



DYNAMIC LOAD BALANCING IN PARALLEL QUEUEING SYSTEMS: STABILITY AND OPTIMAL CONTROL

Mark E. Lewis

University of Michigan

Department of Industrial and Operations Engineering

melewis@engin.umich.edu

We consider a system of parallel queues with dedicated arrival streams. At each arrival time or service completion a decision-maker can move customers from one queue to another. The cost for moving customers consists of a fixed cost and a linear, variable cost dependent on the number of customers moved. There are also linear holding costs that may depend on which queue customers are stored. We seek a policy that minimizes the long-run average expected cost. In this talk, we motivate the problem above by describing similar scenarios in parallel processing in computer networks and work re-allocation in supply chain management.

We develop stability (and instability) conditions for the most general system via a fluid model. In typical operations researcher fashion, we then divide the problem into several smaller problems and consider each separately. The one-server case yields optimal control limit policies. In the case of two-servers the optimal control policy is shown to prefer to store customers in the lowest cost queue. Furthermore, under an exponential assumption, the optimal policy is shown to be an “order up to” policy. These observations are used to suggest several heuristics for the general n -server problem.