CloudLEGO: Scalable Cross-VM-Type Application Performance Prediction

Shicong Meng, Arun K. Iyengar, Ling Liu*, Ting Wang, Jian Tan, Ignacio Silva-Lepe
Isabelle M. Rouvellou

IBM T.J. Watson Research Center
{smeng,aruni,twang,tanji,isilval,rouvellou}@us.ibm.com

*Georgia Institute of Technology
lingliu@cc.gatech.edu

Abstract

Understanding the performance difference of a multi-tier Cloud application between different provisioning plans and workloads is difficult to achieve. A typical IaaS provider offers a variety of virtual server instances with different performance capacities and rental rates. Such instances are often marked with a high level description of their hardware/software configuration (e.g. 1 or 2 vCPUs) which provides insufficient information on the performance of the virtual server instances. Furthermore, as each tier of an application can be independently provisioned with different types and numbers of VMs, the number of possible provisioning plans grows exponentially with each additional tier.

Previous work [10] proposed to perform automatic experiments to evaluate candidate provisioning plans, which leads to high cost due to the exponential increase of candidate provisioning plans with the number of tiers and available VM types. While several existing works [8, 6, 7] studied a variety of performance models for multi-tier applications, these works assume that an application runs on a fixed deployment (with fixed machine type and number for each tier).

We present CloudLEGO, an efficient cross-VM-type performance learning and prediction approach. Since building a model for each possible deployment is clearly not scalable, instead of treating each candidate deployment separately, CloudLEGO views them as derivatives from a single, fixed deployment. Accordingly, the task of learning the performance of a targeted deployment can be decoupled into learning the performance of the original fixed deployment and learning the performance *difference* between the original deployment and the targeted one.

The key to efficiently capture performance difference between deployments is to find multiple independent changes that can be used to derive any deployment from the original deployment. CloudLEGO formulates such "modular" changes as VM type changes at a given tier. To capture changes of performance at a tier caused by VM type changes, CloudLEGO uses relative performance models [5] which predict the performance *difference* between a pair of VMs (rather than the absolute performance of a VM) for a given workload. Moreover, training relative performance models requires only performance data from Cloud monitoring services [1, 4] rather than fine-grain data such as per-tier response time which requires application instrumentation.

Training relative performance models with traditional passive learning techniques would require a large amount of training data as performance data are collected uniformly in a single batch. We find that different types of VMs often share similar performance for many "regions" of workloads. To leverage this characteristic and guide the profiling to regions with high performance differences, CloudLEGO uses active learning techniques [2, 3, 9] that split the profiling process into multiple stages where data collected in one stage are used to identify high-value regions for the next profiling stage. As a result, it significantly speeds up the convergence of models and the profiling process due to substantially reduced measurement.

We deploy CloudLEGO in IBM's Research Computing Cloud (RC2), an Infrastructure-as-a-Service Cloud, to evaluate its effectiveness. Our results suggest that CloudLEGO provides accurate predictions for various deployments and workloads with only a fraction of training cost incurred by existing techniques.

SoCC'13, 1–3 Oct. 2013, Santa Clara, California, USA.
ACM 978-1-4503-2428-1. http://dx.doi.org/10.1145/2523616.2525948