CCIT Report No. 391

Scheduling a Multi-Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy Traffic

Rami Atar, Avi Mandelbaum, Martin I. Reiman

ABSTRACT

We consider the problem of scheduling a queueing system in which many statistically identical servers cater to several classes of impatient customers. Service times and impatience clocks are exponential while arrival processes are renewal. Our cost is an expected cumulative discounted function, linear or non-linear, of appropriately normalized performance measures. As a special case, the cost per unit time can be a function of the number of customers waiting to be served in each class; the number actually being served, the abandonment rate, the delay experienced by customers, the number of idling servers, as well as certain combinations thereof. We study the system in an asymptotic heavy-traffic regime where the number of servers n and the offered load R are simultaneously scaled up and carefully balanced: $n \approx R + \beta\sqrt{R}$, for some scalar β. This yields an operation that enjoys the benefits of both heavy traffic (high server utilization) and light traffic (high service levels.)

We first consider a formal weak limit, through which our queueing scheduling problem gives rise to a diffusion control problem. We show that the latter has an optimal Markov control policy, and that the corresponding Hamilton-Jacobi-Bellman (HJB) equation has a unique classical solution. The Markov control policy and the HJB equation are then used to define scheduling control policies which we prove are asymptotically optimal for our original queueing system. The analysis yields both qualitative and quantitative insights, in particular on staffing levels, the roles of non-preemption and work-conservation, and the tradeoff between service quality and servers' efficiency.