differing in approach and in detail, a common factor in all these descriptions is that some kind of differentiation (in the packet handling) is introduced.

In this paper we put focus on the delay and delay variation experienced in an IP domain for which some form of delay (variation) commitment is intended. The commitments may vary in strictness, ranging from a strict guarantee to a more loosely defined objective. Especially, for network operators it is important to have some estimates of the end-to-end delay for traffic crossing their network domain. These estimates will be an important part of the delay guarantees offered between operators through the Service Level Agreements (SLAs).

The least known factor in the end-to-end delay of an IP packet is the delay contribution due to queuing in the network elements. This contribution is also important because, in most cases, the delay variation introduced by the network is to be removed by the receiving application (de-jittering), thus introducing additional de-jitter delay which is necessarily at least as large as the maximum (or suitable quantile) of the delay variation. Other factors, such as the contribution of the propagation delay and the variation in router's routing look-up latency are expected to either be much easier to assess or to be negligible in comparison to the queueing delay due to statistical multiplexing in the routers.

We therefore address some methods to calculate the queuing delay in a network where several interactive IP flows are mixed with best-effort IP flows and where the interactive packets have strict priority over best effort traffic. The objective is to provide both a suitable model and suitable assessment techniques (e.g. calculation methods) to arrive at a suitable estimate of the queuing delay in a given situation. By 'suitable' we mean a model which is sufficiently close to the real world to have a practical meaning and, at the same time, is sufficiently easy to allow the calculations to be carried out. Earlier work [4], [5] usually addressed some elements of this problem, for example by providing separate estimates for the delay contribution due to interactive packets and due to best-effort packets or by providing separate estimates for the contribution of the core network and the access network.

2 Queueing models

2.1 Some preliminary considerations

In the following we shall consider a DiffServ scenario for the end-to-end delay for typical Real Time (RT) traffic in a large scale IP-network. We consider a path in the network consisting of a given number of (say) K nodes and the aim is to calculate the distribution of the queueing delay for that particular path. We assume that each node may be considered as a non-preemptive priority queueing system with two priority classes where the RT traffic is scheduled as highest priority and the Best Effort (BE) type traffic is scheduled as lower (second) priority. To calculate the distribution of the delay of a particular path we make the following assumption (approximation): All nodes in the end-to-end path are statistically independent. This is the key assumption for the model and makes it possible to obtain the end-to-end delay by convolution. Under which conditions the independent assumption applies is not quite clear, but it seems to be reasonable for rather thin streams where the aggregate flows split at each node and are mixed with traffic from different nodes.

We shall therefore take the M/G/1 non-preemptive queueing system as the model to obtain the waiting time distribution (for the high priority RT packets) in each node, and then we apply convolution to get the end-to-end waiting time distribution. If we let W_k denote the waiting time in the k'th node for the RT-packets then the total delay may be written $W_{NP}^{H} = W_1 + ... + W_K$, and the Laplace-Stieltjes Transforms (LST) of the sum is given as the product of the LST of the waiting times in the individual

the waiting time for a single M/G/1 queue, then simple partial derivative with respect to the parameter ρ yields:

$$\tilde{W}^T(s,\rho) = \left(\tilde{W}(s,\rho)\right)^K = \frac{(1-\rho)^K}{(K-1)!}\frac{\partial^{K-1}}{\partial\rho^{K-1}}\left\{\frac{\rho^{K-1}}{1-\rho}\tilde{W}(s,\rho)\right\} \tag{4}$$

The same result will also apply for the DF (and also PDF) of the convolution, hence we may obtain the convolution by the following formula:

$$W^T(t,\rho) = \frac{(1-\rho)^K}{(K-1)!}\frac{\partial^{K-1}}{\partial\rho^{K-1}}\left\{\frac{\rho^{K-1}}{1-\rho}W(t,\rho)\right\} \quad \text{where} \tag{5}$$

$W(t,\rho)$ denotes the DF of the waiting time in a M/G/1 queue with load ρ. Similar and more extended results may be found in the thesis [9].

The DF of the end-to-end queueing delay can therefore be written as the sum obtained by inverting the LST (2) and (3):

$$W_{NP}^H(t) = (1-p)^K W^T(t,\rho^H) + \sum_{r=1}^{K} b_r(p,K) W^T(t,\rho^H)(*)\hat{b}^L(t)^{*(r)} \tag{6}$$

where $b_r(p,K) = \binom{K}{r} p^r (1-p)^{K-r}$ is the binominal probabilities with parameters $p = \frac{\rho^L}{1-\rho^H}$, and $\hat{b}^L(t)^{*(r)}$ is the r-times convolution of the Probability Density Function (PDF) of the remaining service times for the low priority packets ($(*)$ denotes convolution). The expression (6) represents the general case without any specific assumptions on the actual service time distributions. To carry the analysis any further specific choices on the service time distributions therefore have to be made.

2.3 Deterministic service times for low and high priority packets

In the following we assume that both the high and low priority packets have constant service times given by b^H and b^L, respectively. In this case we have the well-known results for the M/D/1 queue:

$$W(t,\rho^H) = q\left(\frac{t}{b^H},\rho^H\right) \quad \text{where} \quad q(x,\rho) = (1-\rho)\sum_{k=0}^{\lfloor x \rfloor}\frac{\lceil\rho(k-x)\rceil^k}{k!}e^{-\rho(k-x)} \tag{7}$$

is the DF of the waiting time in a M/D/1 queue with service times scaled to unity. The K-fold convolution of $W(t,\rho^H)$ is found from relation (5) by differentiation [8]:

$$W^T(t,\rho^H) = q^K\left(\frac{t}{b^H},\rho^H\right) \quad \text{where} \quad q^K(x,\rho) = (1-\rho)^K\sum_{k=0}^{\lfloor x \rfloor}\sum_{l=0}^{K-1}\frac{(-1)^l}{l!k!}\binom{K+k-1}{K-l-1}\left(\rho(k-x)\right)^{k+l}e^{-\rho(k-x)} \tag{8}$$

Further we have that the remaining service times for the low priority packets are uniform distributed over the interval $(0,b^L)$, and we find the r-time convolution $\hat{b}^L(t)^{*(r)}$ on the following form:

$$\hat{b}^L(t)^{*(r)} = \frac{r}{\left(b^L\right)^r}\sum_{m=0}^{r}\frac{(-1)^m}{m!(r-m)!}H(t-mb^L)(t-mb^L)^{r-1} \tag{9}$$

$$W_{NP}^{H}(t)=\left(\frac{b^H}{b^L}\right)^K\sum_{k=0}^{\left\lfloor\frac{t}{b^H}\right\rfloor}\sum_{m=0}^{\left\lfloor\frac{t-kb^H}{b^L}\right\rfloor}(-1)^m\binom{K}{m}\binom{k+K-1}{K-1}\left(q^{K,K-1}(\frac{t-kb^H-mb^L}{b^H},\rho^H)+(-1)^K\frac{(1-\rho)^K}{\rho^K}\right) \quad (14)$$

C. *Saturated system and the fraction between high a low priority service times are an integer*.

We find:

$$W_{NP}^{H}(t)=\sum_{k=0}^{\left\lfloor\frac{t}{b^H}\right\rfloor}c_K(k,l)\left(q^{K,K-1}(\frac{t}{b^H}-k,\rho^H)+(-1)^K\frac{(1-\rho)^K}{\rho^K}\right) \quad (15)$$

3 Approximative methods

To this end some different types of approximations exist for sums of independent random variables in general. The most famous one is the "ordinary" Normal Approximation (NA) quoting that a sum identically independent random variables approach a normal distribution when the number of variables increases. Since the normal distribution is characterized by its two first (lowest) moments, this leads to the following approximation for the PDF and DF of the convolution:

$$w_{NP}^{HNA}(t)=\frac{1}{\sqrt{K}\sigma}\varphi(\frac{t-Km}{\sqrt{K}\sigma}) \text{ and } W_{NP}^{HNA}(t)=1-\phi(\frac{t-Km}{\sqrt{K}\sigma}) \text{ where} \quad (16)$$

$m=m^H+m^L$ and $\rho^2=\rho^{H\,2}+\rho^{L\,2}$. Further $m^H=E\left[W^H\right]$ and $\rho^{H\,2}=E\left[W^{H\,2}\right]-m^{H\,2}$ are the mean and variance of the queueing delay in the corresponding single server M/G/1 queue with only high priority traffic present, given in terms of the three first moments of the service time distribution (and also the load) through:

$$m^H=\frac{E\left[B^{H\,2}\right]}{2E\left[B^H\right]}\frac{\rho^H}{1-\rho^H} \quad \text{and} \quad \sigma^{H\,2}=\frac{E\left[B^{H\,3}\right]}{3E\left[B^H\right]}\frac{\rho^H}{1-\rho^H}+m^{H\,2} \quad (17)$$

and the influence from the low priority traffic is given through the remaining service times for a low priority packet by:

$$m^L=p\frac{E\left[B^{L\,2}\right]}{2E\left[B^L\right]} \text{ and } \sigma^{L\,2}=p\frac{E\left[B^{L\,3}\right]}{3E\left[B^L\right]}-m^{L\,2}. \quad (18)$$

Further $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ is the standard normal density and $\phi(x)=\frac{1}{\sqrt{2\pi}}\int_{t=x}^{\infty}e^{-\frac{t^2}{2}}dt$ is the standard normal integral.

A second and quite different approximation (though also involving normal distributions) is based on Large Deviation (LD) theory. Formally this approximation is obtained as the asymptotic behaviour of $w_{NP}^{H}(Kx,\rho)$ for large values of K and fixed x. This can be reflected by considering the inversion integral of the LST of the end-to-end delay:

$$w_{NP}^{H}(Kx,\rho)=\frac{1}{2\pi i}\int_{\gamma}e^{-Kg(s,x)}ds \quad \text{where} \quad g(s,x)=\log(1-\rho^H\hat{B}^H(s))-\log(1-p+p\hat{B}^L(s))-sx-\log(1-\rho^H) \quad (19)$$

classes and we focus on the end-to-end delay for the high priority traffic. The high priority class will typically be real time traffic like voice and video that will have constraints on the maximum end-to-end delay. Under the assumption that the load from the high priority traffic is limited we would like to find the effect the low priority traffic will have on the performance of the high priority real time classes. This is a typical situation in IP networks deploying DiffServ. In an IP network the end-to-end delay for a particular stream will be the sum of the delay obtained in a cascade of routers (from the sender to the receiver). The total end-to-end delay will then consist of the waiting times in each node plus service times (transmission times onto the links). In the examples below we consider a particular chain of routers in a packet network and we assume that the routers have output buffers with two priority classes with no extra internal delay due to processing of the packets.

To apply the results in chapter 2 we must assume that all the nodes in the chain have identical parameters, which means that the link capacity is equal for all the routers and that packets for the two priority classes arrive according to Poisson processes with parameters that are equal in each router. In addition we assume that the packet lengths for the high and low priority class is constant mean P_H and P_L respectively. (All the numerical results are scaled according to the transmission time for a high priority packet.) We shall assume that routers are saturated, this means that there will always be low priority packets to be transmitted, implying that the low priority load $\rho_L = 1 - \rho_H$. By the last assumption we may use the somewhat simplified formula given by the equations (14) and (15) in chapter 2 to obtain the DF of the end-to-end delay distribution. With these definitions we get mean service times for high and low priority packets $b_H = P_H / C$, $b_L = P_L / C$ where C is the link capacity. In the examples below we have chosen scenarios among the following parameter values: The ratio between low and high priority packets P_L / P_H is either 1,5 or 10, and the load from high priority traffic ρ_H is either 0.4 or 0.6, and further the number of hops K is either 10 or 15.

In Figure 1 and Figure 2 we have depicted some results for the case when the low priority packet lengths are assumed to be constant. This case could represent the case when we have the packet length more or less limited by an Ethernet frame of 1500 bytes and in addition by assuming rather short real time packet lengths of around 200 bytes. We observe rather strong impact from the ratio of the low and high priority packet lengths. The influence of the load from the high priority traffic is not that strong and this is more or less expected since we assume that the high priority load is limited to say less than 60%, which seems to be reasonable keeping in mind the need to reserve some part of the capacity also for low priority traffic. If we for instance take an example with STM-1 links of approximately 150 Mbit/s and assume that the real time packet lengths are 200 bytes, this will give packet transmission time of around 10 μsec. By assuming a path of 15 hops and assuming packet length ratio of 10, then Figure 2 provide us with the appropriate quantile. If we take the $1\text{-}10^{-3}$ quantile for the highest load we find the appropriate value to be around 125 (high priority packet transmission time), and this leaves us with a value of 1.25 ms for this particular case. This tells us that if a core network deploying DiffServ is properly engineered so that the high priority load is limited to say 60% then the end-to-end queueing delay will be limited to a few milliseconds. (One has to add the contributions from the access part of the particular path to get the complete picture, and this contribution could be larger due to slower links in the access network.)

bound. To apply this method one has to locate the saddle point for each value one would calculate the distribution function of. The UAA is an excellent approximation and gives a uniform approximation of the distribution over the whole range of the distribution function. The relative error is very small in all the cases we have considered (less than 3%) and it is nearly impossible to make the distinction in the graphs. It also seems to give accurate estimates for values where the asymptotic is not fulfilled, i.e. for chains consisting of only one or two queues.

5 Conclusion

We have given and discussed some methods to calculate the end-to-end queueing delay in a packet network where real time traffic has strict priority over other classes of traffic. The described method could for instance be applied to estimate typical end-to-end delay in a core IP network domain deploying DiffServ. The proposed methods are tested against known approximation such as the saddle point method. Especially the UAA gives very accurate results. Compared with the exact methods proposed in this paper the UAA requires that the corresponding saddle points have to be located for each single value under consideration.

We have also demonstrated by the numerical examples that by deploying DiffServ in a core network with STM-1 links or links with higher bit-rate and by limiting the load from the real time traffic to less than 60% it is possible to guarantee the corresponding end-to-end queueing delay to just a few milliseconds with very high probability (e.g. $1\text{-}10^{-3}$ quantile).

References

[1] RFC 2212, *Specification of Guaranteed Quality of Service*, IETF, September 1997.

[2] RFC 2598, *An Expedited Forwarding PHB*, IETF, June 1999.

[3] ITU-T, *Network Performance Objectives for IP-based Services*, ITU-T Recommendation Y.1541 (02/2002); Geneva, February 2002.

[4] Mandjes, M., Wal, K. van der, Kooij, R., Bastiaansen, H., *End-to-end delay models for interactive services on a large-scale IP network*, Seventh IFIP workshop on Performance Modelling and Evaluation of ATM Networks: IFIP ATM'99; Antwerp, Belgium; June 28-30, 1999. Paper 42.

[5] De Vleeschauwer, D., Janssen, J., Petit, G.H., *Voice over IP in Access Networks;* Seventh IFIP workshop on Performance Modelling and Evaluation of ATM Networks: IFIP ATM'99; Antwerp, Belgium, June 28-30, 1999.

[6] Takagi, H., *Queueing Analysis, Volume 1: Vacation and Priority Systems, Part 1*. Amsterdam, North-Holland, 1991.

[7] Kleinrock L., *Queueing Systems, Volume II: Computer Applications.* New York,John Wiley & Sons. 1976.

[8] Østerbø O. *Models for Calculating End-to-end Delay in Packet Networks;* ITC-18, Berlin , Germany, August 31- September 5, 2003, pp 1231-1240.

[9] Østerbø O. *Mathematical Modelling and Analysis of Communication Networks: Transient Characteristics of Traffic Processes and Models foe End-to-end Delay and Delay-jitter;* dr. philos thesis 2003, Department of Telematics, Faculty of Information Technology, Mathematics and Electrical Engineering, NTNU Trondheim Norwegian University of Science and Technology.

[10] Wong R., *Asymptotic Approximations of Integrals*, Academic Press, Inc. San Diego, CA., 1989.