CSC0056 Data Communication

The Queueing Model and Operational Laws

Instructor: Chao Wang

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

Oct. 4, 2024

NATIONAL TAIWAN NORMAL UNIVERSITY

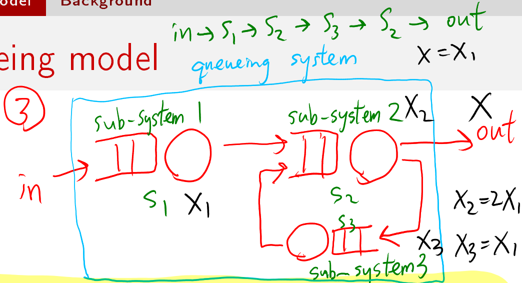
References

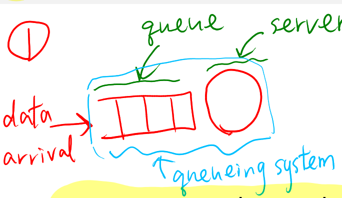
1. Harchol-Balter, Mor. Performance modeling and design of computer systems: queueing theory in action. Cambridge University Press, 2013. ISBN 9781107027503. (Chapters 1, 2, and 6)
2. Bertsekas, Dimitri and Gallager, Robert. Data networks (2nd edition). Prentice Hall, 1992. ISBN 0132009161. (Sections 3.1, 3.2.1, 3.2.3, and 3.3 up to 3.3.1)
3. Little, John DC. "A proof for the queuing formula: L= λ W." Operations research 9.3 (1961): 383-387.
4. Ho, Yao-Hua; Tai, Yun-Juo; Chen, Ling-Jyh. 2021. "COVID-19 Pandemic Analysis for a Country's Ability to Control the Outbreak Using Little's Law: Infodemiology Approach" Sustainability 13, no. 10: 5628.
5. Wang, C., Gill, C., & Lu, C. (2020, April). Adaptive Data Replication in Real-Time Reliable Edge Computing for Internet of Things. In 2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI) (pp. 128-134). IEEE.

Outline

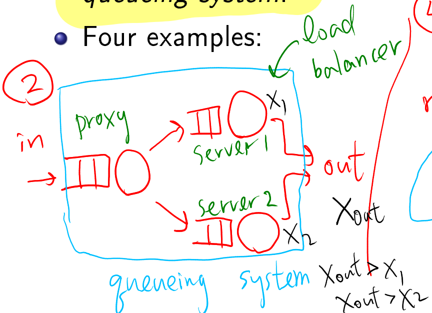
1. The Queueing Model
   - Background
   - Terminologies
   - Performance metrics
   - Open networks vs. closed networks
2. Little's Law and Utilization Law
3. Forced Flow Law
4. Moving forward and takeaways

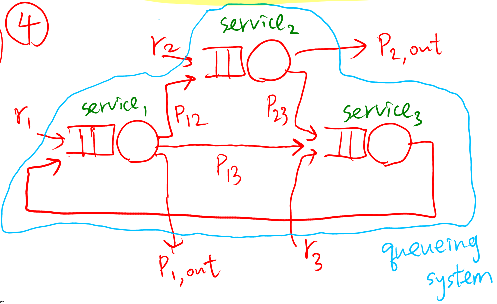
General notion of the queueing model



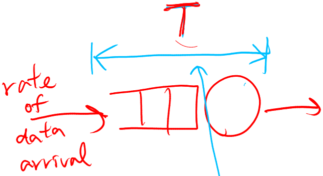


- A system can be modeled as a set of queues (buffering data) and a set of servers (processing data), and the system thus modelled is called a *queueing system*.
- Four examples:





Motivations to learn how to analyze a queueing system

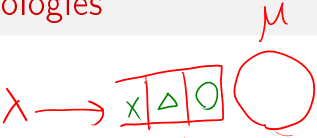


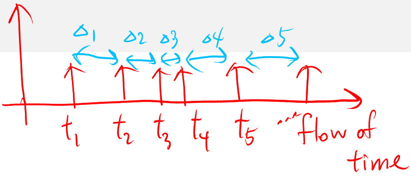
- Motivations:
  1. Predicting the system performance
  2. Driving the system design
- Example of performance prediction:
  - A data communication system can be modelled as a queueing system. For a queueing system, often we knew the rate of arrivals, and we want to predict the length of time each arriving item spent in the system. The length of time spent (T) is related to the number of items (N) currently in the system. Let p_n be the steady-state probability of n data items in the system. Then the expected value of N is

$$E[N] = \sum_{n=0}^{\infty} n \cdot p_n$$

  - Question: could we estimate the value of T as a function of $E[N]$?

Terminologies



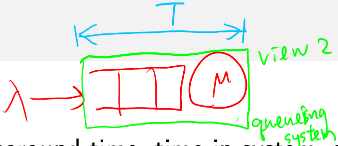


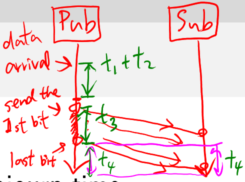
e.g., FCFS   scheduling policy

- Service order (queueing discipline ; qdisc)
- Average arrival rate (λ)
- Mean interarrival time $= \sum_{i=1}^{5} \Delta i / 5$
- Service requirement (S), i.e., Size of a job
- Mean service time (expected value of S, i.e., $E[S]$)
- Average service rate (μ) $= \frac{1}{E[S]}$

$$\mu = \frac{1}{E[S]}$$

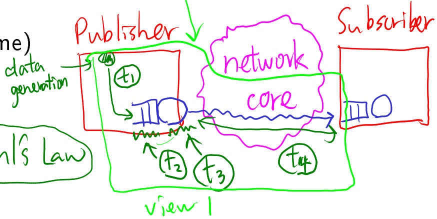
Performance metrics





- **Response time (T)**
  - also known as: turnaround time, time in system, sojourn time
  - generally called latency, and can be broken down into
    - processing time t_1
    - queueing time (aka waiting time) t_2
    - transmission time t_3
    - propagation time t_4
- Ways to *view* a system:
  - zoom-in (view 1)
  - zoom-out (view 2)

Amdahl's Law



- **Throughput (X)** (e.g., 100 Mbps)
  - throughput of the whole system
  - throughput of a device in the system

$$X = \frac{C}{t} = \frac{C}{B} \cdot \frac{B}{t} = \frac{C}{B} \cdot \rho = \frac{1}{E[S]} \cdot \rho = \mu \cdot \rho$$

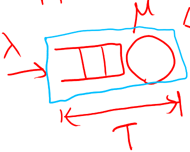
$$\rho = \frac{B}{t}$$

$$\frac{B}{C} = E[S]$$

- **Utilization (ρ)** (e.g., 90 CPU%)
- $X = \mu \cdot \rho$, or, equivalently, $\rho = X \cdot E[S]$ (the Utilization Law)

Queueing analysis examples

Suppose that we have these



if $\lambda \to 2\lambda$, then in order to keep T the same, should we double μ?

Ans: no

- Example 1: maintaining the mean service time
- Example 2: estimating the system throughput

$X = \mu \cdot \rho$

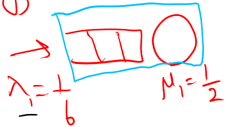
$\Rightarrow X_1 = \mu_1 \cdot \rho_1 = \lambda_1$

$X_2 = \mu_2 \cdot \rho_2 = \lambda_2$

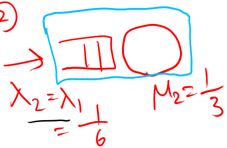
$\boxed{\rho = \frac{\lambda}{\mu}}$

Question: which of the following queueing system has a higher throughput?

①



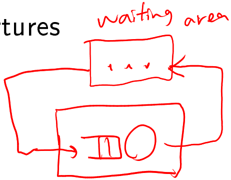
②



Ans: $X_{①} = X_{②}$

Open networks vs. closed networks

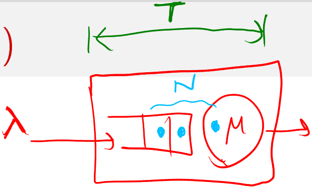
- Open networks: only external arrivals and departures
  - The two examples on the previous page
- Closed networks: no external arrivals and departures
  - Two types
    - Interactive systems (terminal-driven)
    - Batch systems
  - MPL: multiprogramming level



- Questions for you:
  - Shall we consider a web service system open or closed?
  - Shall we consider a data communication system open or closed?
  - Hybrid networks (the arrivals partially depend on the departures)?
- In this course, if not explicitly defined, we assume our subject is an open network.

Little's Law (aka Little's Theorem)

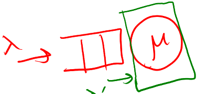
$N = \lambda T$ in average



- Little's Law[1]: $E[N] = \lambda \cdot E[T]$
  - $E[N]$: the average number of customers (data items) in a system
  - λ: the customer arrival rate
  - $E[T]$: the average delay of a customer in the system (i.e., time spent in the system) i.e., response time
- See the textbooks for its derivation.
- Versatility of Little's Law:
  - 1 distribution independent
  - 2 applicable to both the whole and part of a system

[1] Necessary assumption: the system is *ergodic*. If the system is ergodic, then the *time average* equals the *esemble average*.

Example applications of Little's Law



$N' = \lambda' \cdot T'$

where $T' = \frac{1}{\mu}$

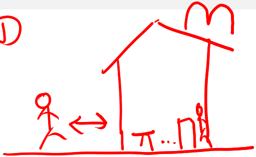
$\lambda' = \lambda$

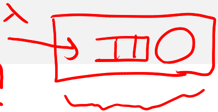
$N' = \rho$

$\Rightarrow \rho = \lambda \cdot \frac{1}{\mu}$

3

1





$N = \lambda \cdot T$

- Example 1: number of seats in a McDonald's restaurant N
- Example 2: average year-of-study for a graduate student
- Example 3: the utilization law, revisited: $\rho = X \cdot E[S] = X \cdot \frac{1}{\mu} = \frac{\lambda}{\mu}$
- Example 4: a finite buffer system (a killer restaurant)

2

$N = \lambda \cdot T \Rightarrow T = \frac{N}{\lambda} = \frac{200}{80} = 2.5$ yrs

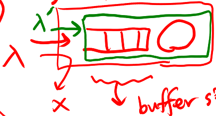
N: total # of students currently

λ: # of enrollments each year

T: average year-of-study

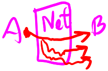
4

e.g. router



$T = \frac{N}{\lambda'}$

bufferbloat



$N = \lambda' \cdot T$

$= \lambda(1 - P_{k+1}) \cdot T$

Another application of Little's Law

Ho, Yao-Hua; Tai, Yun-Juo; Chen, Ling-Jyh. 2021. "COVID-19 Pandemic Analysis for a Country's Ability to Control the Outbreak Using Little's Law: Infodemiology Approach" Sustainability 13, no. 10: 5628.

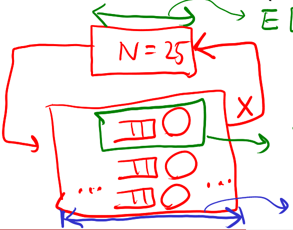
- Applying Little's Law: $E[N] = \lambda \cdot E[T]$
  - $E[N]$: the average number of confirmed COVID-19 patients
  - λ: the rate of confirmed cases
  - $E[T]$: the average recovery time of a COVID-19 patient
- See the textbooks and Wikipedia for some more examples.

Forced Flow Law and its application

$$E[R] = E[T] - E[Z] = \frac{E[N]}{X} - E[Z] = 25 \cdot \frac{1}{X} - 18$$

$$= 25 \cdot \frac{E[V_i]}{X_i} - 18 = 25 \cdot \frac{E[V_i]}{\mu_i \cdot \rho_i} - 18 = 25 \cdot \frac{E[V_i]}{\frac{1}{E[S_i]} \cdot \rho_i} - 18 = 23.67 \text{ seconds}$$

- Forced Flow Law: $X_i = E[V_i] \cdot X$
  - X: the throughput of the whole system
  - X_i: the throughput of device i in the system
  - V_i: the number of visits to device i per job (i.e., the visit ratio)
  - Note that $E[V_i]$ could be larger than 1, equal to 1, or smaller than 1.
- Example application (Section 6.9): closed network



$E[Z] = 18$ seconds

$E[V_i] = 20$

$\rho_i = 30\%$

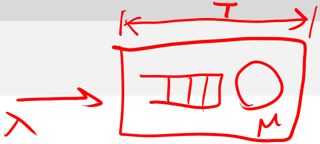
$E[S_i] = 0.025$ seconds

$E[R] = ?$

Little's Law for a closed system is

$E[N] = X \cdot E[T]$

Revisiting the latency analysis



- In a data communication system, given the arrival rate λ, we want to predict the average response time of the system.
- Let p_n be the steady-state probability of n data items in the system. Then the expected value of N is

$$E[N] = \sum_{n=0}^{\infty} n \cdot p_n$$

- From Little's Law, we have

$N = \lambda \cdot T$

$$E[T] = \frac{E[N]}{\lambda}$$

- Question: how could we obtain p_n in the first place?

Takeaways today, and some TODOs

- The queueing model
  - Terminologies
  - Performance metrics
  - open networks vs. closed networks
- Operational Laws and their applications
  - Little's Law (aka Little's Theorem)
  - Utilization Law
  - Forced Flow law
- TODOs
  - Starting next week, we will dive into queueing theory, which will require some background knowledge in basic probability. Chapters 3–4 in the textbook reviews those materials.
  - Optional reading: study Chapter 5 for the definition of ergodicity and related ideas needed for the proof of Little's Law.
  - Optional reading: study Section 6.10 for yet another operational law, the *Bottleneck Law*.