MIDDLEWARE FOR AUTONOMOUS RECONFIGURATION OF VIRTUAL MACHINES

By

Qingling Wang

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

Approved:

_____

Carlos Varela, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

July 2011
(For Graduation August 2011)

ABSTRACT

Cloud computing brings significant benefits for service providers and service users because of its characteristics: *e.g.*, on demand, pay for use, scalable computing. Virtualization management is a critical component to accomplish effective sharing of physical resources and scalability. Existing research focuses on live Virtual Machine (VM) migration as a VM consolidation strategy. However, the impact of other virtual network configuration strategies, such as optimizing total number of VMs for a given workload, the number of virtual CPUs (vCPUs) per VM, and the memory size of each VM has been less studied. This thesis presents specific performance patterns on different workloads for various virtual network configuration strategies. We conclude that, for loosely coupled CPU-intensive workloads, memory size and number of vCPUs per VM do not have significant performance effects. On an 8-CPU machine, with memory size varying from 512MB to 4096MB and vCPUs ranging from 1 to 16 per VM; 1, 2, 4, 8 and 16VM configurations have similar running time. The prerequisite of this conclusion is that all 8 physical processors be occupied by vCPUs. For tightly coupled CPU-intensive workloads, the total number of VMs, vCPUs per VM and memory allocated per VM become critical for performance. We obtained the best performance when the ratio of total number of vCPUs to processors is 2. Doubling memory size on each VM, for example from 1024MB to 2048MB, brings at most 15% improvement of performance when number of VMs is greater than 2. Based on the experimental results, we propose a framework and a threshold-based strategy set to dynamically refine virtualization configurations. The framework mainly contains three parts: *resources monitor*, *virtual network configuration controller* and *scheduler*, which are responsible for monitoring resource usage on both virtual and physical layers, controlling virtual resources distribution, and scheduling concrete reconfiguration steps respectively. Our reconfiguration approach consists of four strategies: VM migration and VM malleability strategies, which are at global level, vCPU tuning and memory ballooning, which are at local level. The strategies evaluate and trigger specific reconfiguration steps (for example, double the

number of vCPUs on each VM) by comparing current allocated resources and corresponding utilizations with expected values. The evaluation experimental results of threshold-based strategy show that reconfiguration in global level works better for tightly coupled CPU-intensive workloads than for loosely coupled ones. Local reconfiguration including dynamically changing number of vCPUs and memory size allocated to VMs, improves the performance of initially sub-optimal virtual network configurations, even though it falls short of performing as well as the initially optimal virtual network configurations. This research will help private cloud administrators decide how to configure virtual resources for a given workload to optimize performance. It will also help service providers know where to place VMs and when to consolidate workloads to be able to turn on/off Physical Machines (PMs), thereby saving energy and associated costs. Finally it let service users know what kind of and how many VM instances to allocate in a public cloud for a given workload and budget.