BeeGFS

The Parallel Cluster File System

Container Workshop ISC 28.7.18

thinkparQ

Marco Merkel
VP ww Sales, Consulting

www.beegfs.io

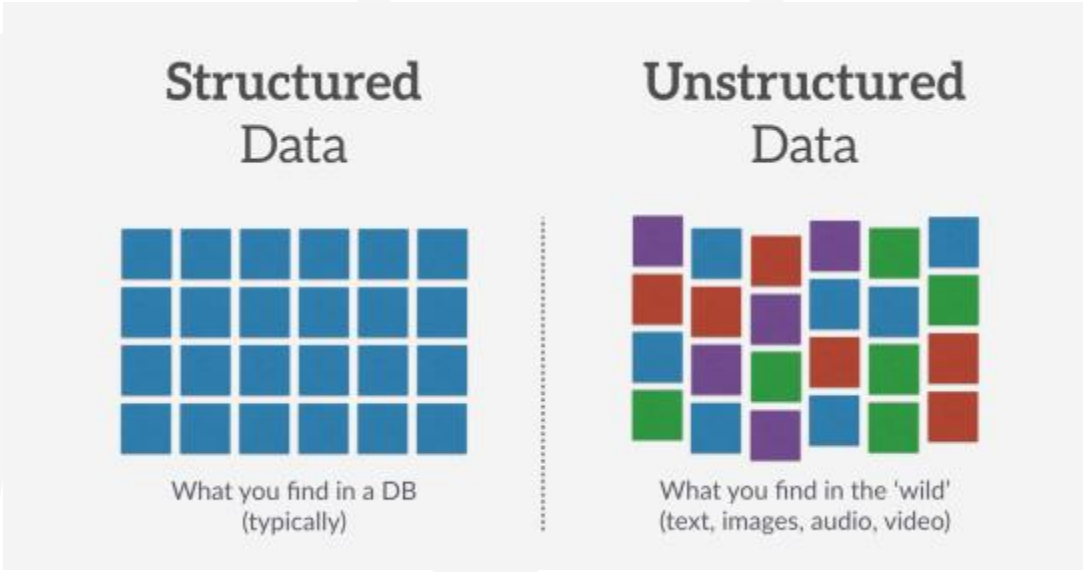
July 2018

HPC & Cognitive Workloads Demand Today

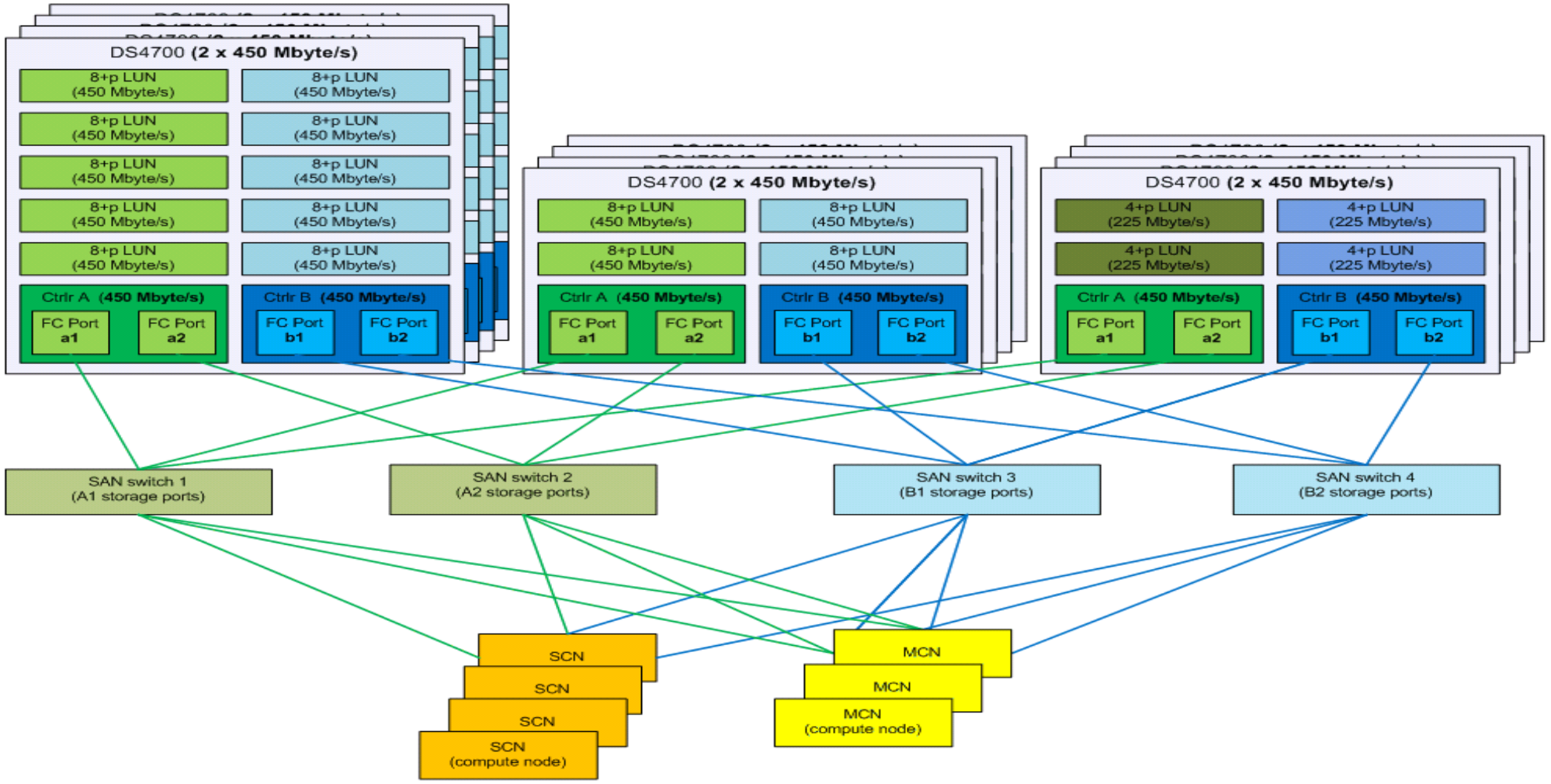
- Flash Storage
- HDD Storage
- Shingled Storage & Cloud

- GPU

- Network



Traditional System Architecture(s)



BeeGFS Inventor

- **The Fraunhofer Gesellschaft (FhG)**
  - Largest organization for applied research in Europe
  - 66 institutes in Germany, research units and offices around the globe
  - Staff: ~24000 employees
  - Annual Budget: ~2 Billion €

Inventor of mp3 Audio Codec

- **The Fraunhofer Center for High-Performance Computing**
  - Part of Fraunhofer Institute for Industrial Mathematics (ITWM)
  - Located in Kaiserslautern, Germany
  - Staff: ~260 employees

ThinkParQ

- A Fraunhofer spin-off
- Founded in 2014 specifically for BeeGFS
- Based in Kaiserslautern (right next to Fraunhofer HPC Center)

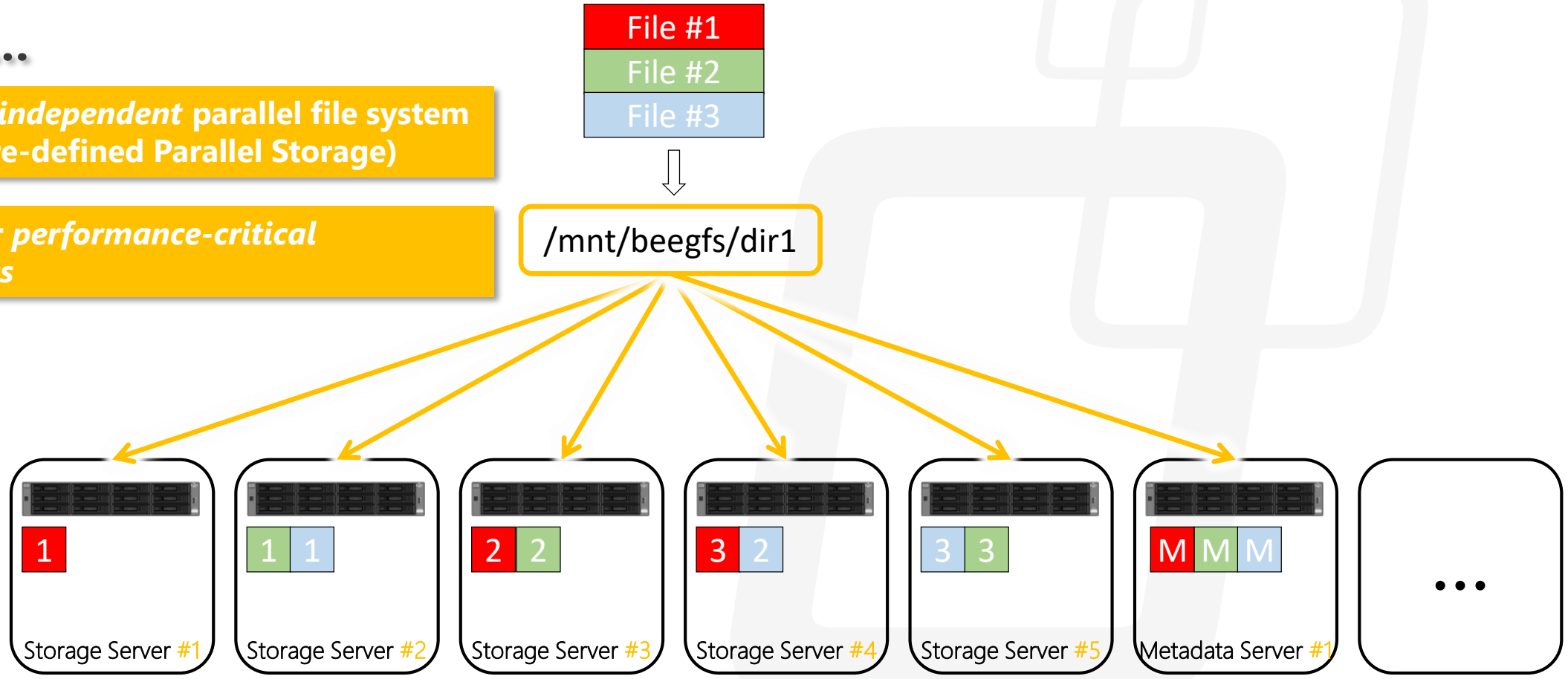
- Consulting, professional services & support for BeeGFS
- Cooperative development together with Fraunhofer
  (Fraunhofer continues to maintain a core BeeGFS HPC team)

- First point of contact for BeeGFS

What is BeeGFS?

BeeGFS is…

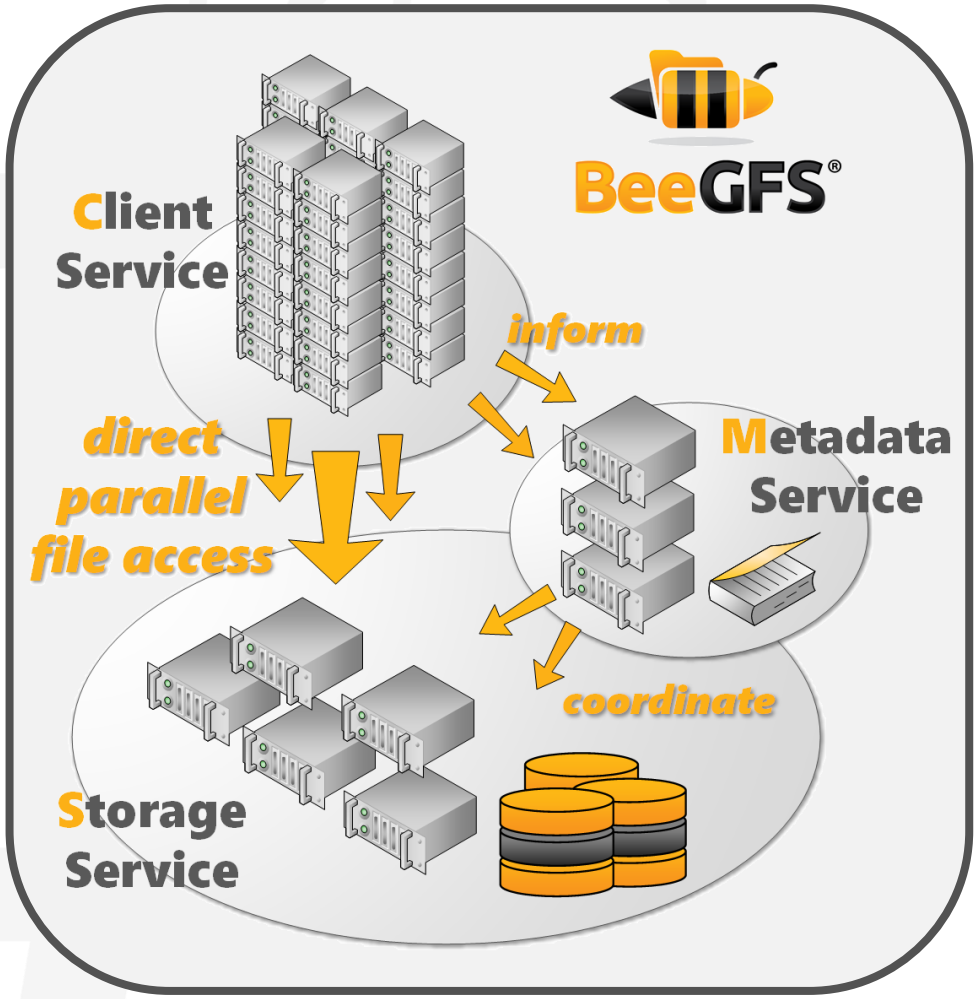
- A *hardware-independent* parallel file system (aka Software-defined Parallel Storage)
- Designed for *performance-critical environments*



- 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“



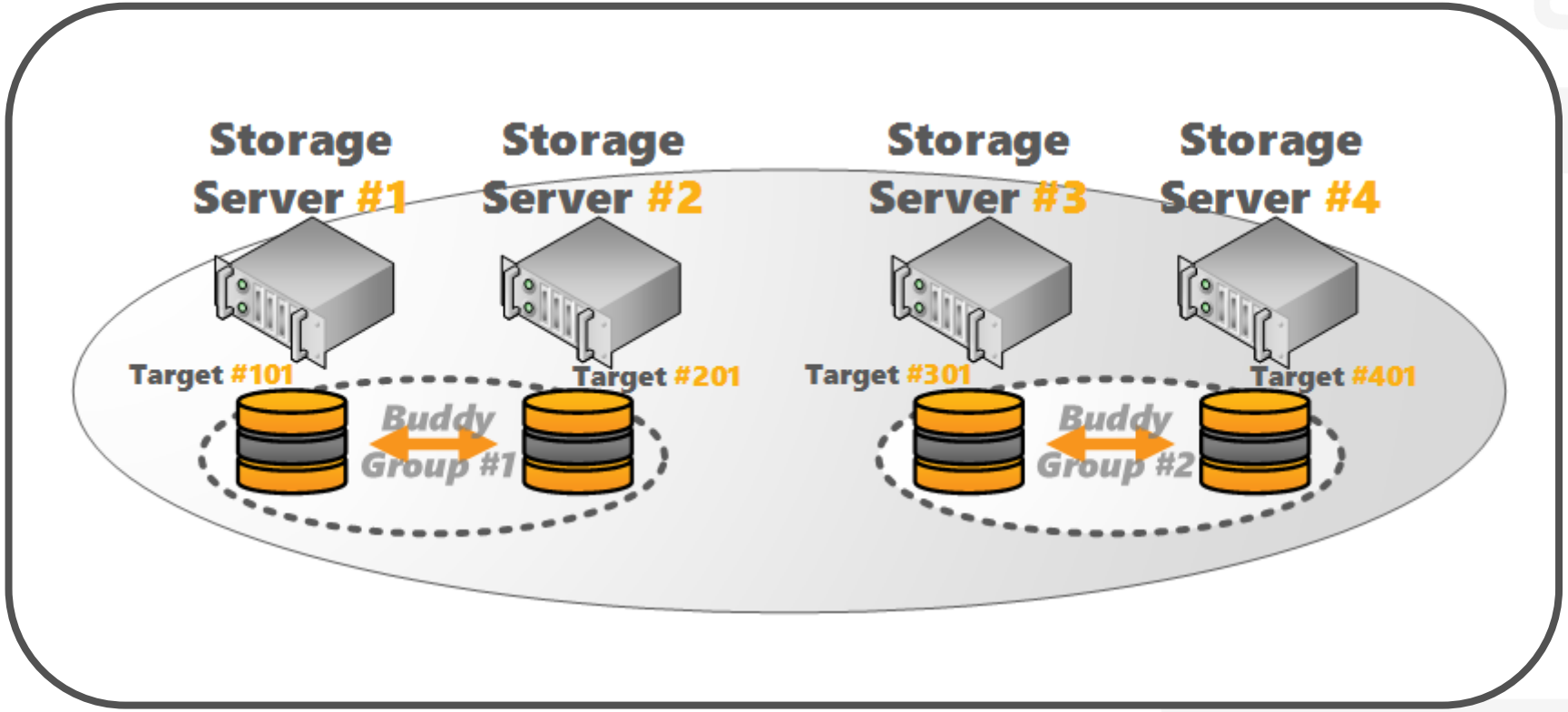
Enterprise Features

- BeeGFS Enterprise Features do require a support contract:
- High Availability
- Quota Enforcement
- Access Control Lists (ACLs)
- Storage Pools
- Burst buffer function with Beeond

Support Benefits:

- Competent level 3 support (next business day)
- Access to customer login area
- Special repositories with early updates and hotfixes
- Additional documentation and how to guides
- Direct contact to the file system development team
- Guaranteed next business day response
- Optional remote login for faster problem solving
- Fair & simple pricing model

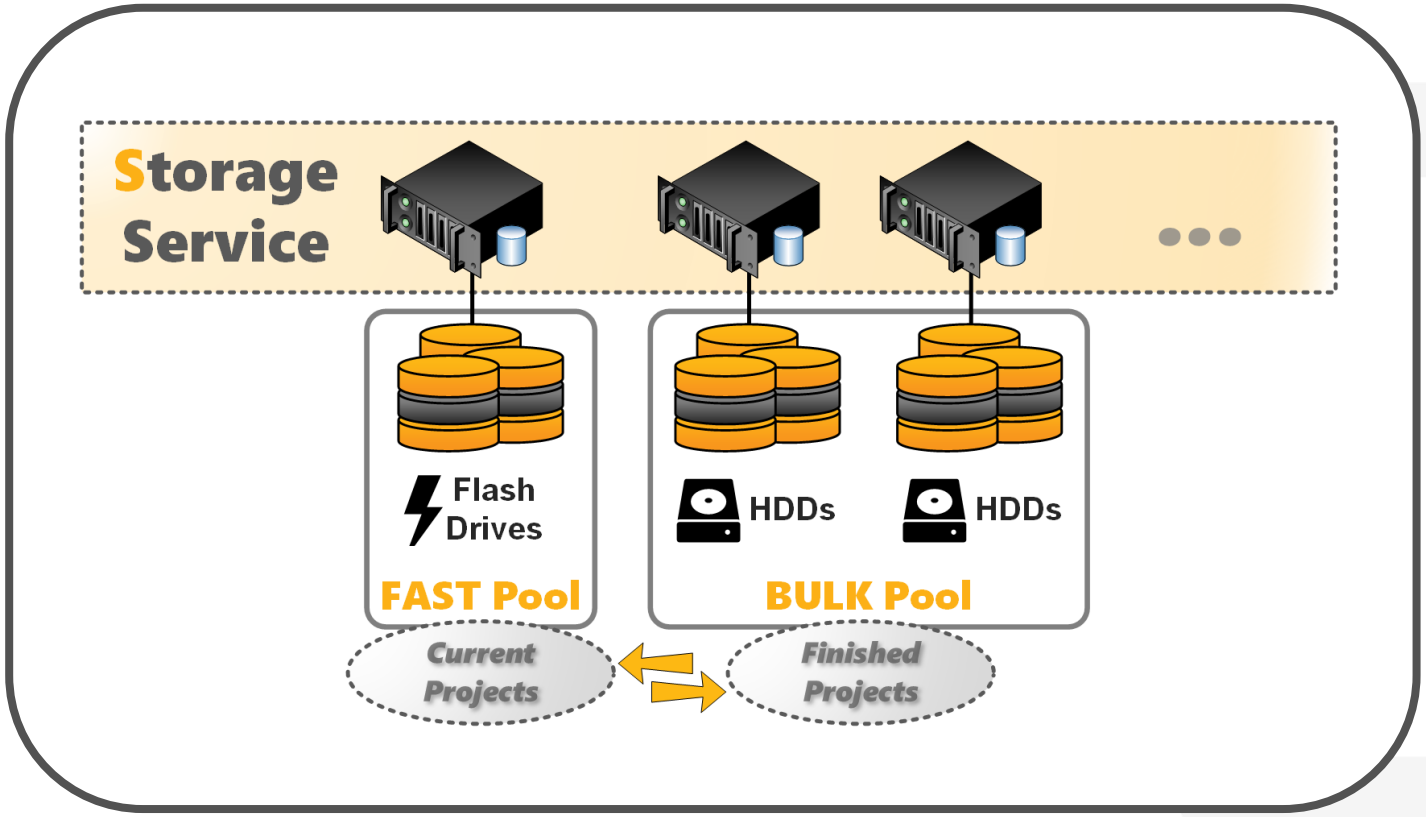
BeeGFS v6: Buddy Mirroring



New in BeeGFS v6:

- 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.

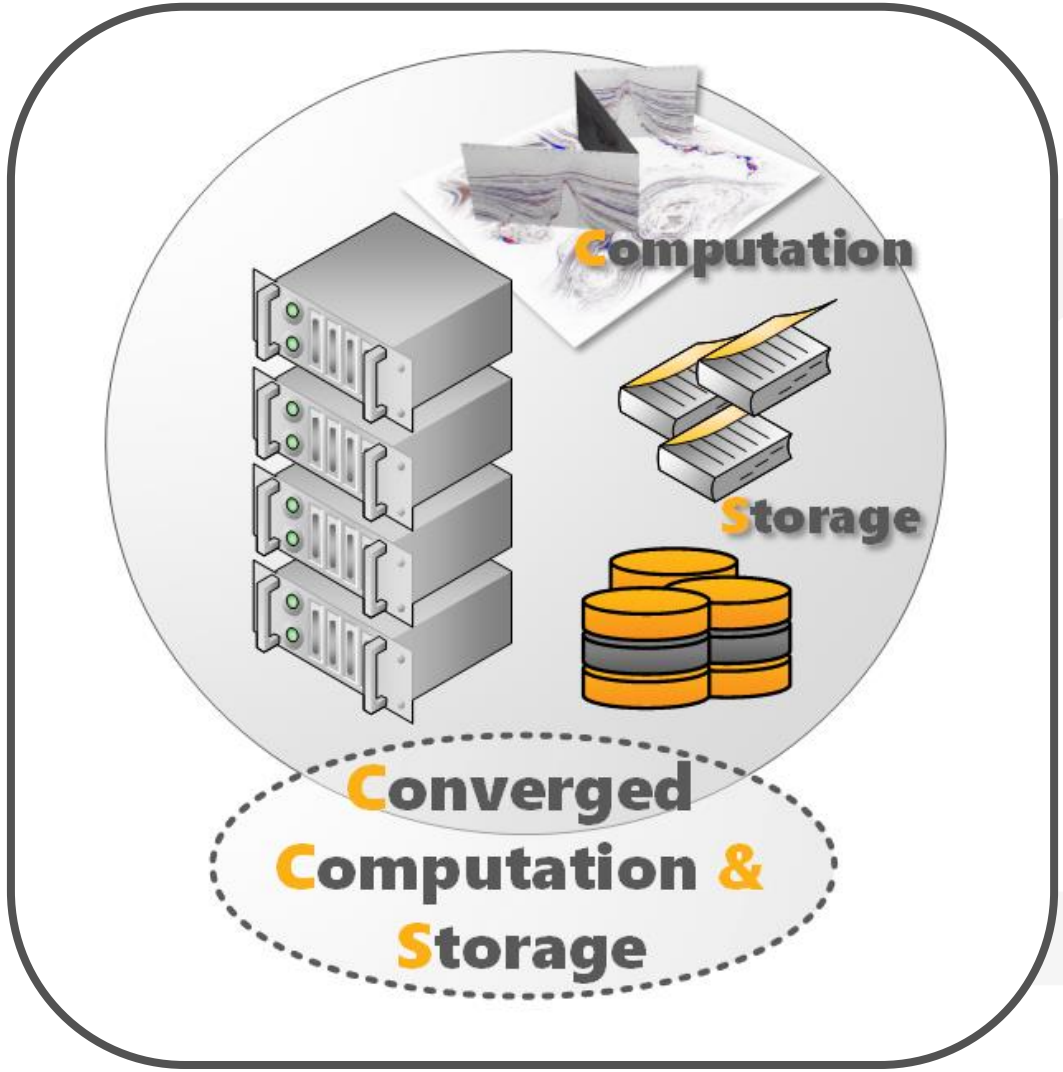
BeeGFS v7: Storage Pools



New in BeeGFS v7:

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

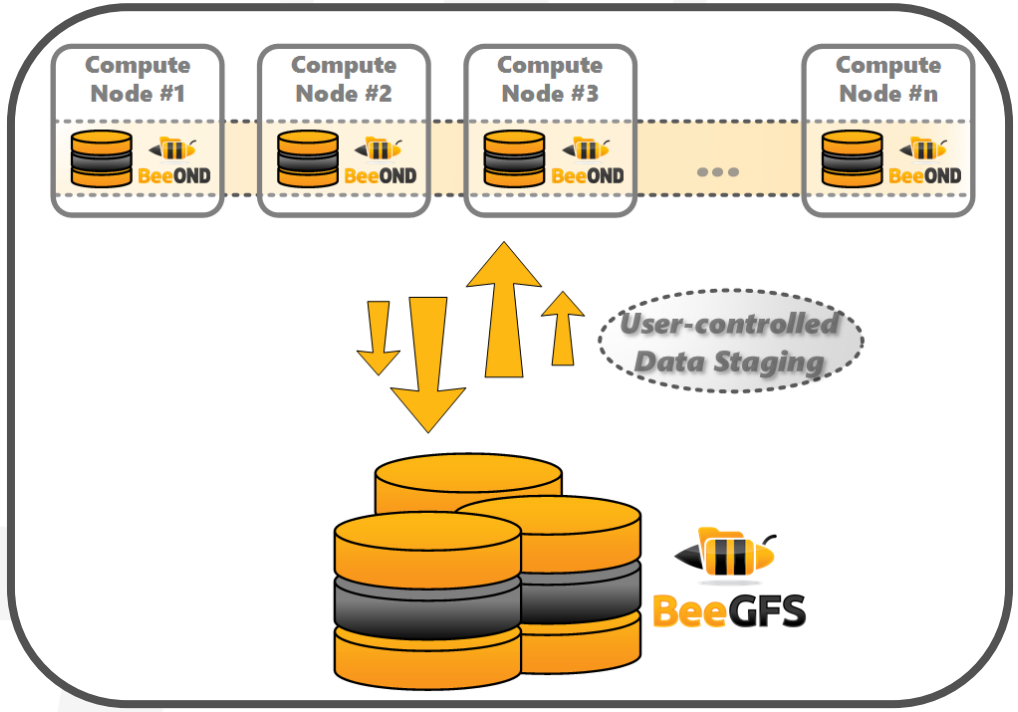
Storage + Compute: Converged Setup



Compute nodes as storage servers for small systems

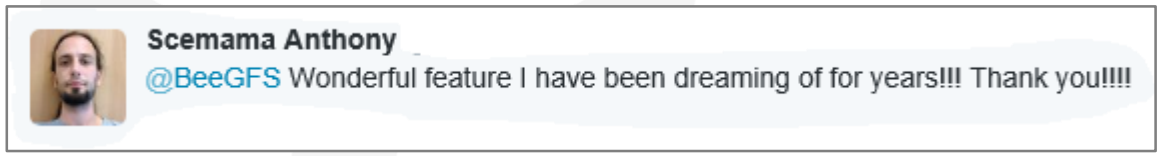
BeeOND – BeeGFS On Demand

- **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. Univa Grid Engine)**
- **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 "nasty" I/O patterns

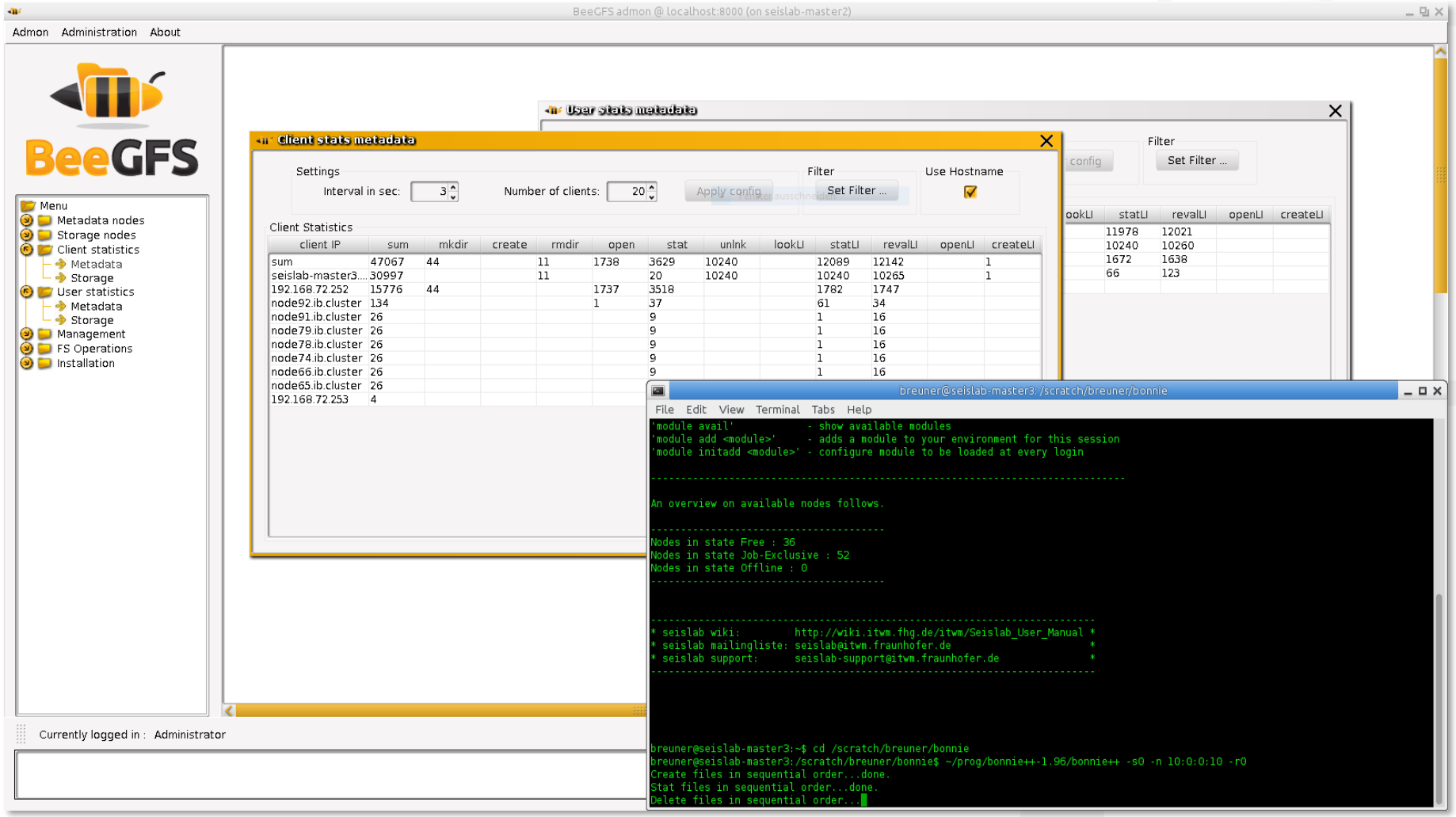


The easiest way to setup a parallel filesystem…


```
# GENERAL USAGE…
$ beeond start -n <nodefile> -d <storagedir> -c <clientmount>

----------------------------------------------

# EXAMPLE…
$ beeond start -n $NODEFILE -d /local_disk/beeond -c /my_scratch

Starting BeeOND Services…
Mounting BeeOND at /my_scratch…
Done.
```



Live per-Client and per-User Statistics

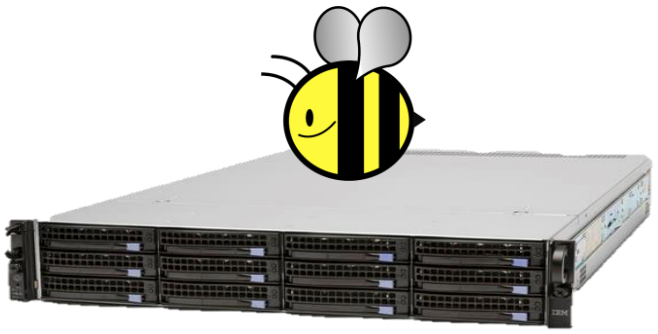


Typical Hardware for BeeGFS

Different building blocks for BeeGFS

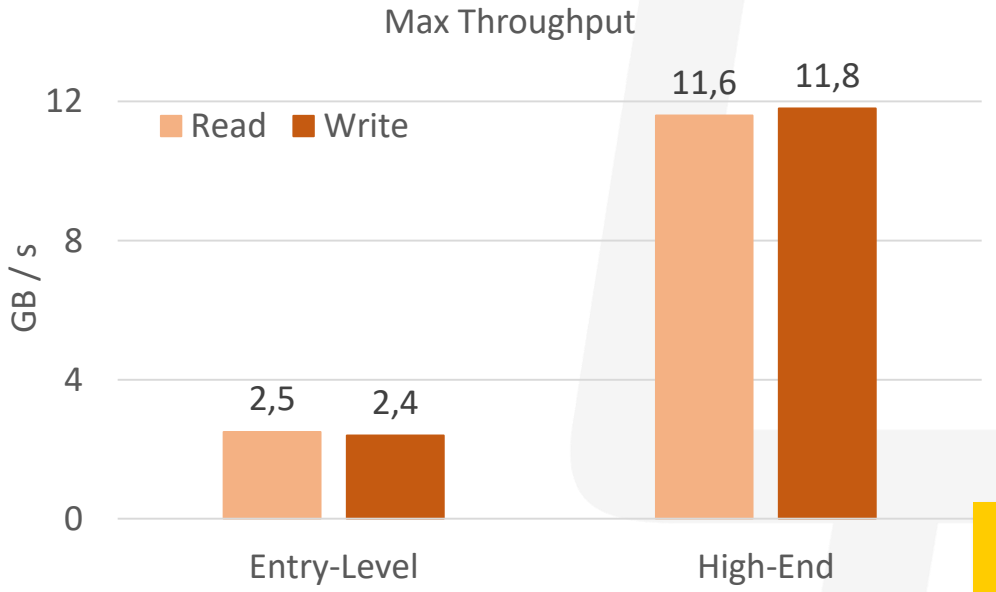
➢ Entry-Level Building Block

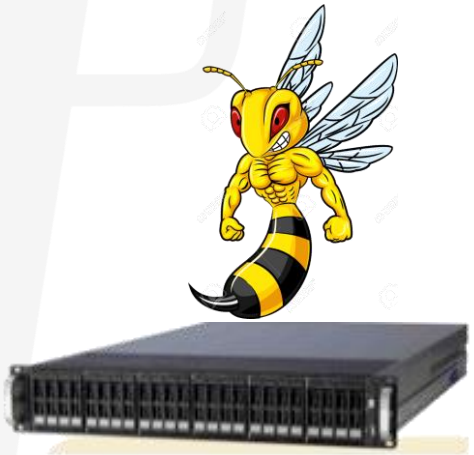
- 8 CPU Cores @ 2GHz
- 2 x 400 GB SSD for OS and BeeGFS Metadata
- 24x 4 TB HDDs
- 128 GB RAM



➢ High-End Building Block

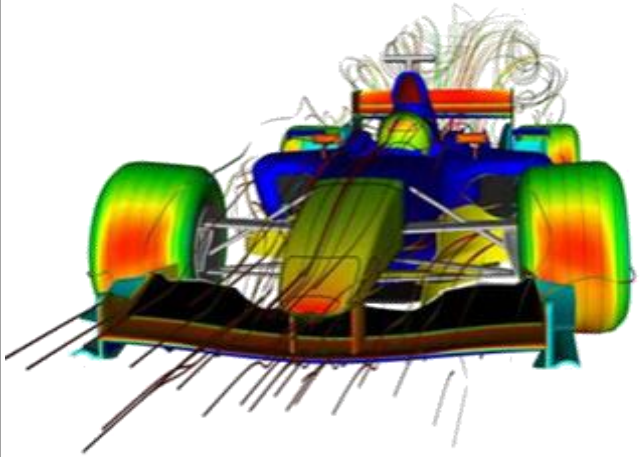
- 36 CPU Cores
- 24x NVMe drives with zfs
- 1 TB RAM





85% of BeeGFS customers use InifiniBand

Key Aspects



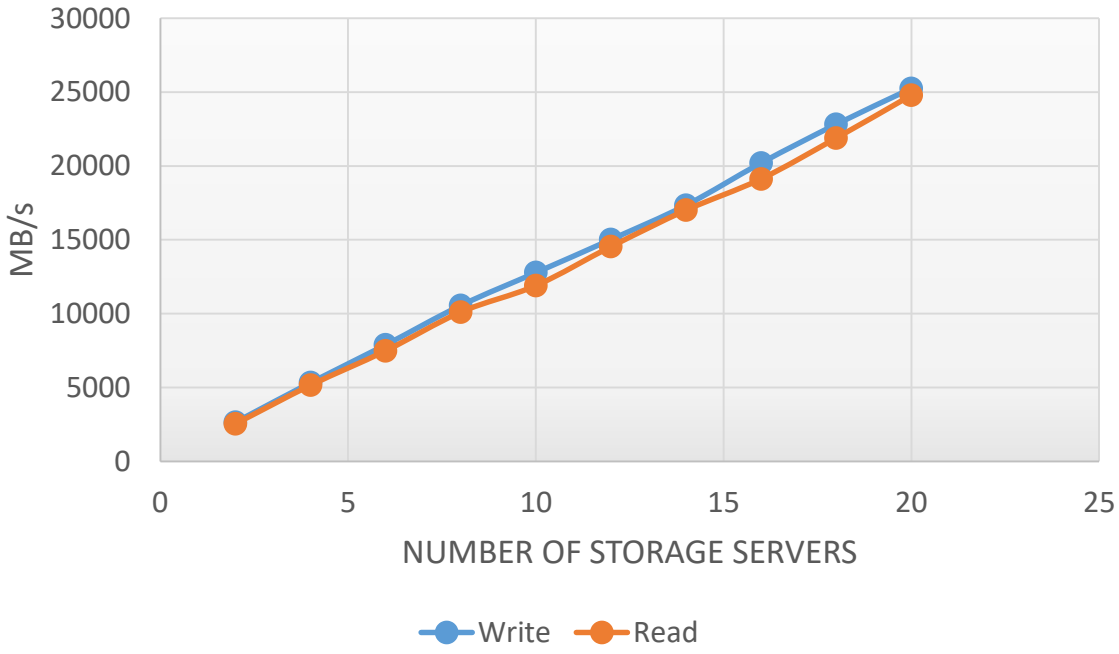
Maximum Performance & Scalability

High Flexibility

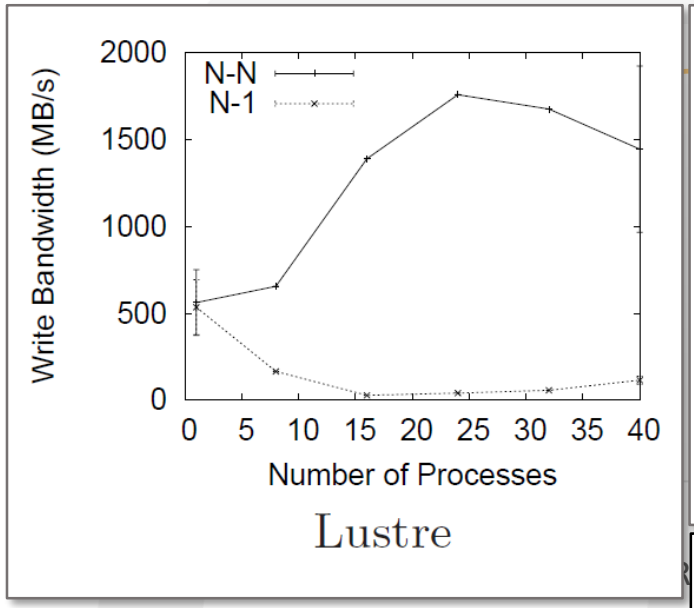
Robust & Easy to use

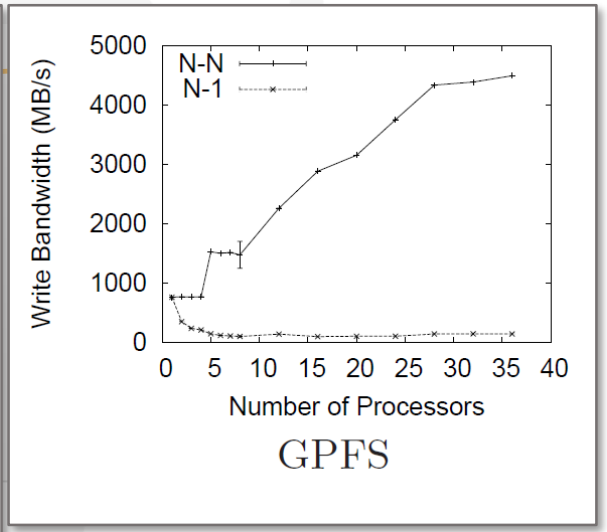
Linear throughput scalability for any I/O pattern

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



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

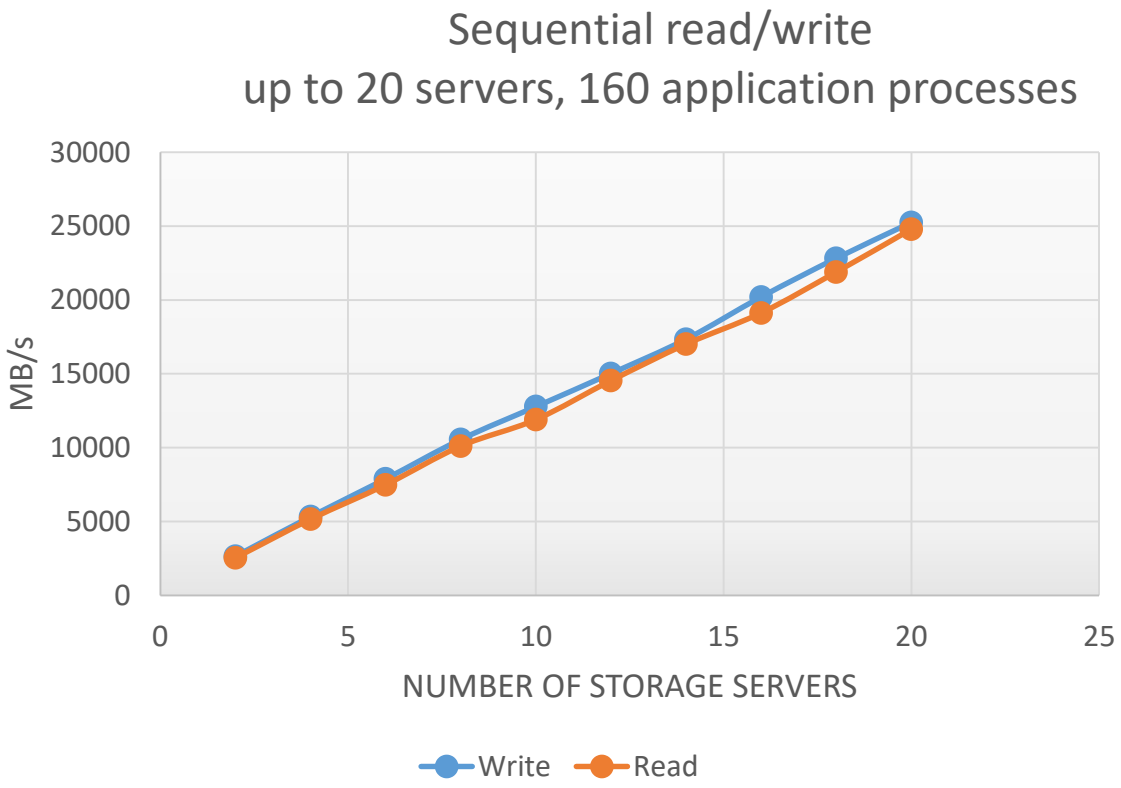


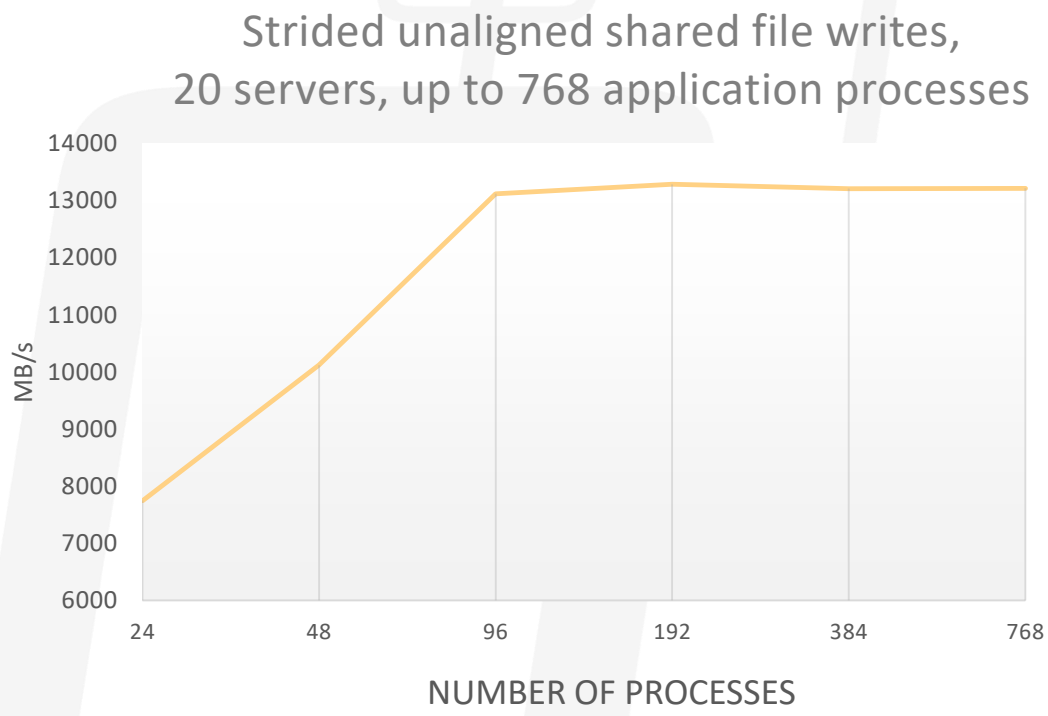


Shared File Write Throughput decreases with more processes for Lustre & GPFS

Source: John Bent,
Los Alamos National Lab,
PLFS Whitepaper

Linear throughput scalability for any I/O pattern





Key Aspects

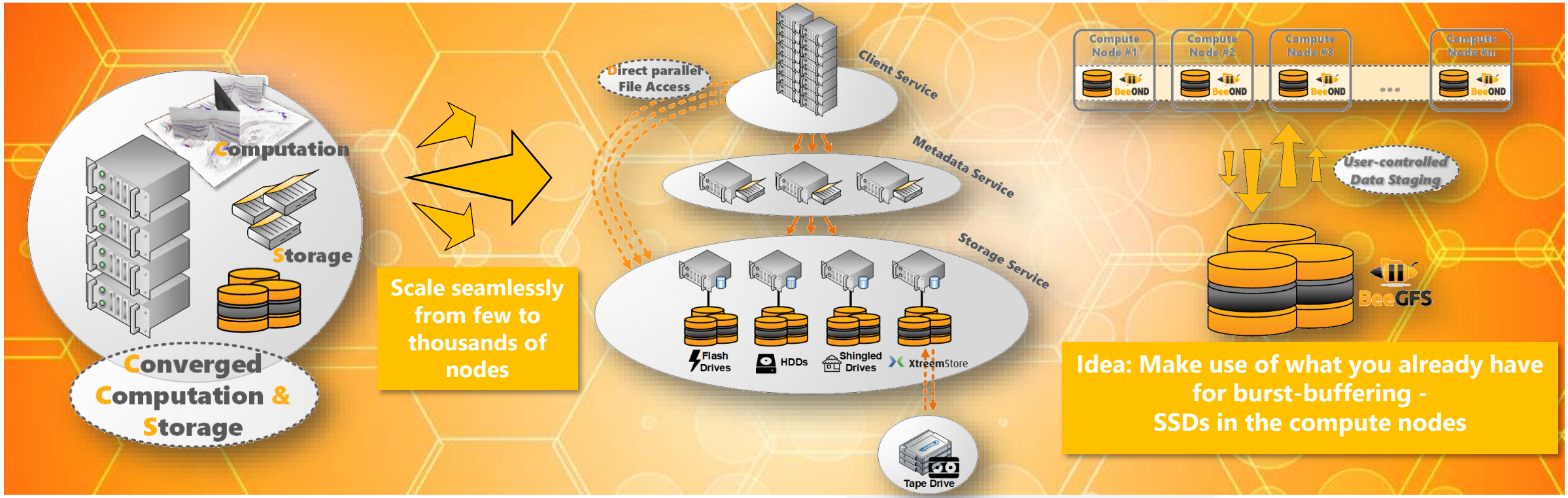
- ✓ **Performance & Scalability**
- ✓ **Flexibility**
- **Robust & Easy to use**
  - Very intensive suite of release stress tests, in-house production use before public release
    - The move from a 256 nodes system to a 1000 nodes system did not result in a single hitch, similar for the move to a 2000 nodes system.
  - 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 kernel patches
    - Kernel, system and BeeGFS updates are trivially simple
  - Packages for Redhat, SuSE, Debian and derivatives
  - Runs on shared-nothing HW
  - Graphical monitoring tool
    - Admon: Administration & Monitoring Interface

Key Aspects

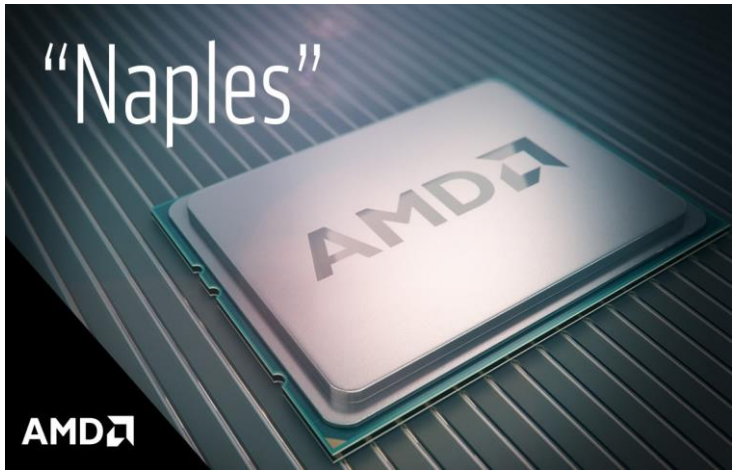
- ✓ **Performance & Scalability**
- **Flexibility**
  - Multiple BeeGFS services (any combination) can run together on the same machine
  - Flexible striping per-file / per-directory
  - Add servers at runtime
  - On-the-fly creation of file system instances (BeeOND)
  - Runs on ancient and modern Linux distros/kernels
  - NFS & Samba re-export possible
  - Runs on different Architectures, e.g.
    - …

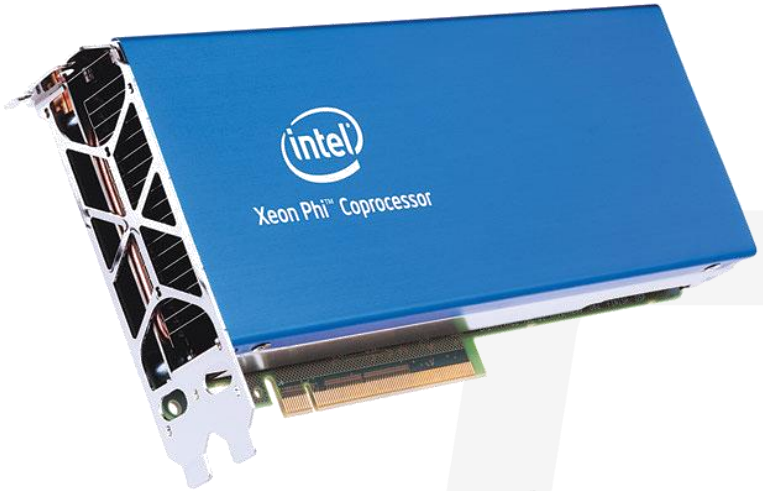
BeeGFS Flexibility Spectrum

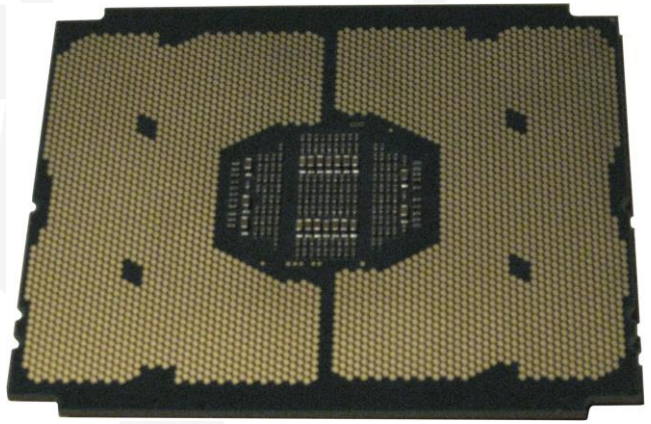
BeeGFS Is Designed For All Kinds And Sizes:

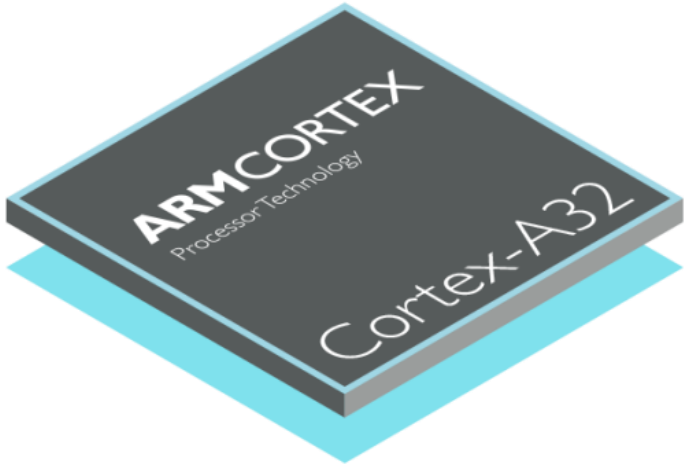


Flexibility: CPU Architectures



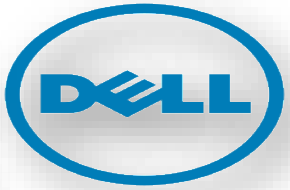


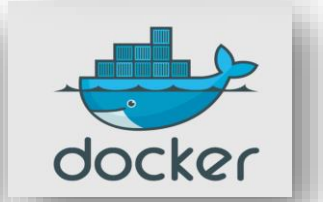


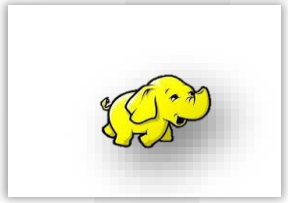


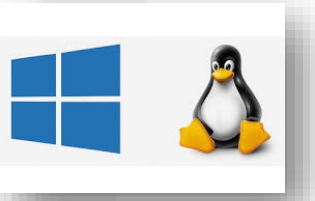
OpenPOWER™

Installations, certified vendors, plugins & integrations

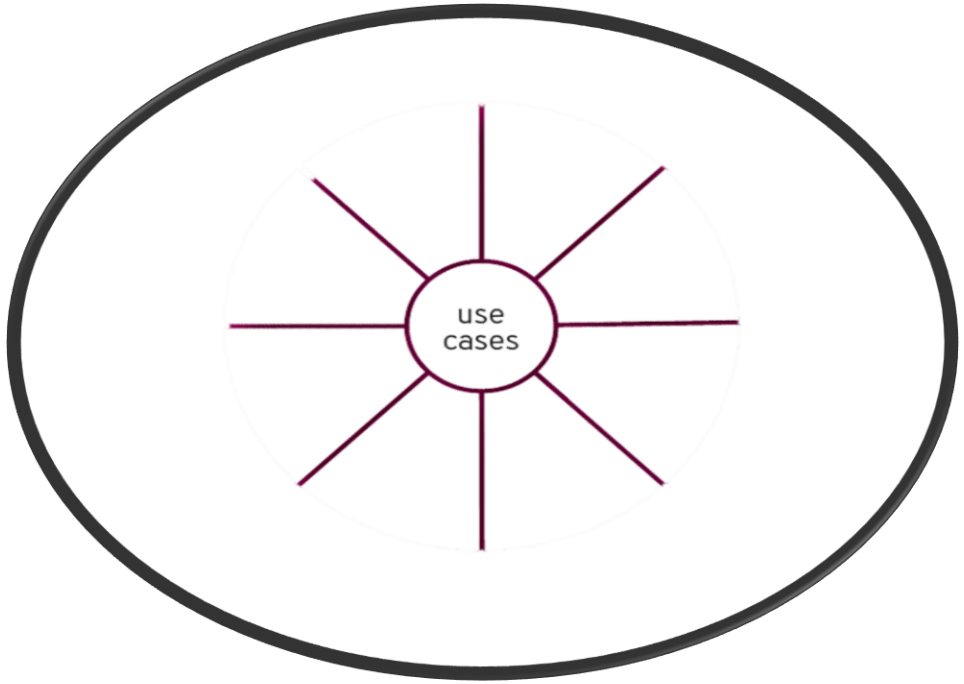


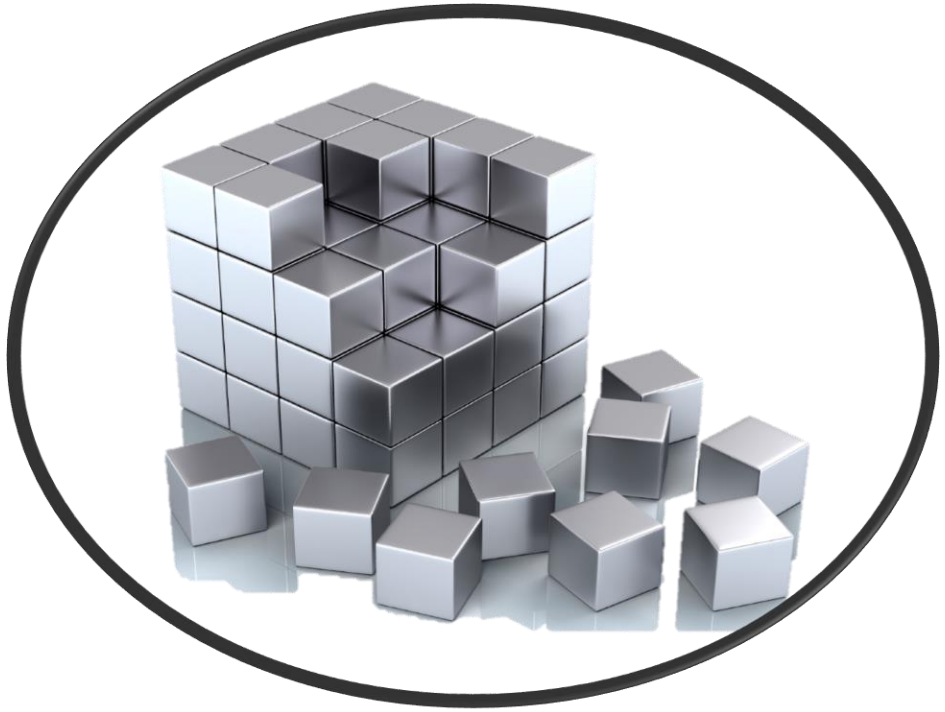






Use Cases & Building blocks





Test Result @ Juron OpenPower NVMe Cluster

- **<u>The cluster consists of 18 nodes</u>. Each node had the following configuration:**
  - 2 IBM POWER8 processors (up to 4.023 GHz, 2x 10 cores, 8 threads/core)
  - 256 GB DDR4 memory
  - Non-Volatile Memory: HGST Ultrastar SN100 Series NVMe SSD, partitioned:
    - partition of 745 GB for storage
    - partition of 260 GB for metadata
  - 1 Mellanox Technologies MT27700 Family [ConnectX-4], EDR InfiniBand.
    - All nodes were connected to a single EDR InfiniBand switch
  - CentOS 7.3, Linux kernel 3.10.0-514.21.1.el7

- **<u>Results:</u>**
  - 2 weeks test time
  - Measured values are equivalent to 104% write & 97% (read) of NVMe manufacturer specs.
  - For file stat operations the throughput is increasing up to 100,000 ops/s per metadata server
  - 16 metadata servers, the system delivers 300,000 file creates & over 1,000,000 stat ops per sec.

  - **IO 500 Annual Winner Chart place # 4 with only 8 nodes** - have check on other setups being above #4 and let me know the price...

Test Result @ Juron OpenPower NVMe Cluster

- **The cluster consists of 18 nodes. Each node had the following configuration:**

#	information				io500		
	system	institution	filesystem	client nodes	score	bw	md
					sqrt(GiB*kIOP)/s	GiB/s	kIOP/s
1	Oakforest-PACS	JCAHPC	IME	2048	101.48	471.25	19.04
2	Shaheen	Kaust	DataWarp	300	70.90	151.53	33.17
3	Shaheen	Kaust	Lustre	1000	41.00	54.17	31.03
4	JURON	JSC	BeeGFS	8	35.77	14.24	89.81
5	Mistral	DKRZ	Lustre	100	32.15	22.77	46.64
6	Sonasad	IBM	Spectrum Scale	10	21.63	4.57	102.43
7	Seislab	Fraunhofer	BeeGFS	24	18.75	5.13	68.55

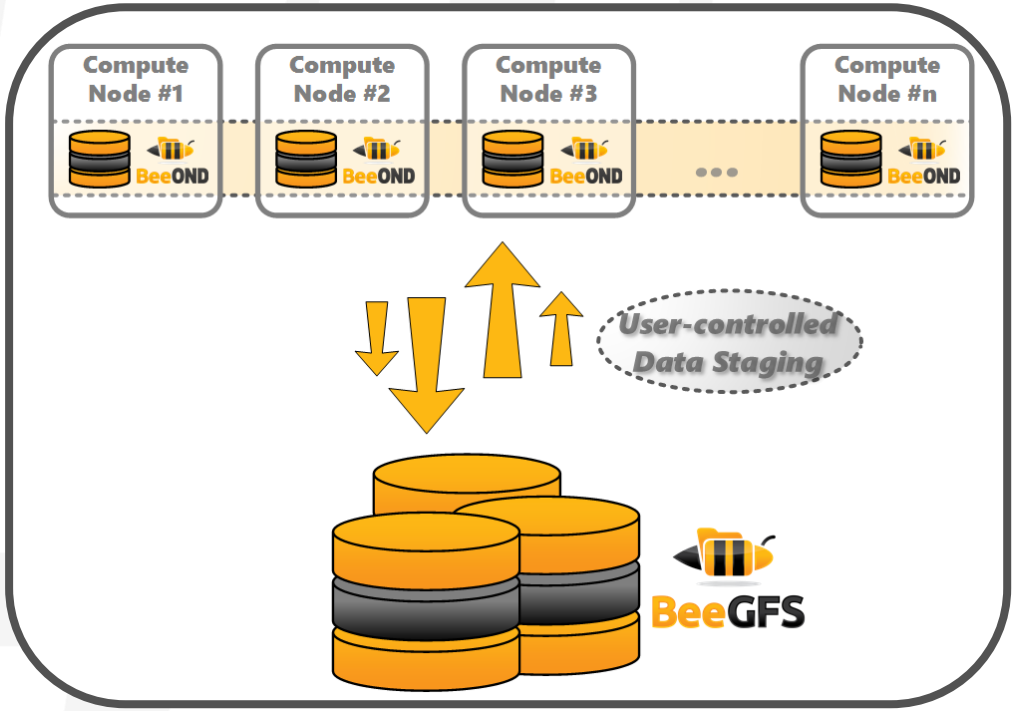
- **IO 500 Annual Winner Chart place # 4 with only 8 nodes** - have check on other setups being above #4 and let me know the price...

BeeOND at Alfred-Wegener-Institute

- **Example: Omni-Path Cluster at AWI**
  - Global BeeGFS storage on spinning disks (***4 servers @ 20GB/s***)
  - 300 compute nodes with a 500MB/s SSD each

    150GB/s aggregate BeeOND speed "for free"
- **Create BeeOND on SSDs on job startup**
  - Integrated into Slurm prolog/epilog script
- **Stage-in input data, work on BeeOND, stage-out results**

ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG



Cray CS400 at Alfred Wegner Institut

BeeGFS developed by Fraunhofer

BeeOND

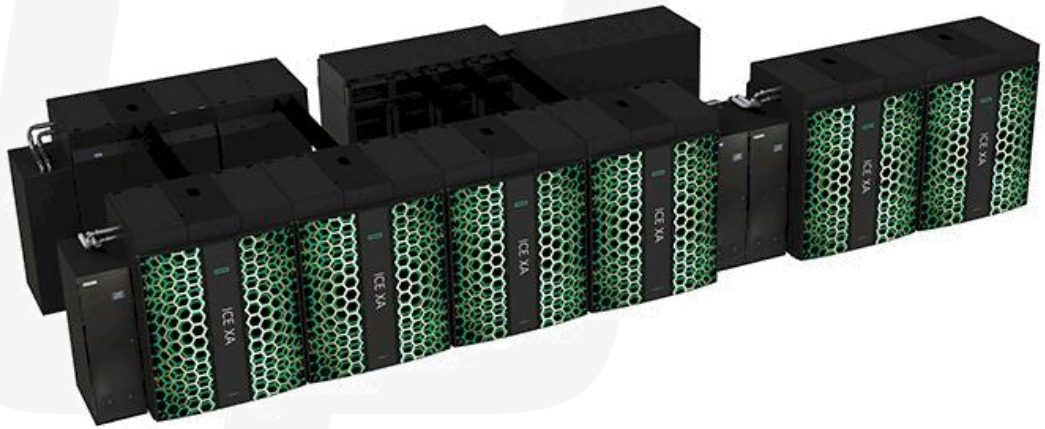
Broadwell CPU
Omnipath Interconnect
0,5 TB SSD in each node

BeeOND IOR 50TB	Stripe size 1, local	Stripe 4	Stripe size 1, any
308 Nodes write	160 GB/sec	161 GB/sec	160 GB/sec
308 Nodes read	167 GB/sec	164 GB/sec	167 GB/sec

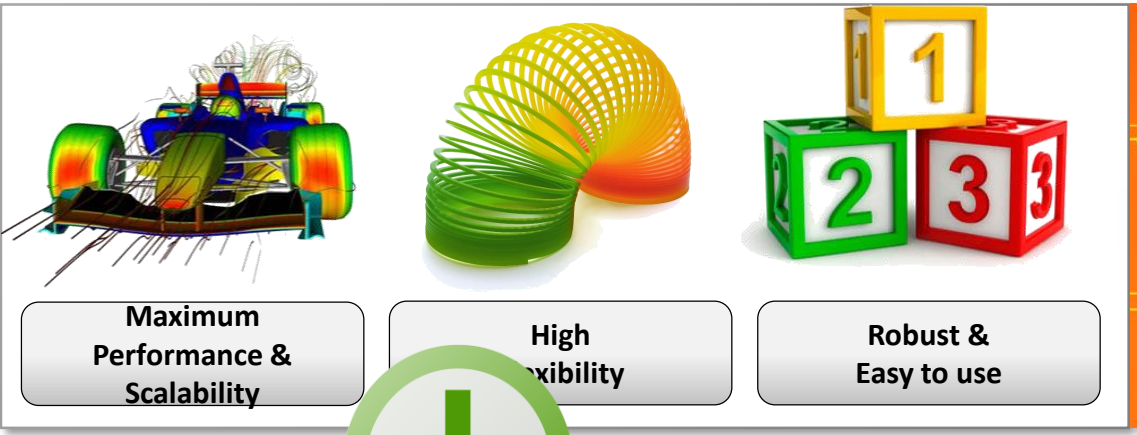
Omnipath Streaming Performance into BeeGFS :
1 Thread : 8,4 GB/sec 2 Threads : 10,8 GB/sec 4 Threads : 12,2 GB/sec

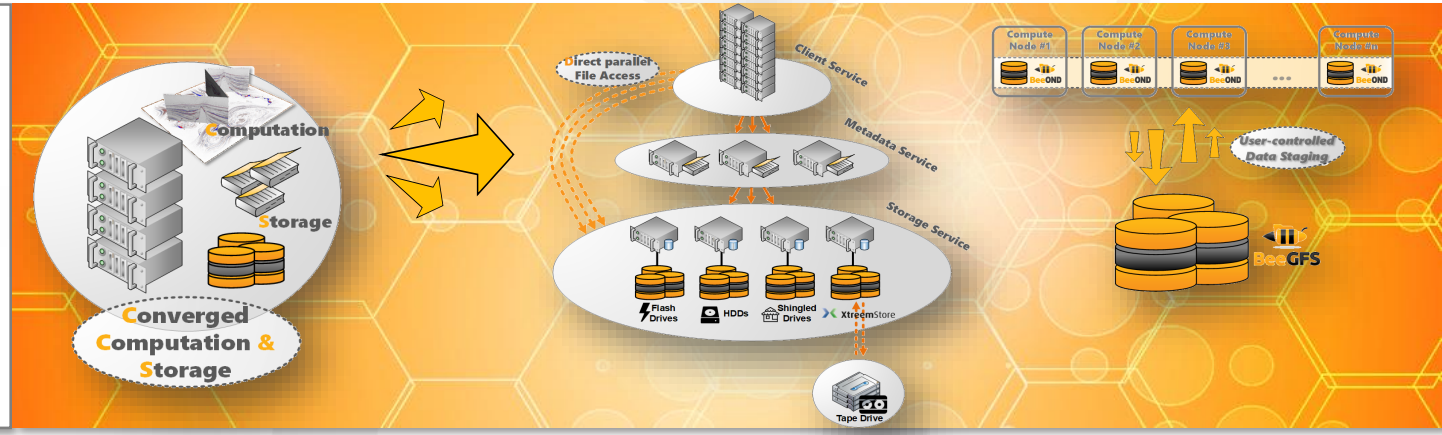
TSUBAME 3.0 in Tokyo

Tsubame 3.0 for hybrid AI & HPC with BeeOND on 1PB NVMe (raw device bandwidth >1TB/s)



BeeGFS Key Advantages





- Excellent performance based on highly scalable architecture
- Flexible and robust high availability options
- Easy to use and easy to manage solutions
- Suitable for very different application domains & IO pattern
- **Fair L3 pricing model (per storage node per year)**

Turnkey Solutions

Questions? // Keep in touch

- **Web** www.beegfs.io
  www.thinkparq.com
- **Mail** training@beegfs.io
  info@thinkparq.com
  support@beegfs.io
- **Twitter** @BeeGFS
- **LinkedIn** www.linkedin.com/groups/8476818
- **Newsletter** www.beegfs.io/news

marco.merkel@thinkparq.com