HOW QUEUEING-THEORETIC MODELS FOR DATA CENTER POWER MANAGEMENT DIFFER FROM REALITY

Anshul Gandhi, Carnegie Mellon University, Pittsburgh, anshulg@cs.cmu.edu
Mor Harchol-Balter, Carnegie Mellon University, Pittsburgh, harchol@cs.cmu.edu
Michael A. Kozuch, Intel Labs, Pittsburgh, michael.a.kozuch@intel.com

Power management in data centers is an important research topic that has received considerable attention from queueing theorists recently. While queueing theory provides invaluable input into the design and analysis of power management policies for data centers, there are often many simplifying assumptions in the underlying queueing model that discourage data center operators from implementing these policies in practice.

In this work, we focus on three popular assumptions: (i) the time it takes to power up a server is negligible, (ii) the mean service rate of a server does not depend on the number of jobs, and (iii) the mean arrival rate for a given workload is constant. We first demonstrate, via actual implementation results, that the above three assumptions often do not hold in reality. We then propose simple modifications to the queueing model that allow us to relax the above three assumptions to more accurately reflect the reality.