CAPACITY MANAGEMENT ON CLOUD

Twisha, Amit Kr. Gautam, Jayant Sharma

1, 2 Assistant Professor, Cluster Innovation Centre, University of Delhi, India
3 Assistant Professor, University of Delhi, India

ABSTRACT

Cloud Computing allows users to modify number of resources in real time thus focusing on computing as a utility that can be access over the Internet. We saw various phases starting from Centralized Bulky systems to Distributed Networks and now the emergence of Cloud Computing which is based on Centralized Systems with Internet as the access. Here the vendor i.e. the owner of Cloud is authenticated to provide services to user as per his demand. These Clouds are location-independent and have a flexible pricing model for the user. But for efficient use of these services, these should be properly managed using Capacity Management as a tool. Cloud Computing is only advantageous if Cloud Services are efficiently used.

Keywords: Cloud Computing, Capacity Management, Pay-as-you-go, Capacity Planning and Virtual Cloud.

I. INTRODUCTION

Cloud Computing provides resources via internet, thus appearing to user as if the Cloud has infinite capacity and pay as per the user demands. It mainly aims at the feature of ‘Shared Infrastructure’, ‘On-Demand-Service’ and ‘Pay-As-You-Go’. The advantages of Cloud Computing are:

- ‘Doing-more-With-Less’, this is with regard to infrastructure.
- User can use the service without knowing the underlying technology
- Enabling customer to use the resources as per need
- ‘Pay-for-What-you-use’

Deployment models in cloud computing are Private Model; Public Model; Hybrid Model.
II. ECONOMICS

Cloud Computing is highly expensive if proper monitoring is not adapted. As per the economic statistics available Cloud Computing is almost 2.83 times expensive as compared to in-house server without incorporating capacity management.

Let’s compare an actual cost of a cloud server as compared to the equivalent in-house infrastructure running within an enterprise. As an example of ‘extra-large’ server on EC2 cloud having 15GB of memory and 8 compute units. It costs $0.68 per CPU-hour to the customer. Purchasing an equivalent server for in-house deployment costs $9500(approx). This price if for an x86 server with 3 quad-core CPUs i.e. 12 processing cores, 12GB memory and 300GB disk. If we run this in-house server for 3years i.e. 26260 hrs (approx), then the cost per hour will be $0.36/hr over 3years. This cost compared to the cost of cloud server ($0.68 /hr), which is 2.83 times the cost of an in-house server. Even if we discount the fact that in-house server has 12 cores compared to EC2 server’s 8 cores, the cloud server costs 1.88 times expensive.

As per the Table 1 cloud server is more expensive than an equivalent in-house server. But now we consider the fact that most servers are heavily underutilizes, say about 40 percent, whereas cloud server has much more utilization, say 80 percent. Factoring this in, the Cost/effective-hour comes out to be $0.90 (i.e. cost /hr divided by the efficiency), whereas the same for a server deployed in cloud comes out to be $0.85. Than adding the management cost, power & cooling cost, etc in In-house server and the Management Cost in Cloud server, we can conclude that In-house Server is 1.5 times (approx.) expensive than the cloud server. But this benefit can only be achieved if we incorporate proper Capacity Management & Monitoring in the cloud server.

Table 1: Economics of Servers

<table>
<thead>
<tr>
<th></th>
<th>In-House server</th>
<th>Cloud Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purchase Cost</td>
<td>$9,600</td>
<td></td>
</tr>
<tr>
<td>Cost/hr(over 3yrs)</td>
<td>$0.36</td>
<td>$0.68</td>
</tr>
<tr>
<td>Price(Cloud/In-House)</td>
<td>$1.88</td>
<td></td>
</tr>
<tr>
<td>Price of core-hr</td>
<td>$0.03 (0.36/12)</td>
<td>$0.09</td>
</tr>
<tr>
<td>Efficiency</td>
<td>40%</td>
<td>80%</td>
</tr>
<tr>
<td>Cost/Effective-hr</td>
<td>$0.90</td>
<td>$0.85</td>
</tr>
<tr>
<td>Power &amp; cooling</td>
<td>$0.36</td>
<td></td>
</tr>
<tr>
<td>Management Cost</td>
<td>$0.10</td>
<td>$0.01</td>
</tr>
<tr>
<td>Total Cost/Effective-hr</td>
<td>$1.36</td>
<td>$0.86</td>
</tr>
<tr>
<td>CostRatio(In-house/Cloud)</td>
<td></td>
<td>1.58 times expensive</td>
</tr>
</tbody>
</table>

III. PRACTICAL ISSUES

Amazon is the first cloud provider and it faced the following issues:

1. Complexity of the application suite
2. Peaks and throughs of seasonal retail business necessities were there. So, regular load monitoring and provisioning additional capacity on demand was required.
Using a public cloud offering astrict the need to plan ahead for peak load and converts fixed costs into variable costs that change with actual usage, thereby eliminating wastage. In contrast, provisioning for peak capacity is a necessity within private data centers. Without adequate scale there is limited opportunity to amortize capacity across different applications.

IV. SOLUTION: CAPACITY MANAGEMENT

Vendors can address client’s challenges by:

- Prioritizing workloads
- Maximizing business return by identifying well suited applications
- Addressing problematic workloads
- Identifying potential problems & addressing them before migration
- Important driving force for cloud economics is the way hardware capacity is planned for and purchased

Using a public cloud offering therefore obligates the need to plan ahead for peak load and converts fixed costs into variable costs that change with actual usage, thereby eliminating wastage. Thus this can be achieved by Capacity Management. Capacity Management is concerned with the following:

1. Monitoring the performance, throughput and load on the system.
2. Analyzing the performance of the system.
3. Tuning the performance of the system as the existing requirements
4. Future planning/ Prediction as per the demands
5. Capacity Planning

Time and Capacity are the two potential constraints in Capacity Management. Considering Cloud Computing, both these are major constraints to be taken care of.

For an example, suppose we have a cloud server which is used by 5 users (as shown below).

We have the free space of 30%, but now the user B and user D leaves the space they have. So, the space has to be lent out to other customers for cost profit. For this forecasting of the amount of data which will be available in future will be predicted by the Capacity Management team. It will monitor the present load on the system and will plan for future. If this analysis and planning is not there, then the cloud server will have the following view:
As per the chart, 70% of the space is free, thus major part of the server is wasted. It is often argued that using virtualization and automatic provisioning technologies in-house it should be possible to create a virtual cloud where these inefficiencies would disappear and benefits of cloud computing could be achieved on premises. But if there is insufficient planning then shifting to cloud server is a total waste, as can be seen in the example above.

V. TECHNIQUE USED

Let $x_1, x_2, x_3 \ldots x_n$ be the respective demands of applications ‘1’ to ‘n’. Also let $X$ be the summation of demands of all running applications, i.e.

$$X(t_1) = x_1 + x_2 + x + \ldots + x_n \text{ at a particular time } t_1$$

& the maximum of $X()$ be $X_{\text{max}}$, i.e.

$$X_{\text{max}} = \max(X(t(x))) \text{ over time ‘t’}$$

Variation in $X$ and $X_{\text{max}}$ is affected by running of various applications. So if the difference between $X$ and $X_{\text{max}}$ is not much at maximum number of time then there is no use of providing dynamic availability of the demands by various applications, so in this case $X_{\text{max}}$ can be provided by the cloud server.

On the other hand if there is major difference between $X$ and $X_{\text{max}}$ at maximum number of times then the demand should be fulfilled dynamically which should be the actual demand at that time with summation of a fixed variable (i.e. time taken to provide the demand) and the rate of change of $X$. 

**Figure: Cloud Distribution**

**Figure: Cloud Over-Provisioning: Best Case**
Figure: Cloud Over-Provisioning: Worst Case

So, the total virtual capacity that should be provided by the cloud service provider should be:

\[ VC(t) = X(t) + \delta \times \frac{dX}{dt}, \text{where } '\delta' \text{ is the provisional delay} \]

VI. CONCLUSION

When dealing with a set of resources in cloud, capacity management can be metaphorically viewed as making sure you have a big enough bucket to hold all of the water the business needs to slake its thirst. There are boundaries on this pool of resources. Capacity management under this "bucket" approach tends to center on predicting peak usage needs and ensuring that in-house data centers can meet those demands.

REFERENCES