BeeGFS

BeeOND

BeeGFS.io

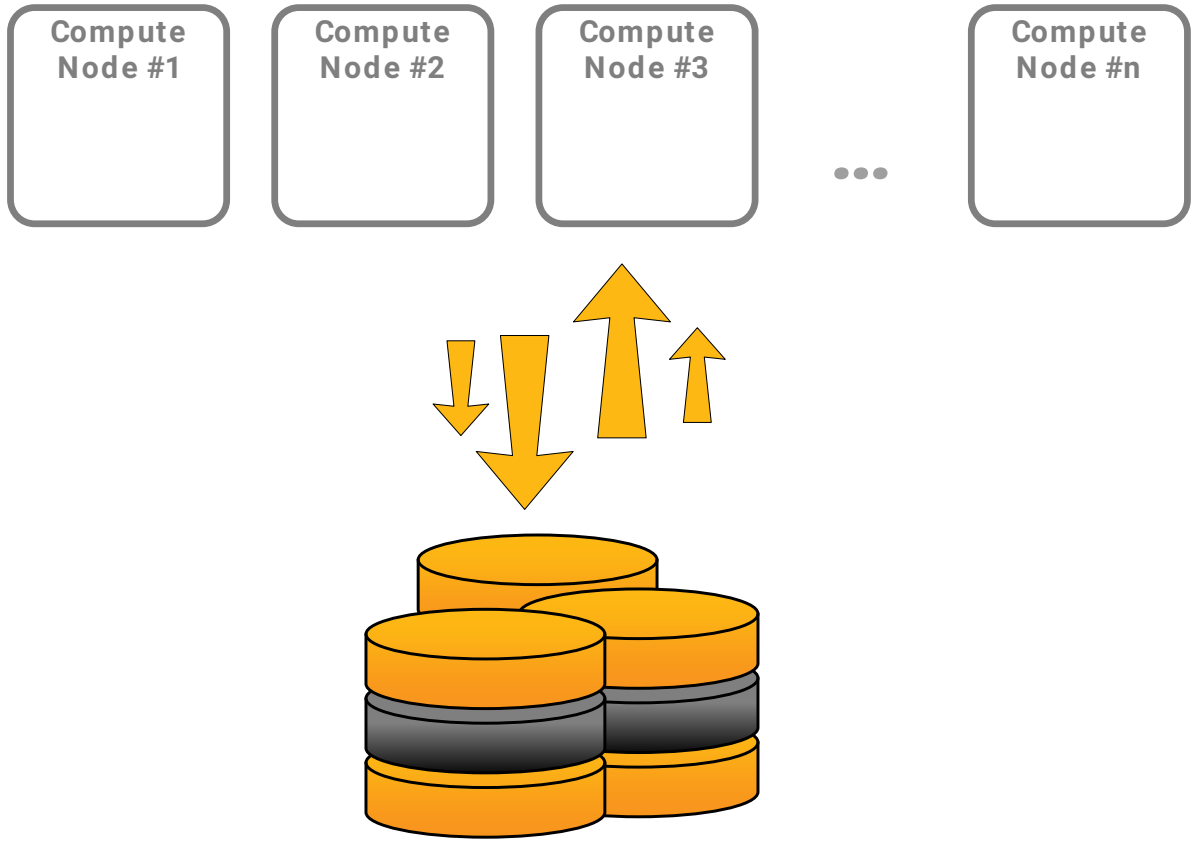
2022

Agenda

- **Basic Concepts**
  - Motivation
  - Overview
  - Common use cases
- **How it Works**
  - Commands
  - Basic setup
- **How we use BeeOND in our own Cluster**
  - Fraunhofer Seislab
  - Torque prologue and epilogue scripts
- **Benchmark Results**

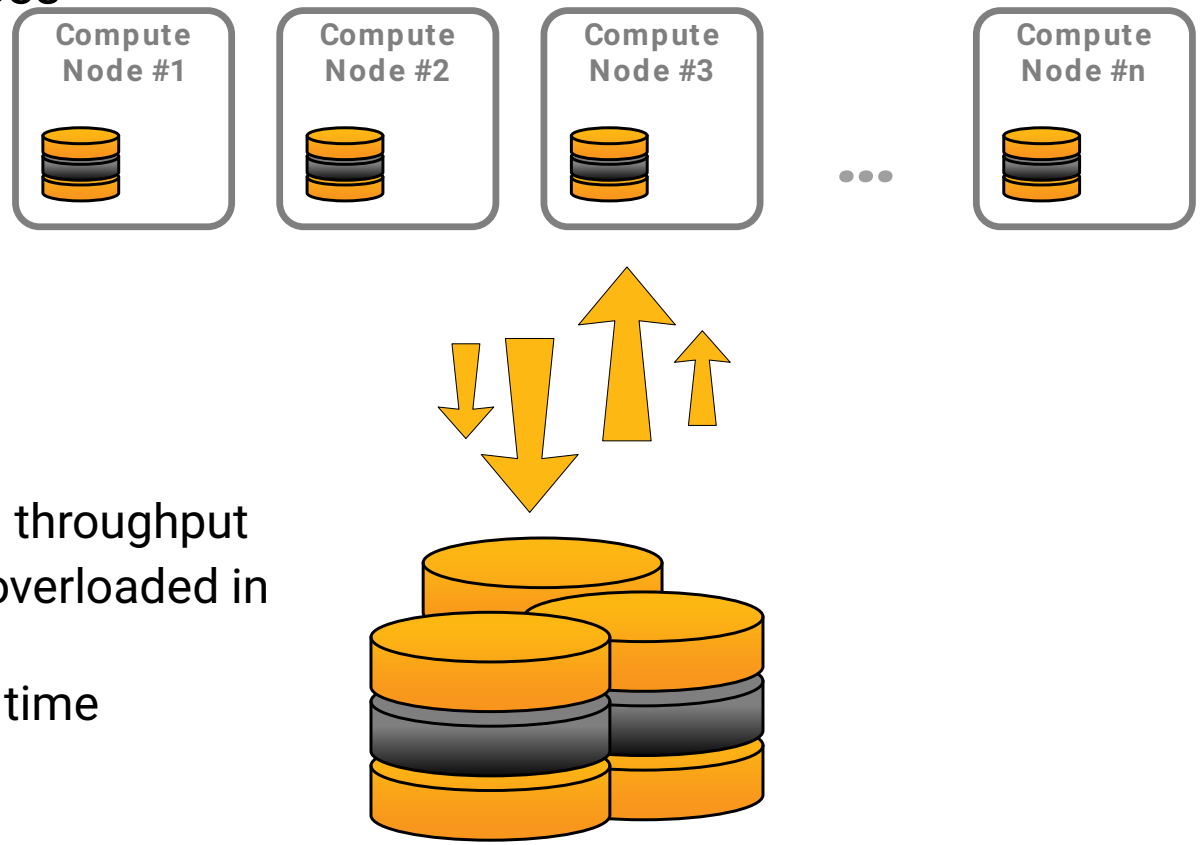
Motivation

- Typical HPC Cluster
  - Multiple compute nodes
  - One global file system
  - Problems
    - Global file system overload
    - Nasty IO patterns
    - Temporary files



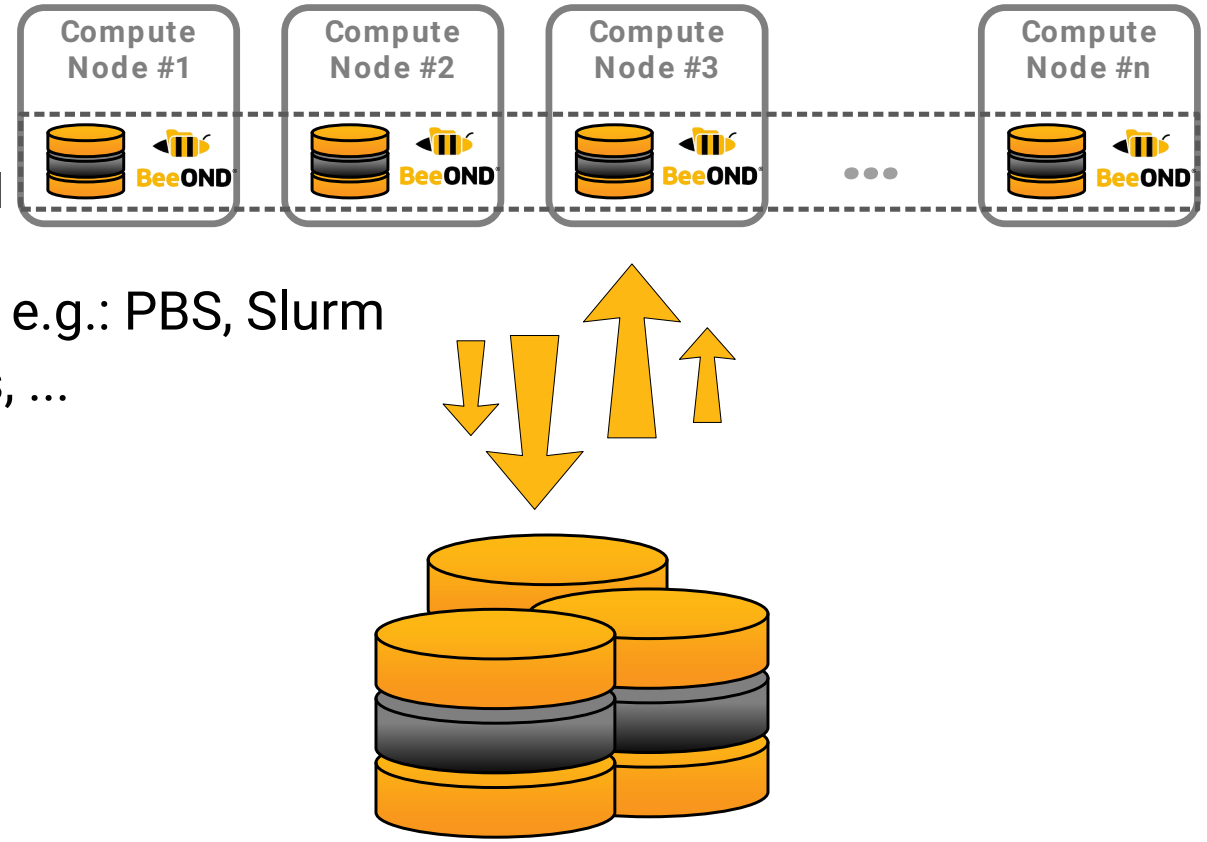
Motivation

- Typical HPC Cluster
  - Compute nodes with local SSDs or NVMe devices
    - Intermediate work storage tier
  - Jobs work on these local devices
    - Copy final results to the global file system
    - Discard temporary files
    - Speed up "dirty" I/O patterns
    - Take load from global storage
  - Problems
    - Local devices have limited storage capacity and throughput
    - SSDs may be idle in some compute nodes and overloaded in other nodes
    - Copying a large file to compute node SSDs take time



BeeOND – What is it?

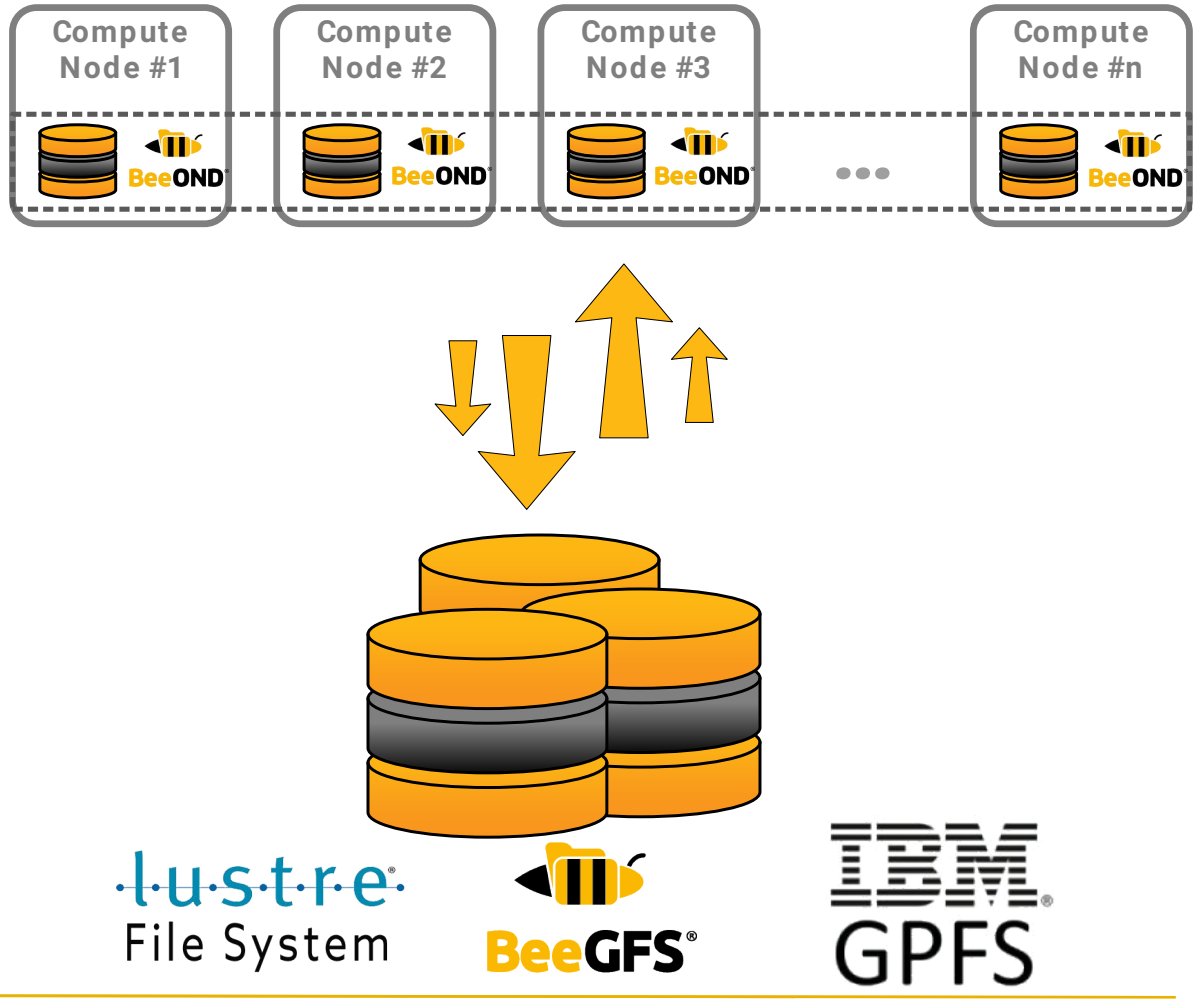
- Characteristics
  - Create a parallel file system instance on-the-fly
  - Common use case: "per-job parallel file system"
    - Aggregate the performance and capacity of local SSDs/disks in compute nodes of a job
  - Can be integrated into the cluster batch system, e.g.: PBS, Slurm
  - Other use cases: cloud computing, test systems, ...



BeeOND – What is it?

- Characteristics
  - Wrapper of BeeGFS
  - All BeeGFS features available
    - Buddy Group Mirroring
    - Data striping
    - Etc.
  - Simpler to use

- Global file system can be powered by
  - Other technologies (Lustre, GPFS, etc)
  - BeeGFS (ideally)



BeeOND – How do I get it?

- Package part of BeeGFS repository

```
root@compute-node01 ~>yum install beeond
```

- Startup with a single command line
  - nodefile: contains all hostnames to run BeeOND on (one host per line)
  - /data/beeond: path, where BeeOND will save its raw data on each node
  - /mnt/beeond: mountpoint of the filesystem on each node
  - /etc/beegfs/beeond: directory for additional tuning options

```
root@compute-node01 ~> beeond start -n nodefile -d /data/beeond -c /mnt/beeond -f /etc/beegfs/beeond -p 500
```

- Stopping BeeOND

```
root@compute-node01 ~> beeond stop –n nodefile -d
```

BeeOND – Security

- BeeOND uses the same security features like BeeGFS

```
root@compute-node01 ~>vi /etc/beegfs/AuthFile
```

- Create a subdirectory beeond in /etc/beegfs/

```
root@compute-node01 ~> mkdir -p /etc/beegfs/beeond
```

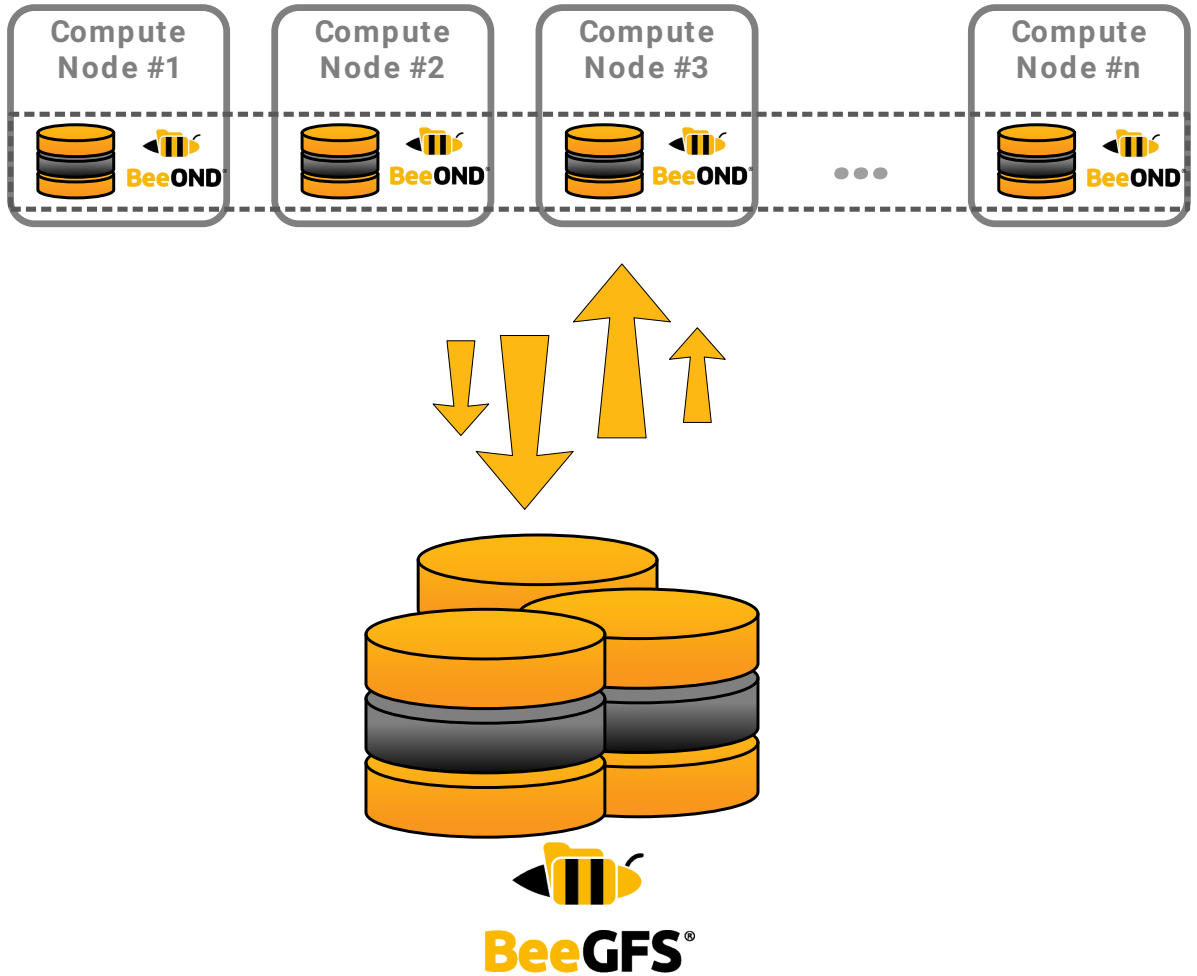
- Create *.conf files for all BeeGFS Services

```
root@compute-node01 ~> for j in mgmtd meta storage helperd client; do \
root@compute-node01 ~> echo „connAuthFile  = /etc/beegfs/AuthFile > beegfs-${j}.conf; done
```

- For BeeGFS 7.3.1 and 7.2.7 add connAuthFile to /etc/beegfs/beegfs-meta.conf and /etc/beegfs/beegfs-storage.conf

Use in Fraunhofer ITWM

- Fraunhofer Seislab
  - In-house cluster of CC-HPC at Fraunhofer ITWM
  - 92 compute nodes with 1 TB of SSDs each
  - Global BeeGFS system with 10 servers, SSD and SATA drives
- BeeOND Integrated into Torque startup
  - Create BeeOND on SSDs on job startup
- Job workflow
  - Stage-in input data
  - Work on BeeOND
  - Stage-out results



Example: Integrate into Torque

BeeGFS®

- Torque Prologue Script

```
sub startBeeOND($jobid, $username)
{
  my $hostfile = $ENV{"PBS_NODEFILE"};
  my $nodefile = $hostfile;
  $hostfile =~ s/\s//g;
  if ($hostfile =~ /(.+)[0-9]+/) {
    my $mountpoint = "/scratch_local";
    my $datapath = "/data/beegfs_ondemand";
    my $cmd = "/opt/beegfs/sbin/beeond stop -n $nodefile -i /data/fod_status -d -L -P -c -q -b /usr/local/bin/pdsh";
    print "$cmd\n";
    system("$cmd");
    my $cmd = "/opt/beegfs/sbin/beeond start -F -n $nodefile -i /data/fod_status -d $datapath/$username
          -c $mountpoint -p 1000 -P -b /usr/local/bin/pdsh -f /etc/beegfs/beeond/";
    print "$cmd\n";
    system("$cmd");
    my $chown_cmd="chown -R $username $mountpoint";
    system("$chown_cmd");
    my $chmod_cmd="chmod -R 770 $mountpoint";
    system("$chmod_cmd");
    print "Mounted BeeOND at $mountpoint.\nPlease be aware that all data in $mountpoint will be deleted at the end of the job.\n";
  }
}
```

Example: Integrate into Torque

- Torque Epilogue Script

```
sub stopFhgfsOndemand($jobid)
{
  my $nodefile = $ENV{PBS_NODEFILE};

  my $cmd = "/opt/beegfs/sbin/beeond stop -n $nodefile -i /data/fod_status -d -L -P -c -q -b /usr/local/bin/pdsh";
  system("$cmd");
}
```

Example Startup on Fraunhofer Seislab


```
[lucy@seislab-master ~]$ getnode -n 4
...
Please be aware that all data in /scratch_local will be deleted at the end of your job.
qsub: waiting for job 36659.master.global.cluster to start
qsub: job 36659.master.global.cluster ready

/opt/beegfs/sbin/beeond start -F -n /cm/local/apps/torque/var/spool/aux//36659.master.global.cluster -i /data/fod_status -d
/data/beegfs_ondemand/lucy -c /scratch_local -p 1000 -P -b /usr/local/bin/pdsh -f /etc/beegfs/beeond/
INFO: Using status information file: /data/fod_status
INFO: Checking PDSH availability on the following hosts: node17,node18,node19,node20
INFO: Number of storage servers automatically set to 4

INFO: Starting beegfs-mgmtd processes
INFO: Management daemon log: /var/log/beegfs-mgmtd_20170626-111433.log
INFO: Starting beegfs-mgmtd on host: node17
INFO: Starting beegfs-storage processes on the following hosts: node17,node18,node19,node20
INFO: Storage server log: /var/log/beegfs-storage_20170626-111433.log
INFO: Starting beegfs-meta processes on the following hosts: node17
INFO: Metadata server log: /var/log/beegfs-meta_20170626-111433.log
INFO: Starting beegfs-client processes on the following hosts: node17,node18,node19,node20
INFO: Client log: /var/log/beegfs-client_20170626-111433.log

Mounted BeeGFS at /scratch_local.
This BeeGFS is running on the SSD drives of your allocated nodes.
Please be aware that all data in /scratch_local will be deleted at the end of your job.
[lucy@node17 ~]$
```

Example on Fraunhofer Seislab

```
[lucy@node17 ~]$ mount
/dev/sda2 on / type ext4 (rw,noatime,nodiratime,user_xattr)
none on /proc type proc (rw,nosuid)
none on /sys type sysfs (rw)
none on /dev/pts type devpts (rw,gid=5,mode=620)
none on /dev/shm type tmpfs (rw)
/dev/md127 on /data type ext4 (rw,noatime,nodiratime,user_xattr,discard)
none on /proc/sys/fs/binfmt_misc type binfmt_misc (rw)
master:/cm/shared on /cm/shared type nfs (rw,rsize=32768,wsize=32768,hard,intr,nfsvers=3,addr=192.168.48.254)
nfsd on /proc/fs/nfsd type nfsd (rw)
sunrpc on /var/lib/nfs/rpc_pipefs type rpc_pipefs (rw)
beegfs_nodev on /scratch type beegfs (rw,_netdev,cfgFile=/etc/beegfs/beegfs-client.conf,)
beegfs_ondemand on /scratch_local type beegfs (rw,sysMgmtdHost=node17,connPortShift=1000,cfgFile=/etc/beegfs/beeond//beegfs-
client.conf)


[lucy@node17 ~]$ beegfs-ctl --listnodes --nodetype=storage --mount=/scratch_local
node17 [ID: 1]
node19 [ID: 2]
node20 [ID: 3]
node18 [ID: 4]
```

Parallel Copy Tool

- Command beeond-cp to stage-in and stage-out data

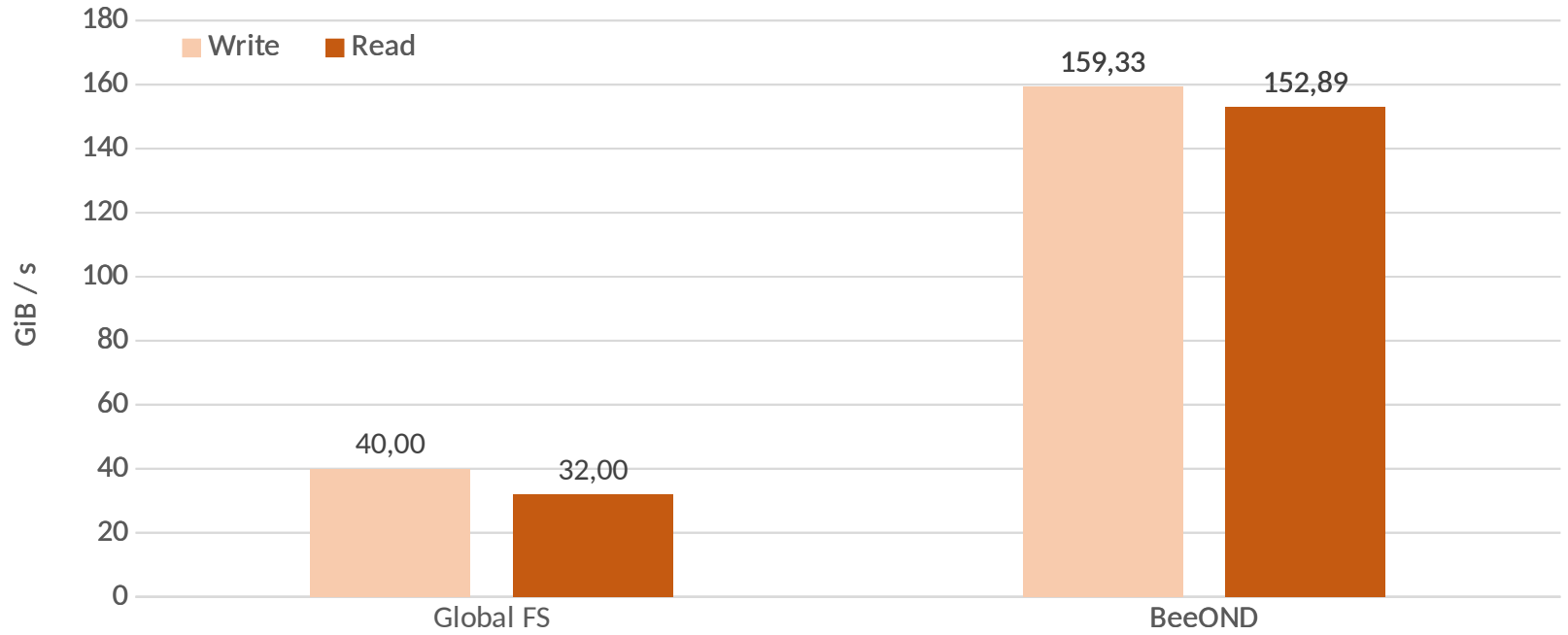
```
[lucy@node17 ~]$ beeond-cp copy –n $PBS_NODEFILE /scratch/lucy/simulations/ /scratch_local/
INFO: BeeOND copy...
INFO: Concurrency: 4
INFO: Nodes for parallel copy: node17, node19, node20, node18

INFO: Path to scan: /scratch/lucy/beegfs/
INFO: Scanning sources...

INFO: Generating target directory structure...
mkdir: created directory `/scratch_local/simulations
mkdir: created directory `/scratch_local/simulations/./2014
mkdir: created directory `/scratch_local/simulations/./2014/january
mkdir: created directory `/scratch_local/simulations/./2014/february
...
```

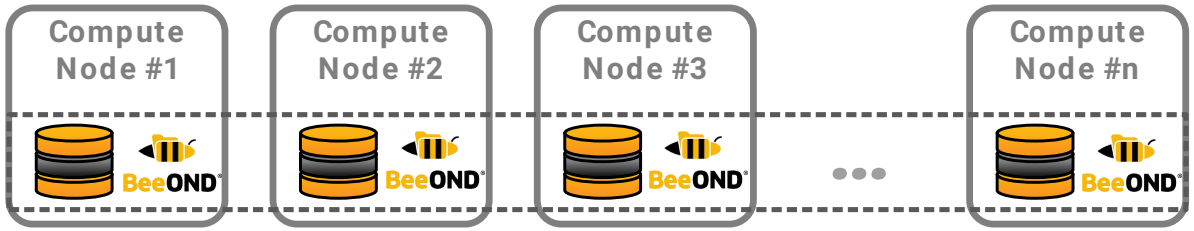
Benchmark Results

- System with 308 compute nodes, each with:
  - CPU: 2 x cores 2.3 GHz
  - RAM: 128 GB
  - 1 x 480 GB SSD



Homework

- How could BeeOND help increase performance of your customers' systems?
  - Are the SSDs of the compute nodes underused?
  - Are their global file systems overloaded?
  - How much more work could be done on their systems if BeeOND is used?
- Test BeeOND
  - Uninstall BeeGFS services from your test environment
  - Install BeeOND on them. Start, use and stop the service.
  - Try to use BeeGFS commands on your BeeOND installation.



What's Next?

- Typical Administration Task ✓
- Sizing & Tuning ✓
- BeeGFS Resiliency Tools ✓
- BeeOND: BeeGFS on Demand ✓
- BeeGFS Hive Index
- Conclusion

Thank You

Follow BeeGFS: