Enhancing Live Migration Process for CPU and/or memory intensive VMs running Enterprise applications

Benoit Hudzia
CEC Belfast / SAP Research
08/2011
With the contribution of Aidan Shribman and Petter Svard
Agenda

• Background: Enterprise Applications and Live Migration
• Warm Up
• Delta Compression
• Page Priority
• Future Works
Background

Migrating Enterprise Class applications
Enterprise application and Live Migration

Issues

• **Enterprise class application:**
  • Bigger than average resource requirement
  • Average SAP ERP 16GB + per VM with 32 GB of swap more than common
  • OLTP system such as ERP are very sensitive to time variation.
  • Rely heavily on precise scheduling capabilities, triggers, timers and on the ACID compliance of the underlying

• **Challenge when migrating such application:**
  • Disconnection of services:
    • Gigabit Ethernet timeout \( \approx 5 \text{ seconds} \) (>500 MB memory left in stop and copy phase)
    • Downtime is workload dependent
  • Disruption of services:
    • Migration progressively increasing the amount of resource dedicated to itself \( \Rightarrow \) gradually degrade performance of the coexisting systems / VMs.
  • Difficulty to maintain consistency and transparency
  • Unpredictability and rigidity
Warm Up for Live Migration

Increasing the flexibility of Live Migration
Warm Up
Increasing flexibility

Extended adaptive Pre-copy phase without triggering actual migration
Increased flexibility:
- “just in time” triggering of live migration
- Reduce down time
- Dynamic adaptive bandwidth allocation
- Manual and automatic
- Allow “hot standby”
- Facilitate WAN link transfer
Experimental Results: Warm-up Summary
SAP Sales and Distribution Benchmark

<table>
<thead>
<tr>
<th>Without Warmup</th>
<th>With Warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Service</td>
<td>Memory Transfer before stopping</td>
</tr>
<tr>
<td>VM Stopped</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VM size</th>
<th>4GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMP</td>
<td>2 vCPU</td>
</tr>
<tr>
<td>Users</td>
<td>150</td>
</tr>
<tr>
<td>Load</td>
<td>~80%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CPU</th>
<th>Avg Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>60%</td>
<td>2.18 sec</td>
</tr>
<tr>
<td>Warm-up</td>
<td>73%</td>
<td>2.16 sec</td>
</tr>
</tbody>
</table>

Downtime under load: <1 sec
Success ratio: ~99%
Delta Compression of Page

Limiting the impact of resending Page
Dirty Page Delta Compression

- Cache page with highest dirtying rate during send operation
- Compression Algorithm:
  - XBRLE: XOR + binary run length encoding

Vanilla (no compr.)

Delta compression
Evaluation
Benchmark

• Memory write benchmark (lm_bench)
  • 1 GB RAM, 1 vcpu VM
  • Near ideal case
  • Downtime reduced by a factor of 100
  • Throughput increased by 63%

• Transcoded HD Video (VLC)
  • 1 GB RAM, 1 vcpu VM
  • Real-world, non-ideal case
  • UDP downtime reduced from 8 s to 1
  • Migration is transparent using XBRLE
  • 31% faster, 51% less data sent
Evaluation - SAP ERP
Sales and Distribution benchmark, load 100%

- Non-responsive on resume with vanilla algorithm
- Survived using XBRLE
- >0.5s of downtime = risk of damaging the system
- Measured downtime was 0.2s for XBRLE and 2s for vanilla
- Live Migration Cpu usage directly impact (limit) the available resource for the ERP

SV:8 GB RAM, 4 vcpus VM
App: SAP ERP 7.0 / S&D Benchmark
Page Prioritization

Dynamic page transfer reordering
Dynamic page transfer reordering

Prioritizing page sends (similar to writable working set concept in Xen)
Dynamic page transfer reordering
Prioritizing page sends

Transfer order

Vanilla
Prioritized

- Streaming HD video migration

<table>
<thead>
<tr>
<th></th>
<th>Total migration time</th>
<th>Transferred data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>22.1 s</td>
<td>459 MB</td>
</tr>
<tr>
<td>PRIO</td>
<td>15.4 s</td>
<td>225 MB</td>
</tr>
</tbody>
</table>

- 31% faster, 51% less data sent
Evaluation

Prio vs XBRLE: reveal Cache miss and compression efficiency Issue
Optimizing Compression

Making XBRL more efficient
XBZRL
Increase compression speed /efficiency

• Only compress unmodified data using word aligned encoding and only encodes runs of zeros

• For encoding page diffs XBZRL is:
  • Compression :
    • 20% more efficient than XBRLE
    • 20% less efficient than LZO/Snappy.
  • Speed:
    • Overall 2.5x-5x faster than XOR + LZO/Snappy
    • 11x-9x faster than the original XBRLE

• Doesn’t solve the impact of cache miss
Performance comparison
Synthetic benchmark representing enterprise workload

Higher bandwidth (1778-2286 MB/s)

Lower CPU time

Encoding

Decoding
Performance comparison
Live Migration Benchmark

• Compute capacity used for live migration:
  • **xbzrle**: 50%
  • **vanilla**: between 30%-60%

• Live Migration:
  • **xbzrle**: terminate in seconds
  • **Vanilla**: not able to complete in the allocated time
Future Work
Future Works

• Dynamically disable XBZRLE algorithm if the cache miss ratio is too important

• Combine Page priority algorithm and XBZRLE:
  • Cache page with highest dirtying rate
  • Eliminate unnecessary cache check
  • Eliminate page compression with low potential return
Thank You!

Contact information:

Dr. Benoit Hudzia
Senior Researcher
benoit.hudzia@sap.com
Experimentations Results: S&D Benchmark with/out warm-up

**Response Time (baseline)**

**Response Time (warm-up)**

<table>
<thead>
<tr>
<th></th>
<th>CPU</th>
<th>Avg Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>60%</td>
<td>2.18 sec</td>
</tr>
<tr>
<td>Warm-up</td>
<td>73%</td>
<td>2.16 sec</td>
</tr>
</tbody>
</table>

VM size : 4GB  
SMP : 2 vCPU  
Users : 150  

Downtime under load: <1 sec  
Success ratio : ~99%
Live Migration over emulated WAN Link

Phase 1: Warm-up
- Duration: as long as we want
- Live Migration over emulated WAN Link
- Emulated WAN Link:
  - 10 Mb/s
  - 350 ms latency
  - 50 ms Jitter
  - [1%,5%] packets drop

Phase 2: Pre-copy
- Duration: ~9 min 30 sec
- Not successful (human timeout)
- Probability of Survivability of the SAP system: ~0%

Phase 3: Frozen Transfer
- Duration: ~2.16 sec
- Scenario 1: “Classic Live Migration”
- Scenario 2: “Warm-up + Live Migration”