Models of Network Delay

Ronan Wallace[1], Xabier Garcia Andrade[1], Pedro Kayser[1], Zhao Luo[1], Hrishav Mukherjee[1], Ruan Nunes[1], and Marc Warrior[1]

Amazon Web Services
ronanw@amazon.com

Abstract. In this paper several mathematical models for end-to-end network delay are derived, where exponential wait times at intermediate network routers are assumed. The feasibility of using these models to extract parameters related to the routers is investigated by performing closure tests using synthetic data generated from the models themselves. Data from an experimental test bed is used to compare the different models. The fits are used to estimate the difference between network probe latencies. Specifically, a difference of 4.67 ± 0.27 μs is measured between a distinct pair of measurement probes across an experimental test bed.

Keywords: IP Networks, Packet Delay

1 Introduction

One important characteristic of a communications network is the delay that is experienced in delivering a message from sender to receiver [1]. There are several distinct contributions to this delay: propagation delay, switching delay, queuing delay, transmission delay and application delay.

Propagation delay is a function of distance and the speed of light in the transmission medium. Switching delay is the time it takes intermediate routers to make their forwarding decision. Queuing delay accounts for time that messages may be waiting due to intermediate routers being busy and unable to forward messages. Transmission delay refers to the time between the first and last bytes of the packet are emitted by the router. Application specific delays account for details related to the events at sender and receiver that trigger timestamping.

In this paper we consider the queuing delay at intermediate routers to be the main driver of end-to-end latency variability. We derive probability distributions for end-to-end latency variability by assuming that the queueing delays at intermediate routers are exponentially distributed. The rate at which routers process incoming data depends on the speed of the interfaces; routers with similar/different interface speeds will have similar/different exponential rate parameters. The feasibility of extracting the rate parameters is investigated using synthetic data generated by these models. Real data from an experimental test bed is also fitted using the models.

The rest of this paper is organised as follows. Section 2 gives a short overview of queuing theory and related work; Section 3 outlines our notation; Section 4

derives a probability distribution for the sum of groups of n and m exponential random variables, where the rate parameters are the same within each group; Section 5 discusses fits to real data; and Section 6 concludes the paper.

2 Queuing Theory and Related Work

The work presented here builds on concepts from queuing theory. In this framework, queuing systems are characterised by: the probability distributions of the number of packets arriving at the server; the probability distribution of the packet service time; and the number of servers [1]. In Kendall's notation, the distributions are denoted by M, G or D depending on whether the distribution is exponential, general or deterministic. For example, the M/M/1 queue system is one where: the inter-arrival time of packets is exponential (or equivalently the number of packets arriving in a time interval is Poisson); the service time of packets is exponential; and there is one server. The M/G/n queue is one where the inter-arrival time of packets is exponential; the service time of packets follows a general distribution; and there are n servers in parallel. For these systems, Little's Theorem is used to express steady state quantities like the average number of packets in a queue, the average service time per packet etc [1]. In this work we are considering m G/M/1 systems in series: we are not concerned about the arrival rate, so we designate a general distribution; we are assuming that the service time is exponential; and we model each router as a single server. What we are interested in is overall delay distribution of the m G/M/1 systems in series.

The models described in this paper assume a dependence between the network delay and the number of network hops. The approach was considered before as a way to describe end-to-end internet latency data, but the assumption above was not consistent with this data and another kind of treatment was required [2] [3]. In this work we look at data where this assumption does hold; the data has been collected on an experimental test bed, which is a more controlled environment than the internet.

The main assumption made in this paper is that the queueing delays are exponentially distributed. One consequence of this is that a delay of zero has the highest probability, yet for real systems this delay is non-zero. In other work the exponential wait time is only assumed to hold in the tail of the distribution [4].

Others have modeled end-to-end delay distributions with Normal, Lognormal and Pareto distributions directly [5]. While these often obtain good results, they are not built up from the delay distributions of the underlying routers and so don't provide an interpretation at the per hop level.

In this paper we look at models of end-to-end delay across several routers. We provide explicit derivations and comparisons to real data collected on an experimental test bed.

3 Notation and Convolution

Random variables are represented using capital letters like X, Y and Z. Lower cases x, y, z refer to particular values of the random variable. Probability density functions of random variables are represented by lower case letters such as f, g, and h. It should be obvious from the context when lower cases letters are used to represent functions or values of random variables.

Suppose we have two random variables X and Y. Suppose that their distributions are known. The distribution of their sum $Z = X + Y$ is given by the convolution. In the continuous case it is given by the integral

$$h(Z = z) = f * g(Z = z) \tag{1}$$

$$= \int_{-\infty}^{+\infty} f(X = z - t)g(Y = t)dt \tag{2}$$

The limits at infinity can be reduced to 0, t for functions supported on $[0, +\infty)$.

4 Combination of gamma random variables

We now consider the end-to-end delay distributions that can occur when we assume the router delay distribution to be exponential. There are a number of different scenarios: all of the exponential rate parameters are the same, leading to a gamma distribution (this is a well-known result); all of the exponential rate parameters are different, leading to a hypoexponential distribution; all but one of the exponential rate parameters are the same; and two sets of exponential rate parameters, corresponding to the situation where there are two sets of several network routers, where the queue depths and processing rates are similar within a set but different between sets. To determine the result we convolve two gamma distribution of the form:

$$X \sim g(x) = \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-\beta_1 x} \tag{3}$$

Implementing the convolution

$$h(z) = g * f(z) = \int_0^z g(t)f(z - t)dt \tag{4}$$

$$= \int_0^z \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} t^{\alpha_1 - 1} e^{-\beta_1 t} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} (z - t)^{\alpha_2 - 1} e^{-\beta_2 (z - t)} dt \tag{5}$$

$$= \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} e^{-\beta_2 z} \int_0^z t^{\alpha_1 - 1} e^{(\beta_2 - \beta_1)t} (z - t)^{\alpha_2 - 1} dt \tag{6}$$

Now make the subsititutions $u = t/z$ and $dt = zdu$.

$$h(z) = \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)}\frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)}e^{-\beta_2 z}\int_0^1 (uz)^{\alpha_1-1}e^{(\beta_2-\beta_1)zu}(z-zu)^{\alpha_2-1}zdu \tag{7}$$

$$= \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)}\frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)}e^{-\beta_2 z}z^{\alpha_1-1}z^{\alpha_2}\int_0^1 u^{\alpha_1-1}e^{[(\beta_2-\beta_1)z]u}(1-u)^{\alpha_2-1}du \tag{8}$$

$$= \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)}\frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}e^{-\beta_2 z}z^{\alpha_1-1}z^{\alpha_2}\int_0^1 e^{[(\beta_2-\beta_1)z]u}u^{\alpha_1-1}(1-u)^{\alpha_1+\alpha_2-\alpha_1-1}du \tag{9}$$

$$= \frac{\beta_1^{\alpha_1}\beta_2^{\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}e^{-\beta_2 z}z^{\alpha_1+\alpha_2-1}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 e^{[(\beta_2-\beta_1)z]u}u^{\alpha_1-1}(1-u)^{\alpha_1+\alpha_2-\alpha_1-1}du \tag{10}$$

$$= \frac{\beta_1^{\alpha_1}\beta_2^{\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}e^{-\beta_2 z}z^{\alpha_1+\alpha_2-1}{}_1F_1(\alpha_1;\alpha_1+\alpha_2;(\beta_2-\beta_1)z) \tag{11}$$

This shows that the convolution of two Gamma random variables can be written in terms of the confluent hypergeometric function of the 1^{st} kind, ${}_1F_1$. Note that if $\beta_1 = \beta 2 = \beta$ then this reduces to $Z \sim Gamma(z; \alpha_1 + \alpha_2, \beta)$.

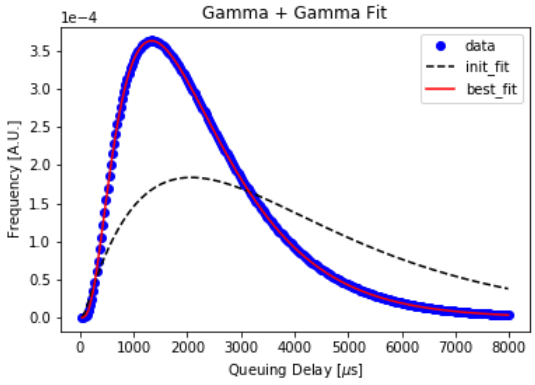

Fig. 1. The fit of the Gamma + Gamma distribution to one hundred million samples generated from the same distribution.

Synthetic data sampled according to a Gamma + Gamma distribition is shown in Figure 1, where the parameters of the underlying exponential distributions are given in Table 1. The first Gamma is generated from 3 independent and identically distributed exponential functions, all with Rate parameter 0.01. The second Gamma is generated from 2 independent and identically distributed exponential functions, all with Rate parameter 0.001. We then fit the same func-

Table 1. Parameters of the Gamma + Gamma distribution. Note that the Mean Delay and Rate are reciprocal.

Parameter	Mean Delay [μs]	Rate [μs]$^{-1}$	Initial	Best Fit
α_1	-	-	3	3
β_1	100	0.01	0.04	0.01
α_2	-	-	2	2
β_2	1000	0.001	0.0005	0.001

tion to the generated data and see that the Best Fit parameters are the same as the true Rate parameters. Thus, the fitting approach satisfies a closure test.

5 Fits to Real Data

5.1 Data Set

Round trip time latency data between servers attached to an experimental test bed was collected on the 1st July 2022 between 2000 and 2100 UTC. Two specific network probes (we shall call them P1 and P2) were configured to collect this data. The two probes have a common source (sharing server and switch) but different destination racks; the network path differs in length by 8 network hops in the complete round trip. Both probes are generated by agents from servers of the same model type. The top-of-rack switches and network routers are also the same type.

There are thus three entity types for this probe - the server, the switch and the router. We can expect the hypoexponential, Gamma + Gamma or Gamma + Exp fits to perform best depending on how different the delays are for these different entity types.

Packets are sent at a rate of 1 packet per second so each of the data sets has 3600 samples from the hour of data taking. The latency measurement is obtained by combining four kernel timestamps, two from each of the sending and receiving servers.

The resulting data is summarised in Figure 2. Both probe latency distributions have four peaks, 1 main peak and 3 smaller peaks. The main peak positions of the distributions are shifted relative to one another. The peak positions of the 3 smaller distributions are not shifted and are understood to be caused by batching/queuing of packets at the server network interface. Understanding the distribution of the main peak is the focus in this note.

5.2 Fits to Data

Before fitting, the minimum value is subtracted from all values. The effect of this is to subtract out the constant delays from each latency measurement, such as the propagation delay. Rather than fitting the latency we are then fitting the variable delay. This simplifies the form of the fitting function.

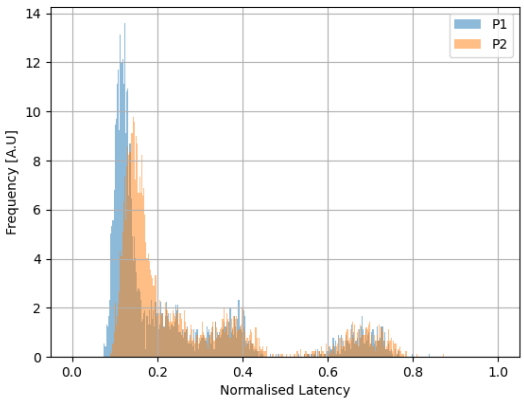

Fig. 2. The histograms of the two latency data sets used in the study. Note that the latency values have been mapped to the dimensionless interval [0,1].

We use the lmfit package with least square and Nelder-Mead methods.

First the data is fit with a combination of 4 Gaussian distributions, which serves as the baseline model. This model is not suitable for a network delay because it is not strictly positive. The Gaussian parameters extracted for the 3 smaller peaks are used for other fits whereas the Gaussian parameters for the main peak are swapped with the parameters from the models, like the ones described in Section 4.

Fits are compared using the reduced χ^2 test statistic. This is the usual χ^2 test statistic divided by the number of degrees of freedom, which in this case is the number of bins of the histogram minus the number of parameters in the model.

Table 2. Results of the fitting methods on the two different data sets.

Method	Data Set	χ^2	Parameters	$\chi^2/ndof$
4 Gaussian	P1	0.002	12	1.99e-06
Hypoexponential + 3 Gaussian	P1	0.011	11	1.14e-05
Gamma + Gamma + 3 Gaussian	P1	0.002	13	2.17e-06
Gamma + Exponential + 3 Gaussian	P1	0.006	13	6.38e-06
4 Gaussian	P2	0.002	12	1.85e-06
Hypoexponential + 3 Gaussian	P2	0.008	11	8.59e-06
Gamma + Gamma + 3 Gaussian	P2	0.002	13	1.98e-06
Gamma + Exponential + 3 Gaussian	P2	0.004	13	3.82e-06

The results of the fits are summarised in Table 2. The best performing model of those proposed in this note is the Gamma + Gamma + 3 Gaussian model.

5.3 Estimating the Latency Shift

Each Gaussian from the 4 Gaussian model has 3 parameters: a mean, a width and a normalisation. The 4 means correspond to the location of the peaks and can be used to quantify the offset between the P1 and P2 data sets.

Table 3. 4 Gaussian peak positions and differences between P1 and P2 data.

Peak	P2 Mean	P1 Mean	Difference	Relative Uncertainty
Main	21.68 ± 0.22	17.01 ± 0.10	4.67 ± 0.24	5 %
Small 1	55.62 ± 1.98	54.03 ± 1.13	1.59 ± 2.28	143 %
Small 2	115.00 ± 1.28	118.25 ± 1.25	-3.25 ± 1.79	55 %
Small 3	240.00 ± 1.69	243.00 ± 1.75	-3.00 ± 2.43	81 %

Uncertaintaines are propagated to the difference by adding in quadrature, which amounts to assuming the uncertainties on the parameters are independent. This is reasonable since the data sets are independent. There is a 4.67±0.24 microsecond difference between the P1 and P2 main peak positions. The differences between the small peak positions are not statistically significant given the uncertainites.

6 Conclusion

We have derived a number of models that can be used to explain the shapes of end-to-end delay distributions. Our derivations assumed exponential wait times at each router. Using these models we have tested the feasibility of extracting parameters related to the routers by performing closure tests using synthetic data generated from the models themselves.

The best fit to real data is with the baseline model of 4 Gaussians. Strictly speaking this model is not a correct description of network delay because the domain of the Gaussian function is the set of real numbers, including the negative ones. The models we have derived are correct in this sense as their domain is the set of positive real numbers. The second best fit is from the Gamma + Gamma + 3 Gaussian model.

We demonstrated a nice application of the fit results where the difference of latency between different probes can be estimated. In our case the measurement is $4.67 \pm 0.27\ \mu s$. Since the only difference between these probes are the 8 networking hops, the difference is a pure measure of network performance. Contrast this with absolute latency measures that can include non-network components.

We have identified potential future work that includes: empirically measuring the router delay with timestamps for packets in and packets out; repeating the

data collection while controlling for the queue lengths in the routers; repeating the data collection with more diverse sets of paths, different path lengths, and paths experiencing congestion.

References

1. Bertsekas, D.P. and Gallager, R.G.: *Data Networks*, Prentice Hall, 1992.
2. Hooghiemstra, G. and Van Miegham, P.: *Delay Distributions on Fixed Internet Paths*, 2001.
3. Van Mieghem, P.: *A Lower Bound for the End-To-End Delay in Networks: Application to Voice Over IP*, In Proceedings IEEE GLOBECOM, 1998.
4. Abate, J. and Whitt, W.: *Exponential Approximations for Tail Probabilities in Queues, I: Waiting Times*, Operations Research 43, 885-901, 1995.
5. Zhang, W. and He, J.: *Statistical Modeling and Correlation Analysis of End-to-End Delay in Wide Area Networks*, Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Volume 3, 968-973, 2007.