QEMU-GlusterFS integration

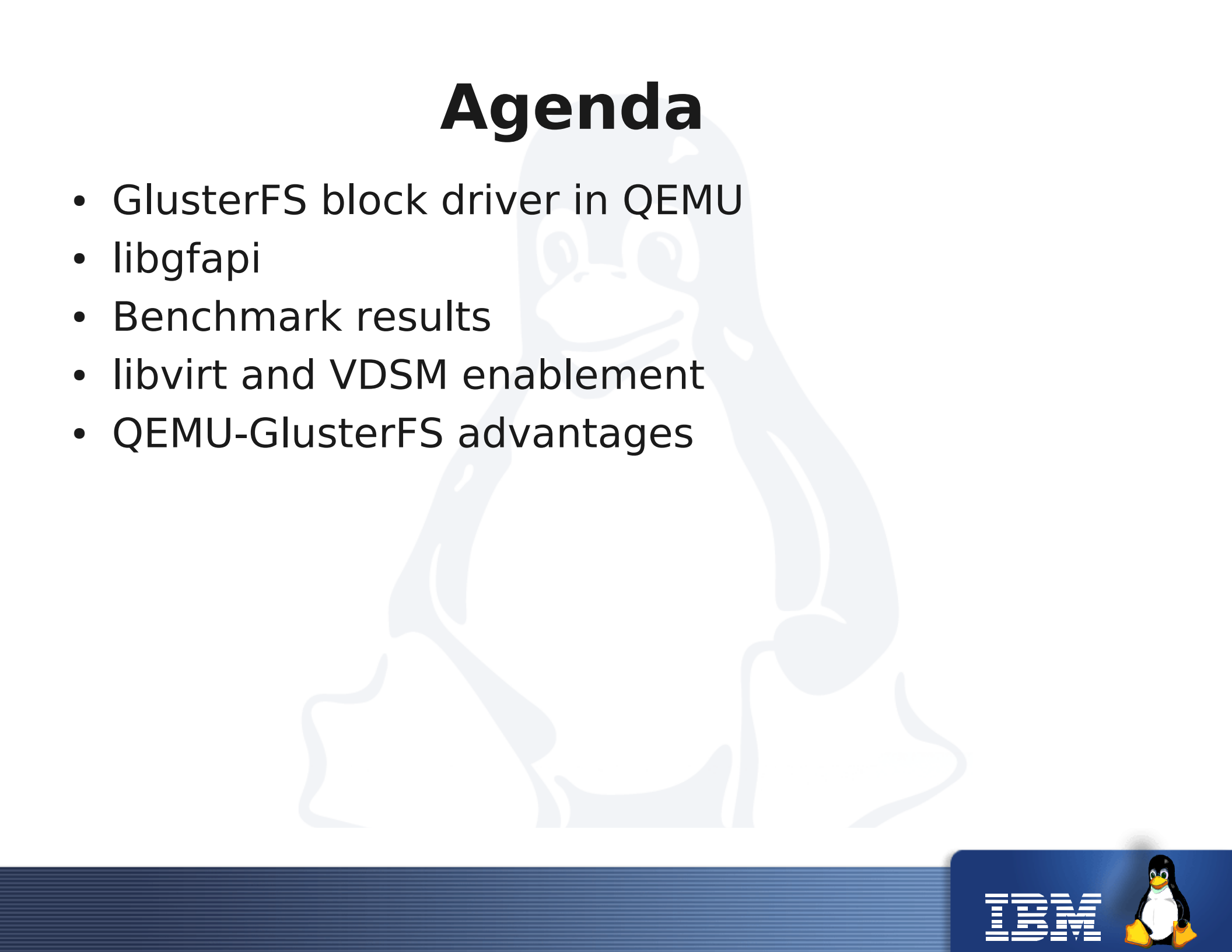
Bharata B Rao
IBM Linux Technology Center, Bangalore
bharata@linux.vnet.ibm.com

Aug 2012

Gluster Workshop2012

Agenda

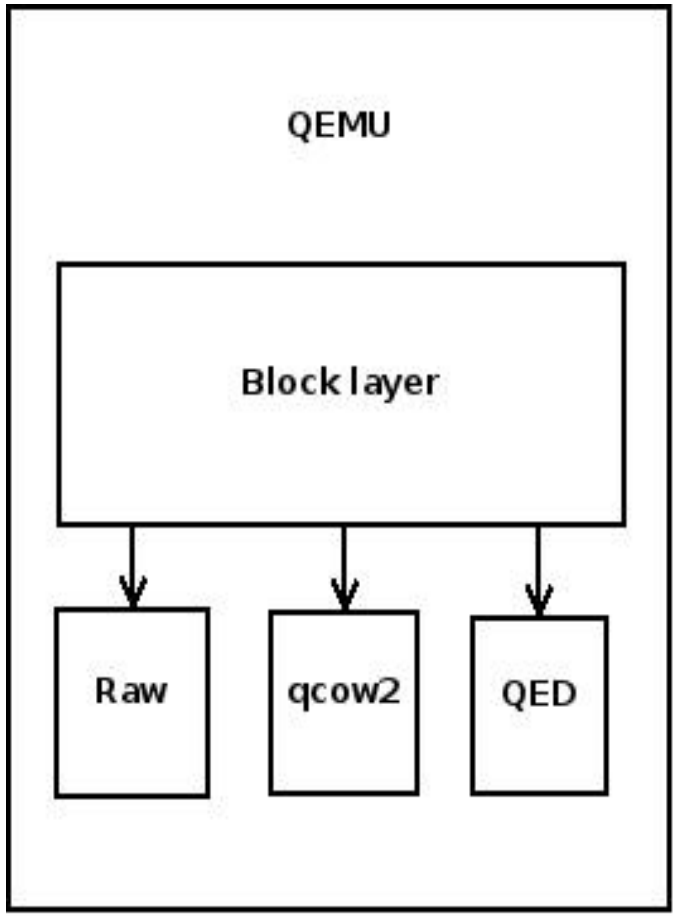
- GlusterFS block driver in QEMU
- libgfapi
- Benchmark results
- libvirt and VDSM enablement
- QEMU-GlusterFS advantages

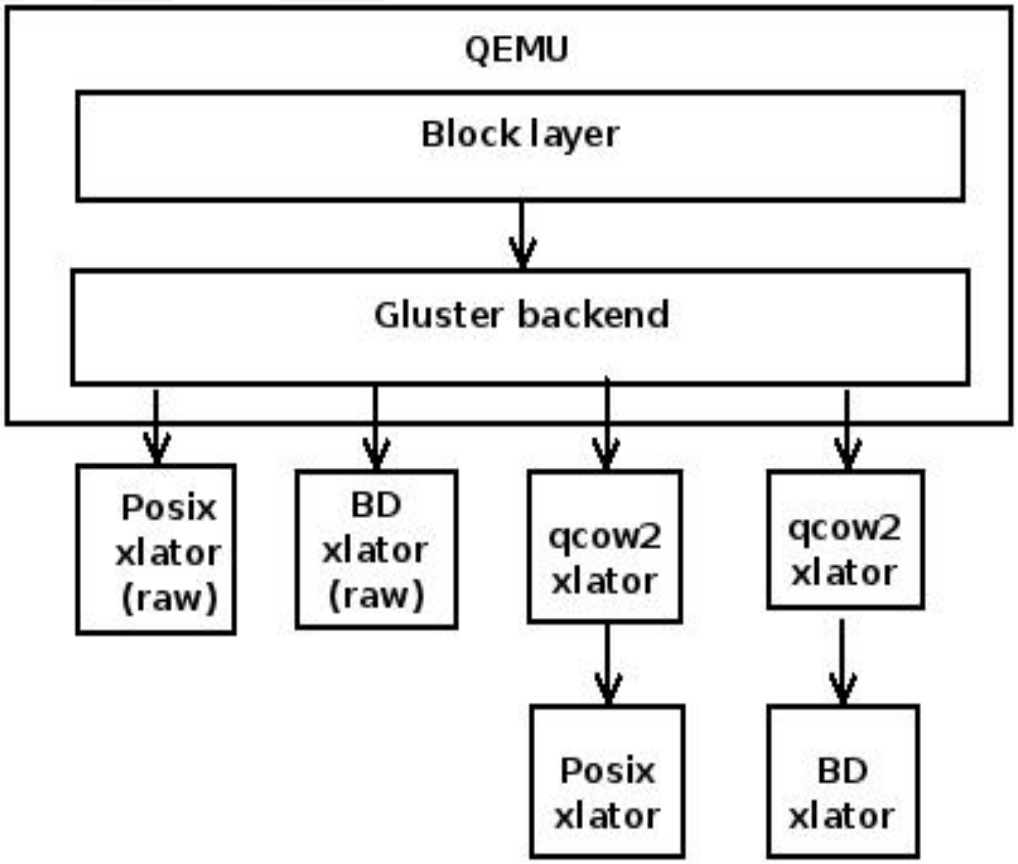


Enabling GlusterFS for Virtualization use

- QEMU-GlusterFS integration
  - Native integration, no FUSE mount
  - Gluster as QEMU block back end
  - QEMU talks to gluster and gluster hides different image formats and storage types underneath
- Block device support in GlusterFS via Block Device translator
  - Logical volumes as VM images

GlusterFS back end in QEMU





QEMU-GlusterFS integration

- New block driver in QEMU to support VM images on gluster volumes
  - Uses libgfapi to do IO on gluster volumes directly
- No FUSE overhead
- Gluster drive specification in QEMU

-drive file=gluster://server[:port]/volname/image[?transport=...]

server – volfile server where glusterd is running

port – port number on which glusterd is listening

volname – name of the gluster volume

image – path to VM image on the gluster volume

transport = Transport type used to connect to glusterd. It can be socket or rdma or unix

libgfapi

- A library to perform IO on gluster volumes directly without FUSE mount
- gluster connection init
  - glfs = glfs_new("volname");
  - glfs_set_volfile_server(glfs, transport, server, port);
  - glfs_init(glfs)
- Interfaces used by QEMU
  - glfs_preadv_async(), glfs_writev_async()
  - glfs_open(), glfs_creat(), glfs_close()
  - glfs_fruncate(), glfs_fstat()
  - glfs_fsync_async()
- gluster connection termination
  - glfs_fini()

Benchmark numbers

- FIO Numbers (Seq **read**, 4 files with direct io, QEMU options: if=virtio, cache=none)

	Aggregate BW (kB/s)	Min BW (kB/s)	Max BW (kB/s)
Base	63076	15769	17488
FUSE mount	29392	7348	9266
QEMU-GlusterFS native integration	53609	13402	14909
QEMU-GlusterFS native integration with trimmed client side xlators	62968	15742	17962

...Benchmark numbers

- FIO Numbers (Seq **write**, 4 files with direct io, QEMU options: if=virtio, cache=none)

	Aggregate BW (kB/s)	Min BW (kB/s)	Max BW (kB/s)
Base	189667	47416	107944
FUSE mount	43028	10757	13382
QEMU-GlusterFS native integration	150635	37658	49238

...Benchmark numbers

- FIO Numbers (Seq read, 4 files with direct io, QEMU options: if=virtio, cache=none)
- FIO run on a data drive

	Aggregate BW (kB/s)
Data drive supplied as a FUSE mount point to QEMU	20894
Data drive FUSE-mounted from inside guest VM	36936
Data drive specified as gluster drive in QEMU	47836

libvirt and VDSM support

- RFC patches out on libvirt mailing list to support gluster drive specification in QEMU
  - https://www.redhat.com/archives/libvir-list/2012-August/msg01625.html
- Libvirt XML specification

```
<disk type='network' device='disk'>
  <driver name='qemu' type='raw'/>
  <source protocol='gluster' name='volume/image'>
    <host name='example.org' port='6000' transport='socket'/>
  </source>
</disk>
```

- Patches to support for QEMU-GlusterFS native integration in VDSM are under review in gerrit
  - http://gerrit.ovirt.org/6856

QEMU-GlusterFS advantages

- VM images as files in all scenarios (esp SAN using BD xlator)
  - Ease of management
  - File system utilities for backup from GlusterFS FUSE mount (Future)
- Off-loading QEMU from storage/FS specific work
  - File system driven snapshots, clones (via BD xlator)
- Storage migration that is transparent to QEMU
  - Driven by GlusterFS (Future)
- Translator advantages
  - User space pluggable VFS, modularity
  - Lean storage-stack

Future

- Get the patches upstream
  - QEMU-GlusterFS
  - BD xlator
  - libvirt support
  - GlusterFS support in VDSM
- QEMU-GlusterFS performance enhancements
  - Zero copy readv/writev

References

- Latest QEMU-GlusterFS patches (v6)
  - http://lists.gnu.org/archive/html/qemu-devel/2012-08/msg01536.html
- Mohan's Block device xlator patches
  - http://review.gluster.org/3551
- Harsh's RFC patches for libvirt support
  - https://www.redhat.com/archives/libvir-list/2012-August/msg01625.html
- Deepak's Patches that add VDSM support
  - http://gerrit.ovirt.org/6856
- Video demo of using QEMU with GlusterFS
  - http://www.youtube.com/watch?v=JG3kF_djclg
- QEMU git tree – git://git.qemu.org/qemu.git
- GlusterFS git tree – git://git.gluster.com/glusterfs.git
- Benchmark details
  - http://lists.nongnu.org/archive/html/qemu-devel/2012-07/msg02718.html
  - http://lists.gnu.org/archive/html/gluster-devel/2012-08/msg00063.html

Legal Statement

- This work represents the view of the authors and does not necessarily represent the view of IBM.
- IBM, IBM(logo) are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries.
- Linux is a registered trademark of Linus Torvalds.
- Other company, product, and service names may be trademark or service marks of others.
- There is no guarantee that the technical solutions provided in this presentation will work as-is in every situation.