Civil Engineering Dimension, Vol. 14, No. 3, December 2012 (Special Edition), 139-146 CED 2012, 14(3), DOI: 10.9744/CED.14.3.139-146
ISSN 1410-9530 print / ISSN 1979-570X online

Queuing Rule of Thumb based on M/M/s Queuing Theory with Applications in Construction Management

Teknomo, K.[1]

Abstract: The current trend of queuing theory development is toward more precision which requires higher mathematical manipulation. In this paper, we attempted to reverve the current trend toward simplification of queuing formulas such that it can be used in more practical purposes, especially in construction industry. Through numerical examples of two case studies on concreting and earth moving, how to model the construction activities as queuing systems is illustrated systematically. Through the numerical examples, it is shown that when the customer cost is much lower than the server cost, queuing system can be simplified only to incorporate the constraint equation. The queueing constraint equation is suggested to be used as queuing rule of thumb. The proposed rule of thumb is rather conservative in term of queuing performance compared to the standard stochastic queuing formula because it is assumed that all the customers arrive at once in the beginning of the service.

Keywords*:* Rule of thumb, constraint equation, concreting, earth moving.

Introduction

Queuing is traditionally defined as a process where people, materials or information need to wait at certain time to get a service. Queuing theory has been developed since the beginning of last century (by A. K. Erlang in 1917). Despite all the advancement of the queuing theory in almost a century, many people still have lack of understanding about how to manage queue. In our daily life, we still face a lot of queuing from school, to bank and from restaurant to toilet. We are still being frustrated with our daily traffic and a long queue in registration and security check. In fact, the queue does not diminish by the advancement of our knowledge in queuing theory, but it grows into more complex reasons. Public and business understanding of queuing is still very far from the science. For instance, instead of distributing the queue over time and space, many managers do the very common mistake by making policy to concentrate the demand over time and space. Instead of solving the queuing problem, they create more severe queues. The lack of understanding on how to solve queueing problems requires public education. However, we cannot educate queuing theory as public common knowledge if the required background knowledge of the mathematics is fairly complicated. We really need a much simpler version of the queuing theory that can be used to educate public about queuing. Hence, a rule of thumb for queuing performance formula is needed. One best part of public education is to educate the engineers. In this paper, I would like to point out how we can use queueing theory in practice of construction field.

[1] Department of Information System & Computer Science, Ateneo de Manila University, PHILIPPINES.
Email: teknomo@gmail.com

It should be noted that this paper is not the first paper which discuss the applications of queuing theory in construction management. In fact, there are a few books and papers [1,2] had discussed about this topic in general. A study by Cheng et al [3] develops a construction management process re-engineering performance measurement model by applying queuing theory to calculate process operation time in order to strike an optimal balance between process execution demand and manpower service capacity.

To understand queueing theory fully, one need to have background in statistics and differential equations and to be able to manipulate Markov chain. Looking at the recent development of vacation queuing theory (example: Tian and Zhang in 2006 [4]) and queuing network (example: Yue et al in 2009 [5]), the trend of queuing theory development is toward more precision which requires higher mathematical manipulation. While higher precision queuing theory computation is great for the body of knowledge, this trend also produces wider gap between the theory and practice. Hall [6] cited an argument that operation research profession could and should be more scientific and less mathematical. We should concern with how the system behave and less concern with abstract symbol manipulation. Therefore, in this paper I attempted to reverse this trend toward simplification of queuing formula such that it can be used in more practical purposes, especially in construction industry. The presentation of the case studies is more qualitative than quantitative. All the formulas are either common sense or well known in the queuing theory that their proofs can be traced back in the references. Construction practitioners and engineers love simplification and rule of thumb.

It would be shown at the end of this paper that we can simplify our treatment of queuing system into a simple queuing rule of thumb.

Most of the traditional literatures on queuing theory are using M/M/s queuing system, which assume Poison arrival distribution, exponential distribution of service time and s number of servers. In this paper, this tradition is also followed. Unlike these literatures, however, this paper describes systematically how to model the construction activities into a queuing system. After showing the formulas, I also extend the treatment of M/M/s queuing model in the existing books and papers into economic analysis and finally to come out with a simple queuing rule of thumb.

Case Study 1: Concreting

In this first case study, we are looking at concreting activity as a queuing system. Unlike queuing in supermarket where there is a single server to serve one customer, concreting operation requires many agents such as crane, hoist or bucket or concrete pump, placement crew, and vibrating crew as well as concrete trucks. The clutters in the agents may become the source of misunderstanding on how to model concreting activity as a queuing system. Which agents will play role as servers and which agent should play role as customers in the queuing system? What kind of optimization we would like to model through the queuing system? How many servers should we provide?

Two main components of a queuing system are customers and servers. Customer is person or thing that demands for service. Customer does not have to be a person and does not necessarily have to wait for service. Servers provide service to the customers.

In the case of concreting activities, which agents are the servers and which agents are the customers? My suggestion to solve this problem is to use a simple technique as the following. First, we identify stakeholders and then we find out the flow of activity in which we say that every type of agents is a server for their immediate customers. After that, we take side with the stakeholder and identify the most expensive agents as the servers.

Let us identify the two stakeholders: contractor and concrete company. From the contractor point of view, the concrete trucks are the customers and what they provide, the crew and crane with bucket or the concrete pumps, are the servers. From the concrete company's point of view, the concrete trucks are the servers while all the concrete pumps equipments and placement crew from the contractor's side are the customers. For the sake of uniformity, let us take side with the contractor point of view. The other point of view would be equivalent anyway.

Now we can look at the flow of the concreting activities and say that the crane and bucket (or the concrete pump or hoist) is the server for the concrete trucks. The placement and vibrating crew are the servers for the crane and bucket. Using this serial system, as we identify the most expensive agent, the concrete pumps should be identified as the servers for both the crew and the concrete trucks.

Queuing theory is rich with optimization. Our next problem is to answer what kind of optimization we would like to model through the queuing system? In our concreting activities, the placement and vibrating crew are usually the ordinary workers that always available on the construction site. These servers must operate together to serve single customers of the concrete truck. If one of the servers is not available, the service of concreting cannot be done. Thus, normally we do not want to optimize the number of the crew. In fact, the number of placement crew is directly related to the number of concrete pumps or the hoists or the cranes. Say, to place and vibrate V cubic meter of concrete per hour, we need x number of crew. If each concrete pump will provide g cubic meter of concrete per hour, we can easily find the number of required total crew to serve the n concrete pumps as:

$$X_n = n \left\lceil \frac{g.x}{v} \right\rceil \tag{1}$$

Figure 1. Concreting Activities as a Serial Queuing System. Concrete Pump is the Server for Both Systems

The notation $\lceil \cdot \rceil$ is ceiling function to get the lowest integer that higher than the argument. The unit time can be set as either hour or minute as convenient of the modeler. For uniformity, we use hourly unit of time.

The direct relation between the number of crew and the number of concrete pumps simplifies our queuing system. Instead of having a serial queues, we can now combine the placement crew with their equipments (such as concrete pump, or crane, with bucket or hoist) as one unit. We would like to optimize the number of servers, which is the number of concrete pumps (or cranes or hoists). Our optimization problem becomes how many concrete pumps (and eventually the crew) we would like to provide such that we can accomplish the concreting activities at minimum cost. Let us assume, just at the moment, that the number of crew and the number of concrete trucks that can be hired are unlimited and the space for the concrete trucks and concrete pumps are also unlimited. Later we will visit and release these assumptions.

If we provide too many concrete pumps, we may accomplish the concreting activities faster but it is also at higher cost of renting the concrete pumps and hiring the crew. In other words, by adding the number of servers, the queuing system incurs higher server cost.

If we provide too few concrete pumps, we may think that the total cost of the queuing system will be lower due to lower cost. However, when we provide less number of servers, the delay of the concrete trucks will be more than necessary. If the waiting time of the concrete trucks is too long, the concrete will be hardened and the overall concreting activities will be delayed and the overall cost will be even higher. Thus, at less number of servers, the queuing system incurs higher customers cost.

As we think in term of system, the total cost of queuing system must include both server side and customer side. Optimal situation happens when both concrete trucks and concrete pumps would be minimal. Let us give notation C_s to a *constant* unit server cost which includes the renting of one unit of concrete pump together with hiring cost of the crew to serve one concrete pump. Notation C_c indicates a unit customer cost *function* which includes the waiting cost of one concrete truck per hour. It should be noted that C_c is a function of time rather than a constant, because the unit cost is higher significantly when the waiting time is longer. Since the behavior of this function over time is gradual increase at low waiting time and sudden increase until infinity at higher waiting time (due to hardening of the concrete), an exponential function with shape parameter beta will serve the purpose of this function. Parameter alpha is used to scale the value of the cost linearly.

$$c_c = \alpha \exp(\beta t) \tag{2}$$

To find the value of the parameters, we need to use non-linear regression. Taking the natural logarithm of both side of Equation 2 produces a simple linear regression equation $\ln c_c = \ln \alpha + \beta t$. If we have at least two points to calibrate the regression, the parameter values can be easily found. For instance, we have range of time t in hour. We would consider the customer cost to be very high (say, \$10,000 or more) if the waiting time is more than 3 hours due to risk of hardening of the concrete. If the waiting time is 30 minutes, the customer cost would be \$1000. Inputting points (0.5, \$1000) and (3.0, \$10,000) into the linear regression equation produces $\alpha = 630.96$ and $\beta = 0.92$.

Then, the total cost that would be minimized is computed as

$$C = s \cdot c_s + W \cdot c_c \tag{3}$$

While we can set schedule for the concrete trucks to arrive at certain regular interval, in practice however due to traffic condition, usually the arrival of the concrete trucks will be stochastic with inter arrival rate at the schedule time. Similarly, since the crew workability is now part of the queuing system, the service time to pour into cast and vibrate certain unit volume of concrete is also stochastic in nature.

Given the input of average and variation of inter arrival time and the service time to place a unit volume of concrete; we can compute the value of average waiting time W. The model to compute the waiting time will be discussed in the next sections of this paper. For a number of server s, we compute the total cost C based on equation (3). The optimum number of server is the one that minimize the total cost.

Case Study 2: Earth Moving

Carmichaela [7] explained the application of queuing theory for earth moving. The paper discussed the assumptions on the service discipline, on steady-state behavior and on the probability distributions for the service times. My treatment here is more qualitative in nature.

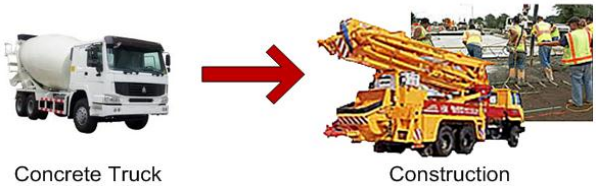

Figure 2. Concreting Activities as a Simple Queuing System.

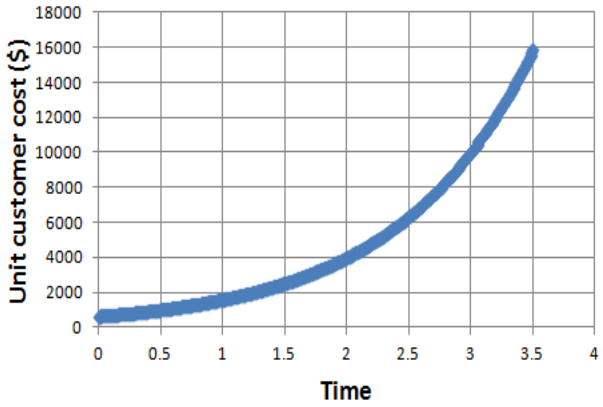

Figure 3. Exponential Function of Unit Customer Cost

In this second case study, we see earth moving activity as a queuing system. Earth moving activities have several agents: the excavators, the dump trucks and the loader (i.e. bulldozer). On the source of the quarry, the excavators cut the top soil and fill into the dump truck. The dump trucks then bring the soils to the construction site, dump the soils and the loader are ready to spread the soils to fill or to create the landscape of the land. We are faced with the same simple questions again: how to model earth moving activity as a queuing system? Which agents are the servers and which agent are the customers? What kind of optimization we would like to model through the queuing system? How many servers should we provide?

First, let us attempt to answer how to model earth moving activity as a queuing system. In earth moving activity, we may have three separate stakeholders: the quarry owner which operates the excavator, the transportation company which operates the dump truck and the contractor who operates the loader. From each stakeholder points of view, they may think that they provide the service to the customer. The quarry owner may think the excavator is the server and the dump truck is the customer. The trucking company will think that the dump truck is the server to serve the excavator and the loader. Similarly, the contractor may think the loader is the server to the dump truck.

It is also possible that the three stakeholders are actually one company. In this case, we will have system point of view. Looking at the flow of the earth moving activity, we have a serial queueing system Excavators ➔ Dump trucks ➔ Loaders.

If the distance between quarry and the construction site is relatively big that the traffic conditions may affect the order of arrival of the dump trucks, we may treat the serial queuing system as two separate queuing systems. First queuing system happens in the quarry where excavators are the server and the dump trucks are the customers. The dump trucks are waiting for the excavators to be filled. The second queuing system happens in the construction site where the loaders are the server to the dump trucks. The trucks are waiting for the loader to clear the filled soil before it can dump the soil to the next slot.

The large distance assumption between quarry and the construction site will greatly simplify our queuing systems. From this point of view, when we identify the most expensive agents as the servers (in term of rental fee per hour), it reveals that the dump trucks are the customers for both queuing systems.

To answer what kind of optimization we would like to model through the queuing system, let us assume for the moment that we can hire unlimited dump truck, and the space for the trucks to queue for both excavators in the quarry and loader in the construction site are also unlimited. Similar to the previous case study, we will visit and release these assumptions later.

Since both queuing system in the quarry and in the construction site has similar characteristics, for simplicity of the explanation, only the quarry site will be discussed.

Figure 4. Earth Moving Activities as a Serial Queuing System. Dump Truck is the customer for both Queuing Systems.

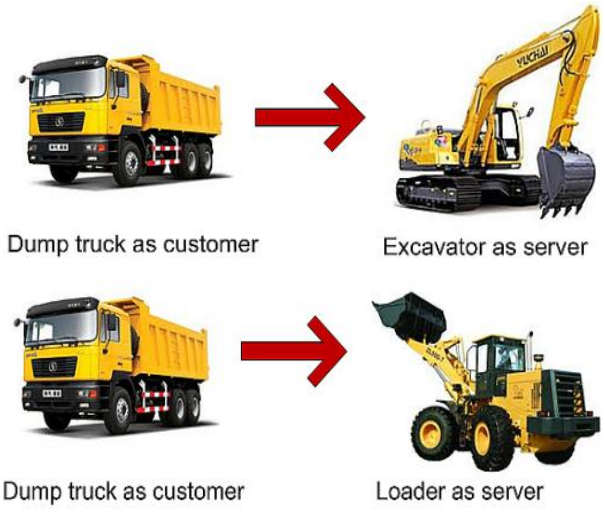

Figure 5. Earth Moving Activities as Two Simple Queuing Systems

If we provide too many excavators, we can finish the earth moving activity faster at higher cost of renting the excavators. Clearly, providing higher number of servers incurs higher server cost. On the other hand, providing too few excavators will create long queue for the trucks to wait and the overall earth moving activities will be delayed. Eventually, in this case the overall cost will be higher. Thus, providing lower number of servers incurs higher customers cost.

With similar reasoning to the first case study of concreting, we can use Equation 2 and Equation 3 to find the optimum number of servers (that is the number of excavators or the number of loaders) that will minimize the total cost of the queuing system. Note, in contrast to that in the literatures (such as Carmichaela [7]), exist what is called Griffis' Application of Queuing Theory to determine the number of trucks to perform earth moving activities. In this paper, we will not consider this application.

Queuing Models

Having the two case studies of concreting and earth moving, in this section we would like to have an integrated treatment of the two case studies. Even though the agents of the two case studies are different, they can be abstracted simply as servers and customers. Having this abstraction, we can now treat them as a simple queuing system.

The servers are characterized by service time distribution. In a simplified queuing theory, we can have either stochastic or deterministic service time. When we have stochastic service time and the variation of the service time is equal or almost equal to the average service time, we say that the service time is having or approaching Markovian distribution. In this case, the service time distribution fits into exponential distribution with mean μ is equal to the variance σ_s^2.

The customers are characterized by arrival distribution. When we have stochastic arrival distribution and the variation is equal or almost equal to the arrival rate, we say that arrival having or approaching Markovian distribution. In this case, the discrete arrival distribution fits into Poisson distribution with mean λ is equal to the variance σ_a^{-2}. Note in this case, the inter arrival rate λ^{-1} is simply an inverse of the arrival rate and the distribution of inter arrival rate would be exponential distribution.

The actual distributions of service time and arrival should be gathered and fit into the closest theoretical statistical distribution and then the appropriate formulas for queuing theory will be used to predict the performance of the queuing system. If the appropriate formulas of the queuing theory for the proper distribution are not available, we need to build simulation model. In practice, however, we often want to simplify this process because building a specific simulation model may require some cost on itself. When the mean is equal to variance for both distributions of service time and arrival, we can use most often used queuing formulas. The Kendal notation of such queuing system would be M/M/s, where the first letter is to indicate arrival distribution, the second letter is to indicate the service time distribution, and the third letter represents the number of servers. For this type of queuing system, we can compute the performance of the queuing system in term of the average number of customer in the system (in waiting line and being served) as:

$$L = L_q + \rho = \frac{P_0\ \rho^{s+1}}{(s-1)!(s-\rho)^2} + \rho \tag{4}$$

where,

$$P_0 = \left(\sum_{i=0}^{s-1} \frac{\rho^i}{i!} + \frac{\rho^s}{s!} \left(\frac{s\mu}{s\mu - \lambda} \right) \right)^{-1} \tag{5}$$

is defined as probability that there is no customer in the system. The ratio of arrival rate and service rate is given as:

$$\rho = \frac{\lambda}{\mu} \tag{6}$$

The average time a customer spends in the system in waiting line and being served is computed using Little's Law

$$W = \frac{L}{\lambda} \tag{7}$$

Now if the mean is not equal to the variance, we need more general type of queuing system, which Kendal notation would be G/G/s. Unfortunately, the formula of the queuing performance for general type queuing system does not exist yet. Only the approximation of G/G/s queuing system is available (Allen & Cunneen's as cited by Hall [6]). Given the coefficient of variation for inter-arrival time C_a and coefficient of variation for service time C_s, we can compute the average customers in waiting line waiting for service as:

$$L_q(G/G/s) = L_q(M/M/s) \cdot \left[\frac{c_a^2 + c_s^2}{2} \right] \tag{8}$$

The average queue length follows the first part of Equation 4 and the average waiting time follows the Little's law in Equation 7.

Earlier in the first case study, we have assumed unrealistically that the space for the concrete trucks

and concrete pumps are also unlimited. Similarly, in the second case study, it was assumed unrealistically that the space for the dump trucks to queue for both excavators in the quarry and loader in the construction site are also unlimited. We can release these assumptions by setting the limiting capacity on the number of customers that can be accommodated in the queuing system. The Kendal notation of this type of queuing system is M/M/s/N, where the last letter indicates the capacity of customers that can enter the queuing system. This type of queuing also means that the trucks that are not allowed to enter the quarry or construction site when the space capacity has been reached.

Another unrealistic assumption involves unlimited number of placement & vibrating crew, the number of concrete trucks or dump truck that can be hired. We can release this kind of assumption on the size of the calling source by limiting the number of customers. The Kendal notation of this type of queuing system is M/M/s/N/N where the last letter indicates the size of the customers.

With the removal of unrealistic assumptions above, we have discussed fully on how to model the examples of construction management case studies into queueing system. The next section will describe simple analysis based on the formulas above.

Queuing Analysis

In this section, I will give numerical illustration of applying the queuing formulas above for the first case study of concreting. The second case study bears similar technique, and therefore shall not be repeated.

In this analysis, we would like to answer two questions on:

What is the typical schedule of the concrete trucks such that the concrete pumps will not get idle 90% of the time?

What is the optimum number of server?

To make the problem more quantitative, the typical values are given as follow. Suppose we would like to cast concrete of one floor of a building with volume of 1500 cubic meters. As typical concrete truck can carry about 6 cubic meter, the activity requires about 250 trucks. A concrete pump has typical capacity of g = 50 cubic meters per hour. Thus, one concrete pump can serve about μ = 8 concrete trucks/hour. The concreting activity of one floor may require 30 concrete pumps to be finished in an hour, or 30 hours using only single concrete pump. Renting cost of a concrete pump is \$150/hour and labor cost to place and vibrate 50 cubic meters of concrete in an hour is about \$100/hour. This gives unit server cost of C_s = \$250/hour. The unit customer cost follows the previous explanation of using exponential function of Equation 2.

To answer the first question of truck schedule, we use Equations 5 and 6 by inputting μ = 8 concrete trucks/hour for various number of server s and customer arrival rate λ such that the idle probability P_0 is less than or equal to 10%. It should be noted that within queuing theory, there is a constraint that the utilization of the queuing system must not be larger than one. If the utilization is larger than one, the queue length and waiting time would be at infinity.

$$U = \frac{\rho}{s} = \frac{\lambda}{s\mu} < 1 \tag{9}$$

Since Equation 5 involves summation with no closed form, it is simpler to build a simulation program than to solve it mathematically. Thus, I made a simple computer program to solve the problem above. Figure 6 shows the result of relationship between idle probability and customer arrival rate. The horizontal line indicates the threshold probability of 10%. The horizontal line intersects the curve at 18.5 trucks per hour. It means that to ensure the concrete pump will not be idle 90% of the time, the concrete trucks should arrive every 3.25 minutes.

The analysis also showed that the minimum number of servers is 5 concrete pumps with average queue length of 2.4 trucks and average waiting time of 7.8 minutes. It should be noted that these last three numbers were obtained without even use the cost values. The minimum number of servers is obtained based on constraint Equation 9 and idle probability (Equation 5) by applying arrival rate of 18.5 trucks per hour and service rate of 8 trucks per hour such that the idle probability is about 10% (we use absolute difference between idle probability and the threshold should be less than a very small positive value).

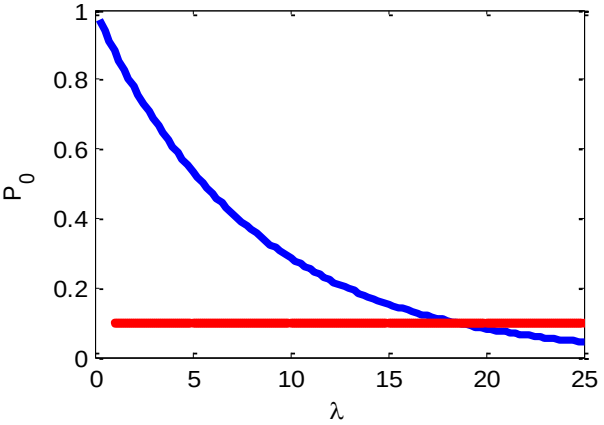

Figure 6. Relationship Between Arrival Rate and Idle Probability of the First Case Study

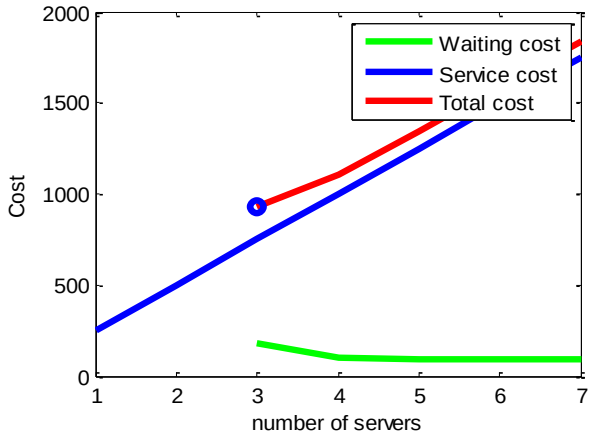

Figure 7. Economic Analysis of Queueing System

Once the idle probability is known, the average queue length and average waiting time follows Equations 4 and 7. Readers who are interested to find the effect of variation may use Equations 8 and 7.

Incorporating the costs Equations 2 and 3 using the same arrival rate of 18.5 trucks per hour and service rate of 8 trucks per hour produces minimum number of servers of 3 concrete pumps with average queue length of 4.33 trucks and average waiting time of 14 minutes. The idle probability is only 6%, and therefore, this is the better solution. Figure 7 shows that the customer cost is much lower than the server cost and therefore the optimum number of server follow the minimum number of servers is following only Equation 9.

This result gives implication that in case the customer cost is much lower than the server cost, the queuing system can be simplified into constrain equation. In other words, we can use the constraint Equation 9 as queuing rule of thumb.

Queuing Rule of Thumb

Most queuing books offer rather complicated mathematical formulas to compute queuing performance similar to equation (4) and (5). While computing these formulas using computer is very straightforward, in practice in the field, engineers often need much simpler formula which does not give accurate results but safe enough for the design. The formula is based on rewriting of the constraint equation (9). The queuing rule of thumb formula I present here has benefit of simplicity. It is so simple that people can even memorize it.

$$s > \frac{N\tau}{T} \qquad (10)$$

Where
T = total time to serve N customers
s = number of servers
N = number of customers
τ = service time

Compared to the actual queuing formula, the proposed queuing rule provides only very rough approximation. The purpose of the rule of thumbs is not to gain precision or optimization. The aim is to gain understanding of the current queuing situation and to give the layman simple tools to solve queuing problem more creatively. The queuing rule of thumb is not the replacement of standard practice and the queuing theory itself. The short survey should only be used as feasibility tool towards more precise standard practice.

In term of performance, the rule of thumb is rather conservative compared to the standard stochastic queuing formula because it is assumed that all the N customers arrive at once in the beginning of the service.

Conclusions and Summary

How to model the examples of construction management case studies into queueing system has been discussed. Through the numerical examples, it was shown that when the customer cost is much lower than the server cost, queuing system can be simplified only to incorporate the constraint equation. I suggest to the constraint equation as queuing rule of thumb when we deal with simplified queuing theory.

To anticipate many queuing problems, we can estimate the number of servers and service performance based on demand that arrive at once and its service time without complicated formulas. The key to solve most queuing problem is on the modeling customers and servers.

Through case studies of concreting and earth moving activities, I will like to encourage Civil Engineers and Construction Managers to use the vast knowledge of queueing theory that have been developed by mathematicians and management scientists for about a century.

References

1. Hendrickson, C. and Tung, A., *Project Management for Construction - Fundamental Concepts for Owners, Engineers, Architects and Builders*, Prentice Hall, 1998. http://pmbook.ce.cmu.edu/04_labor_material,_and_equipment_utilization.html.
2. Oberguggenberger, M., *Queueing Models with Fuzzy Data in Construction Management in: W. Fellin et al (Editors) Analyzing Uncertainty in Civil Engineering*, Springer, 2004. pp. 197-208.

http://www.globalspec.com/reference/34589/203279/ queueing-models-with-fuzzy-data-in-construction-management,

3. Cheng, M.Y., Tsai, H.C., and Lai, Y.Y., *Construction Management Process Reengineering Performance Measurements, Automation in Construction 18*, 2009, pp.183-193. http://www.ct.ntust.edu.tw/ct/filesConstruction%20management%20process%20reengineering%20performance%20measurements. pdf. See also: http://140.118.5.71/bpr%20files/Bpr_ Evaluation.pdf.
4. Tian, N. and Zhang, Z.G., *Vacation Queueing Theory Model – Theory and Application*, Springer, 2006.
5. Yue, W. et al., *Advances in Queueing Theory and Network Applications*, Springer, 2009.
6. Hall, R.W., *Queueing Methods for Services and Manufacturing*, Prentice Hall, 1991.
7. Carmichaela, D.G., Shovel–truck Queues: a Reconciliation of Theory and Practice, *Construction Management and Economics*, 4(2), 1986, pp.161-177. http://www.tandfonline.com/doi/abs/10.1080/01446198600000013.