redhat.

Performance Overview of Samba backed by Gluster

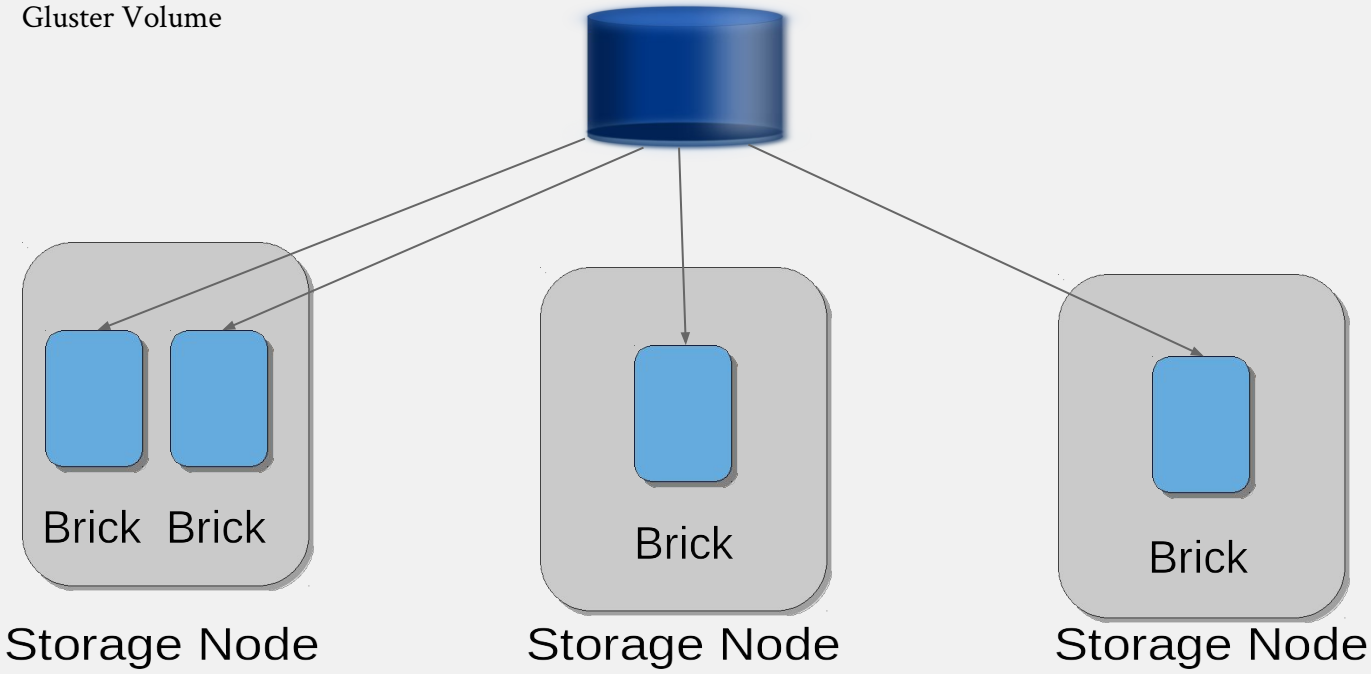
Poornima Gurusiddaiah <pgurusid@redhat.com>
Karan Sandha <ksandha@redhat.com>
Günther Deschner <gdeschne@redhat.com>

Gluster Overview

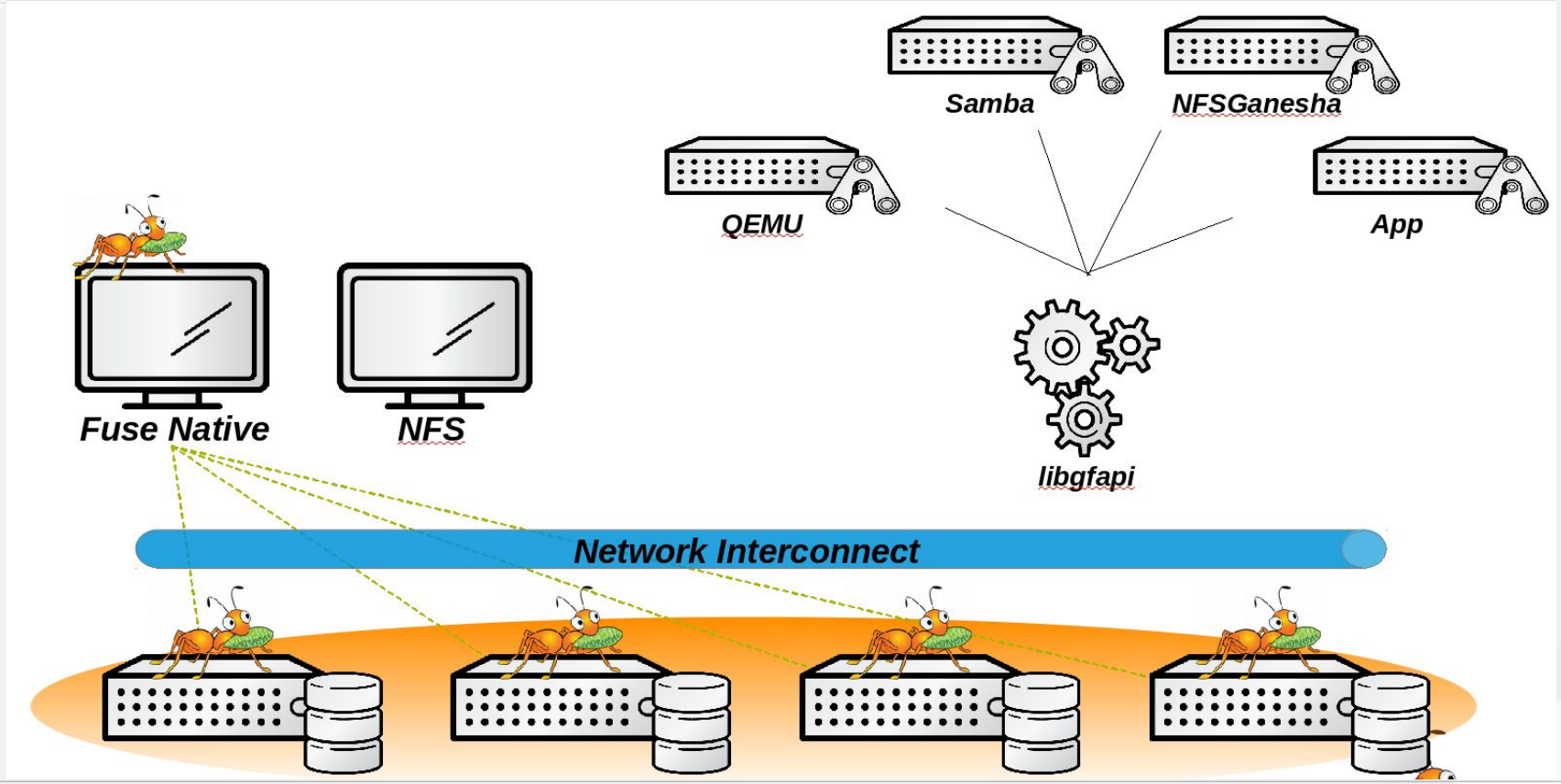
- Open-source general purpose scale-out distributed file system
- Aggregates storage exports over network interconnect to provide a single unified namespace
- Layered on disk file systems that support extended attributes

- No meta-data server
- Modular Architecture for Scale and Functionality
- Heterogeneous commodity hardware
- Scalable to petabytes & beyond

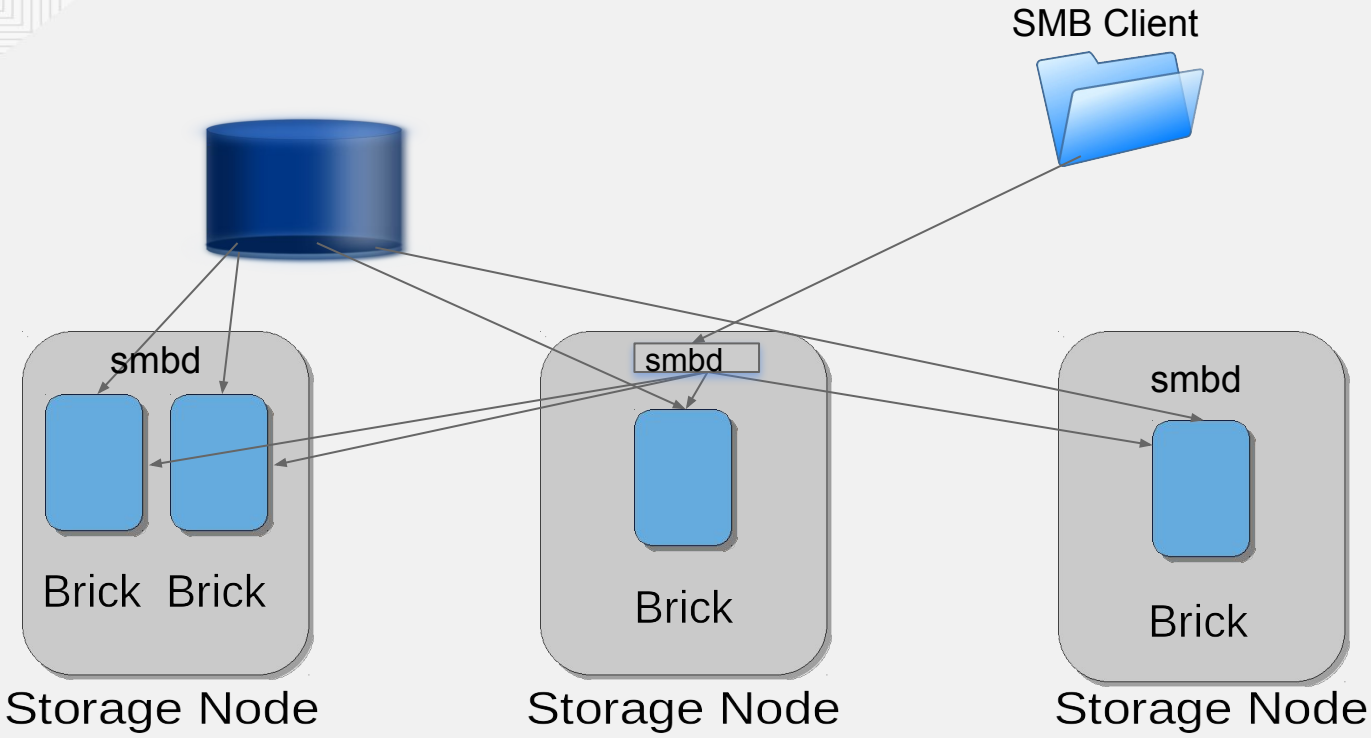
Gluster Overview



Gluster Access Mechanisms



SMB access of Gluster



Smb.conf tunables

kernel share modes = no
kernel oplocks = no
posix locking = no
map archive = no
map hidden = no
map read only = no
map system = no
store dos attributes = yes

Performance numbers as compared to other protocols

- **Large files:**

	SMB (KBps)	GNFS(KBps)	NFS Ganesha	FUSE client(KBps)
Sequential Read	2709019	2861377	xx	2569353
Sequential Write	1350991	1384350	xx	1706948
Random Read	634187	754273	xx	493328
Random Write	319453	303538	xx	376356

Performance numbers as compared to other protocols

- **Small files(64K):**

	SMB(Files/sec)	GNFS(Files/sec)	NFS Ganesha	FUSE client(Files/sec)
Create	90	540	XX	1204
Readdir recursive	502	6214	XX	6982
Rename	29	136	XX	163
Rmdir	209	830	XX	842

Performance numbers as compared to other protocols (contd.)

Large file performance is not bad as compared to other protocols. But the small file performance made it mostly unusable from SMB access. Multiple things contributed to the slow performance:

- One SMB request results in multiple filesystem calls (CREATE ~ 20-30 calls)
- Storing some attributes in XATTR results in lots of set/getxattr calls
- One CREATE results in lots of negative lookups(stat before file create)
- Case insensitive lookup of the filename in the directory
- Lots of stat calls
- Synchronous readdirp can block the poll thread in smbd for long time for large directories.

Performance numbers as compared to other protocols (contd.)

To address the performance issues, we tried couple of things:

- Implement aggressive caching in the Gluster client (loaded in smbd)
  - Md-cache : Cache stat and xattr info
  - Readdir-ahead : Prefetch readdir entries
  - Negative lookup-cache : Cache the negative dentries
- Leverage getrealfilename (in samba vfs)for case insensitive search

- **Small files(64K) with above mentioned improvements:**

SMB	**W/O improvements** (Files/sec)	**W improvements** (Files/sec)	**%age change**
Create	43	88	104 %
Listing	425	4287	900 %
Rename	20	35	75 %
Rmdir	174	252	44 %

WIP Performance Improvements

- Small file put-behind feature in Gluster client
- Make Gluster configurable to be Case insensitive, add option in samba to support backend case insensitive filesystem
- Async bulk readdir(p) call from samba?