Fun queues for 6.041

The importance of queues

- When do queues appear?
  - Systems in which some serving entities provide some service in a shared fashion to some other entities requiring service
- Examples
  - customers at an ATM, a fast food restaurant
  - Routers: packets are held in buffers for routing
  - Requests for service from a server or several servers
  - Call requests in a circuit-oriented system such as traditional telephony, mobile networks or high-speed optical connections

What types of questions may we be interested in posing?

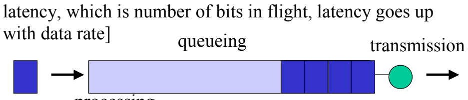
- What is the average number of users in the system? What is the average delay?
- What is the probability a request will find a busy server?
- What is the delay for serving my request? Should I upgrade to a more powerful server or buy more servers?
- What is the probability that a packet is dropped because of buffer overflow? How big do I need to make my buffer to maintain the probability of dropping a packet below some threshold? What is the probability that I cannot accommodate a call request (blocking probability)?
- For networked servers, how does the number of requests queued at each server behave?
- We shall keep these types of questions in mind as we go forward

Analysis versus simulation

- Why can't I just simulate it?
- Analysis and simulation are complementary, not opposed
- It is generally impossible to simulate a whole system- we need to be able to determine the main components of the system and understand the basis for their interaction
- What are the important parameters? What is their effect?
- In many systems simulation is required to qualify the results from analysis, to obtain results that are too complex computationally

Delay components

- Processing delay: for instance time from packet reception to assignment to a queue (generally constant)
- Queueing delay: time in queue up to time of transmission
- Transmission delay: actual transmission time (for instance proportional to packet length)
- Propagation delay: time required for the last bit to go from transmitter to receiver (generally proportional to the physical link distance, large for satellite link) [Not to confuse with latency, which is number of bits in flight, latency goes up with data rate]



Little's theorem

- Rather than refer to packets, calls, requests, etc… we refer to customers
- Relates delay, average number of customers in queue and arrival rate (λ)
- Little's Theorem: average number of customers = λ x average delay
- Holds under very general assumptions

Main parameters of a queueing system

- N(t): number of customers in the system at time t
- P(N(t) = n) = probability there are n customers in the system at time t
- Steady state probability:
$$P_n = \lim_{t\to\infty} P(N(t)=n)$$
- Mean number in system at time t:
$$\overline{N(t)} = \sum_{n=0}^{\infty} n\,P(N(t)=n)$$
- Time average number in the system:
$$N_t = \frac{1}{t}\int_{t=0}^{\infty} N(t)$$
- We assume the system is **ERGODIC**:
$$\lim_{t\to\infty} N_t = \lim_{t\to\infty} \overline{N(t)} = N$$

MIT

Main parameters

- We looked at the system from the point of view if the customers in it, let us now consider the delay of those customers
- T(k): delay of customer k
- α(t): number of customer arrivals up to time t
- β(t): number of customer arrivals up to time t
- Our ergodicity assumption implies that the long-term arrival rate is the long-term departure rate:
$$\lambda = \lim_{t\to\infty}\frac{\alpha(t)}{t} = \lim_{t\to\infty}\frac{\beta(t)}{t}$$
- Our ergodicity assumption implies that there exists a limit:
$$T = \lim_{t\to\infty}\frac{\sum_{k=1}^{\alpha(t)} T(k)}{\alpha(t)}$$

MIT

Little's theorem

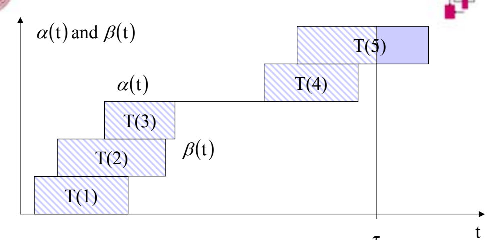
- We have:

$$\lambda T = N$$

- Little's theorem applies to any arrival-departure system with appropriate interpretation of average number of customers in the system, average arrival rate and average customer time in system
- Answers to some extent our first question

MIT

Justification of Little's theorem



$$\sum_{k=1}^{\beta(t)} T(k) \le \text{shaded area} = \int_0^{\tau} N(t)\,dt \le \sum_{k=1}^{\alpha(t)} T(k)$$

Note: a similar picture holds even if we do not have FIFO

MIT

Justification of Little's theorem

- Taking the average over time:

$$\frac{\beta(t)}{\tau}\,\frac{\sum_{k=1}^{\beta(t)} T(k)}{\beta(t)} \le \frac{1}{\tau}\int_0^{\tau} N(t)\,dt \le \frac{\alpha(t)}{\tau}\,\frac{\sum_{k=1}^{\alpha(t)} T(k)}{\alpha(t)}$$

Goes to T in the limit as t → ∞

Goes to λ in the limit as t → ∞

MIT

M/M/1 system

Memoryless arrival

Single server

Memoryless service time

- Poisson process A(t) with rate λ is a probabilistic arrival process such that:
  - number of arrivals in disjoint intervals are independent
  - number of arrivals in any interval of length τ has Poisson distribution with parameter λτ:
$$P(A(t+\tau)-A(t)=n) = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}$$

MIT

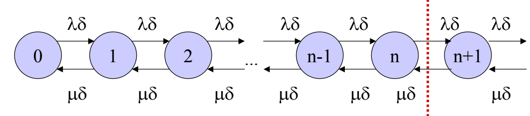
M/M/1

- Single server
- Poisson arrival process with rate λ
- Independent identically distributed (IID) service times $X(n)$ for the service time of user n
- Service times X are exponentially distributed with parameter μ, so
  $P(X(n) \le s) = 1 - e^{-\mu s}$ and $E[X] = 1/\mu$
- Interarrival times and service times are independent
- We define $\rho = \lambda / \mu$, we shall see later how that relates to the ρ we considered when discussing Little's theorem
- Can we make use of the very special properties of Poisson processes to describe probabilistically the behavior of the system?

MIT

Markov chain for M/M/1

- In steady state, across some cut between two states, the proportion number of transitions from left to right must be the same as the proportion of transitions from right to left



- Local balance equations

$$P(N = n)\lambda\delta + o(\delta) = P(N = n+1)\mu\delta + o(\delta)$$

dividing by δ and taking the limit as $\delta \to 0$

$$P(N = n+1) = \rho\, P(N = n)$$

MIT

Balance equations

- We know that $\sum_{n=0}^{\infty} P(N = n) = 1$
- Let us use this fact to determine all the other probabilities

$$P(N = n+1) = P(N = 0)\rho^{n+1}$$

- We have $1 = \sum_{n=0}^{\infty} P(N = 0)\rho^{n+1}$

so $P(N = 0) = 1 - \rho$

- Let us answer the second question:
  - we use the fact that Poisson arrivals see time average (PASTA)
  - the probability of having a random customer wait is ρ

MIT

Mean values

- We can now make use of Little's theorem to answer our first set of questions:

$$\overline{N} = \sum_{n=0}^{\infty} n\, P(N = n) = \sum_{n=0}^{\infty} n\,(1-\rho)\rho^{n+1} = \frac{\rho}{1-\rho}$$

$$\text{so } T = \frac{\overline{N}}{\rho} = \frac{\frac{\lambda}{\mu}}{\lambda\left(1 - \frac{\lambda}{\mu}\right)}$$

- What is the wait in queue, W? Use independence of service times to get $W = T - 1/\mu$

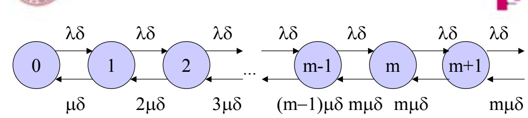
MIT

More queue Scenarios

- A similar type of analysis holds for other queue scenarios:
  - set up a Markov chain
  - determine balance equations
  - use the fact that all probabilities sum to 1
  - derive everything else from there
- M/M/m queue: Poisson arrivals, exponential distribution of service time, m servers
- Similar analysis to before, except now the probability of a departure is proportional to the number of servers in use, because a departure occurs if AT LEAST one of the servers has a departure
- Now $\rho = m\mu$

MIT

Markov chain for M/M/m



$$P(N = n-1) = n\mu\, P(N = n) \text{ for } n \le m$$

$$P(N = n-1) = m\mu\, P(N = n) \text{ for } n > m$$

$$\text{so } P(N = n) = P(N = 0)\frac{m^m \rho^n}{m!}$$

$$\text{where } P(N = 0) = \left[\frac{m^m \rho^m}{(1-\rho)m!} + \sum_{n=0}^{m-1} \frac{m^n \rho^n}{n!}\right]^{-1}$$

MIT

Let us answer our first two questions

- Second question, what is the probability that a customer must wait in queue:Erlang C formula

$$P_Q = \sum_{n=m}^{\infty} P(N=n) = \frac{P(N=0)\, m^m \rho^m}{(1-\rho)m!}$$

- Applying Little's theorem:

$$W = \frac{\rho P_Q}{(1-\rho)\lambda}$$

$$T = \frac{1}{\mu} + W$$

$$\overline{N} = \lambda T$$

One server or many?

- We now have the tools to answer our third question: would I rather have a single more powerful server or many weaker servers?
- Would we rather have a single server with service rate mμ or m servers with service rate μ?

M/M/∞

- Infinite number of servers
- Taking m to go to ∞ in the M/M/m system, we have that the occupancy distribution is Poisson with parameter λ/μ

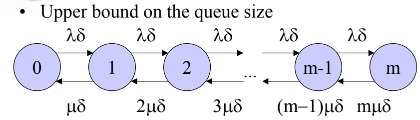
$$P(N=n) = \frac{\left(\frac{\lambda}{\mu}\right)^n e^{-\frac{\lambda}{\mu}}}{n!}$$

$$\overline{N} = \frac{\lambda}{\mu}$$

- T = 1/ μ

M/M/m/m

- Upper bound on the queue size



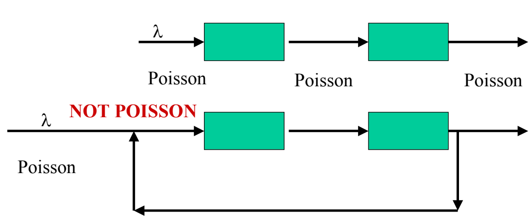
$P(N = n-1) = n\mu\, P(N=n)$ for $n \le m$

$$\text{so } P(N=n) = P(N=0)\frac{\left(\frac{\lambda}{\mu}\right)^n}{m!} \text{ for } n \le m \text{ where } P(N=0) = \left[\sum_{n=0}^{m-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}\right]^{-1}$$

- The answer to our third question is, using PASTA, the probability P(N=m)

Networks of queues

- Closed form solutions are difficult to obtain
- Poisson with feedback does not remain Poisson



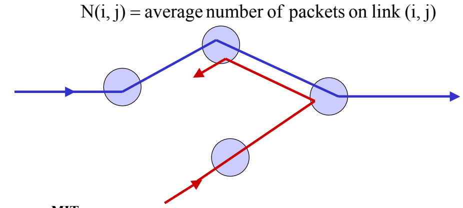
Network of queues

- Several streams, each on a path p, each with rate λ(p)
- Let us look at directed link (i,j):

$$\lambda(i,j) = \sum_{\text{all paths p traversing link (i, j)}} \lambda(p)$$

$$\mu(i,j) = \text{service rate on link (i, j)}$$

$$N(i,j) = \text{average number of packets on link (i, j)}$$



Kleinrock independence assumption

- Assume all queues behave like M/M/1 with arrival rate $\lambda(i,j)$, service rate $\mu(i,j)$, and service/propagation delay $d(i,j)$
- Then

$$N_{i,j} = \frac{\lambda_{i,j}}{\mu_{i,j} - \lambda_{i,j}} + \lambda_{i,j} d_{i,j}$$

average number of packets in the whole network

$$N = \sum_{i,j} N_{i,j}$$

average time in the system (using Little's theorem)

$$T = \frac{N}{\sum_p \lambda_p}$$

How good is it?

- Good for densely connected networks and moderate to heavy loads
- Good to guide topology design before involving simulation, other applications where a rough estimate is needed
- Are there any networks of queues where we can establish analytical results?
- Assuming that:
  - arrival processes from outside the network are Poisson
  - at each queue, streams have the same exponential service time distribution and a single server
  - interarrival times and service times are independent
- Then:
  - the steady state occupancy probabilities in each queue are the same as if the queue were M/M/1 in isolation