

October 2007

Efficient routing in heavy traffic under partial sampling of service times

Rami Atar and Adam Shwartz
Department of Electrical Engineering
Technion–Israel Institute of Technology
{atar}{adam}@ee.technion.ac.il
<http://www.ee.technion.ac.il/people/{adam}{atar}>

October 21, 2007

Abstract

We consider a queue with renewal arrivals and n exponential servers in the Halfin-Whitt heavy traffic regime, where n and the arrival rate increase without bound, so that a critical loading condition holds. Server k serves at rate μ_k , and the empirical distribution of $\{\mu_k\}_{k=1,\dots,n}$ is assumed to converge weakly. We show that very little information on the service rates is required for a routing mechanism to perform well. More precisely, we construct a routing mechanism that has access to a *single sample* from the service time distribution of each of $n^{\frac{1}{2}+\varepsilon}$ randomly selected servers ($\varepsilon > 0$), but not to the actual values of the service rates, the performance of which is asymptotically as good as the best among mechanisms that have the complete information $\{\mu_k\}_{k=1,\dots,n}$.

Keywords: Halfin-Whitt regime; routing policies; service time sampling

MSC2000: Primary: 60F17. Secondary: 68M20, 90B15, 90B22, 60K30, 60K25.