

Queuing Systems

QUEUING PROCESS is a stochastic process $X(t)$, in which

states = # of jobs in the system
(waiting + being served).

| | |
|-----------------------|-----------------------------------|
| <i>Arrival</i> | $\Rightarrow X(t)$ increases by 1 |
| <i>End of service</i> | $\Rightarrow X(t)$ decreases by 1 |

Characteristics:

- Distribution of interarrival times
- Distribution of service times
- Number of servers
- Limited or unlimited capacity

SINGLE-SERVER BERNOULLI QUEUING PROCESS

It is a queuing process with:

- one server;
- unlimited capacity;
- arrivals according to a Bernoulli counting process;

$$P_A = P \{ \text{arrival during any frame} \}$$

- service completions according to a Bernoulli counting process;

$$P_S = P \left\{ \begin{array}{l} \text{completed service} \\ \text{during any frame} \end{array} \middle| \text{server is busy} \right\}$$

- arrivals independent of service times.

Remarks:

- Fastest service = 1 frame.
- P_A, P_S are constant \Rightarrow it is a homogeneous process.

It is a Markov chain.

TRANSITION PROBABILITIES

$$P(0 \rightarrow 0) = 1 - P_A$$

$$P(0 \rightarrow 1) = P_A$$

For all $i \geq 1$,

$$P(i \rightarrow i - 1) = P_S(1 - P_A)$$

$$P(i \rightarrow i + 1) = P_A(1 - P_S)$$

$$P(i \rightarrow i) = P_A P_S + (1 - P_A)(1 - P_S)$$

$$\begin{aligned} \text{ARRIVAL RATE} &= \lambda_A = P_A/\Delta \\ \text{SERVICE RATE} &= \lambda_S = P_S/\Delta \end{aligned}$$

Then

$$\begin{aligned} \mu_A &= E(\text{interarrival time}) = 1/\lambda_A \\ \mu_S &= E(\text{service time}) = 1/\lambda_S. \end{aligned}$$

Modification: LIMITED CAPACITY

That is, $X(t) \leq C$ with probability 1.

Then

$$\left\{ \begin{array}{l} P(i \rightarrow j) \text{ are unchanged for } i \leq C; \\ P(C \rightarrow C - 1) = (1 - P_A)P_S; \\ P(C \rightarrow C) = 1 - (1 - P_A)P_S. \end{array} \right.$$

Example: telephone with 2 lines.

M/M/1 Queuing Process

Let Δ be *small*, then $P_A = \lambda_A \Delta$ and $P_S = \lambda_S \Delta$ are small. Hence,

$$P(0 \rightarrow 0) = 1 - \lambda_A \Delta$$

$$P(0 \rightarrow 1) = \lambda_A \Delta$$

$$P(i \rightarrow i - 1) \approx P_S = \lambda_S \Delta$$

$$P(i \rightarrow i + 1) \approx P_A = \lambda_A \Delta$$

$$P(i \rightarrow i) \approx 1 - \lambda_A \Delta - \lambda_S \Delta$$

M/M/1 PROCESS is a queuing process which is

- continuous time
- Markov
- transition probabilities as above
- with independent increments

Distributions

In an M/M/1 process,

Interarrival times are *Exponential*(λ_A),
mean = $\frac{1}{\lambda_A} = \mu_A$

Service times are *Exponential*(λ_S),
mean = $\frac{1}{\lambda_S} = \mu_S$

Steady-state distribution of $X(t)$

Let $\pi_i = \lim P\{X(t) = i\}$ (i.e., steady-state probabilities), where Δ is small (ignore Δ^2).

Solve

$$\begin{cases} \pi P = \pi \\ \sum \pi_i = 1 \end{cases}$$

For $i = 0$,

$$\pi_0 = \pi_0(1 - \lambda_A \Delta) + \pi_1 \lambda_S \Delta$$

↓

$$\boxed{\lambda_A \pi_0 = \lambda_S \pi_1}$$

M/M/1, steady-state distribution

For $i > 0$,

$$\pi_i = \pi_{i-1} \lambda_A \Delta + \pi_i (1 - \lambda_A \Delta - \lambda_S \Delta) + \pi_{i+1} \lambda_S \Delta$$

⇓

$$(\lambda_A + \lambda_S) \pi_1 = \lambda_A \pi_0 + \lambda_S \pi_2 = \lambda_S \pi_1 + \lambda_S \pi_2$$

⇓

$$\boxed{\lambda_A \pi_1 = \lambda_S \pi_2}$$

etc...

Get "**Balance equations**": $\boxed{\lambda_A \pi_i = \lambda_S \pi_{i+1}}$.

This distribution of $X(t)$ is **shifted Geometric**,
i.e.

$$\frac{\pi_i + 1}{\pi_i} = \frac{\lambda_A}{\lambda_S} = r, \quad \pi_i = (1 - r)r^i, \quad r \geq 0$$

Here

$$r = \frac{\lambda_A}{\lambda_S}$$

is the arrival-to-service ratio, or utilization.

$X(t) + 1$ is Geometric($1 - r$).

Consequences.

For $r < 1$,

- $P(X > x) = r^{x+1}$
- $P(X = x) = (1 - r)r^x$
- $E(X) = \frac{r}{1-r}$
- $STD(X) = \frac{\sqrt{r}}{1-r}$

If $r \geq 1$, the system gets overloaded.

Response time

This is the time a job spends in the system

$$T = \sum_{n=0}^{X+1} (n^{\text{th}} \text{ service time}),$$

so that

$$E(T) = E(X + 1)\mu_S = \frac{1}{(1 - r)\lambda_S}.$$