Edge and Cloud Provider Cost Minimization by Exploiting Extended Voltage and Frequency Margins

Christos KALOGIROU [a], Panos KOUTSOVASILIS [a], Manolis MAROUDAS [a], Christos D. ANTONOPOULOS [a], Spyros LALIS [a] and Nikolaos BELLAS [a]

[a] *Department of Electrical and Computer Engineering, University of Thessaly*
e-mail:{hrkalogi, pkoutsovasilis, emmmarou, cda, lalis, nbellas}@uth.gr

Abstract. Energy costs contribute significantly to the total cost of operation for cloud and edge infrastructure providers. Both conventional (Voltage and Frequency Scaling) and more aggressive (undervolting, overclocking) techniques can be applied to reduce the energy footprint of the infrastructure and thus the cost of the operator. However, these techniques may affect the quality of service (QoS), potentially activating service level agreement (SLA) violation penalties for the provider. In this paper we model and study this tradeoff. We find that undervolting is the most effective of the three techniques in reducing infrastructure operation cost. We identify optimal operating points, and we study the effect of different parameters, such as the severity of SLA penalties, the length of the job and the existence of error protection mechanisms, on the optimal operating point and the extent of potential benefits.

Keywords. Cloud/Edge computing, Energy efficiency, SLAs, Operation Cost, Extended margins, Reliability

1. Introduction

Minimizing energy consumption has become a critical concern in computing, due to technical, economic and environmental reasons. For cloud/datacenter deployments, energy corresponds to a significant percentage of the operation cost and the environmental footprint of the facility. At the same time, power dissipation is the main factor limiting the denser packing of servers. In edge deployments, technical constraints often introduce even tighter power and energy limitations.

Hardware manufacturers have introduced Voltage and Frequency Scaling (VFS) [10] and more recently power throttling as effective mechanisms of power management. VFS enables the concurrent manipulation of supply voltage (V_{dd}) and operating frequency (f) among different nominal (V_{dd}, f) tuples. VFS can result to significant power savings, which however come at the cost of lowering the frequency (and thus typically performance) of the CPU. Due to the associated performance loss in computational workloads, VFS does not necessarily reduce the energy footprint of the workload.

Another, more aggressive, power reduction technique is voltage overscaling (undervolting) [2]. In this case, the CPU is supplied with lower voltage than the nominal for