Building low cost disk storage with Ceph and OpenStack Swift

Paweł Woszuk, Maciej Brzeźniak

TERENA TF-Storage meeting in Zurich
Feb 10-11th, 2014
Low-cost storage – motivations (1)

- Pressure for high-capacity, low-cost storage
  - Data volumes growing rapidly (*data deluge, big data*)
  - Budgets does not extend as quickly as storage
  - Storage market follows the cloud market
  - Virtualisation causes explosion of storage usage (deduplication not always mitigates the increasing number of disk images)
Low-cost storage – motivations (2)

• NRENs under pressure of industry
  – Pricing (see S3 pricelist)...  
  – Features in front of Dropbox, Google Drive
  – Scale-out capability (can we have it?)
  – Integration with IaaS services (VM + storage)

• Issues while building storage on disk arrays
  – Reliatively high invest. cost and maintenance
  – Vendor lock-in
  – Closed architecture, limited scalability
  – Slow adoption of new technologies
Topics covered

• Strategy
• Technology
• Pricing / costs
• Collaboration opportunity
PSNC strategy / approach

• Build a private storage cloud
  – i.e. to build not to buy 😊
  – Public cloud adoption still problematic

• Use object storage architecture
  – Scalable, no centralisation, open architecture
  – HA thanks to components redundancy

• Run a pilot system using:
  – Open source software
  – Cost-efficient server platform

• Test the solutions:
  – Various software / hardware mixtures
  – Various workloads: plain storage, sync&share, VMs, video
Software: open source platforms considered

**OpenStack**

Swift

- User Apps
- Load balancer
- Proxy Node
- Proxy Node
- Proxy Node
- Storage Node
- Storage Node
- Storage Node
- Storage Node

**CEPH**

- APP
- Host / VM
- Client
- Rados
- RadosGW
- RBD
- CephFS
- LibRados
- MONs
- MON.1
- MON.n
- OSDs
- OSD.1
- OSD.n
Software: OpenStack Swift
Software: Ceph

**LibRados**

- **MDS**
  - MDS.1
  - MDS.n

- **MONs**
  - MON.1
  - MON.n

- **Pool 1**
- **Pool 2**
- **Pool X**
- **Pool n**

**CRUSH map**

- **PG 1**
- **PG 2**
- **PG 3**
- **PG 4**
- **PG n**

**Cluster Node [OSDs]**

- **S3**
- **Swift**

- **RadosGW**
- **RBD**
- **CephFS**

- **Client**
- **HOST / VM**
- **APP**

**LibRados**

- **APP**
- **HOST / VM**
- **Client**

**RadosGW**

- **S3**
- **Swift**

**RBD**

- **CephFS**

**Cluster Node [OSDs]**

- **Client**
- **HOST / VM**
- **APP**
Ceph – OSD selection
Ceph – OSD selection + write to replicas
Software: OpenStack Swift vs Ceph

• Scalability:
  • Architecture/features: e.g. load balancing:
    • Swift – external,
    • Ceph – within the architecture
  • Implementation:
    • Swift – python
    • Ceph – C/C++

• Maturity
• User base
• Know-how around
Hardware

• Different people use different back-ends
  – Pan-cakes (1U, 12 drives) vs 'Fat' nodes (4U, 36+ drives)
  – HDDs vs SSDs
  – 1Gbit vs 10Gbit connectivity

• PSNC:
  – 1st stage: regular servers from HPC cluster:
    – 1 HDD (data) + 1 SSD (meta-data, FS journal)
    – 1Gbit for clients, Infiniband within the cluster
  – 2nd stage: pilot installation of 16 servers
    – 12 HDDs: data + meta-data
    – 10 HDD (data) + 2 SSD (meta-data + FS journal, possibly caching)
    – 10 Gbit connectivity
  – Software and hardware comparison tests
Pancake stack storage rack 😊

10 storage servers block:
10 x 12 HDDs = 120 HDDs
120 HDDs x 4TB = 480 TB on HDDs (raw)
1-2 CPU, 32 GB RAM

10 high performance storage servers:
10 x 10 HDDs = 100 HDDs
10 x 2 SSDs = 20 SSDs
100 HDDs x 4TB = 320 TB on HDDs
16 SSDs x 0.4TB = 6.4 TB on SSDs
1-2 CPU, 64 GB RAM

Dual 10gbit switches

10 storage servers block:
10 x 12 HDDs = 120 HDDs
120 HDDs x 4TB = 480 TB on HDDs (raw)
1-2 CPU, 32 GB RAM

10 high performance storage servers:
10 x 10 HDDs = 100 HDDs
10 x 2 SSDs = 20 SSDs
100 HDDs x 4TB = 320 TB on HDDs
16 SSDs x 0.4TB = 6.4 TB on SSDs
1-2 CPU, 64 GB RAM

Quanta Stratos S100-L11SL

TOTAL
400-480 HDDs = 100 HDDs
480 HDDs x 4TB = 1920 TB on HDDs
20-80 SSDs = 20 SSDs
80 SSDs x 0.4TB = 32 TB on SSDs
40-80 CPUs, 1280-2560 GB RAM
A pancake – photos

Photo by PSNC

Photo from: http://www.quantaqct.com/en/01_product/02_detail.php?mid=27&sid=158&id=159&qs=100=
Pancake in action

Diagnostic panel on the server front shows the status of the disk drive (useful while dealing with hundreds of drives)

Photos by PSNC

Server read performance in a throughput mode reaches 1.5GB/s (dstat output under stress test)
Costs (inv./TCO vs capacity)

- Assumptions:
  - Analysis for 5 years long lifecycle of the servers
  - Investment cost includes 5 years warranty
  - Total cost includes:
    - Investment costs
    - Power & cooling, room cost
    - Personnel costs

<table>
<thead>
<tr>
<th>Racks</th>
<th>Config</th>
<th>Servers</th>
<th>investment cost [EUR]</th>
<th>total cost</th>
<th>RAW capacity total [TB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>2CPU/10Gig/12disk/1U</td>
<td>10.00</td>
<td>58,355</td>
<td>292,270</td>
<td>480</td>
</tr>
<tr>
<td>0.50</td>
<td>2CPU/10Gig/12disk/1U</td>
<td>20.00</td>
<td>110,226</td>
<td>398,055</td>
<td>960</td>
</tr>
<tr>
<td>0.75</td>
<td>2CPU/10Gig/12disk/1U</td>
<td>30.00</td>
<td>155,613</td>
<td>497,356</td>
<td>1440</td>
</tr>
<tr>
<td>1.00</td>
<td>2CPU/10Gig/12disk/1U</td>
<td>40.00</td>
<td>194,516</td>
<td>590,173</td>
<td>1920</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>2CPU/10Gig/36disk/4U</td>
<td>4.00</td>
<td>67,432</td>
<td>333,075</td>
<td>576</td>
</tr>
<tr>
<td>0.50</td>
<td>2CPU/10Gig/36disk/4U</td>
<td>6.00</td>
<td>101,148</td>
<td>409,612</td>
<td>864</td>
</tr>
<tr>
<td>0.75</td>
<td>2CPU/10Gig/36disk/4U</td>
<td>8.00</td>
<td>127,372</td>
<td>478,656</td>
<td>1152</td>
</tr>
<tr>
<td>1.00</td>
<td>2CPU/10Gig/36disk/4U</td>
<td>10.00</td>
<td>159,215</td>
<td>553,320</td>
<td>1440</td>
</tr>
</tbody>
</table>
## Monthly TCO / TB

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>2CPU/2x10Gig/12disk/1U</td>
<td>10.00</td>
<td>1</td>
<td>20.30</td>
<td>30.44</td>
<td>1.33</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>2CPU/2x10Gig/12disk/1U</td>
<td>20.00</td>
<td>6.91</td>
<td>13.82</td>
<td>20.73</td>
<td>9.21</td>
<td>7.68</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>2CPU/2x10Gig/12disk/1U</td>
<td>30.00</td>
<td>5.76</td>
<td>11.51</td>
<td>17.27</td>
<td>7.68</td>
<td>6.40</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>2CPU/2x10Gig/12disk/1U</td>
<td>40.00</td>
<td>5.12</td>
<td>10.25</td>
<td>15.37</td>
<td>6.83</td>
<td>5.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>2CPU/2x10Gig/36disk/4U</td>
<td>4.00</td>
<td>8.46</td>
<td>16.91</td>
<td>25.37</td>
<td>11.28</td>
<td>9.40</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>2CPU/2x10Gig/36disk/4U</td>
<td>6.00</td>
<td>7.68</td>
<td>15.36</td>
<td>23.04</td>
<td>10.24</td>
<td>8.53</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>2CPU/2x10Gig/36disk/4U</td>
<td>8.00</td>
<td>7.20</td>
<td>14.39</td>
<td>21.59</td>
<td>9.59</td>
<td>8.00</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>2CPU/2x10Gig/36disk/4U</td>
<td>10.00</td>
<td>6.83</td>
<td>13.66</td>
<td>20.49</td>
<td>9.11</td>
<td>7.59</td>
<td></td>
</tr>
</tbody>
</table>
# Pricing by Amazon

## Storage Pricing

Region: EU (Ireland)

<table>
<thead>
<tr>
<th></th>
<th>Standard Storage</th>
<th>Reduced Redundancy Storage</th>
<th>Glacier Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>First 1 TB / month</td>
<td>$0.085 / GB</td>
<td>$0.068 / GB</td>
<td>$0.011 / GB</td>
</tr>
<tr>
<td>Next 49 TB / month</td>
<td>$0.075 / GB</td>
<td>$0.060 / GB</td>
<td>$0.011 / GB</td>
</tr>
<tr>
<td>Next 450 TB / month</td>
<td>$0.060 / GB</td>
<td>$0.048 / GB</td>
<td>$0.011 / GB</td>
</tr>
<tr>
<td>Next 500 TB / month</td>
<td>$0.055 / GB</td>
<td>$0.044 / GB</td>
<td>$0.011 / GB</td>
</tr>
<tr>
<td>Next 4000 TB / month</td>
<td>$0.051 / GB</td>
<td>$0.041 / GB</td>
<td>$0.011 / GB</td>
</tr>
<tr>
<td>Over 5000 TB / month</td>
<td>$0.043 / GB</td>
<td>$0.034 / GB</td>
<td>$0.011 / GB</td>
</tr>
</tbody>
</table>
Conclusions (1)

- **NRENs can compete on ’pricing’ with industry**
  - At the end we may use similar hardware and software components
  - Can we compete with our SLAs? Can we scale out? How to make it?

- **Cheap storage is not that cheap 😞**
  - Hardware:
    - In the analysis we are not using extremely cheap components
    - We could use even cheaper hardware, but:
      - Do we want it: Operational costs, Know-how cost
      - Are we able to really provide SLAs on top of it?
  - Software:
    - We need RAID-like, e.g. erasure coding mechanisms to increase storage efficiency (in the analysis we assumed 3x replication)

- **There is definitely field to collaborate**
  - Know-how/experience exchange
  - Storage capacity/services exchange?
    - Technically possible, but politics are always difficult
Conclusions (2)

- We should examine possibility to use different hardware solutions
Storage row in PSNC’s data center in 2 years 😊 - see: blog.backblaze.org