Statistical Multiplexing

- Circuit switching
  - Dedicated end-to-end connection
  - Resources are reserved along path
  - Guaranteed constant data rate
  - Achieved through multiplexing (TDM, FDM)

- Packet switching
  - Packets are unit of transmission
  - "Best effort" and no guarantees
  - Switches perform "store-and-forward"
  - Statistical multiplexing incurs queuing delays

- Can we quantify queuing delay?
Simple Queuing Example

- Queuing systems are everywhere
  - Line in bookstore (or Blue Wall)
  - Traffic light
  - Your homework assignments

- Key features
  - “Server” has finite capacity (needs time to process)
    » In network terminology, the server is the link
  - Demand for service (“job” arrival) is unpredictable
    » The jobs are the packets

- Questions
  - How long does a job need to wait before being serviced?
  - How many jobs are in the queue?
  - How high is the utilization of the server?

Notation in Queuing Systems

- Notation introduced by Kleinrock
  - \( C_n \) is \( n \)th customer entering system
    » \( \tau_n \) is arrival time for \( C_n \)
    » \( t_n \) is interarrival time \( (t_n=\tau_n-\tau_{n-1}) \)
    » \( x_n \) is service time for \( C_n \)
    » \( w_n \) is waiting time for \( C_n \)
    » \( s_n \) is system time (waiting plus queuing) for \( C_n \) \( (s_n=w_n+x_n) \)
  - \( N(t) \) is number of customers in system at time \( t \)
  - \( U(t) \) is amount of unfinished work in system at time \( t \)
  - \( \lambda \) is average arrival rate
    » \( E[t_n]=1/\lambda \)
  - \( \mu \) is average service rate
    » \( E[x_n]=1/\mu \)
Basic Queuing Behavior

- $\alpha(t)$ is number of arrivals in $(0,t)$
- $\delta(t)$ is number of departures in $(0,t)$
- Number of customers in system is
  - $N(t) = \alpha(t) - \delta(t)$
- Average system time is
  - Area between $\alpha(t)$ and $\delta(t)$, denoted by $\gamma(t)$
  - $T_t = \gamma(t)/\alpha(t)$

Little’s Law

- Average arrival rate
  - $\lambda_t = \alpha(t)/t$
- Average system time
  - $T_t = \gamma(t)/\alpha(t)$
- Average number of customers
  - $N_t = \gamma(t)/t$
- Substitute $\gamma(t)$ and $\alpha(t)$
  - $N_t = \lambda_t T_t$
- For $t \to \infty$:
  - $N = \lambda T$ (Little’s law)
- Average number of customers in queuing system is average arrival rate times average system time.
Related Results

- Average number of customers in queue
  - \( \bar{N}_q = \lambda W \)

- Relation between waiting and service time
  - \( T = \bar{x} + W \)

- Utilization \( \rho \)
  - \( \rho = \frac{\lambda}{\mu} = \frac{\lambda \bar{x}}{\mu} \)
  - System only stable if \( \rho < 1 \) (why not \( \rho = 1 \)?)
  - Let \( p_0 \) be probability that server idle: \( \rho = 1 - p_0 \)

- So far:
  - Not specific to particular type of queue
  - No quantitative results

Modeling of Queuing Systems

- Any queuing system can be modeled as a “stochastic process”
  - Family of random variables \( X \)
    - \( X(t) \) is indexed by time parameter \( t \in T \)
    - \( X(t) \in S \), where \( S \) is “state space”
  - If \( S \) is discrete, then stochastic is a “chain”

- Each state reflects state of queuing system
  - Probabilities indicate what states are more likely

- Markov chains
  - Probability for any state **only** depends on previous state
  - History of Markov chain is summarized in current state
Discrete Time Markov Chains

- DTMC is defined by
  - \( X_n \) is random variable indicating state in step \( n \)
  - \( p_{ij} \) are transition probabilities between states
    - Probability depends on current state only

- Example:
  - State space \( S=\{0,1\} \)
  - Transition probabilities \( P \)
    - \( S \times S \) matrix
    - \( p_{00}=0.75, \ p_{01}=0.25 \)
    - \( p_{10}=0.5, \ p_{11}=0.5 \)
  - Probability to be in state 0 at step \( n \)
    - \( P[X_n=0] = 0.75 \cdot P[X_{n-1}=0] + 0.5 \cdot P[X_{n-1}=1] \)

Stationary Probability Vector

- What is the probability of being in a particular state?
  - If Markov chain “runs long enough”, initial state irrelevant

- Define \( \pi_i \) as stationary probability of being in state \( i \)
  - \( \pi_i \) is independent of time
    - In matrix form: \( \pi = \pi P \)

- Stationary probability can be solved as set of linear equations:
  - \( \pi_0 = 0.75 \cdot \pi_0 + 0.5 \cdot \pi_1 \)
  - \( \pi_1 = 0.25 \cdot \pi_0 + 0.5 \cdot \pi_1 \)
  - Additional constraint: \( \Sigma \pi_i = 1 \)

- Solution: \( \pi_0=2/3, \ \pi_1=1/3 \)
Continuous Time Markov Chains

- Transition between state may happen at any time
- How should probabilities be represented?
  - Probability for infinitesimally small time steps
  - “Transition rate” is suitable description
- “Infinitesimal generator matrix” Q defines rates
  - \( q_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t} \) (for \( i \neq j \))
  - \( q_{ii}(t) = -\sum_{j, j \neq i} q_{ij} \)
- Example:

\[
Q = \begin{pmatrix}
-\lambda & \lambda & 0 \\
\mu & -2\mu & \mu \\
\lambda & 0 & -\lambda
\end{pmatrix}
\]

- Time in a state is memoryless
  - Exponential distribution is memoryless

Exponential Distribution

- Exponential distribution has one parameter
  - \( \lambda \) if arrival rate
  - \( \mu \) if service rate
- Mean: \( \bar{X} = \frac{1}{\lambda} \)
- CDF: \( F_X(r) = 1 - e^{-r/\bar{X}} = 1 - e^{-\lambda r} \)
- pdf: \( f_X(r) = \lambda e^{-\lambda r} \)
- Variance: \( \text{var}(X) = \frac{1}{\lambda^2} \)
- Convenient properties:
  - Number of arrivals in interval \( t \) is Poisson distributed
    - Poisson parameter \( \alpha = \lambda t \) and \( P[X = k] = \alpha^k e^{-\alpha}/k! \)
  - Rates are additive
    - Combination of two exp. dist. with \( \lambda_1 \) and \( \lambda_2 \) has \( \lambda = \lambda_1 + \lambda_2 \)
Steady-State Probability Vector

- By definition rate of leaving state is rate of staying
  \[ q_i(t) = -\sum_{j \neq i} q_{ij} \]

- Steady state probability vector \( \pi \)
  - In steady state, \( \pi Q = 0 \) or \( \sum_{i \in S} q_{ij} \pi_i = 0 \)
    - Change in probability vector is \( d\pi(t)/dt = \sum_{i \in S} q_{ij} \pi_i(t) \)
    - If steady state, then \( \lim_{t \to \infty} [d\pi(t)/dt] = 0 \)
  - Additional constraint: \( \sum \pi_i = 1 \)

Solution to example:

\[ \begin{align*}
-\lambda \pi_0 + \mu \pi_1 - \lambda \pi_2 &= 0 \\
\lambda \pi_0 - 2\mu \pi_1 &= 0 \\
\mu \pi_1 - \lambda \pi_2 &= 0
\end{align*} \]
- Thus, \( \pi_1 = \lambda / \mu \pi_2 \) and \( \pi_0 = 2 \pi_2 \). With constraint, we get
  - \( \pi_0 = 2/(3 + \lambda / \mu) \)
  - \( \pi_1 = \lambda / (3 + \lambda / \mu) = \lambda / (3\mu + \lambda) \)
  - \( \pi_2 = 1 / (3 + \lambda / \mu) \)

Homework

- Read

- SPARK
  - Assessment quiz