thinkparQ

Sizing and Tuning

Sizing?

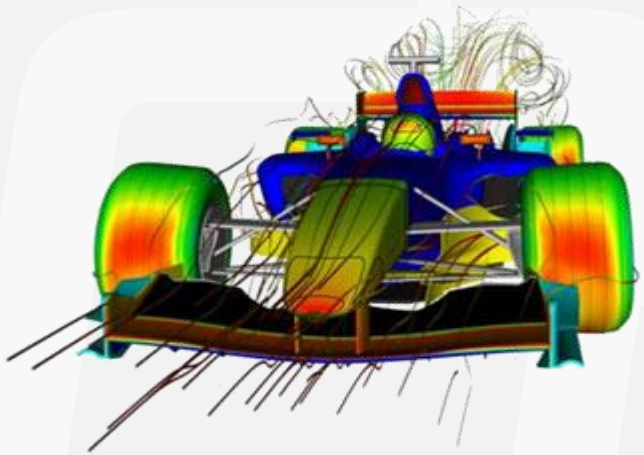
Design a system to meet a certain performance with a given number of pre-conditions and requirements

Tuning?

Get the best possible performance from a given hardware for a defined workload.

Performance?

- single stream bandwidth
- multi stream bandwidth
- (random) IOPS
- MetaData performance
- latency



BeeGFS – main components

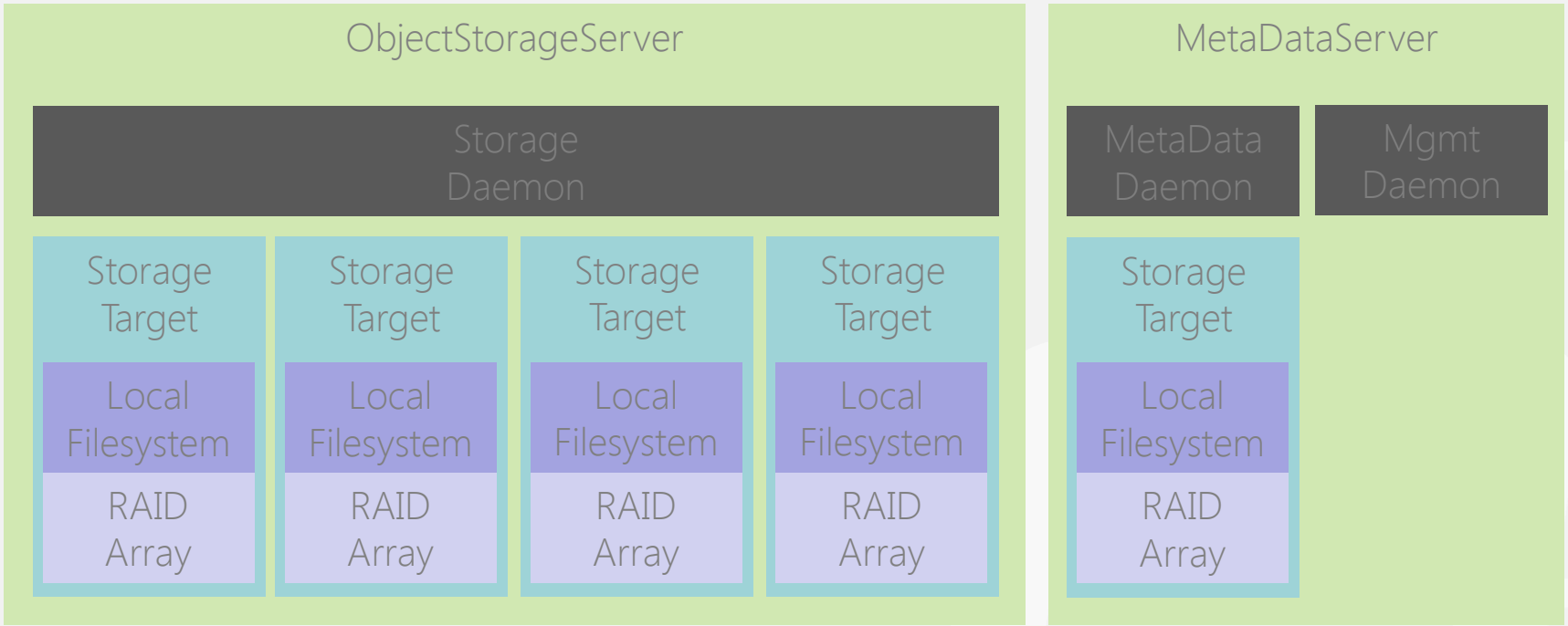
- ManagementServer
  - Rendevouz point for (new) servers and (new) clients
  - not critical for operation
  - a BeeGFS has exctly one ManagementServer

- MetaDataServer (MDS)
  - stores MetaData
  - exports a single MetaDataTarget (MDT)
  - a BeeGFS can have 1 ... 2^16 MDS

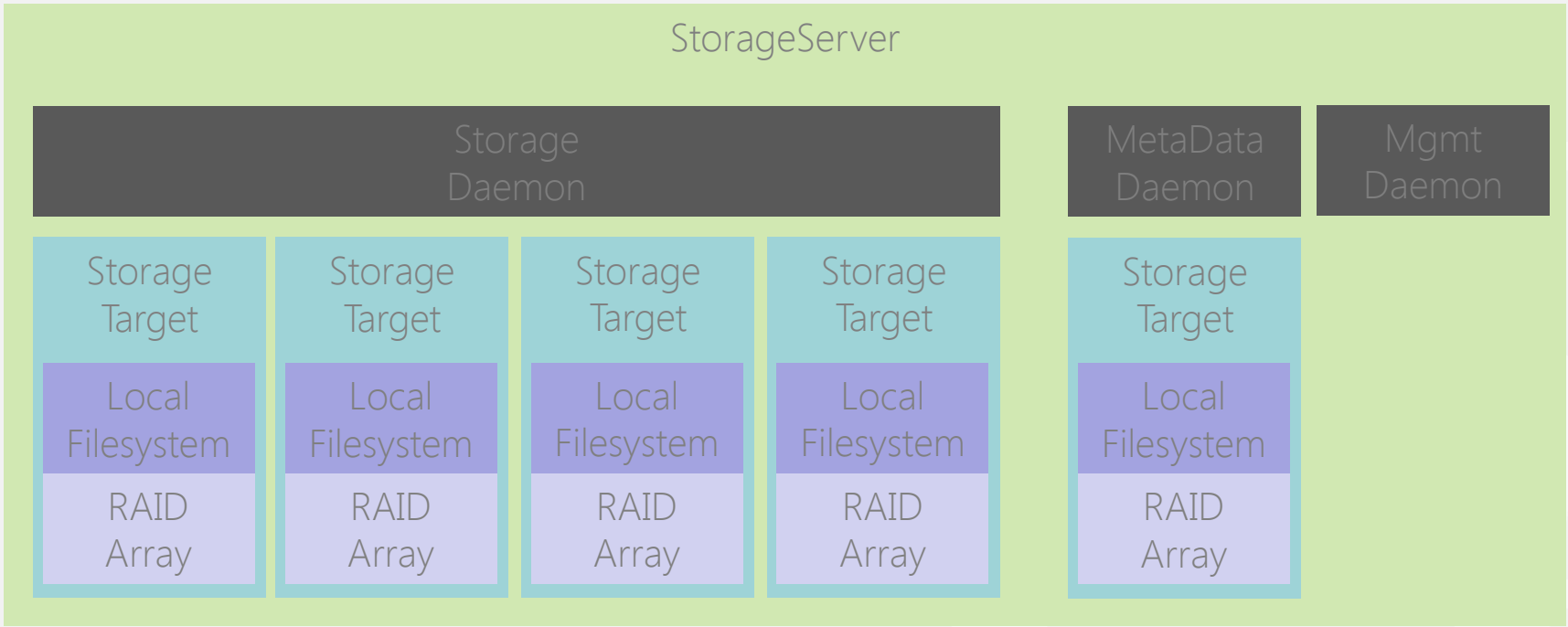
BeeGFS – main components

- ObjectStorageServer (OSS)
  - stores file data
  - exports ObjectStorageTargets
  - a BeeGFS has 1 … 2^16 OSS
- ObjectStorageTarget (OST)
  - is typically a RAID-Array – either internal or externally attached
  - blockdevice formatted with a local filesystem
  - a BeeGFS has 1 … 2^16 OST

storage server



storage server



ManagementServer

- No Special Requirements

MetaDataTargets

- RAID protection
  - no RAID 5 or 6
  - RAID 1 or 10 are common choices
  - chunksizes between 4k and 64k – depending on disk type
  - hardware or software RAID
- Disks
  - SSDs are widely used today
  - 10k or 15k drives can be sufficient for many workloads
  - ext4 as local filesystem
- Capacity
  - 0.3 to 0.5% of the usable object storage space as best practice

MetaDataServer

- CPU
  - single or dual socket
  - fewer cores with higher clock preferred over large number of cores
  - large number of cores required for systems with lots of clients
- RAM
  - RAM is used as cache – more is better
  - server RAM needed for connections:

    per client: connRDMABufSize x connRDMABufNum x 2 x connMaxInternodeNum = 24 MiB
- Network
  - type often defined by existing infrastructure
  - low latency important for MDS – bandwidth less important

ObjectStorageTargets

- RAID protection
  - no RAID 5!
  - RAID6 is the most common choice - RAID60 is possible but better use BeeGFS to stripe
  - not more than 12 disks per RAID6 – HotSpare recommended
  - chunksize between 128 and 512 Kbyte – depending on workload
  - software or hardware RAID
- Disks
  - 7.200 RPM is common choice - 10k and 15k drives rarely used
  - often 2 or 3 TByte drives
  - (nearline) SAS performs better than SATA
  - SSDs for special workload filesystems
  - ext4, xfs or zfs as local filesystem

ObjectStorageServer

- CPU
  - single or dual socket – depending on hardware/software RAID and network
  - with hardware RAID and InfinIband no high CPU load observed – slow CPUs are sufficient
- RAM
  - RAM is used as cache – more is better
  - server RAM needed for connections:

    per client: connRDMABufSize x connRDMABufNum x 2 x connMaxInternodeNum = 24 MiB
- Network
  - type often defined by existing infrastructure
  - Best practice: network should be 25% fast than disk storage

general guidelines

- combine MDS and OSS to scale both performance metrics
- scale up until price/performance breaks down
- define building blocks and scale them out
- ~75MB/s per spindle is a good (conservative) metric
- use hot-spare drives when possible – keep at least cold spares handy
- make use of caching
- disk bandwidth max. of network bandwidth
- redundant PSUs and UPS should be used!

example – capacity configuration

- „Get a system with 2 PByte and 10 GB/s bandwidth over 40GE“
- 1) Number of disks
  - go for 6TB drives to get best capacity per server
  - 2048 Tbyte in 6TB drives = 342 disks
  - RAID6 10+2 => 34,2 targets => round up to 35 targets => 420 drives gross
  - 420 drives x 75 MB/s => 31,5 GB/s expected from drives
- 2) network bandwidth
  - 40GE is ~4 GB/s network bandwidth
  - 75% of 4 GB/s = 3GB/s
  - 10 GB/s should be done with at least 3 interfaces

    (3,3 GB/s per interface = ~83% usage)
  - 3 Interfaces peak at ~12 GB/s

example – capacity configuration

- 3) Number of controllers
  - Highly depending on Hardware Vendor!
  - SAS2 controllers typically peak out at 1.8 to 2 GB/s
  - for 10 GB/s you would need at least 6 controllers

- 4) servers
  - 3 interfaces = at least 3 servers (channel bonded 40GE = bad idea!)
  - 3 servers with each 2 controllers
  - 70 drives per controller (420 / 6)
  - => doesn't compute with RAID6 10+2!
  - Add 12 drives -> 432 drives in total

example – capacity configuration

- 5) MetaDataSpace
  - 0.5% of 2 PByte = 10.24 TByte
  - RAID10 => 20.48 TByte gross needed
- 6) MetaDataServer
  - One dedicated server with 24 x 900GB SAS 10k drives in RAID10
  - MDS in each of the 3 servers with 8 x 900GB SAS 10k drives in RAID10

example – capacity configuration

- 7) Solution
  - 3 Server with 8 x 900GB SAS 10k in RAID10 for MD
  - each server with 144 drives on two controllers
  - each server exports 12 OST + 1 MDT
  - Total of 36 OSTs each with 60TB capacity
  - 2160 TByte usable
  - Total of 10,8 TByte for MetaData => 0,5%
  - expected performance for sequential IO between 10 and 12 GB/s

example – performance configuration

- „Design a system with at least 25GB/s sequential read/writes over QDR InfiniBand and 500TB capacity“
- 1) Number of disks
  - 25 GB/s at 75 MB/s per drive = 334 drives
  - RAID6 10+2 => 27,83 targets => round up to 28 targets
  - 336 drives needed
- 2) network bandwidth
  - QDR InfiniBand does ~3 GB/s
  - 75% of 3 GB/s = 2.25GB/s per link
  - 25 GB/s should be done with at least 11 interfaces

    (~76% usage)
  - 11 Interfaces peak at ~33 GB/s

example – performance configuration

- 3) Number of controllers
  - Highly depending on Hardware Vendor!
  - SAS2 controllers typically peak out at 1.8 to 2 GB/s
  - for 25 GB/s you would need at least 13 controllers
  - 28 targets => at least 14 controllers
- 4) servers
  - 11 interfaces = at least 11 servers (channel bonded InfiniBand = not straight forward!)
  - 14 controllers in 11 servers => doesn't compute
  - Use 14 servers – with 1 controller + 24 drives per controller
  - consider using 2 controllers per server!

example – performance configuration

- 5) Disk Size
  - 500 TB in 280 drives => 2TB per drive
  - 560 Tbyte expected capacity
- 6) MetaDataSpace
  - 0.5% of 0.56 PByte = 2.8 TByte
  - RAID10 => 5.6 TByte gross needed
- 7) MetaDataServer
  - MDS in each of the 14 servers
  - 2 x 200GB SSD in RAID1 is sufficient!

example – performance configuration

- 8) Solution
  - 14 Server with 2 x 200GB SSD in RAID1
  - each server with 24 drives on one controllers
  - each server exports 2 OST + 1 MDT
  - Total of 28 OSTs each with 20TB capacity
  - 560 TByte usable
  - Total of 2,8 TByte for MetaData => 0,5%
  - expected performance for sequential IO between 25 and 28 GB/s

What if you didn't make it?

Tune!

local disk tuning

- optimize your filesystem
  - align your partitions properly!
  - example: mkfs.xfs -d su=128k,sw=8 -isize=512 /dev/sda
  - mount options
    - noatime
    - nodiratime
    - logbufs, logbsize
    - allocsize (xfs)
    - (nobarrier)
    - MDT specific: ext4 with large number of inodes, large inodes, large journal
- optimize your blockdevice
  - pick your scheduler (noop, [deadline], cfq)
  - adjust the number of schedulable requests (128, 4096)
  - set the read-ahead properly (4 … 64 MB)
  - set „max_sectors_kb“ to fit your hardware

memory management

- filesystem cache
  - /proc/sys/vm/dirty_background_ratio [1…5]
  - /proc/sys/vm/dirty_ratio [10…75]
  - /proc/sys/vm/vfs_cache_pressure [50]
- kernel memory
  - /proc/sys/vm/min_free_kbytes [262144]
  - echo never > /sys/kernel/mm/redhat_transparent_hugepage/enabled

tune your daemons (specific)

- fhgfs-meta.conf
  - connMaxInternodeNum -> default 32
  - tuneNumWorkers -> default 0
  - tuneTargetChooser ([randomized], roundrobin, randomrobin)
- fhgfs-storage.conf
  - tuneNumWorkers -> default 12
  - don't touch tuneWorkerBufSize, tuneFileReadSize, tuneFileReadAheadTriggerSize, tuneFileReadAheadSize, tuneFileWriteSize, tuneFileWriteSyncSize

tune your daemons (both)

- fhgfs-{meta,storage}.conf
  - connUseRDMA
  - connAuthFile
  - tuneBindToNumaZone
  - tuneUsePerUserMsgQueues (true, [false])

fhgfs-ctl

- fhgfs-ctl --setpattern
  - --chunksize -> default 1m
  - --numtargets -> default 4
  - --raid10 (true, [false])

- performance monitoring
  - fhgfs-ctl –iostat –interval=2
  - fhgfs-ctl --serverstats --nodetype=meta --interval=3
  - fhgfs-ctl --serverstats --nodetype=storage --perserver --names --interval=1

impact of tuning?

- system set up with
  - sub-optimal choice for MDT
  - sub-optimal raid-arrays
  - no block device tuning
  - no OS tuning
- Internal benchmark
  - Before: 2.3 GB/s
  - After: 3.6 GB/s
- untar a file
  - Before: ~12 minutes
  - After: <5 minutes

conclusion

- hardware design is important for application performance
- changing settings can easily double your performance – and is easily done!
- there is no perfect setting for all workloads – understand what you are doing!
- monitor performance over time to find problems

thinkparQ

Questions?