A General Martingale Approach to Solving a Class of Queueing Problems

Matthew Roughan <matt@serc.rmit.edu.au>
Software Engineering Research Centre
RMIT and Ericsson
WWW: http://www.serc.rmit.edu.au/

PhD supervised by Charles Pearce
supported by Teletraffic Research Centre
at the University of Adelaide

Outline

- Background
  - M/G/1 queue
  - some variants of the M/G/1 queue
- Martingales and stopping times
  - Doob's Optional Stopping Theorem
- Method
  - simple example - Gambler's ruin
  - some results for queueing theory
  - what systems can it be applied to
- Application to the hysteretic threshold overload control
  - Numerical results

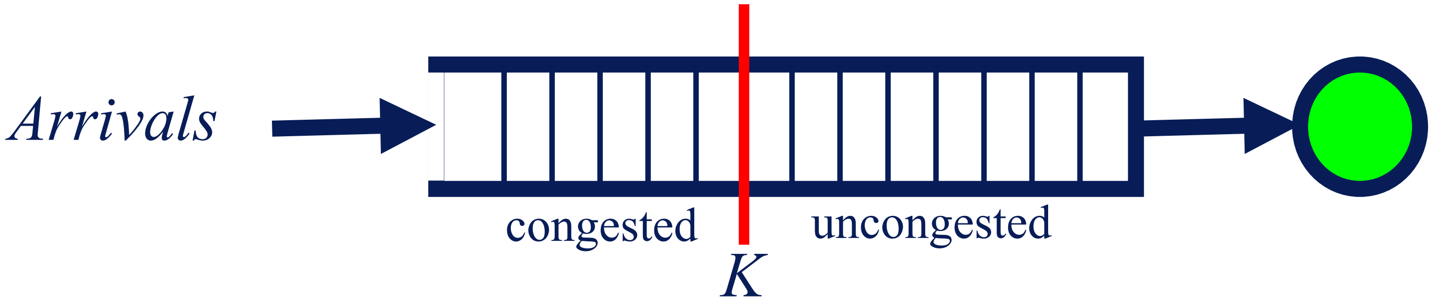
M/G/1 Queue

- **M**arkov Arrivals -- Poisson process rate λ
- **G**enerally distributed service times
- **1** server
- Infinite waiting room

- Service discipline
  - assume FIFO (First In First Out)

- Simple variants
  - generalized vacations

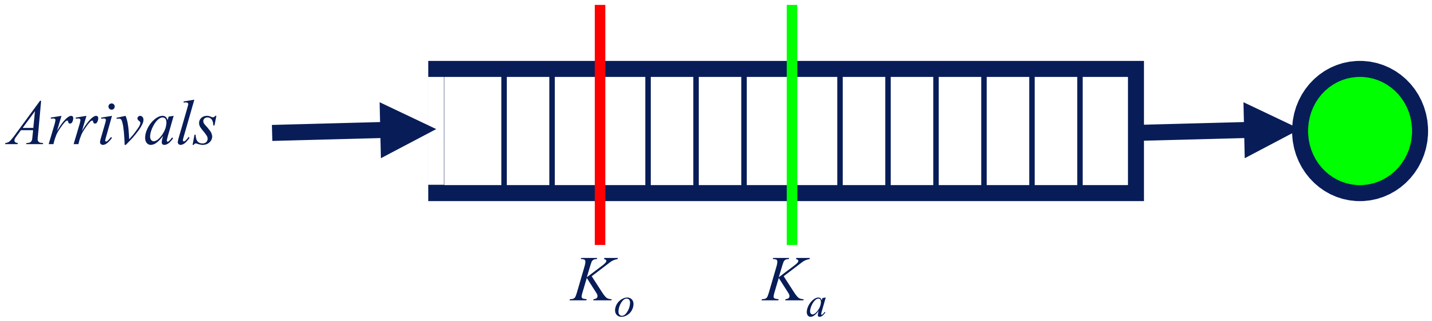
Threshold based overload control



- Queue length threshold K
  - Q≥K queue is congested
  - Q<K queue is uncongested

- When queue is congested slow arrivals, or speed up services
  - automatic call gapping
  - percentage throttling
  - discarding some messages

Hysteretic threshold overload control



- Simple threshold encourages oscillation
  - changing regimes can involve an overhead
  - frequent changes are bad
- Introduce a second threshold
  - congestion onset threshold K_o
  - congestion abatement threshold K_a
- Has been used in real systems, and studied previously
- Congestion state now depends on the history of the queue
  - behavior differs as loads increases or decreases

Martingales and Stopping Times

- Defining properties of martingales

  $E[\ M_{n+1}\ |\ F_n] = M_n$ fair betting process

  $E[\ |\ M_n\ |\ ] < \infty$

- Stopping Times
  - a time T which depends only on the history of the process
  - a R.V. T such that $\{T \leq n\}$ is F_n-measurable
  - times that depend only on the past

    a sensible gambler stops when they run out of money
  - cannot have any dependence on the future

    a gambler can't stop when they hit their maximum

Optional Stopping Theorem

- Under the right conditions for a stopping time T

  $$E[\, M_T \mid F_n] = M_n$$

- Conditions
  - stopping time must be regular for the martingale
    - » important, and non-trivial,
  - often there are simpler sufficient conditions
    - » bounded

Example - Gambler's Ruin

- Gambler starts with $N
- Bets $1 at a time on a fair coin toss
- Stops when
  - runs out of money - "ruin"
  - gets to $K > N

What is the probability of ruin?

How long do you get to play on average?

Example - Analysis of Gambler's Ruin

- Model the problem as a random walk
- X_n is the result of the nth bet (±1)
- After n bets the gambler has \$$S_n$

  $S_0 = N$

  $S_n = S_0 + X_1 + \ldots + X_n$

- Stopping time T is when $S_n = 0$ or K

  $T = \inf\{n \in \mathbb{N} \mid S_n = 0 \text{ or } K\}$

  NB: $S_T = 0$ or K

- Construct a family of exponential martingales

$$M_n(z) = z^{S_n} a(z)^n$$

$$a(z) = \frac{2z}{z^2 + 1}$$

Example - application of OST

- Apply the optional stopping theorem

$$E\left[M_T(z) \middle| F_0\right] = M_0(z)$$
$$E\left[z^{S_T} a(z)^T\right] = z^N$$

- differentiate w.r.t. z, and take z = 1

$$E[S_T] = N$$
$$p\{S_T = K\} = N/K$$

$$a(1) = 1$$
$$a'(1) = 0$$
$$a''(1) = -1$$

- Take $\frac{d}{dz}\left(z\frac{d}{dz}(\bullet)\right)$ and z = 1

$$E[T] = N(K - N)$$

Outline of derivation for M/G/1

- Due to Baccelli and Makowski (1989)
- Consider the system as seen by nth departure from the queue, X_n
- Stopping time $\tau(n)$ is the end of the current busy period at time n
- Result of the Optional Stopping Theorem for M/G/1

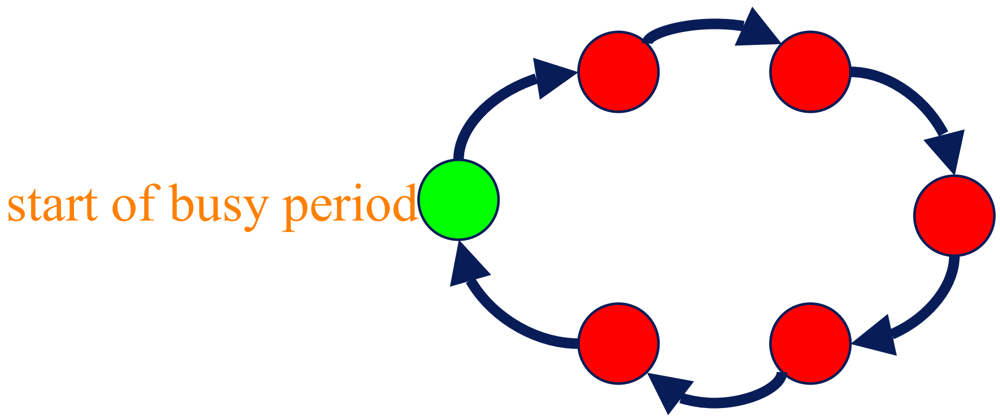
$$E\left[z^{X_n}\right] = E\left[\left(\frac{z}{a(z)}\right)^{\tau(n)-n}\right]$$

  - a(z) is the PGF of the number of arrivals during one service
- The ends of busy periods form an embedded renewal process
  - $\tau(n)$ *-n* is a forward recurrence time
  - can apply the key renewal theorem to obtain limiting distribution of the PGF as n $\to\infty$

$$E\left[z^{X}\right] = (1-\rho)\frac{a(z)(1-z)}{a(z)-z}$$

Classes of systems

- An M/G/1 system that goes through a series of *phases* during one busy period



- The service time distribution is different in each phase
- Properties of phases
  - phase changes occur at service completion times
  - ends of phases must be stopping times
  - length of phases in separate busy period must be independent

Extension of derivation

- Stopping times whenever we switch phases
  - swapping from uncongested to congested or visa versa
- Generalize the renewal process
  - break each renewal time into a series of phases
  - generalize the Key Renewal Theorem
- Obtain a PGF for the equilibrium occupancy distribution in terms of the PGF for the queue length distribution at the end of each phase

Main theorem

- Given stability and regularity conditions, and a M/G/1 type queue which goes through n phases of operation, the PGF of the equilibrium queue length distribution is

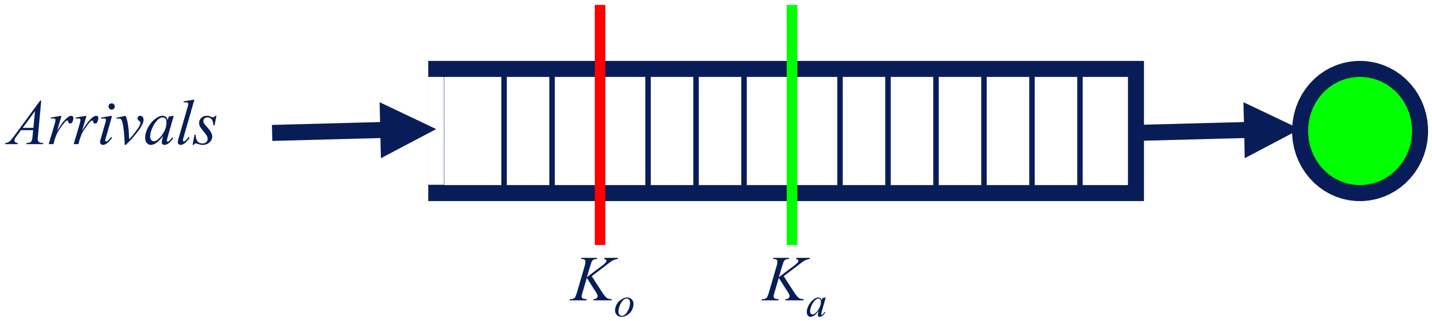
$$E\left[z^X\right] = \frac{1}{m}\left[\frac{E\left[z^{X_{\tau_1(0)}}\right] - z}{1 - z / a_1(z)} + \sum_{j=2}^{n} \frac{E\left[z^{X_{\tau_j(0)}}\right] - E\left[z^{X_{\tau_{j-1}(0)}}\right]}{1 - z / a_j(z)}\right]$$

$\tau_j(0)$ = end of jth phase of the first busy period

$X_{\tau_j(0)}$ = queue length after jth phase of first busy period

$a_j(z)$ = PGF of the number of arrivals during a service of phase j

Two Thresholds



- Two thresholds to control overload control
  - congestion onset threshold K_o
  - congestion abatement threshold K_a
- When congested, discard low priority messages
  - PGF of no. of arrivals during one service when uncongested

$$a_u(z) = \sum a_i^u z^i$$

  - PGF of no. of arrivals during one service when congested

$$a_c(z) = \sum a_i^c z^i$$

Result for two threshold overload control

$$E\left[z^X\right] = \frac{1}{m}\left[\frac{a_c(z)(1-z) + \left(a_c(z) - a_u(z)\right)R_{K_oK_a}(z)}{a_c(z) - z}\right]$$

$$R_{K_oK_a}(z) = \left(\mathbf{e}_1^t + \left(\frac{h_1}{1-h}\right)\mathbf{e}_{K_a}^t\right)\left(\mathbf{I} - \mathbf{P}_{K_o}\right)^{-1}\mathbf{e}_1$$

$$h = 1 - a_0^u \mathbf{e}_{K_a}^t\left(\mathbf{I} - \mathbf{P}_{K_o}\right)^{-1}\mathbf{e}_1$$

$$h_1 = 1 - a_0^u \mathbf{e}_1^t\left(\mathbf{I} - \mathbf{P}_{K_o}\right)^{-1}\mathbf{e}_1$$

$$\mathbf{P}_{K_o} = \begin{pmatrix} a_1^u & a_2^u & a_3^u & \cdots & a_{K_o-1}^u & a_{K_o}^u \\ a_0^u & a_1^u & a_2^u & \cdots & a_{K_o-2}^u & a_{K_o-1}^u \\ 0 & a_0^u & a_1^u & \cdots & a_{K_o-3}^u & a_{K_o-2}^u \\ & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \cdots & a_0^u & a_1^u \end{pmatrix}$$

Other Results

We can obtain a number of other results

- probability of the system being in a particular phase
  - e.g. the congested state

$$p\{\text{congested}\} = \frac{1 + (\rho_u - 1)R_{K_o K_a}(1)}{1 + (\rho_u - \rho_c)R_{K_o K_a}(1)}$$

- mean cycle time (to go from uncongested to congested and back)

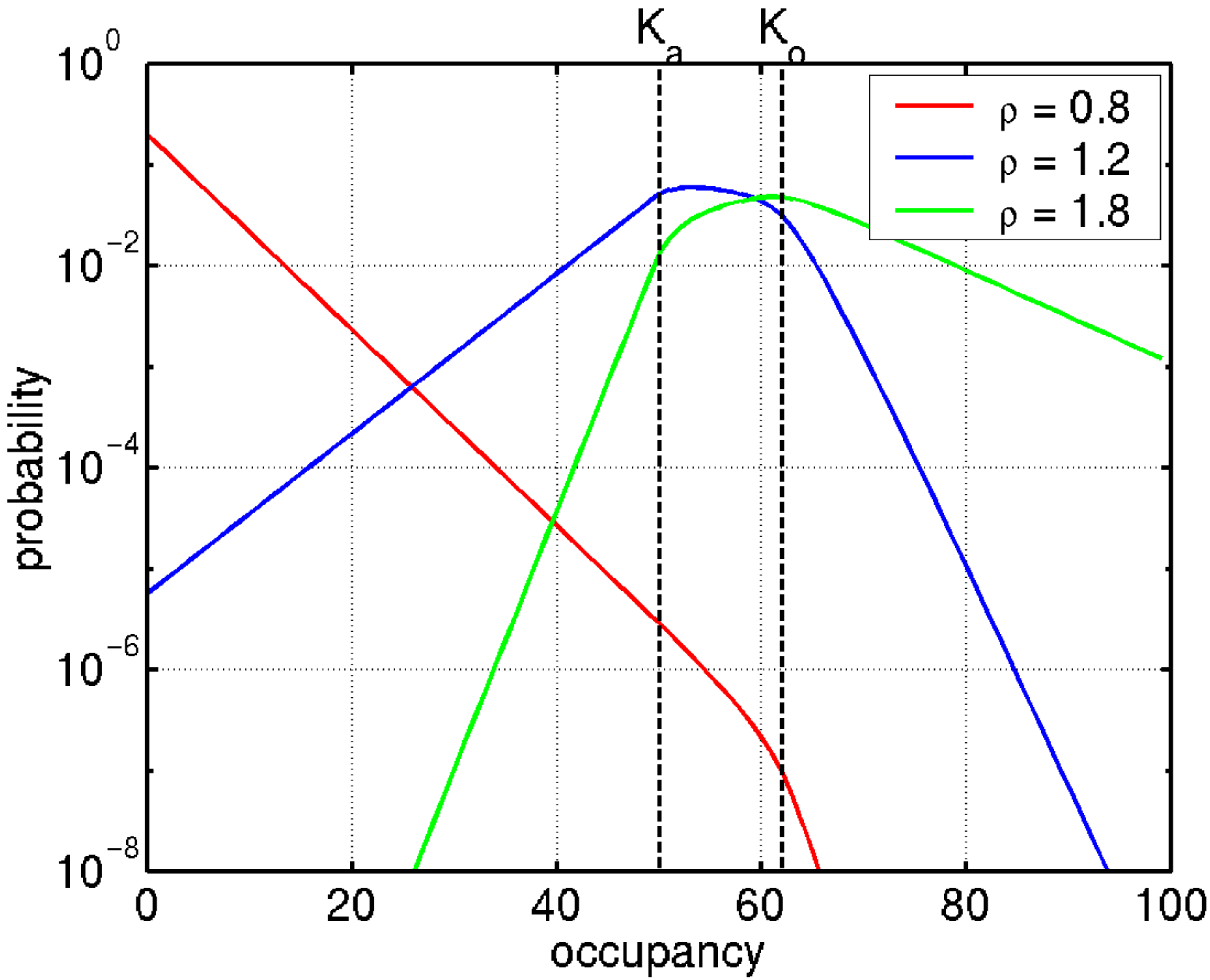
$$E[\nu] = \frac{m(1-h)}{h1}$$

Numerical Results

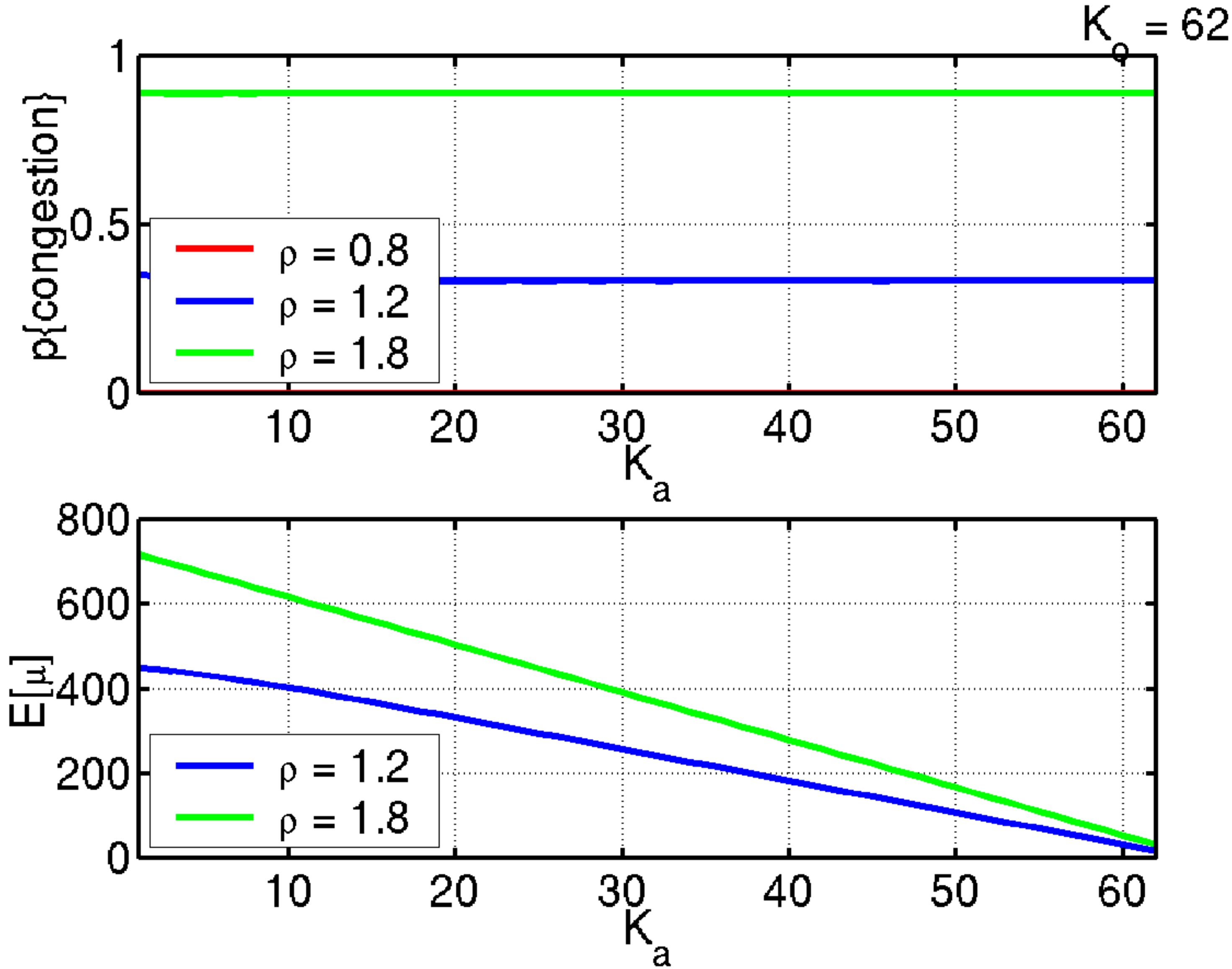
- Scenario
  - congestion onset threshold K_o = 62
  - congestion abatement threshold K_a = 50
  - traffic intensity ρ=0.8, 1.2, 1.8
  - exponential service times
  - 50% random throttling when congested

- Occupancy distribution
  - obtained using FFT based inversion method of Daigle
  - using NEWMAT C++ library

Numerical Results



Numerical Results



Conclusion

- General method for analyzing a set of queueing problems
  - based on martingales, and stopping times
  - M/G/1 queue which goes through a series of phases each with a different service time distribution

- Results M/G/1 queue under overload control
  - Hysteretic overload control does as intended
    - » has little effect when normally loaded
    - » reduce excursions to long queues when overloaded
    - » reduce the effects of oscillation between congested and uncongested regimes