Recent Advances in Modern Datacenter Provisioning and Reliability

Ioannis Manousakis
Dept. of Computer Science

7/14/2017 at 12:00 pm
CoRE B (305)

Abstract

Cloud providers have made significant strides in reducing the cooling capital and operational costs of their datacenters, for example, by leveraging outside air (free) cooling where possible. Despite these advances, cooling costs still represent a significant expense mainly because cloud providers typically provision their cooling infrastructure for the worst- case scenario. The reliability implications of these choices still remains unclear. We first propose to reduce cooling costs by underprovisioning the cooling infrastructure. When the cooling is underprovisioned, there might be (rare) periods when the cooling infrastructure cannot cool down the IT equipment enough. During these periods, we can either (1) reduce the processing capacity and potentially degrade the quality of service, or (2) let the IT equipment temperature increase in exchange for a controlled degradation in reliability. We then explore the reliability implications of the environmental conditions induced by the cooling provisioning and operation. We find that aggressive undeprovisioning and energy-conserving operation strategies can greatly degrade the reliability of hard disks. In particular we collect operational data from from nine Microsoft cloud-scale datacenters, including environmental conditions and failure reports. We find that humidity is the main concern in these datacenters and create a joint lifetime failure model to predict their failure rates.

Finally, we turn our attention to the provisioning strategies of datacenter fleets that enable future internet services with very low latency requirements. We tackle the problem by formulating, implementing and evaluating an optimization framework which provisions small, less reliable datacenters under latency and population coverage constraints. Our results show that our framework provisions datacenter fleets that are capable of supporting services with end-to-end latencies of 50ms.

Defense Committee: Prof. Thu D. Nguyen (Chair), Prof. Ricardo Bianchini, Prof. Abhishek Bhattacharjee, Prof. Anand Sivasubramanian (Penn State University)