STABILIZING QUEUES WITH NON-HOMOGENEOUS AND MULTI-CLASS WORKLOADS IN DATA CENTERS

N. Gautam, Texas A&M University, USA, gautam@tamu.edu
S. Kwon, Texas A&M University, USA, soongeol@tamu.edu

We consider a system with multiple queues and heterogeneous servers where requests from multiple classes arrive in a time-varying fashion. For such a system, our goal is to manage resources so that the resulting performance measures are time-homogeneous. The key benefit of stabilizing the queue length process or sojourn time distribution is that it enables effective analysis by leveraging upon the vast literature for time-homogeneous systems. In addition, systems with smoothed traffic are much more conducive for developing control algorithms. We will first present several scenarios where we tune highly non-homogeneous systems to result in time-stable performance. Then we will illustrate our methodology in the context of energy and performance management in data centers and other distributed computing environments. Our objective is to develop strategies to: (i) assign classes to servers, (ii) determine the number of servers to be powered on, (iii) route requests to appropriate servers, and (iv) create a procedure for speed scaling. The requirement is to develop the aforementioned strategies under: (a) a distributed setting where real-time information is not exchanged between the sub-systems; (b) a necessity for time-stable performance; (c) a preference for simplified operations while maintaining cost-effectiveness and high quality of service. We illustrate our strategies and methodology using numerical examples.