Optimal Surplus Capacity Utilization in Polling Systems via Fluid Models

Ayush Rawal[1]
email: ayush.rawal@iitb.ac.in

Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

WiOpt, 2014
Hammamet, Tunisia

[1] Co-authors: Veeraruna Kavitha and Manu K. Gupta

Outline

1. Introduction
2. Steady State Fluid Model
3. Achievable Region
4. Optimization
5. Conclusions

Polling Systems: Definition and Applications

- Class of queuing system.
- Single server visits a set of queues.
- Non zero time to walk/switch.
- Non work conserving.
- Application areas:
    - Communication Systems,
    - Production Systems,
    - Traffic and Transportation Systems, etc.

What is 'Differential Fairness'?

Providing special treatment to the users/processes, depending upon the requirements.

- Applications of differential fairness:
  - Application driven,
  - Price driven,
  - Market driven.
- Two scenarios:
  - Underprivileged users
  - High demand users
- Two customer classes, one class demands certain Quality of Service (QoS).

Problem Description and Approach

- Primary customers served by a server.
  - System stable $\Longrightarrow$Surplus capacity.
  - Can we get extra revenue? Secondary customers.
- Goal: Allocate resources to secondary customers $\ni$QoS requirements of primary not compromised.
- Can delay priority schedulers help?
  - Achievable Region, $\mathscr{A}$.
  - Complete class of schedulers.
- Optimize performance over $\mathscr{A}$, we propose two constrained optimization problem:
  - Admission control, controlling λ_2.
  - Limited buffer, with loss.

System Description

- Single server polling system, two queues Q_1 and Q_2.
- Infinite buffer capacity.
- Arrival rates $\lambda_i' s$ and Service rates $\mu_i' s$, where $i \in \{1, 2\}$.
- Switching times, IID with mean s.
- FCFS, Non-preemptive queuing discipline.
- Customers leave system only when service is completed.
- Time invariant, state dependent scheduling policies.

β-priority/ Exhaustive Switching Policy

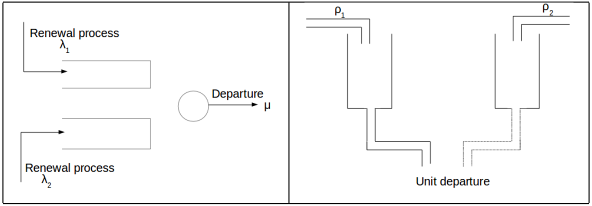
Let $\tilde{w}_i$ and $\tilde{w}_{-i}$ be the waiting time of longest waiting time of customer in Q_i and Q_{-i} respectively.
When in Q_i, the switching rule $\beta = (\beta_1, \beta_2)$ implies the following:

1. If $\beta_i = ex$, switch from Q_i , when Q_i is empty.
2. Else switch from Q_i , when $\beta_i \tilde{w}_i \leq \tilde{w}_{-i}$.

From Discrete to Fluid Model

- Discrete System $(\lambda_1,\lambda_2,\mu) \xrightarrow{\mu\to\infty}$ Fluid System $(\lambda_1,\lambda_2,\mu)$
- Waiting time performance Fluid System $(\lambda_1,\lambda_2,\mu)$

  = Waiting time performance of Fluid System $(\rho_1,\rho_2,1)$, when $\rho_i = \dfrac{\lambda_i}{\mu}$



(a) Discrete System	(b) Fluid Model

Fig. 1: From Discrete System to Fluid Model

Steady State Fluid Model

- SSFM consists of two storage tanks, two inlets and one outlet pipe.
- Fluid flows from inlets to the corresponding tanks at a constant rate.
- Outlet pipe switches between the tanks.
- Consider the following notational analogy.
  - **Switching time**: s, time required to move outlet pipe from one storage tank to another.
  - **Service rate**: μ_1 and μ_2, rate of outflow from tank 1 and tank 2 respectively.
  - **Arrival rate**: λ_1 and λ_2, rate of inflow in tank 1 and tank 2 respectively.
  - **Switching policy parameters**: γ_1 and γ_2 are delay priority parameters based on height of fluid/number of waiting customers.

Switching Cycle

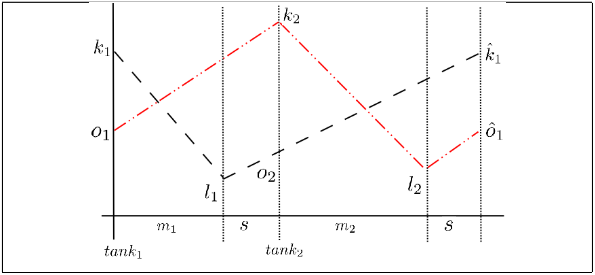


Fig. 2: Fluid level of deterministic system in one cycle

Waiting Time

Average waiting time of fluid in tank 1

$$\bar{w}_1 = \frac{1}{\lambda_1}\left(\frac{c_1+\gamma_2 c_2}{\gamma_1\gamma_2-1}\right)+\varpi_1, \text{ with } \varpi_1 := \frac{s(1-\rho_1)}{1-\rho},$$

$$c_1 = \frac{s\lambda_1(1-\rho_1+\rho_2)}{1-\rho} \text{ and } c_2 = \frac{s\lambda_2(1+\rho_1-\rho_2)}{1-\rho}.$$

Average waiting time of fluid in tank 2

$$\bar{w}_2 = \frac{1}{\lambda_2}\left(\frac{c_2+\gamma_1 c_1}{\gamma_1\gamma_2-1}\right)+\varpi_2, \text{ with } \varpi_2 := \frac{s(1-\rho_2)}{1-\rho}.$$

Theorem (Stability)

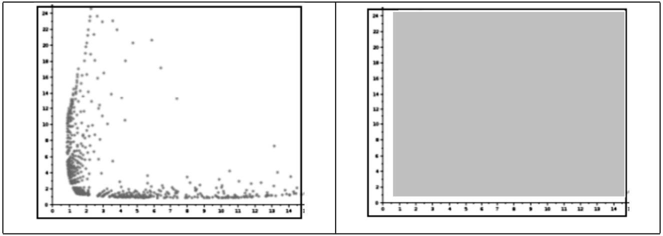
β priority schedulers are stable if:

$(i)\ \rho < 1$

$(ii)\ \gamma_1 \times \gamma_2 > 1$

Achievable Region

Set of all possible performance vectors obtained by all possible combinations of (β_1, β_2) for a set of switching times, arrival and departure rates.



(a) Via Simulation for Discrete system | (b) Computed theoretically for SSFM

Fig. 3: Achievable Region of performance ($\lambda_i = 4.5, \mu = 10, s = 0.1$)

Theorem (Achievable Region)

Achievable region of performance for SSFM is given by:

$$\mathscr{A} = \{[\varpi_1, \infty) \times [\varpi_2, \infty)\}, \text{ where } \varpi_i = \frac{s(1-\rho_i)}{1-\rho},$$

and $\{Exhaustive \cup \beta - priority\} \leftarrow$ complete class.

β versus γ Schedulers

Result: β and γ relationship

β and γ schedulers, achieve same performance when:

$$\frac{\gamma_1}{\beta_1} = \frac{\beta_2}{\gamma_2} = \frac{\lambda_2}{\lambda_1}.$$

			Simulation		SSFM	
μ	β_1	β_2	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_1$	$\bar{w}_2$
100	2	4	0.1848	0.1446	0.2245	0.1775
200	2	4	0.2048	0.1615	0.2245	0.1775
500	2	4	0.2169	0.1712	0.2245	0.1775
100	7	17	0.1420	0.1167	0.1484	0.1246
200	7	17	0.1464	0.1228	0.1484	0.1246
500	7	17	0.1478	0.1242	0.1484	0.1246

TABLE I: Performance of Random System and SSFM ($\lambda_1/\lambda_2 = 0.5, \rho = 0.3$)

Optimizing Surplus Capacity

Two constrained optimization problem to optimally utilize surplus capacity.

1. Admission Control (controlling λ_2):

$$P1: \max_{\lambda_2, \beta \in \mathscr{B}^P} \lambda_2 P(\bar{w}_2) \qquad \text{Subject to: } \bar{w}_1 \leq \eta_1$$

where, $P(\bar{w}_2)$: price paid by secondary customer, monotonically decreasing function in $\bar{w}_2$.

2. Limited Buffer with loss:

$$P2: \max_{B, \beta \in \mathscr{B}^P} \lambda_2 (1-f) P(\bar{w}_2) \qquad \text{Subject to: } \bar{w}_1 \leq \eta_1$$

where, f is fraction of customer lost due to limited buffer capacity,B.

Theorem (Optimization)

Both $P1$ and $P2$, are optimized by exhaustive schedulers, $\beta = (ex, ex)$. To determine optimal λ_2 or B, price function is required.

Contribution and Conclusions

- Idea of **differential fairness** is investigated in case of polling systems.
- We proposed and proved that a class of delay priority schedulers along with exhaustive policy forms the **complete achievable region**.
- Used **Monte Carlo simulations** to show that performance of random systems, converges to that of analyzed limit system with fluid queues.
- We obtained achievable region for **fluid system**.
- We conclude that:
  - achievable region of performance is **unbounded**, and
  - **exhaustive** service discipline is optimal from individual queue perspective.
- Future work: We are trying to prove **convergence** of performance of discrete system to that of SSFM.

Thank you