

GlusterFS performance tuning for small files, replication, distributed, NUFA

Author: André Felipe Machado<andremachado@techforce.com.br>

Small files performance is still the Achilles heel of GlusterFS. Tuning for replication (AFR, mirroring), distributed and NUFA setups is a non-trivial task, and you must know your application behaviour, your hardware and network infrastructure.

For small files performance, **forget GlusterFS prior to v. 2.0.8.**

Since ~~version 2.0.8, introduced~~ [patch 208](#), and this quick-read translator patch proved to be essential. It provides up to 10 times previous small files performance.

Some options at files below were undocumented, realized from gluster-users and gluster-devel mailing lists, my own experiments, and even from the source code. Use them at your risk.

I did not find any stability issues at my setup once it was defined. Mistakes at configuration gave errors and crashes in few minutes under load.

One of the non obvious solutions was to use io-threads at CLIENT side TOO for dealing with parallel file system requests, typical of servers. Performance rocketed under server loads.

The tests were performed using my own Debian 5.0.3 AMD64 GlusterFS compiled packages, based on ~~original~~ [Ubuntu](#) packages.

I used them for php sessions behaviour. Other uses imply other settings.

You must know your application behaviour. For example, too many random reads and writes of

small files makes usage of stats-prefetch useless and read-ahead WORSE than not using it. I/O-threads, cache dimensioning and flushing period are vital. Flush behind only works for files that are read many seconds after they are written, otherwise it can cause lockups. Options like enable-O_SYNC are performance killers and should be used only where needed. Do not forget to mount using noatime or at least relatime mounting options.

glusterfsd.vol

```
### Andre Felipe Machado http://www.techforce.com.br
### 09out2009 10h15min

#####
### GlusterFS Server Volume File ##
#####

#### CONFIG FILE RULES:
### "#" is comment character.
### - Config file is case sensitive
### - Options within a volume block can be in any order.
### - Spaces or tabs are used as delimiter within a line.
### - Multiple values to options will be : delimited.
### - Each option should end within a line.
### - Missing or commented fields will assume default values.
### - Blank/commented lines are allowed.
### - Sub-volumes should already be defined above before referring.

### Export volume "brick" with the contents of /srv/export/php_sessions
directory.
volume posix
    type storage/posix                # POSIX FS translator
    option directory /srv/export/php_sessions # Export this directory
end-volume

volume locks
    type features/locks
    option mandatory-locks on
    subvolumes posix
end-volume

volume iothreads
    type performance/io-threads
    option thread-count 16 # default is 16
    subvolumes locks
end-volume

volume writebehind
    type performance/write-behind
    option cache-size 1000MB # default is equal to aggregate-size
    option flush-behind off # default is 'off'
                                # too aggressive and slow background flush!
                                # do not enable for php sessions behaviour
    subvolumes iothreads
end-volume

volume brick
    type performance/io-cache
    option cache-size 2000MB # default is 32MB
```

```

# option priority *.h:3,*.html:2,*:1 # default is '*:0'
option cache-timeout 1 # default is 1 second
subvolumes writebehind
end-volume

### Add network serving capability to above brick.
volume server
type protocol/server
option transport-type tcp
option transport.socket.nodelay on # undocumented option for speed
# http://gluster.org/pipermail/gluster-users/2009-September/003158.html

# option transport-type unix
# option transport-type ib-sdp
# option transport.socket.bind-address 192.168.1.10 # Default is to
listen on all interfaces
# option transport.socket.listen-port 6996 # Default is 6996

# option transport-type ib-verbs
# option transport.ib-verbs.bind-address 192.168.1.10 # Default is
to listen on all interfaces
# option transport.ib-verbs.listen-port 6996 # Default is 6996
# option transport.ib-verbs.work-request-send-size 131072
# option transport.ib-verbs.work-request-send-count 64
# option transport.ib-verbs.work-request-recv-size 131072
# option transport.ib-verbs.work-request-recv-count 64

# option client-volume-filename /etc/glusterfs/glusterfs-client.vol
subvolumes brick
# NOTE: Access to any volume through protocol/server is denied by
# default. You need to explicitly grant access through # "auth"
# option.
option auth.addr.brick.allow * # Allow access to "brick" volume
end-volume

```

glusterfs.vol.afr.tcpnodelay.noreadahead.iothre
ads.quickread

Andre Felipe Machado <http://www.techforce.com.br>

27nov2009 15h42min

```

#####
### GlusterFS Client Volume File ##
#####

```

```

#### CONFIG FILE RULES:
### "#" is comment character.
### - Config file is case sensitive
### - Options within a volume block can be in any order.
### - Spaces or tabs are used as delimiter within a line.
### - Each option should end within a line.
### - Missing or commented fields will assume default values.
### - Blank/commented lines are allowed.
### - Sub-volumes should already be defined above before referring.

```

```

### Add client feature and attach to remote subvolume
volume debian459140
type protocol/client
option transport-type tcp
option remote-host 10.200.113.170 # IP address of the remote

```

```

brick
# option transport.socket.remote-port 6996      # default server port is
6996
  option ping-timeout 10                        # seconds to wait for a reply
                                           # from server for each request
  option transport.socket.nodelay on           # undocumented option for
speed
  # http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
  option remote-subvolume brick                # name of the remote volume
end-volume

```

```

volume debian459112

```

```

  type protocol/client
  option transport-type tcp
  option remote-host 10.200.113.171           # IP address of the remote
brick
# option transport.socket.remote-port 6996      # default server port is
6996
  option ping-timeout 10                        # seconds to wait for a reply
                                           # from server for each request
  option transport.socket.nodelay on           # undocumented option for
speed
  # http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
  option remote-subvolume brick                # name of the remote volume
end-volume

```

```

volume debian459115

```

```

  type protocol/client
  option transport-type tcp
  option remote-host 10.200.113.172           # IP address of the remote
brick
# option transport.socket.remote-port 6996      # default server port is
6996
  option ping-timeout 10                        # seconds to wait for a reply
                                           # from server for each request
  option transport.socket.nodelay on           # undocumented option for
speed
  # http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
  option remote-subvolume brick                # name of the remote volume
end-volume

```

```

volume debian459111

```

```

  type protocol/client
  option transport-type tcp
  option remote-host 10.200.113.173           # IP address of the remote
brick
# option transport.socket.remote-port 6996      # default server port is
6996
  option ping-timeout 10                        # seconds to wait for a reply
                                           # from server for each request

```

```

option transport.socket.nodelay on          # undocumented option for
speed
# http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
option remote-subvolume brick              # name of the remote volume
end-volume

```

```

volume replicated
type cluster/replicate
subvolumes debian459140 debian459112 debian459115
end-volume

```

Performance translators below

Add IO-Cache feature

```

volume iocache
type performance/io-cache
option cache-size 1000MB          # default is 32MB
# option priority *.h:3,*.html:2,*:1 # default is '*:0'
option cache-timeout 1           # default is 1 second
subvolumes replicated
end-volume

```

Add writeback feature

```

volume writeback
type performance/write-behind
# option aggregate-size 2MB       # deprecated option
option cache-size 500MB          # default is equal to aggregate-size
option flush-behind off         # default is 'off'
                                # too aggressive and slow background flush!
                                # do not enable for php sessions behaviour
subvolumes iocache
end-volume

```

Add quick-read for small files

```

volume quickread
type performance/quick-read
option cache-timeout 1           # default 1 second
option max-file-size 256KB      # default 64Kb
subvolumes writeback
end-volume

```

Add io-threads for parallel requisitions

```

volume iothreads
type performance/io-threads
option thread-count 16 # default is 16
subvolumes quickread
end-volume

```

**glusterfs.vol.nufa.unhashed.tcpnodelay.noread
ahead.quickread.iothreads**

```

### Andre Felipe Machado http://www.techforce.com.br
### 19nov2009 12h02min

```

```

#####
### GlusterFS Client Volume File ##
#####

```

CONFIG FILE RULES:

```

### "#" is comment character.
### - Config file is case sensitive
### - Options within a volume block can be in any order.
### - Spaces or tabs are used as delimiter within a line.
### - Each option should end within a line.
### - Missing or commented fields will assume default values.
### - Blank/commented lines are allowed.
### - Sub-volumes should already be defined above before referring.

### Add client feature and attach to remote subvolume
volume debian459140
    type protocol/client
    option transport-type tcp
    option remote-host 10.200.113.170          # IP address of the remote
brick
# option transport.socket.remote-port 6996    # default server port is
6996
    option ping-timeout 10                    # seconds to wait for a reply
                                            # from server for each request
    option transport.socket.nodelay on        # undocumented option for
speed
    # http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
    option remote-subvolume brick            # name of the remote volume
end-volume

volume debian459112
    type protocol/client
    option transport-type tcp
    option remote-host 10.200.113.171          # IP address of the remote
brick
# option transport.socket.remote-port 6996    # default server port is
6996
    option ping-timeout 10                    # seconds to wait for a reply
                                            # from server for each request
    option transport.socket.nodelay on        # undocumented option for
speed
    # http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
    option remote-subvolume brick            # name of the remote volume
end-volume

volume debian459115
    type protocol/client
    option transport-type tcp
    option remote-host 10.200.113.172          # IP address of the remote
brick
# option transport.socket.remote-port 6996    # default server port is
6996
    option ping-timeout 10                    # seconds to wait for a reply
                                            # from server for each request
    option transport.socket.nodelay on        # undocumented option for
speed
    # http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
    option remote-subvolume brick            # name of the remote volume
end-volume

volume debian459111
    type protocol/client

```

```

    option transport-type tcp
    option remote-host 10.200.113.173      # IP address of the remote
brick
# option transport.socket.remote-port 6996  # default server port is
6996
    option ping-timeout 10                # seconds to wait for a reply
                                        # from server for each request
    option transport.socket.nodelay on     # undocumented option for
speed
                                        # http://gluster.org/pipermail/gluster-users/2009-
September/003158.html
    option remote-subvolume brick         # name of the remote volume
end-volume

```

```

volume nufa
  type cluster/nufa
  option lookup-unhashed off             # off will reduce cpu usage, and
network
  option local-volume-name `hostname`   # note the backquote, so
'hostname'
                                        #output will be used as the option.
  subvolumes debian459140 debian459112 debian459115 debian459111
end-volume

```

Performance translators below

Add IO-Cache feature

```

volume iocache
  type performance/io-cache
  option cache-size 1000MB              # default is 32MB
# option priority *.h:3,*.html:2,*:1   # default is '*:0'
  option cache-timeout 1                 # default is 1 second
  subvolumes nufa
end-volume

```

Add writeback feature

```

volume writeback
  type performance/write-behind
# option aggregate-size 2MB             # deprecated option
  option cache-size 500MB               # default is equal to aggregate-size
  option flush-behind off               # default is 'off'
                                        # too aggressive and slow background flush!
                                        # do not enable for php sessions behaviour
  subvolumes iocache
end-volume

```

Add quick-read for small files

```

volume quickread
  type performance/quick-read
  option cache-timeout 1                 # default 1 second
  option max-file-size 256KB            # default 64Kb
  subvolumes writeback
end-volume

```

Add io-threads for parallel requisitions

```

volume iothreads
  type performance/io-threads
  option thread-count 16 # default is 16
  subvolumes quickread
end-volume

```