Dynamic control of a multi class $G/M/1+M$ queue with abandonments

Alexandre Salch,
Jean-Philippe Gayon, Pierre Lemaire

G-SCOP
Grenoble-INP

24 January 2012

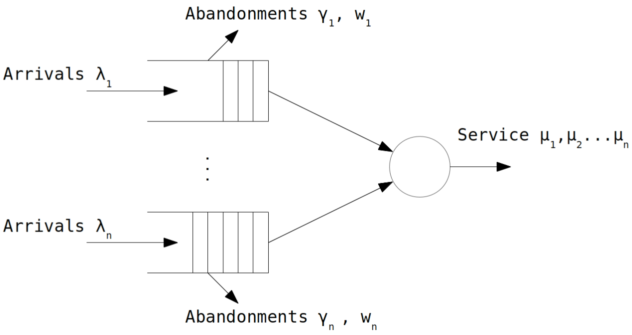
{alexandre.salch,jean-philippe.gayon,pierre-lemaire}@grenoble-inp.fr

Context

- Jobs arrive **randomly**
- They wait until the **end of service**
- If they are not processed, they **abandon with a cost** (no holding costs)

Examples

- Call centers
- Emergency department



Literature review

Down et al. [DKL11]

- Single server
- $n = 2$ classes of jobs
- Poisson arrivals, processing times $X_j \sim exp(\mu_j)$, due dates $D_j \sim exp(\gamma_j)$
- If $\mu_1 = \mu_2$, $\gamma_1 \leq \gamma_2$ and $w_1\gamma_1 \geq w_2\gamma_2 \Rightarrow$ Give priority to class 1

Atar et al. [AGS10]

- n classes of jobs
- Poisson arrivals, processing times $X_j \sim exp(\mu_j)$, due dates $D_j \sim exp(\gamma_j)$
- Many servers fluid scaling

$\Rightarrow$ Give priority to the class of highest $w_j\mu_j/\gamma_j$

Model description

Parameters

- n jobs (n arrivals)
- Processing times $X_j \sim exp(\mu_j)$
- Due dates $D_j \sim exp(\gamma_j)$
- Arrival times R_j : arbitrary
- Abandonment costs w_j

Settings

- Single server
- Dynamic policy with preemption

Objective function

Minimizing the expected abandonment costs : $C = E[\sum_{i=1}^{n}(w_j U_j)]$ with

$$U_j = \begin{cases} 1 & \text{if job } j \text{ is late} \\ 0 & \text{if job } j \text{ is on time} \end{cases}$$

Optimal strict priority rule

Theorem

If jobs can be ordered such that

- $\mu_1 \geq \mu_2 \cdots \geq \mu_n$,
- $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_n$,
- $w_1\gamma_1 \geq w_2\gamma_2 \geq \cdots \geq w_n\gamma_n$,

then it is optimal to give priority to jobs of smallest index

- Generalizes [DKL11]
- Implies the index-rule of [AGS10]

Sketch of the proof (outline)

Progressive generalization

- **Static** priority rule
  - from 2 to n jobs
- **Dynamic** priority rule **without arrivals** and with(out) preemption
- **Dynamic** priority rule **with arrivals** and with preemption

Sketch of the proof (static, $n = 2$ jobs)

Objective: a **pairwise interchange argument** to find a strict priority rule with $n = 2$ jobs

Property 1

Costs improved if $\mu_1 \geq \mu_2$, $\gamma_1 \leq \gamma_2$ and $w_1\gamma_1 \geq w_2\gamma_2$

Sketch of the proof (static, $n = 2$ jobs)

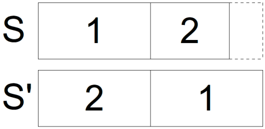
Objective: a **pairwise interchange argument** to find a strict priority rule with $n = 2$ jobs

Property 1

Costs improved if $\mu_1 \geq \mu_2$, $\gamma_1 \leq \gamma_2$ and $w_1\gamma_1 \geq w_2\gamma_2$

The issue of abandonments

- Swapping 2 jobs can delay the process of next jobs
- Conditions improving costs **and** processing time



Sketch of the proof (static, $n = 2$ jobs)

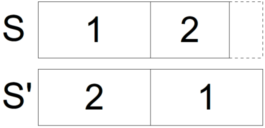
Objective: a **pairwise interchange argument** to find a strict priority rule with $n = 2$ jobs

Property 1

Costs improved if $\mu_1 \geq \mu_2$, $\gamma_1 \leq \gamma_2$ and $w_1\gamma_1 \geq w_2\gamma_2$

The issue of abandonments

- Swapping 2 jobs can delay the process of next jobs
- Conditions improving costs **and** processing time



Property 2

Processing times minimized if $\mu_1 \geq \mu_2$ and $\gamma_1 \leq \gamma_2$

Extensions and Blocking points

1. Same theorem goes for impatience to the **beginning of service**
2. From n jobs to an infinite number of jobs
   - From expected cost to average/discounted cost ?
   - Example: Poisson arrival processes, renewal processes . . .
   - **Is there a method ?**
3. Long run discounted cost ?
4. Has the MDP formulation a chance to work out ?

Abandonment costs

A cost w_j is payed for each class-j job abandonment (with rate γ_j)

Holding costs

A cost h_j is payed per unit of time for each class-j job waiting in the queue

Abandonment costs

A cost w_j is payed for each class-j job abandonment (with rate γ_j)

Holding costs

A cost h_j is payed per unit of time for each class-j job waiting in the queue

Assumptions

- Arbitrary number of jobs
- Arbitrary arrivals
- Arbitrary processing times
- **Exponential due dates** $D_j \sim exp(\gamma_j)$
- Objective: minimizing the expected costs

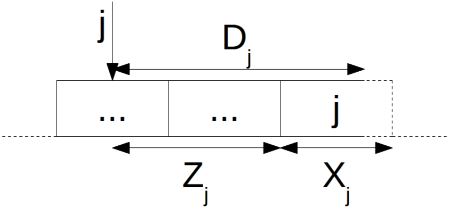
Theorem

If $h_j = w_j \gamma_j$ for all j, the two models are equivalent

Sketch of the proof

Lemma

If $D \sim exp(\gamma)$, then
$E(\min(X, D)) = 1/\gamma \mathbb{P}(X \geq D)$

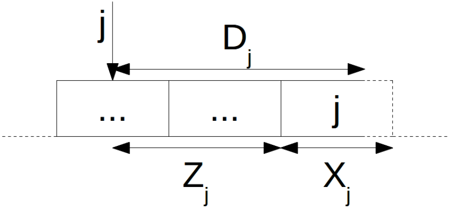


Abandonment costs for job j	Holding costs for job j
$w_j \mathbb{P}(Z_j + X_j \geq D_j)$	$h_j E(\min(Z_j + X_j, D_j))$

Sketch of the proof

Lemma

If $D \sim exp(\gamma)$, then
$E(\min(X, D)) = 1/\gamma \mathbb{P}(X \geq D)$

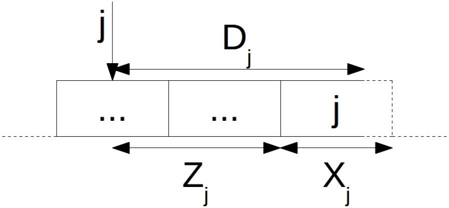


Abandonment costs for job j	Holding costs for job j
$w_j \mathbb{P}(Z_j + X_j \geq D_j)$	$h_j E(\min(Z_j + X_j, D_j))$
$w_j \mathbb{P}(Y \geq D_j)$	$h_j E(\min(Y, D_j))$

Sketch of the proof

Lemma

If $D \sim exp(\gamma)$, then
$E(\min(X, D)) = 1/\gamma \mathbb{P}(X \geq D)$



Abandonment costs for job j		Holding costs for job j
$w_j \mathbb{P}(Z_j + X_j \geq D_j)$		$h_j E(\min(Z_j + X_j, D_j))$
$w_j \mathbb{P}(Y \geq D_j)$		$h_j E(\min(Y, D_j))$
$w_j \mathbb{P}(Y \geq D_j)$	$=$	$h_j/\gamma_j \mathbb{P}(Y \geq D_j)$

$$\text{if } h_j = w_j \gamma_j$$

Conclusion and future research

- Optimal priority rule almost generalizes the results of the literature
  - From expected cost to average/discounted cost ?
  - Numerical study:
    - Which of the three conditions is the most important ?
    - To be compared with the index policy of [AGS10]
- Equivalence of costs models
  - Impatience to the beginning of service ?
  - What happens with a discount factor ?

R. Atar, C. Giat, and N. Shimkin, *The $c\mu/\theta$ rule for many-server queues with abandonment*, Operations Research **58** (2010), 1427–1439 (English).

D.G. Down, G. Koole, and M.E. Lewis, *Dynamic control of a single-server system with abandonments*, Queueing Systems **67** (2011), 63–90.