On 2-moment Completeness of Some Scheduling Policies in Single Class Queues

by

Manu K. Gupta

Joint work with Prof. N. Hemachandra

May 26, 2015

Outline

Introduction

- Single class queueing systems
  - Customers arrive over time and get served.
  - Applications in wireless and computer communications, transportation and job shop manufacturing systems.
- Scheduling discipline and performance measures
  - Set of all non anticipative and work conserving scheduling policy.
  - Mean waiting time is constant in M/M/1 queue.
    - Additional non pre-emptive scheduling assumption is needed for M/G/1 queue.
  - Second moment of waiting time changes with scheduling scheme (FCFS, LCFS etc.).

Problem Description

- Achievable region and completeness for mean waiting time
- Nice geometric structure for mean waiting time driven by Kleinrock's conservation law [1].
- A parametrized policy is mean waiting time *complete* if it sweeps the entire achievable region.
- Some mean waiting time *complete* policies do exist [5].
- Useful tool in solving optimal control problem (see [6], [12]).

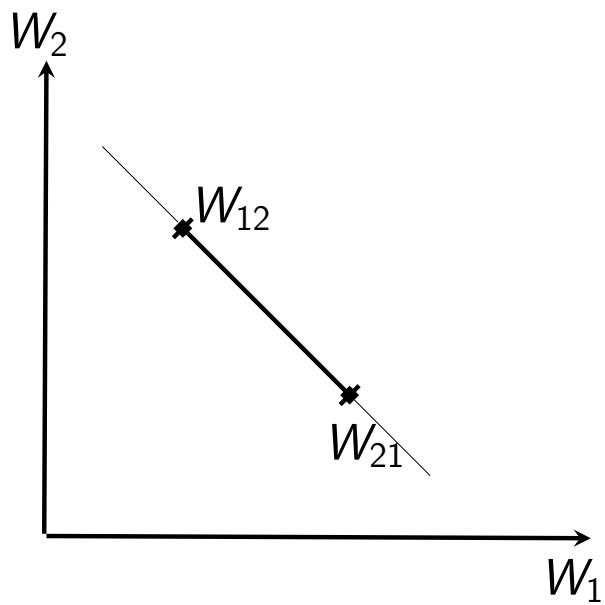


Figure : Achievable region for mean waiting time

Problem Description

Purpose of the talk

To explore the notion of 2-moment completeness for second moment (equivalently, variance) of waiting time.

- One-to-one correspondence between the range of the parameter and all possible second moments of waiting times under non pre-emptive, non anticipative, work conserving scheduling discipline.
- Such polices are good enough for optimization purposes.

2-moment completeness

- Waiting time variance (or second moment) is minimum with FCFS and maximum with LCFS (see [7], [3]).
- Let l and u be the second moment of waiting time associated with FCFS and LCFS.
- The achievable region for second moment of waiting time is the interval $[l, u]$.
- Let $p \in \mathbb{I} \subset \mathbb{R}$ and say class of these policies are denoted by $\{\mathcal{F}\}_{p \in \mathbb{I}}$.

2-moment completeness

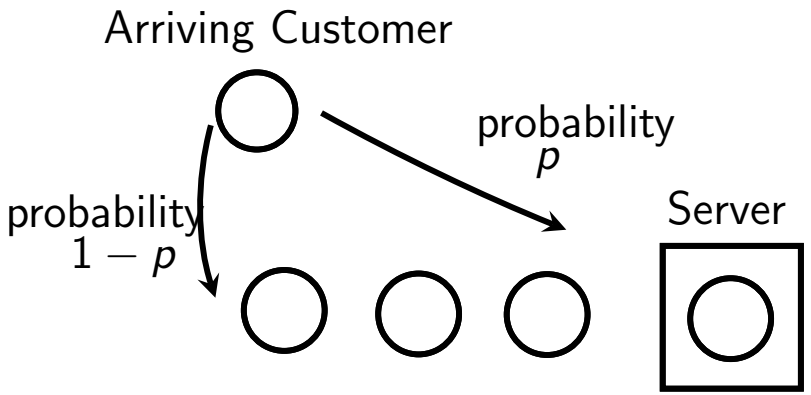
A set of parametrized queue discipline policies $\{\mathcal{F}\}_{p\in\mathbb{I}}$ is called non pre-emptive, non anticipative, work conserving **2-moment complete** if these set of policies satisfy the following conditions.

- Service is non pre-emptive.
- Customers are selected for service in a manner that is independent of their subsequent service time.
- If the service mechanism is ready to receive (serve) a customer at a time when the queue is non empty, then one of the customers present will be immediately served.
- There exists a one-one mapping $V_{\mathcal{F}}(p) : \mathbb{I} \to [l, u]$.

Useful in solving optimal control problem

Impolite Customer class for M/G/1 queue

- Introduced by Suleyman Ozekici, Jingwen Li, and Fee Seng Chou [9].
- An arriving customer joins the front of the queue with probability p and joins in the end of queue with probability $(1 - p)$ and $p \in [0, 1]$.



Theorem 1

Impolite customer class is 2-moment complete.

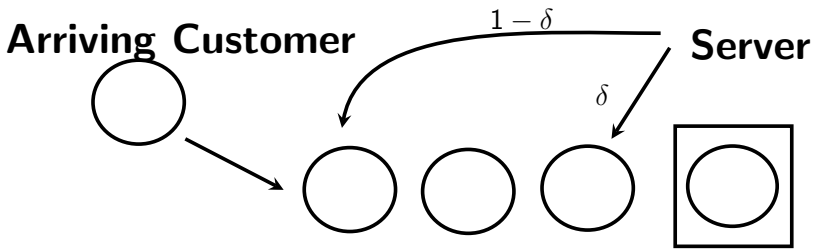
Outline of Proof

$$E(W^2)|_{\text{imp}} = \frac{1}{1-p\rho}\left(\frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2 (E(S^2))^2}{2(1-\rho)^2}\right) \qquad (1)$$

- Let $\mathbb{I} = [0, 1]$ and $\{\mathcal{F}\}_I$ be the impolite parametrized class of policies.
- $p = 0$ and $p = 1$ correspond to FCFS and LCFS service disciplines.
- End points of the achievable region $[l, u]$ are realized.
- $E(W^2)|_{\text{imp}}$ is proportional to reciprocal of an affine function of p.
- $E(W^2)|_{\text{imp}}$ will have one to one mapping from $\mathbb{I} \rightarrow [l, u]$.

A parametrized M/M/1 queue discipline

- Proposed by Vera Dufkova and Frantisek Zitek [2].



Theorem 2

Queue discipline parametrization proposed in [2] is 2-moment complete.

$$E(W^2)|_\delta = \frac{2\lambda}{(\mu-\lambda)^2(\mu-\lambda+\delta\lambda)} \tag{2}$$

$\delta = 1$ and 0 achieves FCFS and LCFS respectively.

Random Order of Service (ROS)

Service Mechanism

If there are $n \geq 1$ customers is the queue, each customer will have equal probability $\frac{1}{n}$ of getting served.

Second moment for M/M/1 queue is (see [11]):

$$E(W^2)_{ROS} = \frac{1}{(\mu-\lambda)^2}\left(\frac{4\rho}{2-\rho}\right) \tag{3}$$

$$E(W^2)|_{\delta} = \frac{1}{(\mu-\lambda)^2}\left(\frac{2\lambda}{\mu-\lambda+\delta\lambda}\right) \tag{4}$$

- On equating above two equations, we get $\delta|_{ROS} = 1/2$.
- ROS queue discipline achieves a single point in interval $[l, u]$ corresponding to $\delta = 1/2$.

Random Insertion (RI)

Introduced by Steven W. Fuhrmann and Ilias Iliadis [4].

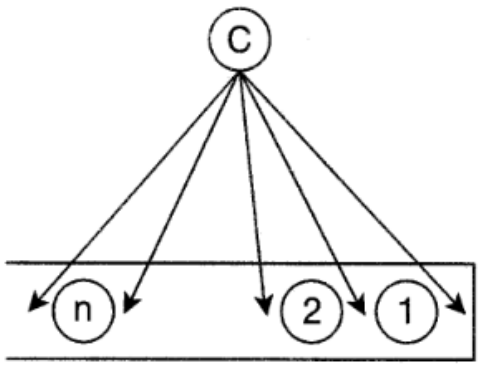


Figure : Random Insertion Mechanism [4]

- RI and ROS has same waiting time distribution (see [4]).
- RI queue discipline will also achieve a single point in interval $[l, u]$ corresponding to $\delta|_{RI} = 1/2$.

RI is 2-moment incomplete policy.

Random Assigned Priority (RAP)

- Independently assign a random value from uniform distribution in $[0, 1]$.
- Serve according to non pre-emptive priority based on their assigned values.
- Smaller values have priority over larger values.

Second moment of waiting time is given by (see [4]):

$$E(W^2)_{RAP} = \frac{\rho(1-\rho)(2-\rho)E(S)E(S^3) + \rho^2(3-\rho)[E(S^2)]^2}{6(1-\rho)^3[E(S)]^2}$$

On equating with parametrized queue for M/M/1 queue,

$$\delta|_{RAP} = \frac{(\mu-\lambda)(3\mu-\lambda)}{3\mu(2\mu-\lambda)+\lambda^2} = \frac{(1-\rho)(3-\rho)}{3(2-\rho)+\rho^2} \qquad (5)$$

RAP contd ..

- Can be easily verified from the stability of queue ($\rho < 1$) that $0 < \delta|_{RAP} < 1/2$.
- $E(W^2)_{RAP} > E(W^2)_{ROS}$
- $\rho \to 1 \Rightarrow \delta|_{RAP} \to 0$
  - In heavy traffic, RAP $\Leftrightarrow$ LCFS
- $\rho \to 0 \Rightarrow \delta|_{RAP} \to 1/2$
  - In low traffic, RAP $\Leftrightarrow$ ROS
- Optimization over RAP policies will result in suboptimal solution.

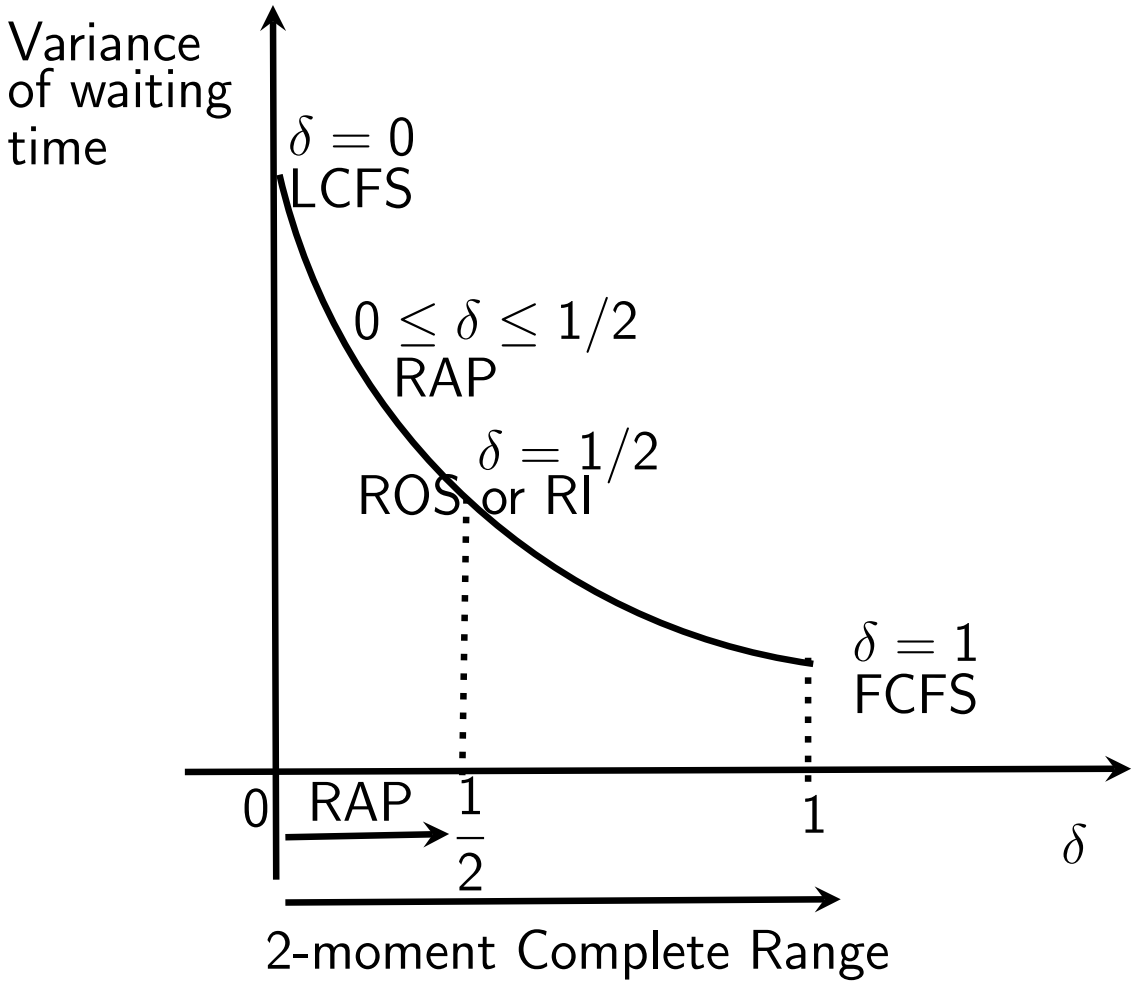


Figure : Variance of waiting time for various queue disciplines vs parameter δ.

Two Level Priority

- Arriving customers are divided in higher and lower priority class with probability p and $(1-p)$.
- Strict static priority to higher class.
- Queue discipline is FCFS within a class.
- Class discrimination is just a way of scheduling customers.

Second moment of above system results in (by conditioning):

$$E(W^2)|_p = \frac{2\lambda p}{\mu(\mu-\lambda p)^2} + \frac{2\lambda(\mu^2-\lambda^2 p)(1-p)}{(\mu-\lambda)^2(\mu-\lambda p)^3} \quad (6)$$

- $p = 1$ and $p = 0$ implies $\delta = 1$ (FCFS).
- Two level priority can never achieve LCFS.

- Some pre-emptive anticipative work conserving queue disciplines.
  - Variance is beyond the 2-moment complete range
- Processor Sharing (PS)
- Pre-emptive Last In First Out (PLIFO)
- Longest Remaining Processing Time (LRPT)

So the conditions in definition of 2-moment completeness are indeed necessary.

- Non pre-emptive, non anticipative and work conserving scheduling policy.

Processor Sharing

- Pre-emptive scheduling discipline.
- Conditional variance of waiting time for M/M/1 queue was derived in [14].

Unconditional variance turn out to be

$$Var(T)_{PS} = \frac{1}{\mu^2(1-\rho)^2}\frac{2+\rho}{2-\rho} \quad (7)$$

$$Var(T)_{\delta} = \frac{1}{(\mu-\lambda)^2}\left[\frac{2\lambda}{\mu-\lambda+\delta\lambda} - \rho^2\right] + \frac{1}{\mu^2} \quad (8)$$

On equating, $\delta = 1 - 1/\rho(3-\rho)$

$\rho \to 1 \Rightarrow \delta \to 1/2$. Hence processor sharing behaves like ROS in high traffic.

Processor Sharing Contd ..

$\delta \geq 0$ results in quadratic $\rho^2 - 3\rho + 1 \leq 0$.

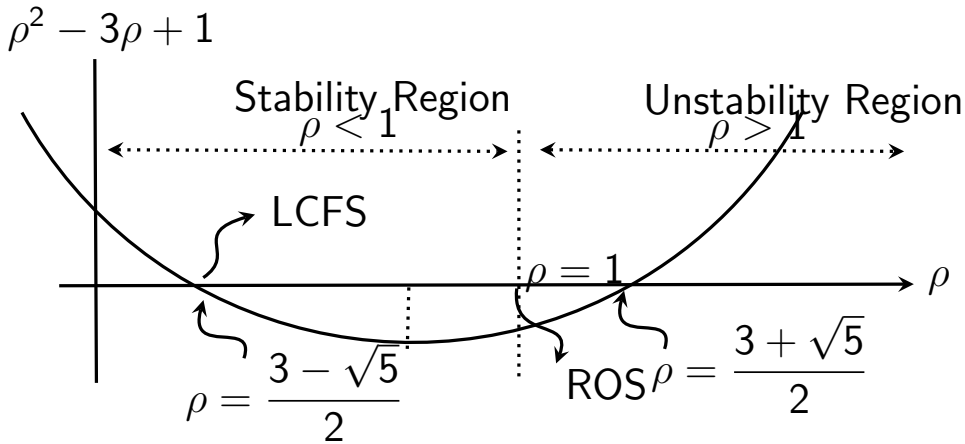


Figure : Change in sign of quadratic $\rho^2 - 3\rho + 1$ w.r.t. ρ

- For $\rho \in (0, \frac{3-\sqrt{5}}{2})$, there is no δ in range $[0, 1]$.
- Variance of PS is beyond 2-moment complete parametrized queue discipline range.

Pre-emptive Last In First Out (PLIFO)

- Server is always working on most recent arrival to the system.

Variance of waiting time for PLIFO is given by [13]:

$$Var(T)|_{PLIFO} = \frac{Var(S)}{(1-\rho)^3} + \lambda \left(\frac{E(S)}{1-\rho} \right)^3 . \qquad (9)$$

On simplifying the expressions for exponential service, we have

$$Var(T)_{PLIFO} - Var(T)_{LCFS} = \frac{2\mu\lambda}{\mu^2(\mu-\lambda)^2} > 0 \qquad (10)$$

Beyond 2-moment complete range for any load factor.

Longest Remaining Processing Time (LRPT)

- Anticipative scheduling discipline.
- Longest remaining size is given pre-emptive priority.
- No job can finish service before the end of a busy period.
- LRPT finishes every job at the last moment possible.

On simplifying the expression for variance, we have

$$Var(T)_{LRPT} - Var(T)_{LCFS} = \frac{2\lambda^3 + \mu^2\lambda + 7\mu\lambda(\mu - \lambda)}{\mu(\mu - \lambda)^4} > 0 \tag{11}$$

Beyond 2-moment complete range for any load factor.

- PLIFO and LRPT have variance beyond 2-moment complete range for any ρ.
- PS is beyond 2-moment complete range for $\rho \in (0, \frac{3-\sqrt{5}}{2})$.

Remark

Variance of waiting time can be beyond 2-moment complete range if scheduling policy violates any of the conditions on queue discipline.

Some Applications

Variance minimization problem with lower bound on variance

P1: $\min_{\mathcal{F}} \quad Var(W)$

Subject to

$$Var(W) \geq \gamma$$

T1: $\min_{0 \leq \delta \leq 1} \quad Var(W)$

Subject to

$$Var(W) \geq \gamma$$

- $\mathcal{F}$ is the set of all non pre-emptive, non anticipative and work conserving scheduling policies for $M/M/1$ queue.
- P1 and T1 are equivalent as parametrized queue discipline is 2-moment complete.
- Problem T1 is easy to solve.

Solution depends on γ.

- If $\gamma > Var(W)|_{LCFS}$
  - Infeasible solution.
- If $\gamma < Var(W)|_{FCFS}$
  - Trivial solution FCFS scheduling.
- If $\gamma \in (Var(W)|_{FCFS}, Var(W)|_{LCFS})$,
  - Pure dynamic optimal scheduling.

Weighted Cost Minimization Problem

$$\textbf{P2:} \quad \min_{\mathcal{F}} \quad c_1 Var(W) + c_2 f(W)$$

- $f(W)$ represents unfairness of a job and unfairness is quantified according to [10].
- Fairness index suggests: FCFS $>$ ROS $>$ LCFS

FCFS is the optimal scheduling policy

Motivated from Markovitz mean-variance model (see [8]).

P3: $\min_{\mathcal{F}'} \quad E(W^3)$

Subject to

$$E(W^2) \leq \beta$$

T3: $\min_{0 \leq \delta \leq 1} \quad E(W^3)$

Subject to

$$E(W^2) \leq \beta$$

Solution of P3

- FCFS as optimal scheduling policy.
- if $\beta_1 \leq E(W^2) \leq \beta_2$, pure dynamic scheduling will be optimal.

Discussion

- Introduced the notion of 2-moment completeness.
- Some parametrized 2-moment complete and incomplete policies.
- Some applications.
- Relax the assumption on non pre-emptive priority.
- Extend these ideas to multi-class queueing systems.
- Optimal control problems in multi-class queue involving higher moments.

References I

EG Coffman Jr and I Mitrani.
A characterization of waiting time performance realizable by single-server queues.
Operations Research, 28(3-part-ii):810–821, 1980.

Vera Dufkova and Frantisek Zitek.
On a class of queue disciplines.
Aplikace Mathematiky, 20:345–357, 1974.

Patrick Eschenfeldt, Ben Gross, and Nicholas Pippenger.
A bound on the variance of the waiting time in a queueing system.
arXiv:1106.0074v1, June 2011.

Steven W. Fuhrmann and Ilias Iliadis.
A comparison of three random discipline.
Queueing Systems, 18:249–271, 1994.

References II


Manu K. Gupta, N. Hemachandra, and J. Venkateswaran.
On mean waiting time completeness and equivalence of EDD and HOL-PJ dynamic priority in 2-class M/G/1 queue.
In *8th international conference on performance methodology and tools (Valuetools)*, 2014.


Refael Hassin, Justo Puerto, and Francisco R Fernández.
The use of relative priorities in optimizing the performance of a queueing system.
European Journal of Operational Research, 193(2):476–483, 2009.


J. F. C. Kingman.
The effect of queue discipline on waiting time variance.
Mathematical Proceedings of the Cambridge Philosophical Society, 58:163–164, 1962.


Harry M Markowitz.
Portfolio selection: efficient diversification of investments, volume 16.
Yale university press, 1968.

References III


Suleyman Ozekici, Jingwen Li, and Fee Seng Chou.
Waiting time in M/G/1 queues with impolite arrival disciplines.
Probability in the Engineering and Informational Sciences, 9:255–267, 1995.


David Raz, Hanoch Levy, and Benjamin Avi-Itzhak.
RAQFM: A resource allocation queueing fairness measure.
In *Proceedings of ACM SIGMETRICS Conference, New York, NY*, 2004.


Michel Scholl and Leonard Kleinrock.
On the M/G/1 queue with rest periods and certain service-independent queueing disciplines.
Operations Research, 31:705–719, 1983.


S. K. Sinha, N. Rangaraj, and N. Hemachandra.
Pricing surplus server capacity for mean waiting time sensitive customers.
European Journal of Operational Research, 205:159–171, August 2010.


Adam Wierman.
Scheduling for today's computer systems: Bridging theory and practice.
PhD thesis, Carnegie Mellon University, 2007.

References IV

S. F. Yashkov.
Processor sharing queues: Some progress in analysis.
Queueing Systems, 2:1–17, 1987.

THANK YOU!!