Low cost, highly dense Storage systems

Designing Storage Architectures Meeting
Library of Congress, September 17, 2018

Pashupati Kumar, Principal SW Eng. Manager
Microsoft
Pashupati.kumar@Microsoft.com
Storage Hierarchy and Technologies

- **Access pattern**
  - **RW occasionally**
- **Write once, Read never**
- **active**

- **SSD**
  - $$$$$
- **HDD**
  - $$$
- **Tape**
  - $

**latency**

- 1 ms
- 10 ms
- seconds
- minutes
- hours
Storage Hierarchy and Technologies

- **Active**: SSD
  - $$$$$
- **Cold-tier Storage System**: HDD
  - $$$
- **Tape**: Cold-tier Storage System
  - $

Access pattern:
- Write once, Read never
  - RW occasionally

Latency:
- 1 ms
- 10 ms (seconds)
- minutes
- hours
Goal

• Build the lowest cost HDD storage possible
• Deliberately trade performance for lower cost
• Avoid stranded storage
• Flexible performance characteristics
• Use commodity components
Driving storage cost down...

Common in the cloud:

- Compute racks
- Storage racks
- Network

Improves performance/cost:
- Independent resource scaling
- Rack hardware specialization

Reduce overheads in Storage racks!

1. Have large number of HDDs for each server
   - Gola is have storage cost same as that of HDD
2. Power off drives that are not currently utilized
   - Put them in lower power mode. E.g. Drives in Standby mode consume 50% less power than in Active Idle state
   - 20 – 25% OPEX saving can be realized
HDD – Power Conditions

- Performing HDD Power off/On is not flexible design options
  - Depends on JBOD enclosure implementation
- HDD supports different Power Conditions, that can be controlled via SW

<table>
<thead>
<tr>
<th>Power Condition</th>
<th>Power (W)</th>
<th>Power Savings (%)</th>
<th>Recovery Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle</td>
<td>2.82</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Idle_A</td>
<td>2.82</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Idle_B</td>
<td>2.18</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Idle_C</td>
<td>1.82</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>STANDBY_Z</td>
<td>1.29</td>
<td>54</td>
<td></td>
</tr>
</tbody>
</table>

- Use standard SBC (START_STOP_UNIT, 0x1B) to go to desired power condition
- Method to determine current power condition are different for SATA & SAS drives
- SCSI Log pages are available for monitoring power transitions counters
  - Start/Stop Cycles counter Log page
  - Power Condition Transitions Log page
Challenges

• Spin-up and down cycle
  • current limitation and future progression
• Disk AFR
  • Need to characterize disk failure rates for this “new workload”
• Drive technology for cold/Archive use cases
• Power surge during standby to Active idle state
Is it ok to do all these spin-ups?

Datasheet spec: 50K per year.
Avoid Stranded Storage

• Software can cope with loss of a server
  • But how much work does that cause?
    • Aggressive re-replication of data consumes lots of resources
    • Gets really worse, as storage server has 10-12 x HDDs

• Suppose data is still accessible
  • Even at a lower performance
  • Software can adjust load balancing
  • Much easier to handle, fewer resources used, lower COGS
Traditional SAS redundancy is expensive

• Traditional method was SAS dual attached disks
  • More expensive disks
  • Dual links to the disks
  • Dual expander hierarchy
  • Dual everything
  • Massively wasteful and expensive

• Not actually what we want
Rack-Scale HDD Storage Disaggregation

• Relaxing the HDD Ownership Principle
  • At a given time, a HDD is managed by one server...
  • ...but it is possible to reconfigure which server it is.

• Enables 4 types of disaggregation:
  • Configuration Disaggregation
  • Failure Disaggregation
  • Dynamic Elastic Disaggregation
  • Complete Disaggregation

No reconfiguration during normal operation
Reconfiguration part of normal operation
Rack Scale HDD Disaggregation

**Rack bandwidth for storage:**

- For the Cloud: low cost components
  - Commodity servers
  - SATA HDDs

**Any HDD connected to any server**

- Server elasticity

For the Cloud: low cost components

- Commodity servers
- SATA HDDs

Any HDD connected to any server

- Server elasticity
Experience with Failure Disaggregation

• Hardware trends impact data availability:
  • HDD and SSD capacities grow
  • Servers can have a LOT of direct-attached storage
  • e.g.: Petabytes of data per Pelican (cold storage) server
  • On failure, amount of data and time to recover increases

• Failure disaggregation improves availability
  • Reduces data unavailability to tens of seconds or less
  • No resources used to rebuild data
  • No reconfiguration overhead for normal operation

Pelican prototype has:
• 1152 HDDs/rack
• 2 servers
Conclusion

• In the cloud today: no disaggregation in storage racks
  • Fixed drive-to-server mapping

• We designed a storage fabric to explore in-rack disaggregation

• Rack-scale storage disaggregation can be useful and affordable
  • Configuration disaggregation
  • Failure disaggregation
  • Dynamic elastic disaggregation
  • Substantial benefits
  • No/small reconfiguration overheads
  • Little or no software/hardware changes

• Can become a challenge
  • Complete disaggregation
  • High reconfiguration overhead
  • Hard to implement and maintain

• Complete disaggregation
Performance

• Design biased for throughput
• User data is striped across many drives in a group
• Drive is assigned to a group with following consideration
  • Across multiple components
  • Minimal contention for storage bandwidth
  • Minimize overall rack vibration and cooling requirement
Configuration

• Breakdown

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Servers</td>
<td>2</td>
</tr>
<tr>
<td>Leases</td>
<td>2</td>
</tr>
<tr>
<td>Classes/Lease</td>
<td>2</td>
</tr>
<tr>
<td>Groups/class</td>
<td>11</td>
</tr>
<tr>
<td>Groups/class</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Disks/group</td>
<td>20</td>
</tr>
<tr>
<td>Total disks</td>
<td>880</td>
</tr>
<tr>
<td>Erasure coding scheme</td>
<td>15+3</td>
</tr>
<tr>
<td>% of disk in Active</td>
<td>9%</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Erasure coding scheme: 15+3 (Overhead = 18/15 = 1.2 )
- % of disk in Active (on loaded system): 80/880 = 9% (72 / 880 = 8.2 %)

• HDD labelling in an enclosure

```
<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>8</th>
<th>9</th>
<th>7</th>
<th>3</th>
<th>4</th>
<th>10</th>
<th>0</th>
<th>0</th>
<th>9</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
```
Another example

• Enclosures with dual cables
• With any single failure one server still has access to at least 7/8 of the disks
TBs transferred

datasheet spec: 60TB/year
Power On Hours

datasheet spec: 3120 POH/year (about 1/3rd of a year)