Improving PaaS Multitenancy

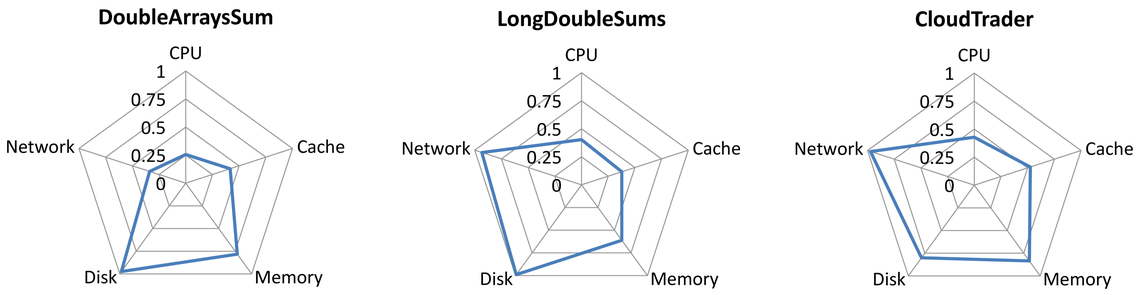
Panagiotis (Panos) Patros, Dayal Dilli,
Stephen A. MacKay, Kenneth B. Kent, Michael Dawson
University of New Brunswick, IBM Canada
Faculty of Computer Science
{patros.panos, stephen.mackay, dayal.dilli, ken}@unb.ca
Michael_Dawson@ca.ibm.com

Background: PaaS Clouds

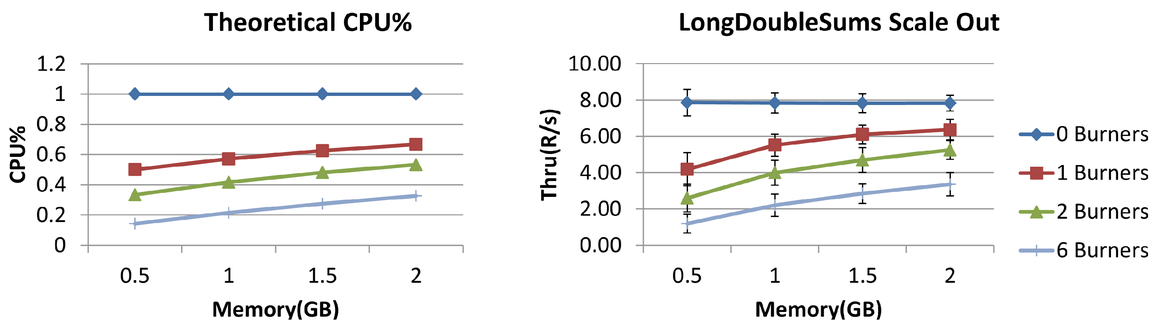
Platform as a Service (PaaS) clouds provide abstraction of the software and hardware stack. Users just upload their code; the cloud then makes the application available online automatically. CloudFoundry is open source PaaS software that IBM Bluemix uses; we run our experiments on it.

Resource Intensiveness of Applications

PaaS clouds can place tenants on the same VM, which makes them able to affect the performance of each other—constrained by certain Service Level Agreements (SLAs). We quantify the slowdown of applications on the presence of specific resource-intensive tenants we call Cloud Burners, which allows resource-intensiveness profiling:



Predicting Throughput after Scaling

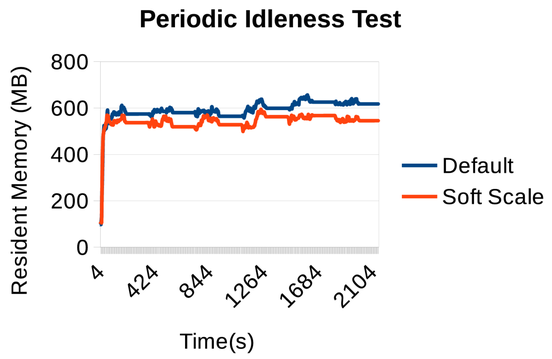


Scaling out adds instances; scaling up increases the resources of the current instances. Both types of scaling increase the CPU shares a tenant gets; however, this is does not necessarily lead to linear performance improvements We propose and evaluate a theoretical model that predicts CPU utilization of tenants depending on the activity of their neighbors.

Fair and Anti-Interference Multitenancy

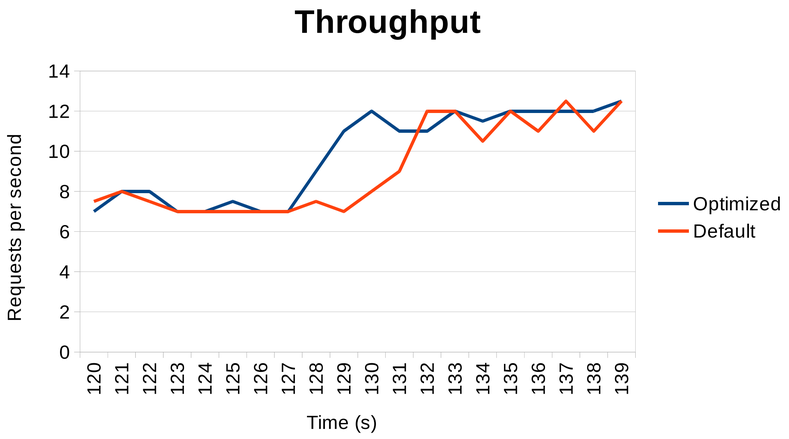
Since PaaS tenants interfere with each other, we are investigating a placement heuristic based on our resource slowdown and interference profiling. Initial experimental results compared all 10 possible placements of 6 tenants on 2 VMs; our heuristically derived plan was the second most performant and the second most fair.

Resident Memory Reduction



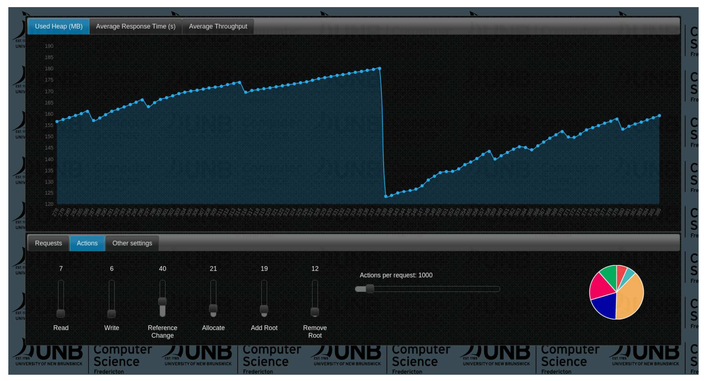
PaaS tenants can utilize idleness periods to perform clean up tasks and also, reduce their resource consumption, which enables other tenants running on the same VM to utilize them and improve their performance.

Faster Scaling and Reduced Downtime



When a PaaS application starts for the first time, it creates some dynamically compiled data. By sharing this data among instances, we reduce the startup and warmup time during scaling or restarting.

The Cloud CG App



We implemented a PaaS cloud application that stresses the Garbage Collector, which is a major bottleneck for any managed runtime, by performing graph actions with customizable ratios. It can be used to simulate multiple types of mutator patterns. Also, in conjunction with its GUI, it can be utilized as an academic tool to demonstrate the effects of operations on the heap in real-time.