QUEUEING SYSTEMS MODELING WITH OPERATIONAL STATISTICS †

J. George Shanthikumar
Krannert School of Management
Purdue University

For Presentation at EURANDOM
June 24, 2010

†This talk is based on joint work with R. Akhavan-Tabatabaei (NCSU, Raleigh & Intel, Chandler), S. Ding (UC, Berkeley), J. Fowler (ASU,Tempe), R. Nurani (KLA Tencor, San Jose), S. Pan (ASU, Tempe), S. Ross (UC, Berkeley), M. Zhang (Intel, Chandler), X. Zhang (UC, Berkeley)

AGENDA

- What is Operational Statistics? Illustrate through three examples
  - accounts for statistical errors
  - accounts for structural errors

  1. GI/GI/1 queue with limited information on data
  2. M/M/1 queue with finite sample - statistical errors
  3. G/G/1 queue with infinite sample - structural errors

- Application of Operational Statistics to a Queueing Model of a Toolset in a Semiconductor Manufacturing System

Stochastic Modeling in Practise

- Observe the real system
- Abstract a stochastic model and identify the INPUTS to fully or partially characterize the model
- Use the observed data to estimate the INPUTS to fully or partially characterize the model
- Predict the performance of the system with new design parameters or control policy
- Find the design parameters or control policy that "optimize" the system performance

Our Consulting Problem

System: Single stage manufacturing system with a single machine

Objective: Speed-up the service rate (say, using a s/w upgrade) so that the total cost of WIP and speed-up is minimized

Model: Single server GI/GI/1 queueing system

Parameters:

1. Full Characterization
   - Inter-arrival time distribution (F_A)
   - Service time distribution function (F_S)
2. Partial Characterization
   - Mean ($1/\lambda$) and squared coefficient of variation (C_A^2) of inter-arrival time
   - Mean ($1/\mu$) and squared coefficient of variation (C_S^2) of service time

Performance:

We are interested in $L(r) + h(r)$, where

r is the speed-up rate ($r > \rho$) [could be slowed down - $h(r)$ could be negative for $\rho < r \leq 1$, due to yield improvement]

$L(r)$ is the average WIP

$h(r)$ is the per unit time cost with speed-up rate ($r > \rho$)

Queueing theory primer:

For single server queueing system: define $\rho = \frac{\lambda}{\mu}$

M/M/1: $L = \frac{\rho^2}{1-\rho} + \rho$

GI/GI/1: $L \approx \frac{\rho^2}{1-\rho}(\frac{C_A^2+C_S^2}{2}) + \rho$

KEY: It can be shown that for some finite constants a and b we have

$$\frac{\rho^2}{1-\rho}(\frac{C_A^2+C_S^2}{2}) + \rho - a \leq L \leq \frac{\rho^2}{1-\rho}(\frac{C_A^2+C_S^2}{2}) + \rho + b$$

For c server queueing system: define $\rho = \frac{\lambda}{c\mu}$

GI/GI/c: $L \approx \frac{\rho\sqrt{2c-1}}{1-\rho}(\frac{C_A^2+C_S^2}{2}) + c\rho$

CASE 1: Limited Partial Information

Parameters: Only limted information such as the average arrival rate, average service time, average WIP, average waiting time etc. is available.

Remark: The manufacturer did not trust our capabiity to solve their problem. So they did not want to spend time collecting data for us – at the very beginning of my consulting experience.

Remark: However the averages that was provided has been computed over a very large sample size for us not to worry about the statistical validy of the estimates of the mean values.

How do we estimate so that we may find the optimal speed-up rate?

Thinking out loud!

- Distributions $F_A$and F_S are not available. Detailed analysis or simulation is out

- Second moments (thus C_A^2 and C_S^2) are not available. Two moment approximations are out.

How can we apply the classical queueing theory results to this problem?

Only the means (thus the λ and μ) are available. What distributions are fully characterized by their mean? OK, we could use the following queues to approximate $L(r)$.

1. D/D/1 (speed-up will not help)

2. D/M/1 (speed-up has very little impact)

3. M/D/1 (speed-up has some impact)

4. M/M/1 (speed-up has the most impact)

We could get away with the M/M/1 model:

But the real issue is – $L(1)$ should represent the observed WIP. None of the four queueing systems provide an estimate of $L(1)$ that even come close to the observed average WIP.

Hence the real question is - Can we live with this level of partial characterization?

A Theoretician's answer - Extremal Analysis

- Let $\hat{L}(r, F_A, F_S)$ be the mean stationary number of customers in a GI/GI/1 queueing system with inter-arrival time distribution F_A and service time distribution F_S

- Define $\mathcal{F}(x)$ be the family of all distribution functions of nonnegtive random variables with mean $\frac{1}{x}$

Then

$$\inf\{\hat{L}(r, F_A, F_S) : F_A \in \mathcal{F}(\lambda); F_S \in \mathcal{F}(r\mu)\} \leq L(r)$$

$$\leq \sup\{\hat{L}(r, F_A, F_S) : F_A \in \mathcal{F}(\lambda); F_S \in \mathcal{F}(r\mu)\}$$

A Big Problem with this IDEA:

$$\inf\{\hat{L}(r, F_A, F_S) : F_A \in \mathcal{F}(\lambda); F_S \in \mathcal{F}(r\mu)\} = \frac{\rho}{r}$$

and

$$\sup\{\hat{L}(r, F_A, F_S) : F_A \in \mathcal{F}(\lambda); F_S \in \mathcal{F}(r\mu)\} = \infty$$

A Practical Approach (or Quick and Dirty Approach)

- Define "queue management efficiency" of the manufacturing facility as follows:

$$\eta_{QM} = \frac{Q_{M/M/1}}{Q_{M/M/1} + Q},$$

where $Q = L - \rho$ is the average WIP in queue. NOTE: η_{QM} is an operational parameter.

Quick and Dirty Approximation

- Now approximate $L(r)$ as follows:

$$L(r) \approx \widetilde{L}(r) = (\frac{1-\eta_{QM}}{1-\frac{\rho}{r}})(\frac{\frac{\rho}{r}}{\eta_{QM}})\frac{\rho}{r} + \frac{\rho}{r}$$

OK! so how good is $\widetilde{L}(r)$ as an approximation to $L(r)$.May be it is as bad as one or all of D/D/1, D/M/1, M/D/1 and M/M/1?

Partial Characterization and Extremal Behavior (of the Operational Estimate $\widetilde{L}(r)$ of $L(r)$)

Use the partial characterization of the INPUT (λ, μ) & OUTPUT (L) of a given system to obtain the extremal behavior!

Specifically we are looking for:

$$L_l(r) \leq \inf\{\widehat{L}(r, F_A, F_S) : \widehat{L}(1, F_A, F_S) = L; F_A \in \mathcal{F}(\lambda); F_S \in \mathcal{F}(r\mu)\}$$

and

$$L_u(r) \geq \sup\{\widehat{L}(r, F_A, F_S) : \widehat{L}(1, F_A, F_S) = L; F_A \in \mathcal{F}(\lambda); F_S \in \mathcal{F}(r\mu)\}$$

The Analysis

It can be shown that there exists constants α and β independent of r such that

$$L_u(r) = (\frac{1-\eta_{QM}}{1-\frac{\rho}{r}})(\frac{\frac{\rho}{r}}{\eta_{QM}})\frac{\rho}{r} + \frac{\rho}{r} + \beta$$

and

$$L_l(r) = (\frac{1-\eta_{QM}}{1-\frac{\rho}{r}})(\frac{\frac{\rho}{r}}{\eta_{QM}})\frac{\rho}{r} + \frac{\rho}{r} - \alpha$$

Furthermore when the inter-arrival times have a decreasing mean residual life distribution, we have $\alpha, \beta \leq 2$.

CASE 2:Detail Data Provided, but the sample size is not very large – Statistical Issues

Given: m observation of inter-arrival times $\{A_1, ..., A_m\}$ and n observations of service times $\{S_1, ..., S_n\}$

Estimate: The average WIP $L(r)$ in the system if we were to speed up the service rate by a factor $r, (r > \rho)$

Suppose inter-arrival times and service times have exponential distributions (very good fit with the data). Define

- $\overline{A} = \frac{1}{m}\sum_{k=1}^{m} A_k$; $\widehat{\lambda} = \frac{1}{\overline{A}}$
- $\overline{S} = \frac{1}{n}\sum_{k=1}^{n} S_k$; $\widehat{\mu} = \frac{1}{\overline{S}}$
- $\widehat{\rho} = \frac{\widehat{\lambda}}{\widehat{\mu}}$

Then an estimate of $L(r)$ is $\widehat{L}(r) = \frac{\frac{\widehat{\rho}}{r}}{1-\frac{\widehat{\rho}}{r}}$, $r > \rho$

Can we realy estimate the average WIP this way?

Observe that to estimate $L(r)$ we need $\frac{\widehat{\rho}}{r} < 1$ for $r > \frac{1}{\rho}$. Let us check this out!

Recall that

$$\frac{1}{2\mathbb{E}[Z]} \sum_{l=1}^{k} Z_l =^d \chi^2_{2k}$$

for $i.i.d.$ exponential random variables $\{Z_1, ..., Z_k\}$

Estimating the Chance of Estimating

Therefore

$$\frac{m\mu}{2\widehat{\mu}} =^d \chi^2_{2m}; \ \frac{n\lambda}{2\widehat{\lambda}} =^d \chi^2_{2n}$$

Hence

$$\frac{\widehat{\rho}}{r} =^d \frac{\rho}{r}\{\frac{\chi^2_{2m}/2m}{\chi^2_{2n}/2n}\} =^d \frac{\rho}{r}F_{2m,2n}$$

[see, Schruben and Kulkarni, OR Letters]

Continue Estimating the Chance of Estimating

Then

$$P\{\frac{\widehat{\rho}}{r} < 1\} = P\{F_{2m,2n} < \frac{r}{\rho}\}$$

Consider the special case $m = n$

Here

$$\lim_{r \downarrow \rho} P\{F_{2n,2n} < \frac{r}{\rho}\} = 0.5$$

That is, we have a 50% chance of not being able to estimate $L(r)$, $r > \rho$.

May be we can almost always estimate $L(1)$, that is, with the current service settings?

The Statistical blues

Percent chance of estimating a utilization large than 1.

$n \backslash \rho$	.90	.93	.96	.99
100	1.09%	2.84%	7.66%	23.92%
200	0.06%	0.35%	2.17%	15.80%
300	0.00%	0.05%	0.67%	10.98%
400	0.00%	0.01%	0.21%	7.81%
500	0.00%	0.00%	0.07%	5.65%

A possible operational approach to this problem - Operational Statistics

Imagine that we have a queueing system with n arrivals and n servie completions. The total time taken to finish all the n jobs is then

$$T_n = \max_{1 \leq k \leq n} \{ \sum_{l=1}^{k} A_l + \sum_{l=k}^{n} S_l \}$$

Note that the server has been busy for a total of $\sum_{l=1}^{n} S_l$ units of time. Hence an operatonal estimate of ρ is

$$\widehat{\rho} = \frac{\sum_{l=1}^{n} S_l}{T_n}$$

The corresponding estimate of the mean queue size is

$$\widehat{L} = \frac{\widehat{\rho}}{1 - \widehat{\rho}}$$

Observe that $\widehat{\rho} < 1$ and hence $\widehat{L}$ is well defined.

Question: Are these estimators strongly consistent?

Recall the following martingale convergence theorem (of Xia, Shanthikuamr and Glynn, Operations Research)

Suppose for some $p > 2$, $\mathbb{E}[|X|^p] < \infty$, $\mathbb{E}[|Y|^p] < \infty$ and $\mathbb{E}[X] < \mathbb{E}[Y]$. Then as $n \to \infty$,

$$\frac{1}{n}\max\{0, \max_{1\leq k\leq n}\{\sum_{l=k}^{n}(X_l - Y_l) + X_k\}\} \to^{a.s.} 0$$

Now noting that

$$\max_{1\leq k\leq n}\{\sum_{l=1}^{k} A_l + \sum_{l=k}^{n} S_l\} = \sum_{l=1}^{n} A_l + \max_{1\leq k\leq n}\{A_k + \sum_{l=k}^{n}(S_l - A_l)\}$$

For $\mathbb{E}[S] < \mathbb{E}[A]$ (that is for $\rho < 1$) we have

$$\frac{1}{n} \max_{1 \leq k \leq n} \{ \sum_{l=1}^{k} A_l + \sum_{l=k}^{n} S_l \} \rightarrow^{a.s.} \mathbb{E}[A]$$

as $n \rightarrow \infty$. Hence $\widehat{\rho} \rightarrow^{a.s.} \rho$ and $\widehat{L} \rightarrow^{a.s.} L$ as $n \rightarrow \infty$.

CASE 3: Sufficiently Large Sample of Inter-Arrival and Service Times

Suppose fitting the data shows that

- inter-arrival times have an exponential distribution with mean 1, and
- service times have an exponential distribution with mean 1

Then $\rho = 1$and $L = \infty$. But we observe that $\bar{L} = 1$.

EXPLAIN the difference!

The Case of A Chatty teller

Consider a server who likes to keep one customer to chat, but one who is extremely efficient. Suppose we have one customer in the system. So the server keeps serving the customer very slowly while chatting happily. As soon as a new customer arrives, the server serves the customer very very fast (I mentioned that the server is very efficient). The server thus keeps only one customer in the system (thus $L = 1$). Here the service time of a customer is the same as the inter-arrival time of the next customer (thus $\lambda = \mu = 1$).

State Dependent Service Rate (Ross, Shanthikumar & Zhang, PEIS)

Model 1: Suppose we have a queueing system with Poisson arrival process with arrival rate λ. Customer requires $i.i.d.$ service with mean 1 and distribution F_S. However, when there are k customers in the system the server provides service at rate rate $\mu(k)$.

Let us assume that there is a large number of data available on the inter-arrival and service times. The exponential distribution with mean $1/\lambda$ will fit the inter-arrival times. Let $\widehat{F}_S$ be the distribution function fit to the service times. If we now use the results for the M/GI/1 queueing system, will the number of customers be larger (in a stochastic sense) or smaller than that in the state dependent service M/G/1 queue?

waiting time Dependent Service times (Buzacott, INFOR)

Model 2: Suppose we have a queueing system with Poisson arrival process with arrival rate λ. Customer requires service times that depends on the time they have waited in the queue.

Let us assume that there is a large number of data available on the inter-arrival and service times. The exponential distribution with mean $1/\lambda$ will fit the inter-arrival times. Let $\widehat{F}_S$ be the distribution function fit to the service times. If we now use the results for the M/GI/1 queueing system, will the number of customers be larger (in a stochastic sense) or smaller than that in the waiting time dependent service M/G/1 queue?

Coordination Efficiency

Consider a G/G/1 and define its co-ordination efficiency (η_c) by

$$\eta_C = \frac{Q_{M/M/1}}{Q_{M/M/1} + \frac{Q}{(\frac{C_A^2+C_S^2}{2})}}$$

For the "chatty teller," example $\eta_c = 100\%$. It can be shown that for Model 1, $\eta_C \leq 50\%$ ($\geq 50\%$) if the service rate ($\mu(k)$) is decresing (increasing) in k. For Model 2, $\eta_C \leq 50\%$ ($\geq 50\%$) if the service time is stochastically increasing (decreasing) in the waiting time.

QUEUEING ANALYSIS OF A TOOLSET IN A SEMICONDUCTOR MANUFACTURING SYTEM (SMS)

- The tool set has seven tools (4 from one specific equipment manufacturer and the other three form another)

We begin with models that are developed for single server queues.

- To apply these models to a toolset in SMS we need to make the assumption that each tool in the toolset works independent of the other six and therefore we have seven separate queueing systems.

- We assume each tool has its own queue of lots waiting to be processed.

To provide input to the model we make the following assumption:

- Lots that are eventually processed by a certain tool are assumed to have arrived to a queue designated to that particular tool. This is a deviation from the real world situation where a single queue is formed in front of the toolset and lots are randomly (or by choice) assigned to tools based on their availability.

- The parameters are estimated through the analysis of historical performance of the toolset over a six month period in relatively steady state conditions.

- Note that all the historical fab data in this paper are normalized to protect the intellectual property of the company that is providing the data.

Table 1: % WIP Processed and Queue Time Mean and Variance

Tool	% WIP Processed	Actual Queue Time	
		Mean	Variance
Tool 1	24.87%	11.09	373.35
Tool 2	25.39%	11.76	382.75
Tool 3	15.08%	13.55	566.87
Tool 4	7.24%	12.11	305.58
Tool 5	12.01%	19.44	1072.45
Tool 6	8.69%	18.38	813.34
Tool 7	6.72%	18.97	965.69

Table 2: Mean and SCV of Inter-arrival Times

Tool	Mean Inter-arrival Time	Squared Coefficient of Variation
Tool 1	0.92	3.02
Tool 2	0.90	2.58
Tool 3	1.51	9.18
Tool 4	3.15	4.43
Tool 5	1.90	3.19
Tool 6	2.62	1.72
Tool 7	3.39	2.17

Table 3: Mean and SCV of Service Times

Tool	Mean Processing Time	Squared Coefficient of Variation
Tool 1	0.92	3.02
Tool 2	0.90	2.58
Tool 3	1.51	9.18
Tool 4	3.15	4.43
Tool 5	1.90	3.19
Tool 6	2.62	1.72
Tool 7	3.39	2.17

Table 4: Availability, Effective Utilization and Effective Processing Time.

Tool	Availability	Effective Utilization *(ρ)*	Effective Processing Time	
			Mean	SCV
Tool 1	0.80	0.85	0.78	3.08
Tool 2	0.78	0.86	0.77	3.48
Tool 3	0.58	0.88	1.34	14.02
Tool 4	0.44	0.93	2.92	8.24
Tool 5	0.69	0.80	1.53	5.16
Tool 6	0.63	0.92	2.40	2.78
Tool 7	0.41	0.94	3.19	10.77

Whitt and Hopp & Spearman approximation for GI/GI/1 queues.

$$t_q = \frac{\rho(C_a{}^2 + C_e{}^2)}{2(1-\rho)} t_e,$$

Table 5: Comparison of Actual Cycle Time and Hopp & Spearman GI/GI/1 Estimation

Tool	Actual CT (hrs)	H&S G/G/1 CT (hrs)	% Difference
Tool 1	11.09	14.63	32%
Tool 2	11.76	15.17	29%
Tool 3	13.55	120.26	787%
Tool 4	12.11	239.84	1880%
Tool 5	19.44	27.78	43%
Tool 6	18.38	62.92	242%
Tool 7	18.97	337.22	1678%

Buzacott and Shanthikumar propose three approximations for G/G/1 queues. Two of these formulas are appropriate for the queues that have $C_a^2 \le 2$ and the third approximation is proposed for cases where $C_a^2 \le 1$. Since in the case of SMS toolset C_a^2 is typically much larger than 1 we apply the more appropriate approximation,

$$t_q = \left\{ \frac{\rho^2 (1 + C_s{}^2)}{1 + \rho^2 C_s^2} \right\} \left\{ \frac{(C_a^2 + \rho^2 C_s^2}{2\lambda(1-\rho)} \right\} + t_e$$

and show the results in Table 6.

Table 6, Comparison of Actual Cycle Time and B&S G/G/1 Estimation

Tool	Actual CT (hrs)	B&S G/G/1 CT (hrs)	% Difference
Tool 1	11.09	15.82	43%
Tool 2	11.76	16.12	37%
Tool 3	13.55	130.92	866%
Tool 4	12.11	248.96	1956%
Tool 5	19.44	30.67	58%
Tool 6	18.38	64.19	249%
Tool 7	18.97	340.72	1696%

Multi-Server Systems with General Arrival and Service Distributions

These models treat the entire tools or similar type of tools in the toolset as a single stage queueing system with parallel servers and general distributions for arrival and service processes and are denoted as G/G/c queues.

Table 5: Toolset data for multi-server systems

Tool Group	Inter-arrival Time		Effective Utilization	Effective Processing Time	
	$(1/\lambda)$	C_a^2	ρ	t_e	C_e^2
1-4	0.31	6.29	0.87	1.10	8.80
5-7	0.83	4.29	0.88	2.19	6.73

Whitt and Hopp & Spearman propose the following formula to approximate the queue time of the G/G/m queues, the result of applying this formula to our case study toolset are presented in Table 8.

$$t_q = \left(\frac{{C_a}^2 + {C_e}^2}{2} \right) \left(\frac{\rho^{(\sqrt{2(m+1)}-1)}}{m(1-\rho)} \right) t_e$$

Table 6: Comparison of Actual CT and H&S G/G/m Approximation.

Tool Group	Actual CT (hrs)	G/G/c Cycle Time (hrs)	% Difference
1 (tools1-4)	11.94	13.50	13%
2 (tools 5-7)	18.99	28.36	49%

Buzacott & Shanthikumar propose the following approximation for the G/G/c queue which is based on the queue time of an M/M/c queue. An M/M/c queue is a multi server system whose arrival and service processes both follow the exponential distributions denoted by M.

$$t_q^{G/G/m} = \frac{C_a^2(1-(1-\rho)C_a^2)/\rho + C_e^2}{2} t_q^{M/M/m}$$

where $$t_q^{M/M/m} = \left(\frac{1}{m\mu - \lambda}\right)\left(\frac{m^m \rho^m}{m!(1-\rho)}\right) p(0)$$

and *p(0)* denotes the steady state probability that we have zero lots in the queue. Applying this formulation to the data in Table 7 results in the cycle time estimates of Table 9 for each tool group.

Table 7: Comparison of Actual CT and B&S G/G/c Approximation.

Tool Group	Actual CT (hrs)	G/G/m CT (hrs)	% Difference
1 (tools1-4)	11.94	9.46	-21%
2 (tools 5-7)	18.99	23.46	24%

Queueing Approximations for Semiconductor Toolsets

Based on the classical queuing approximations some customized models for cycle time estimation in SMS are developed in the literature. The attempt of such models has been to enhance the estimation accuracy of the classical models by making some modifications. In this section we study the performance of such models that are relevant and applicable to our case study toolset data.

Whitt GI/G/m Approximation

Whitt's approximation for the expected waiting time in queue of a GI/G/m system is based on the proportional relationship with the exact values for the M/M/m model. Whitt's formula estimates waiting time as follows.

$$t_q^{G/G/m}(\rho, C_a^2, C_e^2, m) \approx \phi(\rho, C_a^2, C_e^2, m)\left(\frac{C_a^2 + C_e^2}{2}\right) t_q^{M/M/m}$$

where $\phi(\rho, C_a^2, C_e^2, m)$ is given by Whitt (1993) and he concludes that his approximation for the expected waiting time is "fairly accurate because it is relatively robust and extensively studied".

Table 10: Comparison of Actual CT and Whitt G/G/m Approximation.

Tool Group	Actual CT (hrs)	G/G/m CT (hrs)	% Difference
1 (tools1-4)	11.94	84.76	86%
2 (tools 5-7)	18.99	202.40	91%

Morrison & Martin Practical Extensions to Cycle Time Approximation for The G/G/m Queue

Morrison and Martin propose practical extensions to the G/G/m queue time approximations to estimate the cycle time with more accuracy. They propose the Martin Approximation for the cycle time of a G/G/m queue as follows.

$$E(CT) \approx \frac{1}{\mu} + \frac{1}{\mu}\left(\frac{C_a^2 + C_e^2}{2}\right)\left(\frac{\rho^m}{(1-\rho^m)}\right)$$

Applying this formula to the data of our case study toolset approximates the cycle time as shown in Table 11.

Table 11: Comparison of Actual CT and Morrison G/G/m Approximation.

Tool Group	Actual CT (hrs)	G/G/m CT (hrs)	% Difference
1 (tools1-4)	11.94	12.83	7%
2 (tools 5-7)	18.99	27.65	46%

POTENTIAL CAUSES FOR INACCURACY OF CLASSICAL QUEUEING MODELS

We believe this estimation error is due to the fact that the basic assumptions behind such models are far from the reality of semiconductor manufacturing. These assumptions include

1. *The sequence of service times is a sequence of independent and identically distributed random variables.*
2. *Arrival process, service process and WIP level are mutually independent.*
3. *Machine breakdown and maintenance occurs independent of the WIP level.*

PROPOSED MEHODOLOGY FOR CYCLE TIME ESTIMATION OF TOOLSETS WITH CORRELATIONS

Step 1 – Data Collection

For each bucket of time the following variables are estimated from historic data:

- Arrival (IN) - Calculate the total number of lots that have arrived to the system during the course of each time bucket.
- Throughput (OUT) - calculate the total number of lots that have been processed during each time bucket
- Queue Size (WIP) – at the end of each bucket record the WIP level or the total number of lots waiting in the queue at that instance.

Step 2 – Regression and Parameter Estimation

If the length of the buckets are chosen sufficiently short, to quantify the correlation between WIP and service process we need to correlate the WIP of the previous bucket (WIP_Lag) to the output of the current bucket. The reason is that WIP as is calculated through this procedure is a snapshot of the system at the instance when each bucket ends. Hence its effect on the number of outs can only be captured through the following period.

Step 3 – Distribution Fitting

For each value of feasible WIP we can fit a distribution to OUT with the first moment Avg_OUT and the second moment $Avg\_OUT^2 + Std\_OUT^2$. Distribution fitting techniques and goodness of fit tests can be employed to find the best fitting distribution on OUT. Also if we find a significant correlation between WIP_Lag and IN a distribution needs to be fitted in a similar fashion.

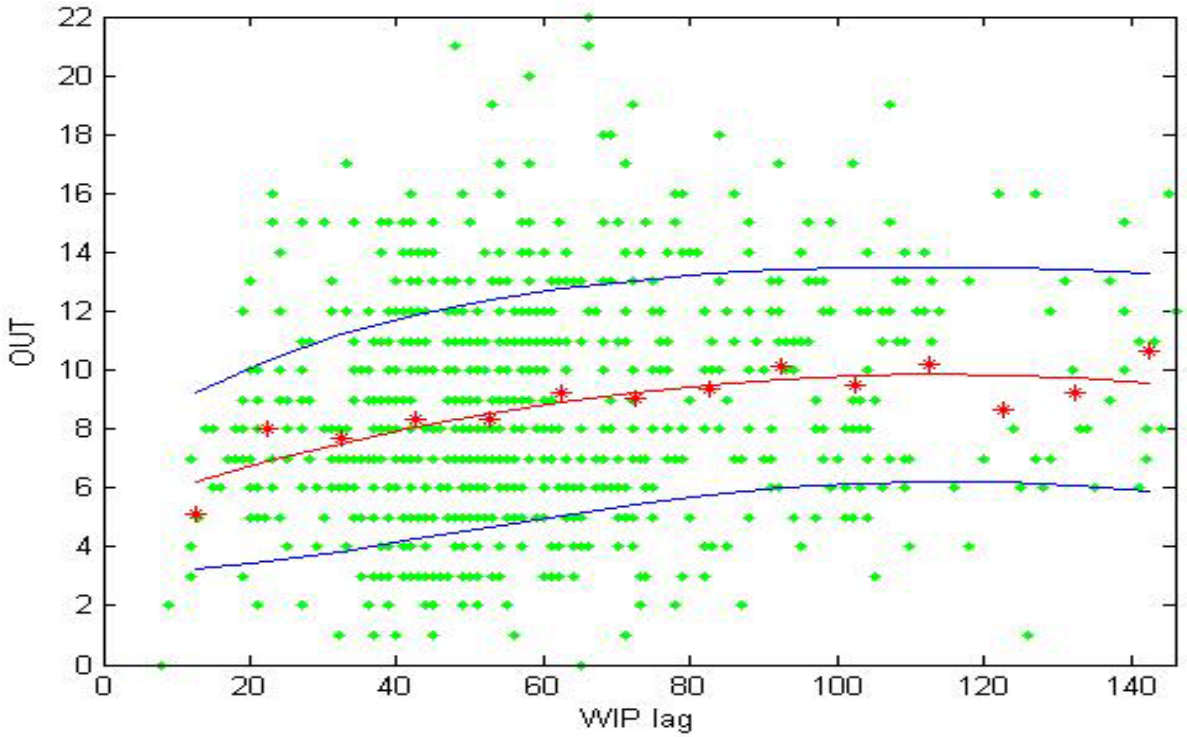

Fig 1: WIP_lag and OUT Relationship

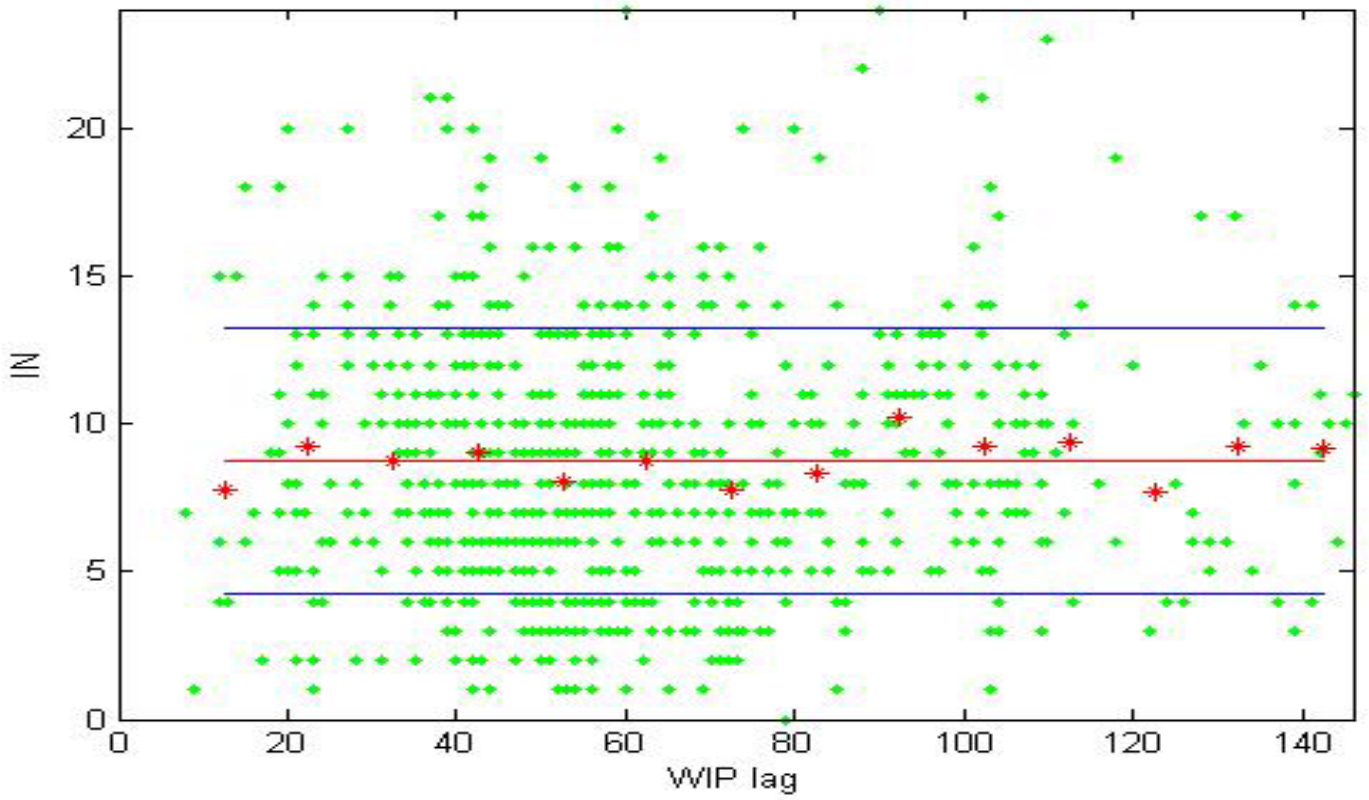

Fig 2: WIP_lag and IN Relationship

Table 12: Results of the iterative algorithm

	Normal Fit		
	WIP	THROUGHPUT	CT (hrs)
Actual	60.94	8.70	13.87
Simulated Mean	64.06	8.81	14.54
Half Width	3.12	0.15	0.71