Rules of Thumb for Response Time Percentiles

Drs. Michael Salsburg and Jeffrey Buzen
Independent Consultants

Whether externally mandated or internally tracked, the enterprise relies on governance of application service response time objectives. In many cases, achieving service objectives in terms of the average response time may not deliver an experience that delights the consumer. There may be stringent limits to the "tail" of response times, even when the mean response time appears reasonable. Service providers want to achieve the promised objectives while avoiding over-provisioning. This paper explores rules of thumb (ROT) that can be applied to estimate 90th or 95th percentiles for service response times, based on the measured or predicted mean. The risk assessment behind these recommendations is described in this paper. Various types of queueing networks were modeled and analyzed. Even though classical queueing models rely on strict assumptions (which are rarely met in the real world), it was found that the classical M/M/1 model provided a low-risk rule of thumb to estimate percentiles for many queueing systems. Another function, Martin's Estimate, was evaluated for tighter accuracy and broader application.

Introduction

Today, a web site lives or dies on its ability to delight the end user. When a company hires an "outside" contractor to create a new web site, service level objectives are critical. Average response times are expected to be under one second, measured at key geographic locations around the world. What makes this more interesting is that quality objectives are being articulated in terms of 90% or 95% of all responses, as opposed to the average.

When a web site is being built and analyzed, in most cases, the measurement and analysis is still focused on the average response time. During the development of the web site, averages are collected, analyzed and various options are proposed and then modeled to determine how to meet the service level objectives. Modeling packages are excellent at predicting average response times, based on previous measurements, and then projecting the improvements gained by upgrading physical hardware or optimization of software.

If percentiles are included within the formal service level objectives, the model is expected to predict the necessary statistics. Sometimes, a simple rule of thumb is applied, based on classical queueing theory. But in most cases, the actual system under development does not meet all of the criteria required to apply these queueing formulae.

This paper explores the risk involved when using the classic formulae. As the authors delved deeper and deeper into the question of risk, a number of different queueing systems were explored and, as results were evaluated, still more were hypothesized and modeled.

Evaluations of these systems were implemented using a discrete event simulator (see [3]), as opposed to amassing the necessary hardware. By using a simulator, classic queueing workloads can be modeled as well as more general workloads that cannot be addressed by classical formulae. Simulations can model various arrival processes, service time distributions and network topologies. At each event, the model can be instrumented to collect additional statistics to gain insights into how the queueing process is proceeding. In many cases, these statistics would not be available from software running on physical systems.

The following sections will provide definitions of the models and assumptions. This will be followed by a number of different queueing topologies and parameters used to explore the robustness of proposed rules of thumb (ROT).

The Classic M/M/1 queueing model

To start, consider a simple queue, such as what you experience when you walk into a bank. Think of one teller with one line in which customers wait. Imagine customers arriving in a process where the times between arrivals are exponentially distributed. This arrival process is called the Poisson process. When the customer gets to the teller, the average service time is also exponentially distributed. This is the classic starting point in modeling queueing networks. The exponential distribution has a specific property of being "memoryless" in that the probability of the current arrival time is unconditionally independent of any past history of arrivals. The same memoryless quality applies to exponentially distributed service times. Think of the "M" as "memoryless". The exponential distribution is the only continuous distribution that has this memoryless quality.

Although this may appear to be an odd distribution, it is surprisingly prevalent as a natural phenomenon, such as service requests of CPUs or arrival processes for telephone calls.

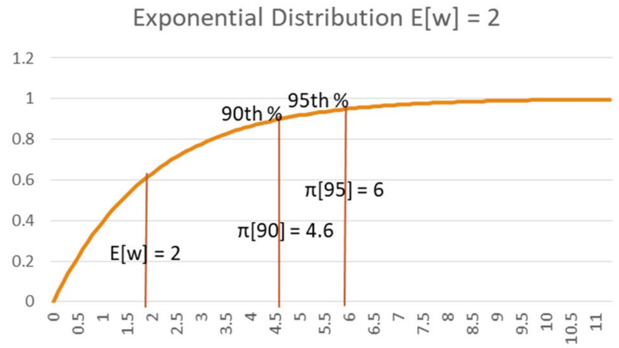
It can be shown that the wait time, W, from entrance into the queue to exit from the server is also exponentially distributed. In the following equation, F[w] is the cumulative distribution function for W and E[w] is the mean.

$$\mathrm{F}(w) = 1 - \mathrm{e}^{-\alpha w}$$

$$\mathrm{E}[w] = \frac{1}{\alpha}$$

There are other statistics worth understanding, such as the standard deviation of the wait time, as well as the 90th and 95th percentile for wait times. For an M/M/1 system, the standard deviation, σ, is equal to E[w]. The percentiles are notated as $\pi[r]$, where ***r*** is the percentile.

The graphic below shows an example of a cumulative exponential distribution with E[w] = 2. It shows the percentiles for r = 90 and 95.



Key Statistics for the Exponential Distribution

This paper defines 90%R and 95%R as the ratio of the 90th and 95th percentiles over the mean. Ratios are very useful in keeping statistics properly scaled. So, for the 90th percentile,

90%R is defined as $\frac{\pi[90]}{E[w]}$

In general, it can be shown that

$$\mathrm{P}\big[\mathrm{X} \le \pi[90]\big] = .9 \Rightarrow$$

$$1 - \mathrm{e}^{-\alpha\pi[90]} = 0.9 \Rightarrow$$

$$\mathrm{e}^{-\alpha\pi[90]} = 0.1 \Rightarrow$$

$$\pi[90] = -\frac{\ln(0.1)}{\alpha} = \mathrm{E[X]}\ln(10) = 2.3\mathrm{E[X]} \Rightarrow$$

90%R ≈ 2.3

Similarly, 95%R ≈ 3.0

To see how these results can be applied in practice, suppose that 90% of the requests processed by a particular app are required to have a response time that is no more than 4 seconds. Since 90%R = 2.3, this is equivalent to requiring that the average response time be less than 1.74 seconds (4/2.3 = 1.74).

Now suppose management decides that it is undesirable for 10% of all requests to have a response time of more than 4 seconds. It is recommended instead that no more than 5% of requests have response times that exceed 4 seconds. What is the impact of this new requirement on average response time?

Since 95%R = 3.0, requiring that 95% of all requests have a response time of no more than 4 seconds is equivalent to requiring that average response time be less than 1.33 seconds (4/3 = 1.33). Thus, it is necessary to reduce average response time by 0.41 seconds to achieve this more stringent service level objective.

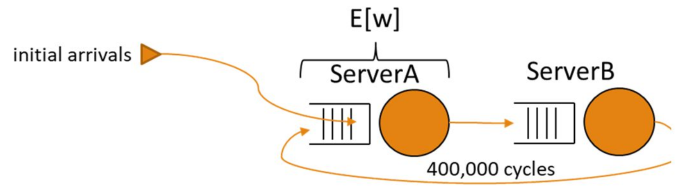
These analyses are all based on the assumption that R90% = 2.3 and R95% = 3. As discussed above, these equations are valid for M/M/1 queues. However, real world systems are unlikely to satisfy the mathematical assumptions of the M/M/1 model with prefect accuracy. This makes it important to understand what happens to the values of R90% and R95% when the assumptions of the M/M/1 model are modified. In particular, how risky is it to assume that R90% = 2.3 and R95% = 3 in cases where M/M/1 assumptions are no longer satisfied?

The next sections of this paper explore this question through a series of simulation models.

For each model, an experimental matrix of parameters were explored to gain insight into various queueing statistics, including mean, standard deviation and percentiles. As will be seen, most of these models are more complex than a single M/M/1 server, since they are intended to address models that are encountered by practitioners in computer performance management.

The Cyclic Queueing Model

This first model reflects a closed queueing system, similar to what one would see with a workload that cycles between CPU processing and storage. This type of queueing model has also been used to model batch workloads where there are a relatively small, finite number of threads that can execute at the same time.



Cyclic Model

In this model, a preset number of "customers" are defined as well as the average service time for Server A. In the simulation, the customers arrive initially at Server A as a Poisson process until the pre-set Customer Count have arrived. Once this number of customers are in the system, they cycle from A to B and back to A. The simulation was run for 400,000 cycles to achieve a high level of accuracy for E[w], which is the response time for customers arriving at the queue for Server A and completing the service. After the initial arrivals, customers arrive at Server A when they exit from Server B. Since random numbers are generated in simulation models, the results are again random numbers, with a finite mean and standard deviation. Accuracy for simulated statistics, such as E[w], is achieved by simulating a sufficient number of samples so that there is a high "confidence" in the mean value. This is described below.

The number of simulated samples are determined as sufficient to yield an "accurate" mean value if, with a 98% confidence level, the mean is bounded by a confidence interval of width ∈, $\left(u-\frac{\in}{2},u+\frac{\in}{2}\right)$, where $\in<\frac{\mu}{10}$. Therefore, there is a 2% chance that the mean does **not** lie within this interval.

The experimental matrix for the cyclic model considered all combinations of the following pre-set numbers of customers and average service times:

- Server A Average Service Time = 1.0
- Customer Count – 2, 4, 8, 16
- Server B Average Service Times – 1, 1.25, 1.5, 1.75, 2, 4, 8 units.

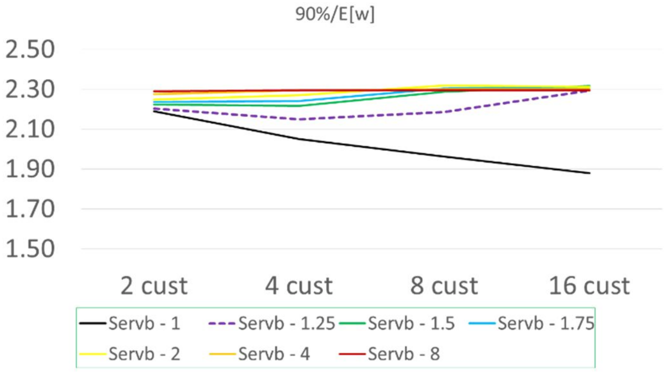
At each arrival to a server, the service time is chosen as a random, exponentially distributed value with an average as shown above.

The next table shows values calculated for 90%R. Note that in the classic M/M/1 model, the rule of thumb would be exactly 2.3 as proved above.

		Average Service Time – ServerB						
		1	**1.25**	**1.5**	**1.75**	**2**	**4**	**8**
Customers	**2**	2.19	2.20	2.22	2.24	2.25	2.28	2.29
	4	2.05	2.15	2.22	2.24	2.27	2.29	2.30
	8	1.96	2.18	2.29	2.30	2.32	2.30	2.29
	16	1.88	2.29	2.32	2.31	2.31	2.30	2.29

Cyclic Model 90%R Values

It should come as no surprise that 90%R does not match our rule of thumb in all cases. As a matter of fact, the results are nonlinear as we keep the service time constant and compare across the various customer counts. For example, look at the nonlinearity of 1.25 second service times across the four different customer counts. The following graph shows this in more clarity.



Nonlinear Effects in Cyclic Model

Going back to the objective of this exercise, if we consider the risk of using the M/M/1 ROT, we could say that, for this model, the rule of thumb is not risky. In most cases, the measured values are less than the 2.30 value and we can safely say that the 90%R has a low risk of exceeding 2.30.

So why does this model deviate at all? The reason is that the cyclic model does not meet all of the assumptions necessary for the M/M/1 model. Specifically, consider the 2 customer model. After 2 customers are in the system, there are no more arrivals. This violates the "memoryless" property.

As we considered other queueing systems, we started to explore models where service times were not exponentially distributed, in which case this would be called an "M/G/1" model, where G stands for "general" distributions. This includes but is not restricted to the exponential distribution. For example, service times for spinning disks are uniformly distributed. Service times for line transmissions in a network can be bi-modal, and so on.

What we found is that there was a high correlation between the 90%R, 95%R values and the standard deviation of the service time distribution. We again chose to work with a dimensionless ratio. In this case, it is the coefficient of variation, CV. For our study, CV is the ratio of the standard deviation over the mean of the waiting times in the queueing system.

$$\boldsymbol{CV} = \frac{\boldsymbol{\sigma}}{\boldsymbol{E}[\boldsymbol{w}]}$$

In general, when the CV < 1, the risk of using our M/M/1 rule of thumb is low but there is a risk of over-configuring the system. When CV > 1, the risk is high that we have underestimated the values of 90%R and 95R using our rule of thumb. If the CV=1, it's "just right".

Following a suggestion from Allen[1], we explored the accuracy of using "Martin's Estimate", as opposed to the M/M/1 rule of thumb. The derivation of this estimate is discussed in the appendix, since the discovery of the estimate is not germane to this discussion.

Martin's Estimates are:

$$\pi[90] \approx W + 1.3\sigma_w \Rightarrow 90\%R = 1 + 1.3CV$$

$$\pi[95] \approx W + 2.0\sigma_w \Rightarrow \ 95\%R = 1 + 2.0CV$$

Note that, when waiting times are exponentially distributed, then CV = 1 and we get our original M/M/1 ROT.

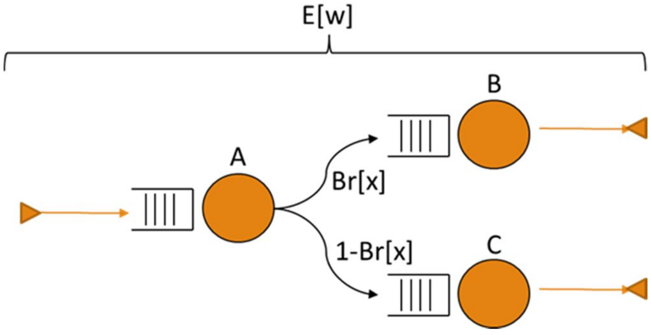
The following table shows the Cyclic Model results, with a comparison of the M/M/1 ROT to Martin's Estimates, using the standard deviation that was calculated within the simulation for the 95%R.

			M/M/1	Measured	Martin's Est
Servb	Customers	CV	E[95%]/E[w]	95%/E[w]	E[95%]/E[w]
1.00	2.00	0.88	3.00	2.75	2.76
1.00	4.00	0.77	3.00	2.49	2.55
1.00	8.00	0.70	3.00	2.30	2.40
1.00	16.00	0.64	3.00	2.13	2.28
1.25	2.00	0.90	3.00	2.78	2.80
1.25	4.00	0.83	3.00	2.64	2.67
1.25	8.00	0.83	3.00	2.65	2.66
1.25	16.00	0.91	3.00	2.83	2.82
1.50	2.00	0.92	3.00	2.82	2.83
1.50	4.00	0.87	3.00	2.76	2.76
1.50	8.00	0.92	3.00	2.85	2.84
1.50	16.00	0.98	3.00	3.05	2.94

Comparison of M/M/1 95%R and Martin's Estimate
For Cyclic Model

An Open Queueing Model with Branching

Moving away from a closed queueing system, we explored some simple open queueing systems.



Branching Model

This first model addressed an experimental matrix where various arrival rates and branching probabilities. The following table shows the 95%R values. In this model, all arrivals requested random service times, exponentially distributed, with a mean of one second. The arrival process was Poisson. The measured wait time, E[w], for each request is the time that elapses from its arrival at Server A's queue to the completion of service at either server B or C.

		Branching Probabilities Br[x]	
		0.333	0.5
Arrivals / Sec	0.1	2.37	2.37
	0.2	2.37	2.38
	0.4	2.39	2.39
	0.8	2.54	2.53
	0.9	2.70	2.79

95%R Values for Branching Model

This next table compared the M/M/1 95%R with Martin's Estimate, along with showing the percent error in these two approaches.

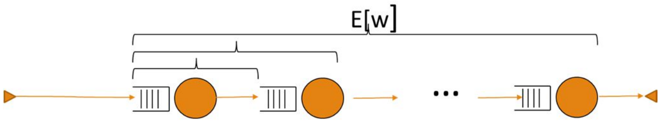
						% Error	
Br[x]	Arrival Rate	CV	M/M/1	Measure	Martin's Estimate	M/M/1	Martin's Estimate
0.33	0.10	0.71	3.00	2.37	2.41	0.26	0.02
0.33	0.20	0.71	3.00	2.37	2.42	0.26	0.02
0.33	0.40	0.72	3.00	2.39	2.44	0.25	0.02
0.33	0.80	0.78	3.00	2.54	2.55	0.18	0.01
0.33	0.90	0.84	3.00	2.70	2.68	0.11	-0.01
0.50	0.10	0.70	3.00	2.37	2.41	0.26	0.02
0.50	0.20	0.71	3.00	2.38	2.42	0.26	0.02
0.50	0.40	0.72	3.00	2.39	2.43	0.25	0.02
0.50	0.80	0.77	3.00	2.53	2.54	0.19	0.01
0.50	0.90	0.90	3.00	2.79	2.79	0.07	0.00

Comparison of M/M/1 95%R and Martin's Estimate
For Branching Model

The Tandem Servers Model

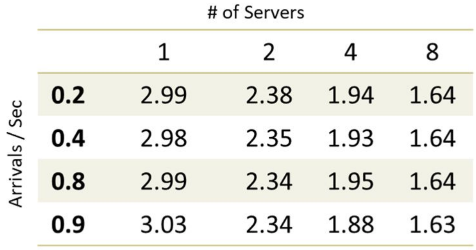
Note in the previous model that the measured wait times show the total wait from entrance to A's queueing to the exit from either B or C. The distribution for wait times through two servers with a Poisson arrival process and exponentially distributed service times is the *Erlang-2* distribution. Similarly, there are *Erlang-k* distributions, for any *k* where *k* is an integer.

From this set of experiments, we went on to explore the impact of the number of servers in Tandem. Eight servers in Tandem, where the arrival process at the first server is Poisson and each service request is exponentially distributed, would be expected to have wait times with an *Erlang-8* distribution.



Tandem Model

The following shows 95%R for the experimental matrix where arrivals/sec and number of servers were modeled in all combinations.

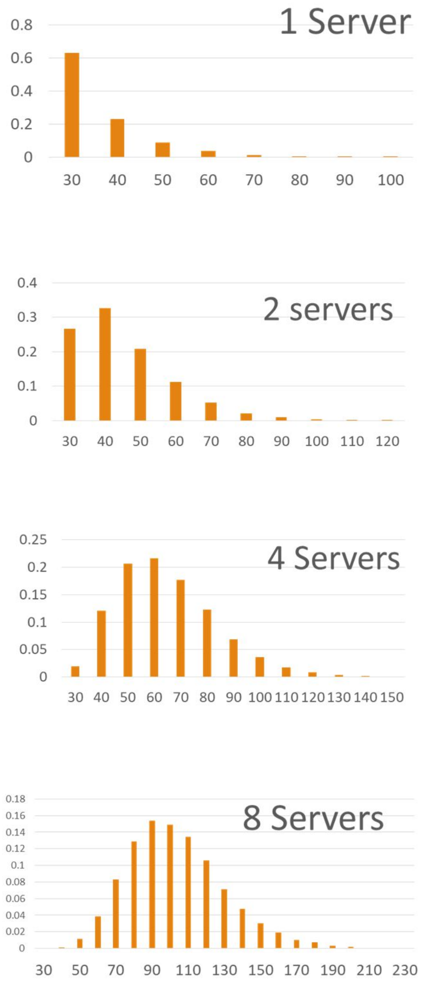


		# of Servers			
		1	2	4	8
Arrivals / Sec	**0.2**	2.99	2.38	1.94	1.64
	0.4	2.98	2.35	1.93	1.64
	0.8	2.99	2.34	1.95	1.64
	0.9	3.03	2.34	1.88	1.63

95%R Values for Tandem Model

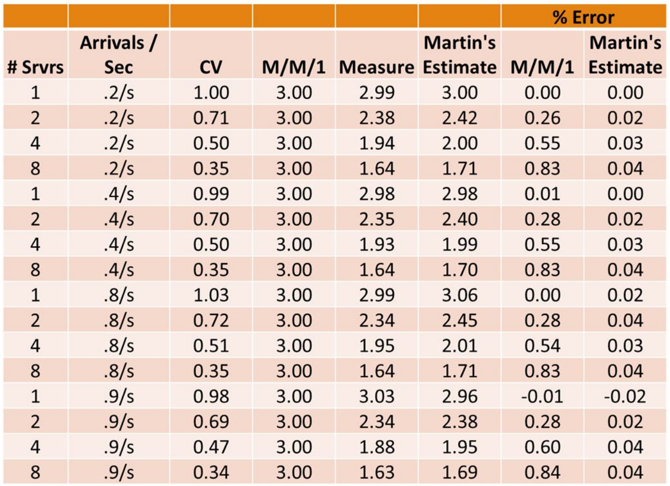
Notice how the values get smaller when more and more servers are traversed by an arrival. It can be shown, using the central limit theorem, that as *k* increases, in the *Erlang-k* distribution, the distribution approaches the normal distribution.

Here are the distributions taken from the 400,000 simulated samples with arrival rate of 0.9/second.



Probability Density Functions for Tandem Models

The following table compares the M/M/1 ROT to Martin's Estimate for the results of the experiments.

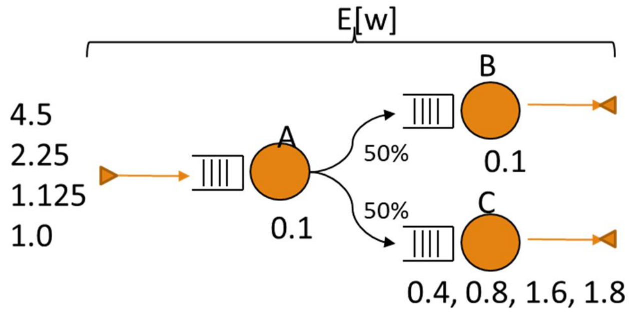


						% Error	
# Srvrs	Arrivals / Sec	CV	M/M/1	Measure	Martin's Estimate	M/M/1	Martin's Estimate
1	.2/s	1.00	3.00	2.99	3.00	0.00	0.00
2	.2/s	0.71	3.00	2.38	2.42	0.26	0.02
4	.2/s	0.50	3.00	1.94	2.00	0.55	0.03
8	.2/s	0.35	3.00	1.64	1.71	0.83	0.04
1	.4/s	0.99	3.00	2.98	2.98	0.01	0.00
2	.4/s	0.70	3.00	2.35	2.40	0.28	0.02
4	.4/s	0.50	3.00	1.93	1.99	0.55	0.03
8	.4/s	0.35	3.00	1.64	1.70	0.83	0.04
1	.8/s	1.03	3.00	2.99	3.06	0.00	0.02
2	.8/s	0.72	3.00	2.34	2.45	0.28	0.04
4	.8/s	0.51	3.00	1.95	2.01	0.54	0.03
8	.8/s	0.35	3.00	1.64	1.71	0.83	0.04
1	.9/s	0.98	3.00	3.03	2.96	-0.01	-0.02
2	.9/s	0.69	3.00	2.34	2.38	0.28	0.02
4	.9/s	0.47	3.00	1.88	1.95	0.60	0.04
8	.9/s	0.34	3.00	1.63	1.69	0.84	0.04

Comparison of M/M/1 95%R and Martin's Estimate
For Tandem Model

The Asymmetric Model

In this model, we explored a queueing system where the CV for the wait times is greater than 1. In his book, Martin[2] is very specific about qualifying his estimates for systems with CV less than or equal to 1. In [2], he admonishes "if the standard deviation is higher than the mean, it is often an indication that some change should be made in the design."



Asymmetric Model

In this model, we carefully keep the utilization of Server C at 90% busy, thus inducing significant queueing times. The arrival process at Server A is Poisson and all service times are exponentially distributed. As different arrival rates varied, so did the mean values for the service times at Server C to maintain the same utilization at that server. The Average service time at Server A and Server B were kept constant at 0.1 seconds.

The following table shows the 95%R as well as the "measured" utilizations of Server B and Server C.

Arrivals/s	ServerB	ServerC	95%R
4.5	22.6%	90.2%	**3.10**
2.25	11.2%	90.1%	**3.84**
1.125	5.6%	89.9%	**4.39**
1.0	5.0%	90.1%	**4.42**

95%R Values for Tandem Model

Note that, as the ratio $\frac{Utilization_B}{Utilization_C}$ deviates from 1, there is more deviation in the overall wait time for all customers. Since the cumulative wait times for all arrivals are under evaluation, there is a higher variance around the mean and 95%R reflects this.

As computer performance practitioners, it may appear that we are actually mixing 2 different workloads or, on the other hand, there is some sort of round-robin system that is not working properly. In an actual engagement, the practitioner may suggest researching this behavior and perhaps correcting it before committing to a service level objective.

In any case, the following table again compares the M/M/1 ROT with Martin's estimate. Even though the system appears to be a bit demented, Martin's estimate still appears to be a useful tool.

					% Error	
Arrival Rate	CV	M/M/1	Measure	Martin's Estimate	M/M/1	Martin's Estimate
4.55	1.03	3.00	3.10	3.06	-0.03	-0.01
2.27	1.37	3.00	3.84	3.73	-0.22	-0.03
1.12	1.62	3.00	4.39	4.24	-0.32	-0.04
1.00	1.71	3.00	4.42	4.42	-0.32	0.00

Comparison of M/M/1 95%R and Martin's Estimate For Tandem Model

Conclusion

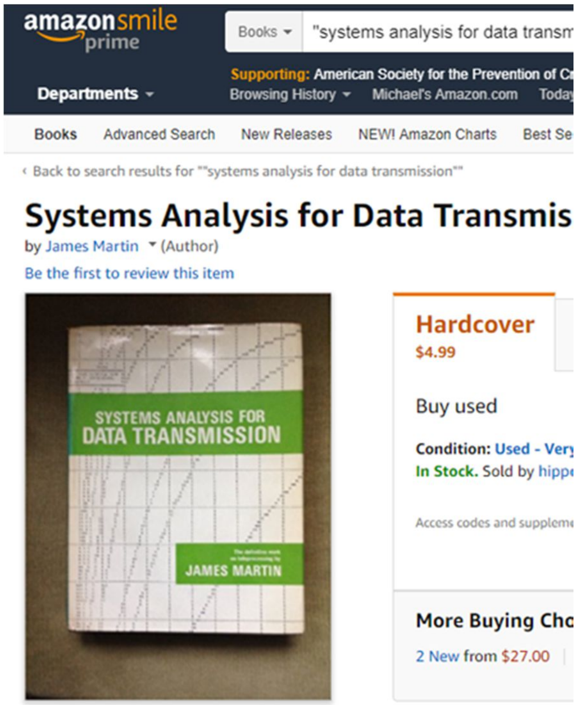
For many computer performance and capacity planning professionals, service level objectives in terms of percentiles have been considered "risky" due to the fact that percentiles are often estimated using the M/M/1 queueing model. Furthermore, we know that there are many queueing systems that do not meet the necessary assumptions for this model.

The authors explored many types of queueing systems, including closed network cyclic models as well as branching, tandem server and asymmetric models. The 90th and 95th percentile estimates, based on the mean value, are often quite conservative and therefore provide a low risk approach.

More accuracy can be gained by including another statistic, standard deviation. The two parameter model, originally proposed by Martin, ensures greater accuracy and also can be useful in indicating situations where the standard deviation is larger than the mean, in which case, the M/M/1 approach is extremely risky.

Appendix A – Whispers from Planet Gamma

During the course of investigation, the authors have considered the rationale behind Martin’s estimates. At first, Allen’s reference in [1] seemed like a simple way to understand the derivation of these estimates. In fact, the actual physical book was readily available at a reasonable cost.



Supporting: American Society for the Prevention of C

‹ Back to search results for ""systems analysis for data transmission""

Systems Analysis for Data Transmis

by James Martin (Author)

Be the first to review this item

Hardcover
$4.99

Buy used

Condition: Used - Very
In Stock. Sold by hippe

Access codes and suppleme

More Buying Cho
2 New from $27.00

The book is 900 pages of excellent suggestions, derivations, proofs and plain common sense. But the actual derivation of the estimates cannot be found. In his second edition, Allen refers to the estimates a number of times. On page 311:

“There is no general formula to calculate percentile values of w for the M/G/1 queueing system, but James Martin[45] gives estimates”.

On page 341 of [1], Allen states:

…it gives Martin’s estimate (see Martin[45], page 461]). The formula was developed by pattern recognition, not by any formal proof.

Looking on page 461, we see the following equation:

The mean waiting time, $E(t_w)$ for a single server queue with nonexponential service times is related to the mean waiting time for exponential service times, $E\left(t_{wexp}\right)$ by the following equation

$$E(t_w) = \frac{E(t_{wexp})}{2}\left\{1 + \left[\frac{\sigma_{ts}}{E(t_s)}\right]^2\right\}$$

He also points out that, in his experience, the upper and lower bounds are 1 and zero respectively, are well known distributions. When standard deviation equals the mean, it is the exponential distribution. When of standard deviation of 0, it is a constant distribution.

On page 438 of [2], Martin describes how an engineer can answer percentile questions, such as:

Will 90 percent of the transactions have a response time of 3 seconds or less?”

His recommendation is to find the mean wait time and its standard deviation. Then select a gamma distribution where the shape parameter is:

$$R = \left[\frac{E(w)}{\sigma}\right]^2$$

The Gamma cumulative distribution function is then shown as:

$$P(t \le T) = \frac{\int_0^T \left[\frac{RT}{E(t)}\right]^{R-1} e^{-RT/E(t)} \frac{R}{E(t)} dt}{\int_0^\infty \left[\frac{RT}{E(t)}\right]^{R-1} e^{-RT/E(t)} \frac{R}{E(t)} dt}$$

This function is not something to be trifled with, but it offers a solid foundation for Martin’s Estimate. Note that R is the inverse of CV^2. When the standard deviation is equal to the mean, R=1. As the standard deviation decreases to 0, R approaches ꝏ.

To determine values for probabilities of $t \le T$, Martin provides tables, generated using simulations of the cumulative distribution function, where the engineer can start with the value for R and then look up the ratio of T/t in the table. Note that this ratio is the same as $\pi[90]/E[w]$ and $\pi[95]/E[w]$ in this paper.

Obviously, Martin’s estimate simplifies things considerably. But there is no formal proof or even a hint of a proof between references [1] and [2]. It was as if these two authors discussed this at a conference in detail but chose to wink knowingly at each other across their literature.

But the Gamma function ties this all together in the following way. A subset of the Gamma functions are the Erlang-k distributions where R in the previous equations is integer valued and R=k in the Erlang-k distribution. The exponential distribution is a gamma function where R = 1. The constant distribution function is a gamma function where R = ꝏ.

Using the Central Limit theorem, it can be shown that, for the Erlang-k distribution, as *k* grows larger, the Erlang-k distribution approaches the normal distribution. Note the distributions of the Tandem Server models and how the density functions appear to be approaching a normal distribution.

So, within the Gamma family, we have proved that

Martin’s estimate holds for the exponential distribution with R=1. Also, for the standard normal distribution, where R can be infinite, we know that the 90th percentile is 1.28 standard deviations from the mean and the 95th percentile is 1.968 standard deviations from the mean. Therefore, the same estimate applies over this whole family of service time distributions. Although we cannot be certain about distributions where R < 1 (i.e., CV > 1), our research so far has shown that the estimates continue to be fairly accurate in all of our models.

References

[1] Arnold O. Allen. 1990. Probability, Statistics and Queueing Theory with Computer Science Applications, Second Edition. Academic Press. Pgs 123-127, 311-312 341.

[2] James Martin. 1972 Systems Analysis for Data Transmission. Prentice Hall Series on Automatic Computation. Pgs 392, 413-481.

[3] User's Guide: CSIM20 Simulation Engine Mesquite Software, Inc.