Low-cost storage @PSNC

Update for TF-Storage

TF-Storage meeting @Uppsala, September 22nd, 2014
Agenda

Motivations – data center perspective
Application / use-case
Hardware components: bought some, will buy more
Software decisions to be taken
Deployment and use
Tender and surprising results
Motivations

We (thought we) make the solution low-cost:
- We need to differentiate disk storage equipment classes we use (vendor-made arrays vs build-your-own-storage-cluster)
- Need to push purchase costs down
- Need to limit TCO down, e.g. avoid paying expensive support

Online capacity needs grow constantly:
- Everything to be online – user expect that
- Even archives… are often used as online repository

Vendor lock-in:
- Can be tolerated in some cases
  + mission-critical systems - production VMs, services, mail, web etc.
  + HPC etc. where we push for throughput/IOPS
- Low cost storage must be free from vendor lock-in
  + non-critical applications
Other storage systems @PSNC *(storage hell)*

**HPC storage:**
- **Scratch:**
  - + DDN 10k / 300TB / FC connectivity +
  - + Lustre, 12 OSD servers
- **Home:**
  - + 4x NetApp E5500 (2x360TB + 2x480TB) + **SSD** (EF540)
  - + GPFS + cNFS: on top of 6 servers
  - + SSD was needed! (latency, backups, I/O waiters…)

**General-purpose storage:**
- **Storage 4 VMs and any other applications:**
  - + 2x NetApp E5400 – 120 drives (360TB)
  - + 1x NetApp E5500 – 90 drives (360TB)
  - + IOPS may be a problem
- **Archival storage**
  - + 3,5 PBs of tape storage + 10PBs in partnering institutions
  - + 200 TB cache
  - + HSM solution (IBM TSM for Space Management)
  - + note – people ARE TRYING TO USE IT AS A DISK!
Applications and use-cases

Application type:
- Sync & share
- Backups / DR / Archives
- Media hosting / content delivery
- VMs

Target aims:
- user base?
  - internal / external – both: users and systems (VMs)
  - university staff / students – first staff, then students
- capacity?
  - PBs…
- number of files / objects?
  - several thousands per user
  - hundreds of millions per storage system
Hardware components and configs

Server platforms:

- Storage nodes: 16x 12-drives pan-cakes / pizza-boxes – WHY:
  + small impact of putting the node down (failure, maintenance)
  + flexibility to build „islands” if we find it feasible at the end

- Options considered:
  + Quanta’s: S100-11D
  + SuperMicro SuperServer (FatTwin) F617H6-FTL+

- Options dropped:
  + SuperMicro’s 36/72-disk servers
  + OpenCompute storage vault
    (not available in small quantities)
Hardware components and configs

Server platforms:

- Redundant PSUs and fans – some traps and tricks

  + by requiring dual PSUs we ’killed’ some interesting options
  + Quanta was not available with 2x PSUs at that time

  + Traps:
    > single PSU
      => 2 large failure zones!!!
      => loss of flexibility
    > an extra PSU takes space inside the server
    > PSUs hot swappable
    > FANs redundant but not hot swappable
      + procedure: „power down the node…” 😊

- SuperMicro FatWin:
  + a common PSUs for all 4 nodes within a chassis
  + Fan failure impacts 4 nodes!
Hardware components and configs

Server configuration – CPU & RAM:

+ 2 CPU/s node
  + *not the cheapest option*
  + high **CPU to disks** – 2 / 12,
  + **flexibility** related to CPU power
  + we may run apps/VMs on the second CPU
  + *real life numbers will verify our decision*

+ architecture: typical x86_64, Xeon E5-2620v2
  + 80Watts, 2,1 GHz, 6 cores/CPU, 1
  + 12 cores/node, **1 core/OSD** (Ceph)

+ RAM: 64 GB
  + enough for Ceph/OpenStack OSDs/Storage Nodes?
  + spare capacity for VMs / apps / Docker images

Computing/application servers configuration – CPU & RAM:

+ 2 CPU – same as storage nodes
+ RAM: 256 GB
Hardware components and configs

Server configuration – disks:

- **HDD:**
  - 10x 4TB / drive per node
  - 2 bays / node used for SSDs (less HDD capacity per node, more IOPS)
  - 160 drives / initial config (640 TB raw capacity)
  - Hitachi UltraStar 7k400 (SATA)
  - heavy duty/enterprise
    + there are different opinions
    + not that big price difference
    + system to run in remote location
    + possibly the next time we will go for cheap rubbish

- **SSDs:**
  - 2x SSD: Intel DC S3700 240 GB
  - XFS / ext3/4 journaling
  - additionally: 1x DOM (Disk On Module) – operating system + logs

- **On-board SATA/SAS controllers:**
  - Intel C602… etc,
  - hardware / hybrid RAID will not be used
Hardware components and configs

Network:

Network interfaces:
- 10Gbit for front-end and back-end production traffic
- Typical 10Gbit Ethernet (not the LLE)
- 1Gbit for management/booting/monitoring
- 2x Chelsio T520-SO-CR / node - dual PCI-E card
- There was an option for on-board 10GbitE

BUT:
+ power consumption of copper option much higher!
+ in this case we had more freedom with purchase cost, less freedom with maintenance costs

Network switch – a part of the overall purchase
- 10Gbit for both front-end and back-end
<table>
<thead>
<tr>
<th></th>
<th>PHYS-micro-cloud (1-8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPS-XL-standard-7</td>
<td>VPS-XL-standard-8</td>
</tr>
<tr>
<td>VPS-XL-standard-5</td>
<td>VPS-XL-standard-6</td>
</tr>
<tr>
<td>VPS-XL-standard-3</td>
<td>VPS-XL-standard-4</td>
</tr>
<tr>
<td>VPS-XL-standard-1</td>
<td>VPS-XL-standard-2</td>
</tr>
<tr>
<td>VPS-XL-high-performance-7</td>
<td>VPS-XL-high-performance-8</td>
</tr>
<tr>
<td>VPS-XL-high-performance-5</td>
<td>VPS-XL-high-performance-6</td>
</tr>
<tr>
<td>VPS-XL-high-performance-3</td>
<td>VPS-XL-high-performance-4</td>
</tr>
<tr>
<td>VPS-XL-high-performance-1</td>
<td>VPS-XL-high-performance-2</td>
</tr>
</tbody>
</table>

- **switch-10GE-2**
- **switch-10GE-1**

<table>
<thead>
<tr>
<th></th>
<th>VPS-XL-standard-storage-3</th>
<th>VPS-XL-standard-storage-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>storage-15</td>
<td>storage-16</td>
<td></td>
</tr>
<tr>
<td>storage-13</td>
<td>storage-14</td>
<td></td>
</tr>
<tr>
<td>storage-11</td>
<td>storage-12</td>
<td></td>
</tr>
<tr>
<td>storage-9</td>
<td>storage-10</td>
<td></td>
</tr>
<tr>
<td>storage-7</td>
<td>storage-8</td>
<td></td>
</tr>
<tr>
<td>storage-5</td>
<td>storage-6</td>
<td></td>
</tr>
<tr>
<td>storage-3</td>
<td>storage-4</td>
<td></td>
</tr>
<tr>
<td>storage-1</td>
<td>storage-2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>VPS-XL-standard-storage-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>storage-11</td>
<td></td>
</tr>
<tr>
<td>storage-9</td>
<td></td>
</tr>
<tr>
<td>storage-7</td>
<td></td>
</tr>
<tr>
<td>storage-5</td>
<td></td>
</tr>
<tr>
<td>storage-3</td>
<td></td>
</tr>
<tr>
<td>storage-1</td>
<td></td>
</tr>
</tbody>
</table>
Hardware components and configs

Installation / deployment:
- Initial configuration in temporary data center:
  + built using container technology within 0,5 year
  + fully covered with systems (cooling, power, etc.)
- Distant location within the city:
  + we will be experimenting with latency impact

Data center B:
===============
20x IT racks
1 MW power
400 kW cooling
Hardware components and configs

Installation / deployment:
- Initial configuration moved to target data center - 2015
- Additional servers put to the new building
Hardware components and configs

Testing / Usage:

- Deployment / configuration tools
  + to be determined – suggestions welcome

- Monitoring: nagios + ganglia
Procurement stuff

Tender:
- Started in July, open in September

Spec:
- We expect 16 nodes fitting 16U
- 10 HDDs + 2SSD + 1DOM / node – explicitly
- CPUs (clock, cores, architecture) – explicitly
- Redundant power to node (dedicated OR shared PSU)

=> 16x1U (e.g. Quanta) OR 4x4U (SM FatWin) allowed

- In addition to storage servers: 4x „computing” servers

Winner:
- SuperMicro FatWin (through Action/Vector integrators)
Procurement stuff

Pricing: Higher than expected

Possible causes:

+ configuration:
  > a lot of CPUs: 2x CPU, 64 GB RAM / node (+20%)
  > E5 CPUs (E3 supports only 32GB RAM)
  > enterprise HDDs, SSDs (also good ones) (+20-30%?)
  > wanted performance, will see if we really got it
+ not enough push on vendor?
Software: what and why

- Initial decision: Object stores
  + Why:
    > promises fault tolerance, cheap scalability,
    > support Erasure Coding

- Implementation:
  + Ceph / RADOS OR OpenStack Swift
  + Performance comparison tests:
    > Ceph: RBD: 150-200MB/s, RADOS-GW: 50-100MB/s
    > Swift: 50-100 MB/s
    > Hardware: 2x Intel® Xeon® E5345, 2.33 GHz, 16 GB RAM
      1x SSD Intrepid drive, 60GB for objects storage

- Alternatives?
  - GlusterFS
  - ...

- I feel a bit uncomfortable without FS 😊
Roadmap

- 4Q2014 – installation and benchmarking of initial config

- 1Q2015:
  + Final decision on Software platform
  + Reporting to TF-Storage (February 2014)
  + 2nd round tender to start

- 2Q2015:
  + Tender opening, agreements, ordering
  + 2nd part of equipment installation & benchmarking

- Usage – decision to be made - experiments needed
  + We will divide final configuration into separate clusters
  + OR put all into one cluster
2nd round options

Target:
- 2 full racks of storage servers
- $2 \times 40 \times (10 - 12) \times 4TB = 3200 - 3840$ TB RAW capacity

Configuration options:

Option 1: go cheaper
- Regular drives, no SSD?
- Less CPUs, RAM

Option 2: go cooler ;)
- Open Compute Storage Vault + OpenRack etc. (costs?)

Option 3: go low-power ;)
- ARM/Atom CPUs (performance?)
Lessons

- It’s not that cheap!
  (but good thing is we have space CPU RAM for VMs)

- Not enough 'courage' / experience
to really relax requirements vs hardware

- We will see the real costs in some years
  (now we may only try to model them,
  which is worth to do – collaboration welcome)

- A lot of platforms do not support
  which might be deeply hidden in the docs (take care)
  + double PSUs / CPUs option
  + hot-swap disks
  + hot-swap fans

- Tendering takes time… (kind of obvious…)