Optimality of a Priority Policy for a Server Scheduling Problem with a Deteriorating Server

Jefferson Huang

School of Operations Research and Information Engineering

Cornell University

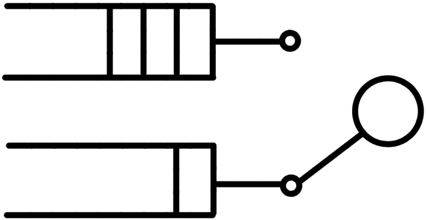
October 22, 2017

INFORMS Annual Meeting

Houston, TX

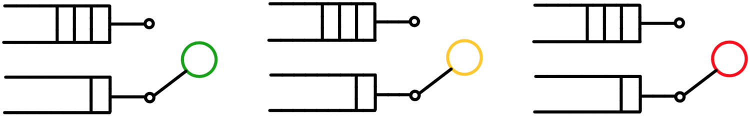
Based on joint work with **Douglas G. Down** (McMaster), **Mark E. Lewis** (Cornell), and **Cheng-Hung Wu** (National Taiwan University)

Server Scheduling: Classic Setting



- Optimality of $c\mu$-rule (Buyukkoc, Varaiya, Walrand 1985), (Nain 1989), (Van Mieghem, 1995)
- Applications to scheduling jobs/customers in production & service systems

Server Scheduling: Our Setting



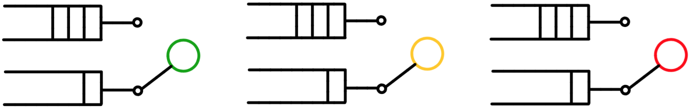
Service rate depends on server state, which can be controlled. (Kaufman, Lewis 2007), (Cai, Hasenbein, Kutanoglu, Liao 2013)

Motivation: scheduling chip testing in semiconductor manufacturing

Question: What is the structure of optimal policies?

- Optimality of $c\mu$-rule?
- Optimality of threshold-type maintenance policies?

The Model



- Independent Poisson arrivals at rates λ_1, λ_2.
- Each job has exponential service requirement with rate 1.
- Server deteriorates according to a *pure-death process* on $\mathcal{S} = \{0, 1, \ldots, B\}$.

$$\begin{aligned} 0 &= \text{server is down and being maintained} \\ B &= \text{server is (like-)new} \\ \mu_k^s &= \text{service rate for class } k \text{ jobs, when server state is } s \end{aligned}$$

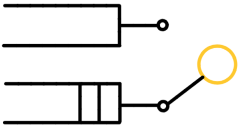
- **Costs:** accrued continuously
  - linear class-dependent holding costs with rates c_1, c_2
  - fixed state-dependent maintenance costs $K(s) > 0$, $s \in \mathcal{S}$.

Policies & Optimality Criteria

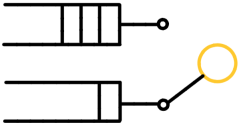
Decision Epochs: arrivals, service completions, server state changes

Service Assumptions: non-anticipative, *non-idling*, preemptive

Decisions: perform service, maintain, or idle; e.g.,



Serve Class 2, or Maintain



Serve Class 1 or 2, or Maintain

Optimality Criteria: total discounted cost, average cost per unit time

> A policy is discounted-cost (resp. average-cost) **optimal** if it achieves the minimal discounted (resp. average) cost for every initial pair of queue lengths and initial server state.

A Key Assumption

Assumption (Constant-Ratio)

For $k \in \{1, 2\}$ and $s \in \mathcal{S}$, let μ_k^s be the service rate at which class k jobs can be served when the server state is s. Then

$$\mu_1^{s-1}\mu_2^s = \mu_1^s\mu_2^{s-1} \qquad \text{for} \quad s = 1, \dots, B.$$

Implication: Either $c_1\mu_1^s \geqslant c_2\mu_2^s \ \forall s \in \mathcal{S}$, or $c_1\mu_1^s < c_2\mu_2^s \ \forall s \in \mathcal{S}$.

(State-Dependent) **$c\mu$-rule:**

If the current server state is $s \in \mathcal{S}$, prioritize class

$$k^* \in \arg\max_k \{c_k\mu_k^s\}.$$

Pure Scheduling Under Deterioration

Theorem (H, Down, Lewis, Wu 2017)

Suppose

- *the **Constant-Ratio** assumption holds, and*
- *the decision-maker has no control over the server state.*

Then the $c\mu$-rule is both discounted-cost and average-cost optimal.

- Proof: adaptation of a classic interchange argument (Nain 1989)
- Holds under more general arrival and server-state processes.
- Fails if the **Constant-Ratio** assumption does not hold.
  - (H, Down, Lewis, Wu 2017) $c\mu$-rule may be unstable, even when a stable policy exists

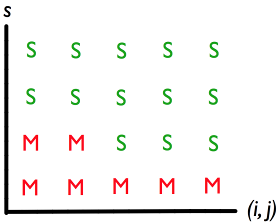
Joint Scheduling & Maintenance

Stationary Policy: At every decision epoch, perform action

$$f(i, j, s) \in \{\text{serve class 1}, \text{serve class 2}, \text{maintain}, \text{idle}\}$$

if there are currently i class 1 jobs, j class 2 jobs, and the server state is s.

A stationary policy is **monotone** in s if it has an associated "switching curve" that is monotonic in the server state s.



(Kaufman, Lewis 2007) provide an example where the optimal switching curve is non-monotone in the number of jobs. (1 class only)

Joint Scheduling & Maintenance

Theorem (H, Down, Lewis, Wu 2017)

There exists a discounted-cost optimal stationary policy with the following properties.

(i) *If it is optimal to perform service at the current decision epoch, and both queues are nonempty, then it is optimal to schedule according to the $c\mu$-rule.*

(ii) *If the maintenance time and maintenance costs $K(s)$ satisfy certain conditions, then the aforementioned policy can be taken to be monotone in s.*

If certain stability conditions hold, then there exists an average-cost optimal stationary policy with the preceding properties.

- Proof: (i) use result on pure scheduling; (ii) dynamic programming, "monotonicity" of discounted-cost value function

Conclusions & Future Work

Contributions:

1. A sufficient condition (**Constant-Ratio**) under which the $c\mu$-rule is optimal for scheduling, when the server deteriorates.
   - If **Constant-Ratio** does not hold, the $c\mu$ rule may not be optimal (or even stable, despite the presence of a stable policy).
2. Extension of results in (Kaufman, Lewis 2007) to two job classes.

Future Work:

Preliminary numerical results indicate that scheduling according to the $c\mu$-rule performs well even when the **Constant-Ratio** assumption does not hold.

- How does the degree of deviation from **Constant-Ratio** affect the optimality of the $c\mu$-rule?