CSC0056 Data Communication

M/M/1 Applications and PASTA

Instructor: Chao Wang

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

Oct. 18, 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.
    1. Chapter 13: topic this week
    2. Chapters 11 and 12: important stuff, the required study materials following our topic last week
2. Bertsekas, Dimitri and Gallager, Robert. Data networks (2nd edition). Prentice Hall, 1992. ISBN 0132009161. (Section 3.3 and Appendix A)

Outline

1. M/M/1 Applications
2. PASTA

Using the analytical results of M/M/1

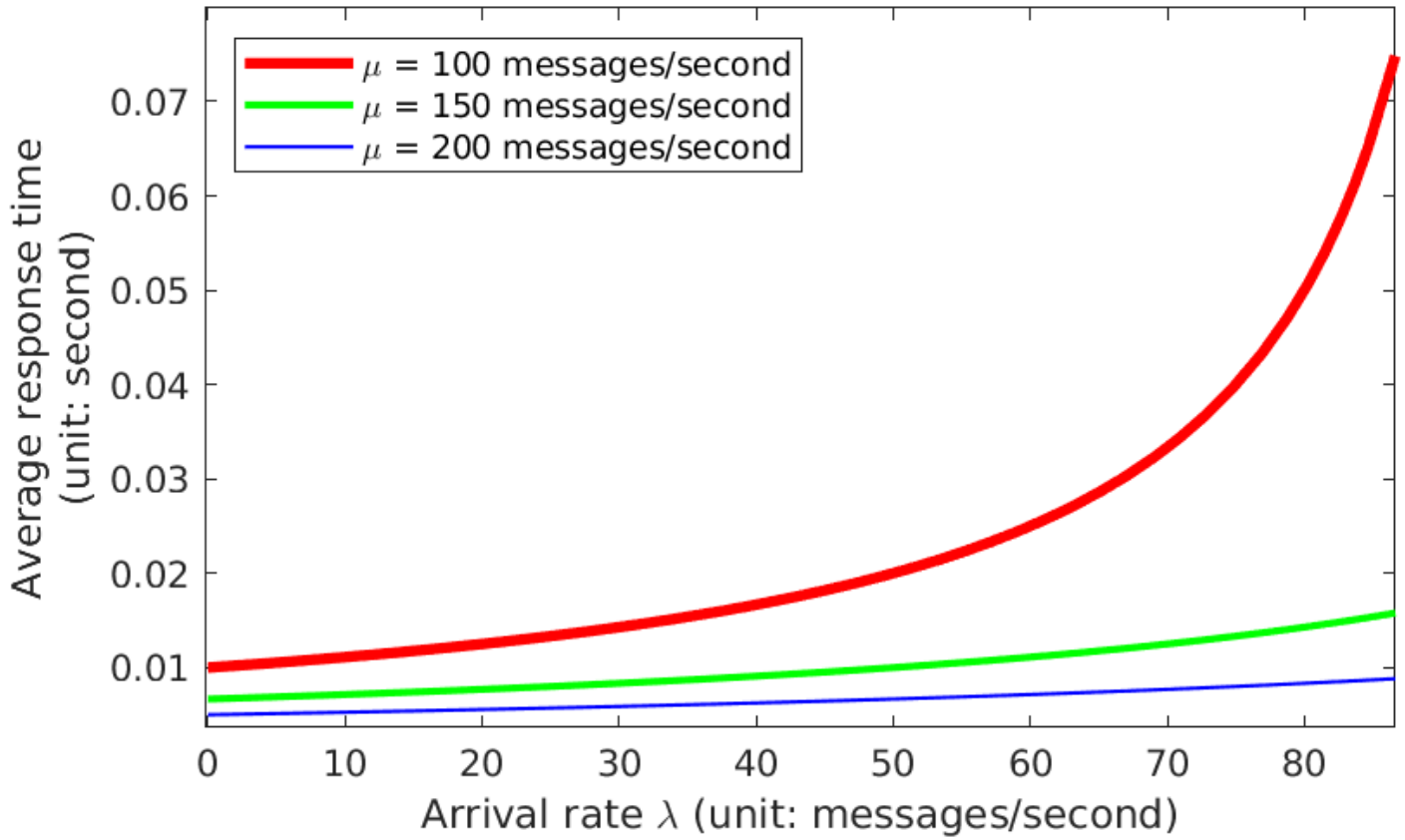
- We have derived, that for M/M/1,

$$N = \frac{\rho}{1-\rho} \quad \text{and} \quad T = \frac{N}{\lambda} = \frac{1}{\mu - \lambda}.$$

- For example, now we may be able to answer the following questions:
  1. For the same arrival rate, how would the improvement in service rate reduce the average response time?
  2. Following 1, under which condition could we significantly reduce the average response time by increasing the service rate?
  3. Suppose the arrival rate will double. To keep the same response time, should we double the service rate?
  4. For an aggregation of m statistically identicial and independent Poisson message streams arriving at a system, each with an arrival rate of λ/m. To have a shorter response time in average, should we apply *statistical multiplexing* (i.e., handle all messages in the FCFS order) or *time-division multiplexing* (i.e., divide the time into rounds of m equal-length slots, each slot$_i$ is reserved for Poisson stream$_i$ only)?

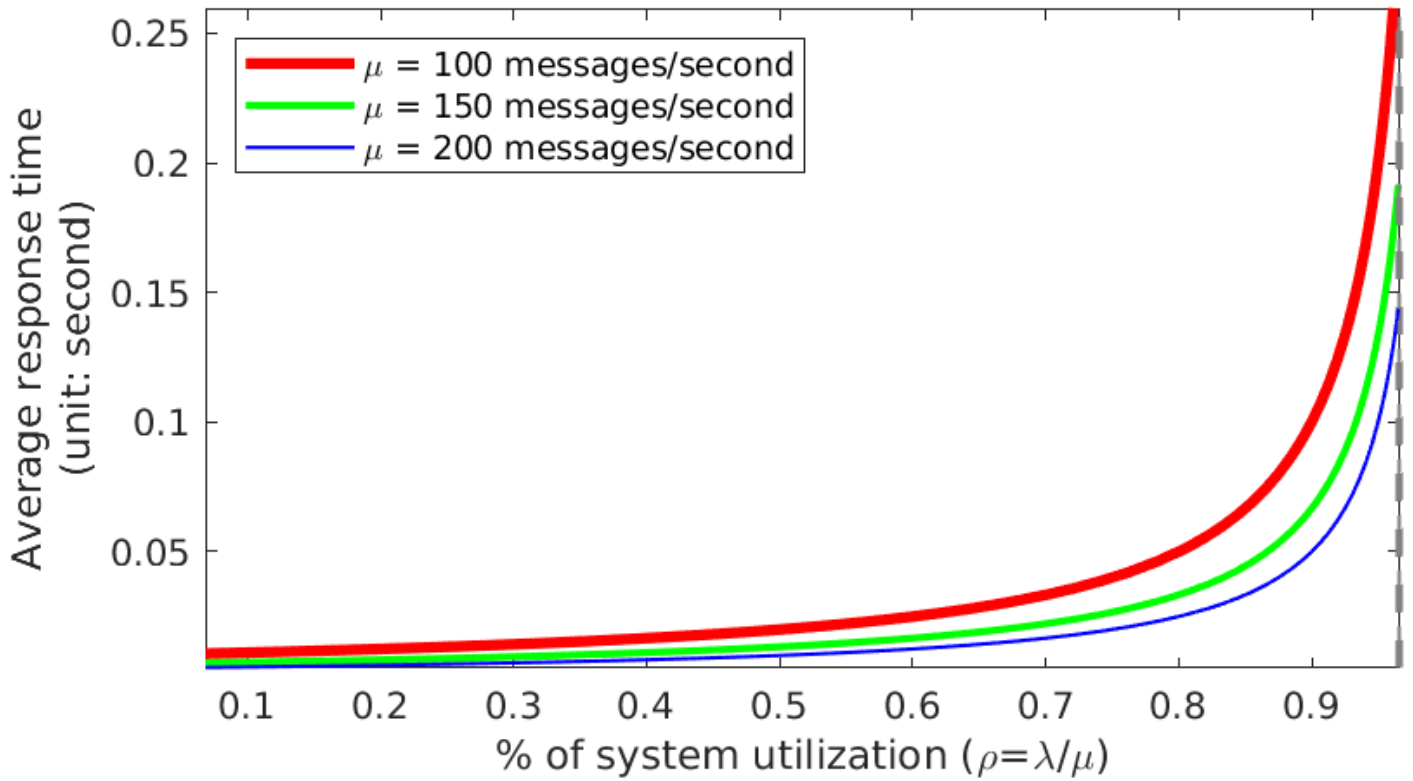
Example 1: Understanding T versus λ and μ

- The increasing of the ratio between the arrival rate and the average response time is nonlinear.
- For the same arrival rate, doubling the service rate may cut off more than 50% of the response time.



Example 2: Understanding T versus $\rho = \frac{\lambda}{\mu}$

- We may model the *system load* in terms of the percentage of time the server is busy
  - Denote the system load by $\rho = \frac{\lambda}{\mu}$.
- Improving the service rate μ may help improve much on the average response time only when the system is under a heavy load.



Example 3: M/M/1 system analysis and design (answer to Question 3 on page 4)

Example 4: comparing two multiplexing strategies (answer to Question 4 on page 4)

Motivation: reasonable measurement of system performance

- To determine the fraction of time the system has n items, can we estimate this percentage by measuring the number of items in the system upon each arrival? If it is generally not true, then under what condition could it be true?
- Formally, let a_n be the limiting probability that an arrival sees n items in the system. And let $p_n = \pi_n$ be the unconditional stationary probability that the system has n items. We want to know under what condition we may have $a_n = p_n$.
- PASTA: Poisson Arrivals See Time Averages
  - $a_n = p_n$ if both (1) the arrival process is Poisson and (2) the interarrival times and the service times are independent.

Deriving PASTA

- Notations:
  - $N(t)$: the number of items in the system at time t.
  - $A(t, t+\delta)$: an event that there is an arrival in the interval $(t, t+\delta)$.
- Derivation: