TASK ASSIGNMENT IN A SERVER FARM WITH SWITCHING DELAYS AND GENERAL ENERGY-AWARE COST STRUCTURE

E. Hyytiä, Aalto University, Finland, esa.hyytia@aalto.fi
R. Righter, University of California Berkeley, USA, rrighter@ieor.berkeley.edu
S. Aalto, Aalto University, Finland, samuli.aalto@aalto.fi

We consider the task assignment problem to parallel servers with switching delay, where servers can be switched off to save energy. However, switching a server back on involves a constant server-specific setup delay. We will use one step of policy iteration from a starting policy such as Bernoulli splitting, in order to derive efficient task assignment (dispatching) policies that minimize the long-run average cost. To evaluate our starting policy, we first analyze a single work-conserving M/G/1 queue with a switching delay and derive a value function with respect to a general cost structure. Our costs include energy related switching and running costs, as well as performance-related costs associated with both means and variability of waiting time and latency. The efficiency of the resulting dispatching policies is illustrated with numerical examples.