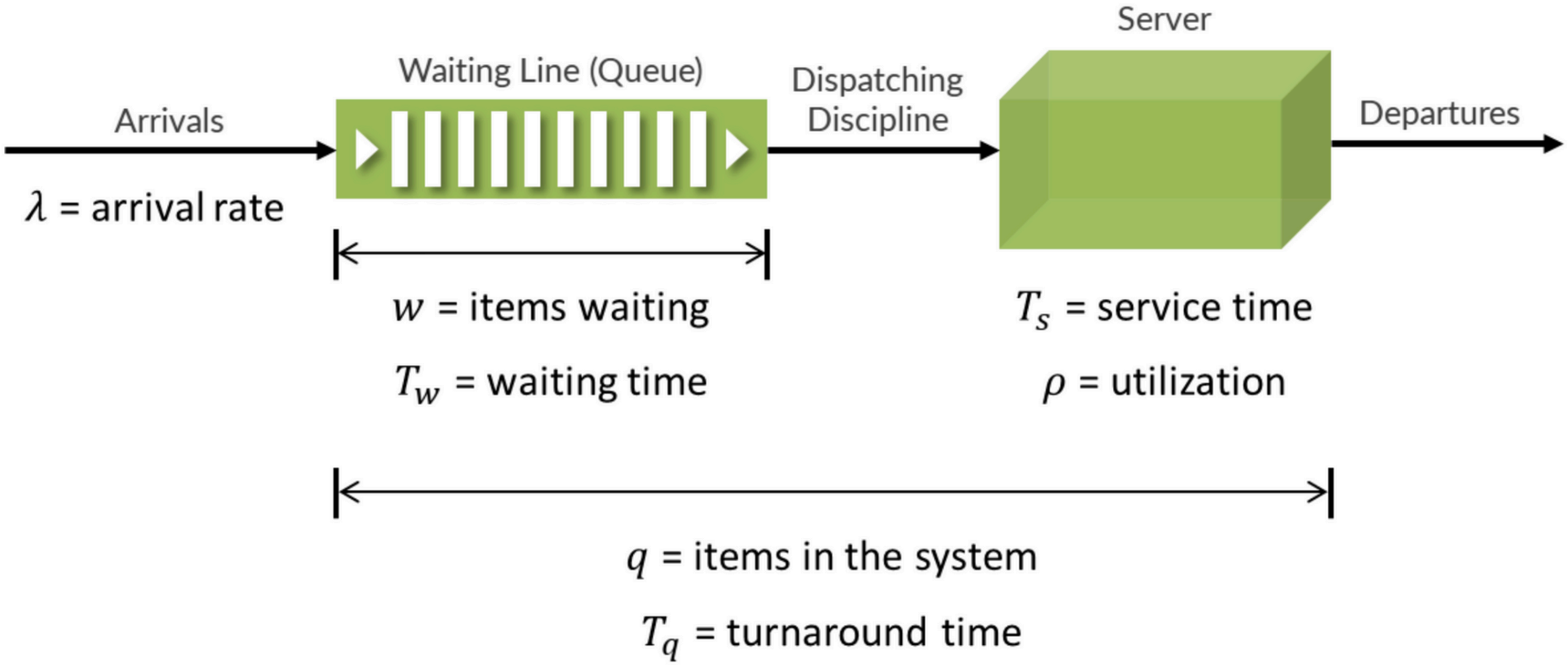
Oct 2, 2024

CS 350 DISCUSSION 4

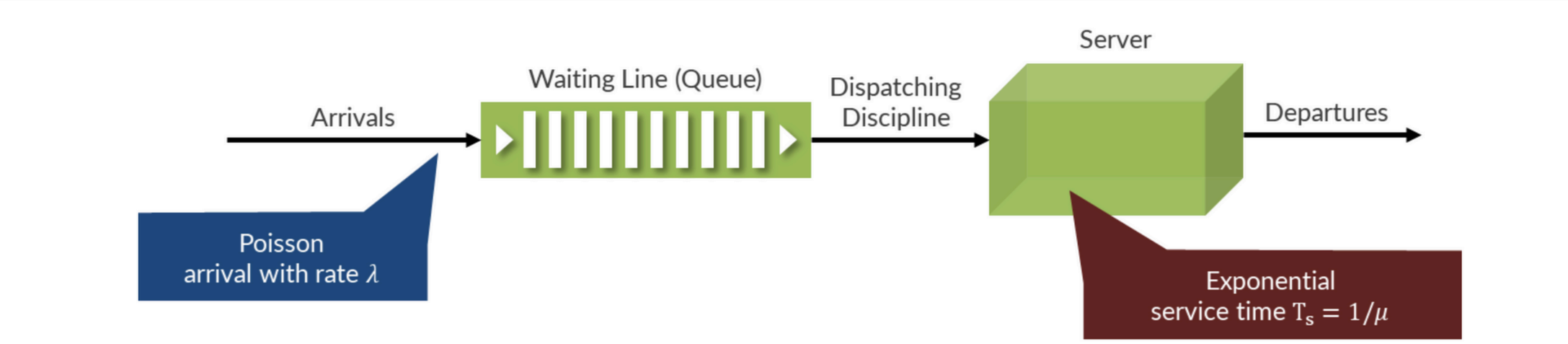
M/M/1

System Architecture Review



Review on M/M/1

- What is a M/M/1 system?
  - single-server queuing system.
  - Infinite single-queue.
  - Arrivals are poisson and service is exponential.



M/M/1 Notation + Formulas

Avg. number of customers waiting for service:

$$w = q - \rho = \frac{\rho^2}{1-\rho}$$

Avg. time in system:

$$T_q = \frac{q}{\lambda} = \frac{1}{\mu \cdot (1-\rho)}$$

Avg. time *waiting* in system:

$$T_w = \frac{w}{\lambda} = \frac{\rho}{\mu \cdot (1-\rho)}$$

Slowdown:

$$Formula_A = \frac{1}{1-\rho} \qquad Formula_B = \frac{T_q}{T_s}$$

Little's laws formulas:

$$W = \lambda T_w$$

$$Q = \lambda T_q$$

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

Avg. number of requests in the entire system For M/M/1 :

$$q = \frac{\rho}{1-\rho}$$

Let’s Practice

Congratulations! You have been seed-funded for your Picvatar startup. Your system works like this. Users submit their pictures and the system generates a cartoon-like avatar for them. Now it’s the time to tune the system. After observing the system for a while you have noticed that users submit images with an average size of 2 MB (1 MB = 1024 KB; and 1 KB = 1024 bytes). Every day, you are already receiving about 30,000 requests. Your investors want to know the following:

a) What is the amount of time it should take to process each individual request so that on average there are never more than 20 requests in the system — i.e. either being currently processed or queued?

Problem 9.2 from the book

Let’s Practice

a) What is the amount of time it should take to process each individual request so that on average there are never more than 20 requests in the system — i.e. either being currently processed or queued?

Problem 9.2 from the book

Let's Practice

a) What is the amount of time it should take to process each individual request so that on average there are never more than 20 requests in the system — i.e. either being currently processed or queued?

$$\lambda = 30,000 \text{ req/day} = 0.35 \text{ req/s and } q = 20 \text{ requests.}$$

$$q = \frac{\rho}{1-\rho} = \frac{\lambda T_s}{1-\lambda T_s} \quad \Rightarrow \quad T_s = 2.7 \text{ seconds}$$

Problem 9.2 from the book

Let’s Practice

b) Assuming you can tune the system to perfectly meet the requirement above, how much would each user end up waiting on average before receiving their avatar?

Problem 9.2 from the book

Let's Practice

b) Assuming you can tune the system to perfectly meet the requirement above, how much would each user end up waiting on average before receiving their avatar?

Each user receives their avatar after T_q seconds on average once a request is submitted. Thus

$$T_q = \frac{q}{\lambda} = \frac{20}{0.35} = 57.1 \text{ seconds}$$

Problem 9.2 from the book

Let’s Practice

c) The manufacturer of the server you are using stated that you should enable the more expensive UV-reactive water-cooling system if your system is going to be busy for more than 51 seconds per minute. Should the investors budget for water cooling?

Problem 9.2 from the book

Let's Practice

c) The manufacturer of the server you are using stated that you should enable the more expensive UV-reactive water-cooling system if your system is going to be busy for more than 51 seconds per minute. Should the investors budget for water cooling?

$$\text{Required utilization of system} = \frac{51}{60} = 0.85, \text{ Current utilization:}$$

$$\rho = \lambda T_s = 0.35 \times 2.7 = 0.95 > 0.85$$

$$\text{Thus, investors should budget for the new water cooling system.}$$

Problem 9.2 from the book

Let’s Practice

d) What is the slowdown of the system?

Problem 9.2 from the book

Let's Practice

d) What is the slowdown of the system?

Since we are assuming this system to be M/M/1, thus the slowdown is:

$$\frac{T_q}{T_s} \approx 21$$

Problem 9.2 from the book

Let’s Practice

e) What is the maximum number of requests per day that the system can sustain with the current tuning?

Problem 9.2 from the book

Let's Practice

e) What is the maximum number of requests per day that the system can sustain with the current tuning?

To sustain the maximum number of requests, utilization would approach infinity. Thus:

$$\lambda T_s = 1 \quad \Rightarrow \quad \lambda = \frac{1}{T_s} = 0.37 \text{ req/s}$$

Problem 9.2 from the book

Let’s Practice

f) How much memory does the system require on average, considering that a request currently being processed still occupies memory.

Problem 9.2 from the book

Let’s Practice

f) How much memory does the system require on average, considering that a request currently being processed still occupies memory.

$$20\ \text{req} \times 2\ \text{MB} = 40\ \text{MB}$$

Problem 9.2 from the book

Let’s Practice

The investors got back to you saying that you either support way more users per day or the deal is off. Overnight, you devise the following optimization. You organize your system in two stages. At the first stage, pictures are compressed down to 250 KB. After compression, they are sent to the main processing server for avatar generation. This way, avatar generation can take as little as 1.2 seconds. Now you worry about the following:

g) How long should compression take so that the system can sustain 60,000 requests per day?

Problem 9.2 from the book

Let's Practice

g) How long should compression take so that the system can sustain 60,000 requests per day?

$$\text{New } \lambda = 60,000 \text{ req/day} = 0.69 \text{ req/s}, T_s^{\text{generate}} = 1.2 \text{ seconds}.$$

Since the system is composed of two stages, we analyze both stages as M/M/1 queuing systems independently. Utilization of each stage must therefore be less than 1 for the entire system to sustain the new λ.

Avatar generation:

$$\rho = \lambda T_s = 0.69 \times 1.2 = 0.83 < 1$$

Compression:

$$\lambda T_s \leq 1 \quad \Rightarrow \quad T_s^{\max} = \frac{1}{\lambda} = 1.45 \text{ seconds}$$

Problem 9.2 from the book

Let’s Practice

h) Assume that compression takes 1 second, what is going to be the average time it takes your whole system (compression+processing) to generate an avatar when your system reaches 60,000 requests per day?

Problem 9.2 from the book

Let's Practice

h) Assume that compression takes 1 second, what is going to be the average time it takes your whole system (compression+processing) to generate an avatar when your system reaches 60,000 requests per day?

$$T_s^{\text{compression}} = 1s,\ \lambda = 60000\,\text{req/day} = 0.69\,\text{req/s}$$

$$\text{Stage 1 Compression:}$$

$$\lambda T_s^{\text{compression}} = 0.69 \times 1 = 0.69, \quad q_1 = \frac{0.69}{1 - 0.69} = 2.23\,\text{req}$$

$$\text{Stage 2 Generation:}$$

$$\lambda T_s^{\text{generate}} = 0.69 \times 1.2 = 0.83, \quad q_2 = \frac{0.83}{1 - 0.83} = 4.88\,\text{req}$$

$$q_{\text{total}} = q_1 + q_2 = 7.11$$

$$\Rightarrow \quad \text{Average time to generate avatar for the whole system} = T_q^{\text{total}} = \frac{q_{\text{total}}}{\lambda} = \frac{7.11}{0.69} = 10.3\,\text{seconds}$$

Problem 9.2 from the book

Let’s Practice

i) With the same assumption on compression time, do you still need cooling at the main processing server?

Problem 9.2 from the book

Let's Practice

i) With the same assumption on compression time, do you still need cooling at the main processing server?

$$\rho = \lambda T_s^{\text{generate}} = 0.69 \times 1.2 = 0.83 < 0.85$$

$$\Rightarrow \quad \text{no cooling needed at the main processing server.}$$

Problem 9.2 from the book