PARVIS - Performance mAnagement of Vlrutualized Systems

Danilo Ardagna

joint work with Mara Tanelli and Marco Lovera

Dipartimento di Elettronica e Informazione, Politecnico di Milano
ardagna@elet.polimi.it

Milan, November 23 2010
Nowadays, large Data Centers provide computational capacity on demand to many customers by sharing a pool of IT resources:

- IBM Cloud
- Amazon EC2
- Windows Azure
- ...

Issues:

- Workload variability and QoS guarantees
- Energy consumption
Workload variability

- Requests rates may change by order of magnitudes with a business day
- Requests rates may change by order of magnitudes with a business day

- Traffic surges
Workload variability

- Requests rates may change by order of magnitudes with a business day
- Traffic surges
- Sport events
Data Center energy consumption: An environmental problem...

About 0.5% of global electric power consumption is due to DC

In developed country:
- UK: 2.2-3.3%
- USA: 1.5%

From the environmental point of view:
- 2% of global CO₂ emissions

Source: EU Commission
Data Center energy consumption: ...but first an economic one

<table>
<thead>
<tr>
<th>Company</th>
<th>Server</th>
<th>Electric Power (TWh)</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>eBay</td>
<td>16K</td>
<td>0.06</td>
<td>$3.7M</td>
</tr>
<tr>
<td>Microsoft</td>
<td>&gt;200K</td>
<td>0.6</td>
<td>&gt;$36M</td>
</tr>
<tr>
<td>Google</td>
<td>&gt;500K</td>
<td>6</td>
<td>&gt;$38M</td>
</tr>
</tbody>
</table>

New Servers costs

IT costs (US$)

Energy and cooling costs

Source: EU Commission

Basic 1U Server - 5 year TCO

- 35% Total equipment cost per server
- 28% Power cost per server
- 25% Data Center cost per server
- 12% Data Center operating cost per server

Amortized Datacenter Infrastructure costs account for ~25% of the Server TCO

Server Power Consumption accounts for ~25% of Server TCO

Source: Microsoft Research
DC Inefficiencies

Courtesy of IBM
Virtualization of Physical Resources

- Virtualization, proposed in early ’70s, is driving again the interest both of industry and academia

- Enabling technology for server consolidation and cloud computing

- Advantages:
  - Physical resources are partitioned among competing running VMs, improved security and reliability, performance isolation
  - Resource allocation parameters can be updated by in few milliseconds without introducing any system overhead
Hardware resources (CPU, RAM, ecc...) are partitioned and shared among multiple **virtual machines** (VMs)

The virtual machine monitor (VMM) governs the access to the physical resources among running VMs
Virtualization of Physical Resources: Research Challenges

- Performance modelling of virtualized environments is challenging

- Traditional queueing network models are inadequate to model virtualized systems performance at a very fine-grained time scale
PARVIS goals

- Develop novel resource allocation policies virtualized cloud infrastructures via an interdisciplinary approach:
  - Performance evaluation and optimization methods for the long-term management of the physical infrastructure
  - System identification and control engineering methods to derive load-dependent black-box models of virtualized systems and to design short-term control systems
PARVIS Data Center: Autonomic resource management

Application\textsubscript{1}  
Application\textsubscript{2}  
Application\textsubscript{3}  

Internet

Free Server Pool
PARVIS Reference framework

- **Virtual Machine Monitor**
  - VM1
  - VM2
  - VMn
  - App1
  - App2
  - Appn
  - S.O.

- **Infrastructure controller**
  - Long term time horizon
  - Queuing network models → Non linear optimization
  - Time scale: ten minutes/hour

- **Local controller**
  - Short term time horizon
  - Dynamic models → Control theory
  - Time scale: minute/seconds

- **Performance metrics**
- **Performance goals**
- **Controller**
- **System**
- **DFS**
- **CPU weights**
- **Admission control**

- **Workload partitioning**
- **Performance goals of individual servers**

- **Fine grained performance and energy consumption goals**
PARVIS Reference framework

- **Infrastructure controller:**
  - Mixed Integer Non Linear Problem
  - Local Search

- **Local controller:**
  - Linear Parameter Varying Models
  - Model Predictive Controllers


Revenues are a function of average response times.

Average response time soft-constraint
Service Center Performance Model

- Open queueing network model: heterogeneous service centers and a delay center
- VMM modelling: GPS (Generalized Processor Sharing) scheduling

Service centers model physical servers which support VMs execution

A class $k$ request $\xi$ becomes a request $k'$ with probability $p_{k,k'}$ or terminates

Service centers model physical servers which support VMs execution

Hexogenous arrival rate

$\lambda_k$ to Dispatcher

Dispatcher to HTTP servers tier, Application servers tier, DBMS servers tier

A class $k$ request $\xi$ becomes a request $k'$ with probability $p_{k,k'}$ or terminates

Service centers model physical servers which support VMs execution
Optimization Problem

- Objective: maximize SLA revenues minus energy costs
- Decision variables:
  - $x_i$: server $i$ ON/OFF (binary variable)
  - $\lambda_{i,j}^k$: server $i$ arrival rate for the VM operating at tier $j$ of request class $k$
  - $\phi_{i,j}^k$: server $i$ CPU capacity fraction devoted to the VM operating at tier $j$ of request class $k$
  - $z_{i,j}^k$: assignment of the VM operating at tier $j$ of request class $k$ to server $i$ (binary variable)
  - $f_{i,h}$: server $i$ operating frequency (binary variable)
Heuristic solution based on problem decomposition:

- Initial solution: Assign VMs to physical servers (problem equivalent to a special case of a CFLP, Capacitated Facility Location Problem)
- Optimum load balancing and capacity allocation (fixed point iteration)
- The solution is then enhanced by a local search:
  - Switch servers ON and OFF
  - Change VMs placement
  - Change servers’ CPU frequency
Linear Parameter Varying (LPV) systems are a class of time-varying systems in discrete-time state space form:

\[
\begin{align*}
x_{k+1} &= A(p_k)x_k + B(p_k)u_k \\
y_k &= C(p_k)x_k + D(p_k)u_k
\end{align*}
\]

“Time varying systems, the dynamics of which are functions of a measurable, time varying parameter vector p.”
LPV state-space models

- Virtualized system identification:
  - Scheduling parameters: arrival rates, requests service times
  - Output variables: requests response times
  - Control variables: VMM parameters
• Real log traces (Politecnico di Milano Web site), 10 requests classes
• Comparison with IBM Tivoli resource allocation policies

Scenario 1: users come from the same time zone

Scenario 2: users come from different time zones
IBM Tivoli comparison – scenario 1

Our solution

IBM Tivoli

Our solution
IBM Tivoli comparison – scenario 2

Our solution

IBM Tivoli

IBM Tivoli

Our solution
System Identification - Experimental setting

- Two reference scenarios:
  - A Micro benchmarking instrumented Web application
  - SPECweb2005 industrial benchmark

- VMM monitor: Xen 3.0 and Xen 3.3

- Validation: Synthetic workload inspired by a real-world. Log trace from a large financial system
**Micro-benchmarking Web Service Application Experiments**

- Number of VMs varied between 2 and 4
- For system identification purposes request arrival rates vary stepwise every 1 minute
- Each request consumes $s_{i_k}$ CPU time varied between 0.06 s and 1.1 s.
- 1,440 intervals (24 hours)
- Parametrization $[s_{i_k} \rho_{i_k}]$
# Micro-benchmarking Web Service Application Experiments

<table>
<thead>
<tr>
<th>2 VMs</th>
<th>VAF on 24h</th>
<th>VAF light-load</th>
<th>VAF heavy-load</th>
<th>$\epsilon_{avg}$ on 24h</th>
<th>$\epsilon_{avg}$ light-load</th>
<th>$\epsilon_{avg}$ heavy-load</th>
</tr>
</thead>
<tbody>
<tr>
<td>VM1</td>
<td>74.85%</td>
<td>74.8%</td>
<td>86.15%</td>
<td>6.5%</td>
<td>8.3%</td>
<td>4.07%</td>
</tr>
<tr>
<td>VM2</td>
<td>70.56%</td>
<td>67.9%</td>
<td>83.94%</td>
<td>3.4%</td>
<td>2.5%</td>
<td>4.17%</td>
</tr>
<tr>
<td>VM3</td>
<td>78.5%</td>
<td>75.19%</td>
<td>83.05%</td>
<td>6.9%</td>
<td>1.5%</td>
<td>12.76%</td>
</tr>
<tr>
<td>VM4</td>
<td>70.75%</td>
<td>69.14%</td>
<td>83.3%</td>
<td>2.76%</td>
<td>16.55%</td>
<td>18.21%</td>
</tr>
<tr>
<td>3 VMs</td>
<td>VAF on 24h</td>
<td>VAF light-load</td>
<td>VAF heavy-load</td>
<td>$\epsilon_{avg}$ on 24h</td>
<td>$\epsilon_{avg}$ light-load</td>
<td>$\epsilon_{avg}$ heavy-load</td>
</tr>
<tr>
<td>VM1</td>
<td>78.46%</td>
<td>75.51%</td>
<td>88.78%</td>
<td>6.5%</td>
<td>3.27%</td>
<td>19.58%</td>
</tr>
<tr>
<td>VM2</td>
<td>78.5%</td>
<td>75.19%</td>
<td>83.05%</td>
<td>6.9%</td>
<td>1.5%</td>
<td>12.76%</td>
</tr>
<tr>
<td>VM3</td>
<td>70.75%</td>
<td>69.14%</td>
<td>83.3%</td>
<td>2.76%</td>
<td>16.55%</td>
<td>18.21%</td>
</tr>
<tr>
<td>VM4</td>
<td>75.97%</td>
<td>66.37%</td>
<td>88.97%</td>
<td>6.91%</td>
<td>10.35%</td>
<td>3.16%</td>
</tr>
<tr>
<td>4 VMs</td>
<td>VAF on 24h</td>
<td>VAF light-load</td>
<td>VAF heavy-load</td>
<td>$\epsilon_{avg}$ on 24h</td>
<td>$\epsilon_{avg}$ light-load</td>
<td>$\epsilon_{avg}$ heavy-load</td>
</tr>
<tr>
<td>VM1</td>
<td>75.97%</td>
<td>66.37%</td>
<td>88.97%</td>
<td>6.91%</td>
<td>10.35%</td>
<td>3.16%</td>
</tr>
<tr>
<td>VM2</td>
<td>72.72%</td>
<td>68.1%</td>
<td>87.23%</td>
<td>9.14%</td>
<td>12.23%</td>
<td>5.24%</td>
</tr>
<tr>
<td>VM3</td>
<td>74.85%</td>
<td>69.22%</td>
<td>82.94%</td>
<td>5.84%</td>
<td>7.29%</td>
<td>4.38%</td>
</tr>
<tr>
<td>VM4</td>
<td>66.80%</td>
<td>62.38%</td>
<td>89.06%</td>
<td>11.99%</td>
<td>16.92%</td>
<td>4.77%</td>
</tr>
</tbody>
</table>
SPECweb2005 Experiment

- Two VMs running the banking and e-commerce loads
- The number of users $N_{ik}$ accessing each of the two VMs varied stepwise every 1 minute, with values between 10 and 220
- Proportional assignment scheme:
  \[ \phi_k^i = \max \left( 0.1, \frac{N_k^i}{N_k^1 + N_k^2} \right) \]
- 1,440 intervals (24 hours)
- Parametrization $[N_k^i, \rho_k]$
SPECweb2005 Experiment

<table>
<thead>
<tr>
<th>Identification data</th>
<th>VAF on 24h</th>
<th>$\epsilon_{avg}$ on 24h</th>
</tr>
</thead>
<tbody>
<tr>
<td>VM1</td>
<td>59.85%</td>
<td>6.85%</td>
</tr>
<tr>
<td>VM2</td>
<td>87.20%</td>
<td>6.97%</td>
</tr>
<tr>
<td>Validation data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VM1</td>
<td>64.51%</td>
<td>7.34%</td>
</tr>
<tr>
<td>VM2</td>
<td>77.60%</td>
<td>10.63%</td>
</tr>
</tbody>
</table>
PARVIS future work

- Analysis of real applications
- Local controller design
- Integration of the two approaches
Mara Tanelli, Marco Lovera, Barbara Panicucci, Marco Bergamasco, Alessandro Barenghi, Alessandro Colleoni, Bernardetta Addis, Giuliana Carello, Antonio Capone, Politecnico di Milano

Folco Bombardieri, Danilo Ghirardelli, Gianluca Pisati, Giovanni Pirotta, Marco Caldirola, Mauro Speroni, Paolo Sala, Marco Casiero, Stefano Vettor, Massimo Bergami, Politecnico di Milano students

Marco Trubian, Università degli Studi di Milano

Li Zhang, IBM Research

Massimo Leoni, Carla Milani, IBM Italy