Capacity Analysis

Undersea Fiber Network

The secret world of submarine cables
Capacity Analysis

• Deterministic Model
  – Assume: each HTTP request takes $T_r$ (request time)
  – and requests arrive at a known time sequence: $A_s$
  – What is service time per request, $T_s$

• No conflicts:

\[ T_r = 1/2 \]
\[ A_s: 0, 1, 2, 3, \ldots \]
(no conflicts)
\[ T_s = 1/2 \]
Sequenced with conflict:

• **Round Robin:**
  - Request 1 arrives at 0, works ‘till 0.25., waits ‘till .5 works ‘till .75 then ends
  - Request 2 arrives at .25, works ‘till 0.5., waits ‘till .75 works ‘till 1.0 then ends
  
  \[ T_s = \frac{(0.75 + 0.75)}{2} = 6/8 \]

• **First-come-first-serve**
  
  \[ T_r = \frac{1}{2} \]
  
  \[ A_s: 0,.25, 1., 1.25,… \]
  
  (same # requests, sequenced differently)
  
  \[ T_s = \frac{(0.5 + 0.75)}{2} = 5/8 \]
Observations

• If server is handling a single page, then deterministic service times is reasonable, otherwise NOT

• The long term behavior depends on the arrival sequence, $A_s$, AND the service rate, $T_s$

• F-C-F-S is a lower bound on the round robin
Probabilistic Model (F-C-F-S)

• Let inter arrival time and the service times be:
  – independent, identically distributed (iid) random variables having an exponential distribution (denoted by M)
  – a r.v. $X$ has an exponential density iff:
    • $f(x) = \alpha e^{-\alpha x}$ $x > 0$
      
    
    So $P_{\text{rob}}\{ X \leq x \} = 1 - e^{-\alpha x}$ $x \geq 0$
    
    $P_{\text{rob}}\{ X > x \} = e^{-\alpha x}$ $x \geq 0$
    
    $E(X) = 1/\alpha$
    
    $\text{var}(X) = 1/\alpha^2$
Assume server has been running for a while and reached steady state

Arrival Rates

States

Service rates

• In steady state (prob. of being in state \( j \), \( j \) waiting to be served): \( P_j \lambda_j = P_{j+1} \mu_{j+1} \)

or \( P_{j+1} = P_j (\lambda_j / \mu_{j+1}) \) (recursion relationship)

Specifically:

\[
P_1 = (\lambda_0 / \mu_1) P_0 \\
P_2 = (\lambda_1 / \mu_2) P_1 = (\lambda_1 / \mu_2) (\lambda_0 / \mu_1) P_0 \\
\vdots \\
P_{n+1} = (\lambda_n / \mu_{n+1}) P_n = ((\lambda_n \lambda_{n-1} \ldots \lambda_0)/(\mu_{n+1} \mu_n \ldots \mu_1)) P_0
\]

Let’s let: \( C_n = (\lambda_{n-1} \lambda_{n-2} \ldots \lambda_0)/(\mu_n \mu_{n-1} \ldots \mu_1) \), \( n = 1, 2, \ldots \)

Then the Steady State Probabilities are \( P_n = C_n P_0 \), \( n = 1, 2, \ldots \)
• What about $P_0$?
  
  – Since $\sum_{n=0}^{\infty} P_n = 1$, then $\sum_{n=0}^{\infty} C_n P_0 = 1$
    
    $$P_0 + \sum_{n=1}^{\infty} C_n P_0 = 1$$
    $$P_0 [1 + \sum_{n=1}^{\infty} C_n ] = 1$$
    $$P_0 = 1 / [1 + \sum_{n=1}^{\infty} C_n ]$$

  Now, the number of requests in the system, $L$
  
  $$L = 1 P_1 + 2 P_2 + 3 P_3 + ... = \sum_{n=0}^{\infty} (n P_n)$$

  Then the expected waiting time (likely time to be served),
  $$W = L / \lambda$$

  $\lambda$ average arrival time
Simple case: M/M/1/FCFS/∞ /∞

Now suppose \( \lambda_n = \lambda \), \( \mu_n = \mu \), \( n = 1, 2, 3, \ldots \)
then \( C_n = (\lambda^n / \mu^n) = (\lambda / \mu)^n \), \( n = 1, 2, 3, \ldots \)
let \( \rho = \lambda / \mu \)
then \( P_0 = 1 / (1 + \sum_{n=1}^{\infty} \rho^n) = 1 / \sum_{n=1}^{\infty} \rho^n \)
Now recall \( \sum_{n=1}^{\infty} x^n = (1 - x^{n+1}) / (1 - x) \) for any \( x \)
and if \( |x| < 1 \) \( \sum_{n=1}^{\infty} x^n = 1 / (1 - x) \)
We need to have \( \lambda < \mu \) or the system will blow up!
So \( \rho < 1 \)
Thus \( P_0 = 1 / (1 / (1 - \rho)) = 1 - \rho = 1 - (\lambda / \mu) \)
and \( P_n = \rho^n P_0 = \rho^n (1 - \rho) \), \( n = 1, 2, 3, \ldots \)
What about $L$ (expected # of requests in queuing system)?

$L = \sum_{n=1}^{\infty} n P_n = \sum_{n=1}^{\infty} n (1-\rho) \rho^n$

$= (1-\rho) \sum_{n=1}^{\infty} n \rho^n$

Now a little trick: Observe that $n \rho^n = n \rho \rho^{n-1}$

So $L = (1-\rho) \rho \sum_{n=1}^{\infty} n \rho^{n-1}$

Why bother? Because $n \rho^{n-1}$ should be familiar. In particular $d/d\rho (\rho^n) = n \rho^{n-1}$

Thus $L = (1-\rho) \rho \sum_{n=1}^{\infty} d/d\rho (\rho^n) = (1-\rho) \rho \frac{d}{d\rho} \left( \sum_{n=1}^{\infty} (\rho^n) \right)$

$= (1-\rho) \rho \frac{d}{d\rho} \left( \sum_{n=1}^{\infty} (\rho^n) \right)$ but $d/d\rho (\sum_{n=1}^{\infty} (\rho^n))$ is a geometric series

$= (1-\rho) \rho \frac{d}{d\rho} \left(1-\rho \right)^{-1}$ but $\frac{d}{d\rho} \left(1-\rho \right)^{-1} = -1(1-\rho)^{-2}$ (-1)

So $L = \rho (1-\rho)^{-1}$ and we can show $L = \lambda (\mu - \lambda)^{-1}$

and expected time in system $W = L / \lambda = \lambda (\mu - \lambda)^{-1} \lambda^{-1} = (\mu - \lambda)^{-1}$
More of M/M/1/FCFS/$\infty$/$\infty$ Queues

- $L_q =$ expected number of customers in line
  
  $L_q = 0 P_0 + \sum_{n=1}^{\infty} (n - 1) P_n = \sum_{n=1}^{\infty} (n P_n) - \sum_{n=1}^{\infty} (P_n)$
  
  $L_q = L - (1 - P_0) = L - \rho = \lambda^2 / (\mu (\mu - \lambda))$

- $L_s =$ expected number of customers in service
  
  $L_s = 0 P_0 + \sum_{n=1}^{\infty} 1 P_n = 1 - P_0 = 1 - (1 - \rho) = \rho = \lambda / \mu$

- $W_q =$ expected time of customer spends in line
  
  $W_q = L_q / \lambda = \lambda / (\mu (\mu - \lambda))$

- $W_s =$ expected time of customer spends in service
  
  $W_s = L_s / \lambda = 1 / \mu$
What about multiple processors:

\[ C_n = (\lambda_{n-1}\lambda_{n-2} \ldots \lambda_0)/(\mu_n\mu_{n-1} \ldots \mu_1), \quad n = 1, 2, \ldots \]

We again let \( \lambda_n = \lambda \) for every \( n \)

Now assume that there are \( s \) processors.

So when \( n \leq s \) we have \( \mu_n = n\mu \), and when \( n > s \) we have \( \mu_n = s\mu \)

So \( C_n = \lambda^n/(\mu_n\mu_{n-1} \ldots \mu_1) \) \( C_n \) has \( s \) terms of \( n\mu \) and \( n-s \) terms of \( s\mu \)

Hence \( C_n = \lambda^n / (n! \mu^n) \) \( n \leq s \)

\[ = \frac{\lambda^n}{(s!)(s^{n-s}) \mu^n} \quad n > s \]

If \( \lambda < s\mu \) as before we can show:

\[ P_n = 1/ \left( \sum_{n=0, s-1} (1/n!) (\lambda/\mu)^n + (1/s!) (\lambda/\mu)^s (1 - (\lambda/(s\mu))) \right) \]
Finally:  \( P_n = P_0 (\lambda / \mu)^n / (n!) \) \quad n \leq s
            = P_0 (\lambda / \mu)^n / ( (s!) s^{n-s} ) \quad n > s

\[
L = P_0 (\lambda / \mu)^s (\lambda / s \mu) / ( s! ((1- \lambda / s \mu)^2 )) + \lambda / \mu
\]

\[
W = P_0 (\lambda / \mu)^s (\lambda / s \mu) / (\lambda (s!) (1- (\lambda / s \mu))^2 ) + ( 1 + \lambda ) / \mu
\]

Where  \( s = \# \) processors
\( \lambda = \) arrival rates
\( \mu = \) processing rates
Realistic Server Performance

- Consider one server connected to the internet

Model each as a single server queue
Realistic Server Performance, cont.

Arrivals

- Assume:
  - File sizes are exponentially distributed
  - Service rate of browser and Internet are exponentially distributed

- Parameters:
  - Arrival rate
  - Mean file size (5275 bytes)
  - Initialization time (independent of size)
  - Bandwidth @ Server ($T_1 = 1.5 \text{ Mbit/s}$, $T_3 = 6\text{Mbit/s}$)
  - Bandwidth @ Client (700Kbit/s)
More Complicated Model

Arrivals

Departures

- Load balancing strategies (above not very smart)