A Scaling Framework for the Many Flows Asymptotic, through Large Deviations.

[Invited Presentation, Extended Abstract]

James Cruise *
University of Bristol
Department of Mathematics
Bristol,United Kingdom
marjrc@bristol.ac.uk

ABSTRACT

We introduce a new parameterization of the many flows asymptotic which allow a better understanding of how current results fit together but also offer more flexibility in the application of this asymptotic regime.

As we move into a world were the internet plays an ever increasing role a better understanding of interactions of various rate control protocols are required. As part of this we need to model the queueing dynamics at core routers which leads naturally to the study of the many flows asymptotic. This new scaling allows the easy exploration of different load and buffer sizing scenarios and there effect on packet loss probabilities.

1. MANY FLOWS ASYMPTOTIC

In engineering there are many queueing systems which we need to understand and model. Unfortunately in most situations we are unable to obtain exact results and so turn to the study of asymptotic regimes. Core network routers see many thousands of traffic flows from individual computers at any point in time, which naturally leads to the consideration an asymptotic regime where we scale the number of flows to infinity. This regime was initially introduced Alan Weiss [3] and is called the many flows asymptotic. Formally this means that in the N^{th} system there will be N independent identical sources.

2. SCALING PARAMETERIZATION

For the many flows asymptotic we are interested in exploring a range of scenarios and their relation to each other. These include considering various load scenarios where load (ρ) is defined to be the ratio of total mean arrival rate to mean service rate. We would like to be able to explore the spectrum of possible loads from heavily loaded where the load is close to one all the way to lightly loaded in which case the load is near zero. Another important class of scenarios to consider is defined by the buffer size relative to the number sources, i.e. whether the buffer size grows faster or slower than the total number of sources.

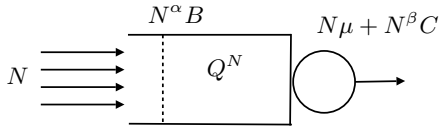

Figure 1: N^{th} system with parameters α and β.

To investigate these scenarios we introduce a new parameterization of the scalings indexed by (α, β). The parameter α is used to control the buffer size scaling, such that in the N^{th} system the buffer is size $N^\alpha B$. For $\alpha > 1$ the buffer grows faster than the number of sources where as for $\alpha < 1$ the buffer grows slower than the number of sources. Secondly β is used to control the excess service rate above the total arrival rate. We are interested in the stationary behavior of the queueing system so the service rate has to be larger than the mean total arrival rate. Under this condition we let Q^N be the stationary queue length of the N^{th} system. Let μ be the mean arrival rate from a single source then in the N^{th} system the service rate is $N\mu + N^\beta C$. For $\beta < 1$ the load tends to one as N increases giving the heavily loaded scenario. In comparison we have for $\beta > 1$ the load decreases to zero as the number of sources increase, the lightly loaded case. In figure 1 we can see a summary of the scaling for the single server queue.

3. SIMPLE MARKOVIAN EXAMPLE

To get a better feel for the parameterization we consider the example of a single server queue where each source produces Poisson traffic with intensity μ and service times are independently exponentially distributed with mean $(\lambda^{(N)})^{-1}$. It is useful to note that the sum of N independent Poisson processes of rate μ is again a Poisson process, of rate $N\mu$. The N^{th} system is an M/M/1 queue with arrival rate $N\mu$ and service rate $\lambda^{(N)}$. In this setting we have $\lambda^{(N)} = N\mu + N^\beta C$ and are interested in

$$\mathbb{P}(Q^N > N^\alpha B).$$

*Research supported by Heilbronn Institute for Mathematical Research

VALUETOOLS October 20-22, 2009 - Pisa, Italy

THEOREM 1. *Let Q^N be the stationary queue length where traffic is produced by N independent Poisson processes of rate μ and each customer has service time distributed exponentially with rate $N\mu+N^\beta C$. For $\alpha > 0$ and $\alpha+\beta-1 > 0$ we get*

$$\lim_{N\to\infty} \frac{1}{f(N,\beta)} \log \mathbb{P}(Q^N > N^\alpha B) = -J(\beta, C, B)$$

where

$$f(N,\beta) = \begin{cases} N^{\alpha+\beta-1} & \text{if } \beta \le 1 \\ N^\alpha \log(N^{\beta-1}) & \text{if } \beta > 1 \end{cases}$$

and

$$J(\beta,C,B) = \begin{cases} BC/\mu & \text{if } \beta < 1, \\ B\log(1+C/\mu) & \text{if } \beta = 1, \\ B & \text{if } \beta > 1. \end{cases}$$

4. TIMESCALES OF INTEREST

An important tool we make use of in exploring the different scalings is the most likely timescale upon which events of interest occur and how it varies with the number of sources. This is best understood by considering the length of time an overflow event takes to occur, i.e. the length of time it takes for the queue to go from empty to overflow given it overflows before emptying again. As we vary the scaling we are interested how the timescale varies in the limit. Often we can make use of the limit timescale in proving results.

To see the effect of the changing the timescales we again consider the simple Markovian example.

THEOREM 2. *Consider a single server fed by traffic produced by N independent identical Poisson sources with arrival rate μ and services times disturbed exponentially with mean $(N\mu+N^\beta C)^{-1}$. Let τ be the length of time it takes to overflow a buffer of size $N^\alpha B$ from empty given an overflow event occurs before the queue empties again.*

Then

$$\mathbb{E}(\tau) = \frac{N^{\alpha-\beta}B}{C}\left(1 - \frac{1}{N^\alpha B} - \frac{2\mu}{N^{\alpha+\beta-1}BC} + \frac{2}{(1+N^{\beta-1}C/\mu)^{N^\alpha B}-1}\right)$$

Furthermore if $\alpha \ge 0$ and $\alpha+\beta > 1$ we have

$$\lim_{N\to\infty} \mathbb{E}(\tau) = \begin{cases} 0 & \alpha < \beta, \\ B/C & \alpha = \beta, \\ \infty & \alpha > \beta. \end{cases}$$

From this we can see that the parameter $\alpha-\beta$ controls the limiting timescale in this case and we find that more generally this is true. So that in general for $\alpha-\beta < 0$ we see the timescale for events of interest tending to zero. This often means that we only need to consider covariance structures over very short intervals and in the case of point process we find that covariance structure does not have any effect [1].

5. INDUCED SAMPLE PATH SCALINGS

In the study of queueing system limits an important approach is that of obtaining results for the scaled sample paths for the arrival process and apply continuous mapping principles to obtain results for the various queueing statistics of interest [2, 4]. The power of this approach comes from the ability to prove results for the arrival process and quickly and easily study many different queueing systems, for example feed forward queueing networks or various service disciplines.

To investigate the sample path scaling induced by the (α,β) many sources scaling we consider the stationary queue length of a deterministic server, i.e. traffic is served at a constant rate. In the N^{th} system the rate of service is $N\mu+N^\beta C$ and we are interested in $\mathbb{P}(Q^N > N^\alpha B)$. We look to find a scaled process of the arrivals, $\tilde{A}^N_{\alpha,\beta}$, such that

$$\mathbb{P}(f_C(\tilde{A}^N_{\alpha,\beta}) > B) = \mathbb{P}(Q_N > N^\alpha B),$$

where f_C is the queueing map defined as

$$f_C(x) = \sup_{t>0}(x(t) - Ct).$$

Using this we find that the sample path scaling to consider is

$$\left\{\tilde{A}^N_{\alpha,\beta}(0,t)\right\}_{t>0} = \left\{\frac{\sum_{i=1}^N A_i(0, N^{\alpha-\beta}t)}{N^\alpha} - N^{1-\beta}\mu t\right\}_{t>0}.$$

6. REFERENCES

[1] R. Cruise. Poisson convergence, in large deviations, for the superposition of independent point processes. *Annals of Operations Research*, 170:79–94, 2009.

[2] A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues*. Springer Verlag, Berlin, 2004.

[3] A. Weiss. A new technique for analyzing large traffic systems. *Advances in Applied Probability*, 18(2):506–532, 1986.

[4] W. Whitt. *Stochastic-process limits*. Springer-Verlag, Berlin, 2002.