ISC High Performance
The HPC Event.

Software Defined Parallel Storage

BeeGFS®

ISC'18 ThinkParQ BoF Session

www.beegfs.io

2018

Frank Herold CEO

The Team today / Agenda

MEGWARE®
SUPERCOMPUTING · TECHNOLOGY

Peter Grossoehme, Head of Engineering

PacificTeck

Howard Weiss, Managing Director

- First point of contact for BeeGFS worldwide
  - Delivers consulting, professional services & support for BeeGFS
  - Founded in 2014
  - Based in Germany
  - as a Fraunhofer spin-off
    - Cooperative development together with Fraunhofer (Fraunhofer continues to maintain a core BeeGFS HPC team)

- 2-tier go to market approach
  - Where partner deliver turnkey solution and 1st & 2nd level support

BeeGFS Design Philosophy

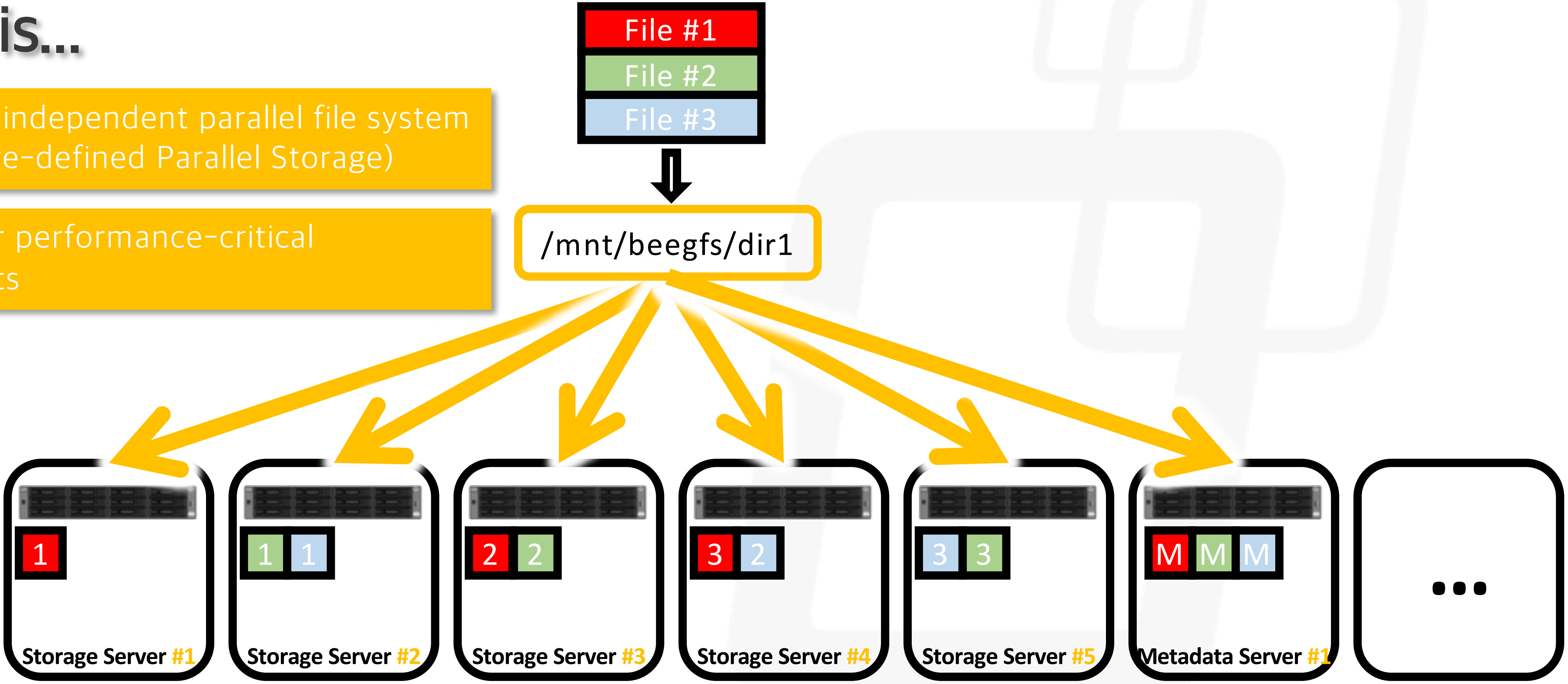
- Designed for Performance and Scalability
  - Distributed Metadata
  - No Linux patches, on top of EXT, XFS, ZFS, BTRFS, ..
  - Scalable multithreaded architecture
- Native IB and Ethernet with dynamic failover (TCP, RDMA)
- Easy to install and maintain (user space servers)
- Robust and flexible (all services can be placed independently)
- High Software Quality

What is BeeGFS?

BeeGFS is...

A hardware-independent parallel file system (aka Software-defined Parallel Storage)

Designed for performance-critical environments



File #1

File #2

File #3

/mnt/beegfs/dir1

1

Storage Server #1

1 1

Storage Server #2

2 2

Storage Server #3

3 2

Storage Server #4

3 3

Storage Server #5

M M M

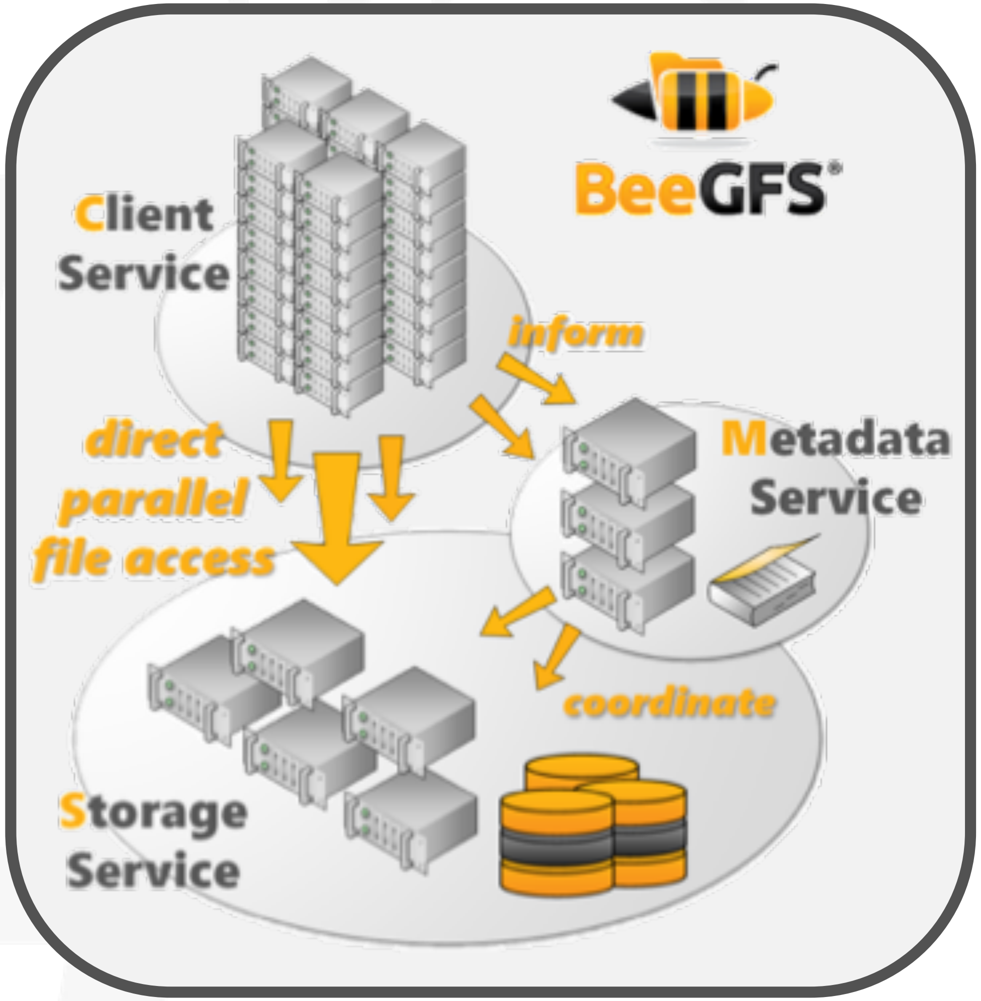
Metadata Server #1

...

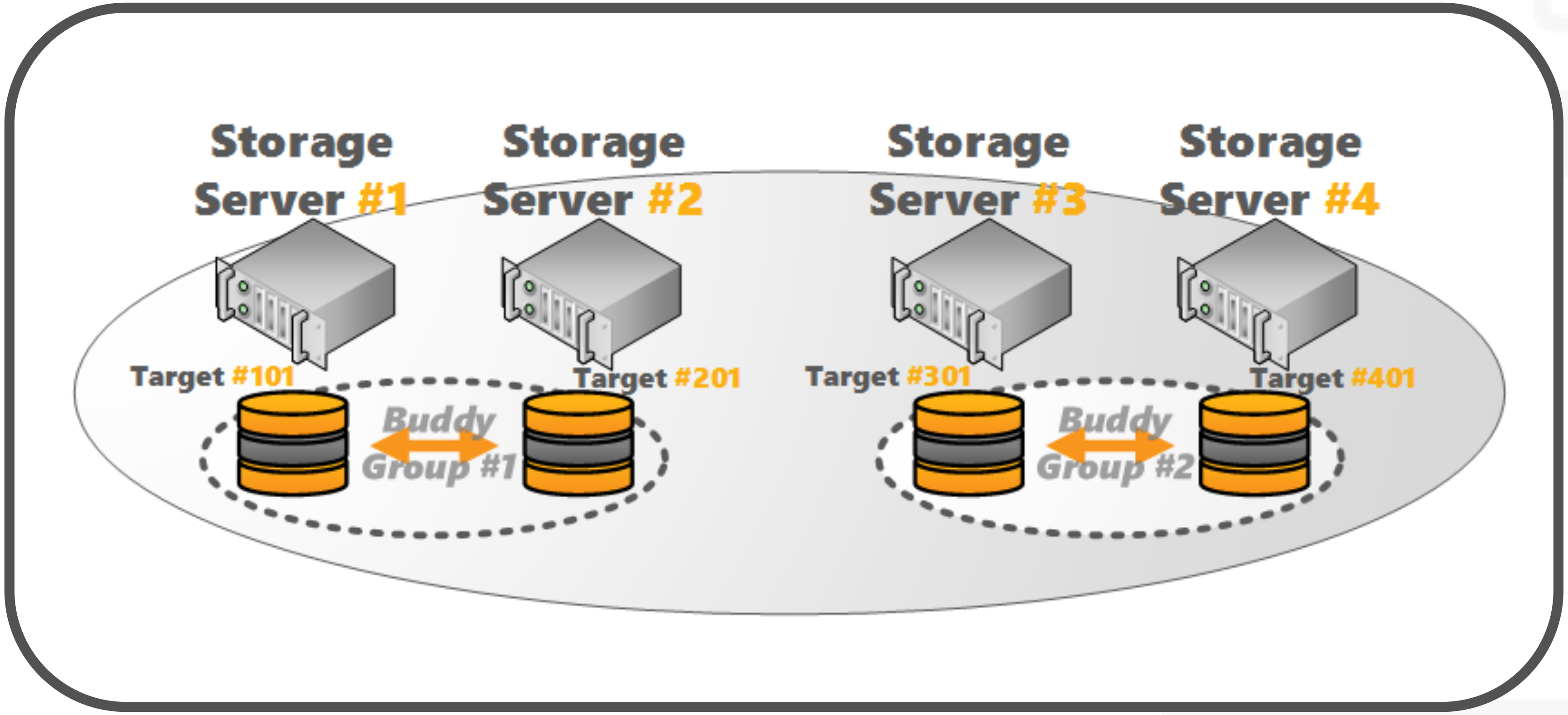
Simply grow capacity and performance to the level that you need

BeeGFS Architecture

- Client Service
  - 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
  - Service registry and watch dog
- Graphical Administration and Monitoring Service
  - GUI to perform administrative tasks and monitor system information
    - Can be used for “Windows-style installation“



Buddy Mirroring



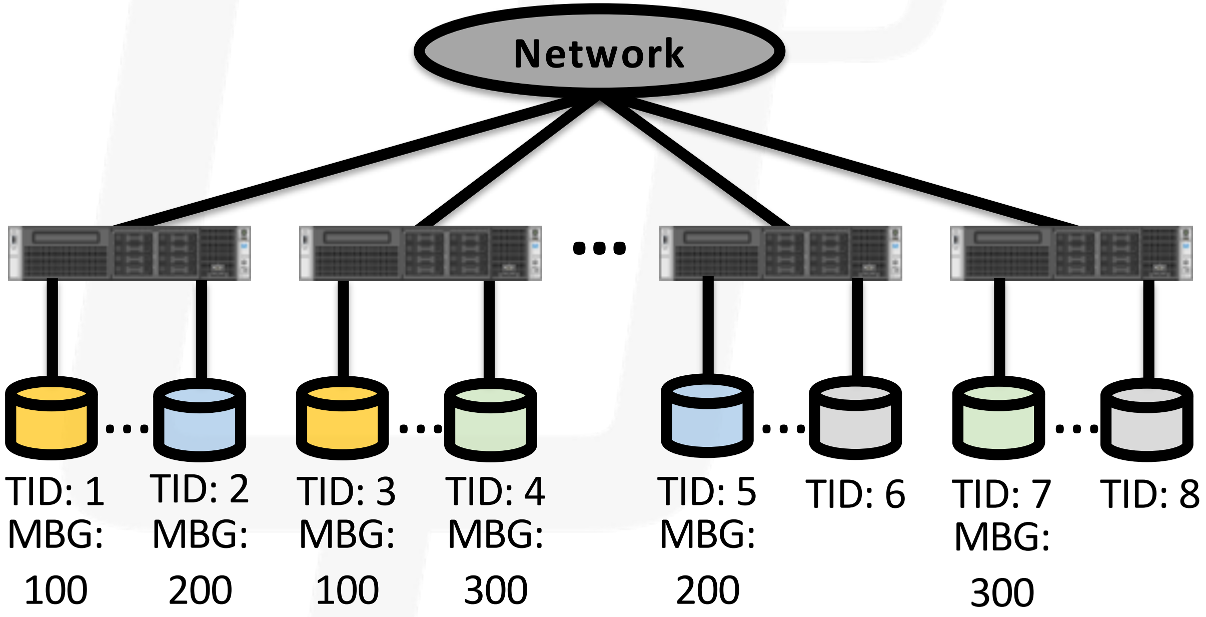
- Built-in Replication for High Availability
- Flexible setting per directory
- Individual for metadata and/or storage
- Buddies can be in different racks or different fire zones.

Built-in Data Mirroring

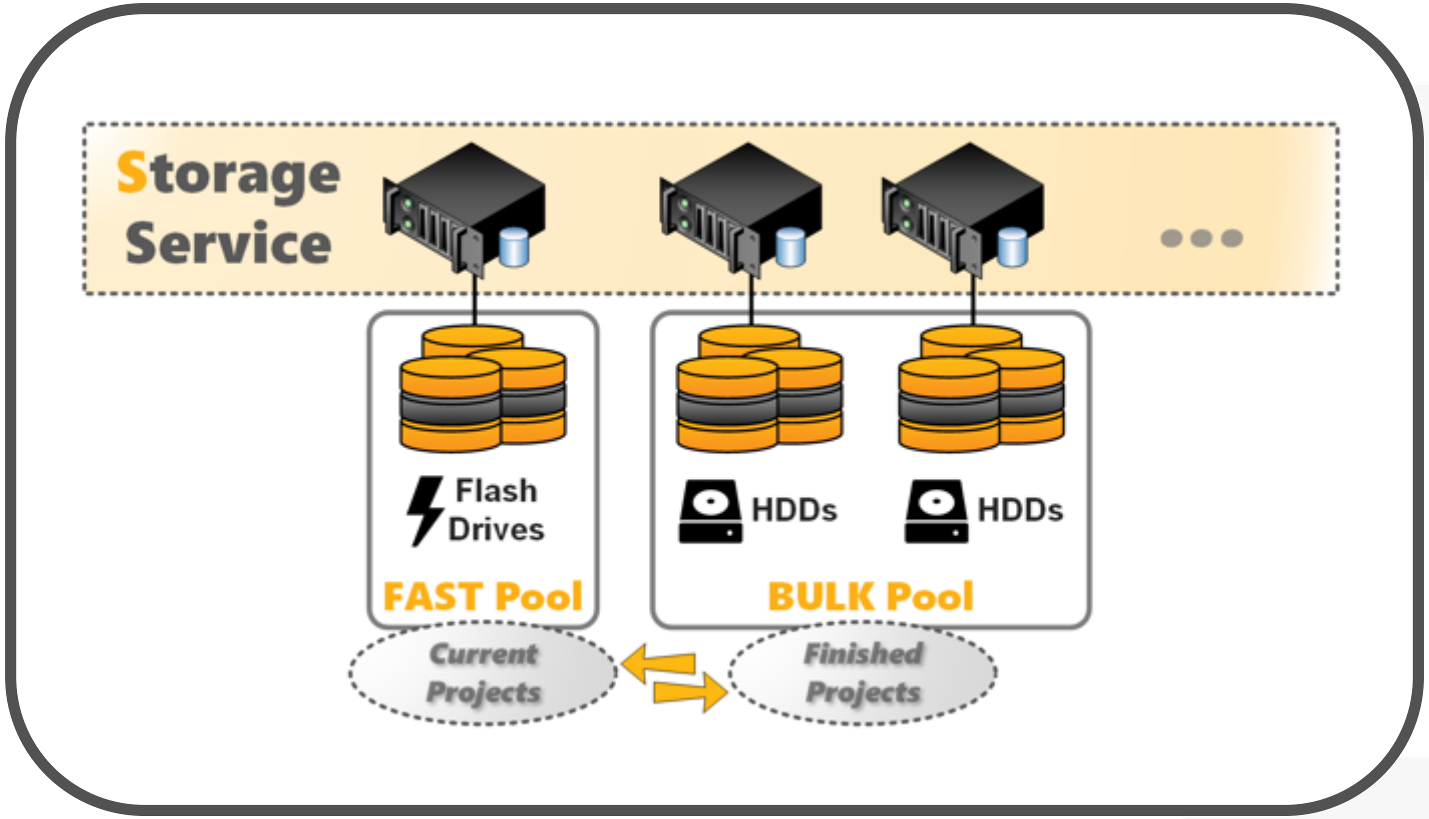
- Based on Mirror Buddy Groups of storage and/or metadata targets
  - Primary/secondary target in a buddy group replicate mirrored chunk
  - But: Targets can still also store non-mirrored chunks
  - Write operations are forwarded for high throughput
  - Read possible from both targets

- Internal failover mechanisms
  - In case primary is unreachable or fails, a switch is performed
  - Self-healing (differential rebuild) when buddy comes back

- Flexible: Can be enabled globally or on a per-directory basis



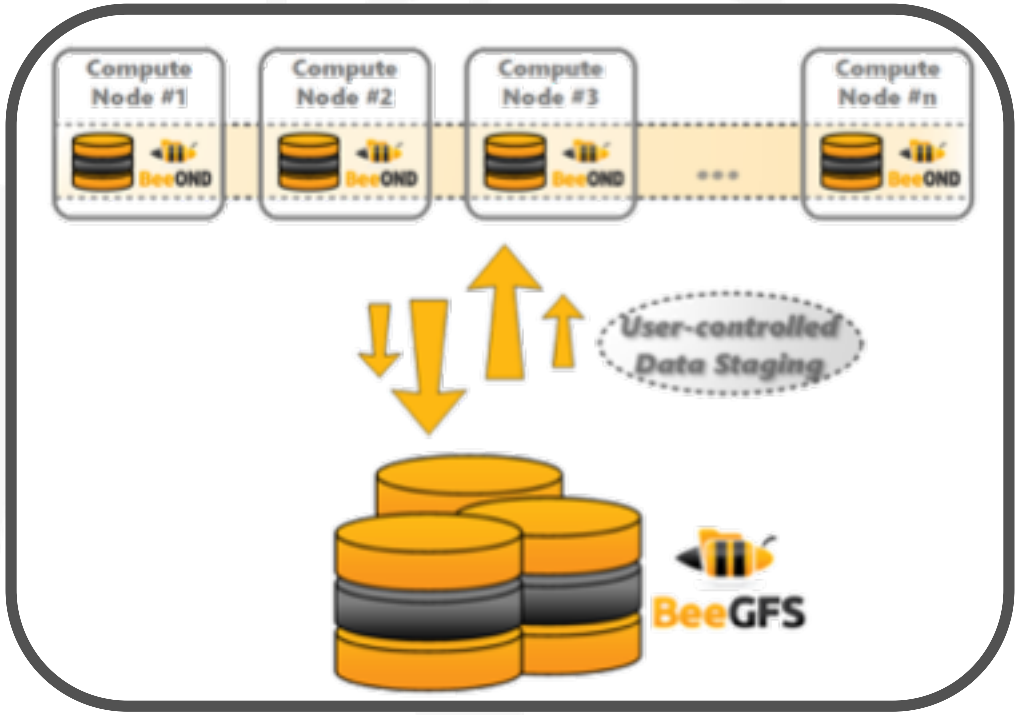
Storage Pools



- Support for different types of storage
- Modification Event Logging
- Statistics in time series database

BeeOND - BeeGFS On Demand

- Create a parallel file system instance on-the-fly
  - Aggregate the performance and capacity of local SSDs/disks in compute nodes
  - Take load from global storage
  - Speed up "nasty" I/O patterns



- Start/stop with one simple command

- Can be integrated in cluster batch system (e.g. Univa Grid Engine)

MEGWARE®
SUPERCOMPUTING · TECHNOLOGY

BeeGFS

Listening to customers

beegfs.io

company profile

- MEGWARE Computer GmbH -

- one of Europe's leading suppliers of High Performance Computing and IT technology solutions
  - established in 1990
  - full-service provider
  - more than over 1100 HPC installations to date
  - several TOP500 projects since 2000

- the only **BeeGFS Platinum partner** EMEA
  - long-term partnership since 2007
  - most BeeGFS installations in Europe

general

customer requirements (1)

- capacity
- performance
  - throughput
  - IOPS / metadata performance
- benchmarks
  - IOR / IOzone
  - Flexible I/O Tester (FIO)
  - MDTest

general

customer requirements (2)

- features
  - Quota-Tracking / -Enforcement
  - NFS- / SMB-Support
  - native InfiniBand- / Omni-Path-Support
  - Performance Monitoring
  - Quality of Service
  - (Auto-) Tiering
  - High Availabilty / Self-Healing
  - Enterprise support (L3)
  - …
- Price–performance ratio

BeeGFS – The Parallel Cluster File System

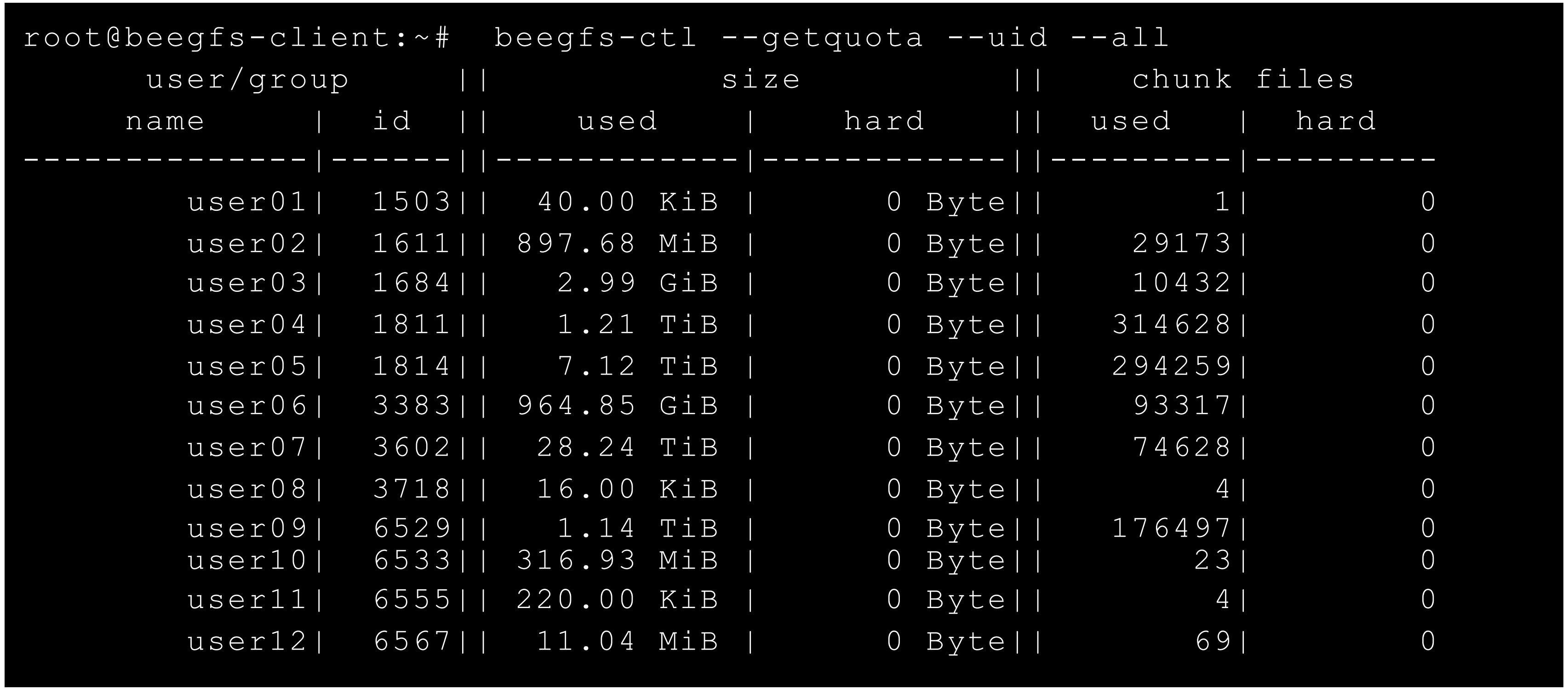
Federal Waterways Engineering and Research Institute (1)

- requirements for HPC-Cluster „Automatix“ in Karlsruhe
  - min. 200 TiB usable storage capacity
  - accessible from all parts of the cluster system
  - export for
    - Linux: NFSv4
    - Windows: SMB (Active Directory Integration)
  - Quota-Tracking / -Enforcement for User and Group(s)
  - min. 150 million files and / or directories
  - robustness against errors
  - benchmarks

BAW
Bundesanstalt für Wasserbau

BeeGFS – The Parallel Cluster File System

Federal Waterways Engineering and Research Institute (2)

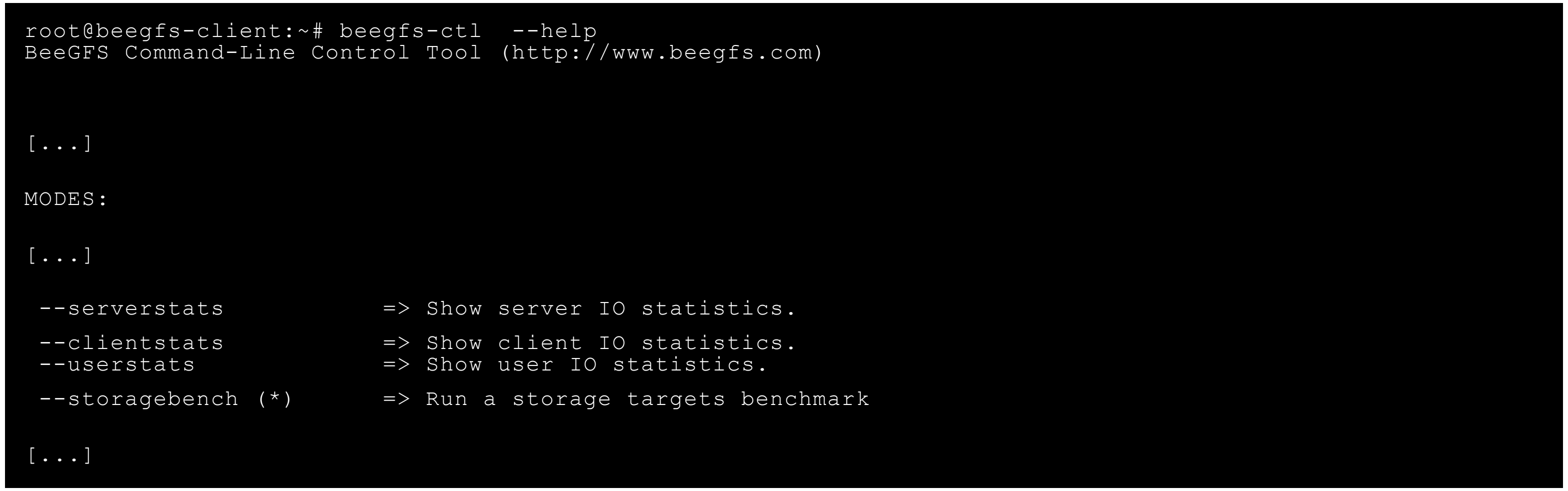


```
root@beegfs-client:~#  beegfs-ctl --getquota --uid --all
      user/group     ||            size          ||    chunk files
     name     |  id  ||    used    |     hard    ||  used   |  hard
--------------|------||------------|-------------||---------|----------
        user01|  1503||  40.00 KiB |       0 Byte||        1|         0
        user02|  1611|| 897.68 MiB |       0 Byte||    29173|         0
        user03|  1684||    2.99 GiB |       0 Byte||    10432|         0
        user04|  1811||    1.21 TiB |       0 Byte||   314628|         0
        user05|  1814||    7.12 TiB |       0 Byte||   294259|         0
        user06|  3383|| 964.85 GiB |       0 Byte||    93317|         0
        user07|  3602||  28.24 TiB |       0 Byte||    74628|         0
        user08|  3718||  16.00 KiB |       0 Byte||        4|         0
        user09|  6529||    1.14 TiB |       0 Byte||   176497|         0
        user10|  6533|| 316.93 MiB |       0 Byte||       23|         0
        user11|  6555|| 220.00 KiB |       0 Byte||        4|         0
        user12|  6567||  11.04 MiB |       0 Byte||       69|         0
```

example: beegfs-ctl – get user quota information

BeeGFS – The Parallel Cluster File System

Federal Waterways Engineering and Research Institute (3)



```
root@beegfs-client:~# beegfs-ctl  --help
BeeGFS Command-Line Control Tool (http://www.beegfs.com)



[...]

MODES:

[...]

 --serverstats            => Show server IO statistics.
 --clientstats            => Show client IO statistics.
 --userstats              => Show user IO statistics.
 --storagebench (*)       => Run a storage targets benchmark

[...]
```

example: beegfs-ctl – get performance metrics

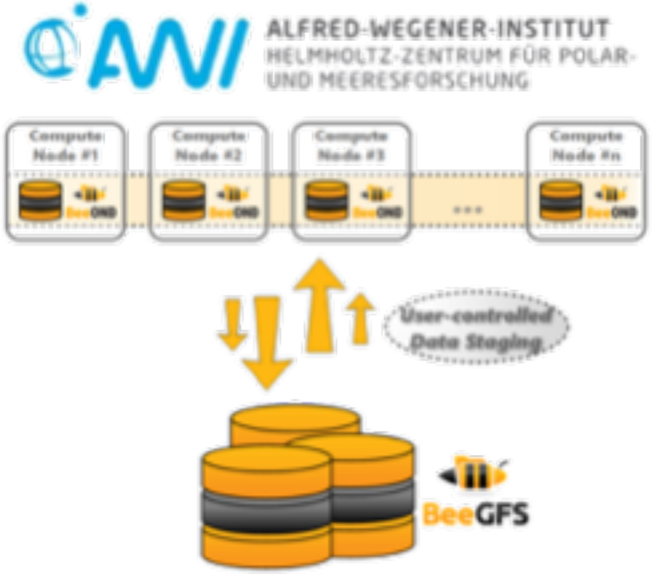
BeeOND – BeeGFS on Demand

- CRAY/Megware CS400 HPC-Cluster at AWI, Bremerhaven -

- environment
  - 308 compute nodes with a 500 MB/s SSD each
    - more than 150 GB/s aggregated bandwidth
    - easy to use (hosts, local data, mount point)

➡ https://www.beegfs.io/wiki/BeeOND

- create BeeOND on SSDs at job startup
  - with SLURM prolog / epilog scripts
    - create and destroy of BeeOND instance
- scripts for
  - stage-in input data, work on BeeOND, stage-out results



BeeGFS customer experiences

- Why are BeeGFS customers so satisfied? -

“Robust and stable, even in a case of unexpected power failure.“

Dr. Malte Thoma

- Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research -
(Bremerhaven, Germany)

- in general
  - **performance & scalability**
  - **robust & easy to use**
  - **flexibility**
  - **compatibility**

… it „just“ works! ☺

How to scale

beegfs.io

Pacific Teck Limited:
HPC/Machine Learning Experts in APAC

- Gold Value Added Reseller for ThinkParQ in APAC
- Located in Tokyo, Japan
- Fluent in English, Japanese and Chinese
- References in the largest computing centers in APAC
- Technical experts in filesystems, interconnects and schedulers

Pacific Teck Products

Univa Corporation

- Univa Grid Engine (workload manager software)
- Docker and Container support

Sylabs

- Singularity Container offering for parallel environments
- Ideal Container solution for Univa Grid Engine

ThinkParQ

- BeeGFS Filesystem
- "BeeOND" - BeeGFS On-Demand

Intel Corporation

- Intel Omni-Path Architecture (interconnect)

Utilizing NVMe with BeeOND

BeeOND

@

TOKYO TECH

産総研

国立研究開発法人 産業技術総合研究所

NATIONAL INSTITUTE OF
ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

Tokyo Institute of Technology: Tsubame 3

Tokyo Institute of Technology

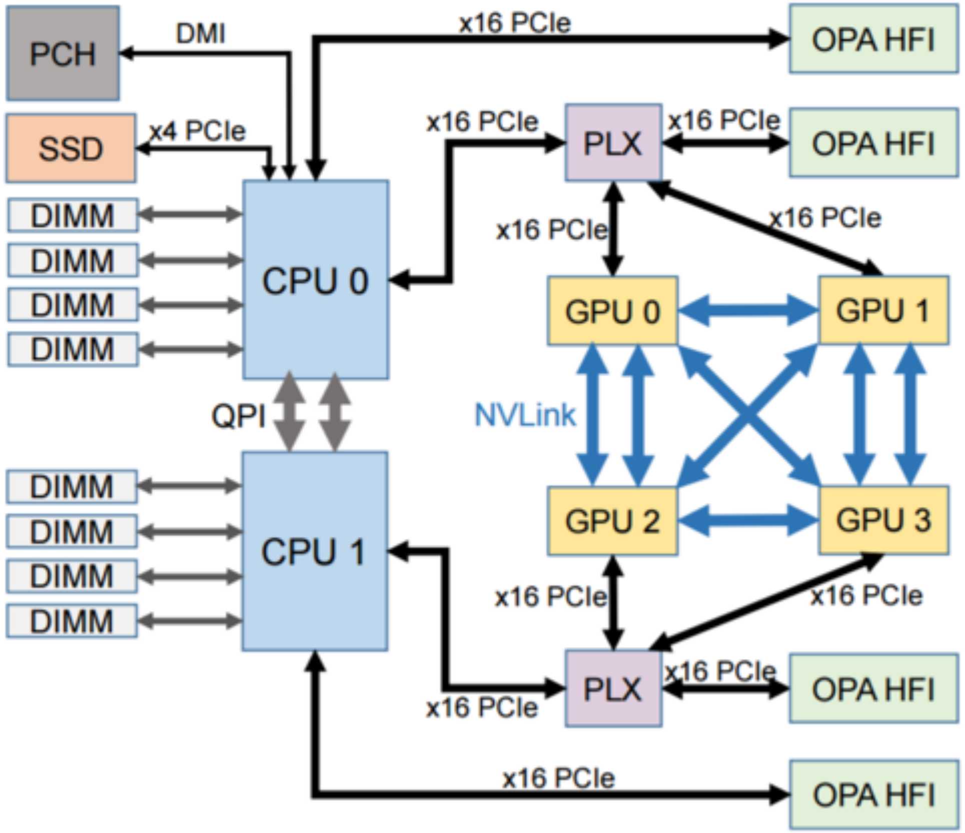
- Top national university for science and technology in Japan
- 130 year history
- Over 10,000 students located in the Tokyo Area

Tsubame 3

- Latest Tsubame Supercomputer
- #1 on the Green500 in November 2017
- 14.110 GFLOPS[2] per watt
- BeeOND uses 1PB of available NVMe

Tokyo Institute of Technology Tsubame 3 Configuration

- 540 nodes
- Four Nvidia Tesla P100 GPUs per node (2,160 total)
- Two 14-core Intel Xeon Processor E5-2680 v4 (15,120 cores total)
- Two dual-port Intel Omni-Path Architecture HFIs (2,160 ports total)
- 2 TB of Intel SSD DC Product Family for NVMe storage devices
- Simple integration with Univa Grid Engine



AIST: ABCI

AIST (National Institute of Advanced Industrial Science and Technology)

- Japanese Research Institute located in the Greater Tokyo Area
- Over 2,000 researchers
- Part of the Ministry of Economy, Trade and Industry

ABCI (AI Bridging Cloud Infrastructure)

- Japanese supercomputer scheduled for production on July 1, 2018
- Theoretical performance is 130pflops – one of the fastest in the world
- Will make its resources available through the cloud to various private and public entities in Japan

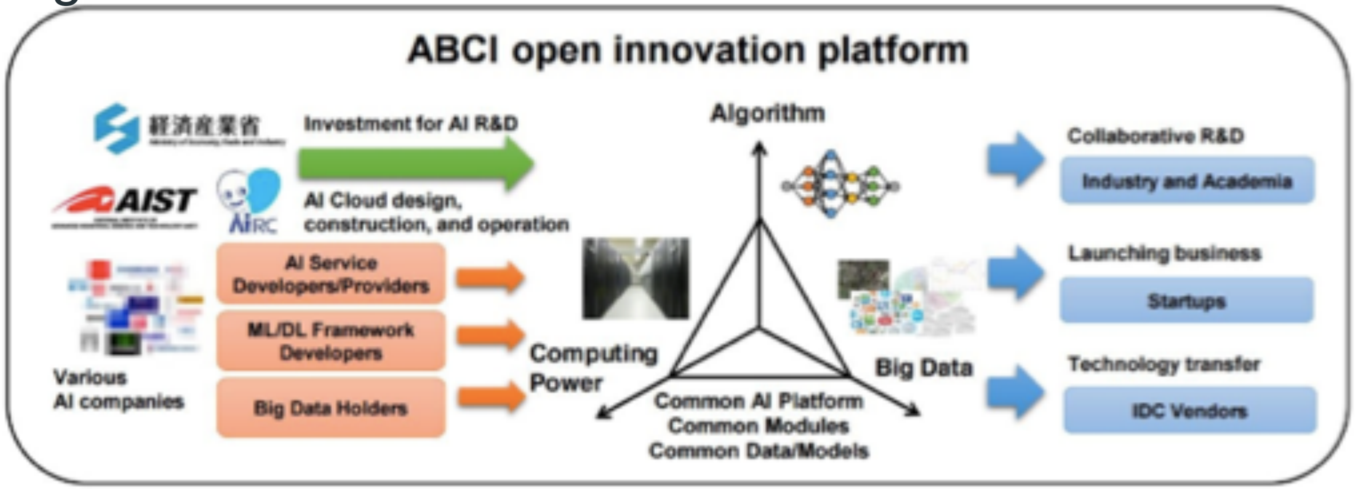
産総研

国立研究開発法人 産業技術総合研究所

NATIONAL INSTITUTE OF
ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

Largest Machine Learning Environment in Japan uses BeeOND

- 1,088 servers
- Two Intel Xeon Gold processor CPUs (a total of 2,176 CPUs)
- Four NVIDIA Tesla V100 GPU computing cards (a total of 4,352 GPUs)
- Intel SSD DC P4600 series based on an NVMe standard, as local storage. 1.6TB per node (a total of about 1.6PB)
- InfiniBand EDR
- Simple integration with Univa Grid Engine



Issues solved with BeeGFS and BeeOND

- Ability to fully utilize NVMe drives in GPU environments with BeeOND
- Converged storage made possible
- Many different OS types supported
- Large and small files supported
- Easy deployment of the BeeGFS into cloud computing environments

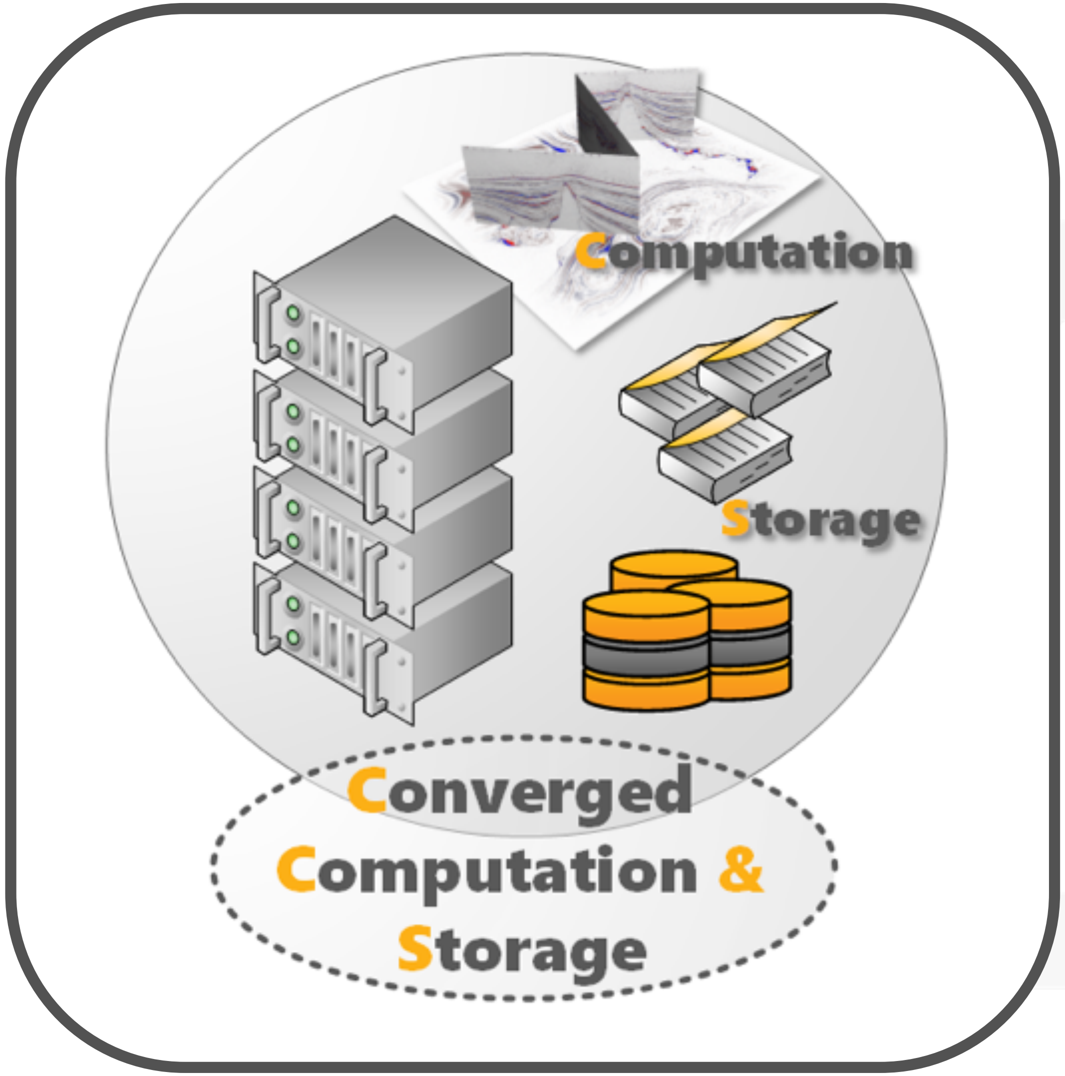
Converged Storage with BeeGFS

BeeGFS®

@

spookfish

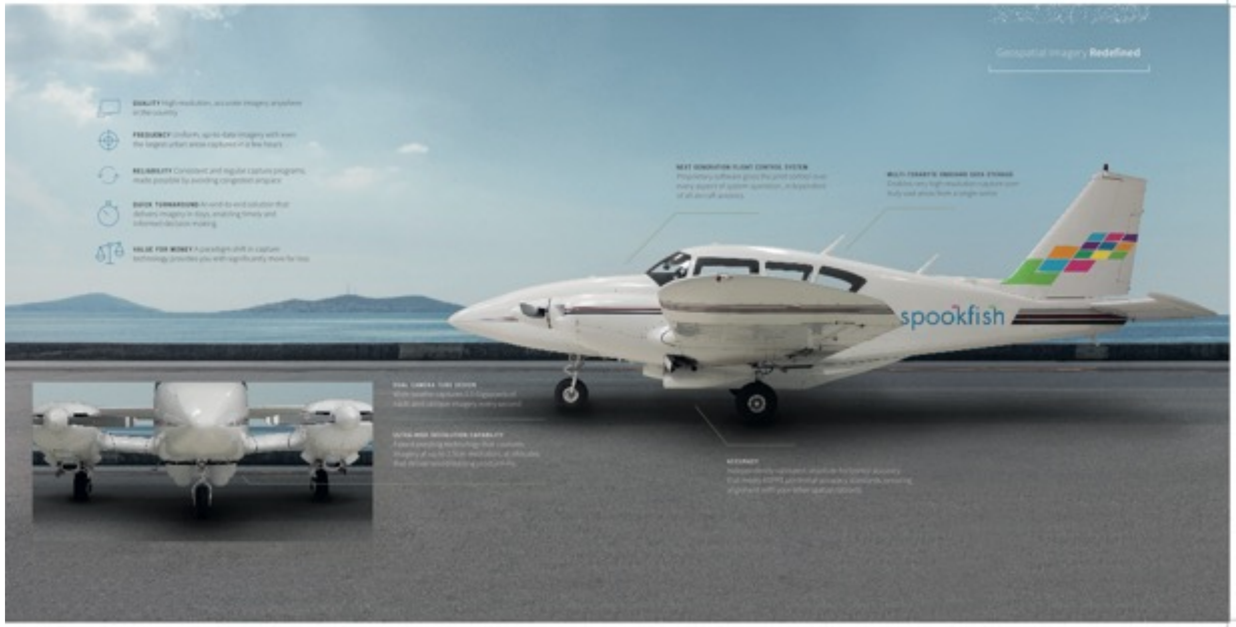
Storage + Compute: Converged Setup



Compute nodes as storage servers for small systems

Spookfish

- Aerial survey system based in Western Australia
- High resolution images are provided to customers who need up to date information on terrain they plan to utilize
- Information can be fed into GIS and CAD applications.



Geospatial Imagery Redefined

QUALITY High-resolution, accurate imagery anywhere in the country

FREQUENCY Uniform, up-to-date imagery with even the largest urban areas captured in a few hours

RELIABILITY Consistent and regular capture programs, made possible by avoiding congested airspace

QUICK TURNAROUND An end-to-end solution that delivers imagery in days, enabling timely and informed decision making

VALUE FOR MONEY A paradigm shift in capture technology provides you with significantly more for less

NEXT GENERATION FLIGHT CONTROL SYSTEM Proprietary software gives the pilot control over every aspect of system operation, independent of all aircraft avionics.

MULTI-TERABYTE ONBOARD DATA STORAGE [illegible]

DUAL CAMERA TUBE DESIGN [illegible]

ULTRA-WIDE RESOLUTION CAPABILITY [illegible]

ACCURACY [illegible]

Spookfish

spookfish

Spookfish System Architecture

- Metadata server x 6
  - Supermicro chassis with 4 x Intel Xeon X7560 and 256GB RAM
  - Only performs MDS Services
  - Metadata target x6 with buddy mirroring
- Converged storage server x 40
  - DELL R730 with 2 x Intel Xeon E5-2650v4 CPU's and 128GB of RAM
  - Storage servers also perform processing for applications
  - Uses Linux cgroups to avoid out-of-memory events
  - cgroups not used for CPU usage and so far no issues of CPU shortage
  - Storage target x 160 with buddy mirroring
- 10GB/s Ethernet

BeeGFS Converged Storage at Spookfish Summary

- Installed BeeGFS in converged storage with application, metadata, and storage all combined in a single server
- 40 converged storage servers ingest map data from cameras in airplanes
- Large and small file types are supported
- Performance exceeded expectations with 10GB/s read and 5-6GB/s write after tuning
- **"The result [of switching to BeeGFS] is that we're now able to process about 3 times faster with BeeGFS than with our old NFS server. We're seeing speeds of up to 10GB/s read and 5-6GB/s write." -Spookfish**

Cloud Compatible BeeGFS

BeeGFS®

@

amazon
web services

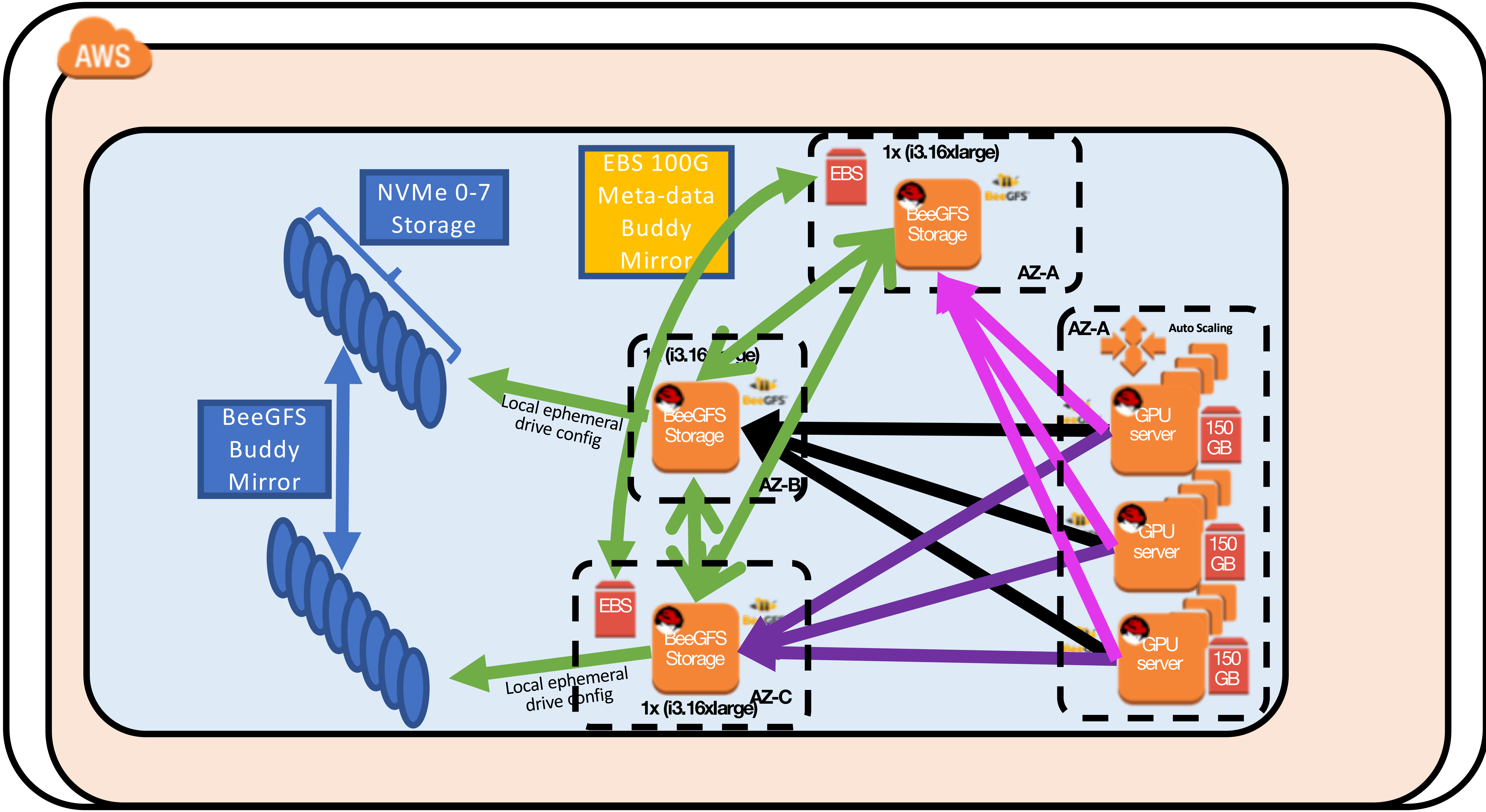
BeeGFS on AWS

- Provisioned BeeGFS in AWS cloud
- This provisioning method is replicable for future AWS users
- Pacific Teck Optimized performance with 8GB/s throughput

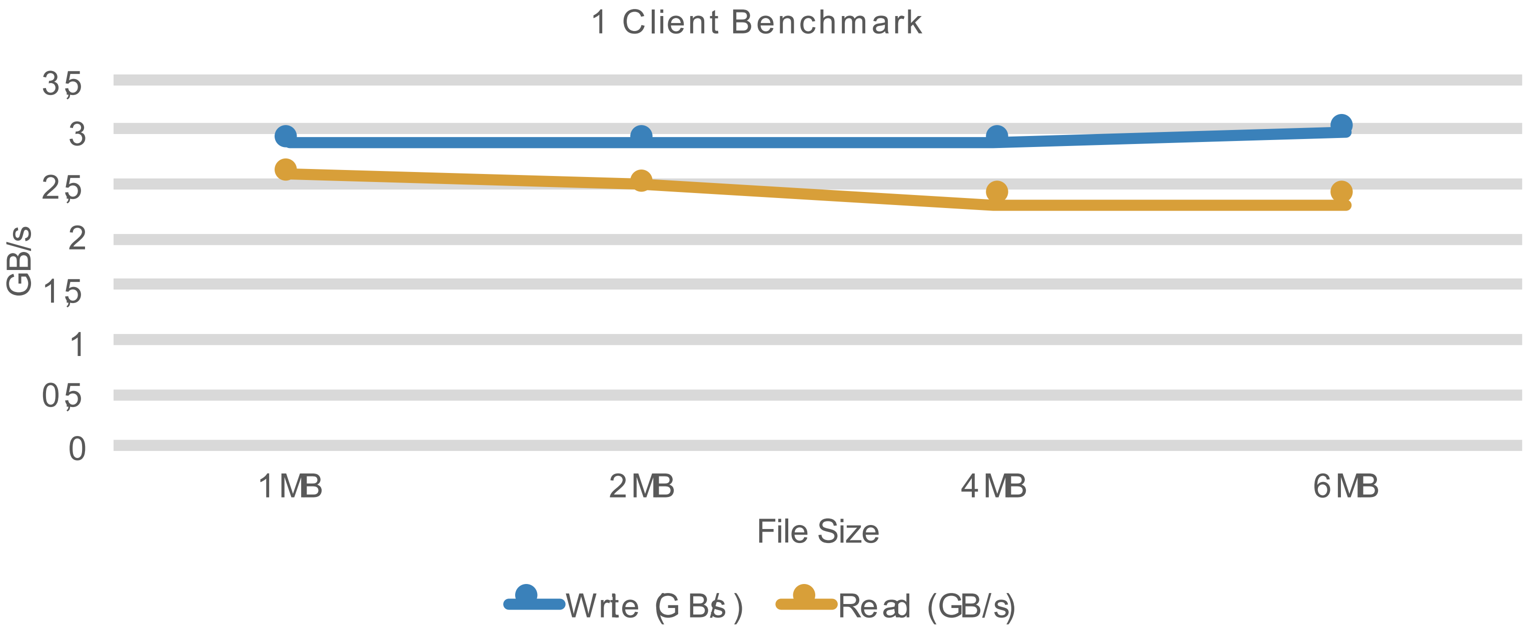
amazon
web services

Sample Architecture

- 3 x IO server i3.16xlarge with 8 nvme disks. Each of the NVMe drives is a storage target
- 100GB metadata with EBS configured LVM RAID 1
- Buddy mirroring with metadata and storage target
- IO server OS is RHEL7.4, client OS is RHEL7.4
- Storage target block size test with 4KiB and test file size 1MB, 2MB, 4MB, 6MB. With one, two,three p2.16xlarge client

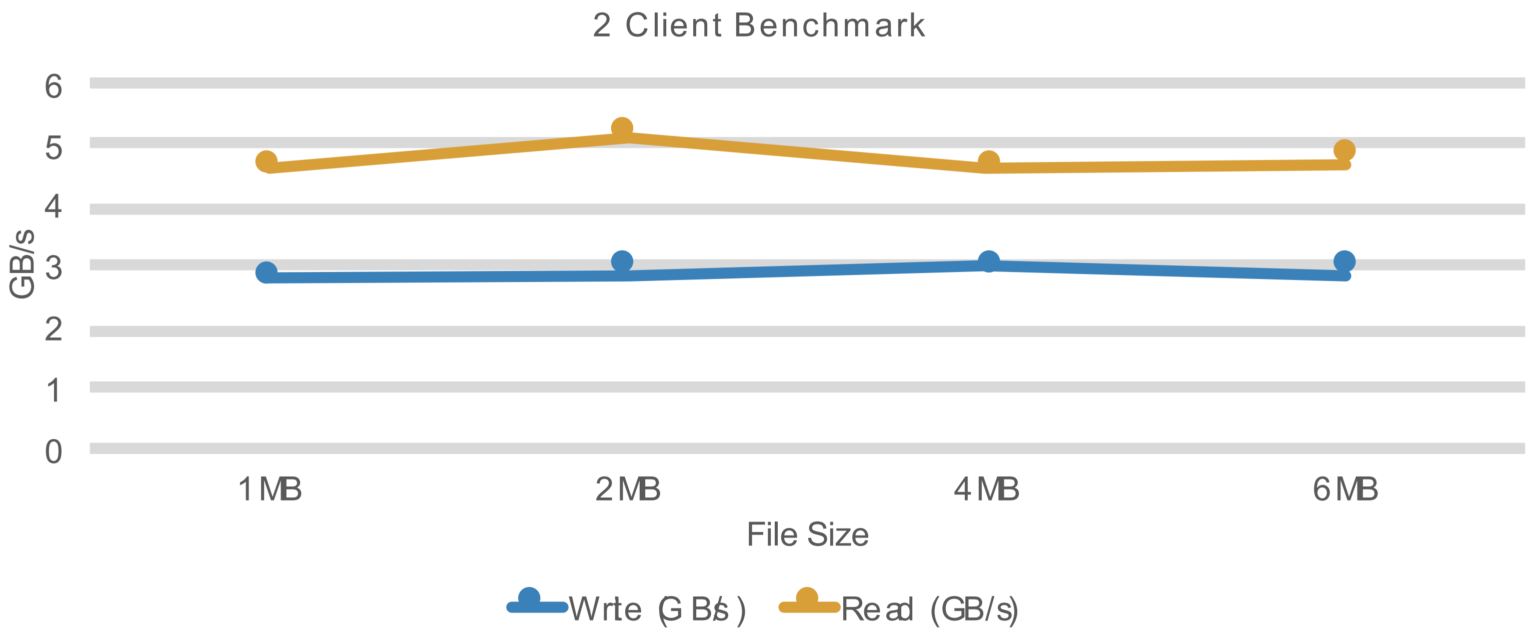

AWS
NVMe 0-7
Storage
EBS 100G
Meta-data
Buddy
Mirror
BeeGFS
Buddy
Mirror
1x (i3.16xlarge)
EBS
BeeGFS
Storage
BeeGFS
AZ-A
1 (i3.16 ge)
BeeGFS
Storage
BeeGFS
AZ-B
Local ephemeral
drive config
EBS
BeeGFS
Storage
AZ-C
1x (i3.16xlarge)
Local ephemeral
drive config
AZ-A
Auto Scaling
GPU
server
150
GB
GPU
server
150
GB
GPU
server
150
GB

Test result for IOZONE with 1 client



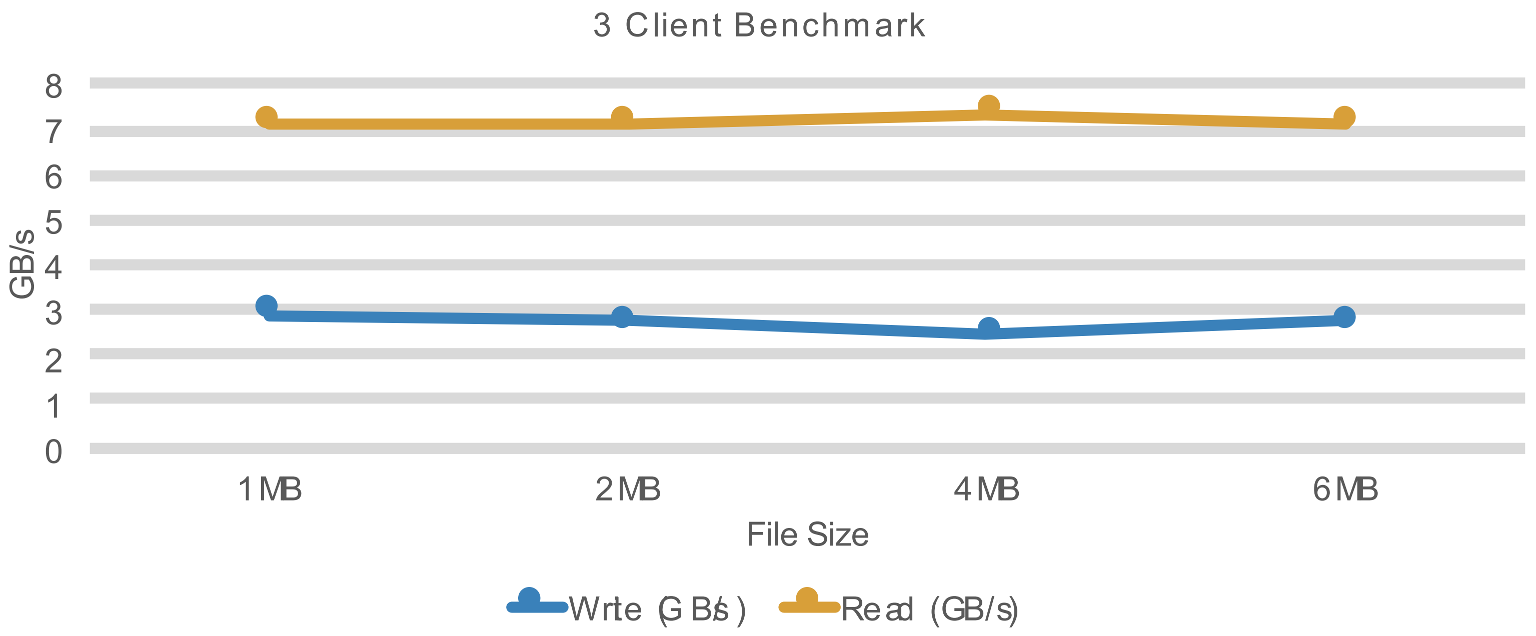
	1MB	2MB	4MB	6MB
Write	2.9 GB/sec	2.9 GB/sec	2.9 GB/sec	3.0 GB/sec
read	2.6 GB/sec	2.5 GB/sec	2.3 GB/sec	2.3 GB/sec

Test result for IOZONE with 2 clients



	1MB	2MB	4MB	6MB
Write	2.8 GB/sec	2.9 GB/sec	3.0 GB/sec	2.9 GB/sec
read	4.6 GB/sec	5.1 GB/sec	4.6 GB/sec	4.7 GB/sec

Test result for IOZONE with 3 clients



	1MB	2MB	4MB	6MB
Write	2.9 GB/sec	2.8 GB/sec	2.5 GB/sec	2.7 GB/sec
read	7.1 GB/sec	7.1 GB/sec	7.3 GB/sec	7.1 GB/sec

Summary

- Pacific Teck is the Gold VAR in APAC with expertise in
  - File systems
  - Interconnects
  - Schedulers
- BeeGFS and BeeOND solve problems in APAC such as
  - Utilizing NVMe
  - Converged storage configurations
  - Providing a high-speed file system in the cloud

Time to listen - your feedback is important

beegfs.io

Please come and visit us @J640

Welcome reception starts now