BeeGFS* in real life

Harry Mangalam, Research Computing, Univ California Irvine

harry.mangalam@uci.edu / 949 824 0084[o] / 949 478 4487[c]

* BeeGFS is a fully distributed parallel filesystem for Linux with parallel clients for Linux.
The context.

Unlike some of the systems & orgs here, we have:

- Limited $  
- Limited People  
- Limited Time  
- Lots of Demands, esp for:

Cheap*, Fast*, Reliable* Storage

(the oxygen of research computing, more than computation)
Why BeeGFS?

Good Judgment comes from Experience.

Experience comes from Bad Judgment.
Why BeeGFS

• GlusterFS :(
• Increase performance, reliability, esp under load.
• So, like Gluster, but it works
• Needed to be cheap.
• Needed to be fairly easy to set up and understand.
• Couldn't eat an FTE
• More resilient to IB upgrades than Lustre
Why we like BeeGFS

- Pretty good docs
- Easy to set up
- Uses vanilla hardware
- Transparent cost
- Trivial to expand
- Mirrored Metadata
- Mirrored directories
- No kernel mods, runs on any Linux distro
- Runs efficiently to >95% capacity.
Why we like BeeGFS (contd)

- Runs on top of ZFS, XFS, ext4
- OK with ZOTfiles; using /dev/shm for ZOTs
- Allows multiple FS instances to co-exist easily
- Allows re-export of FS, to feed desktop clients
- Pretty fast overall, decent recursive perf.
- Keeps working; failback is good → RDMA to TCP, IB → Ethernet, etc.
- Cheap, even with paid support.
- Developed by the Fraunhofer Inst.
Why we DO NOT like BeeGFS

- File locking doesn’t work correctly.
- Berkeley Checkpointing doesn’t run on it.
- Hardlinks across dirs don’t work (?).
- Doesn’t do Erasure Coding on the back end.
- Doesn’t talk to RobinHood directly yet*.
- Doesn’t have a RH-linked visualization tool (tho it does have a simple admin GUI).
OK, end of conclusions
What is a // Distributed FS?

- Stand-alone Network Attached Storage
  - Single RAID, NFS or SMB servers & appliances
- Storage Area Networks
  - Networked block devices aggregated into a FS (network SCSI)
- Distributed Filesystems
  - Distributed on the back end, but not on the client:
    - Isilon, Qumulo
- Fully // Distributed Filesystems
  - Distributed & // on back end & on client
    - Lustre, Gluster, GPFS, BeeGFS
Changing Role of // Distributed FS

- Hot data → intelligent cache flash devices which are getting larger.
- Cold data → cheap object storage.
- // FSs are (slowly) being demoted to storage between hot data and cold data.
- When you can’t afford the cost and overhead of a fully multi-tiered system, what do you do?
- Exploit your existing // FS to provide as much of that functionality as possible.
What should a DFS provide?

- Single namespace (or multiple single namespaces)
- Resilience or robustness
- Very high performance... and
- Increased performance with more servers
- Extreme Scalability
- Open Source preferred
- Reasonably priced Paid Support
- Reputation of supplier
Data Heterogeneity

- Uncompressed ASCII
- Most is unstructured (~80-95%)
- Hierarchical, self-describing formats (HDF5, netCDF4)
- Binary Data from instruments and CPU emulators
- Large compressed archives
- Enormous distribution of file sizes
- Which brings up...
Things we hate

- Zillions Of ...anything in a single dir.
- Dirs, Files, esp Tiny Files (ZOTfiles, usually the result of brilliant students who are bad programmers)
- core dumps and other digital detritus.

Which is why we rely on:

the RobinHood Policy Engine & automated threats requests.
The economics of storage argues against small systems.
Personal Storage

Fig 2a: Buying Storage personally
Social Storage

Fig 2b: Buying Storage in a Cluster
UCI Campus Storage Pool

// Filesystems optimized for..

DFS1: Hi IOPS on SSDs

DFS2: BigData streaming RW on large spinners

DFS3: Sensitive data on a protected, encrypted FS

Erasure-coded Archive

I/O Nodes (// Clients)

SMB

NFS

Web

Science DMZ: rclone, GridFTP

// Filesystems optimized for..

Firewall

Compute Clusters – each node in the cluster can be a // client if needed.
Back End

// Filesystems optimized for..

DFS1: Hi IOPS on SSDs

DFS2: BigData streaming RW on large spinners

DFS3: Sensitive data on a protected, encrypted FS


Erasure-Coded, Multi-tenant Object Archives

rcclone, web
Questions

... or

• Corrections?
• Accusations?
• Me Too's?
Harry Mangalam
Research Computing,
Univ of California, Irvine

harry.mangalam@uci.edu

949 824 0084[o]
949 478 4487[c]