Utility Optimization in Congested Queueing Networks

[Invited Presentation, Extended Abstract]

Neil Walton
Statistical Laboratory
University of Cambridge
n.s.walton@statslab.cam.ac.uk

ABSTRACT

We consider a multi-class single server queueing network as a model of a packet switching network. We discuss how such networks perform a proportionally fair optimization when congested. We discuss the connections between product form queueing networks, insensitivity and proportional fairness. We prove that stationary throughput of a closed multi-class single server queueing network converges to a proportionally fair allocation as the number of packets across routes increases. We then let the rate packets enter different routes of the network be controlled by congestion windows, which record the number of sent but not yet acknowledged packets on each route of the network. By considering a sequence of such congestion windows we allow the network to become congested. We show that these networks maximize aggregate utility subject to the networks capacity constraints. To perform this analysis we require our utility functions to satisfy an exponential concavity assumption. This family of utility functions includes the weighted $\alpha$-fair family of utilities for parameter $\alpha > 1$.

1. DISCUSSION

As a queueing network becomes congested the demand for its resources increases and the network becomes more competitive. The share of the resources received is often expressed in terms of utility optimization subject to the networks capacity constraints.

Initial work on effective bandwidths [3, 6] considered the demand different traffic sources $i \in I$ imposed on a queue $j \in J$ of capacity $C_j$. A large deviations analysis showed the bandwidth required by traffic sources imposed a constraint

$$\sum_{i \in I} \Lambda_i \leq C_j,$$

where $\Lambda_i$ is expressible in terms of the log moment generating function of the traffic load of source $i \in I$.

As a method of allocating resources and introducing fairness, subsequent work of Kelly [7] considered utility optimization of the form

$$\text{maximize } \sum_{i \in I} U_i(\Lambda_i)$$

subject to

$$\sum_{i : j \in I} \Lambda_i \leq C_j, \quad j \in J,$$

over

$$\Lambda_i \geq 0, \quad i \in I.$$

where $U_i$ is an increasing convex utility function. We call this optimization problem the system problem. In addition [7] introduced proportional fairness as the unique solution to the optimization problem

$$\text{maximize } \sum_{i \in I} \bar{m}_i \log \Lambda_i$$

subject to

$$\sum_{i : j \in I} \Lambda_i \leq C_j, \quad j \in J,$$

over

$$\Lambda_i \geq 0, \quad i \in I.$$

We call this optimization problem the network problem or the proportionally fair optimization problem. The paper [7] considered the combined solution of the network problem and the following user problems, for each $i \in I$

$$\text{maximize } U_i(\bar{m}_i / q_i) - \bar{m}_i$$

over

$$\bar{m}_i \geq 0.$$

This combined solution was considered under the relation

$$\bar{m}_i = \Lambda_i q_i, \quad i \in I,$$

where $q_i = \sum_{j \in J} q_j$ and $(q_j : j \in J)$ are the Lagrange multipliers associated with the network problem. Theorem 2 of Kelly [7] found under (5) that the combined solution of the network and user problem gave the solution to the system problem.

This result was constructed to suggest an end-to-end argument for providing optimization and fairness across a communication network. The result provided a method for decomposing the system problem into a user problem that is independent of the network structure except through parameter $q_i$ and a network problem that is independent of users preferences except through parameter $\bar{m}$. Interpreted in the context of a communication network this separated the preferences of users performing end-to-end communication and the network’s preferred optimal behaviour. In [7] the solution is interpreted as setting prices $(q_j : j \in J)$ for sending traffic through the network. With these prices each user, $i \in I$, chooses an amount of money $\bar{m}_i$ it is willing to...
pay per unit of time. From this the user receives an amount of bandwidth \( \lambda_i = \frac{m_i}{d_i} \).

By construction this result considers a static model and the end-to-end argument performed by users is implicit. Subsequent work has successfully used differential equations to add dynamics to this notion of optimization and decomposition \([10, 4, 8, 16]\), other work has considered the form of utility optimization achieved by different protocols \([17, 14, 11]\) and authors have also considered stochastic models of flow across a network \([12, 1, 2]\).

In 1979, Schweitzer \([15]\) studied approximations of closed multi-class queueing networks and considered how asymptotic conditions on such networks might satisfy the Kuhn-Tucker conditions for proportionally fair optimization. In 1989, Kelly \([5]\) studied approximations of closed queueing networks and by an analogous analysis considered a similar optimisation formulation. In 1999, Massoulié and Roberts \([13]\) studied a fluid type queueing model and used these same Kuhn-Tucker conditions to deduce proportional fairness. Using large deviations and heavy traffic analysis, recent work of Walton \([18]\) and Kelly, Massoulié and Walton \([9]\) have provided rigorous formalisations of the relationship between closed queueing networks and proportional fairness.

The connection between multi-class queueing networks and proportional fairness gives a much more literal meaning to the network problem. With this in mind, we construct a queueing system consisting of a multi-class queueing network and congestion windows. This queueing system asymptotically executes Theorem 2 of Kelly \([7]\). This analysis leads us to think of the flow through the network in a similar way to the flows found for effect bandwidths and instead of interpreting \( q \) and \( m \) as prices and wealth we interpret them as round-trip times and congestion window sizes.

In this queueing system congestion windows record the number of sent but not yet acknowledged packets on each route of a multi-class queueing network and sends packets into this network at a rate which is a function of this number. We allow a sequence of congestion windows to congest the multi-class queueing network and we study the large deviations behaviour of the stationary distribution of the queueing system. Noting the user problem (4) is reminiscent of a Legendre-Fenchel transform we allow a sequence of congestion windows to solve a modified user problem. As discussed above when congested the multi-class queueing network will solve the network problem. The relation (5) will be satisfied as it describes Little’s Law for the number of packets on transfer on each route. Thus the queueing system will asymptotically satisfy the network problem, the user problem and relation (5). So given Theorem 2 of Kelly we expect our queueing system to optimize the system problem.

We find in our analysis that we require each utility function to satisfy an \textit{exponential concavity condition}, that the map \( \lambda \mapsto U_i(e^\lambda) \) is concave.

2. REFERENCES


