XvMotion: Unified Virtual Machine Migration over Long Distance

Ali José Mashtizadeh, Min Cai, Gabriel Tarasuk-Levin, Ricardo Koller, Tal Garfinkel, Sreekanth Setty
Stanford University – VMware, Inc.
Live Migration

Source Host

Cloud

Destination Host
Migration Benefits

• Test to production migrations
• Disaster Preparedness Testing
• Cross-Datacenter Load Balancing
• Shared-Nothing Architecture
Migration in Practice

Requires:
Shared Storage
Fast Networking
Single Subnet
This is not what we really want

• Migrations are limited to machines that:
  With shared storage, fast networks, and same LAN

• Technological Changes:
  • Shared nothing application architectures
  • Network mobility possible: LISP, OTV, VXlan, OpenFlow (SDN)
  • Very large virtualized datacenters

• No reason for these limitations any longer
• Customers have new use cases
XvMotion

- XvMotion: First commercially viable WAN migration
- Achieve Low Downtime AND Atomic Switchover
- Uses Asynchronous IO Mirroring

- Principle:
  Flow control mechanisms for memory and disk
## Customer Scenario

<table>
<thead>
<tr>
<th></th>
<th>LAN</th>
<th>WAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td>10 Gbps (sometimes 40 Gbps)</td>
<td>1 Gbps or less</td>
</tr>
<tr>
<td>Latency</td>
<td>&lt;1 ms</td>
<td>~100 ms</td>
</tr>
<tr>
<td>Typical Network</td>
<td>Dedicated NIC(s)</td>
<td>Shared connection between sites</td>
</tr>
</tbody>
</table>
Example Workload

Client \(\quad\) Application \(\quad\) Highly Available Database

HA Timeouts several seconds
TCP Timeouts 120 seconds
Downtime (Switchover Time)

Source Host → Running → Destination Host

Virtual Machine

Goal: Less than 1 Second of Downtime
Atomicity

Source Host

Virtual Machine

Goal: Atomic switchover

Destination Host
XvMotion

• Unifies Memory Migration and Storage Migration
  • Tolerates wide area network bandwidth/latency and reliability
  • Tolerates heterogeneous storage performance
  • Downtimes and workload impact comparable to local migration
  • Atomic switchover

• Deployed in customer metro area networks
• Cross continent migration e.g. Palo Alto to India is practical
Overview

• **Architecture Overview**
• Wide Area Memory Migration
• Wide Area Storage Migration
• Evaluation
Unified Live Migration
XvMotion Architecture

Source

Virtual Machine IO

IO Mirror

Live Migration

IO Buffer

Streams

Destination

Streams

Live Migration

IO Buffer/Writer
Overview

• Architecture Overview
• **Wide Area Memory Migration**
• Wide Area Storage Migration
• Evaluation
XvMotion Architecture: Live Migration

Source
- IO Mirror
- Live Migration
- IO Buffer
- Streams

Destination
- Streams
- Live Migration
- IO Buffer/Writer

Virtual Machine IO → TCP
Live Migration:
Total Time vs Downtime

Source

Memory

Destination
Live Migration:
Total Time vs Downtime

Source

Destination

Memory
Live Migration:
Total Time vs Downtime

Source Memory

Destination
Live Migration: Total Time vs Downtime

- Iterative copy hopefully reduces the working set each iteration
- Depends on Network being faster than Dirty rate
Current Solution: Early Resume/Post-copy

• Problem: Applications can change memory faster than network bandwidth

• Solution:
  • Stop migration copy everything
  • Resume early if downtime is high

• Destination remote page faults on missing pages
Stun During Page Send (SDPS)

- Problems with Early Resume:
  - Remote page faults very expensive for high latency networks
  - Not atomic: Unsafe for WANs where network hiccups are more likely

- Solution:
  Throttle VM just enough to keep dirty rate < network rate
SDPS Downtime Results

<table>
<thead>
<tr>
<th>Users</th>
<th>SDPS OFF</th>
<th>SDPS ON</th>
</tr>
</thead>
<tbody>
<tr>
<td>24-users</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>48-users</td>
<td>2.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

- **SDPS OFF** indicates the downtime when the system is not using the SDPS feature.
- **SDPS ON** indicates the downtime when the system is using the SDPS feature.

The results show a significant decrease in downtime when SDPS is turned on, especially in the 48-users scenario.
Overview

• Architecture Overview
• Wide Area Memory Migration
• **Wide Area Storage Migration**
• Evaluation
XvMotion Architecture: IO Mirroring

Virtual Machine IO

Source

IO Mirror

Live Migration

IO Buffer

Streams

TCP

Destination

Live Migration

Streams

IO Buffer/Writer
Problem: Synchronous mirroring over the WAN slows guest workload

Goal: Hide network latency from VM
XvMotion Buffering: Asynchronous IO

Mirror IO

IO Buffer

Bulk Disk Copy
XvMotion Buffering: Asynchronous IO

Source

VM

Streams Transport Framework

IO Buffer

Destination

Streams Transport Framework

TCP

IO IO IO IO

IO
Problem: Workload may be too fast on source for the destination

Goal: Throttle copy process and workload as necessary
XvMotion Buffering: Congestion Control

Source

VM

Streams Transport Framework

TCP

Destination

Streams Transport Framework

IO Buffer

Send free buffer space

Back Pressure

Limit per-disk OIOs/buffer on destination

Slow Disk
XvMotion Buffering: Congestion Control

- **Source**
  - VM
  - Streams Transport Framework

- **Destination**
  - Streams Transport Framework

- **IO Buffer**
  - Send free buffer space

- **Back Pressure**
  - Limit per-disk OIOs/buffer on destination

- **TCP**

- **Slow Disk**
Overview

• Architecture Overview
• Wide Area Memory Migration
• Wide Area Storage Migration
• Evaluation
Evaluation

- Migration Time: Total end-to-end time
- Downtime: Time machine execution is suspended for final switchover
- Workload Penalty: Average percentage penalty to the VM workload

- OLTP IO Workload (data disk only)
- 10 GB system/12 GB data
XvMotion Downtimes

Take Away: ~1 second downtime independent of distance
Take Away: small linear time increase with distance
California to India Migration

• 1 Gbps network with 214 ms RTT

• OLTP: 68 MB/s disk copy – 89 MB/s memory copy

• ~11% Workload Penalty from Throttling
Summary

• XvMotion frees migration from the need for shared storage and fast local networks
  • Tolerates wide area network bandwidth/latency and reliability
  • Tolerates heterogeneous storage performance
  • Downtimes and workload penalty comparable to local migration
  • Atomic Switchover

• Enables new use cases – e.g. disaster preparedness, cluster upgrade, shared nothing

• On the path to deployment:
  • Deployed in customer metro area networks
  • Cross continent migration e.g. Palo Alto to India is practical
Questions?