Investigating the Effect of Variance in Markovian Shortest Remaining Processing Time Queues

Sean Malter and Dr. Amber Puha, Department of Mathematics, California State University, San Marcos

Queues and Service Disciplines

A queue can be thought of as a waiting line, with jobs waiting to be served and jobs in service. The order in which we process the jobs in the queue is called a service discipline.

First-Come-First-Serve (FCFS)
- Jobs served one at time in order of arrival
- Oldest job gets all of the sever's effort
- Typical for serving people, e.g., at a small retail store with one cash register

Processor Sharing (PS)
- Serves jobs all jobs simultaneously
- Servers effort is divided equal among all jobs
- Idealized computer time sharing

Shortest Remaining Processing Time (SRPT)
- Job with the smallest remaining processing time served first
- Preemptive (the new job with a smaller processing time than the remaining processing time of the job in service gets priority)
- Performance Optimal: Known to minimizes queue length

Heavily Loaded M/M/1 Queue

Interarrival times and processing times are sequences of mutually independent, independent exponential random variables with respective rates λ and μ.
We investigate heavily loaded M/M/1 queues where $\mu = \lambda$.

Performance Processes

- **Queue Length**: $Q(t)$, the number of jobs in the system at time t
- **Workload**: $W(t)$, the total time needed to complete all the work in the system at time t, excluding future arrivals.

Natural Question

Can one quantify how small the queue length process is for an SRPT queue?

An Answer Under Standard Diffusion Scaling

For $n \in \mathbb{N}$ and $t \in [0, \infty)$,

$$\widehat{W}^n(t) = \frac{W(nt)}{\sqrt{n}} \quad \text{and} \quad \widehat{Q}^n(t) = \frac{Q(nt)}{\sqrt{n}}.$$

Established SRPT convergence in distribution results: as $n \to \infty$

$$\widehat{W}^n(\cdot) \Rightarrow W^*(\cdot) \quad \text{and} \quad \widehat{Q}^n(\cdot) \Rightarrow 0.$$

Here $W^*(\cdot)$ is reflected Brownian motion with variance $2/\lambda^2$.
Conlusion: The queue length process is of smaller order of magnitude than the workload process.

Next Natural Question and Suspected Answer

What is the order of magnitude of the queue length process?
We conjecture a correction factor of $\ln(\sqrt{n})$.
We investigated this through simulation.

Non-Standard Diffusion Scaling

For $n \in \mathbb{N}$ and $t \in [0, \infty)$, set

$$\widetilde{Q}^n(t) = \frac{\ln(\sqrt{n})Q(nt)}{\sqrt{n}}.$$

Suspected Behavior

There exists a positive constant C, depending on λ, such that

$$C\widehat{W}^n(\cdot) \approx \widetilde{Q}^n(\cdot), \qquad \text{as } n \to \infty.$$

We used simulations to identify a candidate for C and to explore viability.

Main Outcome

As a result of our investigation, we developed a conjecture:

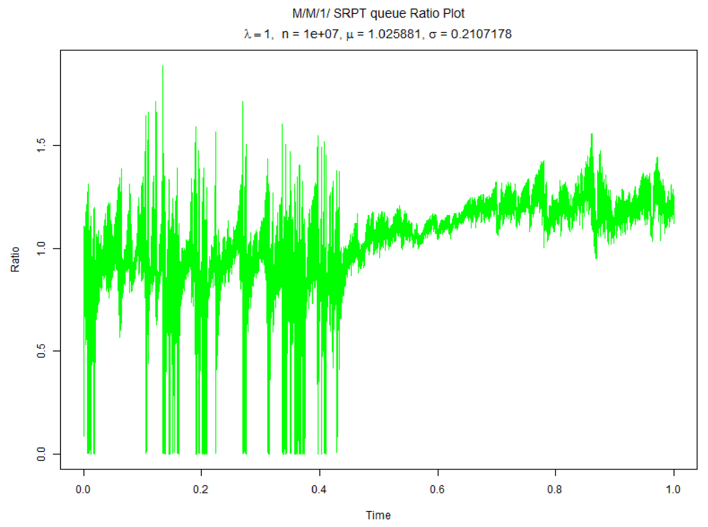
Conjecture (Hunsperger, Malter, Puha)
In an M/M/1 SRPT queue with common processing and incoming rate given by λ, the queue length process appropriately rescaled with a non-standard logarithmic growth factor converges in distribution to a reflected Brownian motion, i.e., $\widetilde{Q}^n(\cdot) \Rightarrow \lambda W^*(\cdot)$, as $n \to \infty$.

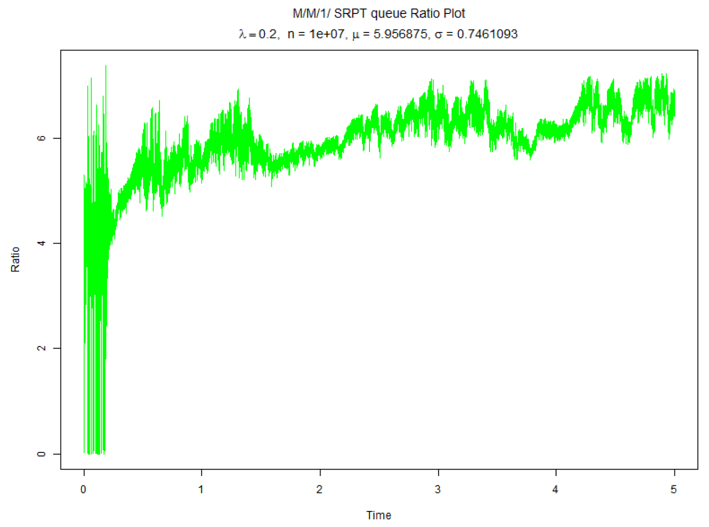
Identifying the Constant C

The following ratio is a natural estimator for $1/C$:

$$\frac{\widehat{W}^n(\cdot)}{\widetilde{Q}^n(\cdot)}$$

Plots of the ratio with $n = 10^7$, and$t \in [0, 1/\lambda]$, with values of t for which $\widetilde{Q}^n(t)(t) = 0$ omitted.





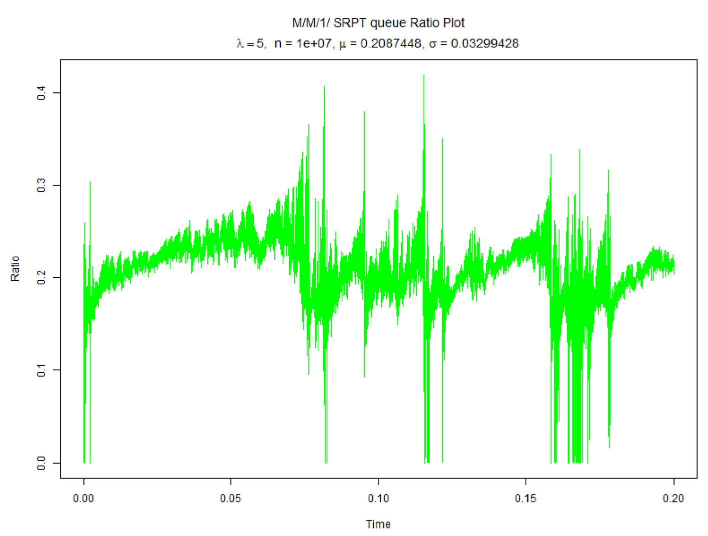


Table of $\widehat{W}^n(\cdot)/\widetilde{Q}^n(\cdot)$, with $n = 10^6$ and $t \in [0, 1/\lambda]$

λ	$1/\mu$	σ
0.1	0.105	2.584
0.5	0.505	0.363
0.7	0.658	0.248
1	0.870	0.155
2	1.972	0.092
5	5.028	0.041
13	13.024	0.014

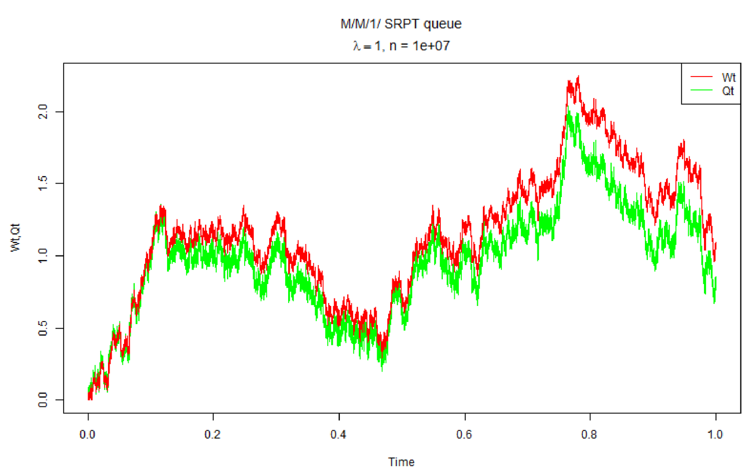
Plot Characteristics
- Relatively Flat
- Bit of randomness
- Fluctuates about the line $y = 1/\lambda$

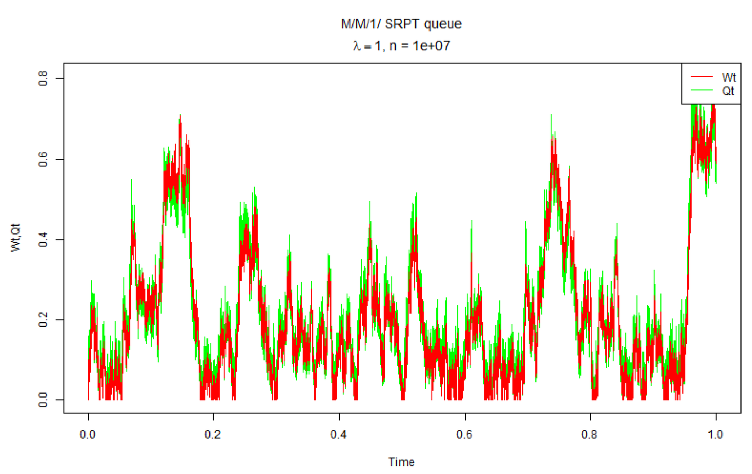
Table Charterisitics
- $1/\mu \approx \lambda$
- Large σ for $\lambda = 0.1$
- σ tends to decrease as λ increases

Prediction: $C = \lambda$

Investigating the Prediction $C = \lambda$

We now display graphics of the rescaled workload $\lambda\widehat{W}^n(t)$ and rescaled queue length $\widetilde{Q}^n(t)$ processes for $n = 10^7$ and $t \in [0, 1/\lambda]$.





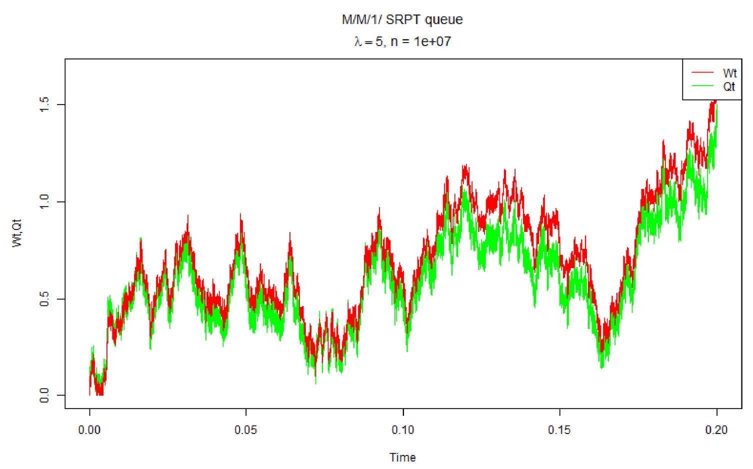
Plot Characteristics for $\lambda = 1$
- Processes are relatively close throughout
- As rescaled workload increases, it often slightly exceed rescaled queue length, as on the left
- Rescaled processes are often very close and frequently zero, as on the right
- Mimics outcome of Hunsperger and Puha '12, where only $\lambda = 1$ was considered.

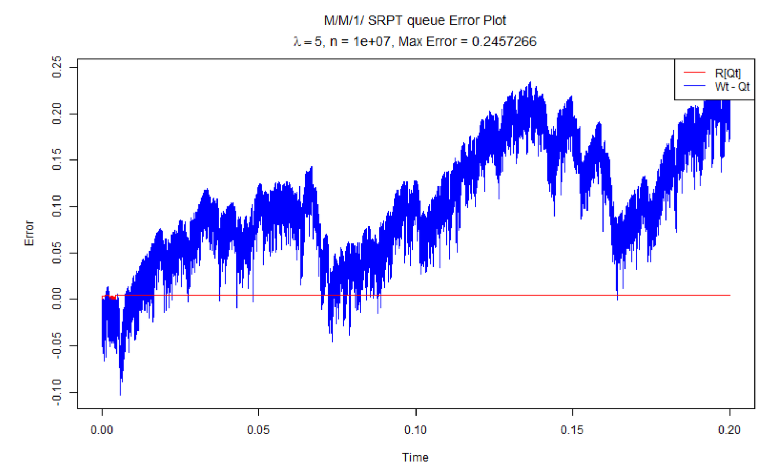
More Detailed Examination of the Error

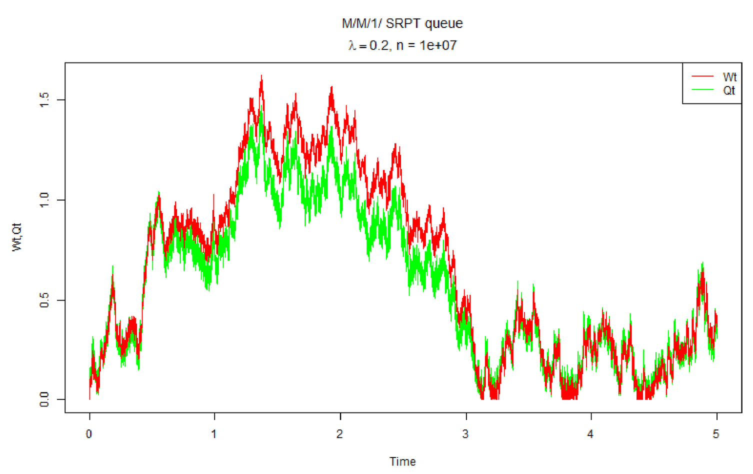
Here we demonstrate that plots for other values of λ have similar characteristics. Also, to more effectively show how close the two processes are, we plot the difference between the two processes, which we define the error,

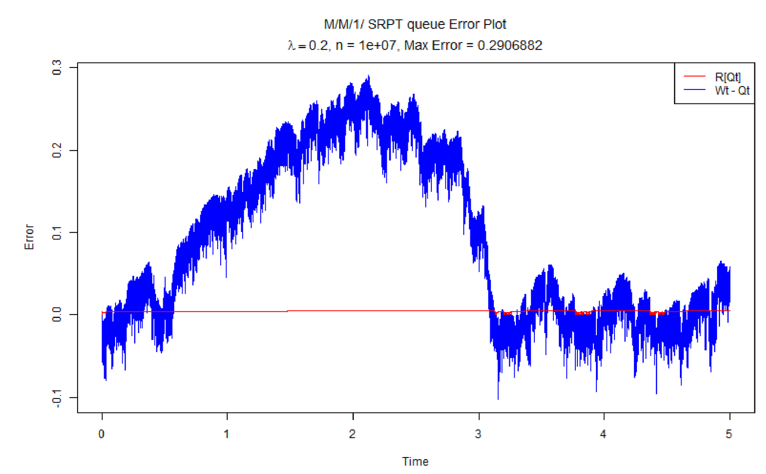
$$E^n(\cdot) = \lambda\widehat{W}^n(\cdot) - \widetilde{Q}^n(\cdot).$$

We plot $E^n(\cdot)$ in blue in a separate figure on the right, along with the largest residual service time $R[Qt]$ rescaled by $1/\sqrt{n}$ in red.









Plots of $\lambda = 5$ and $\lambda = 0.2$ with $n = 10^7$ and $t \in [0, 1/\lambda]$

- Closely related processes with varied λ
- Little separation between processes
- Supports our claim that $C = \lambda$
- Largest errors typically above the y-axis
- Error plots are similarly shaped
- Maximum errors are similar size

Large n Simulations

We created new code to handle larger values of n, in which we track the maximum error defined as

$$M = \max_{t \in [0,1]} \left| \lambda\widehat{W}^n(t) - \widetilde{Q}^n(t) \right|.$$

note we have a new time interval [0,1]

In our tables, μ is the average of M over our samples and σ is the sample standard deviation. $P(M \le x)$ denotes the probability that M is less than or equal to x

$\lambda = 1$ Unbiased Estimators

Sample Size	100	100	100	25
n	10^5	10^6	10^7	10^8
μ	0.347	0.309	0.301	0.270
σ	0.143	0.140	0.150	0.129
$P(M \le 0.1)$	0	0	0.01	0.04
$P(M \le 0.2)$	0	0.31	0.31	0.4
$P(M \le 0.3)$	0.52	0.54	0.56	0.64
$P(M \le 0.4)$	0.77	0.76	0.73	0.84
$P(M \le 0.5)$	0.88	0.88	0.85	0.96

$\lambda = 5$ Unbiased Estimators

Sample Size	100	100	100
n	10^5	10^6	10^7
μ	1.091	0.935	0.870
σ	0.465	0.412	0.435
$P(M \le .2)$	0	0	0
$P(M \le .4)$	0.05	0.05	0.09
$P(M \le .6)$	0.16	0.25	0.31
$P(M \le .8)$	0.28	0.41	0.55
$P(M \le 1)$	0.49	0.6	0.73
$P(M \le 1.2)$	0.61	0.79	0.79

$\lambda = 0.2$ Unbiased Estimators

Sample Size	100	100	100	25
n	10^5	10^6	10^7	10^8
μ	0.211	0.149	0.107	0.086
σ	0.031	0.023	0.020	0.026
$P(M \le .05)$	0	0	0	0
$P(M \le .1)$	0	0	0.39	0.8
$P(M \le .15)$	0	0.57	0.96	0.96
$P(M \le .2)$	0.43	0.97	1	1

Table Characteristics
- μ decreases as n increases
- σ seems to slowly decrease
- $P(M \le x)$ generally increases as n increases
- Supports that M tends to zero, but seems to do so more slowly for larger λ
- Additional evidence supporting "Suspected Behavior" and Conjecture, i.e., the correction factor $\ln(\sqrt{n})$ and constant $C = \lambda$