How Small and Medium Enterprises (SMEs) Should Bid for Spot Instances of Amazon’s EC2 Cloud

Debashis Saha, Indian Institute of Management, Kolkata, India

ABSTRACT

In cloud service provisioning, spot instances are spare slots for which it has no pre-booking, unlike reserved or on-demand instances for which a cloud service provider (CSP) has a priori booking. CSPs like Amazon prefer spot instance approach to sell their “idle” computing resources as and when these idle slots appear. Though they price the spot instances dynamically depending on supply-demand status, usually the spots instances are relatively cheap. Hence, Amazon’s spot instances are an attractive option for IT managers in small and medium enterprises (SMEs) that normally have sporadic requirements for resources. However, SMEs have to win their desired spot instances through the auction mechanism conducted by Amazon. Since the IT manager always looks for finishing her job quickly within some specified budget, finding how to bid for spot instances in order to stay within its limited budget is a challenging task for her. She may continue to consume spot instances as long as her bid exceeds the current spot price. But, if she loses at any point, the unfinished task must be put on hold by some checkpointing mechanism so that the task may resume from the same point when she wins the spot next time. Using simulations for a very popular cloud, namely Amazon EC2, it has been found that, at a lower bid price, OPTIMAL checkpointing leads to a total cost higher than the total HOURLY checkpointing cost on a much higher bid value. Therefore, SMEs should go for higher bid prices when using OPTIMAL checkpointing and lower bid prices with HOURLY checkpointing. In the process, the author has observed some interesting correlation among checkpoint strategy, task reliability and completion time, which is reported here.

Keyword: Bidding, Cloud Computing, EC2, Simulation, SME, Spot Instances, Spot Price

INTRODUCTION

Cloud computing is the utility-way of sharing computing resources – similar to the way we consume our utility services (say, electricity, water, etc.) from the corresponding service providers (Gartner report, 2015). However, a major characteristic of cloud computing is that it is Internet-based utility computing, whereby shared resources (like infrastructure, platform, and services) are provided to end-users’ devices on demand via the shared Internet, whereas the
other utilities are delivered to our home via separate dedicated distribution channels. Among the multitude of benefits that cloud computing offers, arguably the most promising one is to relieve the organizations of planning, purchasing, and maintaining in-house computing facilities, and converts their large fixed costs (CapEx) into much smaller variable costs (OpEx). Nearly 24% of the companies now have adopted “cloud-first” policy.[1]

There are mainly two types of clouds: private and public. In case of public cloud, users (such as individuals, institutions, companies and business houses) purchase just-as-much-required resources, on pay-as-you-go basis, from remote servers hosted by cloud service providers (CSPs), such as Amazon Web Services (AWS), through some form of contract or service level agreement (SLA) (Andrzejak, Kondo, & Songho, 2010). The CSPs build, own, operate and manage these servers at their premises; thus, companies can store data in Google’s cloud (https://cloud.google.com/) or in AWS (http://aws.amazon.com) using its simple storage services (S3). Next, they can use Amazon’s Elastic Compute Cloud (EC2) (http://aws.amazon.com/ec2/) to perform necessary computations on those data and then again store the results in any of the cloud storages. On the other hand, in case of private cloud, organizations create cloud-like environment inside their premises so that various business units (BUs) use that private cloud infrastructure as common shared services over the organization’s intranet/extranet. Usually, the IT department, with the help from 3rd parties, maintains the internal private cloud for the organization.

As reported in the literature (Foster, Zhao, Raicu, & Lu, 2008), there are several advantages of transferring resources from thin client devices to cloud, which may be as thick as possible. Thus, the key attractions of cloud computing (either commercial grade public clouds or enterprise-owned private clouds) are twofold: (i) it is shared, dynamically-configured and market-driven, and (ii) everything in it is provided as a service (XaaS) (Foster, Zhao, Raicu, & Lu, 2008). Naturally, Small and Medium Enterprises (SMEs) are turning to cloud computing in order to eliminate their non-core IT-related activities, thereby allowing them to focus more on their core competence and high impact making strategic activities. Software as a Service (SaaS) is the most common usage for cloud, and CSPs usually charge SaaS on a fixed price basis per month or per year. Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) follow, however, variable pricing models. Cloud users in this case have to deal with several different structures. At the core they have to rent a virtual machine (VM) first. Then they need storage. To calculate storage, the CSPs charge a per-GB price. Since PaaS and IaaS Solutions consume network and Internet traffic too, the users also get charged on incoming and outgoing bandwidth. Here also, the chargeability is based on a per-GB fee.

The CSPs offer on rent their VMs called instances. Each Instance runs for a certain number of hours, called Compute Hours. Often CSPs lease out the instances at a fixed price, such as Amazon EC2 Small Instance (Shao, Yuan, & Li, 2012), and Windows Azure Small Compute Instance (Andrzejak, Kondo, & Anderson, 2010). They also have a dynamically configured price-conscious model with a market-driven mechanism in-built. For example, Amazon EC2 provides three primary models to purchase their instances: (i) On-Demand, (ii) Reserved, and (iii) Spot (Shao, Yuan, & Li, 2012). The models follow different pricing mechanisms and accordingly guarantee three levels of assurance regarding when instances can be launched and terminated. Spot instances, unlike reserved or on-demand instances, are spare slots (or, unused capacity) of VMs for which the CSP has no pre-booking or reservation. Spot instances are typically ideal for flexible tasks, which do not have any stringent constraint on deadline, and also for those unscheduled jobs which pop up suddenly in enterprises. The price for a spot instance changes dynamically based on supply-demand logic and/or some other complex logic followed by CSPs (Macias, & Gitart, 2011; Tsai, & Tsai, 2012; Ben-Yehuda, Ben-Yehuda, Schuster, & Tsafrir, 2011). Being relatively cheap, spot instances are an attractive option for small and medium enterprises
Implement VoIP Based IP Telephony with Open Source Asterisk Architecture
www.igi-global.com/article/implement-voip-based-telephony-open/40959?camid=4v1a

www.igi-global.com/article/international-conference-localization-global-navigation/68999?camid=4v1a