

# Optimal active probing rate for networks

[Extended Abstract]

Ben M. Parker  
Queen Mary, University of London  
Mile End Road  
London, UK  
b.parker@qmul.ac.uk

## Categories and Subject Descriptors

G.3 [Mathematics and Computing Probability and Statistics]: Experimental Design

## Keywords

Optimal Design of Experiments, Active Probing

## 1. INTRODUCTION

Recent research into broadband packet networks has considered the injection of probe packets to measure the packet level performance (e.g. loss, delay); for example whether it is best to probe at a uniform rate, or to send probes according to some renewal process, such as a Poisson process. In general this research has focused on probing of queuing systems as good general models of packet level network performance. Baccelli et al.[2] have shown that probes introduced with inter-arrival times following a gamma distribution have the lowest mean square error in estimating delay and packet loss, amongst all queues which have a convex auto-covariance structure. Roughan[3] has shown that there are fundamental bounds on how accurately network measurements can be made: that no matter how many active probes are used in a time interval, there is a limit to the knowledge we can gather about a queue. However a critical problem is that network traffic and topologies are highly variable and therefore measurement can be prone to very large errors in estimating end-to-end delay (both mean delay and jitter) and packet loss rates.

Increasing the number of probes within a given measurement period will give us more data, and thus increase the precision (lower the variance) of the estimate, but will mean that there are more probe packets to interfere with themselves and data packets, thus potentially lowering the accuracy (increasing the bias) of the estimate. In practice, a probing rate of one per second, or one per minute, is often adopted, although there appears to be little justification for choosing this rate, beyond simplicity of calculation.

In this work the crucial step is to view all network measurements as numerical experiments, in which random processes are sampled, and the effectiveness of the sampling is measured by a utility we place on bias and variance in the resulting estimator. In this way we are then able to apply the statistical principles of Design of Experiments (DOE) to network measurement experiments to develop a methodology which enables us to find an optimal active probing rate. We present a utility function that combines the bias and variance of the estimator with the added congestion in the network caused by probing.

## 2. METHODOLOGY

Suppose we have a two parameter queue into which packets arrive at a rate  $\lambda$ , which is fixed but unknown, and in which they are served and leave the system at a rate  $\mu$ , which is fixed and known. The queue discipline is also known.

We wish to find the value of  $\lambda$ , which we are unable to monitor directly; we can do this by introducing probe packets into the system at a rate  $x$ , and monitoring when these probes emerge. We can then make inference about  $\lambda$  from the amount of time that the probe packets spend in the system. Our goal is to find the value of  $x$  which allows us to best estimate  $\lambda$ .

We define  $S(t)$  as the amount of time required for all packets in the system at time  $t$  to exit the system. We call this the Virtual Waiting Time, because an imagined packet arriving at time  $t$  would spend a time  $S(t)$  waiting in the system.

This is a continuous time right-continuous process, which takes non-negative values. Instantaneous jumps occur when a packet enters the system at arrival times  $a_0, a_1, a_2, \dots$ . The jumps have magnitude depending on the queue discipline, which corresponds to service duration for the packet arriving at that arrival time. For example, in the M/M/1 queue the magnitude of the jumps correspond to the service times and are exponentially distributed with parameter  $\lambda$ .

The jump times,  $a_i$ , and the magnitude of the jumps of  $S(t)$  are random variables, but otherwise the process is deterministic, changing at rate  $-1$  (decreasing) until it reaches 0. Unless we have full knowledge of the queue, we cannot observe  $S(t)$  directly, but we make inference about it by introducing  $N$  probe packets at times  $\tau_1, \tau_2, \dots, \tau_N$ . By introducing new packets into the system, we form a new process  $S^*(t)$ .

We initially restrict ourselves to allowing probes which are Poisson distributed with rate  $x$ , and let the number of probes generated by the process be  $N$  as above (note  $N$  is a random variable). We denote the time between probe packet  $j$  entering and leaving the system as  $y_j$ , i.e. the system time. We are concerned with  $\hat{\lambda}$ , an estimate of  $\lambda$ . Given  $N = n$  packets, we find  $\tau_1 < \tau_2 < \dots < \tau_n$  and observe  $S^*(\tau_1), S^*(\tau_2), \dots, S^*(\tau_n)$  without error. We let  $y_i = S^*(\tau_i)$ , and our data are thus  $y_1, \dots, y_n$ .

In general we wish to minimise the bias and variance of some particular  $\hat{\lambda}$ , an estimator formed from some function of  $\mathbf{Y}$ ,  $x$ , and  $\mu$ . However, we also wish to minimise the disruption of the data packets on the network due to increased probe traffic. We measure this disruption as

$$D(x) = \sum_{i=1}^N c(S^*(a_i) - S(a_i)),$$

where  $c(w)$  is some cost function for a delay of one packet by an amount  $w$ . In general the underlying  $S(a_i)$  will be impossible to observe, and we can only estimate this by simulation.

Thus we are concerned with calculating a utility function  $\psi(\text{Bias}(\hat{\lambda}|x), \text{Var}(\hat{\lambda}|x), D(x))$ , and particularly with finding  $x_\lambda = \arg \min_x \psi(x)$ , our optimal probing rate.

## 2.1 Specifying a utility function

The exact form of the utility function will depend on how much we wish to trade accuracy and precision when estimating  $\lambda$  compared with the disruption caused by being able to measure it thus. When comparing bias and variance, a natural metric is the mean square error,  $\text{MSE}(\hat{\lambda}) = [\text{Bias}(\hat{\lambda})]^2 + \text{Var}(\hat{\lambda})$ . How much to penalise disruption is more subjective, and will depend on the queue under study and the experimenter's view on the relative merits of good estimates versus disruption. For example, in a network carrying SMTP traffic delays are less important than in one carrying VOIP traffic.

We propose a general form of the utility function

$$\psi(x) = k \text{MSE}(\hat{\lambda}) + (1 - k)D(x),$$

where  $D(x)$  is defined as above, and  $0 \leq k \leq 1$ . This framework will suit many applications, although other functions may be useful in particular circumstances.

## 3. AN EXAMPLE: M/M/1 QUEUE

We assume henceforth that our queue is an M/M/1 queue, although we seek to demonstrate an approach rather than provide results for a particular queue. We performed simulations setting  $\mu = 5s^{-1}$  throughout and allowed candidate points  $x$  to be in the range 0.1 to 2.4 seconds, at intervals of 0.1.  $\lambda$  was fixed at  $2.5s^{-1}$  to allow us to assess the bias and variance of our estimator, although we do not use knowledge of this  $\lambda$  when estimating  $\hat{\lambda}$ . Any reasonable estimator may be picked, and different estimators will in general produce different estimates and thus different optimal rates. Following Aigner[1], we picked an estimator

$$\hat{\lambda} = \frac{\frac{1}{N} \sum_{i=1}^N Y_i - \frac{1}{\mu}}{\frac{1}{N} \sum_{i=1}^N Y_i \frac{1}{\mu}}.$$

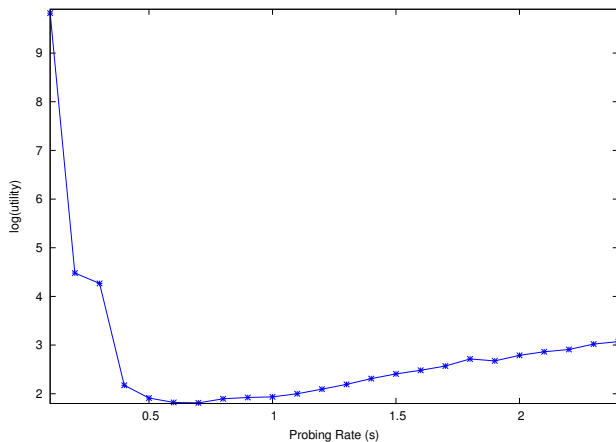


Figure 1: Utility function for M/M/1 queue:  $\log(\psi(x))$  against  $x$

We assumed that we were able to perform the probing experiment only for 10 seconds. We performed 1000 simulations for each candidate point, and by looking at the 1000  $\hat{\lambda}$  generated for each candidate point, were able to estimate the bias and variance of  $\hat{\lambda}$ . Knowing for the simulation the underlying virtual system time process  $S(t)$ , and the altered process after probing  $S^*(t)$ , we were able to estimate the value of our disruption function  $D(x)$ , letting  $c(z) = z$ . In other words we penalise each packet delay linearly. To illustrate a possible utility function, we set  $k = \frac{1}{2}$ .

The results are displayed as Figure 1. The optimal probing rate is shown as the minimum on the graph, here when  $x \approx 0.7s^{-1}$ . A low rate ( $x < 0.4$ ) gives significantly worse results (higher utility) than a high probing rate ( $x > 0.7$ ).

## 4. CONCLUSIONS

DOE techniques have been successfully applied, particularly in biological and industrial contexts. Our work on DOE for network queueing models has provided new insight and results in the design of numerical experiments for network monitoring, and promises to deliver a general framework within which an optimal probing strategy could be determined for any given networking scenario.

## 5. REFERENCES

- [1] D. J. Aigner. Parameter estimation from cross-sectional observations on an elementary queuing system. *Operations Research*, 22 issue 2:422, 1974.
- [2] F. Baccelli, S. Machiraju, D. Veitch, and J. Bolot. On optimal probing for delay and loss measurement. pages 291–302. ACM New York, NY, USA, 2007.
- [3] M. Roughan. Fundamental bounds on the accuracy of network performance measurements. *Proceedings of the 2005 ACM SIGMETRICS international conference on measurement and modeling of computer systems*, pages 253–264, 2005.