

# Sensitivity in Retrial Queue Model

Rui Kang\*      Andrew M. Ross

September 27, 2006

## Abstract

Retrial happens often in call centers and modem banks. If customers are blocked, a large proportion of them would like to try again after a short time. This is called the retrial queue model. We studied the sensitivity of retrial and service time distributions in multiserver retrial queue model. Phase-type service and retrial distribution beyond the mean are used in this paper. We compared the error of the system performance when approximating Phase-type to Exponential distribution and explained a counter-intuitive pattern that happens in the retrial time distribution discussion.

## 1 Introduction

Research of the retrial queue model started in the middle of the twentieth century. A retrial queue has the following features: when a customer can't enter service because all the servers are busy, rather than quit he will retry later with a certain probability. If no servers are idle when he retries, the customer can choose to quit or retry again according to another (different or the same as the previous) probability. The retrial queue model has been widely used in the field of telecommunication networks. The extreme (hypothetical) case where customers always retry and never give up is known as the persistent retrials case. The retrial queue model is applicable in many fields of telecommunication, from the traditional modem bank in ISP dial up service and cellular communication network, to the brand-new satellite Internet access system, where retrial happens because of limited server (modem, or wireless channel). These systems are basically multiserver queue systems with no buffer space.

---

\*Industrial and Systems Engineering Department, Lehigh University, Bethlehem, PA, USA.  
ruk2@lehigh.edu

The random time between two consecutive trials by one customer is called the re-trial time and is usually assumed to have exponential distribution due to the memoryless property. However, exponential distribution is not suitable for some random re-trial time durations. For example, in ISP dial up service, a customer can use a software to automatically redial when this call is blocked due to the capacity of modem bank. Thus, the re-trial time distribution has less variation than exponential distribution. On another side, if the customer pool redials by mixing both manual and automatical redial approach, then this re-trial time distribution is more variable than exponential distribution. In addition, exponential re-trial time distribution is not always an accurate assumption for all re-trial models, since there are memory properties in many re-trial cases, such as manually redialing an ISP access number. However, a general distribution can always be approximated by a phase type distribution [1]. We will discuss the system performance under different phase type distributions in this paper. To make the comparison more focused, we assume persistent retrials, which means all customers continue to retry after each failure to enter service with probability one and nobody quits before successfully entering service. By using only persistent retrials, we reduce our number of parameters.

Research of the re-trial queue model started in middle of 19th century. Besides the research papers published in the US, many papers about re-trial queue are published in European journals, such as Spanish and Russian journals. Falin [2] and Yang [3] respectively gave two detailed surveys about re-trial queue research which discussed the available aspects of current re-trial queue research and presented the main results. Artalejo [4, 5] listed the complete bibliography in re-trial queue research from 1990 to 1999. Because of the increase of dimension, it is more difficult to solve re-trial queue model than the queue models without re-trial. Falin and Templeton [6] introduced thorough analysis for both single and multiple server re-trial queues. For the multiple server re-trial queue, Falin and Templeton [6] gave explicit form for the system performances of the model with 2 servers where the interarrival time, service time, and re-trial time are all exponentially distributed. They didn't provide analytical solution for the re-trial queue systems with general service time and more than 2 servers. Numerical approximation approaches for "truncated" models are often used in multiple server re-trial queue model because it's comparatively easier to handle. Falin and Templeton [6] discussed three truncation approaches. Stepanov [7] created solutions by introducing four different truncating algorithms which are assumed to have different transitions around the truncating boundary. Le Gall [8] also analyzed the multiple server re-trial queue with general service time and provided an approximation approach to obtain the relationship between output system performances. Frederichs and Reisner [9] reduced the two dimension state space of multiple server re-trial queue to a state-dependent, one dimension birth-death process which simplified the working load and they numerically showed the approximation was very close to the two-dimension solution. Neuts [10] and Anisimov [11] also talked about the numerical solution approach of multiserver re-trial queue model.

The server number in a queueing system under a certain traffic load can be decided by a few approaches. One of the approximation approaches to evaluate the server number is to set  $s = \rho + z\sqrt{\rho}$ , where  $s$  is the server number,  $\rho = \lambda/\mu$  is the offered traffic load, and  $z$  is the constant from normal approximation. In practice,  $z$  is often set as 1 or 2. When the target probability of no delay is 95%,  $z$  value is about 1.96 which we can approximate as 2, and when the target probability of no delay is 92.5%,  $z$  value is about 1. This approximation method is usually called “QED” (Quality Efficiency Domain) scheme. Since Halfin and Whitt first proposed this approximation in heavy traffic limit [12], it is sometimes called “Halfin and Whitt scheme”. Grassmann in [13] described that QED scheme is the optimal scheme when traffic  $\rho$  is large and Poisson distribution can be approximated as Normal distribution. This normal approximation are also discussed by Grassmann [14] and Kolesar [15] where Kolesar demonstrated the accuracy of “QED” approximation scheme thoroughly. Below is a simple example to illustrate QED scheme. If traffic load  $\rho$  is 9, then the server number is  $s = 9 + 1 \times \sqrt{9} = 12$  when  $z = 1$ , and  $s = 9 + 2 \times \sqrt{9} = 15$  when  $z = 2$ . Traffic loads are set to 9, 16, 25 in this research. Table 1 shows the traffic, server, utilization settings we will use.

Table 1: system setting of traffic, server and utilizations

	traffic	Server Number	Utilization
$z=1$	9	12	0.75
	16	20	0.8
	25	30	0.833
$z=2$	9	15	0.6
	16	24	0.6667
	25	35	0.7143

In this paper, we will discuss the effect of general service time distribution and retrial time distribution on the system performance and we will discuss the system performances of the average waiting time in orbit and blocking probability of the new arrival. Depending on the properties of different probability distributions, system performance also changes accordingly. We are interested in the effect of variation and the third moment of the different service and retrial time distributions. We first fix the mean (first moment) of the different distributions, and then study the system performance under the different variations influence. Because of the Markovian properties of phase type distribution, we will use two-phase phase type distributions, including Erlang-2, Coxian-2 and Hyperexponential, as well as the Exponential distributions, to study the service time and retrial time. Erlang distribution has the smallest coefficient of variance ( $< 1$ ) in the above phase type distributions and HyperExponential distribution has the largest value ( $> 1$ ). The coefficient of variance values of Coxian-2 ( $< 1$ )<sup>1</sup> and Exponential distribution ( $= 1$ ) lie

---

<sup>1</sup>It is also possible to use Coxian-2 distributions with  $CV > 1$ , but here we use them only with  $CV < 1$

in the middle. If there is more variation of the retrial time, the customer waiting time in the orbit will have more fluctuation and it will affect the system performance better. Similarly, with a higher variation of the service time, the system performance of customers in service will also be influenced. Note that the service rate is set up as 4 in all the experiments of this chapter. It's because we imagine that the average service duration is 15 minutes, and our time units are hours. It coordinates with the previous assumption of a service rate of 1. In addition, the retrial rates used in this chapter are 0.2, 0.5, 1, 2, 5, 60, 180, 500 and 1000, which correspond to average retrial times of 5 hours, 2 hours, ... 20 seconds, 7.2 seconds and 3.6 seconds.

## 2 Phase Type Distribution

For Hyperexponential-2 distributions with same mean and variance, the third moment could be different according to the two respective phase probabilities and retrial rates, and the system performance varies accordingly. Table 2 lists the blocking probability and the average waiting time in orbit with different third moment but same mean and variance of the retrial rate. In table 2, all data are obtained with  $z = 1$ ,  $\rho = 9$  and server number  $s = 12$ .

Table 2: 3rd moment effect to the system performance with Hyperexponential-2 retrial distribution. retrial rate  $\alpha = 5$ ,  $\rho = 9$ , server number = 12,  $CV = 1.5$

3rd moment	blocking prob.	avg. waiting time
0.1268	0.1605	0.0775
0.127	0.1622	0.0766
0.128	0.1671	0.0735
0.13	0.1706	0.0702
0.14	0.173	0.0652
0.15	0.173	0.0632
0.2	0.172	0.0606
0.3	0.1713	0.0593
0.4	0.1709	0.0589
0.5	0.1707	0.0586
1	0.1703	0.0581
1.5	0.1701	0.058

The minimum possible third moment for Hyperexponential-2 with  $CV = 1.5$  is 0.1268. We can see from Table 2 that the average waiting time in orbit decreases continuously with the increase of the third moment, but the blocking probability increases to a peak point first and then decreases. This is our first hint that the blocking probability does not always behave monotonically as we might expect. The maximum value of the blocking

---

as an interpolation between Erlang and Exponential

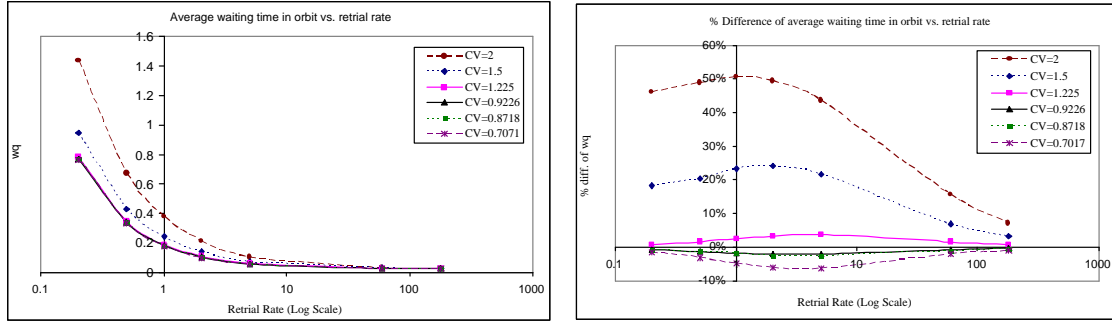


Figure 1: average waiting time in orbit for retrial time distribution with various coefficients of variance

probability has a third moment of 0.14, which is very close to the third moment of Gamma distribution. In this paper we use the Hyperexponential distribution with the same third moment as Gamma distribution, which is also called "Gamma Normalization".

### 3 Retrial time distribution

#### 3.1 Retrial Rate Effect

We first discuss the retrial time distribution. Let service times be exponentially distributed, and the retrial times be Erlang-2, Coxian-2, Exponential and Hyperexponential-2 distribution respectively with coefficients of variance from 0.7071 to 2. Retrial rate has important influence on the system performance. When retrial rate is close to 0, the retrial system can be seen as the Erlang-B loss system because customers will come back only after a near infinite time, which is like the customer is lost. In this case, the retrial time distribution does not matter because no retrials happen. But there is some difference. Erlang-B customers will leave the system forever if all servers are busy, but retrial customers with zero retrial rate will come back eventually and be served. When retrial rate goes up to infinity, the retrial system performs like the Erlang-C delay system with infinite buffers since the redialling customers have the priority to enter service, and they look like waiting in a queue (although with a randomly selected order of service). The distribution of retrial time will not have effect on the system performance in this case either. We are interested in the actual value and the percent difference of the system performance of various phase type distributions from the standard exponential distribution. We use  $z = 1$ , traffic  $\rho = 9$  and server number  $s = 12$  in this subsection. In addition, all experiments in this paper use the same mean of service rate ( $\mu = 4$ ).

Figure 1 shows the actual value and the percent difference from exponential retrial time

of the average waiting time in orbit. The first graph of Figure 1 tells us that the average waiting time in orbit is increasing with the growing of variance, and Erlang distribution has the smallest value and Hyperexponential with  $CV = 2$  has the largest one. However, there is obvious difference only when  $CV = 1.5$  and  $2$ , and the other curves are very close to each other. It means that variation will affect the waiting time in orbit only when the retrial time distribution has comparatively large variance.

The percent difference from exponential retrial time distribution tells us the accuracy if the phase-type retrial time distribution is approximated to exponential distribution which is much easier to handle. To make the results explicit, we use the percent difference to show the comparison. The percent difference is defined as the result difference of phase type (Erlang and Hyperexponential) and exponential distribution divided by the phase type result to get the percent difference from the true performance. Since the average waiting time in orbit for Erlang and Hyperexponential retrial distributions are respectively smaller and larger than exponential distribution, the percent differences will be negative and positive respectively. The second graph of Figure 1 displays the result of percent difference. We can see from it that all curves first increase from 0 when retrial rate is close to 0, then decrease to 0 again when retrial rate approaches to infinity. It agrees with our previous analysis of the retrial system. The actual value of the percent difference rises when the coefficient of variance moves farther from 1 and it means that the approximation of phase type to exponential distribution is better when the phase type distribution has a variance close to the exponential distribution. The difference is small (below 10 percent) when the coefficient of variance is between 0.5 and 1.225. However, when coefficient of variance is over 1.5, the peak of the error increases greatly (about 50% when  $CV$  is 2). That means that phase type retrial distribution with comparatively small variance can be approximated very well by exponential retrial distribution, but we have to be careful when the coefficient of variance of the phase type distribution is more than 1.5. It can be seen from Figure 1 that the peaks of the percent difference for all curves (except when  $CV$  is 1.5 and 2) occur at approximately the same position as retrial rate 4. Note our service rate  $\mu$  is set up to be 4, therefore the average waiting time in orbit has the largest percent error when retrial rate is close to service rate. If we zoom in the retrial rate between 2 and 6.5, as shown in Figure 2, we can see that the average waiting time percent difference varies in a small range, and the worst position happens at the point with retrial rate as around 3 for all the curves. It's again very close to the service rate 4. Note that to see clearer trend, we do not include the curves of  $CV$  are 1.5 and 2 in Figure 2, since they have bigger value and if we include them, the other curves will be compassed too small to see.

Figure 3 is the actual value and percent difference of the blocking probability. Similar to the average waiting time in orbit, it is not significant for the effect of variance to blocking probability when the coefficient of variance is smaller than 1.5, since the curves in the actual value graph are very close to each other and the percent difference have peaks

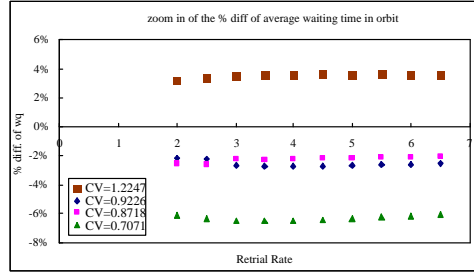


Figure 2: Zoom in of the % difference of waiting time in orbit

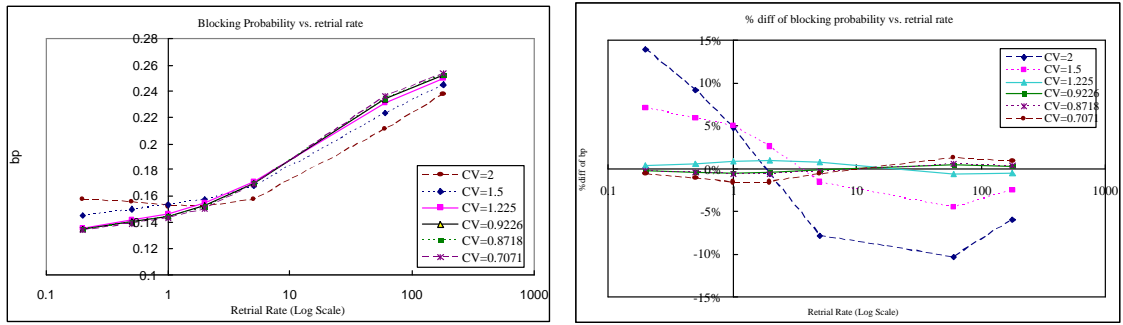


Figure 3: Blocking probability for retrial time distribution with various coefficients of variance

of at most 3%. Therefore, approximating phase type retrial distribution to exponential distribution has insignificant influence to the blocking probability when variance is less than 1.5. The percent difference of blocking probability shows an interesting conclusion. We can see from the second graph of Figure 3 that the percent differences change signs at about the point where retrial rate equals 10, as compared to the service rate 4. On the left of this point, Erlang retrial time distributions give smaller blocking probabilities than exponential distribution, and so the percents are negative values; while on the right of the transition point, the percent difference from Erlang are all positive values. Hyperexponential distribution has the opposite conclusions.

The cross-curve pattern of blocking probability is counter-intuitive. Generally intuition tells us that system performance gets worse when the variance of distribution is greater. However, 3 shows when variance of retrial time distribution gets larger from Erlang to Hyperexponential distribution, blocking probability first is better then worse with the growing of retrial rate, which is counter intuitive. Atkinson [16] listed some famous paradoxes in queueing systems and explained them. Atkinson indicates such paradox happens basically under two essential conditions: i) the coefficient of variance of the arrival stream (denoted as  $C_A$ ), and ii) some insensitivity in the system (for example,  $M/G/1$  is

insensitive to the service time). Some paradox happens when  $C_A > 1$  and some happens when  $C_A < 1$ . Paradox of variance happens to  $p_w$ , the probability that a customer waits before service, in G/G/1 queue, since  $p_w = \rho$  and it is insensitive to the service time variance. In particular, there is paradox in H2/G/1 system since H2 arrival has  $C_A > 1$  and the queue is insensitive to service time. Also paradox happens in the M/G/ $\infty$  system with batch arrivals since  $C_A > 1$  for batch arrival and M/G/ $\infty$  is insensitive to service time. Most importantly, The waiting time in queue,  $W_q$ , is sensitive to the variance in G/G/1 queue, but insensitive to the third moment of service time. Thus paradox of the third moment also exists for  $W_q$  (when  $C_A$  grows,  $W_q$  has crossover pattern with the variation of the third moment of service time). Atkinson [16] also guessed that paradox might happen in some similar multiserver cases, but unfortunately not any experiment results and explains are provided. Besides, Atkinson [16] did not discussed anything about retrial queue model either. But Atkinson did mention in [16] that paradox happens fairly often in queueing models, and we have to admit it instead of doubting our program and wasting time debugging.

Our retrial model has the similar paradox as the model mentioned in [16]. The arrival rate is a key point. Our arrival stream consists of new arrivals plus repeating customers. Intuitively there is more variance of the arrival stream, since the original arrival is Poisson, and now we add some new random stream from orbit. Therefore the coefficient of variance of arrival stream varies accordingly. To find the reason of the counter intuitive pattern of our retrial model, we first check the coefficient of variance for the arrival stream by simulation. Table 3 shows the coefficient of variance for Erlang and Hyperexponential retrial time distribution. We listed 3 simulation results and all simulations have sample size of 10000.

Table 3: Coefficient of Variance of Arrival Stream

<b>Erlang</b>								
CV	Retrial Rate	0.2	0.5	1	2	5	60	180
	Simulation 1	1.3439	1.3053	1.2553	1.2511	1.1098	0.4359	0.5378
	Simulation 2	1.3474	1.2933	1.2423	1.2314	1.1371	0.5131	0.485
	Simulation 3	1.3422	1.294	1.2449	1.1913	1.0763	0.525	0.5256
<b>Hyperexponential</b>								
CV	Retrial Rate	0.2	0.5	1	2	5	60	180
	Simulation 1	1.3226	1.2914	1.3218	1.3582	1.2816	0.792	0.8396
	Simulation 2	1.319	1.3131	1.296	1.3393	1.3293	1.0741	0.8512
	Simulation 3	1.3226	1.2914	1.3218	1.3582	1.2816	0.792	0.8396

From Table 3, we can see that when retrial rate is less than 60, the coefficient of variance for both Erlang and Hyperexponential retrial distribution have values of greater than 1, and when retrial rate is greater or equal to 60, CV decreases to less than 1. Retrial



rate does affect the input stream. When retrial rate is low, the arrival stream is more variable, which makes  $C_A > 1$ . But when retrial rate is high, the arrival isn't affected much, thus  $C_A$  is less than 1.

Now we discuss the insensitivity. Our cross-over pattern happens in the blocking probability, but not  $Wq$ , in the retrial time distribution discussion. In earlier this session, we concluded that the retrial time is practically ignorable, or in another word, practically insensitive. Moreover, this paradox happens depending on the retrial rate, which is that blocking probability decreases with the increase of variance when retrial rate grows larger than a certain value (10 approximately). We have mentioned that when retrial rate is large, the retrial system is like a queueing system, where the repeating customers has priority to enter service, although their order to enter service is random, but not first come first serve. Retrial distribution does not affect system performance in queueing system, thus retrial time distribution is insensitive when retrial rate is large. Therefore, another essential condition of insensitivity is satisfied. From the above discussion, the counter-intuitive phenomenon is explained.

In conclusion, our paradox pattern is as follows:

- When retrial rate is small,  $C_A > 1$ , and  $p_w$  increases with the increase of variance of retrial time distribution.
- When retrial rate is large,  $C_A < 1$ , and  $p_w$  decrease with the increase of variance of retrial time distribution.

### 3.2 Traffic Traffic and System Size Effect

We now study the effects of traffic and system size to the system performance under different phase type distributions. As we increase the traffic, we increase the number of servers according to the QED scheme. The retrial time distributions of Erlang-2 and Hyperexponential with  $CV = 1.5$  are used to make the graph explicit to show the difference.

The actual value and percent difference from exponential distribution of the average waiting time in orbit is shown in Figure 4. It is easy to tell from Figure 4 that when traffic grows from 9 to 25 (and the system size increases accordingly), the actual value and the percent difference of the average waiting time in orbit both go down. Because we set up our traffic and server number based on the QED scheme, this conclusion agrees with the feature of QED which tells us that system performance is getting better when server number increases, even though the relative load (traffic divided by server number) rises. However, the curves in Figure 4 are very close to each other and it means that the traffic effect to the retrial system is very tiny and negligible. Also, as the system sizes grow the percent errors shrink, if only a small amount, which leads us to believe that even

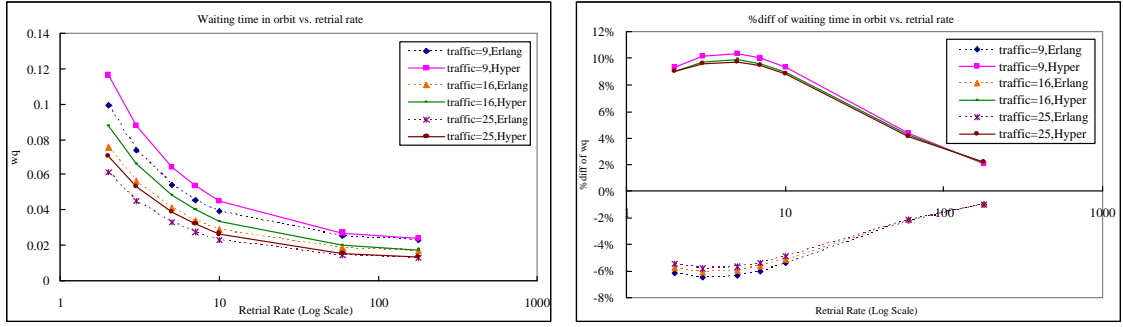


Figure 4: Traffic effect to the average waiting time in orbit

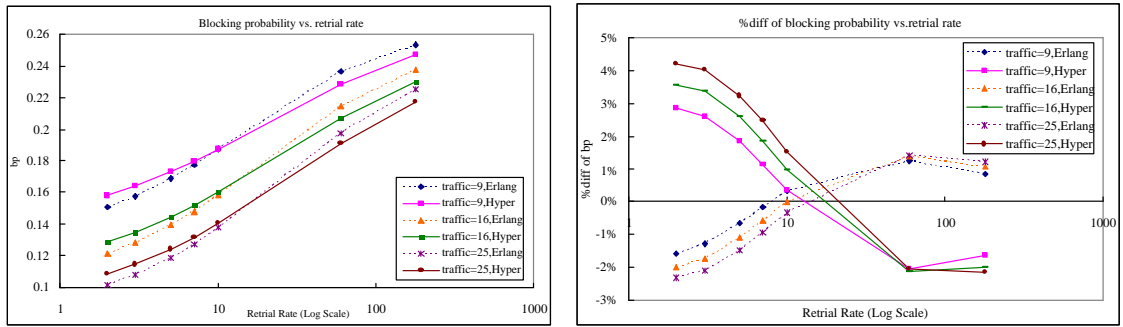


Figure 5: Traffic effect to the blocking probability

larger systems will be even less susceptible to approximating the average waiting times by using exponential retrials.

Figure 5 is the actual value and percent difference from exponential distribution of the blocking probability. We can see from the first graph of Figure 5 that as traffic and systems get larger, the blocking probability gets smaller. This echoes what happens in the ordinary Erlang-C model under the QED scheme which tells us that system performance is getting better when server number increases in the QED scheme, even though the relative load (traffic divided by server number) increases. However, the absolute values of percent difference from both the Erlang and Hyperexponential distribution always increases greater when traffic grows larger from 9 to 25, as seen on the second graph of Figure 5. That means, when traffic grows larger, the error of approximating Erlang and Hyperexponential retrial distribution to Exponential distribution is getting worse. We can see there is still a transition point (about 10 of retrial rate) after which the sign of percent difference changes.

If we look at the absolute value of the average waiting time in orbit from the left

graph of figure 5 which shows the actual values of the blocking probability instead of the percentage difference, we can observe that when traffic grows heavier, from 9 to 25, the blocking probability is getting smaller which means a larger traffic gives us a better system performance. It accords for the feature of QED scheme which tells us that system performance is getting better when server number increases in the QED scheme, even though the offered load (traffic divided by server number) increases.

## 4 Service Duration Distribution

### 4.1 Retrial Rate Effect

Service time duration distribution plays an important role in system performance. We will discuss the system performance difference among various phase type service duration distributions with the different coefficients of variance, like retrial time distribution discussion in the previous section. In addition, the traffic and third moment influences are also discussed in this section.

Figure 6 is the actual value and percent difference of the average waiting time in orbit from Exponential distribution when service duration has various phase type distributions associated with different coefficients of variance. From the first graph of Figure 6, we can see that when retrial speeds up, the actual value of waiting time in orbit is getting smaller for all distributions. It means that the waiting time in orbit of customers will be decreased if they retry faster, since they will check more often whether or not there is any idle server and have a higher chance to get into service. The graph also shows that when coefficient of variance is greater than 1, the percent difference of average waiting time in orbit increases with the increase of CV, but when CV is smaller than 1, percent difference of the average waiting time in orbit decreases with the increase of CV. It tells us that the approximation will be better when the coefficient of variance is close to 1, which is the CV of exponential distribution.

The second graph of Figure 6 is the percent difference from exponential distribution of the waiting time in orbit. Opposite to the actual value, the percent difference is getting bigger from zero with the increase of retrial rate and approaches a constant when the retrial rate is fast enough. When retrial is slow, the system can be seen as an Erlang-B loss system which is insensitive to the service duration distribution. Therefore, the difference approaches zero for all phase type service distributions with small enough retrial rate. On the other side, the percent difference approaches constant when the retrial rate goes to infinity. This is because when retrial speeds up, the system is like an Erlang-C delay system except that the service duration has various distributions. Therefore, the difference is the stationary average waiting time in queue between the several delay systems with different service duration distribution where the queue buffer capacities are infinite. The

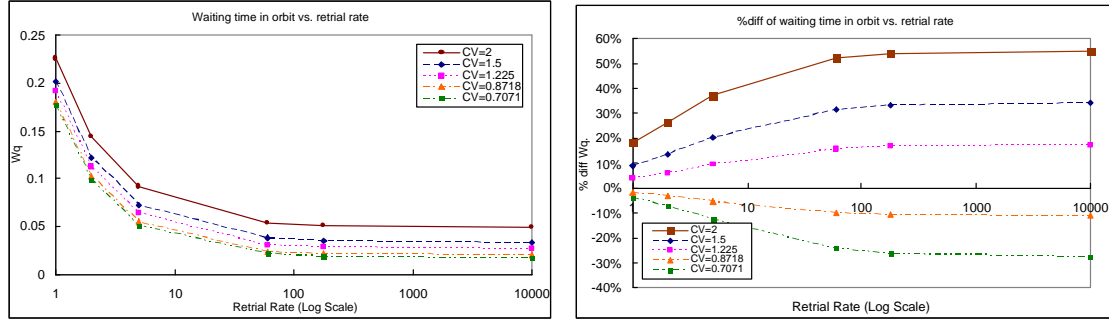


Figure 6: Actual value and percent difference of the avg. waiting time in orbit from Exponential service duration

curves of percent difference, however, are increasing for all coefficients of variance and opposite to the actual value curves which are decreasing. Therefore, when retrieval rate gets faster, it will incur more error if we approximate phase type service time distribution to exponential service time distribution which is much easier to handle.

Moreover, compared to the error of retrieval time distribution with coefficient of variance (less than 15% for all retrieval rates), service time distribution gives much larger errors even if the retrieval rate is very small (about 10% when retrieval rate is 1).

Figure 7 is the blocking probability with different service duration distributions. The first graph of Figure 7 is the actual value and the second graph is the percent difference from exponential service duration distribution. From the first graph of Figure 7, it can be shown that with the increase of variation (increasing of CV), the blocking probability grows larger. Therefore, the system performance for the incoming new customer is hurt when the coefficient of variance of the service duration distribution grows. However, if we look at the second graph about the percent difference from exponential distribution, we can see that the curve first goes up from zero until a peak and then goes down to reach a constant. The explanation of the two ends of the curves is similar as the analysis of average waiting time, system being considered as Erlang-B loss system when retrieval rate is slow and Erlang-C delay system when retrieval rate is large. The largest error occurs at the peak. Blocking probability doesn't have as big error as the average waiting time, therefore it can be approximated better for blocking probability when the phase type service duration distribution is replaced by exponential distribution.

Figure 6 and Figure 7 shows that the peak of percent difference of waiting time in orbit occurs at the point where retrieval rate is about 4. It is close to the service rate and agrees with the peak retrieval rate in PH-retrial model. It is more clear to see the peak of percent difference from Figure 8 which zooms in the retrieval rate from 2 to 6.5.

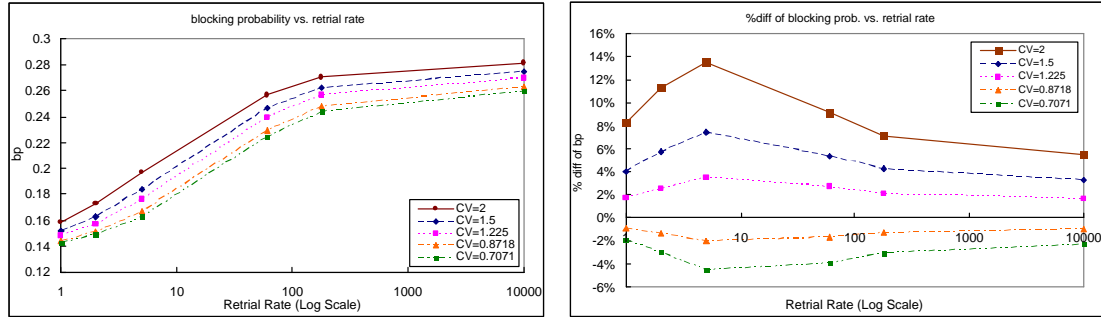


Figure 7: Actual value and percent difference of blocking probability from Exponential service duration

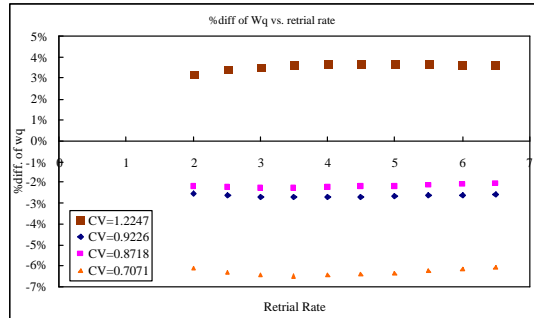


Figure 8: Zoom in of % difference of waiting time in orbit

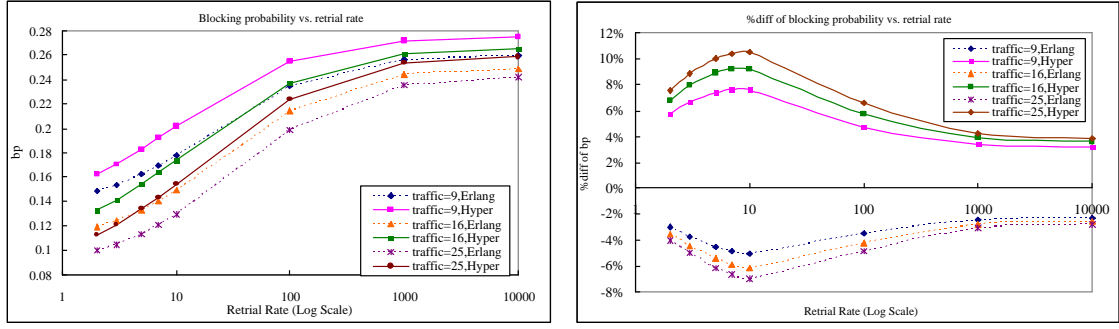


Figure 9: Actual value and percent difference of blocking probability from Exponential service duration

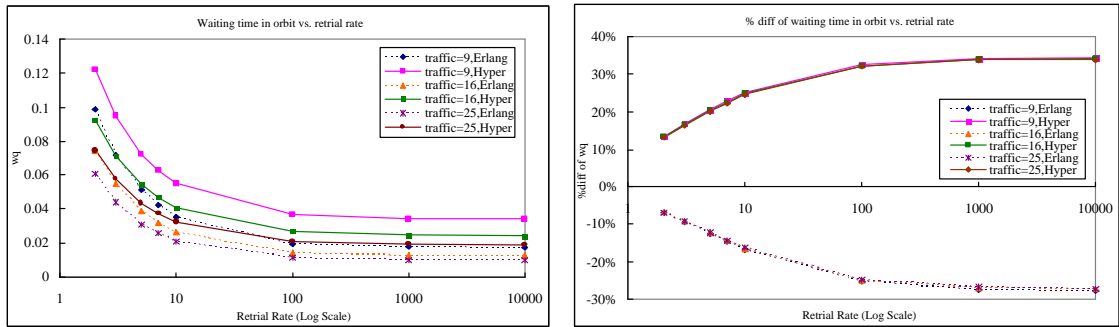


Figure 10: Actual value and percent difference of the waiting time in orbit from Exponential service duration

## 4.2 Traffic and System Size Effect

Again service duration distributions of Erlang-2 and Hyperexponential with  $CV = 1.5$  are used to study the traffic effect to the different service duration distributions. Again as we increase the traffic, we increase the number of servers according to the QED scheme.

Figure 9 and Figure 10 shows the actual value and percent difference from exponential service distribution of the blocking probability and average waiting time in orbit respectively. The blocking probability and average waiting time in orbit both decrease when traffic grows from 9 to 25 (and the system size increases accordingly) for Erlang and Hyperexponential distributions. As we discussed in the previous section, this conclusion agrees with the QED scheme which says that when traffic increases, the server number also increases such that the relative load will decrease and the system performance is getting better.

However, if we look at the second graph of Figure 9, the percent difference of the

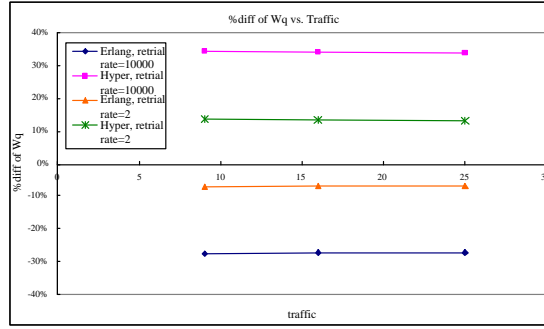


Figure 11: Percent difference of the waiting time in orbit with retrial rate  $\alpha=2$  and  $\alpha=10000$

blocking probability increases when traffic goes up which indicates that it gets worse to approximate phase type service duration distribution to exponential distribution when traffic rises. The second graph of Figure 10 is the percent difference from exponential distribution of the waiting time in orbit. The curves in this graph are very close to each other which means that traffic has very tiny effect on the difference of waiting time in orbit. To make the graph more explicit to see, we draw another graph of the percent difference of the waiting time in orbit, as shown in Figure 11. The x-axis of Figure 11 is traffic (9, 16 and 25), and the curves in Figure 11 are percent differences of Erlang and Hyperexponential service time with retrial rate of 2 and 10000. We can see that the curves are almost flat lines, which means there is very tiny effect of traffic on the approximation error of waiting time in orbit.

Again, service duration distribution has an influence of about 40% at most to the average waiting time in orbit and exponential distribution can not be easily approximated to the other phase type distributions, as we discussed in the previous subsection.

## 5 Conclusion

In this paper, we studied the sensitivity of phase type distribution retrial time and service time on the system performance under QED scheme. Retrial time distribution effect could be practically ignored since the errors of approximating general retrial distribution to exponential distribution are acceptable. In particular, when coefficient of variance is less than 1.5, the system performance of retrial time distribution has errors less than 5%. There is counter-intuitive pattern for the blocking probability in the retrial time distribution discussion. Namely, the blocking probability is better (or worse) when coefficient of variance of retrial distribution changes larger (or smaller). This paradox is because of the arrival stream variance and the insensitivity of blocking probability with retrial time distribution. Theoretical proof of this explanation is a future research topic.

Service time distribution is comparatively significant, especially when coefficient of variance is more than 1.5. Traffic load affects system performance according to the general conclusion of QED scheme. With the increasing of traffic load, system performance is even better, since server numbers are increased. However, the approximation error for both retrial and service duration distributions are getting worse with traffic.

## References

- [1] Satoshi Nojo and Hitoshi Watanabe. A new stage method getting arbitrary coefficient of variation by two stages. *The Transaction of the IEICE*, 70(1):33–36, January 1987.
- [2] G. I. Falin. Survey of retrial queues. *Queueing Systems*, 7:127–167, 1990.
- [3] T. Yang and J. G. C. Templeton. Survey on retrial queues. *Queueing Systems*, 2:201–233, 1987.
- [4] J. R. Artalejo. Accessible bibliography on retrial queues. *Mathematical and Computer Modelling*, 30:1–6, 1999.
- [5] J. R. Artalejo. A classified bibliography of research on retrial queues: Progress in 1990-1999. *TOP*, 7(2):187–211, 1999.
- [6] G. I. Falin and J. G. C. Templeton. *Retrial Queues*. Chapman & Hall, 1997.
- [7] S. N. Stepanov. Increasing the efficiency of numerical methods for models with repeated calls. *Problems of Information Transmission*, 22(4):313–326, 1987.
- [8] P. Le Gall. The repeated call model and the queue with impatience. In *Proceedings of Third International Seminar on Teletraffic Theory*, pages 278–289, 1984.
- [9] A. A. Fredericks and G. A. Reisner. Approximation to stochastic service systems, with an application to a retrial model. *The Bell System Technical Journal*, 58(3):557–576, March 1979.
- [10] M. F. Neuts and B. M. Rao. Numerical investigation of a multiserver retrial model. *Queueing Systems*, 7:169–189, 1990.
- [11] Vladimir V. Anisimov and Jesus R. Artalejo. Approximation of multiserver retrial queues by means of generalized truncated models. *TOP*, 10(1):51–66, 2002.
- [12] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 1981.



- [13] Winfried K. Grassmann. Finding the right number of servers in real-world queuing-systems. *Interfaces*, 18(2):94–104, Mar–Apr 1988.
- [14] W.K.Grassmann. Is the fact that the emperor wears no clothes a subject worthy of publication? *INTERFACES*, 716(2):43–51, March–April 1986.
- [15] Peter J. Kolesar and Linda V. Green. Insights on service system design from a normal approximation to Erlang’s formula. *Production and Operations Management*, 7(3):282–293, 1998.
- [16] J.B. Atkinson. Some related paradoxes of queuing theory: new cases and a unifying explanation. *Journal of the Operational Research Society*, 51:921–935, 2000.