Optimizing appointment driven systems via IPA

with applications to health care systems

BMI Paper

Aschwin Parmessar

VU University Amsterdam

Faculty of Sciences

De Boelelaan 1081a

1081 HV Amsterdam

September 2010

Preface

Part of the Master's program in Business Mathematics and Informatics is writing a paper on a subject chosen by the student. The topic, however, must cover a part of the study.

During my study I became interested in Optimization of Business Processes and Stochastic Optimization. That is when I came into touch with Professor Sandjai Bhulai. He came up with the idea to study how to plan the arrivals of patients at general practitioners and other similar systems effectively by using a simulation technique called Infinitesimal Perturbation Analysis.

In this paper you can find a way to plan the arrivals of patients or customers in these type of systems. The discoveries in this paper are applicable to many different fields as long as there are appointment-driven schedules. Another example would be scheduling in GRID systems, in this case a customer is assigned a specific time slot by a computer for a certain service.

Furthermore, I would like to thank Professor Sandjai Bhulai for his support in the research.

Finally, I hope the reader will enjoy this paper.

Amsterdam, September 2010

Parmessar, Aschwin

Executive summary

In this paper we will look at appointment driven systems and try to analyze them with a simulation technique called IPA, Infinitesimal Perturbation Approximation. This technique takes one schedule and tries to come up with one which has a lower expected waiting time of the customers. Furthermore we will also look at the expected waiting time of the server, also known as the idle time, because most of the times it is expensive to have the system idle for a long time.

Also a comparison will be made with the Bailey-Welch rule, a rule what says that two customers should be planned at the beginning of the day to minimize the expected waiting time of the server. Since IPA and the Bailey-Welch try to minimize two different things we will also look in this paper at the average of both expected waiting times to check which results would be better if the waiting time for customers would be as expensive as the waiting time of the server.

IPA gives us better numbers for the expected waiting time of the customers and the following rules of thumb can be derived from our results:

1. *Plan customers slightly before the expected finishing time of the previous customer at the beginning of the day and slightly later at the end of the day.*
2. *Customers with larger variances in the arrival time should be planned as late as possible.*
3. *Customers with a larger probability of not showing up should be planned as late as possible.*
4. *Customers with larger variances in the service time should be planned as late as possible.*
5. *Customers with larger expectations of the service time should be planned as late as possible.*

The Bailey-Welch rule gives better numbers for the expected waiting time of the server as expected.

The best averages of waiting times are derived from IPA, so it might be beneficial to look at the costs of waiting for the customers and compare these to the costs of waiting for the server. If the latter are lower than or equal to the costs of waiting for the customers than it would be beneficial to apply a schedule like the one used in this paper.

Furthermore, it should be researched if the IPA-technique can be applied on the average waiting time of both the customer and the server simultaneously.

Table of Contents

1 Introduction

In the last decade companies have been trying to work more efficiently and effectively. However, for most companies there is still room to work more effectively, especially in Health Care. Health Care is already one of the biggest industries worldwide and it is still growing according to Wikipedia[1]. According to the American Bureau of Labor Statistics[2], ten of the twenty fastest growing occupations are healthcare related. This happens largely in response to the rapid growth in elderly population. Since almost everybody has health insurance, everybody pays for the healthcare industry. So it is extremely important to organize healthcare in an efficient way. To be able to do this a lot of research is needed. A lot of research is already done in Health Care. However, there is a lot that we still have not investigated yet. This paper will cover such a subject.

In general practitioner offices patients in need of consultancy or treatment come in all day. Most of them come with an appointment, but even with an appointment a lot of patients still have to wait when they arrive on the agreed time. To give a feeling about the waiting time, look at the following example: If a general practitioner has 25 patients arriving on a day and they all have to wait for five minutes, then all the customers combined would have to wait together 125 minutes. That is only for one general practitioner. There are approximately 9000 general practitioners in the Netherlands. If they all let their patients wait for 125 minutes then all the patients in the Netherlands would have to wait approximately 18750 hours per day. This is a lot of hours waiting for the patients that could be used in a meaningful way if these patients would not have to wait. However, in practice, not only the patients have to wait but also the general practitioner.

The reason that the patients and the general practitioners have to wait is because of the randomness of the arrival times of patients, the show-up percentage of patients and service durations. There are already some models that are able to deal with the randomness of the service durations and the occurrence of no-shows, this happens when a patient does not come to the general practitioner's office. It becomes a problem when patients are arriving with appointments. Most models deal with patients arriving on random times, because these are easier to solve. For arrivals with appointments the system is difficult to model, and it becomes impossible to solve mathematically. These systems are called appointment driven systems.

Let us look at an example just to give an idea why this model is so hard to solve: In a certain general practitioner office 25 patients arrive on one day. To be able to treat as many patients possible and to keep the general practitioner from waiting all patients are scheduled to arrive just after the expected end time of the previous patient. If one patient now arrives five minutes too late all the patients that will arrive later than this patient are expected to wait at least five minutes. However if one of the later patients does not show up than the general practitioner would have to wait some time and then the next patient can be helped immediately. A delay can also come from an unexpected long service time of a patient, let us say the service time takes ten minutes longer than expected. Then again we expect all patients that arrive later than this patient to wait ten minutes. This is what happens if there is only one deviation from the planning. Consider now that in reality it is more likely that multiple deviations will occur. Some patients will arrive too late, some will not show up and

[1] See http://en.wikipedia.org/wiki/Health_care_industry#Growth
[2] See http://www.bls.gov/oco/cg/cgs035.htm

others will have unexpectedly long waiting times. Then it is hard to try to predict what will happen and how to plan effectively.

Currently the best solution is the Bailey-Welch rule. This rule says that if all patients have the same waiting time distribution it is best to plan two patients at the beginning of the day and plan the other patients at intervals equal to the average service time. The reason this is done is to balance the waiting time of the patients and the idle time of the general practitioner.

As you can see part of planning this system is to balance the general practitioner's waiting time and the patient's waiting time.

The focus of this paper will be on studying the planning of appointment driven system by using a simulation technique called Infinitesimal Perturbation Analysis, also known as IPA. IPA allows us to come up with (sub-) optimal solutions of difficult models, like the appointment driven system. With IPA we will try to come up with some planning rules of thumb. Examples of these rules might be to plan patients at the same time if they both have a large probability of not showing up, think of overbooking in the airline industry where there are more seats sold than there are on the airplane. In this paper only rules of thumb for planning will be given, not the optimal way of planning. Again the reason for this is that the system is too complicated to come up with exact solutions. However, with some good rules of thumb it is still possible to come up with a more efficient planning than currently is used. To check if IPA gives us better solutions we will check some currently used planning methods like the Bailey-Welch rule.

Appointment driven systems are not only applicable to the general practitioner planning problem, but to many other fields too. The only requirements for other applicable fields is that they have to deal with appointment driven arrivals in need of a service, like the dental practice, GRID systems or even planning systems for managers where the managers might have a part of the week reserved for talking with the employees individually. However, the Health Care industry is currently the most important field, since it is such a large part of the industry and since almost everybody of us pays for the Health Care and thus also pays for the inefficiency. To keep this paper as general as possible we will only talk about customers who are in need of a service, but keep in mind that these customers can also be patients who need treatments. Furthermore we will also only talk about servers and again a practitioner can also be seen as a server, a thing or person which is able to provide services to customers.

This paper is constructed as follows: First of we start by describing the queuing system, the single server model, in Section 2.1. Next, appointment driven arrivals in queuing systems will be discussed in Section 2.2. The IPA technique will be explained in Section 2.3. In Section 3 the experiments will be described. And in Section 4 we will discuss the results. Finally, the paper is wrapped up with a conclusion in Section 5.

2 Theory

2.1 Queuing models, the single server model

Queuing models deal with systems where customers arrive with a certain rate and need service from the server. In this research we will focus on a specialization of the single server model, which has, as the name mentions, only one server. In this paper customers do not arrive with a certain rate but with appointments and the exact time of arrival is decided by the time of the appointment and some distribution.

Before we start explaining this special single server model we will first look at what can be obtained by looking at these models and what we are interested in.

Queuing models are used to model real systems. The reason this is done is most of the times queuing models are less complex than the original system and thus it is easier to obtain the necessary parameters. The parameters we are most interested in are the total waiting time for the customers over a day, denoted as EW_{cust}, and the total idle time of the server over a day, denoted as EW_{server}.

In the simpler forms of queuing systems there are formulas for these expressions. However in this paper a more complex version of the single server model is being used, so we are forced to calculate the necessary values through simulation.

Now we know what kind of parameters we want from the queuing model we can look at the single server model with appointment driven arrivals.

2.2 Appointment driven arrivals

As already mentioned, in this paper appointment driven arrivals will be applied to the single server model. These arrivals can only arrive during the operating times of the server. It is common for practitioners to work eight or nine hours a day, so we will say that servers also work about eight or nine hours a day. Customers are allowed to arrive only during this time period. Let us say that the server is open for only T hours.

Since we are considering appointment driven arrivals we know the number of scheduled arrivals in advance and define M as the number of customers that are scheduled for one day. Now let us look at the n-th customer. Let d_n, for n = 1, … , M, denote the appointment time of the n-th customer. We assume that customers are scheduled to arrive in order of their index number by assuming $d_n \leq d_m$ if $n< m$. Remember that we also have to deal with no-shows, the event that a certain customer does not show up. Define p_n as the probability that customer n shows up and thus $1-p_n$ is the probability this customer does not show up. So it is possible for a customer not to show up, but we also have to deal with this in our system. Since we do not know upfront if a customer will show up, we will say a customer can show up between a certain time interval, say between $d_n - \tau_n^l$ and $d_n + \tau_n^u$. Denote by D_n the general distribution function of the arrival of customer n. This distribution function has to take

in account the interval we discussed before. Denote e_n as number drawn from D_n, which is the arrival time of a customer. Note that we assumed that customer n is scheduled to arrive before customer n+1, for all n. However we will not make the assumption that $e_n < e_{n+1}$. This means that customer n does not have to arrive before customer n+1, even though he is scheduled to arrive before customer n. We do this to keep the model as realistic as possible. So customer n+1 might have to wait before customer n arrives and is being served. Unfortunately it is not doable to let customer n+1 to be served before customer n, not only because this is not realistic, but also because the model becomes too difficult to describe mathematically. Next we will look at the service duration of this customer, in most fields the exponential distribution is used for the service duration with mean service $1/\mu_n$. However in this paper we will also include other distributions, e.g., the normal distribution. The moment at which customer n is done with its service will be denoted by s_n. This is the moment at which customer n+1 can go into service if the customer has already arrived at the system.
In Figure 1 you can see the overview of customer n.

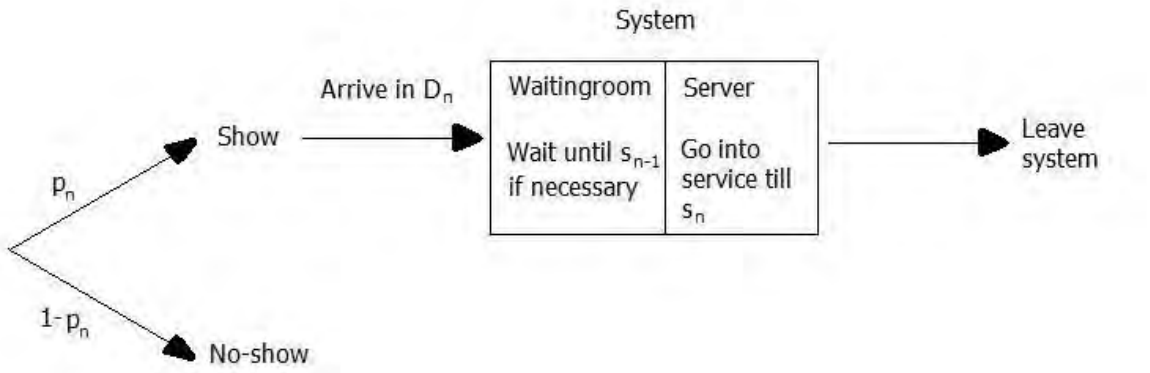


Figure 1: an overview of the route of customer n

Again we are interested in EW_{cust} and EW_{server}. However, now it is very hard to give a precise mathematical expression for EW, since we have extended the standard M/M/1 model by not only allowing all kinds of distributions, but also no-shows and appointment driven arrivals. Luckily, it is still possible to calculate parameters with the help of simulation. Since we simulate the system, we can keep track of the simulated customers and their waiting times. When we take the average of these waiting times, we have a good estimate of both parameters.

For more information on Appointment driven arrivals see Jouini and Benjaafar (2010)[3].

[3] See Chapter 7 "Literature" more information

2.3 Infinitesimal Perturbation analysis

2.3.1 What is Infinitesimal Perturbation analysis?

Infinitesimal Perturbation Analysis, also known as IPA, is a simulation technique used for stochastic discrete event systems. This means that the state of the system will only change at discrete, random time instants. Examples include queuing systems, like the single server model described in Section 2.1.

The benefit of using simulation is that we are able to derive satisfying (sub-)optimal solutions without the use of a mathematical derivation. The mathematical derivation gives us the benefit of having the global optimal solution, but also has the disadvantage that it is hard to derive in practice. Most mathematical proofs need assumptions resulting in a perfect world, e.g., customers will arrive in time for a service. So we will use simulation to come up with a solution that is optimal or nearly optimal.

The reason IPA is used in this research is because most of the simulation techniques need to analyze all parameters individually. This means that almost only small trivial problems can be solved. For larger systems it will take too long to analyze the system. This is where IPA enters. IPA looks at the initial state and changes every system parameter, i.e., the decision variable, a little. This allows us to approximate the derivative. When we have the derivative we can see which parameter is best to change and apply the change and use this as the initial state. Then we calculate the derivative again and again choose which parameter is best to update. We keep repeating the process until an (sub-)optimal solution is found. The difference with other simulation techniques is that we have a method for searching for an (sub)optimal value that works for these type of systems.

For more information on IPA see Glasserman(1991)[4].

2.3.2 Infinitesimal Perturbation analysis in detail

The IPA technique looks as follows. First, we start off with an initial state, say state Y. This initial state contains the planned arrival times, service durations and show-up probabilities of all the customers. For this state we calculate the expected waiting time for all customers, denote this as $EW_{cust}(Y)$, and the expected waiting time for the server, denoted as $EW_{server}(Y)$. The latter is also known as the idle time. Next, we will try to change all arrival parameters one by one by a small amount, say θ_1. Let us say we change the i-th parameter so Y will look like $Y + e_i * \theta_1$, where e_i is a vector with only zeros except for the i-th element which is equal to one, and calculate the expected waiting time for all the customers, $EW_{cust}(Y+e_i * \theta_1)$, and the expected waiting time for the service provider again. Now we can approximate the derivative to θ as follows:

$$dEW_{cust}/d\theta_1 = (EW_{cust}(Y+e_i * \theta_1) - EW(Y)) / \theta_1$$

This formula gives us the derivative for each parameter. Now choose the parameter with the highest derivative, say that this parameter has index i. Then i will denote the i-th arrival and tells us to update the arrival time.
We will update the state Y as follows: $Y = Y - e_i * a_n * dEW_{cust}/d\theta_1$. This allows us to move into a direction that minimizes $EW_{cust}(Y)$.
a_n is a variable to multiply the derivative with, also called the step size. In the beginning this number

[4] See Chapter 7 "Literature"

should be high such that it will push the system to its optimal state as quickly as possible. However over time a_n should decrease such that the algorithm converges to an optimal solution. To make sure a_n satisfies the first restriction, we demand that $\sum_{n=1}^{\infty}(a_n) = \infty$
To make sure a_n will respect the latter restriction, we need that $\sum_{n=1}^{\infty}(a_n)^2 < \infty$

A possibility for $a_n = 1/n$. a_n is dependent only on n. n stands for the number of iterations so far. So now we will have n=1.

Now we will repeat all the steps above with $\theta_n = \theta_{n-1} - a_{n-1}\, dEW_{cust}/d\theta_{n-1}$. Again n is the number of iterations. We replace all θ_1 with θ_n and $dEW_{cust}/d\theta_1$ becomes:

$dEW_{cust}/d\theta_n = (EW_{cust}(Y+e_i * \theta_n) - EW_{cust}(Y)) / (\theta_n - \theta_{n-1})$

Finally we will update Y again: $Y = Y - e_i * a_n * dEW_{cust}/d\theta_n$.
We repeat this process until θ hardly changes, say: $\theta_n - \theta_{n-1} < \epsilon$. Where Є is a small number like 10^-4. By repeating the process we can try derive to derive a better schedule than we had before with a lower $EW_{cust}(Y)$.

3 Experiment

3.1 Setting up the experiment

An easy way to set up the experiment is by using Crystal Ball, Excel and VBA (Visual Basic for Applications). Crystal Ball allows us to use the Monte-Carlo method. The Monte-Carlo method is a technique for simulation that allows us to repeatedly draw samples from distributions. We can use these samples for our experiment, like calculating the waiting time. When we draw multiple samples we will get an average for the expected waiting time. Crystal Ball is an add-in for Excel. We can set up the experiment in Excel and then draw samples from Crystal Ball. VBA gives us the possibility to let Excel run multiple iterations of the IPA technique. VBA is a programming language specialized for applications, in this case Excel.

3.2 Which experiments are done?

In this paper we will try to discover useful rules of thumb that can be applied to appointment driven systems, especially health-care related appointment driven systems. Due to the fact that in reality there is no perfect model, the experiment can be divided in partial experiments. The basic partial experiments are the following:

- The first partial experiment will cover only systems in which patient arrivals are random. The blocking probability for each customer will be zero in this experiment and the service times are fixed.
- The second experiment is the same as the first one except for the fact that the service time now is random and the arrival times are fixed.
- Third is the experiment where arrival and service times are fixed and the no-show probability will be non-zero.

From this the following more complex experiments can be created by combining two of the basic partial experiments. These problems are:

- This experiment is a combination of the first and second where arrival and services times are random and the no-show probability is zero.
- Next is the experiment where arrival times are random and the no-show probability is non-zero. The service times will be fixed in this case.
- The last combination of the two basic experiments is the one where service times are random, the no-show probability is non-zero and the arrival times are fixed.

Finally, there is a highly complex system where all three basic partial experiments are combined:

- The final experiment is the one where arrival and service times are random and the no-show probability is non-zero.

In this paper we will only focus on the most complex system. To come up with some rules of thumb we will try to answer the following questions:

1. How does more uncertainty in arrivals affect the system, i.e., how does a greater variance in arrivals affect the waiting time of the patients, EW_{cust}, and the waiting time of the server, EW_{server}.
2. How does the uncertainty in service times affect the system performance.

3. How does the show-up percentage affect the system performance.

Since we are trying to come up with rules of thumb which can be applied to scheduling of patients we will try to answer the system by just allowing one patient to deviate from all the other patients. This means that all patients except one will have the same profile: variance in deviation of the planned arrival time, variance in the service time and a show-up percentage. The patient left will be the one with a higher variance in deviation of the planned arrival time, higher variance in the service time and a different show-up percentage. This patient will be planned at different moments of the day, i.e., beginning of the day, and the results can be compared. From these results conclusions can be drawn and rules of thumbs can be given.

Next to that we will compare the results IPA gives us with the Bailey-Welch Rule. As mentioned before, the Bailey-Welch Rule will plan two patients at the beginning of the day and plan the other patients at intervals equal to the average service time. Remember that this rule is based on minimizing the idle time of the server. However in this paper we are mostly focused on minimizing waiting times of patients, therefore it will be hard to draw conclusions on comparing the experiments of this paper with the Bailey-Welch Rule.

Furthermore, we will also look at what happens when we look at a special scenario with rush-hour traffic in the morning. This will lead to an expected arrival time that will be later than the planned arrival time or to a larger variance in the arrival time. Since we already researched the latter we only need to look at the first, thus the case in which the expected arrival times are later than the planned arrival times.

Note that we again used the term patients. This is because the experiment is based on a general practitioner's office and to keep this link clear the term patients is used in this subsection. The customers are patients and the server is a practitioner. This practitioner works between eight and nine hours a day. Furthermore the practitioner is expecting patients to arrive during the day. From now on we will again talk only about customers and a server.

4 Results

4.1 General

In this chapter we will discuss the results of our experiments. Because of the size of the experiments we will first give an overview of the different types of experiments and explain the general set-up of the experiments.

All the different experiments have the same general set-up. For each type of experiments we will look at the following distributions for the arrivals:

- Uniform distribution
- Exponential distribution
- Triangular distribution
- Normal distribution

The service times have the following distributions:

- Uniform distribution
- Exponential distribution

Each customer will have the following no-show distribution:

- Bernoulli-distribution

A setting of an experiment looks as follows: there will be 25 customers. Of these, 24 will have exactly the same properties, i.e., the same type of distributions for arrivals and service times with the same parameters and the remaining customer will have the same type of distributions however one of the parameters will be different than the parameter of the other customers. The standard properties we will use in this research are:

- Expected service time of 20 minutes
- No-show probability is equal to 0.07[5]
- Service can only take twice as long as the average, so for standard customers the service can take up to 40 minutes
- Customers are allowed to arrive between the planned arrival time of the previous and the next customer, so standard customers can arrive at most 20 minutes earlier or later than planned.

Now we can study the effect of scheduling the deviating customer at the beginning of the day, at the middle of the day, and at the end of the day. Next to that we will also look at the experiment where all 25 customers have the same type of distributions and parameters.
Since IPA will try to improve the current schedule it makes sense to use a schedule which already provides us with a good number for EW_{cust}. This allows us to use fewer results and obtain either better or the same number for EW_{cust}. One schedule which seems to give a reasonable number for EW_{cust}, is the one where the planned interarrival times of customers are equal to the average times

[5] According to Annabel patients do not show up between 5% and 10% of the time for their appointments in hospitals. See: http://www.ac-outbound.nl/over-ons/sectoren/zorg-sector/

of the service duration of the previous customer. Now we have the settings for the experiments we can apply the IPA approach to the schedule. Due to the long running time we will keep the number of iterations fixed. So after a number of iterations we will have the new schedule with planned arrival times, EW_{cust} and EW_{server}. Also we will look at what happens if we apply the Bailey-Welch rule to the schedule. From this new schedule we again have new planned arrival times, EW_{cust} and EW_{server}. So we will also look at the differences between IPA and the Bailey-Welch rule. Finally, we will also look at the EW, the average of EW_{cust} and EW_{server}. We do this to compare IPA with the Bailey-Welch rule fairly. Remember that IPA will minimize EW_{cust} and the Bailey-Welch rule mainly minimizes EW_{server}.

We will have the following types of experiments:

- A larger variance in arrivals will be discussed in Section 4.2
- A higher probability of the no-show probability will be treated in Section 4.3
- Section 4.4 will cover a larger variance in service times
- A larger expectation of service times is covered in Section 4.5
- Section 4.6 will cover a special topic, there we will look at what happens when customers are expected to arrive late

4.2 Larger variance in arrivals

In this section we will look at what happens when there is one customer with a larger variance in arrivals.

4.2.1 Expected waiting time of customers

First let us look at the expected waiting time of the customer: EW_{cust}. In Figure 2 we can see what happens to EW_{cust} if one customer has a larger variance in the arrival time for the exponential distribution. We can see that EW_{cust} will slightly decrease if the customer with a larger variance is placed later on the day. This holds for both IPA and the Bailey-Welch rule. However, if the customer with a larger variance is placed at the beginning of the day, the Bailey-Welch rule gives us a higher EW_{cust} than if there was no customer with a larger arrival time. IPA, however, seems to be capable of keeping EW_{cust} almost the same when dealing with a customer with a larger variance in his arrival time who is scheduled to arrive at the beginning of the day compared to when this same customer has the same variance as all the other customers.

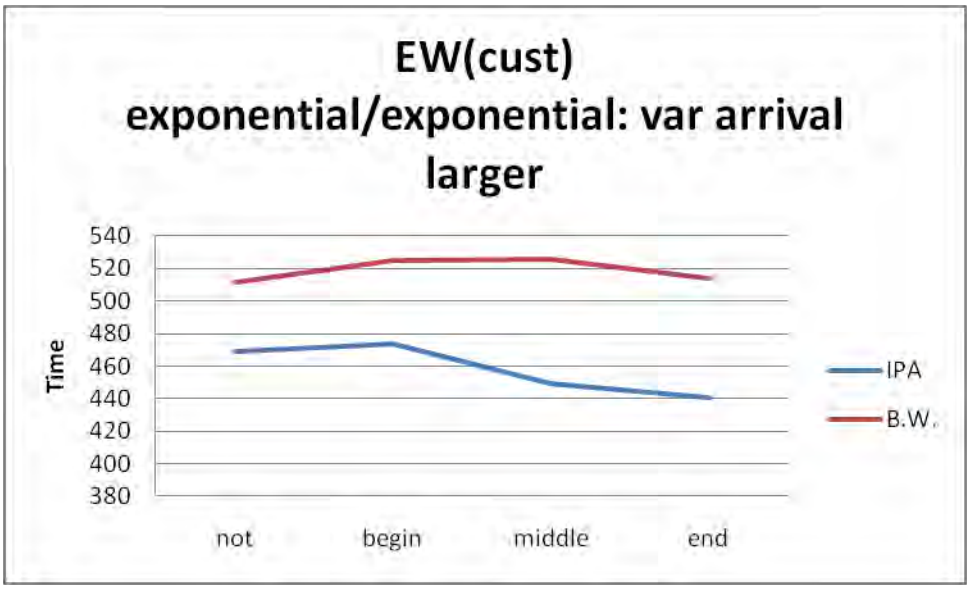


Figure 2: EW_{cust} with a larger variance in arrival times

In Appendix 1 we can see that this also holds for arrivals with other types of distributions. However, there are some discrepancies: when we look at arrivals with a triangular distribution or a normal distribution we can see that the customer scheduled at the end of the day and the customer scheduled at the middle of the day have a larger EW_{cus} for IPA, respectively. This is probably because these numbers have not yet converged.
Finally, we can see that IPA gives us lower numbers for EW_{cust} than the Bailey-Welch rule.

4.2.2 Expected waiting time of the server

Next we will look at EW_{server}. In Figure 3 we can see that for an exponential distribution in arrivals this number increases when there is a customer with a larger variance.

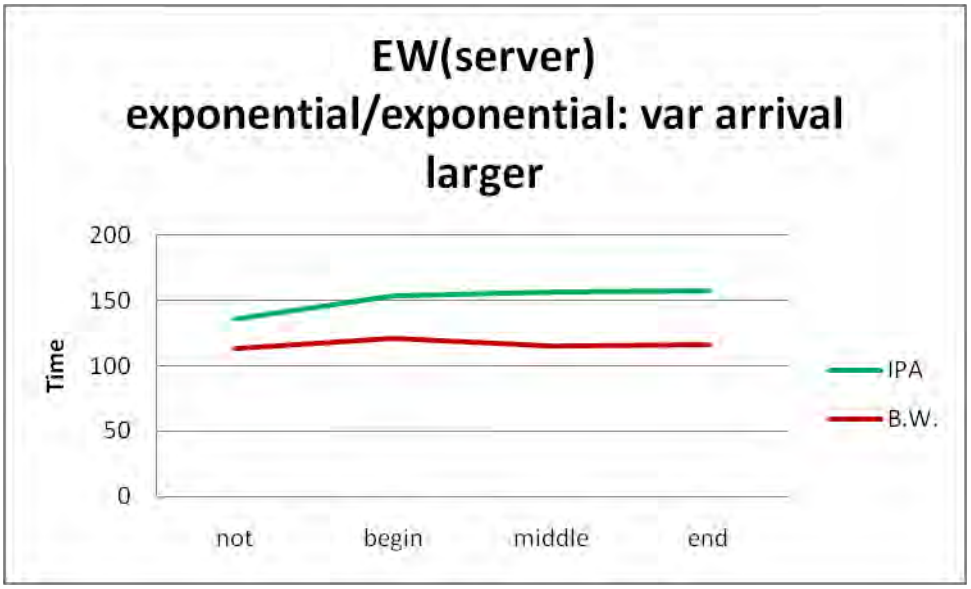


Figure 3: EW_{server} with a larger variance in arrival times

In Appendix 2 we can see it is not always true that when we have a larger variance in arrivals EW_{server} will also increase, you can see this by looking at arrivals with a normal or a triangular distribution. So

it does not mean that when the variance in arrivals increases EW_{server} also increases, at least not for all distributions.
However what we can see is that IPA gives us higher numbers for EW_{server} than the Bailey-Welch rule. This does seem to hold for all distributions and is also quite logical because this is what the Bailey-Welch rule minimizes.

4.2.3 Average expected waiting time of both the customer and the server

To fairly compare the results of IPA and the Bailey-Welch rule we can look at EW, the average expected waiting time for both the customer and the server. In Figure 4 we can see the results for the exponential distribution again. We can see that IPA gives us a lower EW than the Bailey-Welch rule.

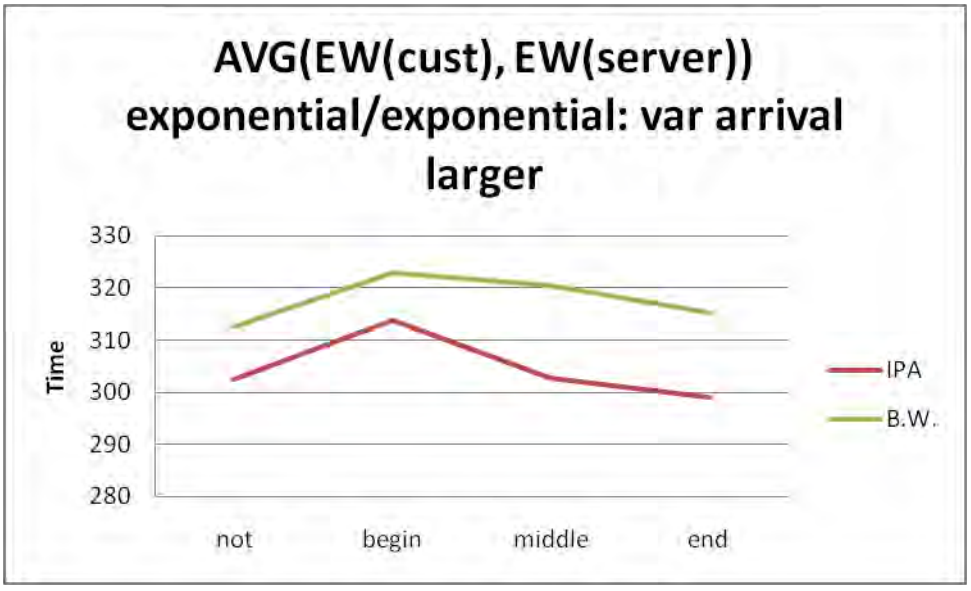


Figure 4: EW with a larger variance in arrival times

In Appendix 3 we can see that IPA gives us equal or better results when handling with a larger variance in the arrival time.

4.3 Higher probabilities of no-shows

In this section we will look at what happens when a customer has a greater probability of not showing up.

4.3.1 Expected waiting time of customers

Let us look at what will happen with EW_{cust} when a customer has a greater probability of not showing up. In Figure 5 we have an overview of EW_{cust} under exponentially distributed arrivals and service times. We again see that IPA gives us lower waiting times for customers than the Bailey-Welch rule. Also we can see that EW_{cust} decreases when the customer has a higher probability of not showing up.

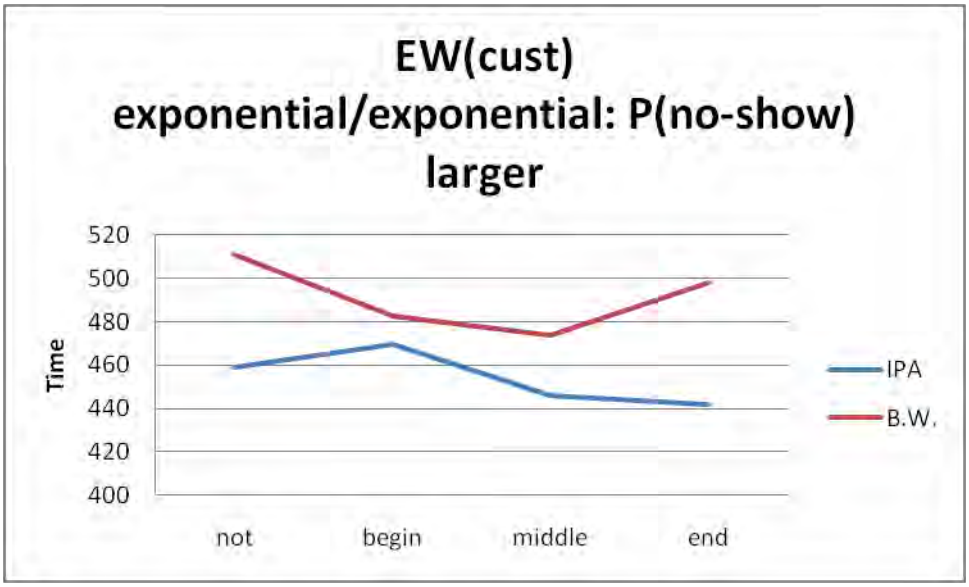


Figure 5: EW_{cust} with a larger probability of no-shows

However, in Appendix 4 you can see for triangular distributed arrivals and exponential service times, that when a customer with a higher probability of not showing is planned at the end of the day, it might not give the best results. It might be because the number of iterations still was not large enough.

Finally, we see that IPA also gives us better results for EW_{cust} than the Bailey-Welch rule when dealing with a greater probability of no-shows.

4.3.2 Expected waiting time of the server

Next we will look at what happens with EW_{server}. In Figure 6 we see EW_{server} for exponential arrivals and exponential service times. We can see that a higher probability of no-shows slightly increases EW_{cust}.

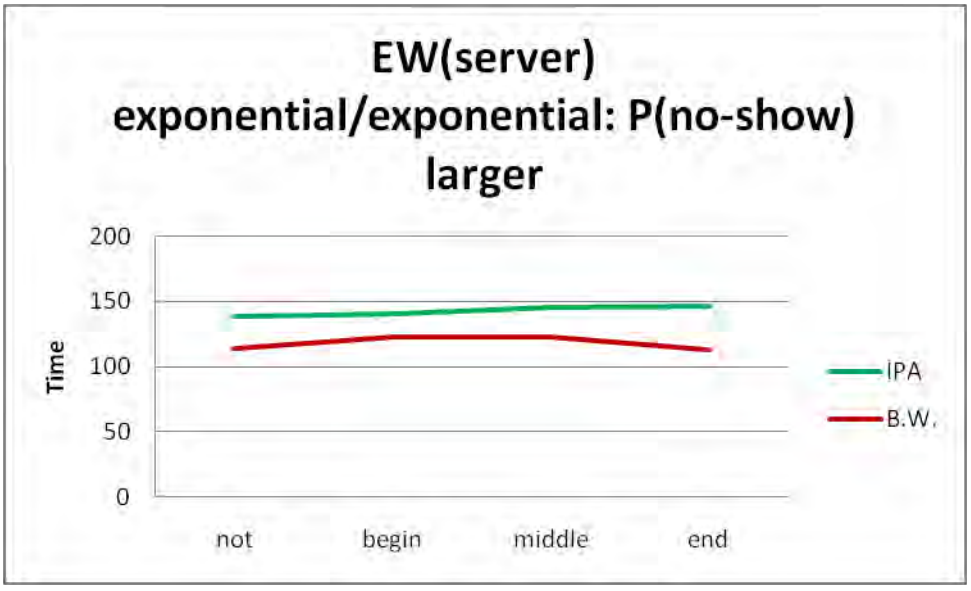


Figure 6: EW_{server} with a larger probability of no-shows

We can see in Appendix 5 that also for other distributions of arrivals and service times that a higher probability of no-shows negatively affects EW_{server}. This seems logical because the server has planned a certain time interval for a customer and now he will likely have to wait for the next customer.

4.3.3 Average expected waiting time of both the customer and the server

Finally, we will look at the average waiting time of both the customer and the server. This is shown for exponential arrivals and service times in Figure 7. We can see that IPA again gives us better average waiting times.

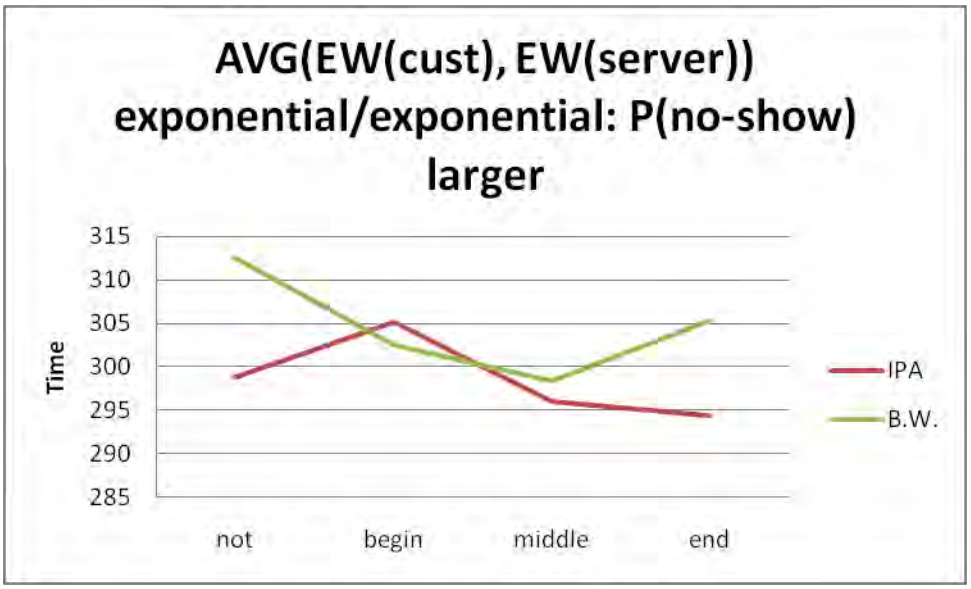


Figure 7: EW with a larger probability of no-shows

IPA also provides us most of the times better numbers EW, this can be seen in Appendix 6. However the difference of EW between IPA and the Bailey-Welch rule might not be always as large as in Figure 6.

4.4 Larger variance in service times

Next, we will look at what happens when a certain customer has a larger variance in service times. In this section we will only look at distributions that have different parameters for the mean and the variance, which is only the uniform distribution in this experiment. The exponential distribution cannot have a larger variance without changing the expected service time, which will be covered in the next section.

4.4.1 Expected waiting time of customers

Again we will start to look at EW_{cust}, which can be seen in Figure 8. Figure 8 shows the results of EW_{cust} under uniformly distributed arrivals and service times. We can see that a customer with a larger variance in service times should be planned as late as possible to minimize EW_{cust}. IPA is able to slightly improve EW_{cust} when dealing with a larger variance. However for the Bailey-Welch rule EW_{cust} slightly increases when there is a customer with a larger variance in the service time.

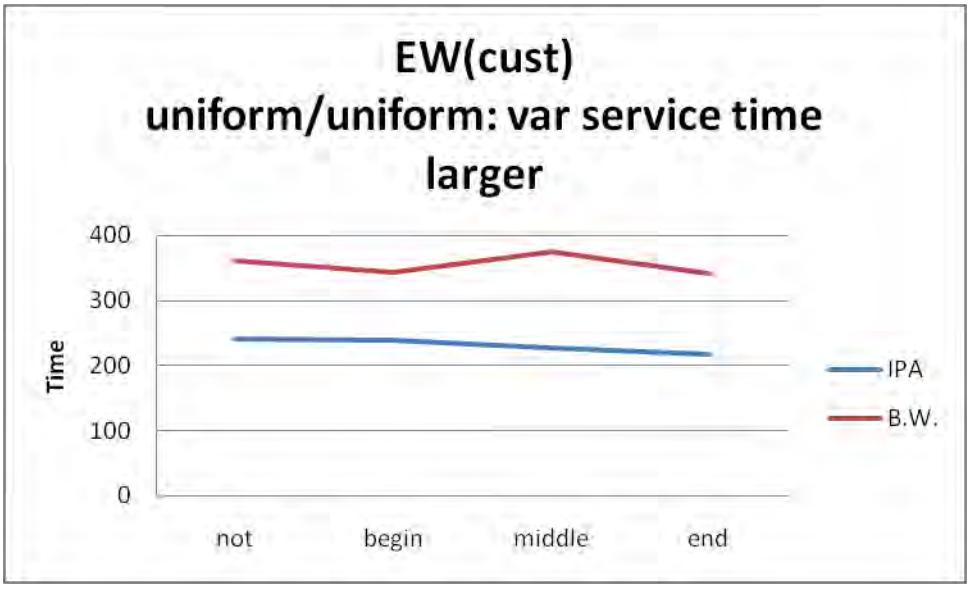


Figure 8: EW_{cust} with a larger variance in service times

4.4.2 Expected waiting time of the server

Next up is the effect of a larger variance in service times on the expected waiting time of the server, EW_{server}. These numbers are shown in Figure 9 for uniform arrivals and service times. When looking at IPA, we can see that scheduling a customer with a larger variance in service times should not be planned at the beginning of the day. If this does happen and a customer has a short service time, the server has to wait a long time until the next customer arrives. We can now also see the effect of the Bailey-Welch rule clearly, by scheduling two customers at the beginning of the day, we are sure that if one service is ended quickly the server can already start with the next and there is still a reasonable amount of time left for another customer to arrive and thus EW_{server} is not affected much by having a customer with a larger variance in service time.

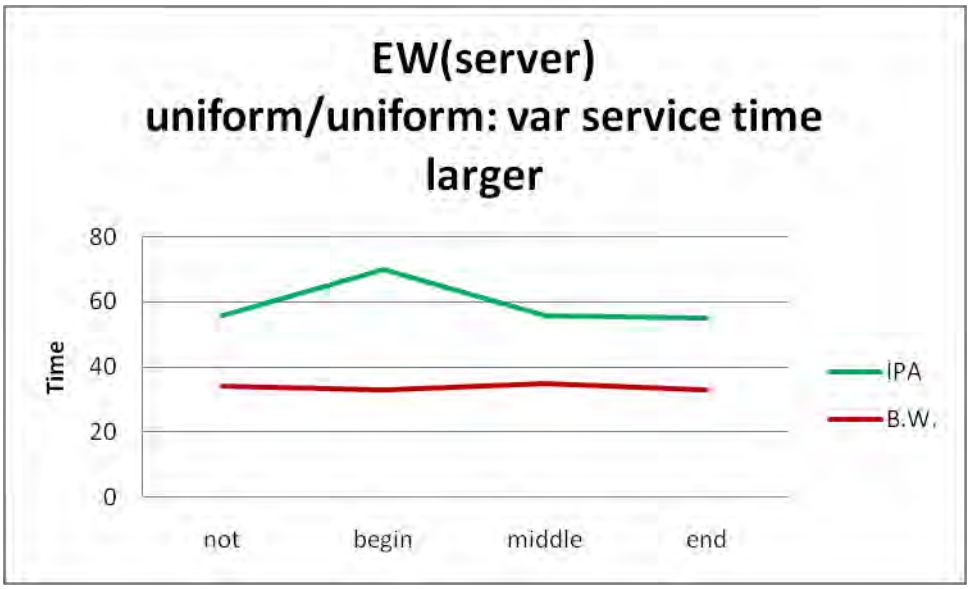


Figure 9: EW_{server} with a larger variance in service times

4.4.3 Average expected waiting time of both the customer and the server

To finish the section, we look at a customer with a larger variance in service times. The results are shown in Figure 10. Again we see that IPA provides a better EW than the Bailey-Welch Rule. Also we can see the numbers for EW slightly decrease if a customer with a larger variance is planned later on the day, this seems to hold for both IPA and the Bailey-Welch rule.

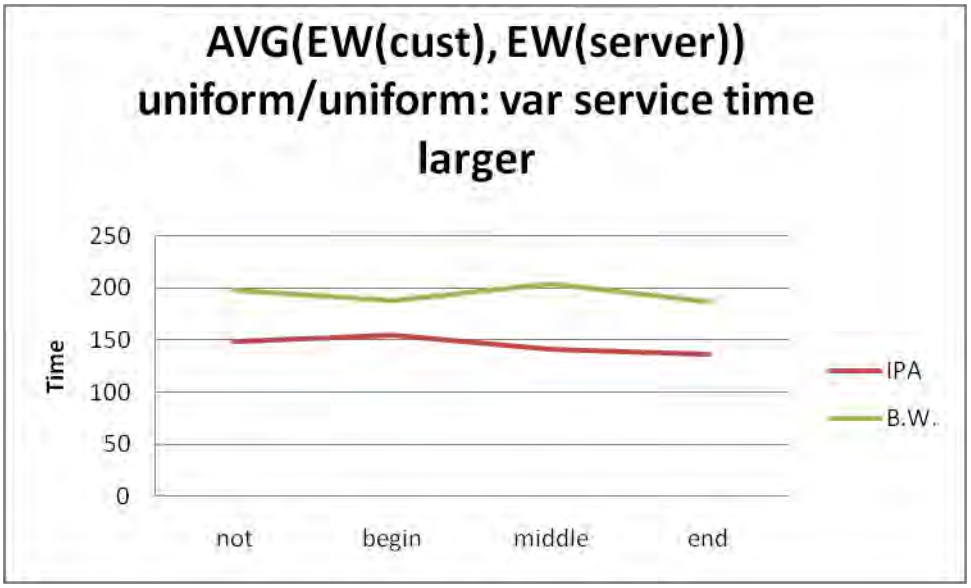


Figure 10: EW with a larger variance in service times

4.5 A larger expectation of service times

Next, we will deal with larger expectations of service times. Remember that each customer is planned to arrive at the expected finishing time of the service of the previous customer. Thus when there is a customer with a service duration twice is long, the planned interarrival time between this and the next customer will also be approximately twice as long compared to the others.

4.5.1 Expected waiting time of customers

Now we can look at EW_{cust}. The results for exponentially distributed arrival and service times are shown in Figure 11. As we can see, having a customer with an expected longer service time negatively affects EW_{cust}, the effect can however be negated by placing this customer at the end. This holds for both IPA and the Bailey-Welch rule and is quite remarkable since one would expect a longer service times is actually what we want to have, however when we are looking at longer service time expectations we also automatically look at a larger variance in case of the exponential distribution, this is because both the mean and the variance use the same parameters.

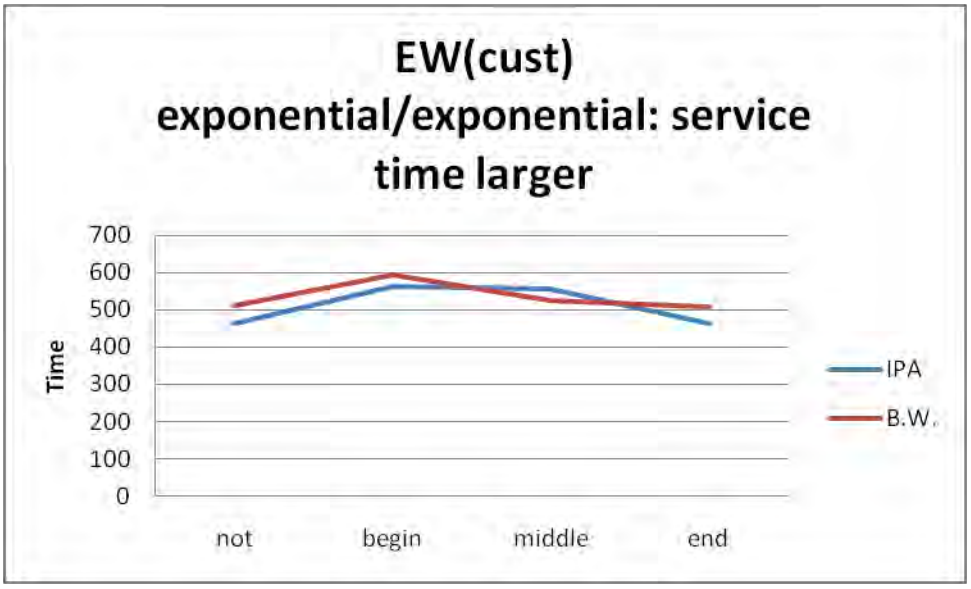


Figure 11: EW_{cust} with a larger variance in service times

In Appendix 7 we can see that when services are not exponentially distributed and the mean and variance thus use different parameters, we can increase the variance without increasing the mean, larger service times do not really have an effect on the results that IPA gives us. For the Bailey-Welch Rule, however, placing someone with a larger expected service time at the beginning of the day increases the overall waiting time EW_{cust} for customers.

4.5.2 Expected waiting time of the server

Next, we will look at how larger expected service times affect EW_{server}. Figure 12 shows the results for exponential arrival and service times. We can see that EW_{server} slightly decreases when we place a customer at the beginning of the day with a larger expected service time and that the Bailey-Welch rule gives us better numbers for EW_{server}.

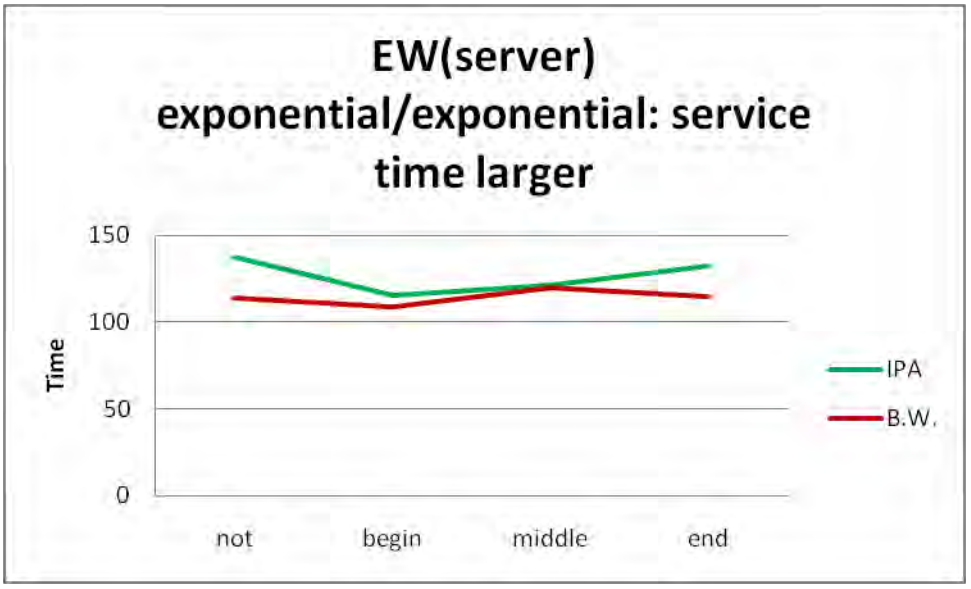


Figure 12: EW_{server} with a larger variance in service times

In Appendix 8 we can see that placing a customer with a larger expected service time at the beginning of the day does not necessarily gives us a lower number of EW_{server} for IPA.

4.5.3 Average expected waiting time of both the customer and the server

The average of the waiting times and how they are affected by larger service times can be seen in Figure 13. We can see that EW decreases as the customer with the longest expected service time is placed later on the day.

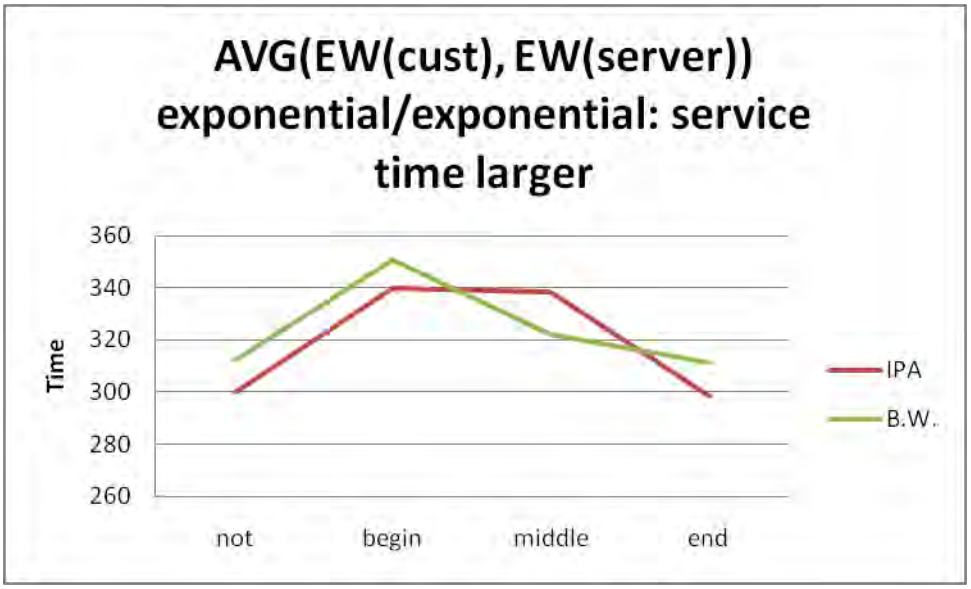


Figure 13: EW with a larger variance in service times

In Appendix 9 we can see our findings hold for another distribution of the service times as well.

4.6 When customers are expected to arrive late

In this final subsection of the results we will look at what happens when customers are expected to arrive late. This might be because of rush-hour traffic or that some customers just have a tendency to arrive late. In this case we will only look at morning rush-hour, this means that the first customer has a tendency to arrive ten minutes later than expected. The second, third and fourth customer are expected to arrive five minutes late. We will try to solve this problem with IPA by using a normal distribution for the arrivals and an exponential distribution for the service times. Since the server can only start at the beginning of the day, say T=0, no customer can be planned to arrive sooner than T=0. So we will expect the remaining customers, the 5^{th} customer till the 25^{th} customer, to be planned later.

Now let us look at the result. The second, third and fourth customer are planned to arrive earlier than in the original schedule and the customers hereafter will be planned later than in the older schedule. See appendix 10 for the planned arrival times. In Table 1 we can see the waiting times for the scenario with rush-hour in the morning and the original scenario.

	With rush-hour in morning	Original scenario
EW_{cust}	353.16	348
EW_{server}	124.79	126.5
EW	238.98	237.25

Table 1: waiting times for the different scenarios

So we can see that IPA is smart enough to deal with deviations of the mean, even though they are not directly visible to IPA.

5 Conclusions

Finally we can draw conclusions from the results.

5.1 Conclusions on the expected waiting time of customers

Concerning the expected waiting time of customers, EW_{cust}, we can draw the following conclusions and rules of thumb for IPA-like schedules.

First of all, IPA provides us better numbers than the Bailey-Welch rule. It is easy to see why this is true because IPA focuses on EW_{cust} while the Bailey-Welch rule is supposed to minimize EW_{server}. So whenever we are working with systems where only EW_{cust} should be minimized we can use IPA or the rules of thumb that we derived from IPA to schedule all customers. If all customers are similar, this comes down to planning the people earlier on the day slightly before the expected finishing time of the previous customer and the people later on the day slightly after the expected finishing time of the previous customer.
Rule 1: Plan customers slightly before the expected finishing time of the previous customer at the beginning of the day and slightly later at the end of the day.

Larger variances in arrival times do not have a negative effect on the overall waiting time of customers, unless the Bailey-Welch rule is used. Customers with larger variances in arrival times should be placed as late as possible to minimize EW_{cust}.
Rule 2: Customers with larger variances in the arrival time should be planned as late as possible.

When we are dealing with customers with a bad reputation of often not showing up we should look at what schedule we are using when we are planning these customers. When we are using a roster like the one IPA provides us, we are better off when these customers are planned last. However, when we are using the Bailey-Welch rule customers with a higher probability of not showing up should be planned between the beginning and the middle of the day. It could also be noted that when a customer has a high probability of not showing up it has almost always a positive effect on the waiting time of the customers.
Rule 3: Customers with a larger probability of not showing up should be planned as late as possible.

It is hard to tell what happens when there are customers whose service time has larger variances. They seem to have a positive effect on the IPA method when they are placed at the end of the day. For the Bailey-Welch rule there is only a negative effect which can be nullified when the customer with the larger variance is placed at the end of the day.
Rule 4: Customers with larger variances in the service time should be planned as late as possible.

Customers with larger expected service times should be placed at the end of the day, both for IPA and the Bailey-Welch rule. Larger expectations of the service time do not have an positive effect on EW_{cust}, although it again should be mentioned that a larger expectation also provides a larger variance for the exponential distribution.
Rule 5: Customers with larger expectations of the service time should be planned as late as possible.

5.2 Conclusions on the expected waiting time of the server

It would only seem logical that the Bailey-Welch rule provides better numbers now for EW_{server}.

Customers with larger variances do not really have a large effect on EW_{server} for the Bailey-Welch rule. For IPA however, EW_{server} will increase since it will try to keep EW_{cust} as low as possible.

Higher probabilities of not showing up slightly increases the waiting time of the server.

Larger variances of the service time do not affect EW_{server} much.

Whenever service times are expected to take longer EW_{server} decreases when these customers are placed at the beginning of the day.

5.3 Conclusions on the average expected waiting time

It is clear that whenever we take the average of the expected waiting time of the customer and of the server, IPA will provide us most of the times with better results. So if the waiting time of customers has the same costs as the waiting time of the server, IPA or the rules of thumb given by IPA should be used.

The lowest number for EW can be derived when customers with larger variances in arrival times, larger variances in service times and larger averages of service times are placed at the end of the day just as the rules of thumb say. Also for customers with higher probabilities of not showing up, these customers should be placed last for IPA. For the Bailey-Welch rule better results will be derived when customers with a high probability of no-shows are planned in the middle.

6 Suggestions for further research

In this paper we have researched the behavior of appointment driven systems through IPA, however, there are still some unanswered questions that have not been addressed in this paper due to lack of time. Remember that this paper only studied if it is possible to apply IPA to appointment driven systems and if it is possible to come up with rules of thumb. The following questions can be answered in further research.

Is it possible to add the expected waiting time of all customers and the server and approximate this derivative and apply IPA to this total expected waiting time? If this is possible it will lead to new possibilities, we can then also add weights to the expected waiting time of the customer and the expected waiting time of the server. This allows us to balance both types of expected waiting time and give an importance level to each of them.

Another idea that can be investigated is what will happen when during every iteration we will look at the customer that affects the waiting time the most, or least, and switch this customer with the previous or next customer depending on whether the derivative is positive or not. If the switch improves the expected waiting time the switch will be permanent and we will go the next iteration, if the expected waiting time does not improve we will switch the customers back and go to the next iteration.

Also lunch breaks or short breaks can be inserted into the systems. It might be that the lunch break is at a given time say 13.00 hour or just that the lunch break takes one hour but the timing of the lunch break does not really matter. Both can be studied in appointment driven systems.

Finally, a last idea for further research is to mix the different distributions such that part of the customers arrives according through a different distribution than the other customers.

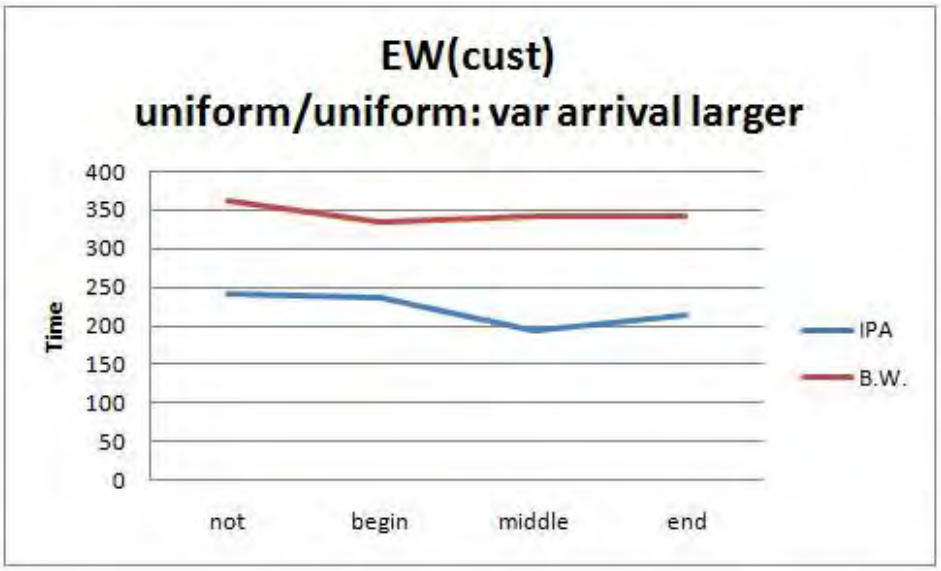
7 Literature

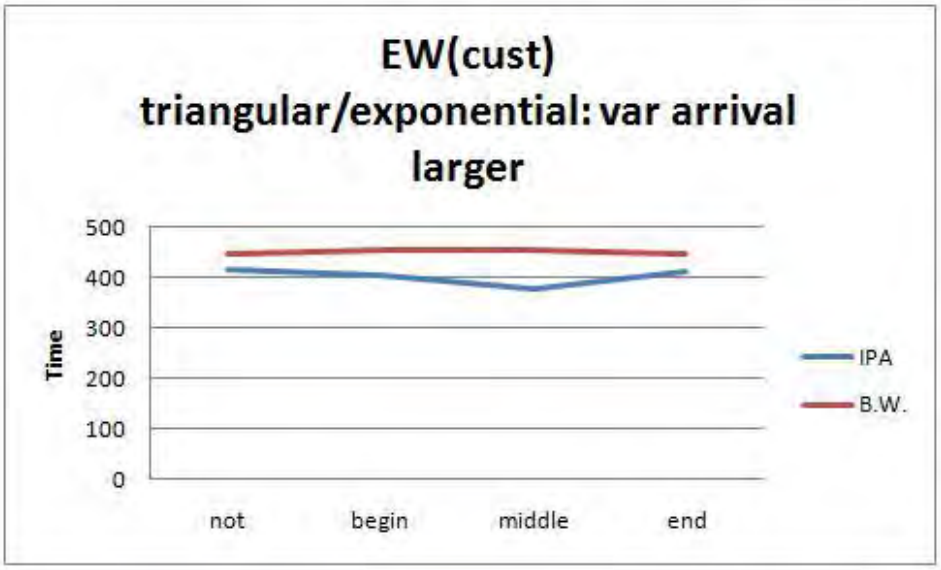
The following literature was used for this paper:

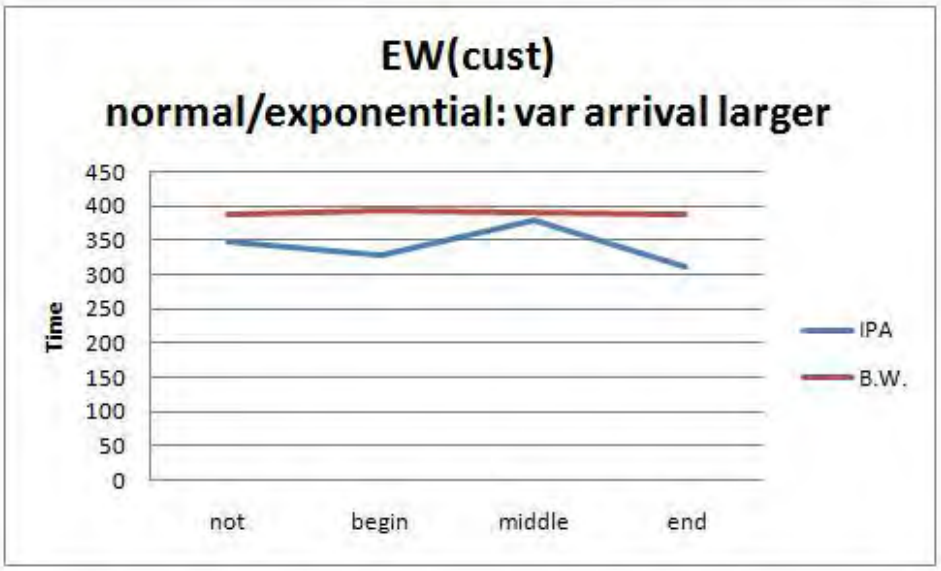
- Hutzschenreuter A. (2004), *Waiting Patiently*, BWI-werkstuk (http://www.few.vu.nl/nl/Images/werkstuk-hutzschenreuter_tcm38-91363.pdf)
- Heidelberger P., Xi-Ren, Zazanis A. M. and Suri R. (1988), *Convergence properties of Infinitesimal Perturbation Analysis estimates*
- Xi-Ren C. (1994), *Infinitesimal Perturbation Analysis of Generalized Semi-Markov Processes: A Tutorial*
- Johnson M. E. and Jackman J. (1989), *Infinitesimal Perturbation Analysis: A Tool for Simulation*
- Jouini O. and Benjaafar S. (2010), *Queuing Systems with Appointment-Driven Arrivals, Non-Punctual Customers, and No-Shows*
- Glasserman P. (2010), *Gradient Estimation via perturbation analysis*

8 Appendix

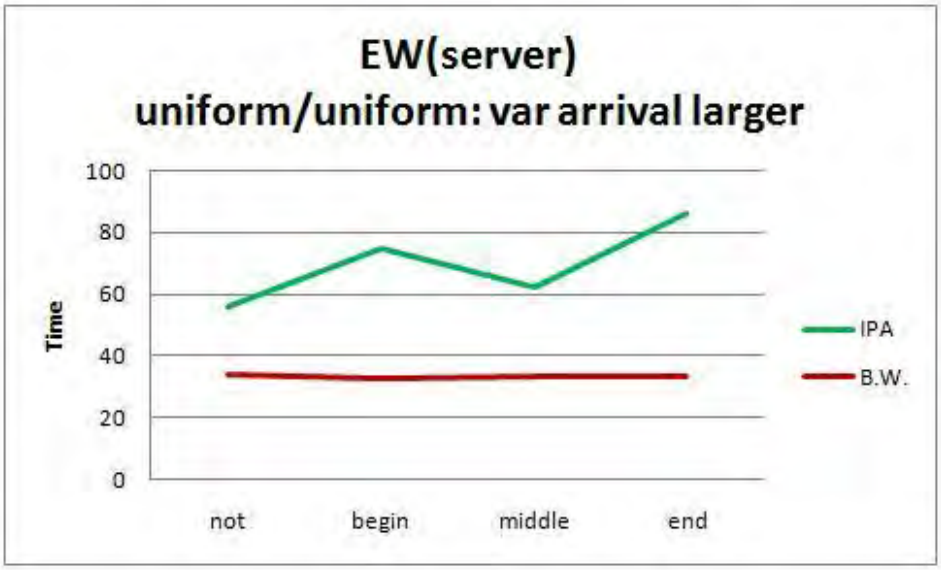
8.1 EW_{cust} with a larger variance in arrival times

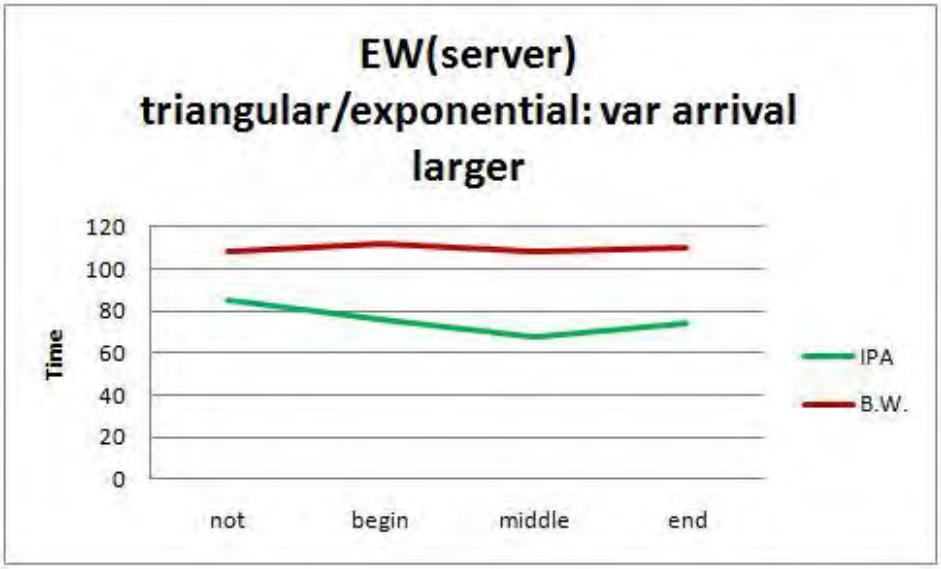


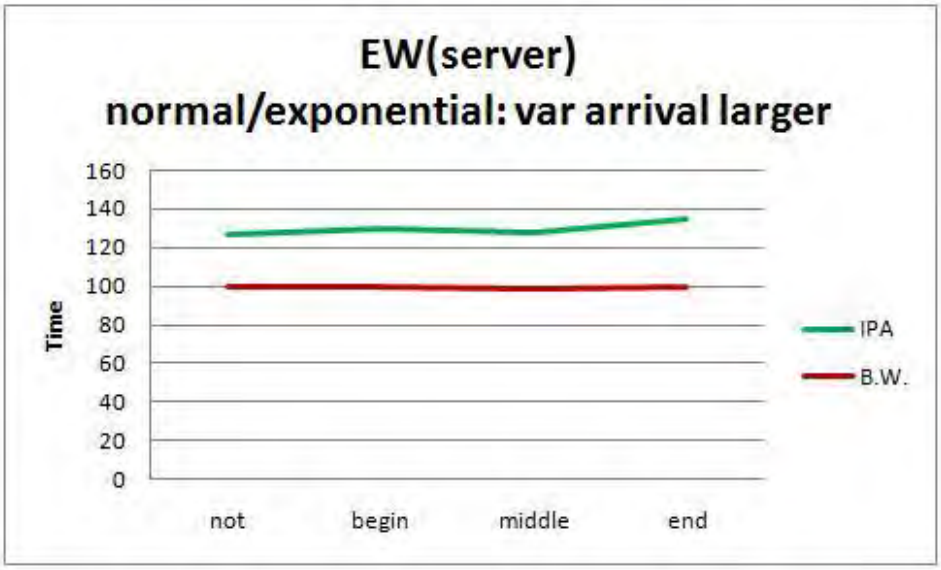




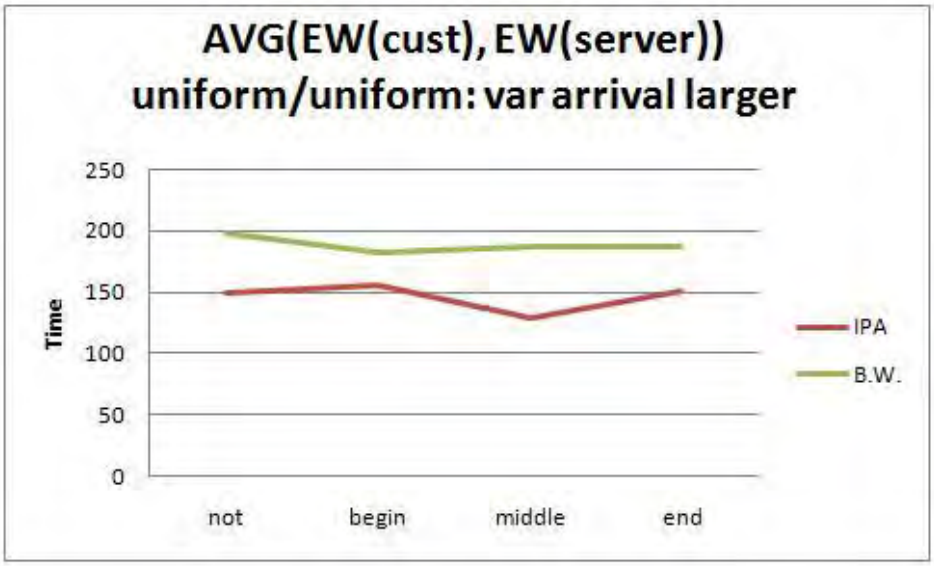
8.2 EW_{server} with a larger variance in arrival times

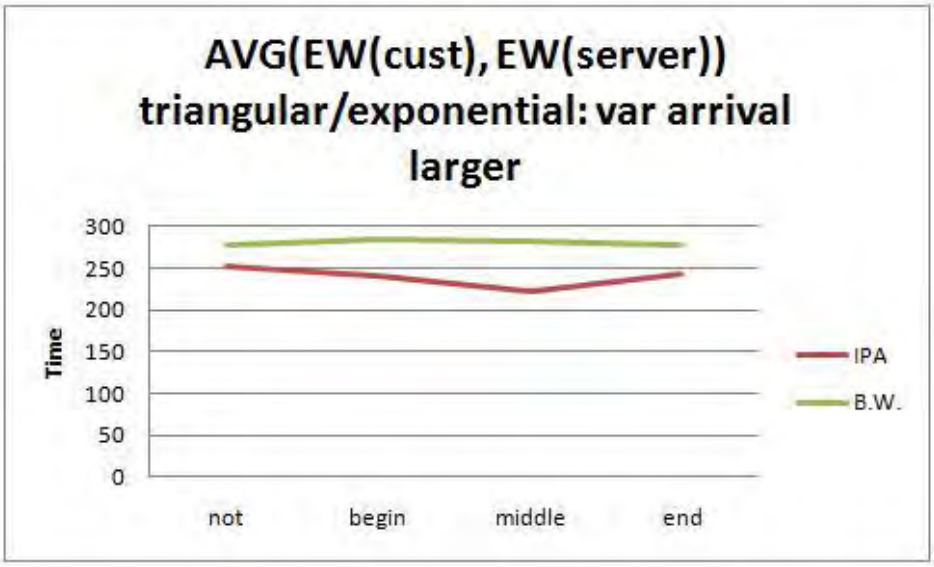


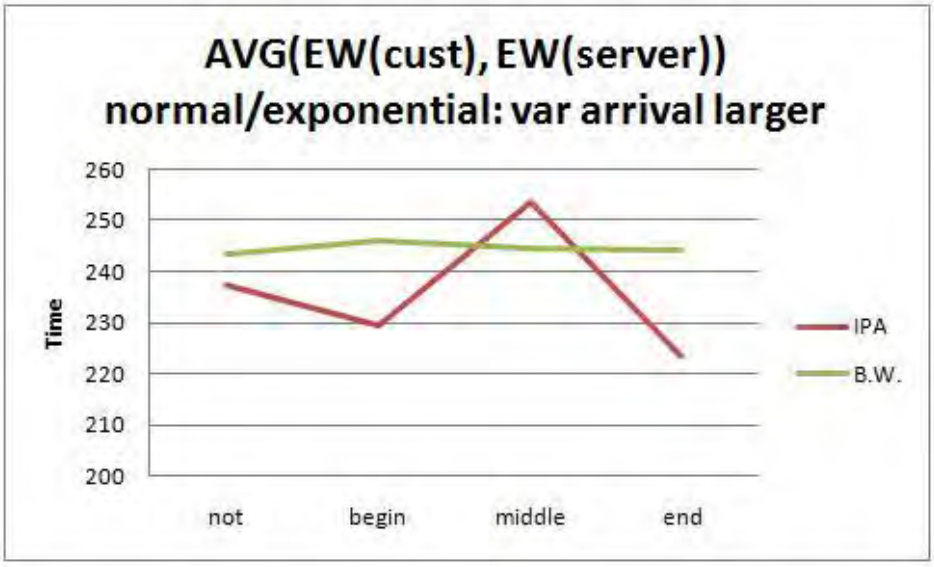




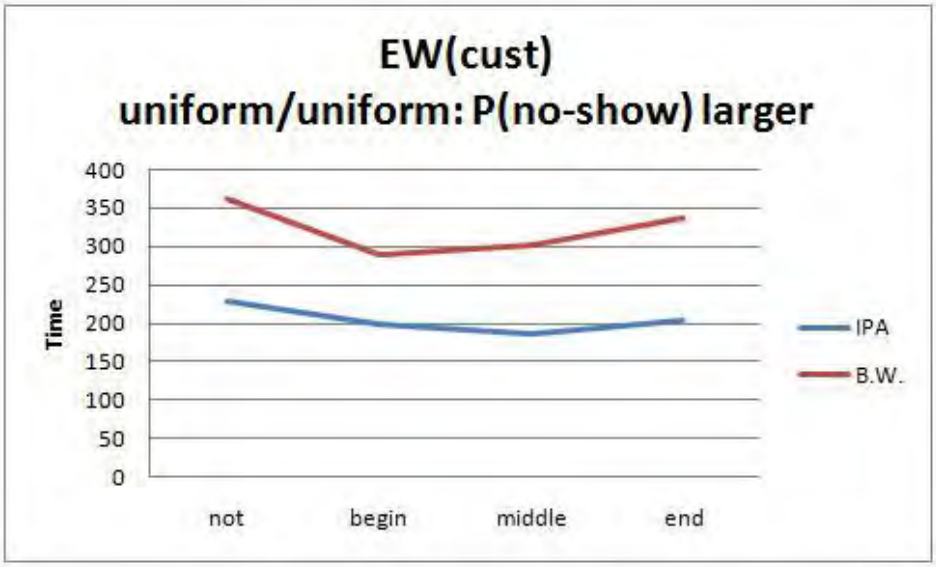
8.3 EW with a larger variance in arrival times

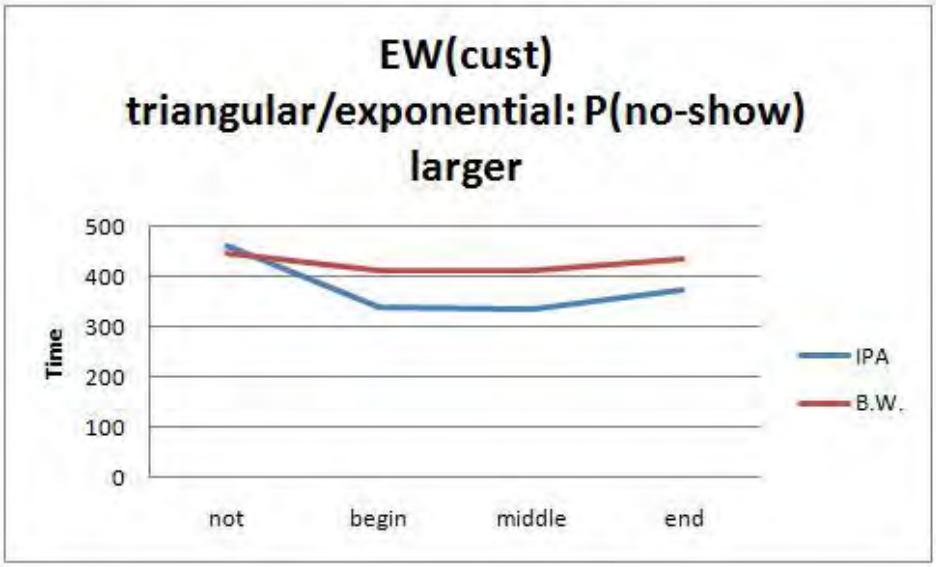


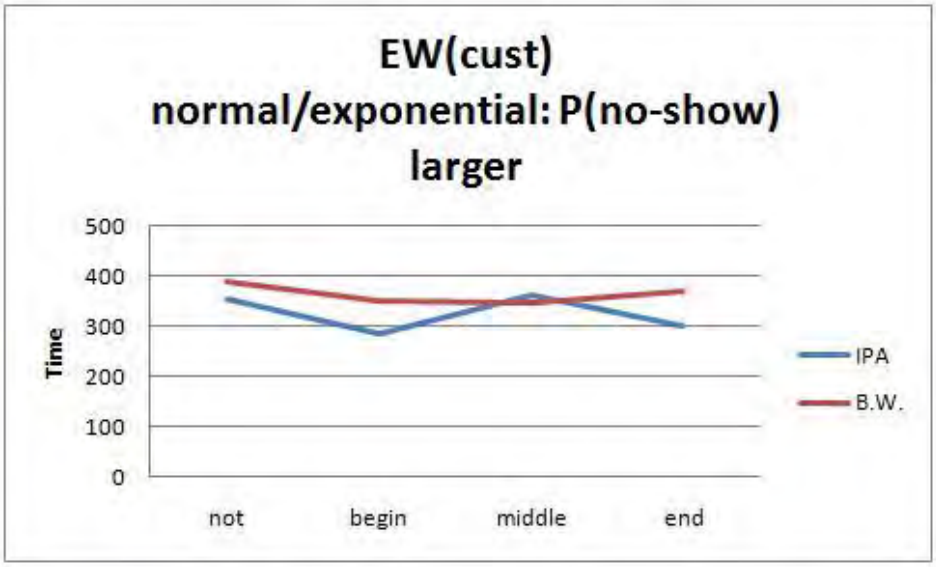




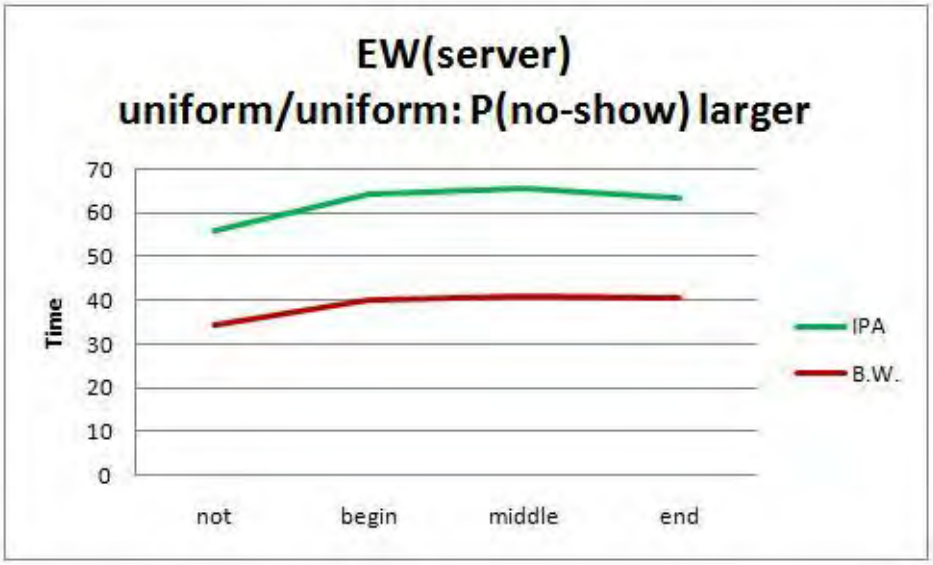
8.4 EW_{cust} with a larger probability of no-shows

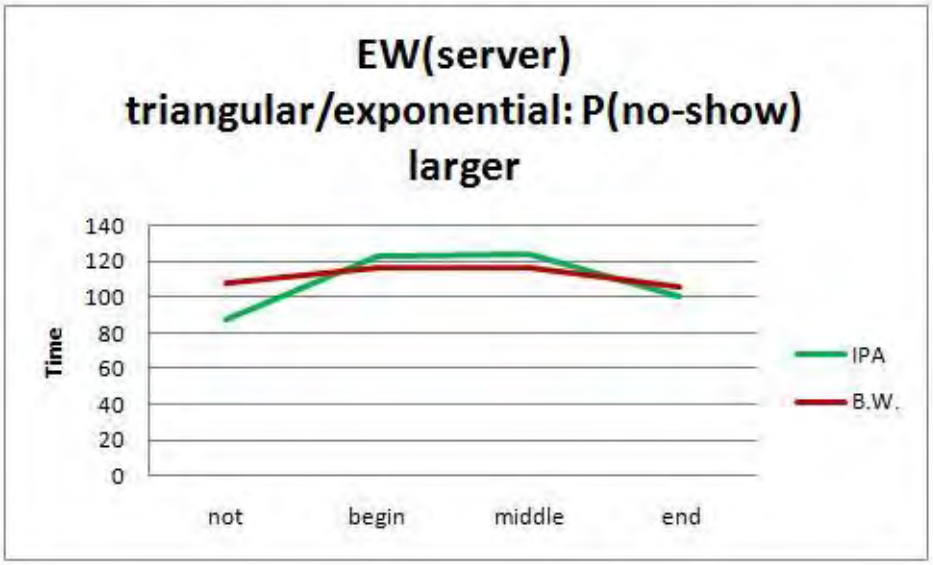


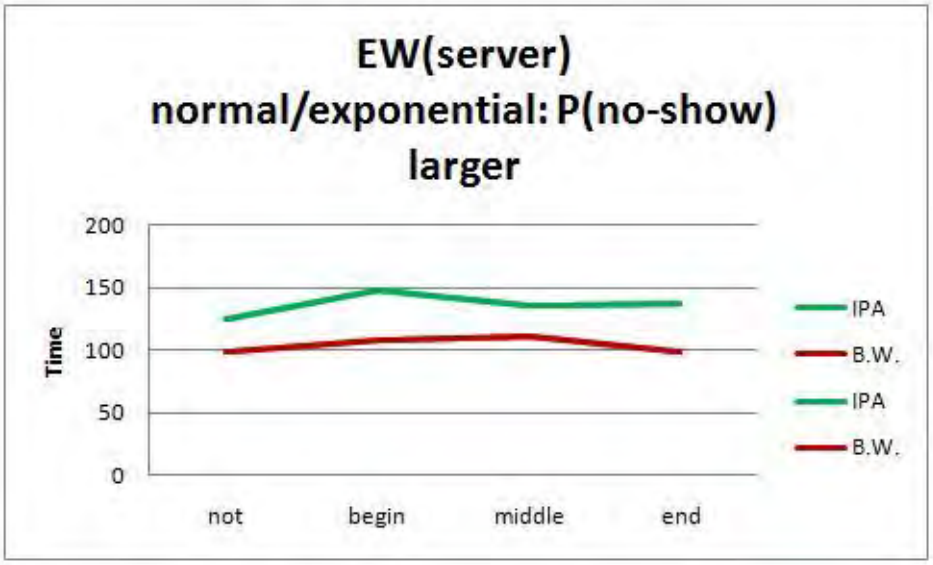




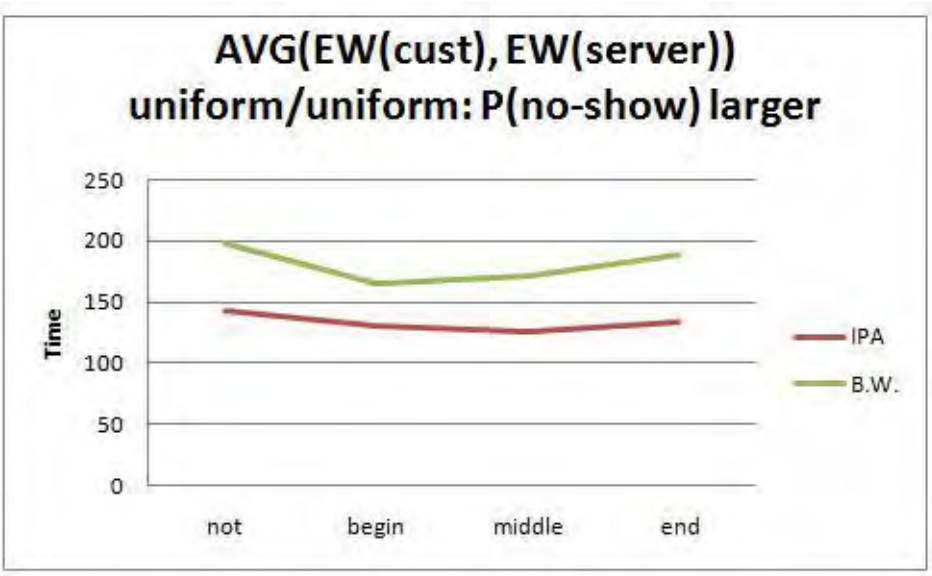
8.5 EW_{server} with a larger probability of no-shows

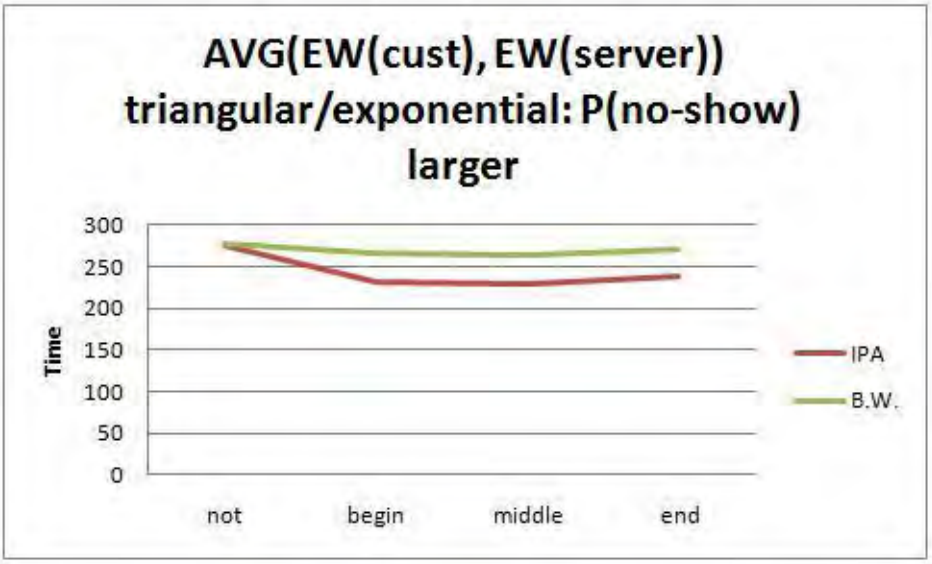


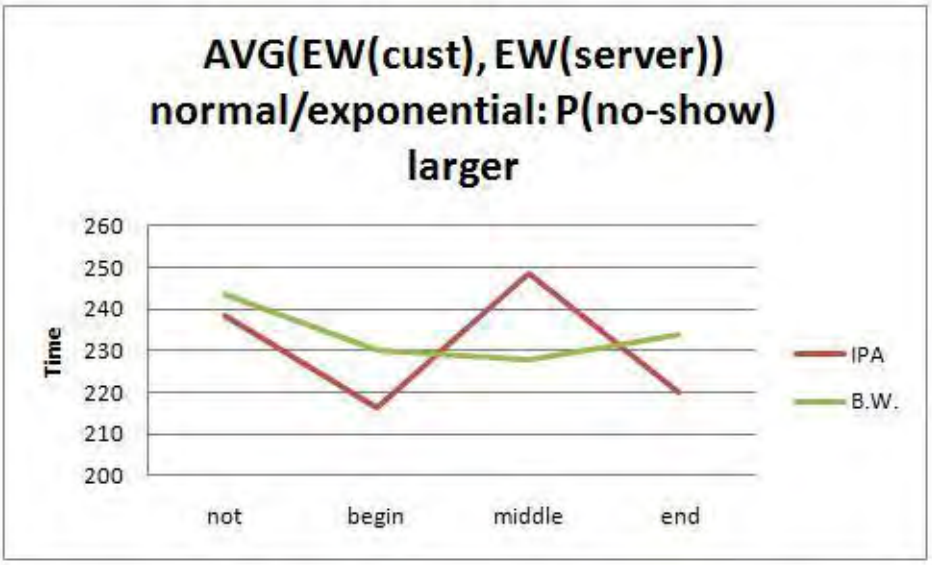




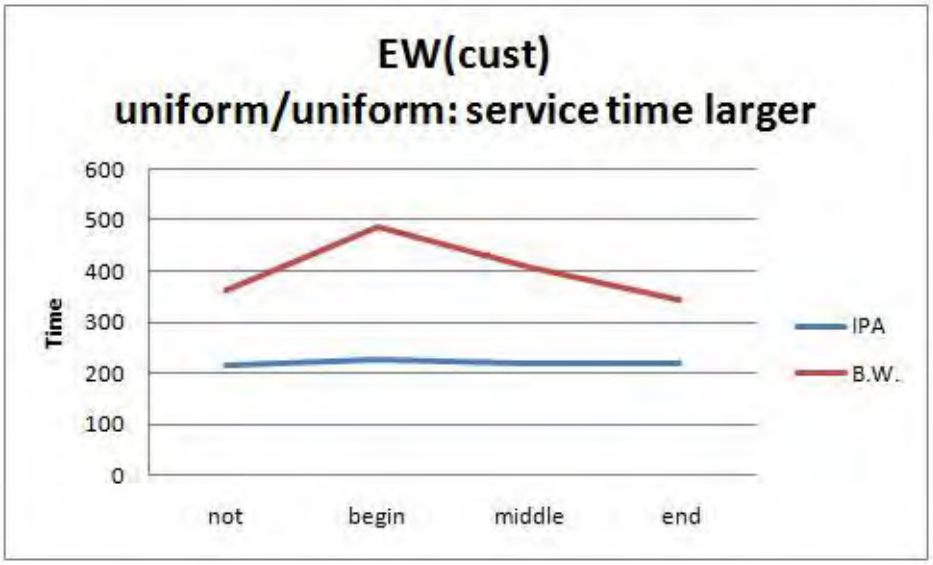
8.6 EW with a larger probability of no-shows

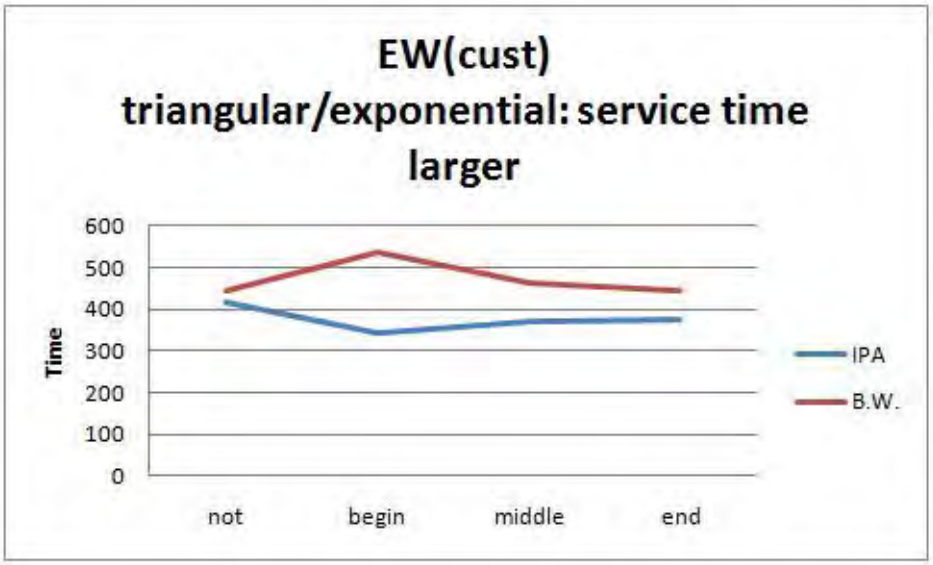


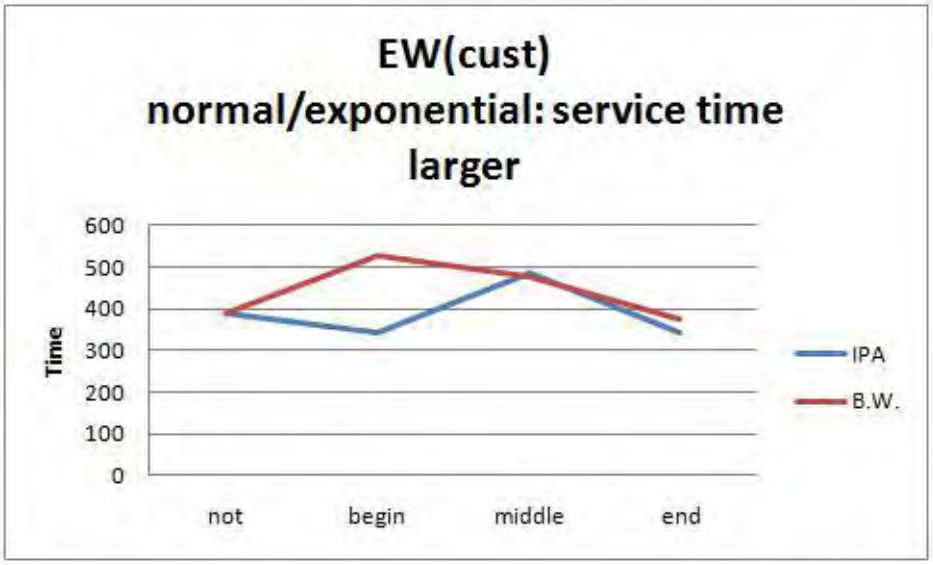




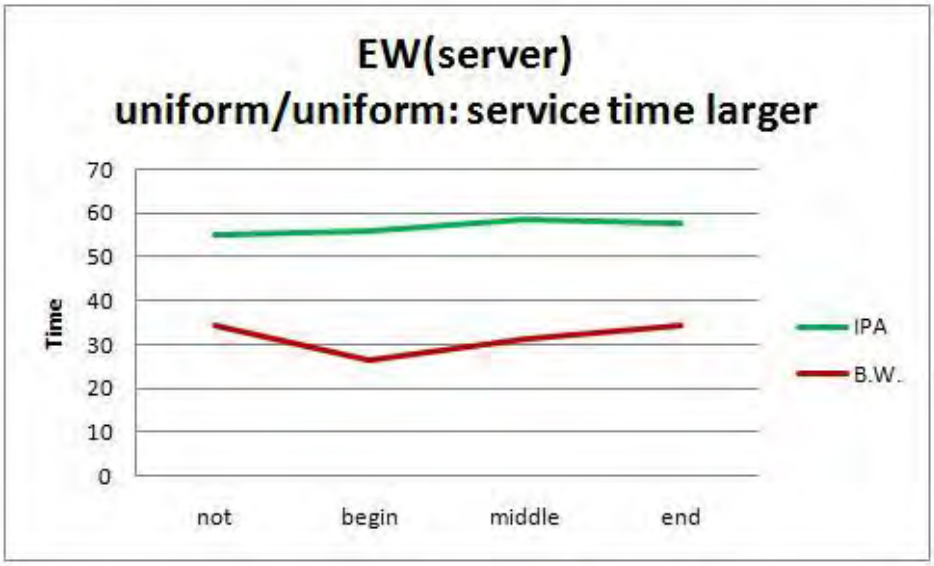
8.7 EW_{cust} with a larger variance of service times

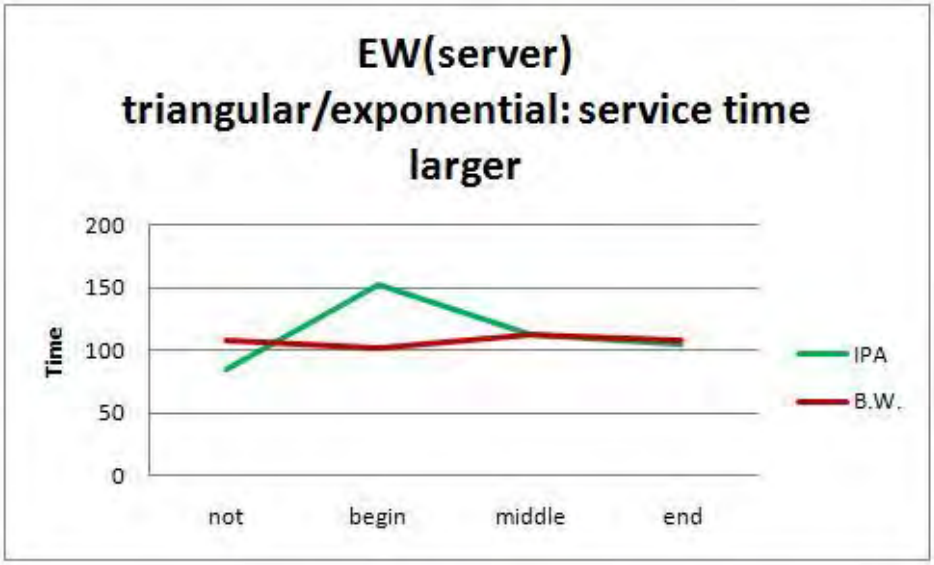


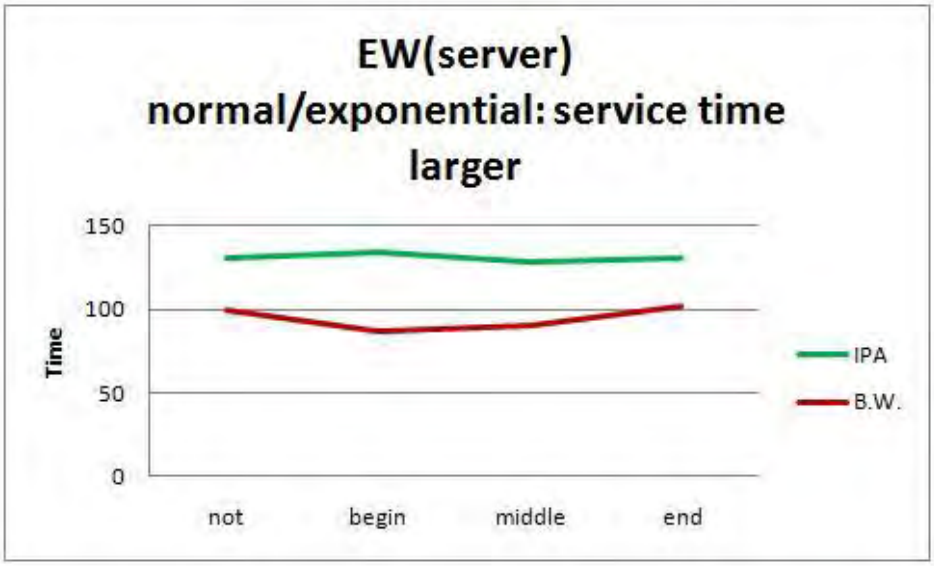




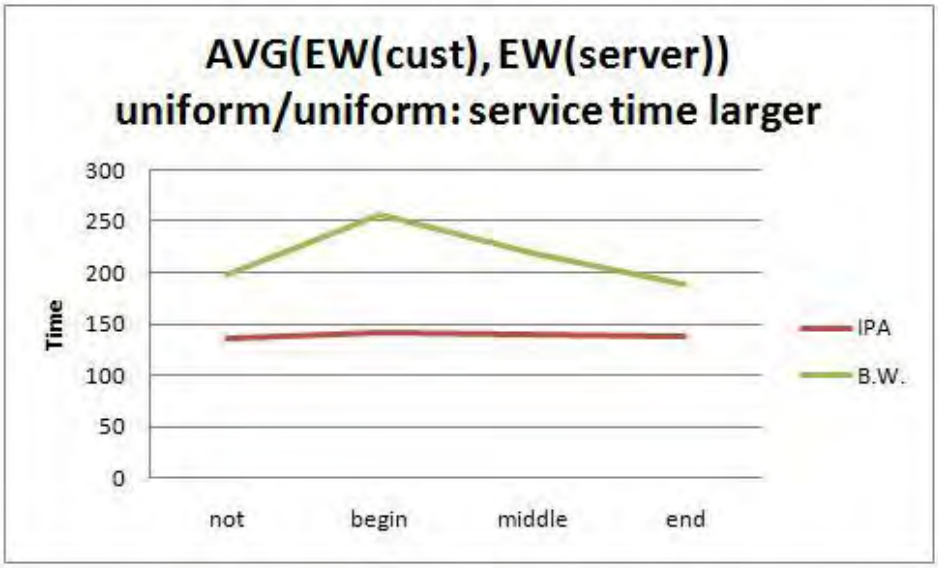
8.8 EW_{server} with a larger variance of service times

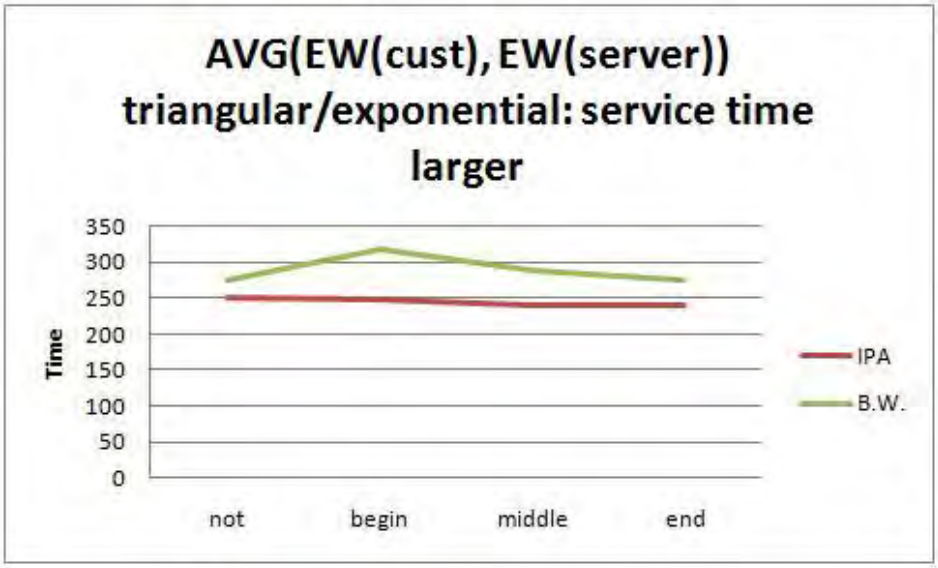


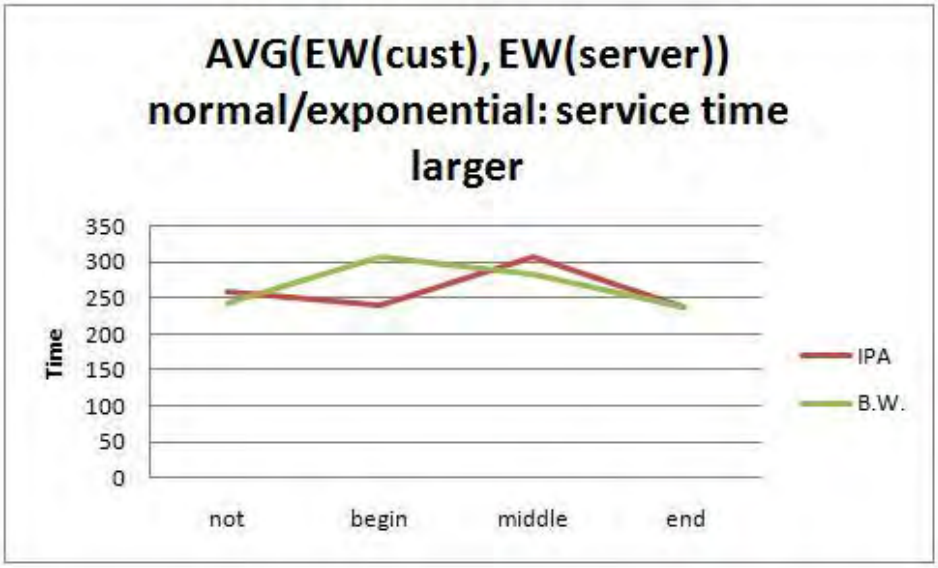




8.9 EW with a larger variance of service times







8.10 Planned arrival times - When customers are expected to arrive late

Customer number	Expected minutes late	Planned arrival time
1	10	0
2	5	16.59
3	5	39.42
4	5	59.87
5	0	80.69
6	0	102.03
7	0	125.03
8	0	144.25
9	0	164.64
10	0	179.66
11	0	206.72
12	0	220.99
13	0	243.13
14	0	267.80
15	0	288.32
16	0	302.96
17	0	328.95
18	0	346.41
19	0	367.01
20	0	383.40
21	0	402.03
22	0	398.95
23	0	441.94
24	0	452.08
25	0	492.17