Skynet: Performance-driven Resource Management for Dynamic Workloads

Yannis Sfakianakis[1,2#*], **Manolis Marazakis**[2], and **Angelos Bilas**[1,2]

[1] Computer Science Department, University of Crete, Greece

[2] Institute of Computer Science, Foundation for Research and Technology - Greece

Presenting author: Yannis Sfakianakis, email: jsfakian@ics.forth

* Corresponding author: Yannis Sfakianakis, email: jsfakian@ics.forth.gr

ABSTRACT

A primary concern for cloud operators is to increase resource utilization while maintaining good performance for applications. This is particularly difficult to achieve for three reasons: (1) users tend to overprovision applications, (2) applications are diverse and dynamic, and (3) their performance depends on multiple resources.

In this paper, we present *Skynet*, an automated and adaptive cloud resource management approach that ad- dresses all three concerns. *Skynet* uses performance level objectives (PLOs) to capture user intentions about required performance more accurately to remove the user from the resource allocation loop. Then, *Skynet* estimates the resources required to achieve the target PLO. For this purpose, we employ a Proportional Integral Derivative (PID) controller per application and adjust its parameters on the fly. Finally, to capture the dependence of applications on different or multiple resources, *Skynet* extends the traditional one-dimensional PID controller to estimate CPU, memory, I/O throughput, and network throughput.

Essentially, *Skynet* builds a model on-the-fly to map target PLOs to resources for each application, considering multiple resources and changing input load. We implement *Skynet* as an end-to-end, custom scheduler in Kubernetes and evaluate it using real workloads on both a private cluster and AWS. *Skynet* decreases PLO violations by more than 7.4x and increases resource utilization by more than 2x, compared to Kubernetes.