

EEC 686/785 Modeling & Performance Evaluation of Computer Systems

Lecture 22

Wenbing Zhao

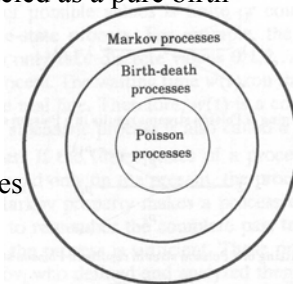
Department of Electrical and Computer Engineering
Cleveland State University

wenbing@ieee.org

(based on Dr. Raj Jain's lecture notes)

Relationship Among Stochastic Processes

- **Markov process** is broader than **birth-death process** and **birth-death process** is broader than **Poisson process**
- The Poisson process can be modeled as a pure birth process with constant birth rate
- All birth-death processes are Markov processes with the restriction that the transitions are restricted to neighboring states



23 November 2005

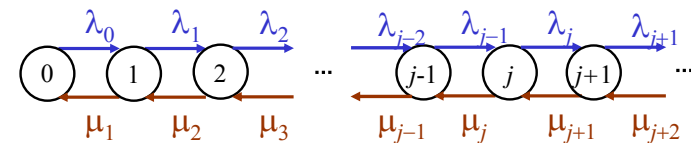
EEC686/785

Outline

- Review of lecture 21
- Analysis of a single queue (part II)

Birth-Death Processes

- State-transition diagram



- In state n :
 - New arrivals take place at a rate λ_n
 - The service rate is μ_n
- Both the interarrival times and service times are assumed exponentially distributed

23 November 2005

EEC686/785

Wenbing Zhao

23 November 2005

EEC686/785

Wenbing Zhao

Theorem: State Probability

- The steady-state probability p_n of a birth-death process being in state n (i.e., there are n jobs in the system) is given by:

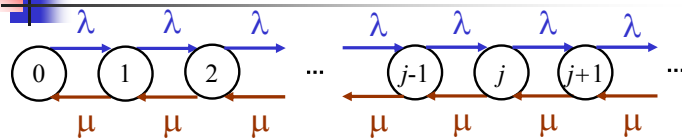
$$p_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} p_0 \quad n = 1, 2, \dots, \infty$$

- Here, p_0 is the probability of being in the zero state (no job in system, i.e., system is idle)

Results for M/M/1 Queues

- Utilization of the server: $U = 1 - p_0 = \rho$
- Mean number of jobs in the system: $E[n] = \frac{\rho}{1 - \rho}$
- Variance of the number of jobs: $Var[n] = \frac{\rho}{(1 - \rho)^2}$
- Mean response time: $E[r] = \frac{1/\mu}{1 - \rho}$
- Cumulative distribution function of the response time: $F(r) = 1 - e^{-r\mu(1-\rho)}$

Results for M/M/1 Queues



- Birth-death processes with $\lambda_n = \lambda \quad n = 0, 1, 2, \dots, \infty$
 $\mu_n = \mu \quad n = 1, 2, \dots, \infty$

- Probability of n jobs in the system

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \quad n = 1, 2, \dots, \infty$$

$\rho = \lambda/\mu$: **traffic intensity**

Results for M/M/1 Queues

- Cumulative distribution function of the waiting time

$$F(w) = 1 - \rho e^{-w\mu(1-\rho)}$$

- This is a truncated exponential distribution. Its q -percentile is given by

$$w_q = \frac{1}{\mu(1-\rho)} \ln\left(\frac{100\rho}{100-q}\right)$$

- The above formula applies only if q is greater than $100(1-\rho)$. All lower percentiles are 0

$$w_q = \max\left\{0, \frac{E[w]}{\rho} \ln\left(\frac{100\rho}{100-q}\right)\right\}$$

Results for M/M/1 Queues

- Mean number of jobs in the queue

$$E[n_q] = \sum_{n=1}^{\infty} (n-1)P_n = \sum_{n=1}^{\infty} (n-1)(1-\rho)\rho^n = \frac{\rho^2}{1-\rho}$$

- **Busy period:** the time interval between two successive idle intervals

Example 31.1

- Arrival rate $\lambda = 125$ pps
- Service rate $\mu = 1/0.002 = 500$ pps
- Gateway utilization $\rho = \lambda/\mu = 0.25$
- Probability of n packets in the gateway = $(1-\rho)\rho^n = 0.75(0.25)^n$
- Mean number of packets in the gateway = $\rho/(1-\rho) = 0.25/0.75 = 0.33$
- Mean time spent in the gateway = $(1/\mu)(1-\rho) = (1/500)/(1-0.25) = 2.66$ milliseconds

Example 31.1

- On a network gateway, measurements show that the packets arrive at a mean rate of 125 packets per second (pps) and the gateway takes about two milliseconds to forward them. Using an M/M/1 model, analyze the gateway
 - What is the probability of buffer overflow if the gateway had only 13 buffers?
 - How many buffers do we need to keep packet loss below one packet per million?

Example 31.1

- Probability of buffer overflow =
 $P(\text{more than 13 packets in gateway}) = \rho^{13} = 0.25^{13} = 1.49 \times 10^{-8} \approx 15$ packets per billion packets
- To limit the probability of loss to less than 10^{-6} :
 $\rho^n \leq 10^{-6}$ or: $n > \log(10^{-6})/\log(0.25) = 9.96$
 We need about 10 buffers
- The last two results about buffer overflow are approximate

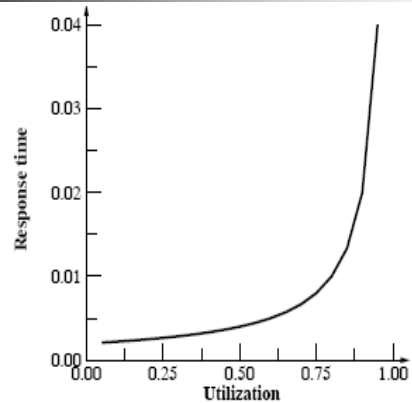
Example 31.1

- Strictly speaking, the gateway should actually be modeled as a finite buffer M/M/1/B queue
- Since the utilization is low and the number of buffers is far above the mean queue length, the results obtained are a close approximation
- For an M/M/1 queue to be stable, the traffic intensity ρ must be less than 1

M/M/m Queue

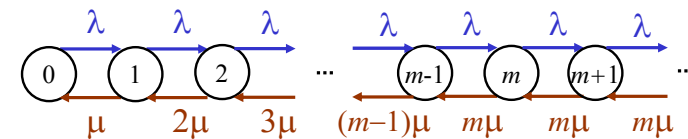
- Used to model multiprocessor systems or devices that have several identical servers and all jobs waiting for these server are kept in one queue
- m identical servers each with a service rate of μ jobs per unit time
- The arrival rate is λ jobs per unit time
- If any of the m servers are idle, the arriving job is serviced immediately
- If all m servers are busy, the arriving jobs wait in a queue

Example 31.1



M/M/m Queue

- The state of the system is represented by the number of jobs n in the system



Analysis of M/M/m Queue

- Number of jobs in the system is a birth-death process:

$$\lambda_n = \lambda \quad n = 0, 1, 2, \dots, \infty$$

$$\mu_n = \begin{cases} n\mu & n = 1, 2, \dots, m-1 \\ m\mu & n = m, m+1, \dots, \infty \end{cases}$$

Analysis of M/M/m Queue

- Probability of zero jobs in the system:

$$\sum_{n=0}^{\infty} p_n = 1$$

$$\Rightarrow p_0 + p_0 \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + p_0 \frac{(m\rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m} = 1$$

$$\Rightarrow p_0 = \left[1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

Analysis of M/M/m Queue

- Probability of n jobs in the system:

$$p_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} p_0 & n = 1, 2, \dots, m-1 \\ \frac{\lambda^n}{m! m^{n-m} \mu^n} p_0 & n = m, m+1, \dots, \infty \end{cases}$$

In terms of the traffic intensity $\rho = \lambda/m\mu$, we have:

$$p_n = \begin{cases} \frac{(m\rho)^n}{n!} p_0 & n = 1, 2, \dots, m-1 \\ \frac{\rho^n m^m}{m!} p_0 & n = m, m+1, \dots, \infty \end{cases}$$

Analysis of M/M/m Queue

- Probability \mathcal{G} that an arriving job has to wait in the queue (\mathcal{G} really should be ρ , can't input it in PointPoint):

$$\mathcal{G} = P(\geq m \text{ jobs})$$

$$= p_m + p_{m+1} + p_{m+2} + \dots$$

$$= p_0 \frac{(m\rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m} = p_0 \frac{(m\rho)^m}{m!(1-\rho)}$$

This is known as **Erlang's C formula**. Notice that for $m = 1$, $\mathcal{G} = \rho$

Analysis of M/M/m Queue

- Mean number of jobs in the queue $E[n_q]$:

$$\begin{aligned} E[n_q] &= \sum_{n=m+1}^{\infty} (n-m)p_n \\ &= p_0 \frac{(m\rho)^m}{m!} \sum_{n=m+1}^{\infty} (n-m)\rho^{n-m} \\ &= p_0 \frac{(m\rho)^m \rho}{m!(1-\rho)^2} = \frac{\rho\mathcal{G}}{1-\rho} \end{aligned}$$

Analysis of M/M/m Queue

- Expected number of jobs in the system:

$$E[n] = E[n_q] + E[n_s] = m\rho + \frac{\rho\mathcal{G}}{1-\rho}$$

- Variance of n and n_q can similarly be shown as

$$\text{Var}[n] = m\rho + \rho\mathcal{G} \left[\frac{1 + \rho - \rho\mathcal{G}}{(1-\rho)^2} + m \right]$$

$$\text{Var}[n_q] = \frac{\rho\mathcal{G}(1 + \rho - \rho\mathcal{G})}{(1-\rho)^2}$$

Analysis of M/M/m Queue

- Expected number of jobs in service

$$\begin{aligned} E[n_s] &= \sum_{n=1}^{m-1} np_n + \sum_{n=m}^{\infty} mp_n \\ &= 1p_0 \frac{(m\rho)}{1!} + 2p_0 \frac{(m\rho)^2}{2!} + \dots + (m-1)p_0 \frac{(m\rho)^{m-1}}{(m-1)!} \\ &\quad + m(p_m + p_{m+1} + p_{m+2} + \dots) \\ &= m\rho(p_0 + p_0 \frac{(m\rho)}{1!} + p_0 \frac{(m\rho)^2}{2!} + \dots + p_0 \frac{(m\rho)^{m-2}}{(m-2)!}) + m\mathcal{G} \\ &= m\rho(p_0 + p_1 + p_2 + \dots + p_{m-2}) + m\mathcal{G} \\ &= m\rho(1 - p_{m-1} - \mathcal{G}) + m\mathcal{G} \\ &= m\rho - m\rho p_{m-1} + m\mathcal{G}(1 - \rho) \\ &= m\rho \quad \text{since } m\mathcal{G}(1 - \rho) = mp_m = m\rho p_{m-1} \end{aligned}$$

Analysis of M/M/m Queue

- In T seconds

- Total number of jobs arriving and getting service will be λT
- Total busy time of m servers to service these jobs will be $\lambda T / \mu$
- Utilization of each server:

$$U = \frac{\text{Busy time per server}}{\text{Total time}} = \frac{(\lambda T / \mu) / m}{T} = \frac{\lambda}{m\mu}$$

Analysis of M/M/m Queue

- Mean response time using Little's law

$$E[r] = \frac{E[n]}{\lambda} = \frac{1}{\mu} + \frac{\mathcal{G}/m\mu}{1-\rho} = \frac{1}{\mu} \left(1 + \frac{\mathcal{G}}{m(1-\rho)} \right)$$

- Mean waiting time

$$E[w] = \frac{E[n_q]}{\lambda} = \frac{\mathcal{G}}{m(1-\rho)}$$

Analysis of M/M/m Queue

- Cumulative distribution function of the waiting time:

$$F[w] = 1 - \mathcal{G}e^{-m\mu(1-\rho)w}$$

since w has a truncated exponential distribution function, its q -percentile:

$$w_q = \max \left\{ 0, \frac{1}{m\mu(1-\rho)} \ln \left(\frac{100\mathcal{G}}{100-q} \right) \right\}$$

If the probability of queueing \mathcal{G} is less than $1-q/100$, the second term in the above equation can be negative. The correct answer in those cases is 0

Analysis of M/M/m Queue

- Cumulative distribution function of time response time:

$$F[r] = \begin{cases} 1 - e^{-\mu r} - \frac{\mathcal{G}}{1-\rho} e^{-m\mu(1-\rho)r} - e^{-\mu r} & \rho \neq (m-1)/m \\ 1 - e^{-\mu r} - \mathcal{G}\mu r e^{-\mu r} & \rho = (m-1)/m \end{cases}$$

($r > 0$)

- Notice that the response time r is not exponentially distributed unless $m=1$
- In general, the coefficient of variation of r is less than 1
 - C.o.v = standard deviation / expected value

Example 31.2

- Students arrive at the university computer center in a Poisson manner at an average rate of ten per hour. Each student spends an average of 20 minutes at the terminal and the time can be assumed to be exponentially distributed. The center currently has five terminals. Some students have been complaining that waiting times are too long. Let us analyze the center usage using a queueing model
 - M/M/5 queueing system with $\lambda = 1/6$ per minute
 - $\mu = 1/20$ per minute

Example 31.2

- Traffic intensity = $\rho = \lambda/m\mu = 0.167/(5 \times 0.05) = 0.67$
- Probability of all terminals being idle is:

$$p_0 = \left[1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

$$p_0 = \left[1 + \frac{(5 \times 0.67)^5}{5!(1-0.67)} + \frac{(5 \times 0.67)^1}{1!} + \frac{(5 \times 0.67)^2}{2!} + \frac{(5 \times 0.67)^3}{3!} + \frac{(5 \times 0.67)^4}{4!} \right]^{-1}$$

$$= 0.0318$$

Example 31.2

- Average number of students waiting in the queue is:

$$E[n_q] = \frac{\rho g}{1-\rho} = \frac{0.67 \times 0.33}{1-0.67} = 0.65$$

- Average number of students using the terminals is:

$$E[n_s] = E[n] - E[n_q] = 4 - 0.65 = 3.35$$

Example 31.2

- Probability of all terminals being busy is:

$$g = p_0 \frac{(m\rho)^m}{m!(1-\rho)} = 0.0318 \frac{(5 \times 0.67)^5}{5!(1-0.67)} = 0.33$$

- Average terminal utilization = $\rho = 0.67$
- Average number of students in the center is:

$$E[n] = m\rho + \frac{\rho g}{1-\rho} = 5 \times 0.67 + \frac{0.67 \times 0.33}{1-0.67} = 4.02$$

Example 31.2

- The mean and variance of the time spent in the center are:

$$E[r] = \frac{1}{\mu} \left(1 + \frac{g}{m(1-\rho)} \right) = \frac{1}{0.05} \left(1 + \frac{0.33}{5(1-0.67)} \right) = 24$$

$$\text{Var}[r] = \frac{1}{\mu^2} \left(1 + \frac{g(2-g)}{m^2(1-\rho)^2} \right) = \frac{1}{0.05^2} \left(1 + \frac{0.33(2-0.33)}{5^2(1-0.67)^2} \right) = 479$$

Thus, each student spends an average of 24 minutes in the center = 20 minutes working + 4 minutes waiting

Example 31.2

- Mean waiting time:

$$E[w] = \frac{\rho}{\mu m (1 - \rho)} = \frac{0.33}{5 \times 0.05 \times (1 - 0.67)} = 4$$

- 90-percentile of the waiting time is:

$$\max\left\{0, \frac{E[w]}{\rho} \ln(10\rho)\right\} = \max\left\{0, \frac{4}{0.33} \ln(10 \times 0.33)\right\} = 14$$

- 10% of the students wait more than 14 minutes

Example 31.3

- With $m = 6$ terminals:

- Traffic intensity: $\rho = 0.167 / (6 \times 0.05) = 0.556$
- Probability of all terminals being idle = $p_0 = 0.0346$
- Probability of all terminals being busy = $\rho = 0.15$
- Average waiting time = $E[w] = 1.1$ minutes
- The 90-percentile of waiting time is:

$$\max\left\{0, \frac{1.1}{0.15} \ln(10 \times 0.15)\right\} = 3.0$$

⇒ With just one more terminal we will be able to satisfy the students' demands

Example 31.3

- The students would like to limit their waiting time to an average of two minutes and no more than five minutes in 90% of the cases. Is it feasible? If yes then how many terminals are required?
- $\lambda = 0.167$ and $\mu = 0.05$

Example 31.4

- Consider what would have happened if the five terminals in example 31.2 were located in five different locations on the campus, thereby needing a separate queue for each
- Five separate M/M/1 queues
- $m = 1$, $\lambda = 0.167/5 = 0.0333$, and $\mu = 0.05$
- Traffic intensity: $\rho = 0.0333/0.05 = 0.67$

Example 31.4

- The mean time spent in the terminal room is:

$$E[r] = \frac{1/\mu}{1-\rho} = \frac{1/0.05}{1-0.67} = 60$$

- The variance of the time spent in the terminal room is:

$$\text{Var}[r] = \frac{1/\mu^2}{(1-\rho)^2} = \frac{1/0.05^2}{(1-0.67)^2} = 3600$$

- Compare this to the mean of 24 minutes and a variance of 479 in example 31.2 => **single queue alternative is better**

M/M/∞ Queue

- **Infinite servers.** Jobs never wait
- Response time = service time
- Mean response time = mean service time regardless of the arrival rate
- => Called **delay centers**
- Used to represent dedicated resources, such as terminals in timesharing systems
- Properties of such queues can be easily derived from those for M/M/m queues
- Also see results for M/G/∞ queues

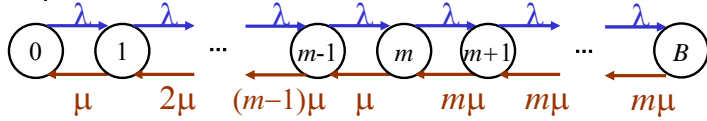
Example 31.4

- Walking time was ignored. Having several terminal rooms distributed across the campus may reduce the walking time considerably
- If all jobs are identical, it is better to have just one queue than to have multiple queues
- If some students need very short terminal sessions and others need very long sessions, separate queues may be better

M/M/m/B Queue with Finite Buffers

- Similar to the M/M/m queue but number of buffers B is finite
- All arrivals, after B buffers are full, are lost
- B is greater than or equal to m; otherwise some servers will never be able to operate => M/M/B/B queue

Analysis of M/M/m/B Queue



- Birth-death process:

$$\lambda_n = \lambda \quad n = 0, 1, 2, \dots, B-1$$

$$\mu_n = \begin{cases} n\mu & n = 1, 2, \dots, m-1 \\ m\mu & n = m, m+1, \dots, B \end{cases}$$

Analysis of M/M/m/B Queue

- Probability of zero jobs in the system:

$$\sum_{n=0}^B p_n = 1$$

$$p_0 + p_0 \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + p_0 \frac{(m\rho)^m}{m!} \sum_{n=m}^B \rho^{n-m} = 1$$

$$p_0 = \left[1 + \frac{(1 - \rho^{B-m+1})(m\rho)^m}{m!(1 - \rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

Analysis of M/M/m/B Queue

- Probability of n jobs in the system:

$$p_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} p_0 & n = 1, 2, \dots, m-1 \\ \frac{\lambda^n}{m! m^{n-m} \mu^n} p_0 & n = m, m+1, \dots, B \end{cases}$$

In terms of the traffic intensity $\rho = \lambda/m\mu$:

$$p_n = \begin{cases} \frac{(m\rho)^n}{n!} p_0 & n = 1, 2, \dots, m-1 \\ \frac{\rho^n m^m}{m!} p_0 & n = m, m+1, \dots, B \end{cases}$$

Analysis of M/M/m/B Queue

- Mean number of jobs in the system $E[n]$:

$$E[n] = \sum_{n=1}^B n p_n$$

- Mean number of jobs in the queue $E[n_q]$:

$$E[n_q] = \sum_{n=m+1}^B (n-m) p_n$$

Analysis of M/M/m/B Queue

- Variance and other statistics on n and n_q can be similarly computed
- All arrivals occurring when the system is in the state $n=B$ are lost. Rate of the jobs actually entering the system, called **effective arrival rate**, is:

$$\lambda' = \sum_{n=0}^{B-1} \lambda p_n = \lambda \sum_{n=0}^{B-1} p_n = \lambda(1 - p_B)$$

The difference $\lambda - \lambda' = \lambda p_B$ represents the **loss rate**

Analysis of M/M/m/B Queue

- Utilization of each server:

$$U = \frac{\text{Busy time per server}}{\text{Total time}} = \frac{(\lambda' T / \mu) / m}{T} = \frac{\lambda'}{m\mu} = \rho(1 - p_B)$$

- Probability of the full system = p_B

Analysis of M/M/m/B Queue

- Mean response time (using Little's law):

$$E[r] = \frac{E[n]}{\lambda'} = \frac{E[n]}{\lambda(1 - p_B)}$$

Similarly, the mean waiting time is:

$$E[w] = \frac{E[n_q]}{\lambda'} = \frac{E[n_q]}{\lambda(1 - p_B)}$$

Analysis of M/M/m/B Queue

- For a $M/M/m/m$ system ($B=m$), loss probability is:

$$p_m = \frac{(m\rho)^m}{m!} p_0 = \frac{(m\rho)^m / m!}{\sum_{j=0}^m \frac{(m\rho)^j}{j!}}$$

This is **Erlang's loss formula**

- Used to compute the probability of lost phone calls. Valid also for $M/G/m/m$ queues

Example 31.5

- Consider the gateway of example 31.1 again. Let us analyze the gateway assuming it has only two buffers. The arrival rate and the service rate, as before, are 125 pps and 500 pps, respectively.
- In this case: $\lambda=125$, $\mu=500$, $m=1$, and $B=2$
- Traffic intensity: $\rho=\lambda/m\mu=125/(1\times 500)=0.25$

Example 31.5

- Mean number of jobs in the system:

$$E[n] = \sum_{n=1}^B np_n = 1 \times 0.19 + 2 \times 0.0476 = 0.29$$
- Mean number of jobs in the queue:

$$E[n_q] = \sum_{n=m}^B (n-m)p_n = (2-1) \times 0.0476 = 0.0476$$
- Effective arrival rate in the system:

$$\lambda' = \lambda(1-p_B) = 125(1-p_2) = 125(1-0.047) = 119\text{pps}$$

Example 31.5

- p_n for $n=1,2,\dots,B$ are:
 - $p_1 = \rho p_0 = 0.25p_0$
 - $p_2 = \rho^2 p_0 = 0.25^2 p_0 = 0.0625p_0$
 - p_0 is determined by summing all probabilities:
 $p_0 + p_1 + p_2 = 1 \Rightarrow p_0 + 0.25p_0 + 0.0625p_0 = 1$
 $p_0 = 1 / (1+0.25+0.0625) = 0.76$
 Substituting p_0 for in p_n , we get:
 $p_1 = 0.25p_0 = 0.19$
 $p_2 = 0.0625p_0 = 0.0476$

Example 31.5

- Packet loss rate = $\lambda - \lambda' = 125 - 119 = 6\text{pps}$
- Mean response time

$$E[r] = \frac{E[n]}{\lambda'} = \frac{0.29}{119} = 2.40 \times 10^{-3} \text{ seconds}$$
- Mean time waiting in the queue

$$E[w] = \frac{E[n_q]}{\lambda'} = \frac{0.0476}{119} = 4.0 \times 10^{-4} \text{ seconds}$$



Example 31.5

- Variance and other statistics for the number of jobs in the system can also be computed since the complete probability mass function p_n is known. For example:

$$\begin{aligned}\text{Var}[n] &= E[n^2] - (E[n])^2 \\ &= (1^2 \times 0.19 + 2^2 \times 0.0476) - (0.29^2) = 0.2963\end{aligned}$$