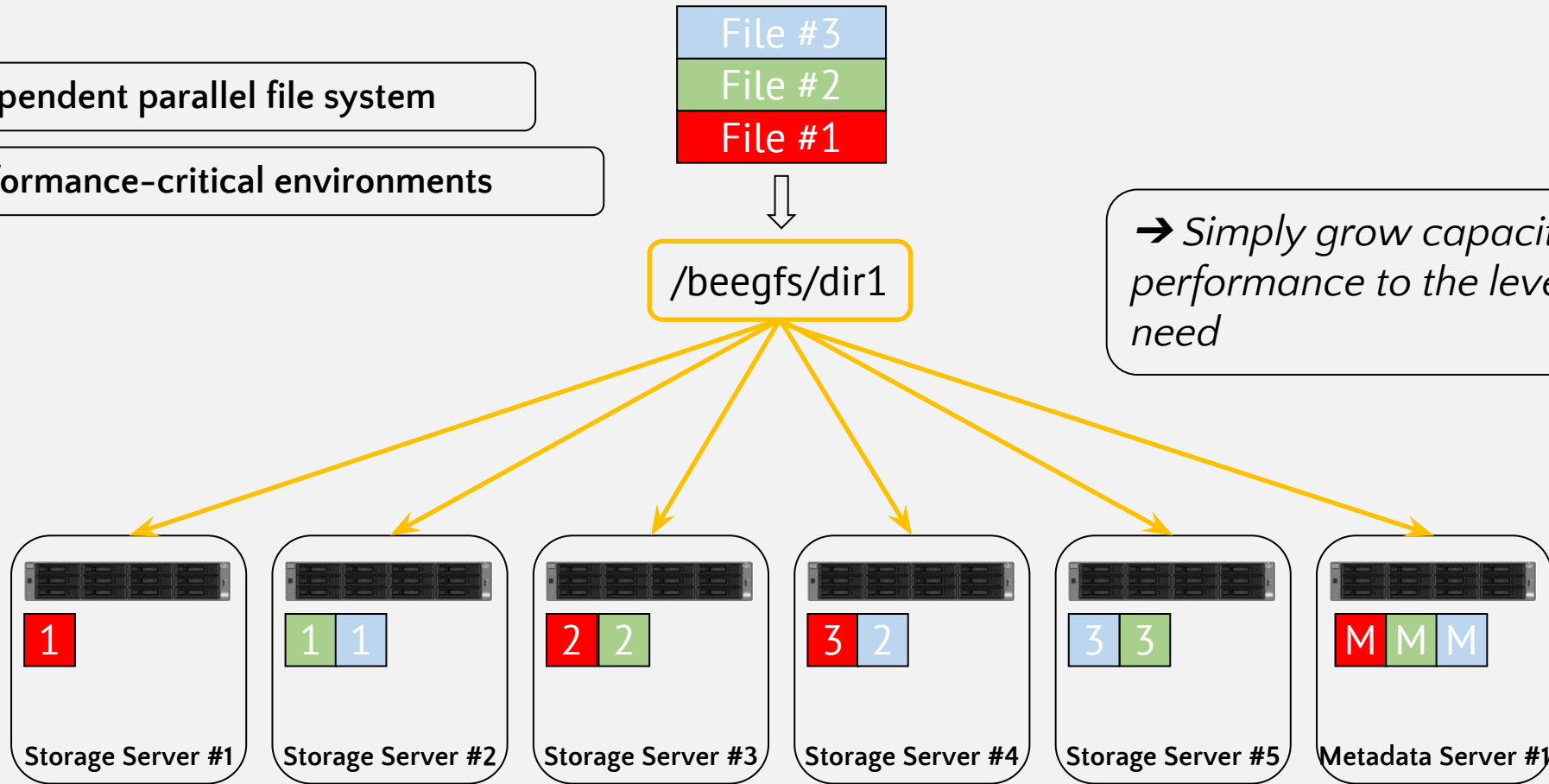
SIE



BeeGFS is...

...a hardware-independent parallel file system

...designed for performance-critical environments



→ Simply grow capacity and performance to the level that you need

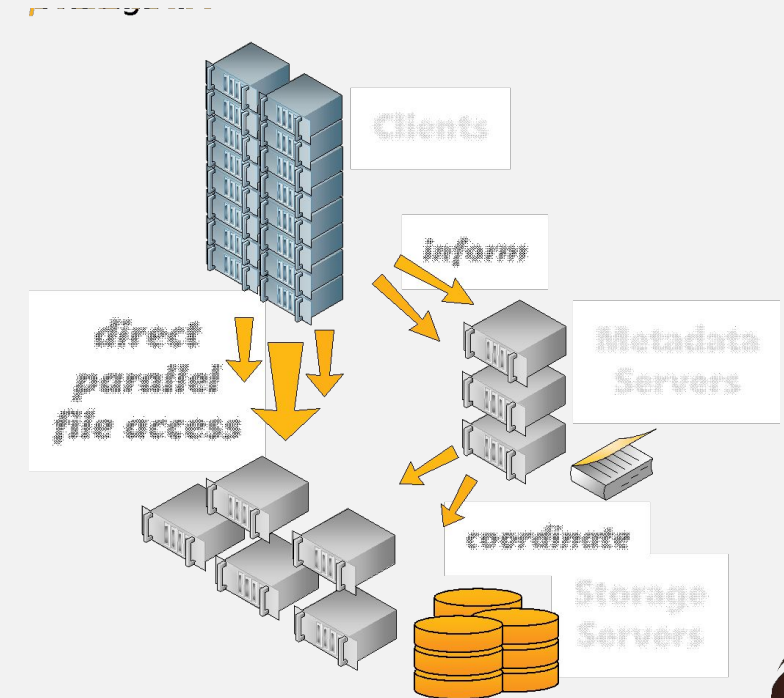
BeeGFS Architecture



S I E



- Client
 - Native Linux module to mount the file system
- Storage Service
 - Store the (distributed) file contents
- Metadata Service
 - Maintain striping information for files
 - Not involved in data access between file open/close
- Management Service
 - Glue everything together and watch services
- Graphical Administration and Monitoring System
 - GUI to perform administrative tasks and monitor system information
 - Can be used for “Windows-style installation”



Key Aspects



S I E



- **Performance & Scalability**

- Initially optimized for HPC
- Completely multi-threaded – lightweight design
- Supports RDMA/RoCE and TCP (Infiniband, 40/10/1GbE, ...)
- Distributed file contents:
aggregated throughput of multiple servers
- Distributed metadata across multiple servers
- High single stream performance (multiple GB/s)



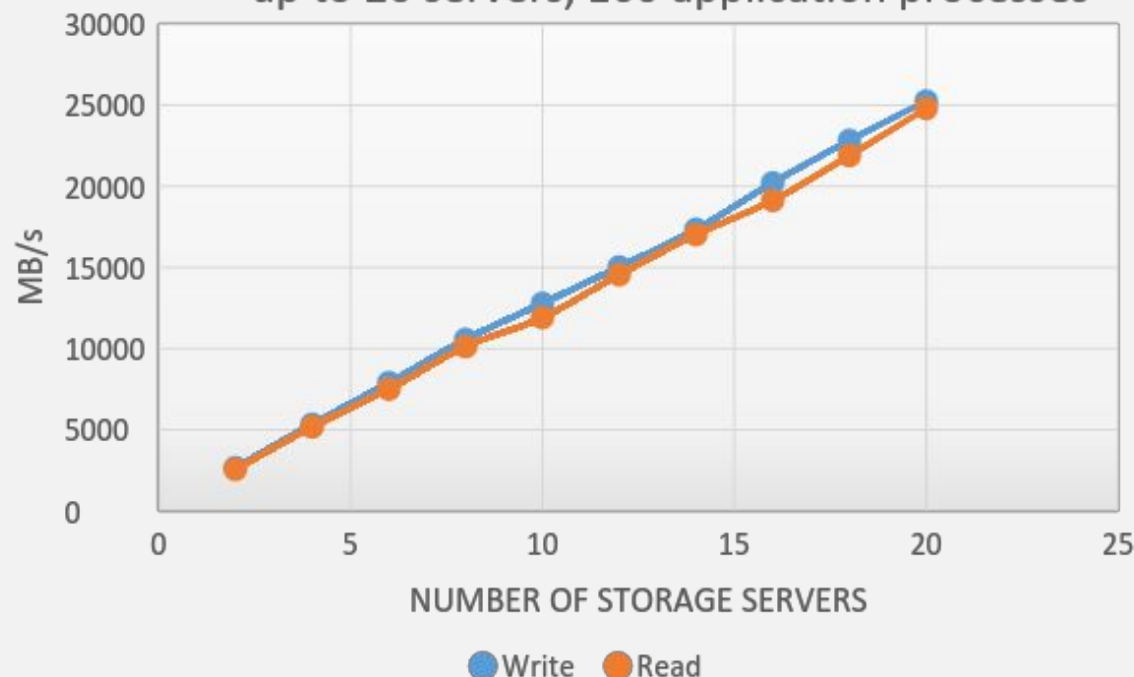
Throughput Scalability



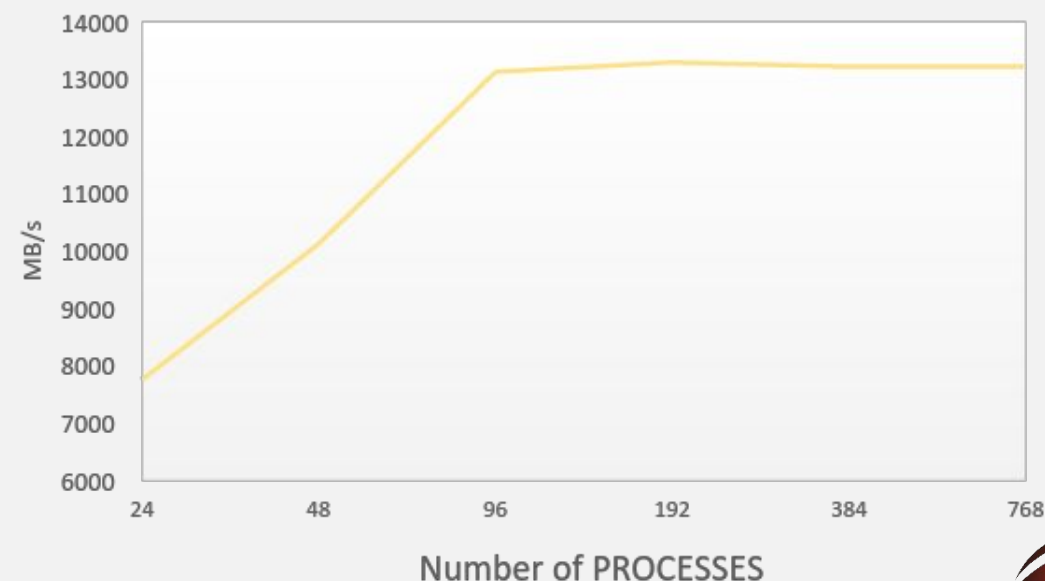
SIE



Sequential read/write
up to 20 servers, 160 application processes



Strided unaligned shared file writes,
20 servers, up to 768 application processes



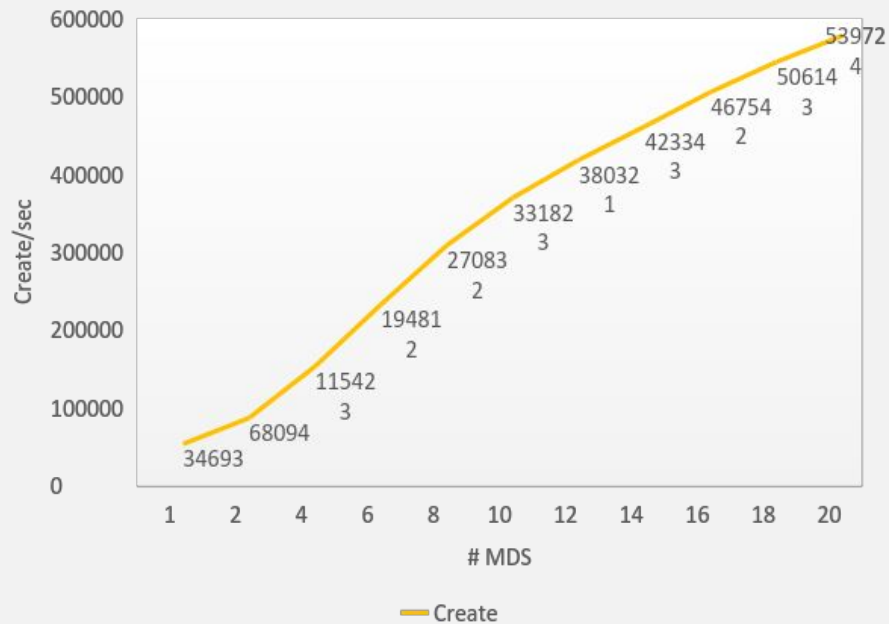
Metadata Scalability



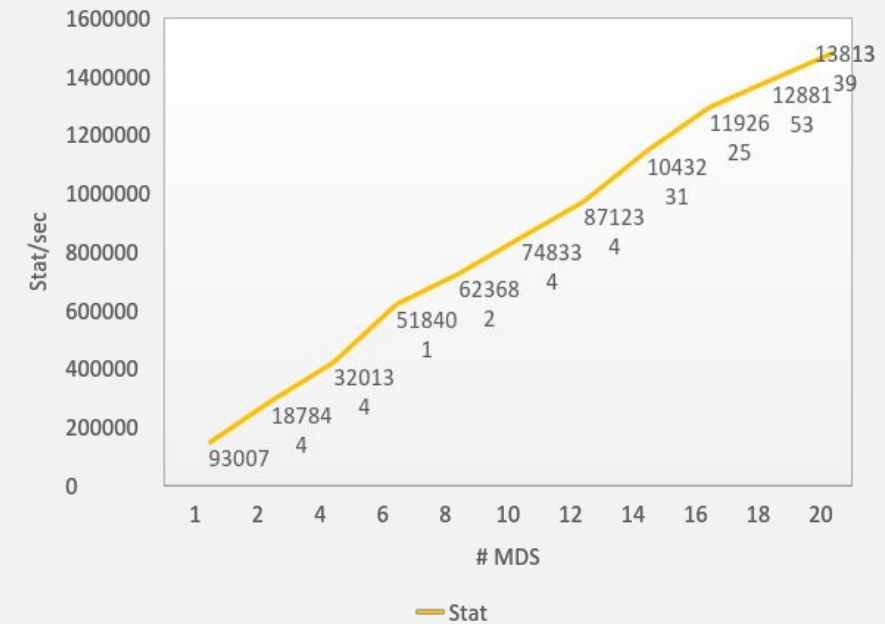
SIE



File creation scalability with increasing number of metadata servers



File stat (attribute query) scalability with increasing number of metadata servers



Key Aspects



S I E



- Performance & scalability
- **Flexibility**
 - Multiple services (any combination) can run together on the same machine
 - Flexible striping per-file / per-directory
 - Add servers at runtime
 - On demand filesystem „per job“ possible (BeeOND)
 - Runs on ancient and modern Linux distros/kernels
 - Runs on different Architectures, e.g.
 - ARM, Xeon Phi, Power, Tiler, ...
 - NFS & SMB/CIFS re-export possible



Key Aspects



S I E



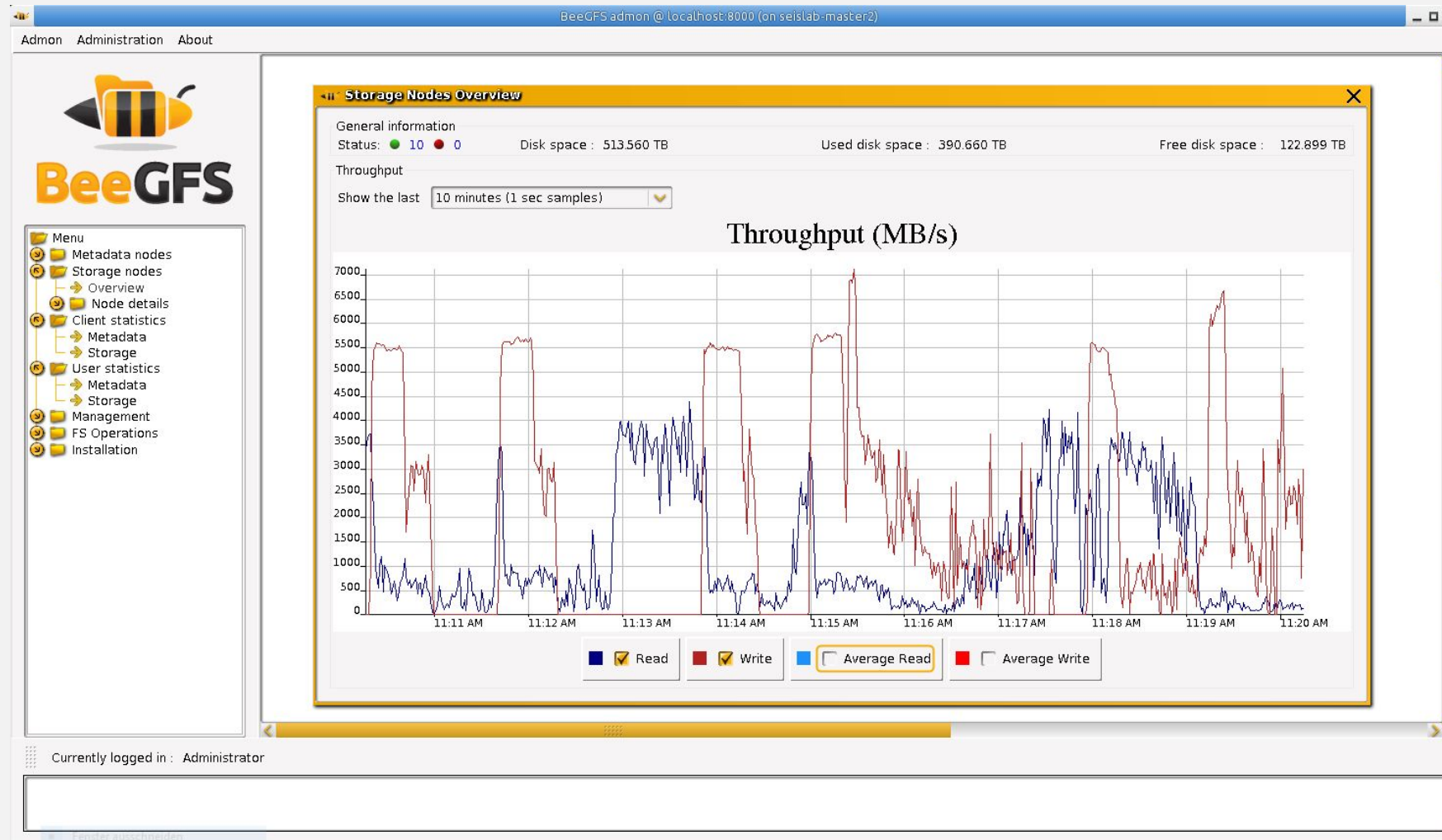
- Performance & scalability
- Flexibility
- **Robust & easy to use**
 - Applications access BeeGFS as a “normal” (very fast) file system mountpoint
 - Applications do not need to implement a special API
 - Servers daemons run in user space on top of standard local filesystems (ext4, xfs, zfs, ...)
 - No client kernel patches, kernel updates are trivially simple
 - Packages for Redhat, SuSE, Debian and derivatives
 - Hardware independent (runs on shared-nothing HW)
 - Graphical monitoring tool



Live Throughput Overview



SIE



Live per-Client and per-User Statistics



S I E



BeeGFS admin @ localhost:8000 (on seislab-master2)

Admon Administration About

BeeGFS

Menu

- Metadata nodes
- Storage nodes
- Client statistics
 - Metadata
 - Storage
- User statistics
 - Metadata
 - Storage
- Management
- FS Operations
- Installation

Currently logged in : Administrator

Client stats metadata

Settings

Interval in sec: 3 Number of clients: 20 Apply config

Filter

Use Hostname ☒

Set Filter ...

client IP	sum	mkdir	create	rmdir	open	stat	unlnk	lookLI	statLI	revalLI	openLI	createLI
sum	47067	44		11	1738	3629	10240		12089	12142		1
seislab-master3...	30997			11		20	10240		10240	10265		1
192.168.72.252	15776	44			1737	3518		1782	1747			
node92.ib.cluster	134				1	37		61	34			
node91.ib.cluster	26					9		1	16			
node79.ib.cluster	26					9		1	16			
node78.ib.cluster	26					9		1	16			
node74.ib.cluster	26					9		1	16			
node66.ib.cluster	26					9		1	16			
node65.ib.cluster	26					9		1	16			
192.168.72.253	4											

User stats metadata

Filter

Set Filter ...

lookLI	statLI	revalLI	openLI	createLI
11978	12021			
10240	10260			
1672	1638			
66	123			

breuner@seislab-master3: /scratch/breuner/bonnie

```
File Edit View Terminal Tabs Help
'module avail' - show available modules
'module add <module>' - adds a module to your environment for this session
'module initadd <module>' - configure module to be loaded at every login

An overview on available nodes follows.

Nodes in state Free : 36
Nodes in state Job-Exclusive : 52
Nodes in state Offline : 0

* seislab wiki: http://wiki.itwm.fhg.de/itwm/Seislab_User_Manual *
* seislab mailinglist: seislab@itwm.fraunhofer.de *
* seislab support: seislab-support@itwm.fraunhofer.de *

breuner@seislab-master3:~$ cd /scratch/breuner/bonnie
breuner@seislab-master3:/scratch/breuner/bonnie$ ~/prog/bonnie++-1.96/bonnie++ -s0 -n 10:0:0:10 -r0
Create files in sequential order...done.
Stat files in sequential order...done.
Delete files in sequential order...
```

GUI for Windows-style Installation



S I E



Fraunhofer FS

File Windows About

FhGFS admin @ mrfuji:8000

Installation -> Configuration

Define roles | Create basic configuration | Configure Infiniband

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 admin daemon.

Management daemon: lenny64

Metadata server

- suse11164
- sl664

Storage server

- lenny64
- suse11132
- squeeze64
- suse11164
- sl664
- lenny32

Clients

- lenny64
- suse11132
- squeeze64
- suse11164
- sl664
- lenny32

Save | Reload from server

Install FhGFS

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

Reload | Install

Currently logged in: Administrator



The easiest way to setup a parallel FS... S I E



```
# EXAMPLE...
```

```
$ beeond start -n $NODEFILE -d /local_disk/beeond -c /my_scratch
```

```
Starting BeeOND Services...
```

```
Mounting BeeOND at /my_scratch...
```

```
Done.
```

```
-----
```

```
# GENERAL USAGE...
```

```
$ beeond start -n <nodefile> -d <storagedir> -c <clientmount>
```



Scemama Anthony

@BeeGFS Wonderful feature I have been dreaming of for years!!! Thank you!!!!



BeeOND – BeeGFS On Demand



S I E



- Create a parallel file system instance on-the-fly
- Start/stop with one simple command
- Use cases: cloud computing, test systems, cluster compute nodes,
- Can be integrated in cluster batch system (e.g. PBS)
- Common use case: "per-job parallel file system"
 - Aggregate the performance and capacity of local SSDs/disks in compute nodes of a job
 - Take load from global storage
 - Speed up "dirty" I/O patterns



BeeOND[®]



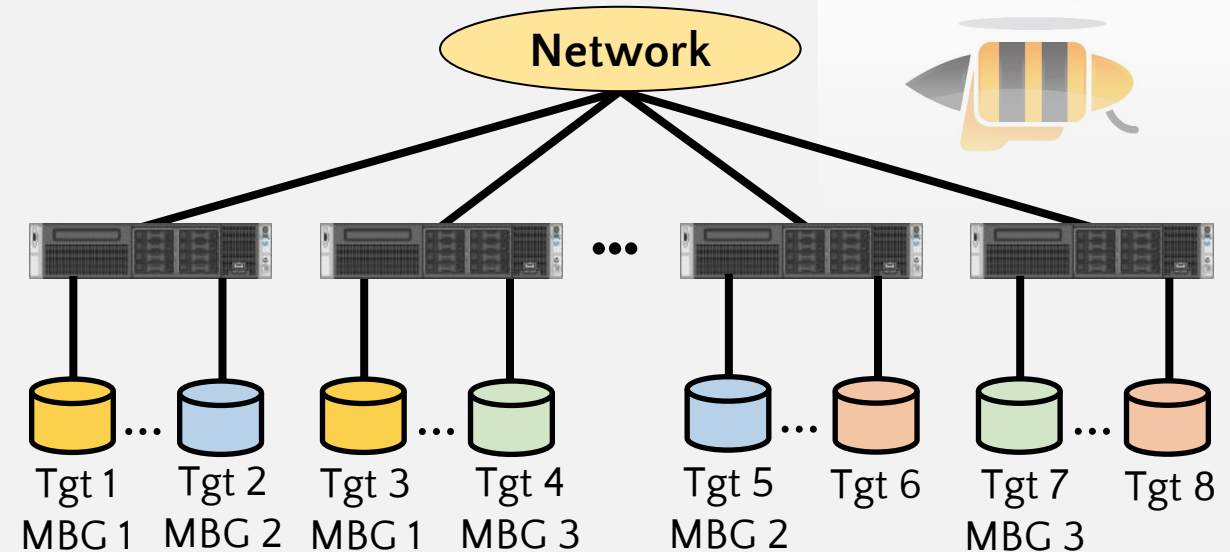
Built-in Data Mirroring



SIE



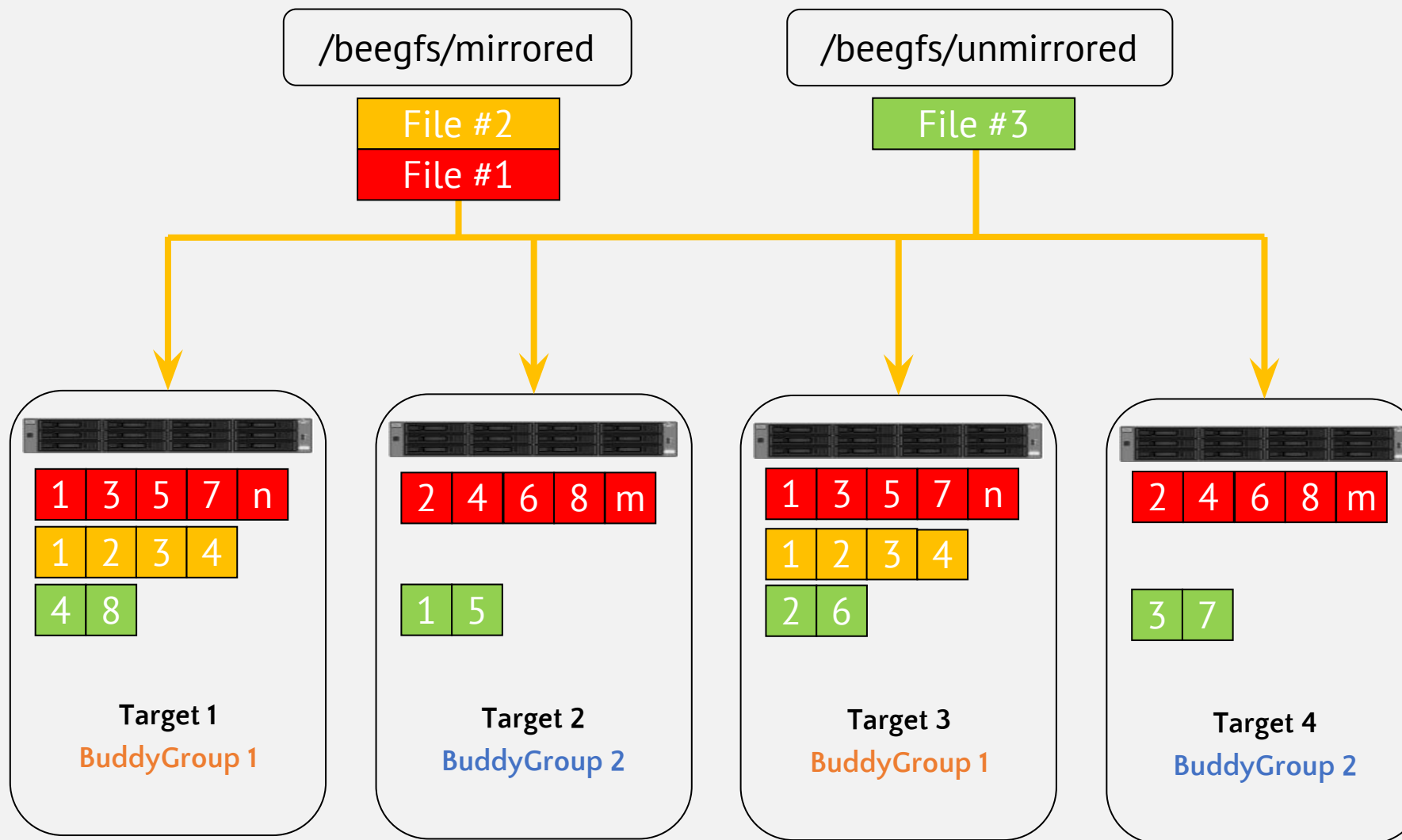
- Based on „mirror buddy groups“ of storage targets
 - Primary/secondary target in a buddy group internally replicate chunks
 - But: Targets can still also store non-mirrored chunks
 - Write operations are forwarded for high throughput
 - Read possible from both targets
- Internal failover mechanisms
 - In case primary is unreachable or fails, an automatic switch is performed
 - Self-healing (differential rebuild) when buddy comes back
- Flexible: Can be enabled globally or on a per-directory basis



Buddy Mirroring per Directory



SIE



Business Model



S I E



- **BeeGFS is free to use for end users: www.beegfs.com/download**

- Ready-to-install binaries, complete source code also available

- **System integrators/partners for turn-key solutions**

- System setup and tuning
- First point of contact (1st- and 2nd-level support)
- Partners make back2back contract with ThinkParQ for 3rd-level support



transtec

BeeGFS allows us to easily deliver petascale turn-key storage solutions
- transtec

- **Professional 3rd-level support**

- Pricing based on number of servers and timeframe (e.g. 3 or 5 years)
- Access to enterprise features (Buddy Mirroring, ACLs, quota enforcement)
- Special customer website area: www.beegfs.com/customerlogin



What's new in the 2015 Release Series?



S I E



- BeeOND (BeeGFS On Demand)
- Trinity:
 - Quota Enforcement
 - Access Control Lists (ACLs)
 - Built-in data mirroring
- Per-User Statistics in Admon GUI
- New manual setup tools (/opt/beegfs/sbin/beegfs-setup...)
- BeeGFS C API



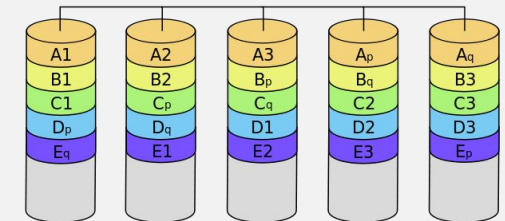
Topics for 2016



S I E



- Buddy mirroring for metadata
 - Work in progress, expected Q2/2016
- BeeGFS as a service on Amazon Cloud
 - Received funding from Amazon
 - Also in touch with Microsoft for Azure cloud
- Target pools for different hardware (e.g. fast vs big)
- Striping with parity across servers
 - Tolerate server failures with less capacity overhead compared to mirroring
 - Configurable on a per-directory basis
- Object interface for HTTP put/get style access
 - To support applications that were written for such interfaces





SIE



- LIVE TRAINING / DEMO!!!!

Thank you!

