Journal of Organizational and End User Computing
Volume 33 • Issue 2 • March-April 2021

Placement for Intercommunicating Virtual Machines in Autoscaling Cloud Infrastructure:

Autoscaling and Intercommunication Aware Task Placement

Sridharan R., National Institute of Technology, Tiruchirappalli, India

Domnic S., National Institute of Technology, Tiruchirappalli, India

ABSTRACT

Due to pay-as-you-go style adopted by cloud datacenters (DC), modern day applications having intercommunicating tasks depend on DC for their computing power. Due to unpredictability of rate at which data arrives for immediate processing, application performance depends on autoscaling service of DC. Normal VM placement schemes place these tasks arbitrarily onto different physical machines (PM) leading to unwanted network traffic resulting in poor application performance and increases the DC operating cost. This paper formulates autoscaling and intercommunication aware task placements (AIATP) as an optimization problem, with additional constraints and proposes solution, which uses the placement knowledge of prior tasks of individual applications. When compared with well-known algorithms, CloudsimPlus-based simulation demonstrates that AIATP reduces the resource fragmentation (30%) and increases the resource utilization (18%) leading to minimal number of active PMs. AIATP places 90% tasks of an application together and thus reduces the number of VM migration (39%) while balancing the PMs.

KEYWORDS

Autoscaling, Cloud Computing, Communication Latency, Elasticity, Intercommunication, Network Latency, Tasks Migration, VM Placement

INTRODUCTION

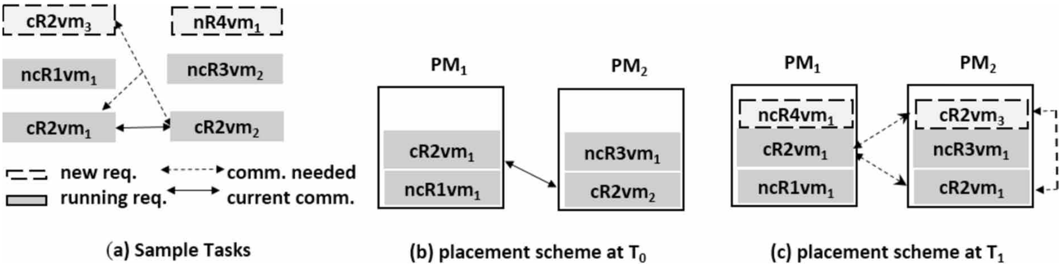
The rate at which data is generated by modern day applications is growing exponentially and is unpredictable with respect to time. Amount of information contained in the data are also disruptive for predictive data processing and analytics functionality. This data needs to be analyzed timely for building business intelligence leading to better decisions. Hence, these applications depend heavily on cloud computing paradigm. Using pay-per-use mode, cloud computing provides *autoscaling* services to manage the sporadic resource requirement of modern day applications. As industries desire reduced time to market for their applications, they adopt to cloud. Cloud operates via DC (datacenters), has huge computing and storage resources, resulting in heavy electric power consumption. This leads to higher operational cost for the CSPs (cloud service provider) and eventually the same will be passed to the customers. Hence, the current focus of researchers is to identify efficient DC management

DOI: 10.4018/JOEUC.20210301.oa2

leading to reduced power consumption while delivering high quality service to the customers. This presents several research challenges including resource management (Manvi, & Shyam, 2014). These challenges need to be addressed after considering multiple parameters related to application types, cloud services and networks of DC.

Application types could be multi-tier web applications, big data processing applications, machine learning and multimedia applications, and high-performance applications. Applications such as video surveillance, real time gaming, real time streaming and augmented reality have intercommunication tasks[1] which are too sensitive towards latency and also need more computational power. For example, VMs hosting database tasks and hosting application tasks need intercommunication in three-tier web applications. Similarly, VMs that hosts object tracker task in a video surveillance system that uses multiple cameras covering multiple areas, need inter task communication. As these applications are sporadic resource requester by nature, they depend heavily on elastic capability. Typically, these applications having group of jobs (could be long-duration jobs needing autoscaling), execute their intercommunicating or non-intercommunicating tasks using VMs (Virtual Machines instance) in a DC. Autoscaling service operates on an adjustable capacity (minimum and maximum number of VM), (Boucher Jr, et al, 2018; Barclay, 2016) within which the tasks need to be executed. Without violating this capacity range, addition or deletion of VM instances shall be carried out, upon meeting the resource threshold value specified by the user.

Figure 1. Sample VM and its placement using FCFS



(a) Sample Tasks (b) placement scheme at T_0 (c) placement scheme at T_1

Virtual Machine Placement (VMP) (Lopez-Pires & Baran, 2015) is the process of selecting suitable Physical Machine (PM) to place the requested VM. VMP taking into account the auto-scalability of the applications along with intercommunication of theirs tasks is a challenging cloud problem.

Consider an example of three application requests that are currently (at time T_0) executing in a DC. Let $ncR1vm_1$ and $ncR3vm_1$ be the VMs of first and third application respectively, which are non-intercommunication type. Let the second application have two inter-communicating VMs ($cR2vm_1$ and $cR2vm_2$) that needs autoscaling. Assume that at time T_1, fourth and fifth requests arrive to the DC. While fourth is a fresh non-intercommunicating type request ($ncR4vm_1$), the fifth request ($cR2vm_3$) is an additional VM of the second request that needs inter communication with $cR2vm_1$ and $cR2vm_2$. Further, let these tasks follow the communication pattern represented by arrows, as shown in Figure 1(a). Assuming that we have two PMs to place these tasks, using First Come First Serve (FCFS) VMP algorithm along with load balancing, placements of VMs at T_0 resembles to that of Figure 1(b). Placement of tasks at T_1 is shown Figure 1(c). If all the tasks are placed onto single PM itself, the other PM could be switched off to reduce the overall power consumption. Else placing all the intercommunicating tasks ($cR2vm_1$, $cR2vm_2$ and $cR2vm_3$) on to the same PM or on to a network