Priority Queueing

Professor Yeali S. Sun
Department of Information Management
National Taiwan University
Introduction

- Nobody likes to wait in line.
- Priority queueing system
  - A structure that provides preferential treatment at the expense of others.
Objectives of this chapter

- To discuss a few priority systems
- Issues
- Methods of useful approach
- The basis for some computer and network applications.
A Queueing System

- Arrival process: customers to be served
  - Inter-arrival time distribution
- Queue: finite or infinite capacity
- Queueing discipline
- Service time distribution
- Number of servers
The Model (1/2)

- Queueing discipline
  - A means for choosing which customer in the queue is to be serviced next.

- Decision can be based on any or all of the following
  - Some measure related to the relative *arrival times* for those customers in the queue;
    - E.g., first-come-first-serve (FCFS), last-come-first-serve (LCFS).
The Model (2/2)

- Some measure (exact value, estimate, probability density function) of the *service time* required or the services so far received.
  - E.g., shortest-job-first (SJF), longest-job-first (LJF), similar rules based on average, etc.

- Or some function of *group membership* (i.e. the order of service based on an externally imposed priority class structure)
  - e.g., priority queueing
Formulation of Priority Queueing System

- Arriving customers belong to one of a set of $P$ different priority classes, indexed by the subscript $p$ ($p=1, 2, \ldots, P$)
- The larger the class index, the higher the associated priority level.
- Consider equilibrium results of those groups that reach a limiting stable behavior
  - Note that some priority groups may have no stable behavior.
The model

- An arriving customer is assigned a set of parameters.
- The position of a customer in the queue may vary as a function of time owing to the appearance of customers of higher or lower priority in the queue.

\[ q_p(t) \] – priority function of a customer at time \( t \)
  - It is a function of the customer’s assigned parameters, service time, and his time in the system.

- Decision always selects the customer with the largest \( q_p(t) \)
  - It ties, use FCFS.
The model based on the system \( M/G/1 \) (1/2)

- **Arrival process**
  - Customers from priority group \( p \) arrive in a Poisson stream at rate \( \lambda_p \) (customers per second)

- **Service time**
  - Independently from \( B_p(x) \) with mean \( \bar{\chi}_p \) seconds

- \( \rho \) - The fraction of time the server is busy (so long as \( \rho < 1 \))
The model based on the system \( M/G/1 \) (2/2)

\[\lambda = \sum_{p=1}^{P} \lambda_p \]  
\[x = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} x_p \]  
\[\rho_p = \lambda_p x_p \]  
\[\rho = \lambda x = \sum_{p=1}^{P} \rho_p \]
The system is **preemptive** if the process of being served customer is liable to be ejected from service and returned to the queue whenever a higher-priority customer appears in the queue.

Otherwise, non-preemptive
Approach to calculate average waiting times (1/2)

Notation

- $W$ – average waiting time
- $T$ – average system time (queue and service)
- $T = W + \bar{x}$
- $W_p \overset{\Delta}{=} E[\text{waiting time for customers from group } p]$  
- $T_p \overset{\Delta}{=} E[\text{total time in system for customers from group } p]$  
  \[= W_p + x_p \]
Approach to calculate average waiting times (2/2)

Basic observation

- Queueing delay – any delay due to the customer found in service upon arrival; any delay due to the customers found in the queue upon arrival; and any delay due to customers who arrive after he does.

Assumptions

- Consider non-preemptive systems

- From the viewpoint of a newly arriving customer from priority group p (the tagged customer)
1. delay due to the customer found in service

- The in-service customer’s residual service time
- Depending on the service time distribution of the priority group of the in-service customer
- \( W_0 \)
- \( \rho_i \) – the fraction of time that the server is occupied by customers from group i

for the case of Poisson arrival process, \( \rho_i \) is the probability that our tagged customer finds a type-i customer in service.

- The mean residual service time as observed by an arrival is as follows (3.7).

\[
W_0 = \sum_{i=1}^{P} \rho_i \frac{x_i^2}{2x_i} = \sum_{i=1}^{P} \frac{\lambda_i x_i^2}{2}
\]
2. delay due to the customers found in the queue upon arrival

- Customers who receive service before the tagged customer

\[ N_{ip} = \Delta \]

The number of customers from group \( i \) found in the queue by our tagged customer (from group \( p \)) and who receive service before our tagged customer does.
2. delay due to the customers found in the queue upon arrival

- \[ E[N_{ip}] \equiv N_{ip} \]

- Total average delay (3.9)

\[
\sum_{i=1}^{P} x_i N_{ip}
\]
3. delay due to customers who arrive after the tagged customer

- \( M_{ip} = \Delta \)  
The number of customers from group \( i \) who arrive to the system while our tagged customer (from group \( p \)) is in the queue and who receive service before he does

- Average total delay (3.11)

\[
W_p = W_0 + \sum_{i=1}^{P} x_i (N_{ip} + M_{ip}), \quad p = 1,2,..., P
\]
3. delay due to customers who arrive after the tagged customer (cont’d)

Solution procedure

1. Find the averages of $N_{ip}$ and $M_{ip}$
2. find solution of (3.11)
Analysis of the Delay Cycles

Goal:

• Calculate the Laplace transform of the pdf of “generalized” busy periods
• Obtain the Laplace transform for the waiting time density
The Delay Cycle

- $Y_0$ – the initial delay
  - E.g., in-service residual service time, or some “special” task
- $Y_b$ - delay busy period duration
  - The period to service “ordinary” customers;
  - May view as a sequence of sub-busy periods
Busy Period (1/3)

Unfinished work

Server

Queue

or by any means without the prior written permission of the author.
Busy Period (2/3)

(a) Decomposition of the busy period

Service time for $C_1$

Sub-busy period generated by $C_4$

Sub-busy period generated by $C_3$

Sub-busy period generated by $C_2$

Busy period generated by $C_1$
Busy Period (3/3)

(b) Number in the system

(c) Customer history

N(t)

Copyright 2005 Yeali S. Sun. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, or transmitted in any form, or by any means without the prior written permission of the author.
**Delay busy period duration**

- The time interval that services
  - an ordinary customer and
  - all those customers (his descendents) who enter the system during his service time
  - or during the service time of any of his descendents (each descendent will generate his own sub-busy period).

- The pdf for the duration of a sub-busy period is the same as that for a busy period.
Assumptions

Assumption

- Poisson arrival process: □

Notation

- $B^*(s)$: transform for the pdf of service time
- $G^*(s)$: transform for the pdf of busy period durations
- $G_0(y)$: PDF of $Y_0 \rightarrow G_0^*(s)$
- $G_b(y)$: PDF of $Y_b \rightarrow G_b^*(s)$
- $G_c(y)$: PDF of $Y_c \rightarrow G_c^*(s)$
Delay Cycle Analysis: $G_b^*(s)^{(1/3)}$

- $N_0$ - the number of customers arrivals in $Y_0$
- We have $G^*(s) = E[e^{-sY_b}]$
- Assume $n$ sub-busy periods (one is generated by each of the arrivals during $Y_0$) are independent and have identical distribution as a busy period.

$$E[e^{-sY_b} \mid Y_0 = y, N_0 = n] = [G^*(s)]^n$$
Delay Cycle Analysis: $G_b^*(s)$ (2/5)

- Uncondition $N_0$

$$E[e^{-sY_b} \mid Y_0 = y] = \sum_{n=0}^{\infty} [G^*(s)]^n \frac{(\lambda y)^n}{n!} e^{-\lambda y}$$

$$= e^{-[\lambda - \lambda G^*(s)]y}$$

- Uncondition $Y_0$

$$E[e^{-sY_b}] = G_b^*(s) = \int_{y=0}^{\Delta} e^{-[\lambda - \lambda G^*(s)]y} dG_0(y)$$

$$G_b^*(s) = G_0^*(\lambda - \lambda G^*(s))$$

Poisson arrival
Delay Cycle Analysis: \( G_c^* (s) \) (3/5)

\[
E[e^{-s Y_c} \mid Y_0 = y, N_0 = n] = e^{-sy} [G^* (s)]^n
\]

\[
G_c^* (s) \overset{\Delta}{=} E[e^{-s Y_c}] = \int e^{-sy} \sum_{n=0}^{\infty} [G^* (s)]^n \frac{(\lambda y)^n}{n!} e^{-\lambda y} dG_0 (y)
\]

\[
G_c^* (s) = G_0^* (s + \lambda - \lambda G^* (s))
\]

\[
G^* (s) = B^* (s + \lambda - \lambda G^* (s))
\]
Conservation Laws

Observations

- For priority system, “preferential treatment” given to one class of customers is afforded at the expense of other customers”

- “Work Conserving system” – no work (service requirement) is created or destroyed within the system.

- Wish to find the invariants of the system.
Systems with queueing disciplines independent of customer service time

- We know “waiting time distribution depends on the order of service”.
- For queueing systems with queueing disciplines independent of customer service time, then

  the distributions of number of customers in the system and average waiting time are invariant to the order of the service.
The distribution of the number of customers in the system

- For M/G/1 system having FCFS, LSCF, etc.

\[ Q(z) = B^* (\lambda - \lambda z) \frac{(1 - p)(1 - z)}{B^* (\lambda - \lambda z) - z} \] (3.15)
M/G/1 Conservation Law

For any M/G/1 system with any non-preemptive work-conserving queueing discipline, the conservation law (unfinished work)

\[ \sum_{p=1}^{P} \rho P W_p = \frac{\rho W_0}{1 - \rho} \quad \rho < 1 \]

or

\[ \infty \quad \rho \geq 1 \]

- \( W_0 \): residual service time seen upon an arrival
Conservation Law: derivations (1/2)

- **Unfinished work**

  \[ U(t) = x_0 + \sum_{p=1}^{P} \sum_{i=1}^{n_p} x_{ip} \]

- **Expected unfinished work**

  \[ E[U(t)] = W_0 + \sum_{p=1}^{P} \sum_{n_p=0}^{\infty} P[N_p(t) = n_p] \sum_{i=1}^{n_p} E[x_{ip}] \]

- **Limiting average \( t \to \infty \)**

  \[ \overline{U} = W_0 + \lim_{t \to \infty} \sum_{p=1}^{P} \sum_{n_p=0}^{\infty} n_p P[N_p(t) = n_p] \overline{x_p} = W_0 + \sum_{p=1}^{P} x_p E[N_p] \]
By Little’s Result, we have $E[N_p] = \mathbb{E}_p W_p$

$$\bar{U} = W_0 + \lim_{t \to \infty} \sum_{p=1}^{P} \sum_{n_p=0}^{\infty} n_p P[N_p(t) = n_p] x_p = W_0 + \sum_{p=1}^{P} x_p E[N_p]$$

$$\bar{U} = W_0 + \sum_{p=1}^{P} \rho_p W_p$$
Consider FCFS

- For Poisson arrivals, **average unfinished work** is the average waiting time

\[
W = \frac{\lambda x^2 / 2}{1 - \rho}
\]

- We know

\[
\bar{x}^2 = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} \bar{x}_p^2 = \frac{2W_0}{\lambda}
\]

- Thus

\[
\bar{U} = W = \frac{W_0}{1 - \rho}
\]
M/G/1 Conservation Law

- For any M/G/1 system with any non-preemptive work-conserving queueing discipline, the conservation law (unfinished work)

\[ \sum_{p=1}^{P} \rho_p W_p = \frac{\rho W_0}{1 - \rho} \quad \rho < 1 \]

or

\[ \infty \quad \rho \geq 1 \]

- \( W_0 \): residual service time seen upon an arrival
**M/G/1: Average Unfinished Work**

- For priority queueing discipline

\[
\sum_{p=1}^{P} \rho_p W_p = \frac{\rho W_0}{1 - \rho}, \quad \rho < 1 \quad (3.16)
\]

- Any attempt to modify the queueing discipline to reduce a \( W_p \) will force an increase in some of the other \( W_p \)
- Not an “even trade” since the \( p \) may be distinct.
The $G/G/1$ Conservation Law

Assumption – equilibrium distributions exist.

\[
\sum_{p=1}^{P} \lambda_p W_p = \overline{U} - W_0 \quad (3.22)
\]
Waiting time for LCFS (1/3)

- Consider a simple queueing system with one single priority class
- Has the same average queue size and average waiting time with FCFS discipline
  - Why?
- For waiting time distribution
  - Larger variance
  - Good for analysis
  - Independent of the customers found in the queue on arrival
  - ascribed to Customers who arrive the system after he does but prior to his initiation of service.
Waiting time for LCFS (2/3)

- Use delay cycle analysis

\[
E[e^{-s\hat{w}} \mid \text{system busy upon arrival}] = G_c^*(s)
\]

\[
= G_0^* \left( s + \lambda - \lambda G^*(s) \right)
= \frac{1 - B^* \left( s + \lambda - \lambda G^*(s) \right)}{[s + \lambda - \lambda G^*(s)]x}
= \frac{1 - G^*(s)}{[s + \lambda - \lambda G^*(s)]x}
\]

\[
G_0^*(s) = \frac{1 - B^*(s)}{sx}
\]

\[
G^*(s) = B^* (s + \lambda - \lambda G^*(s))
\]
Waiting time for LCFS (3/3)

\[ E[e^{-s\bar{w}} \mid \text{system busy upon arrival}] = \frac{1 - G^*(s)}{[s + \lambda - \lambda G^*(s)]} \]

\[ W^*(s) = E[e^{-s\bar{w}}] = 1 - \rho + \frac{\lambda[1 - G^*(s)]}{s + \lambda - \lambda G^*(s)} \]

- Probability found system is empty
- Probability \( \bar{W} \) found system is busy
Compare

- **FCFS**
  
  \[ W^*(s) = \frac{s[1-\rho]}{s + \lambda - \lambda B^*(s)} \]

- **LCFS**
  
  \[ W^*(s) = 1 - \rho + \frac{\lambda[1-G^*(s)]}{s + \lambda - \lambda G^*(s)} \]

- **Larger second moment** (variance)
Head-of-the-line priority queue (1/5)

- aka. ”Strict priority queueing” or “fixed priority queueing,” i.e., priority value remains constant in time.

Want to find the average waiting time $W_p$
Head-of-the-line priority queue (2/5)

- The average waiting time $W_p$

$$W_p = W_0 + \sum_{i=p+1}^{P} x_i (N_{ip} + M_{ip}) + x_p \cdot N_{pp}$$

$$= W_0 + \sum_{i=p}^{P} x_i \lambda_i W_i + \sum_{i=p+1}^{P} x_i \lambda_i W_p$$

- Rewrite

$$W_p = \frac{W_0 + \sum_{i=p+1}^{P} \rho_i W_i}{1 - \sum_{i=p}^{P} \rho_i}$$

N: come before tagged customer
M: come after

By Little’s Result
Head-of-the-line priority queue (3/5)

- A set of equations
- Solve the equations recursively
- Find $W_P$, $W_{P-1}$, … $W_1$

$\sigma_p = \sum_{i=p}^{P} \rho_i$

$$W_p = \frac{W_0}{(1-\sigma_p)(1-\sigma_{p+1})} \quad p = 1,2,...,P \quad (3.31)$$
Head-of-the-line priority queue (4/5)

- Finite queueing time for priority group $p$ iff
  \[ \rho - (\rho_1 + \rho_2 + \rho_3 + \ldots + \rho_{p-1}) < 1 \]
i.e. $\rho < 1 + (\rho_1 + \rho_2 + \rho_3 + \ldots + \rho_{p-1})$

- Different groups may have different service time distribution

- Conservation law holds

- Different distribution of the number of customers in the system from that for FCFS.
Head-of-the-line priority queue (5/5)

Waiting time distribution

\[ W_p^*(s) = \frac{(1-p)[s + \lambda_H - \lambda_H G_H^*(s)] + \lambda_L [1 - B_L^*(s + \lambda_H - \lambda_H G_H^*(s))]}{s - \lambda_p + \lambda_p B_p^*(s + \lambda_H - \lambda_H G_H^*(s))} \]

\[ \lambda_H = \sum_{i=p+1}^{P} \lambda_i \]

\[ \lambda_L = \sum_{i=1}^{p-1} \lambda_i \]

\[ B_H^*(s) = \sum_{i=p+1}^{P} \frac{\lambda_i}{\lambda_H} B_i^*(s) \]

\[ B_L^*(s) = \sum_{i=1}^{p-1} \frac{\lambda_i}{\lambda_L} B_i^*(s) \]

\[ G_H^*(s) = B_H^*(s + \lambda_H - \lambda_H G_H^*(s)) \]
HOL with nonpreemptive: \( P=5, \quad \lambda_p = \lambda / 5, \quad x_p = x \)
HOL with nonpreemptive: \( P=5, \ \lambda_p = \lambda / 5, \ 
\bar{x}_p = \bar{x} \)
Shortest Job First (1/2)

Assumptions

- Each job arrives with exact known service time.
- The shorter the service time, the higher the priority class.
- Customers whose service times $x < x^* \leq x + dx$ are in one group.
- The fraction of customers in the group is $b(x)dx$.
- $b(x)$ – pdf of service time.
Shortest Job First (2/2)

- Average waiting time \( W(x) \) for a customer of service time \( x \)

\[
\sum_{i=p}^{P} \rho_i \rightarrow \int_{y=0}^{x^+} p(y)dy
\]

\[
W(x) = \frac{W_0}{[1 - \lambda \int_{y=0}^{x^-} yb(y)dy][1 - \lambda \int_{y=0}^{x^+} yb(y)dy]}
\]

\[
W_p = \frac{W_0}{(1 - \sigma_p)(1 - \sigma_{p+1})} \quad p = 1, 2, \ldots, P \quad (3.31)
\]

\[
\lim_{x \to \infty} W(x) = \frac{W_0}{(1 - \rho)^2}
\]

\[
\lim_{x \to 0} W(x) = W_0
\]
HOL with Pre-emptive Queueing (1/2)

- Preemptive-resume type
- $T_p$ – avg system time of the tagged customer of priority $p$

$$T_p = x_p + \sum_{i=p}^{p} \frac{\lambda_i x_i^2}{2} + \frac{\sum_{i=p+1}^{p} \rho_i T_p}{1-\sigma_p}$$

- The average unfinished work found upon arrival (by Conservation Law)
- The average work incurred after my arrival due to higher priority customers
HOL with Pre-emptive Queueing (2/2)

- Average system time of customer of group p

\[
T_p = \frac{x_p (1-\sigma_p) + \sum_{i=p}^{p} \lambda_i x_i^2 / 2}{(1-\sigma_p)(1-\sigma_{p+1})}
\]

\[
\sigma_p = \sum_{i=p}^{p} \rho_i
\]

Compare with “Non-preemptive” case

\[
T_p = \frac{x_p (1-\sigma_p)(1-\sigma_{p+1}) + \sum_{i=1}^{p} \lambda_i x_i^2}{2(1-\sigma_p)(1-\sigma_{p+1})}
\]
Non-preemptive HOL with cost (1/3)

- Consider a system cost (rate) of $C_p$ for each customer from priority group $p$.
- Average system cost per second

$$C = \sum_{p=1}^{P} C_p \overline{N}_p$$

$$C = \sum_{p=1}^{P} \rho_p C_p + \sum_{p=1}^{P} C_p \lambda_p W_p \quad (3.40)$$
Non-preemptive HOL with cost (2/3)

- Objective – find an appropriate queueing discipline to minimize $C$

$$C = \sum_{p=1}^{P} \rho_p C_p = \sum_{p=1}^{P} \left( \frac{C_p}{x_p} \right) (\rho_p W_p)$$

- $C_p$ and $x_p$ are given
Non-preemptive HOL with cost (3/3)

Answer

\[ C - \sum_{p=1}^{P} \rho_p C_p = \sum_{p=1}^{P} (C_p / x_p)(\rho_p \overline{W}_p) \]

\[ \sum_{p=1}^{P} g_p = \text{constant with respect to queue discipline} \]

\[ f_1 \leq f_2 \leq \cdots \leq f_p \quad (3.42) \]
Homework

- Pages 150-154: 3.2, 3.9, 3.11 and 3.16