Using Distributed Fault Tolerant Memory in Virtualized Data Centers

Woon Jung
PernixData
Agenda

- Background
- Motivation for DFTM
- Survey of challenging problems with DFTM
- Q&A
Background
It started with FVP…

Hypervisor-based non-disruptive data acceleration

Hypervisor-based non-disruptive data acceleration

Hypervisor

PernixData FVP

FVP Cluster

Primary Storage
Clustered Data Tier

Server flash aggregation into cluster-wide pool

VM

Hypervisor

PernixData FVP

FVP Cluster

Primary Storage
Decoupled Architecture

Storage Performance

FVP

Storage Capacity

Data Services
Does the exact acceleration media type matter?
- Perfect physics
- Resource pooling
- Fault tolerance

Flash is great. Why bother with Memory?
- Denser and cheaper memory (overprovisioned memory)
- Easy to test, deploy and manage (business operation).
My policy: At least one WB replica in #1 and one in #2
Challenge: Software Overhead

- Software deficiencies show up… with ultra low latency devices

<table>
<thead>
<tr>
<th>Time to issue an IO via storage stack</th>
<th>15us ~ 20us</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledging IO completion</td>
<td>20us ~ 350us</td>
</tr>
<tr>
<td>4KB IO to local PCIe flash device</td>
<td>15us ~ 75us</td>
</tr>
<tr>
<td>4KB IO to new NVMe device</td>
<td>1us ~ 3us (Claimed)</td>
</tr>
<tr>
<td>4KB IO to DRAM</td>
<td>1us</td>
</tr>
</tbody>
</table>

- Dedicated contexts to issue and complete IOs
- Scrutiny over every single lock.
- Memory Allocation.
Challenge: Memory is Precious

- Dynamic re-sizing
  - Memory is a “flexible” resource.

- Reduce Metadata overhead
  - This works great even with Flash.

- DFTM-z (Compression)
  - Do more with the same amount of RAM.
Hello, I am FVP with DFTM

<table>
<thead>
<tr>
<th>Name</th>
<th>Acceleration Resources</th>
<th>Capacity</th>
<th>Datastores/VMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVP_01</td>
<td>4 Resources (from 4 of 4 Hosts)</td>
<td>1 TB</td>
<td>5 Datastores, 24 VMs</td>
</tr>
</tbody>
</table>

**FVP_01**

- **Overview**
- **Configure**

**Resources**

- **Hosts**: 4 of 4 with devices
- **Resources**: 4 Memory

**Status**

- **OK**: The FVP Cluster is functional

**Performance**

- **Realtime**
  - VM IOPS (Sum)
  - VM Throughput (Sum)
  - VM Latency (Avg)
  - Acceleration Hit Rate

**Consumers**

- **Eligible**: 24 VMs (24 VMs Active)
- **Not Eligible**: 0 VMs
- **Blacklisted**: 0 VMs

---

2015 Data Storage Innovation Conference © PernixData. All rights reserved.
RAM based I/O Acceleration

4KB random reads
RAM based I/O Acceleration

4KB random reads
RAM based I/O Acceleration

4KB random reads
RAM based I/O Acceleration

64KB sequential reads
64KB sequential reads
RAM based I/O Acceleration

64KB sequential reads
Potential Impact

- Allows up to 1TB of RAM per host
  - Cluster of 32 host, we can create a 32TB DFTM tier

- Infrastructure Level play
  - Applies to all type of applications running inside a VM, without any modification.

- Extremely easy to configure and manage.
Thank You!

@unnojung
whj@pernixdata.com