Dynamic Simulations

Yoo-Suk Hong <yhong@snu.ac.kr>
Dept of Industrial Engineering
Seoul National University

Seila *et. al.*, Chapter 4

Introduction

Static *vs.* Dynamic Simulations

Static: observed at a single point in time.

Dynamic: describes the behavior of a system *over time*.
e.g. Queueing system

Mechanisms for Time Advancing

Fixed time advancing: moves time by a fixed amount Δt.
The time between t and $t + \Delta t$ is called a *period*.

Dynamic time advancing: moves time by a variable amount.
e.g. discrete-even simulation

Queueing — Variability Interactions

$$\left.\begin{array}{l}\text{Process-time varaibility}\\ \text{Flow variability}\end{array}\right\} \Rightarrow \text{Performance} \left\{\begin{array}{l}\text{WIP } (L)\\ \text{Cycle time } (W)\\ \text{Throughput } (\lambda)\end{array}\right.$$

A SIngle-Server Queueing System (SSQS)

1. An arrival process
2. A service process
3. A queue

Queueing Theory

Characterizing *performance measures* in terms of *descriptive parameters*.

- Descriptive parameters: λ (arrival rate), m (number of parallel machines), b (max number of jobs allowed), μ (service rate), etc.
- Performance measures: W_q, W, L, L_q, etc.

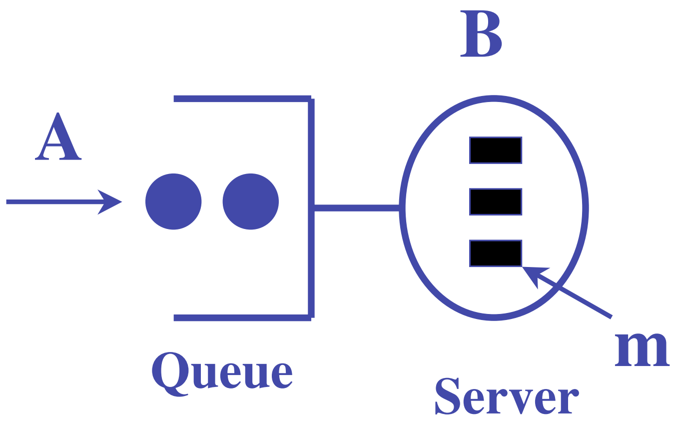
Kendall's Notation

$$\boxed{A/B/m/b}$$

$$A/B = \begin{cases} D \text{ (Deterministic)} \\ M \text{ (Markovian)} \\ G \text{ (General)} \end{cases}$$

m = Number of parallel machines

b = Buffer size



Fundamental Relations

Holds for all *single-station* systems
(i.e., regardless of the assumptions about arrival and process time distributions, number of machines, etc.).

$$\text{Prob of server being busy:} \quad \rho = \frac{\lambda}{\mu} \qquad (1)$$

$$\text{Average time in the system:} \quad W = W_q + \frac{m}{\mu} \qquad (2)$$

$$\text{Average jobs in the system:} \quad L = \lambda \times W \qquad (3)$$

$$\text{Average jobs in the queue:} \quad L_q = \lambda \times W_q \qquad (4)$$

M/M/1 Queue

Assumptions:

- Exponential interarrival times
- A single machine with exponential process times
- FCFS
- Unlimited space for jobs waiting in queue

Memoryless Property: *What information is needed to characterize the future (probabilistic) evolution of the system?*

- $\left.\begin{cases} \text{time since the last arrival} \\ \text{time the current job has been in process} \end{cases}\right\}$ irrelevant!!
- Only the number of jobs currently in the system matters.
- State of the system: n.

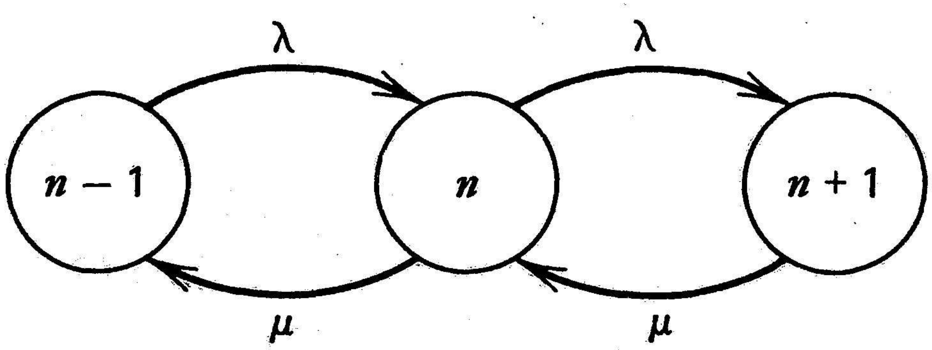
State Transition Analysis — M/M/1

Transition Rates:

- Conditional rates (i.e., given the system is in state n):

$$\begin{cases} n \rightarrow (n+1): & \lambda \\ n \rightarrow (n-1): & \mu \end{cases}$$

- Unconditional (steady-state) rates: $\boxed{p_{n-1}\,\lambda = p_n\,\mu}$



Average Number of Jobs in the System (L)

$$\left\{ \begin{array}{l} p_n = \dfrac{\lambda}{\mu}\, p_{n-1} = \rho\, p_{n-1} \\ p_0 = 1 - \rho \quad \text{(machine idle)} \end{array} \right\} \quad \Rightarrow \quad p_n = \rho^n (1-\rho)$$

$$\begin{aligned} L &= \sum_{n=0}^{\infty} n\, p_n \\ &= \sum_{n=0}^{\infty} n\, \rho^n (1-\rho) \\ &= \rho\,(1-\rho) \sum_{n=1}^{\infty} n\, \rho^{n-1} \qquad \Leftarrow \left(\textstyle\sum_{n=1}^{\infty} n\, \rho^{n-1} = \frac{1}{(1-\rho)^2} \right) \\ &= \frac{\rho}{1-\rho} \end{aligned}$$

Performance Measures — M/M/1

$$
\begin{aligned}
L(M/M/1) &= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \\
W(M/M/1) &= \frac{L(M/M/1)}{\lambda} = \frac{\frac{\lambda}{\mu-\lambda}}{\lambda} = \frac{1}{\mu-\lambda} \\
W_q(M/M/1) &= W(M/M/1) - \frac{1}{\mu} = \frac{\lambda}{\mu\,(\mu-\lambda)} \\
L_q(M/M/1) &= \lambda \cdot W_q(M/M/1) = \frac{\lambda^2}{\mu\,(\mu-\lambda)} \\
&= \frac{\rho^2}{1-\rho}
\end{aligned}
$$

Performance Measures — M/M/1 (Continue)

Observations:

1. L, W, W_q, and L_q are all increasing in ρ.

   Busy systems (ρ) $\Rightarrow$ More congestion (L, L_q)

2. Slower machine (μ) $\Rightarrow$ More waiting time (W, W_q)

3. $\dfrac{1}{1-\rho}$ terms $\Rightarrow$ All measures explode as $\rho \rightarrow 1$.

Waiting Times in a SSQS: Lindley's Formula

X_n = service time of the nth customer
Y_n = time between the arrivals of the nth and $(n+1)$st customers
W_n = waiting time in the queue for the nth customer

$$W_{n+1} = \max(0, W_n + X_n - Y_n)$$

If the $(n+1)$st customer arrives

(a) at the same time as the nth customer, i.e., $Y_n = 0$, he/she has to wait $W_n + X_n$.

(b) after the nth customer has left, i.e., $Y_n > W_n + X_n$, he/she is served right away (that is, does not have to wait).

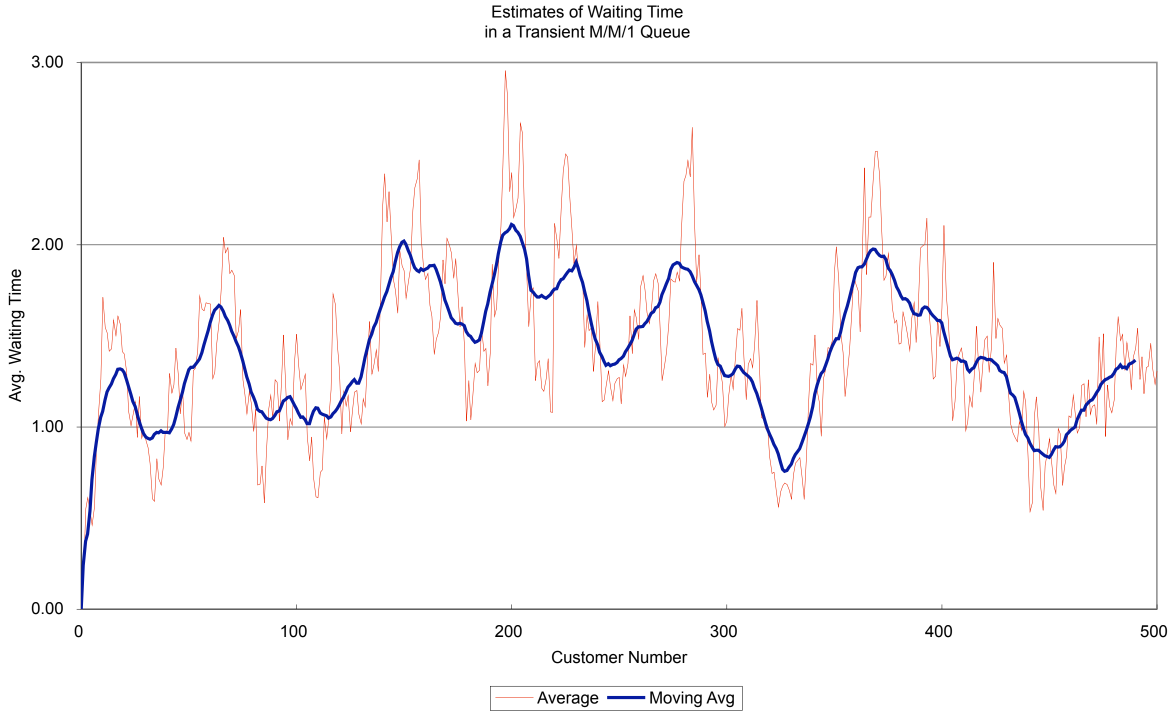
Spreadsheet Simulation of an M/M/1 Queue

	A	B	C	D	E	F	G
3							
4	Mean service time:		0.7				
5	Mean interarrival time:		1.0				
6							
7							
8		Customer number	Waiting Time	Service Time	Interarrival Time		Busy/Idle
9		(n)	(W_n)	(X_n)	(Y_n)	W_n+X_n-Y_n	
10		0	0	0.4328	0.3852	0.0476	0
11		1	0.0476	0.1770	0.0454	0.1792	1
12		2	0.1792	1.3494	0.1230	1.4056	1
13		3	1.4056	0.4659	1.6279	0.2436	1
14		4	0.2436	0.3996	1.0828	-0.4395	1
15		5	0.0000	1.9593	2.8404	-0.8811	0
16		6	0.0000	0.1485	1.7174	-1.5689	0
17		7	0.0000	0.5877	0.2410	0.3467	0
18		8	0.3467	1.8964	1.6122	0.6310	1
19		9	0.6310	1.9285	0.0878	2.4717	1
20		10	2.4717	0.1926	1.5142	1.1500	1
21							

```
D10 =$C$3*LN(RAND()) F10 =C10+D10-E10
E10 =$C$4*LN(RAND()) G10 =IF(C10>0,1,0)
C11 =IF(F10>0,F10,0)
```

Simulation Results (Transient)

Average waiting times over 20 replications consisting of 500 observations each:



Initial Transient Period ("Warm-up Period")

Mean waiting time for a stable M/M/1 queue:

$$W_q = \frac{\lambda}{\mu\,(\mu-\lambda)} = \frac{1.0}{\frac{1}{.7}\,(\frac{1}{.7}-1.0)} = 1.633$$

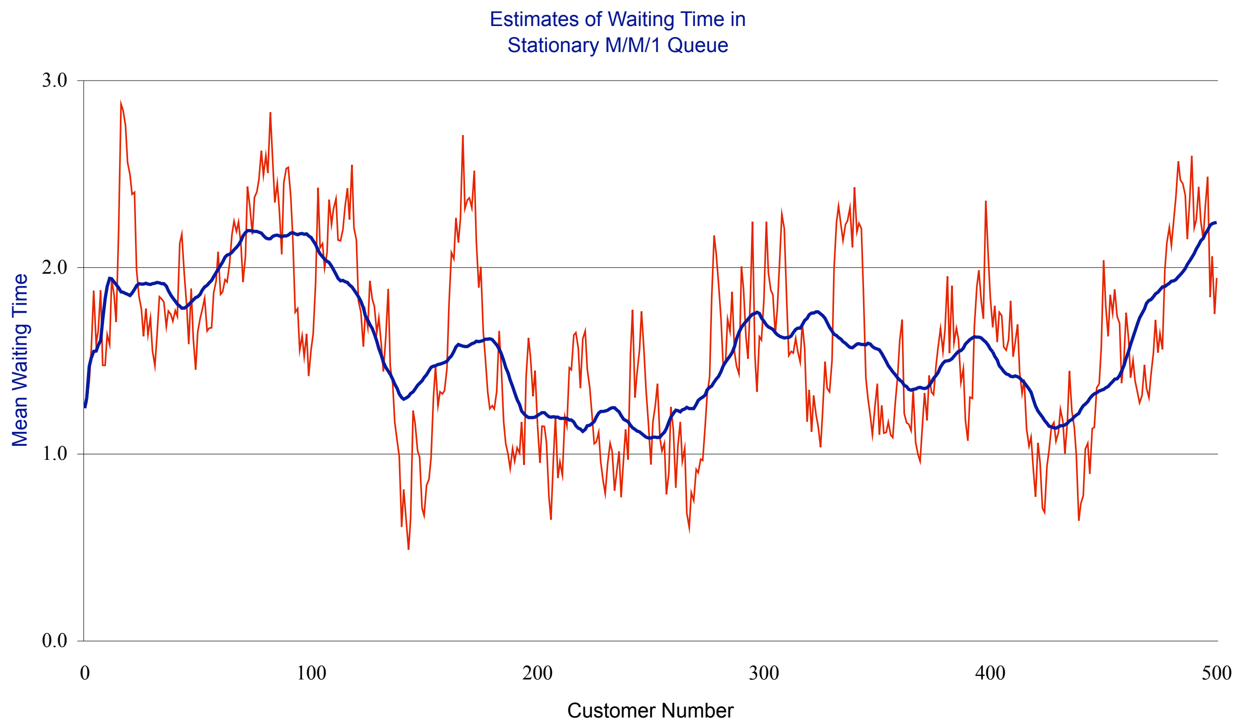
In the first 100 observations, much evidence of stationary behavior. However we cannot be sure exactly where!

Stationary distribution of waiting time:

$$P(W \le w) = \begin{cases} 1-\rho & \text{if } w = 0 \\ 1-\rho\, e^{-(\mu-\lambda)w} & \text{if } w > 0 \end{cases}$$

If the *first* waiting time is chosen from above, all subsequent waiting times will be from the stationary distribution.

Simulation Results (Stationary)



⇒ Same general appearance as much of the previous results.
⇒ The previous graph was fairly close to the stationary behavior within the first 100 observations.

Characteristics of Data from Dynamic Simulations

1. **Initial condition bias**
   - Selecting the appropriate starting condition
   - Discarding the observations recorded during the transient period of simulation
   - Making very long runs (However, in general, longer runs, fewer replications)

2. **Autocorrelated observations**

   In general, the data set $\{y_1, y_2, \ldots\}$ are *not* i.i.d.

   - Replication
   - Batching
   - etc.

Terminating *vs.* Non-terminating Simulations

Terminating: There is a natural event that ends the simulation.

- We typically do not run the terminating simulations long enough for any convergence to take place.
- All the collected data come from the *transient* distribution.
- *Example*: Bak with 5 tellers
  - Opens at 9:30 am and closes at 4:30 pm.
  - Stays open until all customers served.
  - Arrival rate of 1 per min; Service times 4 min.
  - Performance measure: Average customer delay.

Non-terminating: No natural event that ends the simulation.

- We are not interested in the transient distribution.
- We are interested in the *steady-state* distribution.

General Replication (Batching) Structure

Run Number	Y_1	Y_2	$\cdots$	Y_m		Replication Statistic
1	y_{11}	y_{12}	$\cdots$	y_{1m}	$\longrightarrow$	X_1
2	y_{21}	y_{22}	$\cdots$	y_{2m}	$\longrightarrow$	X_2
$\vdots$						$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nm}	$\longrightarrow$	X_n

- Rows are *not* IID, i.e., the data values $\{y_{i1}, y_{i2}, \cdots, y_{im}\}$ are autocorrelated.
- However, columns are IID. The data $\{y_{1j}, y_{2j}, \cdots, y_{nj}\}$ can be considered to be an independent sample from the distribution of random variable Y_j.

$\Rightarrow$ We have independence *across* runs.

Output Analysis for Terminating Simulations

Suppose we take n independent replications, each with the same initial condition and the same terminating event.

Let X_i be a *replication statistic* computed from the ith run:

$$X_i = f(Y_{i1}, Y_{i2}, \cdots, Y_{im}).$$

For example, X_i can be average, sum, maximum, minimum value of the observations obtained from the ith run.

Then we can get a random sample $X_1, X_2, \cdots, X_n$ of size n.

Q: How can we assume the "independence" of X_i's?

A: Simulation was replicated, each time with independent random numbers.

Obtaining a Specified Precision: Background

We've learned in Ch. 2 that the confidence interval for mean μ is

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \quad \text{where} \quad S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}.$$

If we want a $(1-\alpha)$ probability that one estimate of μ differs by an amount no greater than β, how many replications do we need?

$$1-\alpha = P(|\overline{X} - \mu| \le \beta) = P(\overline{X} - \beta \le \mu \le \overline{X} + \beta)$$

Thus, we want our $(1-\alpha)100\%$ CI half width to be β.

$$\beta = t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \quad \Rightarrow \quad n = S^2\left(\frac{t_{\alpha/2,n-1}}{\beta}\right)^2$$

Note: Both S and $t_{\alpha/2,n-1}$ are dependent on n.

Obtaining a Specified Precision: Procedure

From the initial n replications ($n \geq 30$), compute S^2.

And then, assume S^2 is fixed, i.e., it will not change significantly with additional replications.

<u>Procedure</u>

(a) Perform $n \geq 30$ replications and compute S^2.

(b) Compute $n^* = \min_{i \geq n} \left\{ t_{\alpha/2, i-1} \dfrac{S}{\sqrt{i}} \leq \beta \right\}$.

(Increase i by 1 until a value of i is obtained.)

(c) Take additional $(n^* - n)$ replications and compute the CI.

Obtaining a Specified Precision: Example

From the initial 10 replications, the values for replication statistics X_i's were obtained as
$\{1.53, 1.66, 1.24, 2.34, 2.00, 1.69, 2.69, 2.86, 1.70, 2.60\}$.

Find n such that, with 95% probability, the absolute error is no greater than $\beta = 0.25$.

We calculated $\overline{x} = 2.03$ and $s = 0.555$.

i	$t_{0.05/2,i-1}$	Half width
11	1.833	.322
⋮		
16	1.753	.253
17	1.746	**.235**

$n^* = 17$.

Output Analysis for Non-Terminating Simulations

We'd like to focus on the *steady-state* behavior of the system.
$\Rightarrow$ Data are usually obtained from a *single long* simulation run.
$\Rightarrow$ *Individual observations* are used, rather than replication stats.
$\Rightarrow$ The independence of data *cannot* be easily assumed.

3 approaches for analyzing non-terminating simulations:

- Replication
  - How to deal with the initial condition bias?
  - Wider confidence intervals due to fewer replications
- Batching
  - Pseudo-independence (if the batches are sufficiently large)
  - Robustness in deleting data for initial condition bias
- Using individual raw observations with autocorrelation
  - Auto-correlogram, etc.

Batch Means Method

(1) Group observations into n equal, non-overlapping batches, each of size m.

(2) Compute the sample mean of each batch. The ith batch mean is

$$X_i = \frac{\sum_{j=1}^{m} Y_{ij}}{m}, \quad i = 1, 2, \cdots, n$$

If batches are sufficiently large, X_i's are approximately independent, even though the observation at the end of batch i are correlated with the one at the beginning of batch $i+1$.

(3) Compute the confidence interval from these batch means. (Traditional statistical methods can apply thanks to pseudo-independence.)

Some Remarks on Batch Means Method

Batch size should be

1. large enough that batch means are approximately uncorrelated.
   $\Rightarrow$ *Batch size* to be at least 10 times as large as the largest lag
2. small enough that the maximum number of batches is formed.
   $\Rightarrow$ *Number of batches* to be at least 10

Replication *vs.* Batching

1. Independent replications run through the transient period in *each replication*. Batch means method requires this *only once*.
2. Errors in determining the transient period will cause the sample mean in each replication to be biased. Batch means is robust in that the bias will reduce in successive batches.