Notes on Statistical Multiplexing

Introduction

Statistical multiplexing refers to the phenomenon whereby sources with statistically varying rates are mixed or input into a common server or buffer. Typical sources that occur are "bursty"- there are periods when they generate bit or packets at a high rate (ON state) while there are other periods when they generate a few or no packets (OFF state). Typical sources are depicted in the Figure 1:

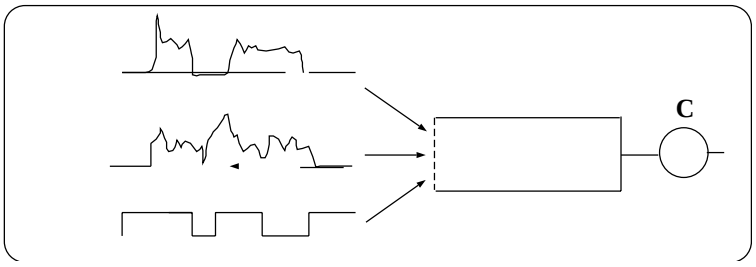

Figure 1: Statistical multiplexing of bursty sources

Because of statistical independence it is a very unlikely scenario when all sources will be simultaneously in the ON state (especially when there are many), and thus to design a server to serve at a rate corresponding to the maximum sum rate of all the sources would be very wasteful. If we allow for a fraction of the offered traffic to be lost then we will see that it is possible that a link of given capacity can carry much more traffic than would be the case if we assumed that the traffic was synchronized.

A caricature of a typical scenario whereby one could exploit the statistical independence is shown in Figure 2:

In these notes we will study the issue of statistical multiplexing and show that there is a very important concept that emerges, namely the notion of *effective bandwidths*. This has been one of the major conceptual advances that emerged in the 1990s when the issue of providing Quality of Service (QoS) became important in ATM networks. The attractiveness of this idea is that it allows us to map a packet level phenomenon to a flow or connection level phenomenon i.e. allows us to ignore a queueing model and convert it into a loss model. This will be expanded in the course of our discussions.

Quality of Service (QoS) is a buzzword that became popular in the late 1980s when the idea of Asynchronous Transfer Mode (ATM) networks emerged. The idea here was that the transport of packets or bits within a network was going to be assured a an architecture that resembled a circuit-switched scenario (fixed paths) except that rather than reserving resources along a path fixed at a given level, the traffic streams were to be split into packets of fixed size called cells and the statistical variation in cell generation as in Figure 1 was to be exploited to "pack in" many more connections by allowing the variability in rates to be used. However since one could not predict the OFF moments there would be periods when there would be temporary overload of the capacity and cells would be lost as in Figure 2. So if one could tolerate the loss of some bits (not critical in many voice and video applications) then one could pack in many more sources. In the ATM context

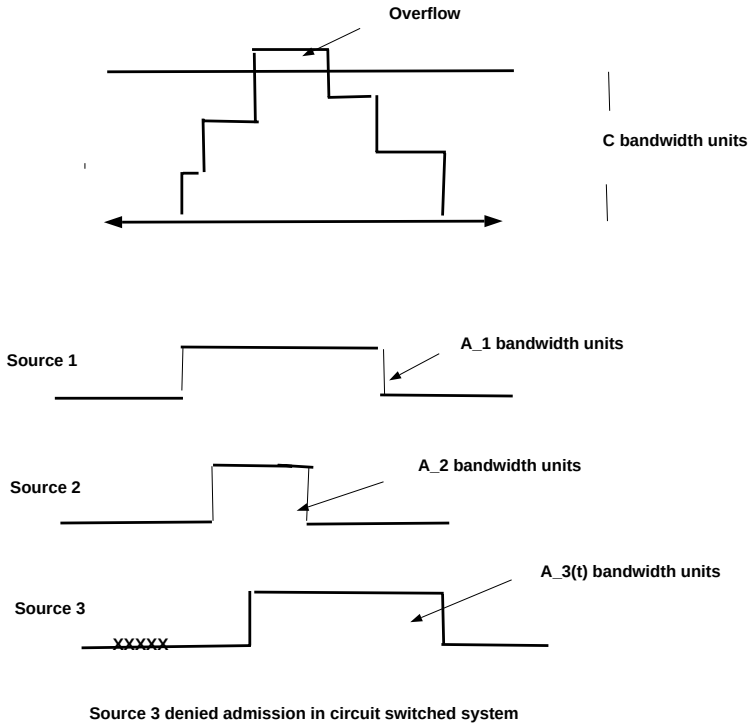

Figure 2: How statistical multiplexing works

this was set at a probability of packet loss being of the order of 10^{-6}. This is indeed a very small probability and thus the question arose is how does one determine that indeed such a criterion is being met. Simulations are one way but that would be extremely cumbersome given the extremely small probabilities. Thus arose an industry in developing methods to estimate quantities like this. We will discuss these issues later. We first begin by defining the various QoS metrics of interest.

1 Performance metrics for Quality of Service (QoS)

The key performance metrics in a network are the notions of average throughput, delay characteristics and loss. In the context of statistical multiplexing we will restrict ourselves to two of these measures namely delay and packet or bit loss.

In the context of delays, the key statistical measures are the average delay and the delay distribution. QoS requirements are usually specified by bounds on the average delay or by giving bounds on the probability of the delay exceeding a certain level. Thus if D represents the delay, the performance measures are usually $\mathbb{E}_A[D] < D_{max}$ or the $\Pr(D > D_{max}) \leq \varepsilon$ where ε is a very small number. Here $E_A[.]$ denotes the mean with respect to the arrival distribution.

In the case of packet or bit loss we are usually interested in $\Pr(Packet(bit)\ is\ lost) \leq \epsilon$. We will see that both the packet loss probability or the delay distribution are related to computing the tail probabilities associated with buffer or capacity exceedence.

Let us first define the various quantities. We begin with a simple scenario when there is no buffering. Let X_t denote the instantaneous rate of a source transmitting on a link of capacity C bits per second. Assume $\mathbb{E}[X] < C$. Then the total number of bits lost in an interval of length T is: $Bits\ lost = \int_0^T (X_t - C)^+ dt$. If the source is stationary and ergodic then the average number of

bits lost is

$$\begin{aligned} Loss &= \lim_{T\to\infty} \frac{1}{T}\int_0^T (X_t - C)^+ dt = \mathbb{E}[(X_0 - C)^+] \\ &= \int_C^\infty (1 - F(x))dx \\ &\approx const.\mathbb{P}(X > C) \end{aligned}$$

where the last approximation is when C is large and $F(x)$ denotes the distribution of X_t.

A related quantity of interest is the fraction of bits lost which is just:

$$\lim_{T\to\infty} \frac{\int_0^T (X_t - C)^+ dt}{\int_0^T X_t dt} = \frac{\mathbb{E}[(X_0 - C)^+]}{\mathbb{E}[X_0]}$$

When queueing is involved the commonly considered performance measure is the overflow probability that is approximated by the tail distribution of the buffer occupancy assuming the buffer is infinite given by:

$$\mathbb{P}(W_0 > B) \qquad \text{(Overflow Probability)}$$

Here B represents the finite buffer size that we would like to consider. When B is large the tail distribution is a good approximation for the overflow probability.

Thus. in both cases for measures of packet loss, we see the quantity of interest is related to the computation of tail or the complementary distribution function.

2 Statistical multiplexing and effective bandwidths- Motivation

To understand the idea statistical multiplexing we will consider a very simple case. This clearly brings out the ideas.

Consider the case when the desired performance measure is the average delay. Consider the following $M/G/1$ model where there are N sources that are transmitting at a Poisson rate of λ packets per second. The sever serves at a rate of C bits per second. The packet sizes are variable and uniformly distributed in $[0, M]$ where M represents the maximum packet size in bits.

For stability of the queue the maximum number of sources that could be supported is given by: $N\lambda\frac{M}{2} < C$ or $N < \frac{2C}{\lambda M}$.

Now consider the maximum number of sources that could be admitted if the packets were all of size M . This is given by $\frac{C}{\lambda M}$ or the fact that the packets are uniformly distributed allows the system to carry twice the number of sources.

Now let us calculate the maximum number of sources that can be admitted in the system if the bound on the mean delay is D. From the Pollaczek-Khinchine formula we obtain by noting that $\mathbb{E}[\sigma^2] = \frac{M^2}{3C^2}$ and $\mathbb{E}[\sigma] = \frac{M}{2C}$:

$$\mathbb{E}[W] = \frac{\lambda\mathbb{E}[\sigma^2]}{2(1-\rho)} = \frac{N\lambda\frac{M^2}{3C^2}}{2(1 - N\lambda\frac{M}{2C})} \leq D$$

or

$$N\left(\frac{\lambda M^2}{6CD} + \frac{\lambda M}{2}\right) \leq C \tag{2.1}$$

Hence the max number of sources that can be admitted for meeting the average delay constraint is :

$$N \leq \frac{C}{\frac{\lambda M^2}{6CD} + \frac{\lambda M}{2}}$$

Clearly as $D \to \infty$ which corresponds to an unconstrained system we see that the maximum number coincides with the number corresponding to the maximum number to guarantee stability namely $\frac{2C}{\lambda M}$.

A similar caculation assuming all of the packets are of maximum size M gives the maximum number as:

$$N \leq \frac{C}{\frac{\lambda M^2}{CD} + \lambda M}$$

Let us define the multiplexing gain denoted by MG as follows:

$$MG = \frac{\text{Max number of admitted sources with statistical variation}}{\text{Max number of sources with maxpacket size}}$$

Then we see that:

$$MG = \frac{\frac{\lambda M^2}{CD} + \lambda M}{\frac{\lambda M^2}{6CD} + \frac{\lambda M}{2}}$$

Now as $C \to \infty$ or as the link capacity grows we see that $MG \to 2$ which is the ratio of the number of sources that can be admitted for stability assuming that the packets are uniformly distributed in length to the number of sources that can be admitted assuming all the packets are of maximum size.

In other words, the statistical multiplexing gain approaches the ideal gain possible as the capacity C of the server increases and thus the more the number of sources we can multiplex the better are the gains.

Now, equation 2.1 , can be viewed as $NA \leq C$ where the quantity

$$A = \frac{\lambda M^2}{6CD} + \frac{\lambda M}{2} \tag{2.2}$$

can be thought of as the *effective bandwidth* in analogy with loss systems. Note $\frac{\lambda M}{2}$ is just the mean load ρ and thus $\rho \leq A$ and $A \to \rho$ as $C, D \to \infty$ or the effective bandwidth is the same as the mean when the capacity becomes large or there is no delay bound.

With this definition of A the probability that the delay bound is exceeded can be computing by considering the Erlang loss formula with $\lambda = N\lambda$, bandwidth requirement A and mean holding time 1 and taking capacity C. So the effective bandwidth can be thought as a way to map a queueing system characteristics into an equivalent loss system.

The above discussion helps to understand the notion of effective bandwidths and the fact that we can achieve gains for the same delay constraint when the capacity is large.

It can readily be seen that if we multiplex$\{n_k\}_{k=1}^{N}$ independent inputs, each of which generate packets at rates λ_k and have different packet length distributions, then the delay bound will be met if:

$$\sum_{k=1}^{N} n_k A_k \leq C \tag{2.3}$$

where $A_k = \frac{\lambda_k M_k^2}{6CD} + \frac{\lambda_k M_k}{2}$ is the effective bandwidth of packets.

Thus we define the admission region Ω as the region where the number of sources such that the delay constraint is met as:

$$\Omega = \left\{ \{n_k\} : \sum_{k=1}^{N} n_k A_k \leq C \right\} \tag{2.4}$$

In the above discussion we took the packets to be uniformly distributed but it is clear such an analysis can be carried out for any general distribution.

A similar effect takes place when we consider the packet loss characteristics. We look at the main ideas here.

3 QoS -Flow Level Models

The previous example considered packet arrivals as a Poisson process. However in modern high-speed networks such models are inadequate and a more appropriate model is one that models session arrivals as discrete events but the flow of information within a session is modelled as a continuous process whose rate is random. This leads to a fluid view of the queueing process rather than a customer-centric view.

Let us recall in such a viewpoint, the total work in the buffer or workload is now the quantity of interest. Let W_t denote the buffer content at time t. Then, for $t \geq 0$

$$W_t = W_0 + A(t) - C \int_0^t \mathbb{1}_{[W_s > 0]} ds \tag{3.5}$$

where $A(t)$ denotes the total number of bits arriving in the interval $(0, t]$ and C denotes the server rate. Note C need not be a constant or even deterministic (as in wireless links where the rate at which the link operates is random depending on the interference, channel conditions, etc.). In such a case we replace $C \int_0^t \mathbb{1}_{[W_s>0]} ds$ by $C(t)$ where $C(t)$ denotes the dotal number of bits served during $(0, t]$.

It can be shown that if the input has stationary increments and $E[A(0,1] = \rho < C$, where ρ is the average work brought in a unit of time, then the stationary workload is given by:

$$W_0 = \sup_{t \geq 0} \{A(-t, 0] - Ct\}$$

When time is discrete a similar relation

$$W_0 = \sup_{k \geq 1} \{A[-k+1, 0] - Ck\} \tag{3.6}$$

where $A[-k+, 0]$ denotes the number of bits arriving in $[-k+1, 0]$

It is very easy to see how this relation arises. Let $-k$ denote the last time before 0 when the queue is empty. Then:

$$W_0 = A[-k+1, 0] - Ck$$

But if the queue is stable then a k (which is a random time) must exist.

Hence:

$$W_0 = \bigcup_{k \geq 1} \{A(-k+1, 0) - Ck\}$$

since k can take any value and the union is to be understood as $A \subset B$ if $B \geq A$, and hence taking the union over all the values will definitely account for the k. Now the event of the union over all k is just equivalent to equation 3.6.

We now focus our attention on the discrete-time model. Now to compute the overflow probability we need to compute:

$$\mathbb{P}(W_0 > B) = \mathbb{P}(\sup_{k\geq 1}\{A[-k+1,0] - Ck\} > B)$$

Noting the interpretation of the sup we obtain using the union bound ($\mathbb{P}(\cup A_i) \leq \sum \mathbb{P}(A_i)$) that:

$$\mathbb{P}(W_0 > B) \leq \sum_{k\geq 1} \mathbb{P}(A[-k+1,0] - Ck] > B)$$

So we need to estimate probabilities $\mathbb{P}(A[-k+1,0] - Ck > B)$.

Now to estimate these probabilities we use ideas from the theory of Large deviations (assuming Band C are such that the probabilities are small so that the rhs of the union bound leads to a quantity that is non-trivial knowing that there are an infinite number of terms.

It is convenient to see this by first considering $A(-K-1,0] = \sum_{j=-k+1}^{0} r_j$ where r_j are i.i.d and denote the number of bits that arrive at time k. Then if we use the Chernoff bound

$$\mathbb{P}(A[-k_1,0] > Ck + B) \leq e^{-\theta B} e^{-k(\theta C - \Gamma(\theta))}$$

where $e^{\Gamma(\theta)}$ is the moment generating function of r_j. Now suppose $\frac{\Gamma(\theta)}{\theta} < C$ and calling $\alpha = C\theta - \Gamma(\theta) > 0$ we see that:

$$\mathbb{P}(W_0 > B) \leq e^{-\theta B} \sum_{k=1}^{\infty} e^{-k\alpha} = const.e^{-\theta B} \tag{3.7}$$

or we have bounded the overflow probability by a negative exponential.

One can generalize this result to the non i.i.d case by assuming that the input has stationary and ergodic increments and by defining:

$$\Gamma(\theta) = \lim_{n\to\infty} \frac{1}{n} \log\left(\mathbb{E}[e^{\theta A[-n+1,0]}]\right)$$

and carrying out the same type of argument. Also the result given by equation **??** and the moment generating function now viewed as the cumulative input in an interval that is continuous can be shown when we consider the continuous-time fluid queueing model. It is only a bit more technically difficult since the union bound cannot be simply applied.

Now the quantity $\frac{\Gamma(\theta)}{\theta}$ is now identified as the *effective bandwidth*. Let us study some of its properties.

- If $r_j \in [0,R]$ with $\rho = E[r_j] < R$, here R is the peak rate, then

  $$\rho \leq \frac{\Gamma(\theta)}{\theta} < R$$

  This follows from the fact that $E[e^X] \geq e^{EX}$ and so $\Gamma(\theta) \geq \theta\rho$ and $\Gamma(\theta) \leq \theta R$ trivially.

  So as in (2.2) we see the definition of the effective bandwidth gives a number between the mean and peak rate.

- The second property is that the effective bandwidths are additive. Let $A[-k.0] = \sum_{j=1}^{N} A_i[-k,0]$ where A_i denote independent inputs. Then it is easy to see that $\Gamma(\theta) = \sum_{j=1}^{N} \Gamma_j(\theta)$ and so the effective bandwidths are additive.

- Finally let us show the statistical multiplexing property, as we multiplex more and more sources then the effective bandwidth converges to the mean. For simplicity let us consider the case when there are N identical statistically independent sources accessing a buffer whose buffer size is NB and capacity is NC. Note the worst case delay given by $\frac{B}{C}$ is constant for all N. Now suppose the overflow probability bound is $\epsilon = e^{-\delta}$. Then clearly we need $\theta NB = \delta$ or $\theta = \frac{\delta}{NB}$. Now clearly $\Gamma_N(\theta) = N\Gamma(\theta)$ by independence. Hence suppose θ is such that: $N\frac{\Gamma(\frac{\delta}{NB})}{\frac{\delta}{NB}} < NC$ then we see that as $N \to \infty$ we have

$$\lim_{N\to\infty} \frac{\Gamma(\frac{\delta}{NB})}{\frac{\delta}{NB}} \to \Gamma'(0) = \rho$$

  by definition of the moment generating function. This shows that the statistical multiplexing gain that matches the max number of sources that can be admitted for stability is obtained in large systems as in the mean delay discussion above.

As above one can readily define the admission region of the number of sources that could be supported by a buffer of size B that is drained at rate C assuming that there are N types of sources, by:

$$\Omega = \left\{ \{n_k\} : \sum_{k=1}^{N} n_k \frac{\Gamma_k(\theta)}{\theta} \leq C \right\} \tag{3.8}$$

where $\theta = \frac{\delta}{B}$ where $\delta = -\ln(\varepsilon)$ and ε is the bound on the overflow probability.

Note by the definition of Ω the admission region has the same form of the state-space of a multi-rate Erlang model, and so if we know the rate of arrival of the sources say λ_k for type k, assuming that they are Poisson(which can be justified and verified in practice), by computing the blocking probability for that type, we obtain the probability that the QoS constraints cannot be met.

Thus, one of the great advantages of the effective bandwidth is we can map the queueing model to an equivalent loss model. This is important for service providers as they can install routers with appropriate buffers and speeds to meet QoS requirements without getting into details about the queueing. Indeed, the only difficulty is that one needs the moment generating functions $\Gamma_k(\theta)$ which might be difficult to obtain in practice. So one way is to *shape* or regulate traffic flows to conform to a profile whose moment generating function is known.

The Figure 3 indicates the type of gain in the admission region defined by the number of sources that can be admitted while meeting statistical QoS as opposed to the worst case (deterministic) results and the stability region is also indicated. Of course, if the system is large the stability region very closely approximates the admissible region as we have seen. Thus there can be substantial gains to be had by considering statistical criteria for admission control. The drawback is that more information is required about the sources.

Final remarks:

The ideas of effective bandwidths can be traced to [2]. Kelly [3] provides a very general way of looking at the effective bandwidths issue for both buffered and unbuffered systems. Finally in [4] the application of these ideas as well as rigorous results on the effective bandwidths, the ideas of statistical multiplexing, and admission control schemes for large bandwidth links is given.

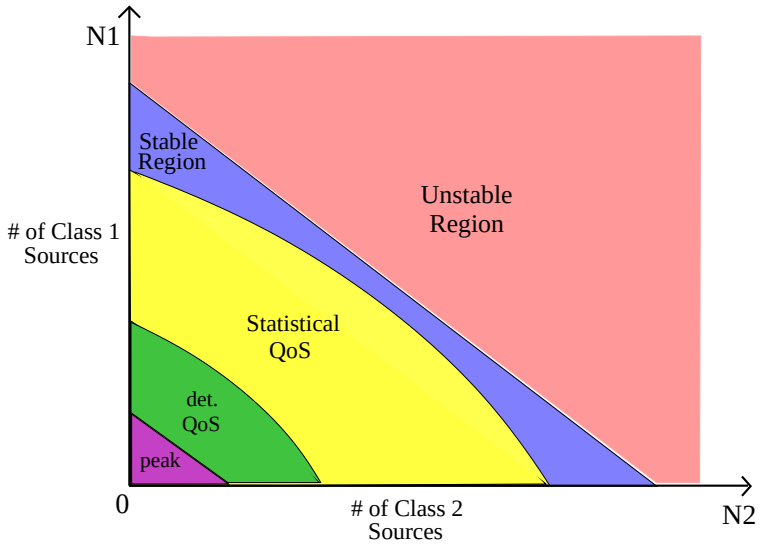

Figure 3: Admission regions under various criteria

References

[1] A. Kumar, D. Manjunath, and J. Kuri: *Communication Networking: An analytical approach*, Morgan-Kaufman (Elsevier), 2004.

[2] Hui, J.Y. (1988). Resource allocation in broadband networks,. *IEEE Journal of Selected Areas in Communications*, 6:1598–1608.

[3] Kelly, F. P. (1996). Notes on effective bandwidths. In Kelly, F.P., Zachary, S., and Zeidins, I., editors, *Stochastic Networks*. Oxford University Press.

[4] N. B. Likhanov, R. R. Mazumdar, and F. Theberge Providing QoS in Large Networks: Statistical Multiplexing and Admission Control In E. Boukas and R. Malhame , editors, *Analysis, Control and Optimization of Complex Dynamic Systems*, Springer 2005, pp 137-168.