SIMULATION AND ANALYSIS OF TANDEM QUEUES WITH BLOCKING

Constantinos Boulis, Laura Ing

ABSTRACT

In this work we describe simulation as well as analytic results for the problem of tandem queues with blocking. We give an answer to a number of interesting questions, like what is the expected delay of a customer, what is the capacity of the system, if the order of servers is important and given that the sum of service rates is constant, under what service rates is the expected delay minimized. We present exact analysis for the M/M/2 case and further support our findings with simulation results. Although exact analysis is difficult to extend for the G/G/n case, we present some simulation results under non-Poisson arrival and service rates.

PROBLEM DESCRIPTION

Tandem or serial or cascade connection of $N$ queues refers to a configuration of a group of queues where customers enter queue 1, served in queue 1, then forwarded to queue 2, served in queue 2, forwarded in queue 3 and so on until they finish service in queue $N$. The total time a customer spends in a system will depend on the time it spends waiting for service in all queues and being serviced in all queues. The problem is trivial if all the queues have independent Poisson arrival and service rates, because in this case the output of each queue is a Poisson process and since it will be the input to the next queue, the total system is an interconnection of M/M/1 queues. But in the general case, if the queues have a general distribution for inter-arrival and service times the output of each queue is not a renewal process and analytical tracking of results becomes progressively more difficult. In [1] and [2] simulation results are reported and a number of interesting questions are answered. Specifically it is shown that the ordering of queues can have important impact on the expected delay and a heuristic to order the queues is given. By fitting functions to experimental data it is shown that tighter bounds can be established than by using analytic tools.

Tandem servers with blocking, refers to a similar problem but with an important difference. Each customer is serviced in just one server but she is not allowed to exit the system until all customers in front of her have exited the system. Therefore each server can block, meaning that the server is not servicing any customer but also cannot accept new customers. A practical scenario where this problem arises is a gas station with just one lane for entering/exiting the station and two gas pumps in serial connection with each other, as it is shown in Figure 1.

![Figure 1 Tandem servers with blocking](image_url)
There are two blocking cases for the two tandem servers problem shown in Figure 1. If customer 1 finishes before customer 2 then she can exit the system the moment she finishes but customer 3 cannot be forwarded to server 2 because she is blocked by customer 2. Similarly if customer 2 finishes before customer 1 then she cannot exit the system until customer 1 finishes.

These blocking cases show that under heavy traffic the utilization of each server is going to be less than 1. But since at some times we will have some parallel servicing of customers we can readily infer that the expected delay of a customer is going to be less than the expected delay of a two parallel server system but more than a single server system. We will show exactly where the total expected delay lies and compare it with other queue configurations in the next sections.

Notice that even when the service times are independent with each other and with the arrival process, and even when the inter-arrival and service times are exponential the departing process is not Poisson. This is because we may see \( N \) (where \( N \) is the number of tandem servers) departures at the same moment. This will happen when the service time of the last customer is greater than the service time of all other customers. So when customer in \( N \) exits, every other customer will also exit.

The organization of this paper is as follows. In Section 2, we present analytic derivation for the tandem M/M/2 case, calculating the expected delay of a customer, the capacity of the system, the utilization of each server, if the order of servers is important and under the constraint that the sum of the service rates is constant how do we set each one of the service rates to minimize the expected delay. In Section 3, we present simulation results, backing up our analysis and also show some results under non-Poisson arrival and service rates. In Section 4, we conclude our work showing how these results can be of practical interest to someone designing a gas station (or any problem similarly formulated).

**ANALYTIC RESULTS FOR THE M/M/2 TANDEM QUEUE**

We present analytic results for the M/M/2 tandem queue. That is we assume that the inter-arrival time is distributed exponentially with rate \( \lambda \), the service times for server 1 and 2 are distributed exponentially with rate \( \mu_1 \) and \( \mu_2 \). Service times are assumed independent of each other and with the inter-arrival time. The queue is assumed to have infinite capacity (no customers are rejected) and follow the FIFO scheduling policy.

Based on these assumptions we give answer to a number of interesting questions:

*What is the maximum arrival rate to have a finite expected delay per departed customer?*

If we assume that the system operates at the unstable region (arrival rate greater than service rate) then with probability 1 there will always be customers in the buffer after an infinite amount of time, and so customers will always enter the servers by increments of two. From Figure 1 we see that new customers will enter the servers only when the customer in 1 finishes. The system is unstable if the expected number of customers that arrived during the expected occupation time of customer in 1 is greater than two. But the expected occupation time of customer in 1 is the expected maximum of the service times in 1 and 2 because if 1 finishes before 2 she has to wait until 2 finishes, to exit the system. That is:

\[
\lambda E[T] > 2 \iff \lambda E[\max\{T_1, T_2\}] > 2 \iff \lambda \int_0^\infty \left(1 - P(T_1 \leq t, T_2 \leq t)\right) dt > 2 \iff \\
\lambda \int_0^\infty \left(1 - P(T_1 \leq t)P(T_2 \leq t)\right) dt > 2 \iff \lambda \int_0^\infty \left[1 - (1 - e^{-\mu_1 t})(1 - e^{-\mu_2 t})\right] dt > 2 \iff \\
\lambda \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} - \frac{1}{\mu_1 + \mu_2}\right) > 2
\]
If we assume that $\mu_1 = \mu_2 = 1.0$ then we get that the system will have finite delay as far as $\lambda \leq 1.33$. Note that for a parallel M/M/2 queue the system is stable if $\lambda \leq 2.0$ and for a M/M/1 queue stability implies $\lambda \leq 1.0$. So indeed the critical point$^1$ of the tandem M/M/2 queue is between the critical points of the parallel M/M/2 and the M/M/1, as is intuitively appealing. But note that the critical point is closer to M/M/1 rather than M/M/2 so by doubling the number of servers we are only able to increase the arrival rate by 33%.

**Does the order of the servers affect the critical point?**

No. Switching $\mu_1$ with $\mu_2$ and $\mu_2$ with $\mu_1$ does not change the critical point.

**Given that the sum of service rates is constant, what service rates maximize the critical point?**

The maximum point will have a derivative equal to zero (since it’s a saddle point) and therefore we have:

$$\frac{d}{d\mu_1} \left( \frac{1}{\mu_1} + \frac{1}{c - \mu_1} + \frac{1}{c} \right) = 0$$

$$\frac{-1}{\mu_1^2} + \frac{1}{(c - \mu_1)^2} = 0$$

$$\frac{c}{2} = \mu_1 = \mu_2$$

That’s a very interesting result, stating that under the constraint that the sum of the service rates are constant, the critical point is greatest when all the service times are i.i.d. So $\lambda = 1.33^-$ is the maximum arrival rate to have finite expected delay given that the sum of the service rates is equal to 2.0. With $1.33^-$ we denote the number, which is as close as possible to 1.33.

**What is the utilization factor of each server under heavy traffic?**

Again we assume that the queue operates at the unstable region so with probability 1 there are infinite customers in the buffer after infinite time. Let $T_{i,j}$ be the service time of the $j^{th}$ customer in the $i^{th}$ server, and let $U_i$ be the entry time of the $j^{th}$ customer into the server, where $i = 1,2$ and $j = 1,2,3\ldots$ Since we operate in the unstable region two customers will always enter simultaneously the servers. We start with the utilization percentage for Server 2:

---

$^1$ The critical point of a queue is defined as the maximum arrival rate to have finite expected delay.
\[
\lim_{n \to \infty} \frac{\sum_{i=1}^{n} T_{2,i}}{U_{2,n+1}}
\]

\[
\lim_{n \to \infty} \frac{\sum_{i=1}^{n} T_{2,i}}{n}
\]

\[
\frac{E[T_2]}{E[\max\{T_1, T_2\}]}
\]

\[
\frac{1}{\mu_2}
\]

\[
\frac{1}{\mu_1 + \mu_2} - \frac{1}{\mu_1 + \mu_2}
\]

\[
\frac{\mu_1 (\mu_1 + \mu_2)}{\mu_1^2 + \mu_1 \mu_2 + \mu_2^2}
\]

The same formulation is used for server 1 and we finally get the set of equations:

\[
\rho_1 = \frac{\mu_2 (\mu_1 + \mu_2)}{\mu_1^2 + \mu_1 \mu_2 + \mu_2^2}
\]

\[
\rho_2 = \frac{\mu_1 (\mu_1 + \mu_2)}{\mu_1^2 + \mu_1 \mu_2 + \mu_2^2}
\]

Where \( \rho_1, \rho_2 \) are the utilization factors of servers 1 and 2 respectively. If we further assume that \( \mu_1 = \mu_2 = 1.0 \) then \( \rho_1 = \rho_2 = 2/3 \). Again the results are intuitively appealing, since we know that because of blocking the utilization factor for each server will never be unity. These results are valid only under heavy traffic.

**What is the Markov Chain for the 2-server tandem system?**

The discrete time Markov chain for a standard M|G|1 system uses small increments of time (\( \delta \)) to calculate the probabilities of transitioning from one state to another. Using \( \delta \) as the time unit, the probability that an arrival happens in \( \delta \), but Server 1 and Server 2 do not finish their service, is as follows:
\[ P\{Y_{\text{arrival}} \leq \delta, Y_1 > \delta, Y_2 > \delta\} \]
\[ P\{Y_{\text{arrival}} \leq \delta\} \cdot P\{Y_1 > \delta\} \cdot P\{Y_2 > \delta\} \]
\[ (1 - e^{-\lambda \delta}) \cdot (e^{-\mu_1 \delta}) \cdot (e^{-\mu_2 \delta}) \]
\[ [1 - (1 - \lambda \delta + o(h))] \cdot (1 - \mu_1 \delta + o(h)) \cdot (1 - \mu_2 \delta + o(h)) \]
\[ \lambda \delta + o(h) \]

\( Y_{\text{arrival}} \) is the excess lifetime of the next arrival, and \( Y_1 \) and \( Y_2 \) are the excess lifetimes until Server 1 and 2 (respectively) finish processing. Similarly, \( P\{Y_{\text{arrival}} > \delta, Y_1 \leq \delta, Y_2 > \delta\} = \mu_1 \delta + o(h) \) and \( P\{Y_{\text{arrival}} > \delta, Y_1 > \delta, Y_2 \leq \delta\} = \mu_2 \delta + o(h) \).

From Poisson process theory, we know that the probability of 2 arrivals occurring in a sufficiently small \( \delta \) is zero. Using a similar approach, we can argue that the probability of more than one of the three described events (an arrival, Server 1 finishes or Server 2 finishes) occurs in \( \delta \) is 0. A brief outline follows:

\[ P\{Y_{\text{arrival}} \leq \delta, Y_1 < \delta, Y_2 > \delta\} \]
\[ P\{Y_{\text{arrival}} \leq \delta\} \cdot P\{Y_1 \leq \delta\} \cdot P\{Y_2 > \delta\} \]
\[ (1 - e^{-\lambda \delta}) \cdot (1 - e^{-\mu_1 \delta}) \cdot (e^{-\mu_2 \delta}) \]
\[ [1 - (1 - \lambda \delta + o(h))] \cdot [1 - (1 - \mu_1 \delta + o(h))] \cdot (1 - \mu_2 \delta + o(h)) \]
\[ o(h) \]

To summarize, one and only one event can happen in \( \delta \).

Given these transition probabilities, we can draw the discrete-time Markov chain as the following:

![Figure 2. Discrete time Markov chain for the M|M|2 tandem system](image-url)
system decreases by one is equal to the event that server 2 finishes in \( \delta \), server 1 does not finish (\( T_1 > T_2 \)), and there were no arrivals.

\[
P_{n,n-1} = P\{Y_2 < \delta, T_1 > T_2, Y_{arrival} > \delta\} \iff P_{n,n-1} = P\{Y_2 < \delta\} P\{T_1 > T_2\} P\{Y_{arrival} > \delta\} \iff P_{n,n-1} = (1 - e^{-\mu_2 \delta}) \left( \frac{\mu_2}{\mu_1 + \mu_2} \right) \left( e^{-\delta} \right) \iff P_{n,n-1} = \left[ 1 - \left( 1 - \mu_2 \delta + o(h) \right) \right] \left( \frac{\mu_2}{\mu_1 + \mu_2} \right) \left( 1 - \lambda \delta + o(h) \right)
\]

Similarly, the probability of decrementing by 2 is equivalent to the event that \( T_1 < T_2 \), server 2 finishes in the window \( \delta \), and there were no arrivals in \( \delta \).

\[
P_{n,n-2} = P\{Y_2 < \delta, T_1 < T_2, Y_{arrival} > \delta\} \iff P_{n,n-2} = \left( \frac{\mu_1 \mu_2}{\mu_1 + \mu_2} \right) \delta + o(h)
\]

Given these limiting probabilities, we can redraw the Markov chain, this time with the number of customers in the system as the states.

Figure 3. Discrete time Markov chain with the number of customers in the system as the states.

With a standard M|M|m server, the method to determine the steady-state number of customers in the system is to first find the recursive relationship between \( p_n \) and \( p_{n+1} \), where

\[
p_n = \lim_{t \to \infty} P\{N(t) = n\} \text{, from the Markov Chain [3]. Drawing an imaginary line between any two states, under steady state we can assume that the flow in both directions must be equal. Then, because these limiting probabilities sum to one, } p_n \text{ can be expressed as the product of } p_0 \text{ and some constant to the } n^{th} \text{ power. Using power series, the steady-state number of customers in the system, } \sum np_n \text{, is reduced to a closed form solution.}
\]

However, the difficulty with the tandem server is that we can go from state \( n \) to state \( n-2 \). The standard M|M|m server can only increment or decrement by one, so \( p_n \) and \( p_{n+1} \) are directly proportional. From the recursive relationship of the tandem servers we can in theory calculate the steady state expected number of customers in the system, but a closed form solution is computationally difficult. In this case, a simulation is much more convenient.

**SIMULATION RESULTS**

We designed a simulator for the tandem queues with blocking and used it to compute various quantities. In Figure 4 we show the expected delay of a customer for three different queues assuming...
exponential inter-arrival and service times. We observe that the M/M/2 tandem system becomes unstable for arrival rates greater than 1.33 as was predicted from our analysis. Each point on the graph is the average value of 10 different experiments with 10 million departed customers for each of the experiments. We found that the more we were approaching the stability threshold, the more variant our results were, this is the reason we took many measurements per point.

Since analysis cannot carry us much further than the M/M/2 case we used our simulation for the more general cases. We run experiments for the U/U/2 case where U denotes uniform distribution. Therefore a U/U/2 queue is a queue where both the inter-arrival and service times are uniform. To have a common ground of reference between the M/M/2 and U/U/2 cases we matched the mean of the distributions, that is if the exponential distribution has mean $1/\lambda$ then the uniform distribution will also have mean $1/\lambda$. The results are shown in Figure 5. Interestingly enough, we observe a fundamental change in the behavior of the system. Firstly, the M/M/1 appears to have higher expected delay than the U/U/1 case revealing that in the general case higher moments are also important. Secondly, the stability of the U/U/2 tandem queue is now at 1.5 instead of 1.33, which was the case with the M/M/2 tandem. These results are intuitively appealing and especially for the stability of the G/G/2 case we can say that it lies somewhere between 1 and 2 but the exact point will depend on the inter-arrival and service distributions. For example, if the inter-arrival and service times are deterministic then the stability can be as high as 2 (customers entering at the same time and requiring the same time of servicing) or as low as 1 (customers entering with a time difference of $d$ and their service time is $d$).

Another interesting question is what is the behavior of a tandem queue when the number of servers is increasing. We can find the answer to this question by using the same formulation we used for the derivation of the stability point for the 2-server case. Assuming M/M/n queues, $\lambda$ the arrival rate and that all servers have service rate equal to 1 we want:

$$\lambda E \left[ \max \{T_1, T_2, \cdots, T_n \} \right] < n \iff \lambda < \frac{n}{\sum_{i=1}^{n} C(n,i) \frac{(-1)^{i+1}}{i}}$$

Where $C(n,i)$ is the binomial coefficient. A plot for various numbers of tandem servers is shown in Figure 6.
Figure 5. Comparison of expected delay for M/M/2 and U/U/2 queues

Figure 6. Comparison of N-parallel and N-tandem servers with all service rates equal to 1
CONCLUSIONS

Our paper formulated the problem of tandem servers with blocking and gave answer to a number of interesting questions. We presented analytic results for the M/M/2 case and also simulation results for the more general case of U/U/2. We have also provided analytic derivation for the M/M/n and M/M/∞ cases.

From a practical point of view these results may help design a gas station. Suppose that a gas owner does not have enough space to have two parallel gas pumps and also suppose that the inter-arrival and service times are exponentially distributed. The gas station owner faces two choices. Choice one is to have a single gas pump with service rate 0.66 which costs C_1 and choice two, is to have two tandem gas pumps each one with service rate 0.5 and cost C_2 . Both systems have the same total expected delay per customer but the cost of the first one is C_1 while the cost of the second one is 2C_2 . Depending on what is the relative difference of these costs one solution may be more cost-effective than the other.

The analysis that we have presented also suggests some reasons why the drive-through ordering that many fast food restaurants are using has only a single queue and not tandem queues. If we assume that space does not allow for a multi-lane drive-through system (parallel system) then why we do not see 2-tandem servers per drive-through. Of course there are many answers to this question and probably the most important ones cannot be found in engineering (for example customers may be really upset if they have finished their service and they cannot exit the system because of the preceding customer). But our analysis shows that the expected delay is minimized only when the two service rates are equal. This is an important constraint not easily achievable in practice. If a new employee comes his/her performance for the first weeks will be worse than the more experienced employees and this will increase the expected delay. Given that the renewal rate of employees in fast food restaurants is quite high this would suggest that the performance of a tandem system would often be much worse than the optimum case. Even if the expected delay will be lower than the single server system, the whole system will not be as cost-effective as the single server system.

Further topics of study outside the scope of this paper include finding the closed form solution for the long-term probabilities and expected number of customers. Although much less mathematically tractable than the M/M/2 case, the U/U/2 was only minimally explored. We also had questions about the behavior of the tandem system as the number of servers approached infinity that the simulations were unable to resolve.

REFERENCES

