Predicting Web Service Levels During VM Live Migrations
Helmut Hlavacs and Thomas Treutner
Research Group Entertainment Computing, University of Vienna, Austria

Introduction
► VM live migration important for energy efficiency
► Enables us to establish energy efficient target distribution of VMs
► Supposedly no perceivable service downtime while live migrating
► Live migration is resource intensive (iterative page copying)
► Experiments: Influence on service levels while migrating?
► Modelling: Predict service levels based on utilization?

Scenario
► Virtualized data center, static consolidation (P2V)
► Provisioning for peak load, still bad energy efficiency
► E.g., 9-5-cycles, very low utilization at night
► Live migration enables dynamic consolidation
► But: Seldom used, fear of possible side effects
► Identify and quantify effects on (web) service levels
► Find most influencing utilization metrics

Experiment
► Two servers, a single VM, migrating forth and back
► VM disk image on central node (Gbit, open-iscsi)
► qemu-kvm VM: Linux, Apache2, PHP5, MediaWiki
► VM disk image on central node (Gbit, open-iscsi)
► Live migration enables dynamic consolidation
► But: Seldom used, fear of possible side effects
► Identify and quantify effects on (web) service levels
► Find most influencing utilization metrics

Data Overview
► We can interpret the UNIX load as approximation to Q, the average number of jobs in a Markovian M/M/1 queue, and the VM's CPU utilization as ρ, the system utilization: Q_{M/M/1} = \frac{1}{1-\rho}
► UNIX load is exponentially averaged by definition and the service times are not necessarily exactly exponentially distributed: Q_{M/G/1} = \frac{1}{1-\rho} \left(1 + \frac{\rho}{\mu}ight)
► For deterministic service times \mu = 0, resulting in Q_{M/D/1} = \frac{1}{1-\rho} \left(1 + \frac{\rho}{2}\right)
► Simple linear regression delivers the coefficient \hat{c}_2 = 0.42
► P(T \leq x) = F_T(x) = 1 - e^{-\mu(1-\rho)x}, the theoretical probability that the response time T is lower than or equal to a limit x for a given service rate \mu

Service Levels for Low/Medium/High Workload Scenarios
► Low (top right): Slightly increased response times during live migration, seldom response time violations
► Medium (bottom left): SLA ratio generally satisfies the 97% limit
► High (bottom right): Often and heavy violations, unacceptable low service levels, typically decreased by 20-25% percentage points

Model Selection
► Stepwise model selection: Akaike Information Criterion
► Finds trade-off between number of parameters (model size) and goodness of fit (model quality)
► For comparison: Exhaustive all-subsets-regression (LEAPS)
► LEAPS: Find best of all possible models for given range of model size
► Computationally intensive even if number of variables is limited

Most Influencing Model Parameters
<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>R^2 Adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td></td>
<td>2.995e-00</td>
<td>5.69e-01</td>
<td>1.00e-00</td>
</tr>
<tr>
<td>m(load)</td>
<td>VM load5</td>
<td>-1.87e-02</td>
<td>5.87e-01</td>
<td>&lt; 2e-10</td>
</tr>
<tr>
<td>s_swap</td>
<td>VM swap used</td>
<td>-7.89e-07</td>
<td>8.83e-06</td>
<td>&lt; 2e-10</td>
</tr>
<tr>
<td>s_swap</td>
<td>reserved_mem</td>
<td>4.85e-04</td>
<td>6.85e-04</td>
<td>&lt; 2e-10</td>
</tr>
<tr>
<td>s_load</td>
<td>Tasks created</td>
<td>-2.5e-03</td>
<td>3.87e-03</td>
<td>&lt; 2e-10</td>
</tr>
<tr>
<td>s_mem</td>
<td>CPU Mem used</td>
<td>1.194e-01</td>
<td>4.13e-01</td>
<td>&lt; 2e-10</td>
</tr>
<tr>
<td>s_swap</td>
<td></td>
<td>4.1e-02</td>
<td>3.18e-02</td>
<td>&lt; 2e-10</td>
</tr>
<tr>
<td>s_load</td>
<td>CPU Mem used</td>
<td>9.97e-03</td>
<td>2.33e-02</td>
<td>&lt; 2e-10</td>
</tr>
<tr>
<td>s_swap</td>
<td></td>
<td>1.5e-02</td>
<td>1.92e-02</td>
<td>&lt; 2e-10</td>
</tr>
</tbody>
</table>

Conclusions
► Impact of live migration on SL depends on amount of workload
► Tighter SLAs can be fulfilled during low and medium workload
► Migrating during high load causes massive decrease of service level
► Service level variance during a live migration to 90% predictable
► Using only a single variable, the UNIX load5 average, models with 12 variables can explain 95% of variance
► Systems using live migration as a mechanism to realize a more energy efficient target distribution and have service level targets need to consider the UNIX load average, but typical hypervisors do not collect/export this information
► Hypervisors should be extended to export load information (cf. free memory)

Future Work
► Influence of additional VMs (idle, utilized, mixed)
► Linux and qemu-kvm: Kernel Semapge Merging (KSM)
► Database VM migration, currently taboo due to potentially severe influence
► qemu-kvm parameters: Bandwidth limits, maximum allowed downtime
► Predict migration delay, energy consumption, service downtime

http://cs.univie.ac.at/research/research-groups/entertainment-computing/