Survey on resource management for Cloud computing environment (*)

Nguyen Quang Hung
Email: hungnq2@cse.hcmut.edu.vn
Outline

• What is Cloud computing (CC)?
• Cloud computing and Grid computing (GC) on 360° comparison
• State-of-the-art
• Challenges on efficient resource management
• Future works
• Conclusions
• References
What is Cloud computing?

• Definitions for the Cloud computing:
  
  o ...a “cloud” refers to an “Infrastructure-as-a-Service” (IaaS) cloud, such as Amazon EC2, where IT infrastructure is deployed in a cloud provider’s datacenter in the form of virtual machines, Ian Foster, et.al. (2010)
What is Cloud computing?

• **Definitions for the Cloud computing**:
  
  o “...a “cloud” refers to an “Infrastructure-as-a-Service” (IaaS) cloud, such as Amazon EC2, where IT infrastructure is deployed in a cloud provider’s datacenter in the form of virtual machines, Ian Foster, et.al. (2010) [26]

  o “…a Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers.”, Rajkumar Buyya, et.al. (2009) [2].
What is Cloud computing?

• Definitions for the Cloud computing [26]:
  o ...a “cloud” refers to an “Infrastructure-as-a-Service” (IaaS) cloud, such as Amazon EC2, where IT infrastructure is deployed in a cloud provider’s datacenter in the form of virtual machines, Ian Foster, et.al. (2010) [26]
  o “…a Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers.”, Rajkumar Buyya, et.al. (2009) [27][2].
  o “A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet....” [1]
Cloud computing has ...(con’d)

• Taxonomies of Cloud Computing [12, 20]:
  o Infrastructure-as-a-Service (IaaS)
    • Amazon EC2 [6], ...
  o Platform-as-a-Service (PaaS)
    • Google AppEngine [17], ...
  o Software-as-a-Service (SaaS)
    • Gmail, SalesForce,...
  o Data-as-a-Service (DaaS)
    • Strikeiron.com, Kognitio.com
## Cloud computing and Grid computing

### 360-degree compared

<table>
<thead>
<tr>
<th>Grid computing/Grid systems</th>
<th>Cloud computing/Cloud systems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Compute &amp; storage is packed as metered services can be charge pay-as-you-go as electricity, telephone network.</td>
</tr>
<tr>
<td>- Lesser scale than Supercomputer and Cloud</td>
<td>- Large scale and Web 2.0 based wholly.</td>
</tr>
<tr>
<td>- Distributed paradigm or infrastructure spans across multiple virtual organizations.</td>
<td>- Single organization (e.g. Amazon S3 &amp; Amazon EC2)</td>
</tr>
</tbody>
</table>
Cloud computing and Grid computing: 360-degree compared
Cloud computing and Grid computing
360-degree compared

<table>
<thead>
<tr>
<th>Criterias</th>
<th>Grids</th>
<th>Clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business model</td>
<td>- Project-oriented, limited CPU hours.</td>
<td>- Unlimited use, pay-as-you-go as electricity, gas, telephone</td>
</tr>
<tr>
<td>Architecture</td>
<td>- 1990s</td>
<td>- 2000s</td>
</tr>
<tr>
<td></td>
<td>- integration of network commodity resources</td>
<td>- Developed to address Internet-scale computing problem. i.e. a large pool of compute and storage resources can be access by common protocols (e.g. web services)</td>
</tr>
</tbody>
</table>
Cloud computing and Grid computing: 360-degree compared

Grid Protocol Architecture

Cloud Architecture

Nguyễn Quang Hùng
Cloud computing and Grid computing
360-degree compared (cont’d)

<table>
<thead>
<tr>
<th>Criterias</th>
<th>Grids</th>
<th>Clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Management:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Compute Model</td>
<td>Batch-job management</td>
<td>Shared resources by all users at the same time</td>
</tr>
<tr>
<td></td>
<td>-Portable Batch System (PBS), Condor, SGE, LSF,…</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Dedicated to queuing system</td>
<td></td>
</tr>
<tr>
<td>- Data Model</td>
<td>-Data Grid</td>
<td>-Data security &amp; privacy, big data problem (TB),…</td>
</tr>
<tr>
<td>- Data Locality</td>
<td>-Even harder than Cloud</td>
<td>-Hardly Google’s MapReduce on Google File System</td>
</tr>
<tr>
<td></td>
<td>- Shared data stores on NFS/GPFS/PVFS/Luster</td>
<td>-Need scheduler to be data-aware</td>
</tr>
<tr>
<td></td>
<td>- Need scheduler to be data-aware</td>
<td>-Need scheduler to be data-aware</td>
</tr>
</tbody>
</table>
Cloud computing and Grid computing
360-degree compared (cont’d)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Grids</th>
<th>Clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Management:</td>
<td>In progress of works on scheduler data-aware</td>
<td>Not yet</td>
</tr>
<tr>
<td>- Combining compute and data management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Virtualization</td>
<td>- Not need as much as Cloud on Virtualization</td>
<td>- Need Virtualization</td>
</tr>
<tr>
<td>- Monitoring</td>
<td>- Mostly physical resource</td>
<td>- Hard to fully monitor resources and services on cloud systems</td>
</tr>
<tr>
<td>- Provenance</td>
<td>- Built-in workflows systems: Chimera, Swift, Keepler, Tavena,…</td>
<td>- More difficult than in Grids</td>
</tr>
</tbody>
</table>
Cloud computing and Grid computing 360-degree compared (cont’d)

<table>
<thead>
<tr>
<th>Criterias</th>
<th>Grids</th>
<th>Clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Management:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Programming Models</td>
<td>- Similar to parallel &amp; distributed computing: MPI, MPICH-G2, GridRPC, Pop-C++,…</td>
<td>- MapReduce - Mash-up and scripting</td>
</tr>
<tr>
<td>- Security Models</td>
<td>- Across many VOs - Single Sign-On</td>
<td>- Clouds mostly is dedicated data centers belong to one organization. - SSL based</td>
</tr>
</tbody>
</table>
Challenges on resource management in Clouds

• On Cloud provider’s view:
  o Provision resources on both HPC batch-job requests w/wo deadlines and advanced reservation requests on the same system.
  o Energy efficient resource management in Data Centers.

• On Cloud service provider’s view:
  o Renting cheapest resources on performance-constraints
  o QoS to their cloud users

• On Cloud user’s view:
  o Renting cheapest resources on performance-constraints
  o Cloud provider guarantees Service Level Agreement
Problems

• Problem 1:
  o Provision resources on both HPC batch-job requests w/wo deadlines and advanced reservation requests on the same system.

• Problem 2:
  o Energy efficient resource management in Data Centers.
The problem of provisioning computational resources

The problem of provisioning computational resources

• Problem:
  o Short time need for computational resources
  o Conflicts on optimized functions: Batch-jobs and Advanced Reservation (AR)
Some cases on need for computational resources

- The need for computational resources many is in short time [3], for examples:
  - “A scientist requires a large number of computers to run a simulation in hours”
  - “A college instructor needs a cluster to teach MPI labs”
  - “A company to host a website which can scale up/down on traffics”
  - Etc.
Challenges on provisioning computational resources

• How to provision shared computational resources efficiently?
• How is resource provisioning model and architecture for multiple scenarios efficiently and simultaneously?
State-of-the-art on provisioning computational resources

• VM-based approaches:
  o Virtual Clusters (Nishimura et al, Yamasaki et al, Nimbus toolkit,...)

• Batch-jobs approaches
  o PBS, SGE, Condor,...

• Lease based approaches [3-6]
  o B. Sotomayzor (2010)

• Energy-efficient resource management [7-8]
  o Green Cloud
B. Sotomayor (2010)

• A lease-based model: [3-6][9]
  o Lease as a fundamental abstraction
  o VM as a implementation vehicle
  o Concern on overhead in using VM
  o Single domain administration

• High performance computing driven

• Batch-job and AR simulationecessarily
  o Preemptive jobs with/without deadlines
  o Non preemptive AR requests
Resource and Leasing model

- Assumptions:
  - Computational resources in a site
  - $P = \{\text{Set of leasable computer, nodes, in a site}\}$
  - Leasable resources within a node can be allocated to one or more leases, up to maximum capacity, and may include processors, memory, disk space, network bandwidth, etc.
  - $\Gamma_R = \{\text{memory, disk, ...}\}$: set of the types of leasable resources in a site
  - $r$ : quantity of a resource
  - A site define as: $\Gamma_R=\{\text{proc, mem, disk, net-in, net-out}\}$
Lease state machine
Scheduling algorithms


Problem 2:
Energy-efficient resource management for Data Center

• Challenges:

• Data center owner:
  o Minimize power/energy consumption within QoS (e.g. performance) constraints.
State-of-the-art on Energy-Efficient resource management


Virtualization

• One of the most important underlying technology in cloud is the use of virtualization.

![Virtual Machine Abstraction](image)
Green computing

- Dynamic voltage and frequency scaling (DVFS)

Fig. 3. Possible energy to performance trade-off. Here you can see a 18% reduction in frequency contributes to only a 5% performance loss. [28]

Green Cloud framework

Fig. 4. Green Cloud Framework. Shaded items represent topics discussed in this paper.
Power-aware scheduling

- An experiment #cores and watts on Intel Core i7 920 CPU [Fig. 5 in [7]]
Power-aware scheduling

• Change in power consumption decreases.
• Greedy-based algorithm
Greed-based algorithm

Algorithm 1 Power based scheduling of VMs

FOR $i = 1$ TO $i \leq |pool|$ DO
    $pe_i =$ num cores in $pool_i$
END FOR

WHILE (true)
    FOR $i = 1$ TO $i \leq |queue|$ DO
        $vm = queue_i$
        FOR $j = 1$ TO $j \leq |pool|$ DO
            IF $pe_j \geq 1$ THEN
                IF check capacity $vm$ on $pe_j$ THEN
                    schedule $vm$ on $pe_j$
                    $pe_j - 1$
                END IF
            END IF
        END FOR
    END FOR
    wait for interval $t$
END WHILE
VM image management

- Idle physical machines can be dynamically shutdown and restarted.
- Live migration
- Wake on LAN

Fig. 6. Virtual Machine management dynamic shutdown technique
VM image design

- VM size
- Boot time: light weight VM
The result

Fig. 7. Illustration of Scheduling power savings
Conclusions

• This slides have presented the state-of-the-art on resource management on cloud computing. Focus on:
  o B. Sotomayor’s work on provision computational resources using Virtual Machines and Leases
  o Energy-efficient resource management for cloud computing environment

• Overview on Cloud computing and Cloud & Grid 360-degree comparison.
References


Thank you!

• Questions?