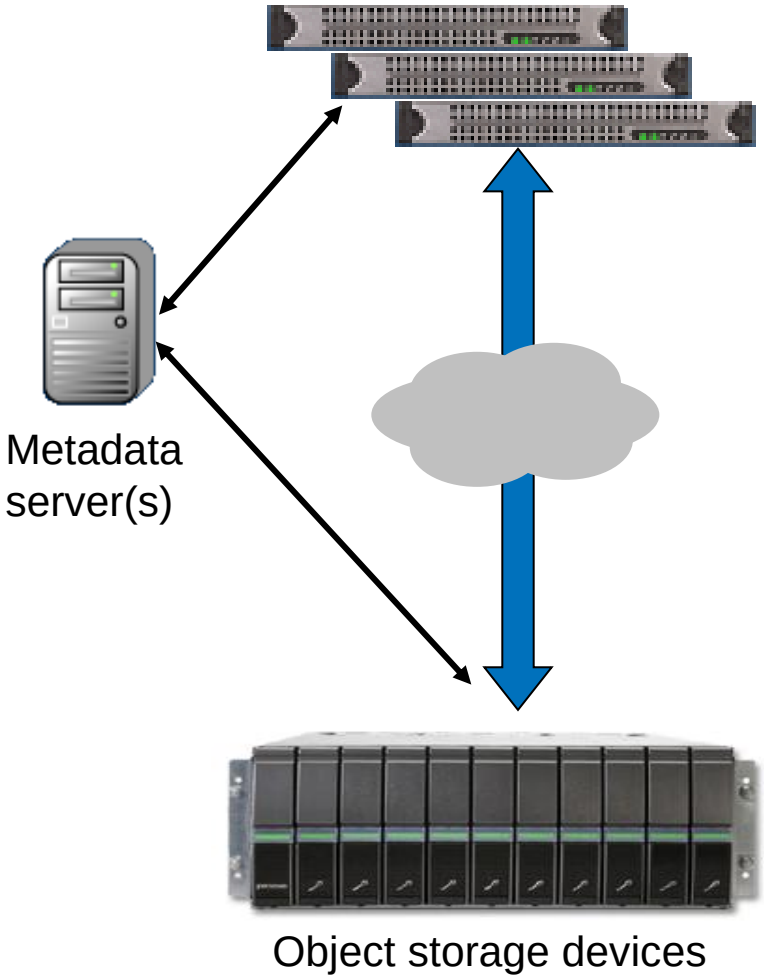
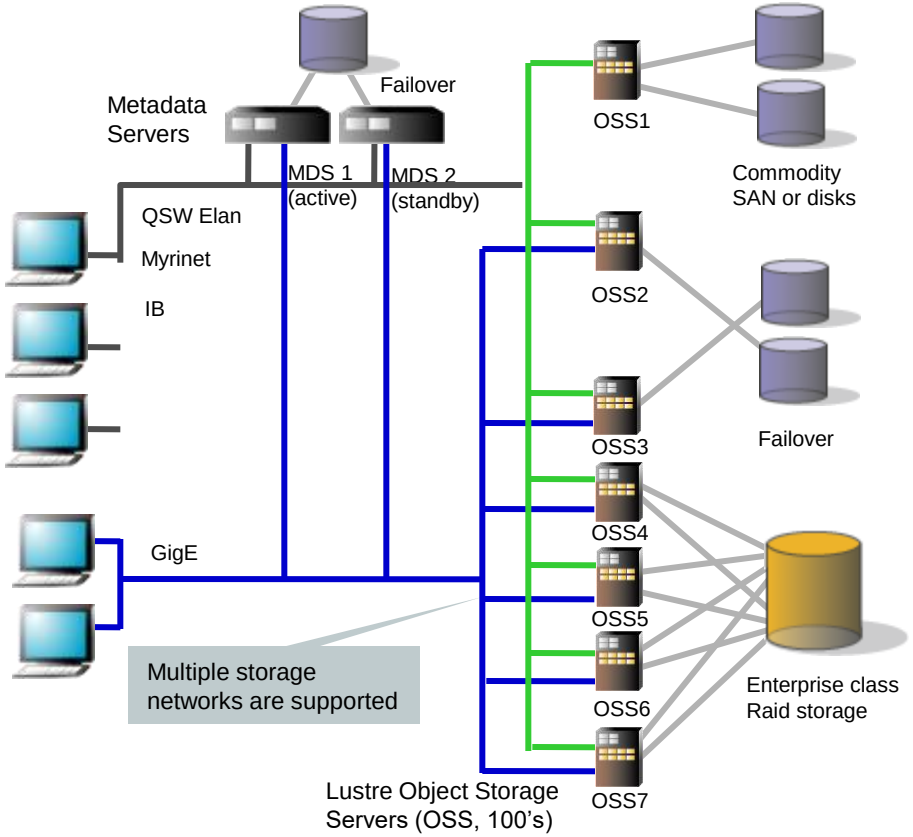
Object-based Storage Clusters

- Lustre, PanFS, Ceph, PVFS
- File system layered over objects
  - Details of block management hidden by the object interface
  - Metadata server manages namespace, access control, and data striping over objects
  - Data transfer directly between OSDs and object-aware clients
- High performance through clustering
  - Scalable to thousands of clients
  - 100+ GB/sec demonstrated to single filesystem



Lustre

- Open source object-based parallel file system
  - Based on CMU NASD architecture
  - Lots of file system ideas from Coda and InterMezzo
  - ClusterFS acquired by Sun, 9/2007
  - Sun acquired by Oracle 4/2009
  - Whamcloud aquired by Intel, 2012
- Originally Linux-based; Sun ported to Solaris
- Asymmetric design with separate metadata server
- Proprietary RPC network protocol between client & MDS/OSS
- Distributed locking with client-driven lock recovery



Lustre and GPFS Data Path

Lustre clients stripe data across Object Storage Servers (OSS), which in turn write data through a RAID controller to Object Storage Targets (OST). OST hides local file system data structures

GPFS has different metadata model but a similar data path
Control protocols to metadata servers are not shown