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 α-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 $i \in \mathcal{I}$ imposed on a queue $j \in \mathcal{J}$ of capacity C_j. A large deviations analysis showed the bandwidth required by traffic sources imposed a constraint

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

where Λ_i is expressible in terms of the log moment generating function of the traffic load of source $i \in \mathcal{I}$.

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

$$\begin{aligned} &\text{maximize} && \sum_{i \in \mathcal{I}} U_i(\Lambda_i) && && (1)\\ &\text{subject to} && \sum_{i:j \in i} \Lambda_i \leq C_j, && j \in \mathcal{J}, && (2)\\ &\text{over} && \Lambda_i \geq 0, && i \in \mathcal{I}, && (3)\end{aligned}$$

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

$$\begin{aligned} &\text{maximize} && \sum_{i \in \mathcal{I}} \bar{m}_i \log \Lambda_i && \\ &\text{subject to} && \sum_{i:j \in i} \Lambda_i \leq C_j, && j \in \mathcal{J},\\ &\text{over} && \Lambda_i \geq 0, && i \in \mathcal{I}.\end{aligned}$$

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 $i \in \mathcal{I}$

$$\begin{aligned} &\text{maximize} && U_i\big(\frac{\bar{m}_i}{q_i}\big) - \bar{m}_i && \\ &\text{over} && \bar{m}_i \geq 0. && (4)\end{aligned}$$

This combined solution was considered under the relation

$$\bar{m}_i = \Lambda_i q_i, \qquad i \in \mathcal{I}, \tag{5}$$

where $q_i = \sum_{j \in i} q_j$ and $(q_j : j \in \mathcal{J})$ 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 $(q_j : j \in \mathcal{J})$ for sending traffic through the network. With these prices each user, $i \in \mathcal{I}$, chooses an amount of money $\bar{m}_i$ it is willing to

VALUETOOLS 2009, October 20-22, Pisa, Italy
Copyright © 2009 ICST 978-963-9799-70-7
DOI 10.4108/ICST.VALUETOOLS2009.7826

pay per unit of time. From this the user receives an amount of bandwidth $\Lambda_i = \frac{\bar{m}_i}{q_i}$.

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 $\bar{m}$ 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 *exponential concavity condition*, that the map $\lambda \mapsto U_i(e^\lambda)$ is concave.

2. REFERENCES

[1] Bonald, T. and Massoulie, L. (2001). Impact of fairness on internet performance. *Proc. of ACM Sigmetrics 29*, 82–91.

[2] Bonald, T. and Proutière, A. (2004). On performance bounds for balanced fairness. *Performance Evaluation 55*, 25–50.

[3] Gibbens, R. J. and Hunt, P. J. (1991). Effective bandwidths for the multi-type uas channel. *Queueing Systems 9*, 17–28.

[4] Johari, R. and Tan, D. K. H. (2001). End-to-end congestion control for the internet: delays and stability. *IEEE/ACM Transactions on networking* **9**, 6, 818–832.

[5] Kelly, F. P. (1989). On a class of approximations for closed queueing networks. *Queueing Systems 4*, 69–76.

[6] Kelly, F. P. (1991). Effective bandwidths at multi-class queues. *Queueing Systems 9*, 5–16.

[7] Kelly, F. P. (1997). Charging and rate control for elastic traffic. *European Transactions on Telecommunications 8*, 33–37.

[8] Kelly, F. P. (2003). Fairness and stability of end-to-end congestion control. *European Journal of Control 9*, 159–176.

[9] Kelly, F. P., Massoulié, L., and Walton, N. S. Resource pooling in congested networks: proportional fairness and product form. *Preprint*.

[10] Kelly, F. P., Maulloo, A. K., and Tan, D. K. H. (1998). Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society 49*, 237–252.

[11] Kunniyur, S. and Srikant, R. (2003). Stable, scalable, fair congestion control and aqm schemes that achieve high utilization in the internet. *IEEE Transactions on Control* **48**, 11, 2024–2028.

[12] Massoulie, L. and Roberts, J. (1998). Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems 15*, 185–201.

[13] Massoulie, L. and Roberts, J. (1999). Bandwidth sharing: Objectives and algorithms. *IEEE Infocom 1999* **10**, 3, 320–328.

[14] Mo, J.and Walrand, J. (2000). Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking 8*, 556–567.

[15] Schweitzer, P. J. (1979). Approximate analysis of multiclass closed networks of queues. *Proceedings of the international conference on stochastic control and optimization*.

[16] Srikant, R. (2004). *The Mathematics of Internet Congestion Control*. Birkhauser.

[17] Vojnovic, M., Boudec, J. Y. L., and Boutremans, C. (2000). Global fairness of additive-increase and multiplicative-decrease with heterogeneous round trip times. *Proc. IEEE Infocom 3*, 1303–1312.

[18] Walton, N. S. (2009). Proportional fairness and its relationship with multi-class queueing networks. *To appear in Ann. Appl. Probab.*.