

CSC0056 Data Communication

Beyond M/M/1

Instructor: Chao Wang

Networked Cyber-Physical Systems Laboratory
Department of Computer Science and Information Engineering
National Taiwan Normal University

Nov. 1, 2024



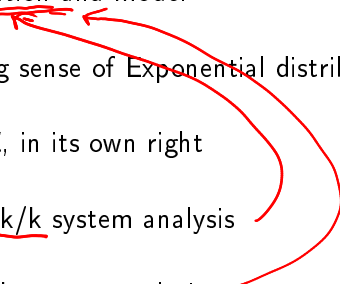
NATIONAL TAIWAN NORMAL UNIVERSITY

References

Harchol-Balter, Mor. Performance modeling and design of computer systems: queueing theory in action. Cambridge University Press, 2013. ISBN 9781107027503.

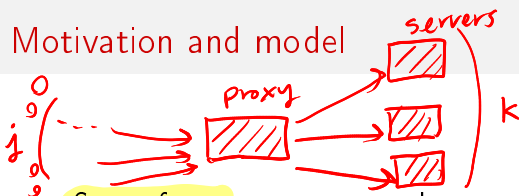
- M/M/k and M/M/k/k
 - Chapter 14
- Background
 - More on exponential distribution
 - Section 11.3
 - More on DTMC and CTMC
 - Sections 9.6 and 9.7; Chapter 12

Outline

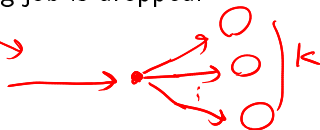
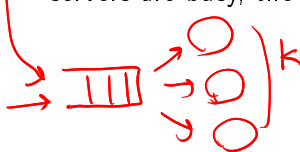
- 1 Motivation and model
 - 2 Making sense of Exponential distribution
 - 3 CTMC, in its own right
 - ✓ 4 M/M/k/k system analysis
 - ✓ 5 M/M/k system analysis
 - 6 Concluding remarks
- 

Motivation and model

buffer bloat



- **Server farms**: one server works as a proxy and redirects jobs to one of a set of servers.
 - Similar setting in broker-based data communication.
 - Two assumptions: the proxy can or **cannot** keep pending jobs.
- **M/M/k**: like M/M/1, but now assume that we have k servers, each having service rate μ .
- **M/M/k/k**: like M/M/k, but now assume that the system has no queue, and thus there are at most k jobs in the system; when all servers are busy, the new arriving job is dropped.



Questions that we will be able to answer

- ① If the proxy cannot keep pending jobs, then given some specific service rate and arrival rate, what would be the fraction of jobs that are lost?
- ② If the proxy can keep pending jobs, then given some specific service rate and arrival rate, what would be the expected number of busy servers at each point of time? What would be the probability that a job has to wait at the proxy? What would be the expected number of jobs waiting at the proxy?
- ③ Design question: Given the same arrival rate λ and total service rate $k\mu$, which of the following three designs would give a shorter mean response time?
 - ① Time-division multiplexing on one single server of service rate $k\mu$.
 - ② Statistical multiplexing on one single server of service rate $k\mu$.
 - ③ Redirecting arrivals to k servers, each having service rate μ .

Exponential distribution, from Geometric distribution (Section 11.3)



- Exponential distribution can be thought of as an approximation of the Geometric distribution (number of coin flips until success):
 - Consider Bernoulli trials. In each trial, let p be the success probability.
 - Consider a parameter $\lambda = np$, where n is the number of trials performed within each unit of time. In this sense, from $p = \lambda/n$, we can think of λ as rate, and that p gets smaller as n gets larger.
 - Let random variable T represent the time until success. Then

$$\begin{aligned}
 P\{T > \underbrace{t}_{\substack{\uparrow t \text{ time units}}} \} &= P\{\text{at least } tn \text{ failures}\} \\
 &= (1-p)^{tn} \quad \text{Geometric distribution} \\
 &= ((1 + (-p))^{-1/p})^{-\lambda t} \quad (\text{since } tn = \lambda t/p) \\
 &= e^{-\lambda t} \quad (\text{since by definition } e = \lim_{y \rightarrow 0} (1+y)^{1/y}).
 \end{aligned}$$

The last line assumes $n \rightarrow \infty$. Since $p = \lambda/n$, we have $p \rightarrow 0$.

CTMC definition (compare it to DTMC defined in Week 6)

Definition (12.2)

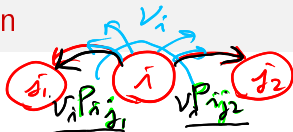
A Continuous-Time Markov Chain (CTMC) is a continuous-time stochastic process $\{X(t), t \geq 0\}$ s.t., $\forall s, t \geq 0$ and $\forall i, j, x(u)$,

$$\begin{aligned} P\{X(t+s) = j | X(s) = i, \underbrace{X(u) = x(u), 0 \leq u \leq s}_{\text{history}}\} \\ = P\{X(t+s) = j | X(s) = i\} \text{ (by Markovian property)} \\ = P\{X(t) = j | X(0) = i\} = P_{ij}(t) \text{ (stationarity),} \end{aligned}$$

- Let τ_i be the time until the CTMC leaves state i , given that the CTMC is currently in state i . Then because of the Markovian property, this implies that the amount of time the process spends in state i before making a transition is Exponentially distributed with some rate (call it ν_i), denoted as $\tau_i \sim \text{Exp}(\nu_i)$.
- When the process leaves state i , it will enter state j with some probability (call it p_{ij}) independent of τ_i .

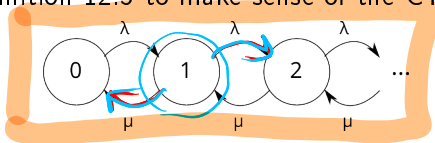
CTMC, alternative definition

Definition (12.5)



Using the variables in the previous page, we may define a CTMC like this:
 Let $X_i \sim \text{Exp}(\nu_i p_{ij})$ represent the time to transition from i to j , $\forall j \neq i$.
 Then $\tau_i = \min_j X_j$. Suppose the minimum is X_m . Then the next state is m .

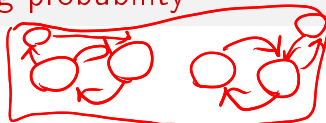
- To prove the equivalence of definitions, apply Theorems 11.3 and 11.4 and the δ -step argument in Section 11.3 in the textbook.
- We may use Definition 12.5 to make sense of the CTMC for M/M/1:



note that $\tau_i \sim \text{Exp}(\lambda + \mu)$.

- The **structure** of the above CTMC, i.e., transitions are defined only between consecutive states, is referred to as a **birth-death process**.

Time reversibility and limiting probability



Definition (9.3)

A Markov chain is **irreducible** if all its states communicate with each other.

Lemma (14.2)

Given an **irreducible** CTMC, let q_{ij} be the rate of transitions from state i to state j given that the current state is i . Suppose we can find x_i 's such that

$$\sum_i x_i = 1 \quad \text{and} \quad x_i q_{ij} = x_j q_{ji}, \quad \forall i, j.$$

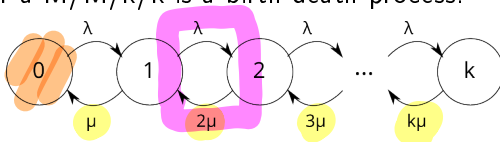
← time-reversibility equations

Then x_i 's are the limiting probabilities of the CTMC. And such a CTMC is said to be *time-reversible*.

Every birth-death process is time-reversible. Therefore, we may use the time-reversibility equations in Lemma 14.2 to find the limiting probabilities.

Analyzing M/M/k/k

- The CTMC for a M/M/k/k is a birth-death process:



State	Time-reversibility equation	Simplified equation
0	$\pi_0 \lambda = \pi_1 \mu$	$\pi_1 = \frac{\lambda}{\mu} \pi_0$
1	$\pi_1 \lambda = \pi_2 2\mu$	$\pi_2 = \left(\frac{\lambda}{\mu}\right)^2 \frac{1}{2!} \pi_0$
2	$\pi_2 \lambda = \pi_3 3\mu$	$\pi_3 = \left(\frac{\lambda}{\mu}\right)^3 \frac{1}{3!} \pi_0$
$k-1$	$\pi_{k-1} \lambda = \pi_k k\mu$	$\pi_k = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \pi_0$

Along with $\sum_{i=0}^k \pi_i = 1$, we may obtain the closed form of π_0 . Try it before you turn the page..

The blocking probability in M/M/k/k

$\pi_0 = 1 / \sum_{i=0}^k \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}$, and therefore we have

$$\pi_i = \frac{\left(\frac{\lambda}{\mu}\right)^i / i!}{\sum_{j=0}^k \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}},$$

for $1 \leq i \leq k$. Here π_k is called the Erlang-B formula.

Now, the fraction of jobs that are lost (i.e., our motivating question 1) is the probability that an arrival finds all k servers busy. By PASTA, this is simply π_k , and people also call this the *blocking probability*, P_{block} .

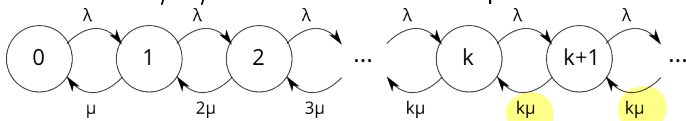
Interestingly, by multiplying both its numerator and denominator by $e^{-\lambda/\mu}$, we will have

$$P_{block} = \frac{P\{X = k\}}{P\{X \leq k\}},$$

where random variable X is poisson distributed with rate λ/μ .

M/M/k system analysis

- The CTMC for a M/M/k is also a birth-death process:



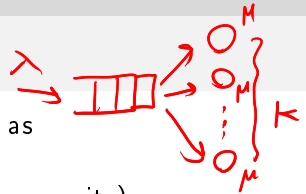
State	Time-reversibility equation	Simplified equation
0	$\pi_0 \lambda = \pi_1 \mu$	$\pi_1 = \frac{\lambda}{\mu} \pi_0$
1	$\pi_1 \lambda = \pi_2 2\mu$	$\pi_2 = \left(\frac{\lambda}{\mu}\right)^2 \frac{1}{2!} \pi_0$
2	$\pi_2 \lambda = \pi_3 3\mu$	$\pi_3 = \left(\frac{\lambda}{\mu}\right)^3 \frac{1}{3!} \pi_0$
$k-1$	$\pi_{k-1} \lambda = \pi_k k\mu$	$\pi_k = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \pi_0$
k	$\pi_k \lambda = \pi_{k+1} k\mu$	$\pi_{k+1} = \left(\frac{\lambda}{\mu}\right)^{k+1} \frac{1}{k!} \frac{1}{k} \pi_0$
$k+1$	$\pi_{k+1} \lambda = \pi_{k+2} k\mu$	$\pi_{k+2} = \left(\frac{\lambda}{\mu}\right)^{k+2} \frac{1}{k!} \frac{1}{k^2} \pi_0$

Minimum resource requirement

For an M/M/k, the system utilization ρ is defined as

$$\rho = \frac{\lambda}{k\mu} \quad (\text{viewed from total service capacity})$$

$$= \frac{\lambda/k}{\mu} \quad (\text{viewed from each single server})$$



Let R be the the expected number of busy servers (motivating question 2), which is also called the *minimum resource requirement*. To get R , we may use formula $R = \sum_{i=0}^k i \cdot P\{i \text{ jobs in service}\}$, or we may use the following:

Theorem (Linearity of Expectation (3.26))

For random variables X and Y , $E[X+Y] = E[X] + E[Y]$.

and therefore

$$R = k \cdot \frac{\lambda}{k\mu} = \frac{\lambda}{\mu}.$$

Stationary probabilities

Back to the CTMC for M/M/k, using the time-reversibility equations and that $\sum_{i=0}^k \pi_i = 1$, we will have

$$\pi_i = \begin{cases} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} \pi_0 & \text{if } \hat{i} \leq k \\ \left(\frac{\lambda}{\mu}\right)^i \frac{1}{k!} \left(\frac{1}{k}\right)^{i-k} \pi_0 & \text{if } \hat{i} > k \end{cases}$$

$$= \begin{cases} \frac{(k\rho)^i}{i!} \pi_0 & \text{if } \hat{i} \leq k \\ \frac{\rho^i}{k!} k^k \pi_0 & \text{if } \hat{i} > k \end{cases}$$

and

$$\pi_0 = \left[\sum_{i=0}^{k-1} \frac{(k\rho)^i}{i!} + \frac{(k\rho)^k}{k!(1-\rho)} \right]^{-1}.$$

Probability for an arriving job to be queued

Let P_Q be the probability that an arriving job has to queue (motivating question 2). We have

$$\begin{aligned}
 P_Q &= P\{\text{An arrival finds all servers busy}\} \\
 &= P\{\text{An arrival sees at least } k \text{ jobs in the system}\} \\
 &= \sum_{i=k}^{\infty} \pi_i \quad (\text{by PASTA}) \\
 &= \frac{(k\rho)^k \pi_0}{k!(1-\rho)}.
 \end{aligned}$$

This equation is called the Erlang-C formula.

(Further study: see textbook Section 14.3, page 261, for a relation between P_Q in M/M/k and P_{block} in M/M/k/k.)

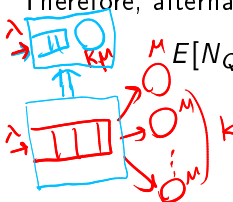
Other statistics in M/M/k

Let N_Q be the number of jobs in the queue (motivating question 2). From the formula of expectation, $E[N_Q] = \sum_{i=k}^{\infty} \pi_i(i - k) = \dots = \frac{\rho}{1-\rho} \cdot P_Q$.

Theorem (Expectation via conditioning (3.25))

For discrete random variables, $E[X] = \sum_y E[X|Y = y] \cdot P\{Y = y\}$.

Therefore, alternatively, we may compute this way:

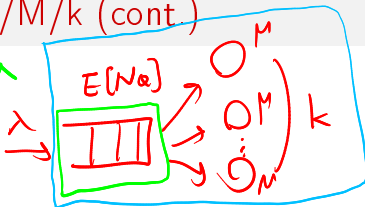


$$\begin{aligned}
 E[N_Q] &= E[N_Q \mid \text{queueing}] \cdot P\{\text{queueing}\} \\
 &\quad + E[N_Q \mid \text{no queueing}] \cdot P\{\text{no queueing}\} \\
 &= \frac{\rho}{1-\rho} \cdot P_Q. \quad (E[N_Q \mid \text{no queueing}] = 0)
 \end{aligned}$$

Note that $E[N_Q \mid \text{queueing}]$ for M/M/k is equal to the expected number of jobs in system for an M/M/1 — all servers in M/M/k are currently busy, and therefore, conceptually, the jobs in the queue are processed at the service rate of an M/M/1 server.

Other statistics in M/M/k (cont.)

$$E[N_Q] = E[T_Q] \cdot \lambda$$



$$E[N] = \lambda \cdot \underline{E[T]}$$

Let T_Q , N , T be the sojourn time in the queue, total number of jobs in the system, and total sojourn time in the system, respectively.

- $E[T_Q]$ can be obtained using Little's Law for the queue itself.
- $E[T] = E[T_Q] + 1/\mu$.
- $E[N]$ can be obtained using Little's Law for the whole system.

Comparing three server organizations

Given the same arrival rate λ and total service rate $k\mu$, which of the following three designs would give a shorter mean response time?

- M/M/1*
- ① Time-division multiplexing on one single server of service rate $k\mu$.
 - ② ✓ Statistical multiplexing on one single server of service rate $k\mu$.
 - ③ Redirecting arrivals to k servers, each having service rate μ .
- M/M/k*

We have compared the first two cases in our study of M/M/1. Now, to compare the second with the third, consider

$$\frac{E[T]^{M/M/k}}{E[T]^{M/M/1}} = \dots = P_Q^{M/M/k} + k(1 - \rho).$$

When $\rho \rightarrow 0$, $\frac{E[T]^{M/M/k}}{E[T]^{M/M/1}} \rightarrow k$;

when $\rho \rightarrow 1$, $\frac{E[T]^{M/M/k}}{E[T]^{M/M/1}} \rightarrow 1$.

another performance metric: jitter

Concluding remarks

server farms

- ✓ Think about how real-world systems can be modeled as M/M/k/k or M/M/k.
- Think about how we extended our understanding of M/M/1 to M/M/k/k and M/M/k.
- Think about how we may answer those questions posed on page 5.
- CTMC, the birth-death process, and how we may find its stationary probabilities:
 - stationary distribution = limiting distribution, extended from our study of DTMC in Week 6.
- **This concludes our study of queueing analysis in this course :)**
 - Curious minds are encouraged to study Chapters 23, 25, and 26 for M/G/1, where the service time is generalized to any distribution (!).