Optimal stochastic scheduling

690PE_16 Class 15
Setting

- Single server, infinite waiting room
- K classes of jobs
- Class-k job has to paid constant fee $c_k$ per time unit it resides in the system, $k = 1,...,K$
- Deterministic interarrival times ($0 = t_1 < t_2 < ....$)
- Deterministic service times
- Server handles one unit of work per unit of time (job with service time $S$ spends $S$ units of time in server)
A Resource Allocation Policy (RAP) determines, at any time, which class of jobs should be served.

Objective: \[ C(t) := \min_{\pi \in \Pi} \sum_{k=1}^{K} c_k E[Q_k^\pi(t)] \text{ for any } t > 0 \]

with \[ Q_k^\pi(t) \] # class-k at time t under \( \pi \)

\( \Pi \) = set of RAPs that may preempt job in server, and may use information regarding future arrivals, service times of future jobs and service times of jobs in the system.
Let \( \gamma \in \Pi \) be RAP that always allocates server to jobs in system with highest holding fee, i.e. jobs of class \( i \) have priority over jobs of class \( i+1, \ldots, R \). Define \( V_k^{\pi}(t) \) total workload of class-\( k \) jobs at time \( t \).

**Result 1:** For \( r_1 \geq \ldots \geq r_K \)

\[
\sum_{k=1}^{K} r_k V_k^\gamma(t) \leq \sum_{k=1}^{K} r_k V_k^\pi(t) \text{ for } t>0, \pi \in \Pi
\]

**Proof.** (\( a_1, \ldots, a_K \), \( b_1, \ldots, b_K \)) s.t. \( \sum_{i=1}^{K} a_i \leq \sum_{i=1}^{K} b_i \) for \( k=1,2,\ldots,K \)

\[
\sum_{k=1}^{K} r_k a_k = \sum_{k=1}^{K} a_k \sum_{i=k}^{K} (r_i - r_{i+1}) = \sum_{k=1}^{K} (r_k - r_{k+1}) \sum_{i=1}^{k} a_i \leq \sum_{k=1}^{K} (r_k - r_{k+1}) \sum_{i=1}^{k} b_i = \sum_{k=1}^{K} r_k b_k \quad (r_{K+1} := 0)
\]

Proof concluded as \( \sum_{i=1}^{K} V_i^\gamma(t) \leq \sum_{i=1}^{K} V_i^\pi(t) \) (equality for all work-conserving policies)
Class-\(k\) service requirements are exponentially distributed with mean \(1/\mu_k\).

Service times are all independent.

Class \(\Gamma \subset \Pi\) of policies which do not know present and future service requirements.

Note that \(\gamma \in \Gamma\).
Result 2: For any $\pi \in \Gamma$, $E[Q^\pi_k(t)] = \mu_k E[V^\pi_k(t)]$

Proof. $V^\pi_k(t) = \sum_{j=1}^{Q^\pi_k(t)} \sigma^\pi_{j,k}(t)$ with $\sigma^\pi_{j,k}(t)$ service time of $j$-th oldest class-$k$ job in system under $\pi$

$$E[V^\pi_k(t)] = \sum_{n=0}^{\infty} \left[ \sum_{j=1}^{n} E[\sigma^\pi_{j,k}(t) \mid Q^\pi_k(t) = n] \right] P(Q^\pi_k(t) = n) = E[Q^\pi_k(t)] / \mu_k$$

as $E[\sigma^\pi_{j,k}(t) \mid Q^\pi_k(t) = n] = 1 / \mu_k$ since

- $\pi$ does not know present and future service requirements (those are typical jobs)
- service times are memoryless
Label classes such that $\mu_1 c_1 \geq \ldots \geq \mu_K c_K$

From

$$\sum_{k=1}^{K} r_k V_k^\pi(t) \leq \sum_{k=1}^{K} r_k V_k^\gamma(t)$$

for all $\pi \in \Pi$ when $r_1 \geq r_2 \geq \ldots \geq r_K$

and

$$E[Q_k^\pi(t)] = \mu_k E[V_k^\pi(t)]$$

for all $\pi \in \Gamma$, $t > 0$

we get

$$\sum_{k=1}^{K} c_k E[Q_k^\gamma(t)] = \sum_{k=1}^{K} \mu_k c_k E[V_k^\gamma(t)] \leq \sum_{k=1}^{K} \mu_k c_k E[V_k^\pi(t)] \leq \sum_{k=1}^{K} c_k E[Q_k^\pi(t)]$$

for all $\pi \in \Gamma$, $t > 0$

⇒ Celebrated $\mu c$-rule that gives preemptive priority
to classes in decreasing order of their $\mu_k c_k$
Performance of preemptive priority policy for M/G/1 queue

- Single server, infinite waiting room
- K classes of jobs:
  - Class-k arrive according to Poisson process, rate $\lambda_k$
  - $G_k(x) = \text{cd of service times of class-k jobs, mean } 1/\mu_k$, 2nd order-moment $\sigma_k^{(2)}$
- Class-i has preemptive priority over class-j, if $i < j$
- All above rvs mutually independent
Consider first exponential service times

\{X(t)=(X_1(t), \ldots, X_K(t)), t>0\} with \(X_k(t)\)\textsuperscript{\#} class-\(k\) at time \(t\) is a CTMC on \(\{0,1,\ldots\}^K\)

**Balance eqns:** \(n = (n_1, \ldots, n_K)\)

\[
\pi(n) \left( \sum_{k=1}^{K} \lambda_k + \sum_{k=1}^{K} \mu_k 1(n_1 = \ldots = n_{k-1} = 0, n_k > 0) \right) = \sum_{k=1}^{K} \pi(n - e_k) \lambda_k 1(n_k > 0)
\]

\[
+ \sum_{k=1}^{K} \pi(n + e_k) \mu_k 1(n_1 = \ldots = n_{k-1} = 0, n_k > 0) \quad \text{with } n_0 = 0
\]

Can be solved using z-transform. Will use different approach (but will get less, no free lunch \ldots)
Service times are no longer assumed to be exponentially distributed

\( \Rightarrow \) Multiclass \( M/G/1 \) queue. Set \( \rho_k = \frac{\lambda_k}{\mu_k} \) \( k=1, \ldots, K \)

Let us calculate

- \( E[X_k] \), expected \# of class-\( k \) jobs
- \( E[W_k] \), expected waiting time of class-\( k \) jobs
- \( T_k \), expected sojourn time of class-\( k \) jobs

**Observation:** \( E[X_1], E[W_1] \) and \( T_1 \) obtained from \( M/G/1 \) queue with arrival rate \( \lambda_1 \) and cdf \( G_1(x) \) for service times
Service times are no longer assumed to be exponentially distributed

\( \Rightarrow \text{Multiclass } M/G/1 \text{ queue} \)

- \( k = 1 \)

\[
\begin{align*}
E[X_1] &= \rho_1 + \frac{\lambda_1^{2}\bar{\sigma}_1^{(2)}}{2(1-\rho_1)}, \\
E[W_1] &= \frac{\lambda_1 \bar{\sigma}_1^{(2)}}{2(1-\rho_1)}, \\
T_1 &= \frac{1}{\mu_1} + \frac{\lambda_1 \bar{\sigma}_1^{(2)}}{(1-\rho_1)}
\end{align*}
\]

when \( \rho_1 < 1 \)
\( k \geq 2 \)

\[ T_k = \frac{1}{\mu_k} + T_{k,1} + T_{k,2} \]

with

- \( T_{k,1} = \text{expected time to serve all class-} i\text{-} \text{jobs, } i=1,\ldots,k \text{ present in the system upon arrival of the class-} k\text{ job} \)

- \( T_{k,2} = \text{expected time to serve all class-} i\text{-} \text{jobs, } i=1,\ldots,k-1 \text{ that will arrive while the class-} k\text{ job is in the system} \)
$T_{k,1} = \text{expected time to serve all class-}i \text{ jobs, } i=1,...,k \text{ present in system upon arrival of a class-}k \text{ job}$

$T_{k,1}$ same as expected waiting in $M/G/1$ queue with

- arrival rate $\Lambda_k = \sum_{1 \leq i \leq k} \lambda_i$
- cdf of service times $H_k(x) = \sum_{1 \leq i \leq k} (\lambda_i / \Lambda_k) G_i(x)$

\[ T_{k,1} = \frac{\sum_{i=1}^{K} \lambda_i \sigma_i^{(2)}}{2 \left(1 - \sum_{i=1}^{k} \rho_i \right)} \text{ provided that } \sum_{1 \leq i \leq k} \rho_i < 1 \]
\( T_{k,2} \) = expected time to serve all class-\( i \) jobs, \( i=1,\ldots,k-1 \) that arrive while a class-\( k \) job is in system

\[ \tau_k := \text{sojourn time of class-} k. \quad E[\tau_k] = T_k \]

\[ T_{k,2} = \sum_{1 \leq i \leq k-1} E[\text{# class-} i \text{ arrived in } \tau_k] \times (1/\mu_i) \]

\[ E[\text{# class-} i \text{ arrived in } \tau_k] = \int_0^\infty E[\text{# class-} i \text{ arrived in } \tau_k | \tau_k = x] d\tau_k(x) = \int_0^\infty \lambda_i x d\tau_k(x) = \lambda_i E[\tau_k] = \lambda_i T_k \]

Poisson arrivals, rate \( \lambda_i \)

\[ T_{k,2} = T_k \sum_{1 \leq i \leq k-1} \rho_i \]
Recap.: \( T_k = \frac{1}{\mu_k} + T_{k,1} + T_{k,2} \) with

\[ T_{k,1} = \sum_{i=1}^{k} \frac{\lambda_i \sigma_i}{2 \left( 1 - \sum_{i=1}^{k} \rho_i \right)} \], \( T_{k,2} = T_k \sum_{1 \leq i \leq k-1} \rho_i \)

Therefore \( T_k = \frac{1}{\mu_k} + \frac{\sum_{i=1}^{k} \lambda_i \sigma_i}{2 \left( 1 - \sum_{i=1}^{k} \rho_i \right)} + T_k \sum_{i=1}^{k-1} \rho_i \)

giving \( T_k = \frac{1}{\left( \frac{1}{k-1} - \sum_{i=1}^{k-1} \rho_i \right)} \left( \frac{1}{\mu_k} + \frac{\sum_{i=1}^{k} \lambda_i \sigma_i}{2 \left( 1 - \sum_{i=1}^{k} \rho_i \right)} \right) \) provided \( \sum_{1 \leq i \leq k} \rho_i < 1 \)
Finally, for $k=1, \ldots, K$

$$T_k = \left( \frac{1}{1-\sum_{i=1}^{k-1} \rho_i} \right) \left( \frac{1}{\mu_k} + \frac{\sum_{i=1}^{k} \lambda_i \sigma_i^{(2)}}{2 \left( 1-\sum_{i=1}^{k} \rho_i \right)} \right), \quad E[W_k] = T_k - \frac{1}{\mu_k}$$

provided $\sum_{1 \leq i \leq K} \rho_i < 1$, and by Little

$$E[X_k] = \frac{1}{\left( \frac{1}{\sum_{i=1}^{k-1} \rho_i} \right)} \left( \frac{\lambda_k \sum_{i=1}^{k} \lambda_i \sigma_i^{(2)}}{\rho_k + \sum_{i=1}^{k} \rho_i} \right)$$
Performance of non-preemptive priority policy for $M/G/1$ queue

- Single server, infinite waiting room
- $K$ classes of jobs:
  - Class-$k$ arrive according to Poisson process, rate $\lambda_k$
  - $G_k(x) = \text{cd of service times of class-}k\text{ jobs, mean }1/\mu_k$, 2nd order-moment $\sigma_k^{(2)}$
- Class-$i$ has preemptive priority over class-$j$, if $i<j$ but preemption from server not allowed
- All above rvs mutually independent
Class-1 jobs:

Similar to analysis of M/G/1 queue with \( R \) the expected residual time (use of PASTA)

\[
\overline{W}_1 = \frac{\overline{R}}{1 - \rho_1}
\]

Class-2 jobs:

\[
\overline{W}_2 = \overline{R} + \frac{1}{\mu_1} \overline{X}_1 + \frac{1}{\mu_2} \overline{X}_2 + \frac{1}{\mu_1} E[Z_1]
\]

- \( \overline{X}_k \) = expected \# of class-\( k \) in waiting room
- \( Z_1 \) = \# class-1 arrivals during wait in queue of class-2

By Little

\[
\overline{X}_1 = \lambda_1 \overline{W}_1, \quad \overline{X}_2 = \lambda_2 \overline{W}_2
\]
Therefore
\[
W_2 = R + \rho_1 W_1 + \rho_2 W_2 + \frac{1}{\mu_1} E[Z_1]
\]

\[
E[Z_1] = \int_0^\infty E[Z_1 | W_2 = x] dW_2(x) = \int_0^\infty \lambda_1 x dW_2(x) = \lambda_1 \overline{W}_2
\]
since class-1 arrivals are Poisson with rate \(\lambda_1\)

Hence,
\[
\overline{W}_2 = \frac{\overline{R} + \rho_1 \overline{W}_1}{1 - \rho_1 - \rho_2} = \frac{\overline{R}}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}
\]
as \[
\overline{W}_1 = \frac{\overline{R}}{1 - \rho_1}
\]

Repeating the argument gives
\[
\overline{W}_k = \frac{\overline{R}}{\left(1 - \sum_{i=1}^{k-1} \rho_i\right) \left(1 - \sum_{i=1}^{k} \rho_i\right)}
\]
provided \(\sum_{i=1}^{k} \rho_i < 1\)
**Calculation of \( \bar{R} \)**

Same calculation as that for \( M/G/1 \) queue with arrival rate \( \Lambda := \sum_{1 \leq k \leq K} \lambda_k \) and 2\textsuperscript{nd} order moment of service times \( (1/\Lambda) \sum_{1 \leq i \leq K} \lambda_i \sigma_i \)

Hence \( \bar{R} = \sum_{i=1}^{K} \lambda_i \frac{\sigma_i^{(2)}}{2} \) and finally

\[
\bar{W}_k = \frac{\sum_{i=1}^{K} \lambda_i \sigma_i^{(2)}}{2 \left(1 - \sum_{i=1}^{k-1} \rho_i\right) \left(1 - \sum_{i=1}^{k} \rho_i\right)}
\]

for \( k=1,\ldots,K \) provided \( \sum_{i=1}^{K} \rho_i < 1 \)
Finally,

\[
T_k = \frac{1}{\mu_k} + \frac{\sum_{i=1}^{K} \lambda_i \bar{O}_i^{(2)}}{2 \left( 1 - \sum_{i=1}^{k-1} \rho_i \right) \left( 1 - \sum_{i=1}^{k} \rho_i \right)}
\]

\[
E[X_k] = \rho_k + \frac{\lambda_k \sum_{i=1}^{K} \lambda_i \bar{O}_i^{(2)}}{2 \left( 1 - \sum_{i=1}^{k-1} \rho_i \right) \left( 1 - \sum_{i=1}^{k} \rho_i \right)}
\]

for \( k=1,\ldots,K \) provided \( \sum_{i=1}^{K} \rho_i < 1 \) by Little
Kleinrock’s conservation law

Class of policies:

1. All customers remain in system until completely serviced
2. There is a non-idling single server equipped with infinite waiting room
3. Pre-emption is allowed only if service time distributions are exponential, and a preempted customer resumes its service
4. Arrivals are Poisson, service time distributions are arbitrary; arrivals and service times are independent
Kleinrock’s conservation law (cont’)

Consider K different priorities classes:

- Class-i arrival rate is $\lambda_i$ units per second
- Each unit of class-i has a generally distributed service requirement with mean $1/\mu_i$ and 2$^{nd}$ order-moment $\sigma_i^{(2)}$

Define

\[
\lambda = \sum_{1 \leq i \leq K} \lambda_i \\
1/\mu = \sum_{1 \leq i \leq K} \left( \lambda_i / \lambda \right) \left(1/\mu_i\right) \\
\rho = \lambda/\mu = \sum_{1 \leq i \leq K} \rho_i
\]
Kleinrock’s conservation law (cont’)

- $W_i$: expected waiting of class-i

Result (Kleinrock’s conservation law)

$\sum_{1 \leq i \leq K} \rho_i W_i$ does not depend on service policy, and

$\sum_{1 \leq i \leq K} \rho_i W_i = R[\rho/(1-\rho)]$ if $\rho<1$ and $\sum_{1 \leq i \leq K} \rho_i W_i$ if $\rho \geq 1$

with $R = (1/2) \sum_{1 \leq i \leq K} \lambda_i \sigma_i^{(2)}$ the expected residual load in server at arbitrary time (or at arrival times since PASTA holds here)
Kleinrock’s conservation law (cont’)

Consequences:

- Improvement in expected waiting for a class will necessarily degrades that of another class.

- If all $\mu_i$’s are equal then $\sum_{1 \leq i \leq K} \lambda_i W_i$, the expected number of units in queue by Little, does not depend on service policy.
Proof: $U(t) = \text{unfinished work at time } t$

$U(t)$ does not depend on order of service as no load created/destroyed and server non-idling by (1)-(3)

$V = \lim_{t \to \infty} \frac{1}{t} \int_0^t U(x) \, dx$ does not depend on order of service
Since PASTA applies by (4)

\[ V = W \]

with \( W \) expected waiting time in FIFO M/G/1 queue with arrival rate \( \lambda \), mean service time \( 1/\mu \) and 2\(^{nd}\) order-moment \( \sum_{1 \leq i \leq K} \lambda_i \sigma_i^{(2)} \)

Therefore \( W = R/(1-\rho) \) for \( \rho < 1 \)

with \( R = (1/2) \sum_{1 \leq i \leq K} \lambda_i \sigma_i^{(2)} \)

Note that \( R \) does not depend on service policy
For any service policy $\pi$

\[
\frac{R}{1-\rho} = W = V = R + \sum_{1 \leq i \leq K} E_{\pi}[\# \text{ class-}i \text{ waiting}]x(1/\mu_i) = \\
= R + \sum_{1 \leq i \leq K} \lambda_i W_i(\pi)x(1/\mu_i) = \\
= R + \sum_{1 \leq i \leq K} \rho_i W_i(\pi)
\]

by Little

Therefore, \(\sum_{1 \leq i \leq K} \rho_i W_i(\pi) = \frac{R}{1-\rho} - R = \frac{R\rho}{1-\rho}\)

Proves that \(\sum_{1 \leq i \leq K} \rho_i W_i(\pi)\) does not depend on $\pi$
Checkpoint

- Today Lecture 15 (10 more to go)
  Lectures 17 and 24 likely given by Bo Jiang
- HW1 & HW2 graded

HW1: average 87.68, standard deviation 10.86
100=2, 7>95, 95≥3>90, 89≥4>80, 79≥3>70, 70≥2

HW2: av. 85.77, st. 13.63
100=1, 99≥5>95, 95≥2>90, 89≥7>80, 79≥2>70, 70≥2
Checkpoint (cont’)

- Grades for HW3 known by end of this week
- Mid-term take-home: Nov. 3, due Nov. 13
- HW4 Nov. 17 due Nov. 29 (Thanksgiving week)
- HW5 Dec. 7 due Dec. 15
- Final class, in class: Dec. 22 (10:30-12:30)