Performance Modeling of Delay-based Dynamic Speed Scaling Systems

Caglar Tunc
caglar@ee.bilkent.edu.tr

Nail Akar
akar@ee.bilkent.edu.tr

Bilkent University
Deparment of Electrical and Electronics Engineering
Ankara, Turkey

June 28, 2016

Outline

- Introduction
- Problem Definition
- Markov Fluid Queues
- Delay-based Dynamic Speed Scaling Model
- Numerical Examples
- Conclusion

Single Server Speed Scaling

- ***Speed scaling***: Adapting the speed of a computer or communication system to tradeoff energy and performance

i. *Static speed scaling*: System is busy ⇒ single speed,
System is idle ⇒ sleep mode

ii. *Dynamic speed scaling*: Speed is continuously adapted based on the system state, i.e., the number of jobs in the system, delay experienced by jobs, etc.

Single Server Speed Scaling

- Low speed ↔ low power
- Takes longer to finish a task with lower speed, BUT generally less energy is consumed
- How to adapt the speed according to the system state in order to obtain energy savings?

Motivating Application Areas

- Adaptive speed in processors and computer systems
  - Change the speed of a processor according to the number of jobs waiting in the system to save energy [1]
- Adaptive link rate (ALR) schemes in Ethernet links
  - Change the rate of an Ethernet link according to the link utilization to obtain energy savings (not standardized) [2]
  - Data rate = $\begin{cases} 100\text{ Mbps}, & \text{if link utilization} < 10\% \\ 1\text{ Gbps}, & \text{if link utilization} \geq 10\% \end{cases}$

[1] F. Yao, A. Demers, and S. Shenker. A Scheduling Model for Reduced CPU Energy. *In Proceedings of FOCS '95*, pages 374-, Washington, DC, USA, 1995. IEEE Computer Society.
[2] C. Gunaratne, K. Christensen, B. Nordman, and S. Suen. Reducing the Energy Consumption of Ethernet with Adaptive Link Rate (ALR). *Computers, IEEE Trans. on*, 57(4):448-461, April 2008.

Motivating Future Applications

- Wireless link that supports different power levels and adaptive coding and modulation (ACM) techniques
- Adjust the link rate according to delays of the jobs in the system
- Save from the power while satisfying QoS constraints

Delay-based Dynamic Speed Scaling

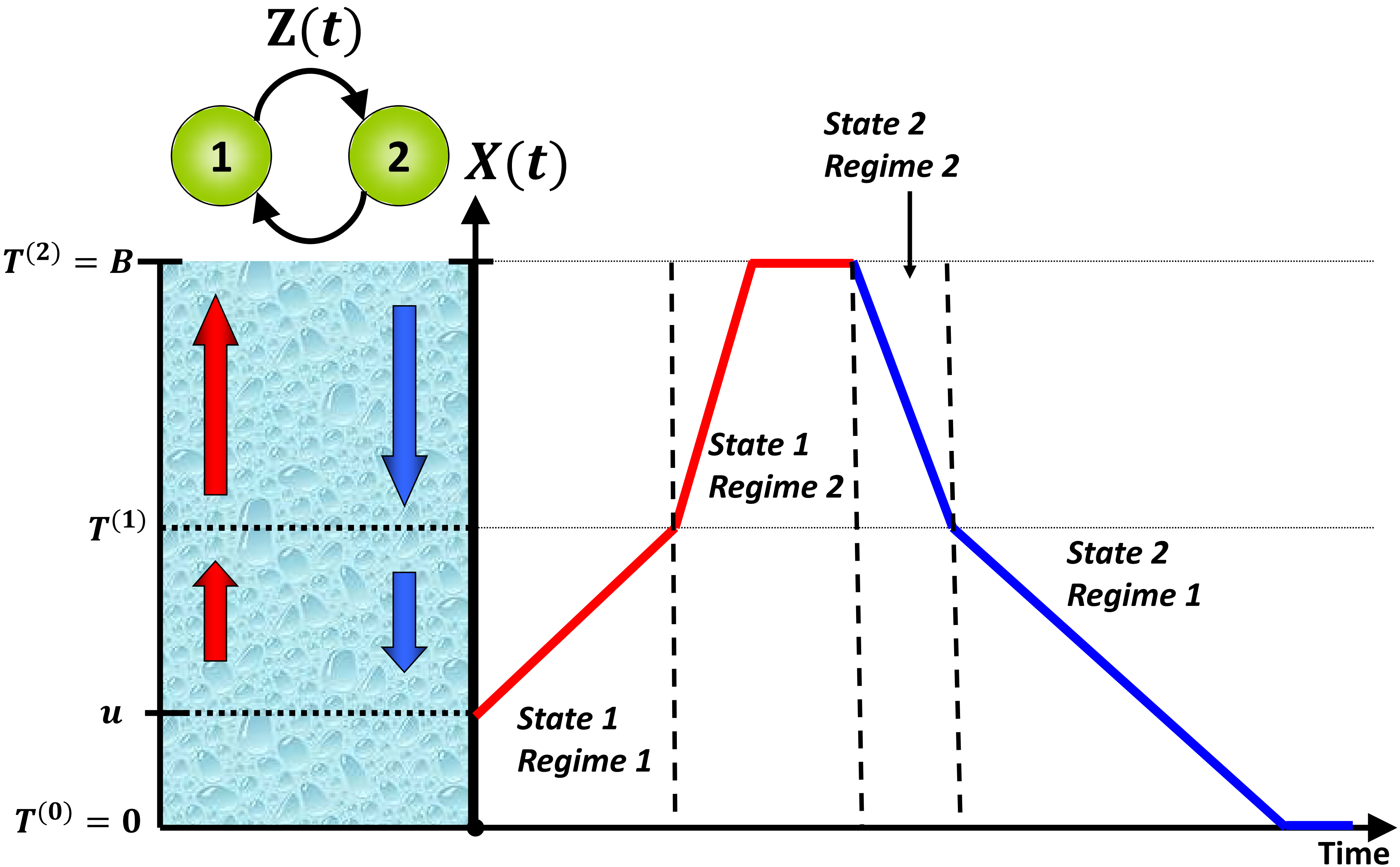
- Assign a service rate for the head-of-the-line (HOL) job of a FIFO queue according to the total delay it has experienced in the system
- Jobs may have strict deadlines
  - Jobs with delays greater than the deadline abandon the system without service

Low service rate $\rightarrow$ Low power $\rightarrow$ Energy saving

Markov Fluid Queues (MFQs)

- Background process determines the rate of change (*drift*) of a buffer
- Finite state space Continuous Time Markov Chain (CTMC)
- Each state has its own drift value
- Infinitesimal generator and drift values
- **Multi-Regime (Multi-Layer/Multi-Threshold) MFQ (MRMFQ)**
  - Buffer is divided into a finite number of regimes
  - Each regime has own infinitesimal generator and drift values

Sample Evolution of an MRMFQ



Multi-Regime Markov Fluid Queues

$$f_i^{(k)}(x) = \lim_{t\to\infty} \frac{d}{dx} \Pr\{X(t) \le x,\, Z(t) = i\},$$

$$f^{(k)}(x) = \left[f_0^{(k)}(x) \;\; f_1^{(k)}(x) \;\; \ldots \;\; f_{N-1}^{(k)}(x)\right],$$

$$c_i^{(k)} = \lim_{t\to\infty} \Pr\{X(t) = T^{(k)},\, Z(t) = i\},$$

$$c^{(k)} = \left[c_0^{(k)} \; c_1^{(k)} \;\; \ldots \;\; c_{N-1}^{(k)}\right],$$

$$\longrightarrow \; \frac{d}{dx} f^{(k)}(x) R^{(k)} = f^{(k)}(x) Q^{(k)}.$$

- $Z(t)$: N-state CTMC, $N < \infty$
- $Q^{(k)}$: Infinitesimal generator of $Z(t)$ for $1 \le k \le K$
- $r_i^{(k)}$: Net drift of the buffer for $0 \le i \le N-1$ and $1 \le k \le K$
- $R^{(k)}$: $diag\left(r_0^{(k)} \; r_1^{(k)} \ldots \; r_{N-1}^{(k)}\right)$, for $1 \le k \le K$

[1] H. E. Kankaya and N. Akar. Solving multi-regime feedback fluid queues. *Stochastic Models*, 24(3):425-450, 2008.

Multi-Regime Markov Fluid Queues

$$f_i^{(k)}(x) = \lim_{t\to\infty} \frac{d}{dx} \Pr\{X(t) \le x,\, Z(t) = i\},$$

$$f^{(k)}(x) = \left[f_0^{(k)}(x)\;\; f_1^{(k)}(x)\;\; \dots\;\; f_{N-1}^{(k)}(x)\right],$$

$$c_i^{(k)} = \lim_{t\to\infty} \Pr\left\{X(t) = T^{(k)},\, Z(t) = i\right\},$$

$$c^{(k)} = \left[c_0^{(k)}\; c_1^{(k)}\;\; \dots\;\; c_{N-1}^{(k)}\right],$$

$$\longrightarrow\; \frac{d}{dx} f^{(k)}(x) R^{(k)} = f^{(k)}(x) Q^{(k)}.$$

- $\left[T^{(0)}\; T^{(1)} \dots T^{(K)}\right]$: Boundary points, $T^{(0)}$=0, $T^{(K)}$=∞
- $\tilde{Q}^{(k)}$: Infinitesimal generator at boundary k for $0 \le k < K$
- $\tilde{r}_i^{(k)}$: Net drift of the buffer at boundary k for $0 \le k < K$
- $\tilde{R}^{(k)}$: $diag\left(r_0^{(k)}\; r_1^{(k)} \dots\; r_{N-1}^{(k)}\right)$, for $1 \le k < K$

Boundary Conditions of MRMFQs

$$c_i^{(0)} = 0, \quad \forall i \in S_+^{(1)}$$

$$c_i^{(k)} = 0, \quad \forall i \in \left(S_+^{(k)} \cap S_+^{(k+1)}\right) \cup \left(S_-^{(k)} \cap S_-^{(k+1)}\right)$$

$$c_i^{(k)} = 0, \quad \forall i \in \left(S_-^{(k)} \cap S_+^{(k+1)}\right) \cap \left(\tilde{S}_+^{(k)} \cup S_-^{(k)}\right)$$

$$f^{(1)}(0+)R^{(1)} = c^{(0)}\tilde{Q}^{(0)}$$

$$f^{(k+1)}(T^{(k)}+)R^{(k+1)} - f^{(k)}(T^{(k)}-)R^{(k)} = c^{(k)}\tilde{Q}^{(k)}$$

$$f_i^{(k)}(T^{(k)}-) = 0 \quad \forall i \in S_-^{(k)} \cup \left(\tilde{S}_0^{(k)} \cap \tilde{S}_+^{(k)}\right)$$

$$f_i^{(k+1)}(T^{(k)}+) = 0 \quad \forall i \in \left(\tilde{S}_0^{(k)} \cap \tilde{S}_-^{(k)}\right) \cup S_+^{(k+1)}$$

$$\left(\sum_{k=1}^{K} \int_{T^{(k-1)}+}^{T^{(k)}-} f^{(k)}(x)dx + \sum_{k=0}^{K-1} c^{(k)}\right)\mathbf{1} = 1$$

Computational Complexity

- An N-state K-regime MFQ system requires
  - a Schur decomposition and a pair of Sylvester equations for each regime: $O(N^3 K)$
  - the solution of a linear matrix equation of at most size $N(2K+1)$
  - Exploiting the block tridiagonal form of the linear matrix equation reduces the computational complexity to $\boldsymbol{O(N^3 K)}$ [1]

[1] M. A. Yazici and N. Akar. The finite/innite horizon ruin problem with multi-threshold premiums: a Markov fluid queue approach. *Annals of Operations Research*, 2016.

System Model

- Server has $K + 1$ available service rates to select
- Exponentially distributed service times with rate μ_k, $k = 1, \dots, K + 1$
- Poisson job arrivals with rate λ
- $D(t)$: Delay already experienced by the HOL job at service start time t
- $A(t)$: Unfinished work (process) in the system at time t
- $X(t)$: Fluid level at time t, obtained by replacing abrupt jumps in $S(t)$ by linear decrements

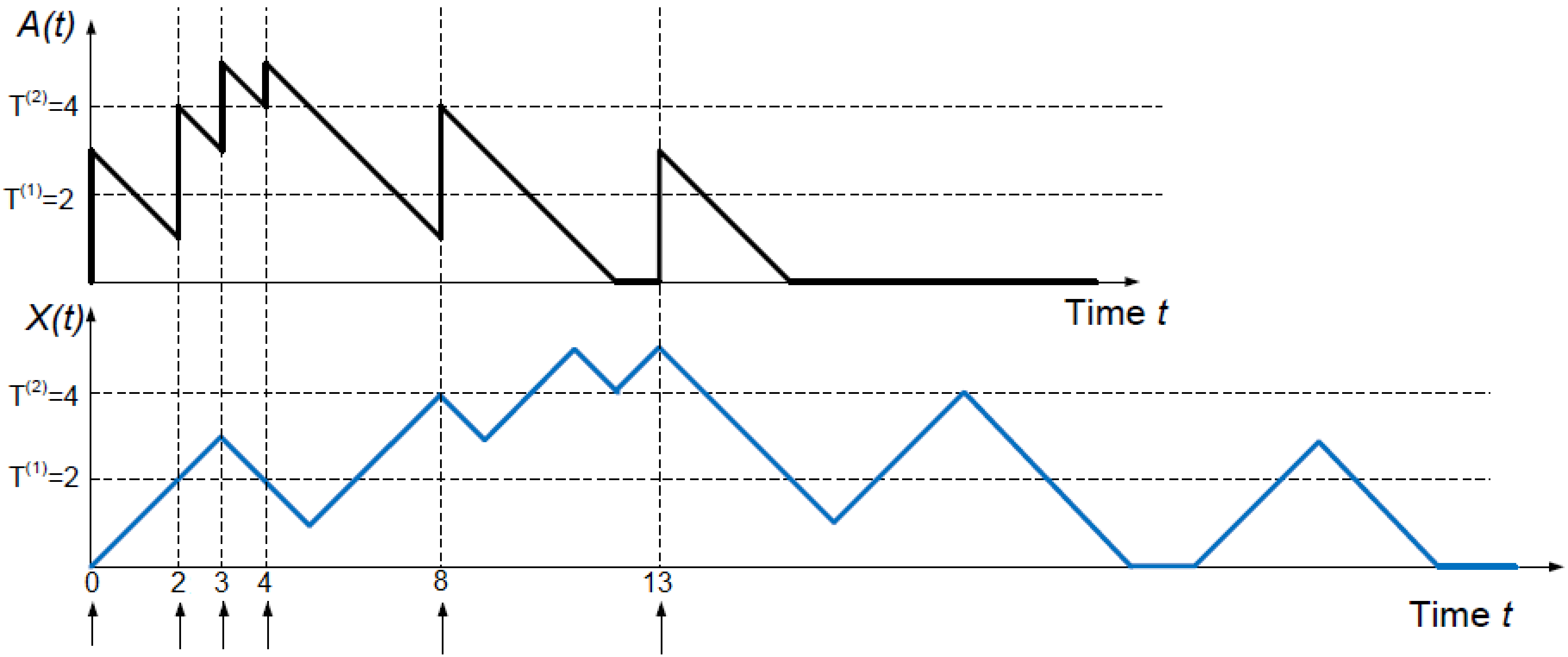
System Model

- Regime boundaries of the MRMFQ model

$$0 = T^{(0)} < T^{(1)} < \cdots < T^{(K)} < T^{(K+1)} = \infty$$

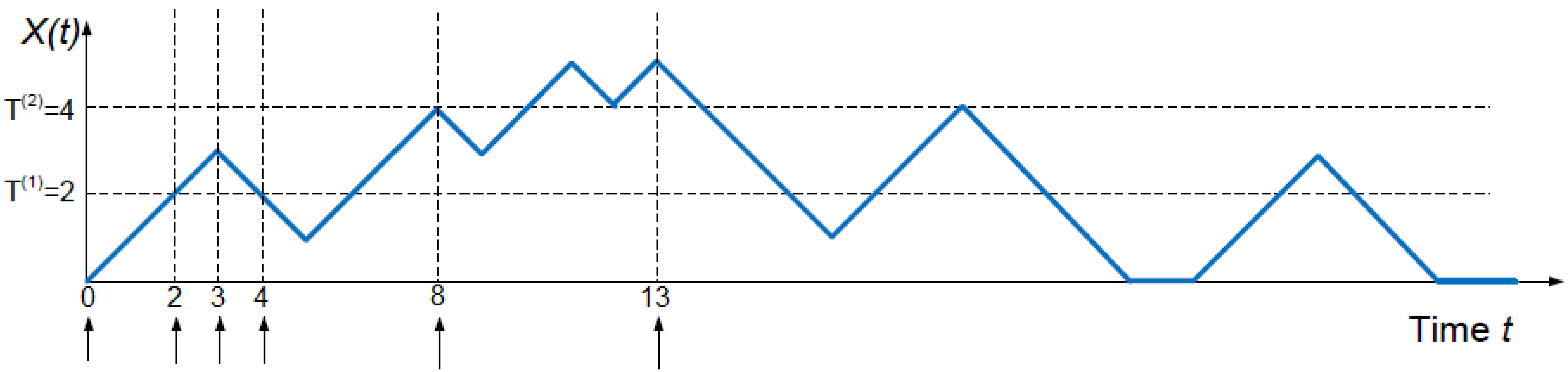
- When $T^{(k-1)} \leq D(t) < T^{(k)}$, the HOL job is served with rate μ_k
- Service rate is fixed during the service of the HOL job.
- Operating power at rate μ_k is P_k.
- If $T^{(K)} \leq D(t)$, the job is either: i) served with rate μ_{K+1}, or ii) blocked.
- $T^{(K)}$ is called the ***deadline*** or ***delay threshold***.

Sample Paths



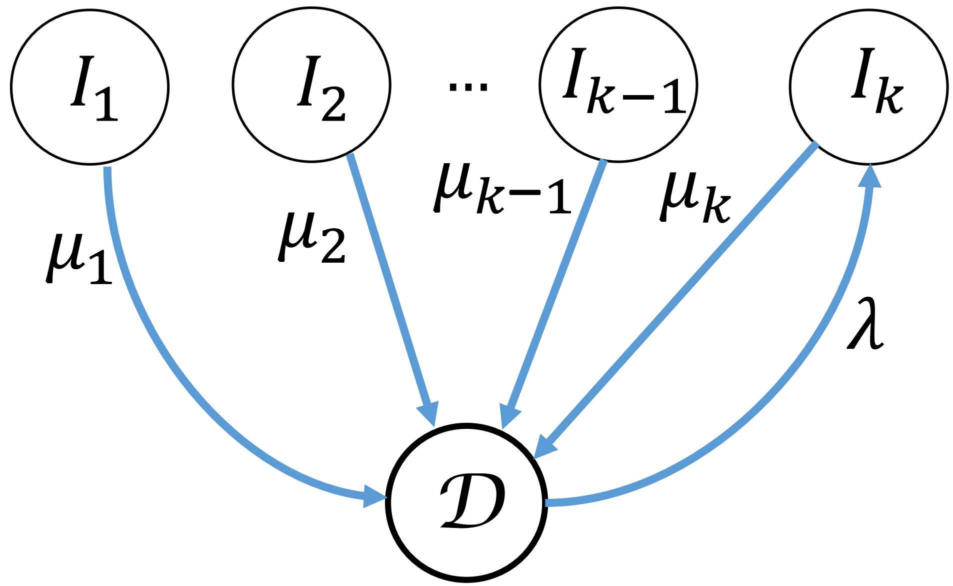
State Space

- I_k: Service state in regime k, $k = 1,2,\ldots,K+1$
  - $I_k \rightarrow \mu_k$
  - $X(t)$ is increased with a drift of 1.
- $\mathcal{D}$: State representing the inter-arrival times
  - $X(t)$ is decreased with a drift of 1.

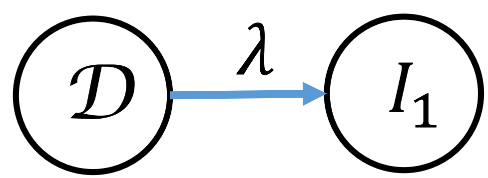


State Transitions

- Regime-k



- $X(t) = 0$



Infinitesimal Generator and Drift Matrices

$$
Q^{(j)} = \begin{array}{c} \\ \begin{array}{c} I_{K+1} \\ \vdots \\ I_{j+1} \\ I_j \\ I_{j-1} \\ \vdots \\ I_1 \\ \mathcal{D} \end{array} \end{array}
\begin{array}{c}
\begin{array}{cccccccc} I_{K+1} & \cdots & I_{j+1} & I_j & I_{j-1} & \cdots & I_1 & \mathcal{D} \end{array} \\
\begin{bmatrix}
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & \cdots & 0 & -\mu_j & 0 & \cdots & 0 & \mu_j \\
0 & \cdots & 0 & 0 & -\mu_{j-1} & \cdots & 0 & \mu_{j-1} \\
\vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & \cdots & 0 & 0 & 0 & \cdots & -\mu_1 & \mu_1 \\
0 & \cdots & 0 & \lambda & 0 & \cdots & 0 & -\lambda
\end{bmatrix}
\end{array},
$$

$\tilde{Q}^{(j)} = Q^{(j+1)}$, except that there is no transition from I_1 to $\mathcal{D}$ in $\tilde{Q}^{(0)}$

$$
R^{(k)} = \boldsymbol{diag}(\boldsymbol{I}, -1)\,,\ 1 \le k \le K+1, \quad \tilde{R}^{(k)} = \begin{cases} R^{(k+1)}, & 1 \le k \le K \\ \mathbf{max}\left(0, R^{(1)}\right), & k = 0 \end{cases}
$$

The Delay Distribution

- $A(t)$ determines the amount of delay that newly arriving jobs will experience.
- By *PASTA* property, average system power, blocking probability and the delay distribution can be calculated from the steady-state probability distribution of state $\mathcal{D}$.

$$\lim_{t \to \infty} \Pr\{A(t) \leq x\} = \lim_{t \to \infty} \frac{\Pr\{X(t) \leq x,\, Z(t) = \mathcal{D}\}}{\Pr\{Z(t) = \mathcal{D}\}}$$

Average Operating Power

- p_k: probability that a newly arriving job finds the system in regime k
- p_0: probability that a newly arriving job finds the system empty

$$p_k = \lim_{t\to\infty} \Pr\left\{T^{(k-1)} < A(t) < T^{(k)}\right\}, \quad 1 \le k \le K+1$$

$$p_0 = \lim_{t\to\infty} \Pr\{A(t) = 0\}$$

- q_k: probability that a job is served with rate μ_k

$$q_k = \begin{cases} p_k, & k \ge 2, \\ p_0 + p_1, & k = 1. \end{cases}$$

$$P_{avg} = p_0 P_I + (1 - p_0) \sum_{k=1}^{K+1} \frac{\frac{q_k}{\mu_k}}{\sum_{i=1}^{K+1} \frac{q_i}{\mu_i}} P_k$$

Blocking Probability

- For the case of abandonments: $\mu_{K+1} \to \infty$, no energy is consumed
- p_b: blocking probability

$$p_b = \lim_{\mu_{K+1} \to \infty} p_{K+1} = \lim_{t \to \infty} \lim_{\mu_{K+1} \to \infty} \Pr\left\{A(t) \geq T^{(K)}\right\}.$$

Numerical Examples

Example I – Case of Abondonments

- $K = 2, T^{(1)} = 10, T^{(2)} = 20, \mu_1 = 0.5, \mu_2 = 1, \eta = \lambda/\mu_2$
- Jobs with delays greater than $T^{(2)} = 20$ abandon the system
- $P_I = 0, P_k = {\mu_k}^2$
- Increase μ_3 in order to model abandonments

μ_3	p_b (%)		P_{avg}	
	$\eta = 0.4$	$\eta = 0.8$	$\eta = 0.4$	$\eta = 0.8$
1e2	0.1123	3.0429	0.2238	0.6662
1e4	0.1118	3.0196	0.2238	0.6664
1e6	0.1118	3.0193	0.2238	0.6664
1e8	0.1118	3.0193	0.2238	0.6664
Sim	0.1118	3.0185	0.2238	0.6664

Table 1: Blocking probability p_b and average system power P_{avg} compared with simulation results for two values of $\eta = 0.4, 0.8$.

Example II – Piecewise Linear Rate Adjustment Policy (PiLRAP)

- Selects service rates from piecewise linear functions of the unfinished work process $A(t)$ from the interval $[\mu_{min}, \mu_{max}]$.
- $\mu_K = \mu_{max}$
- Jobs with $A(t) \geq T^{(K)}$ are blocked.
- (x_0, y_0) point determines the exact service rate function.

Example II – Piecewise Linear Rate Adjustment Policy (PiLRAP)

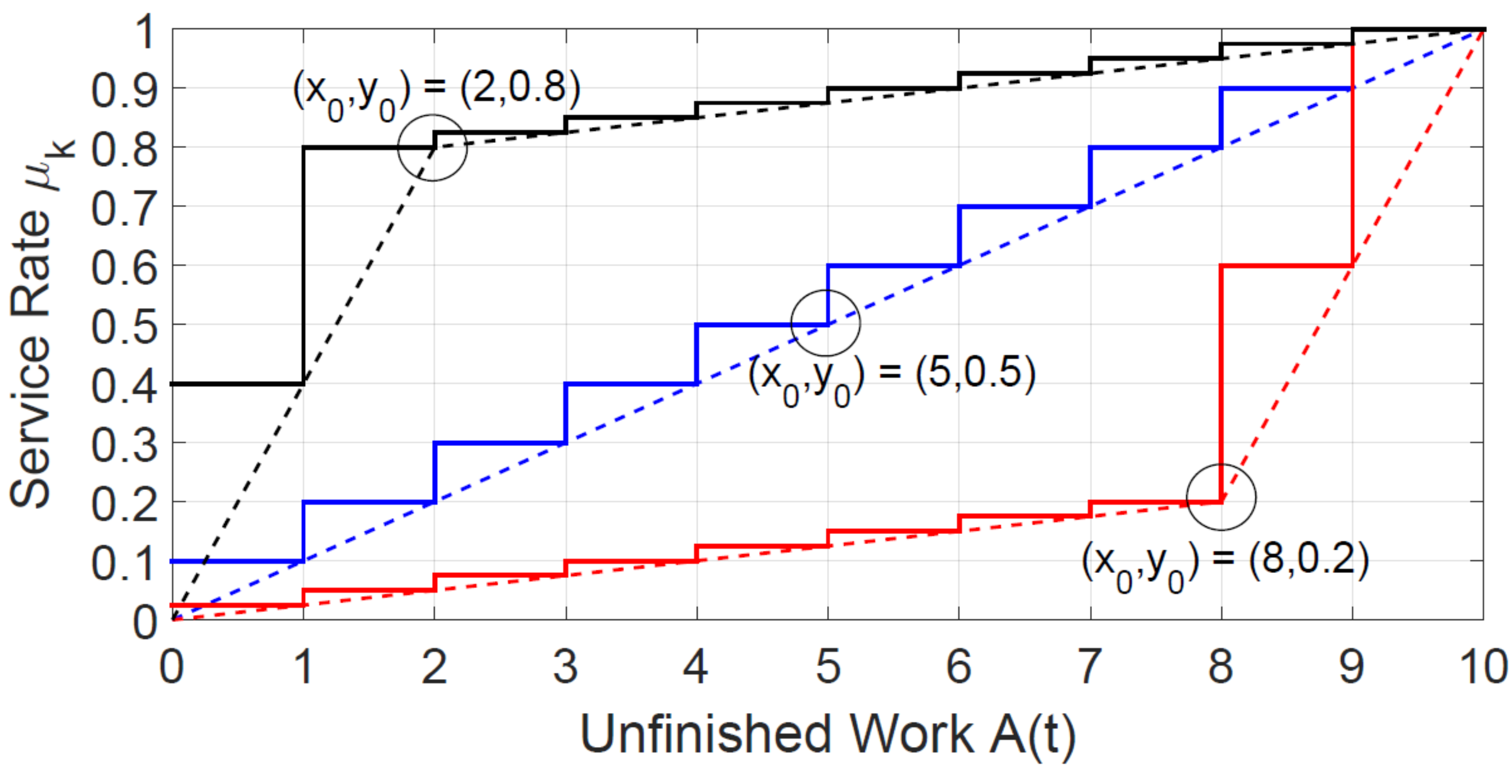


Figure 1: Service rate function (dashed lines) and actual service rate μ_K (straight lines) as functions of $A(t)$ for $\mu_{min} = 0,\ \mu_{max} = 1,\ T^{(K)} = 10,\ K = 10.$

Example II – Piecewise Linear Rate Adjustment Policy (PiLRAP)

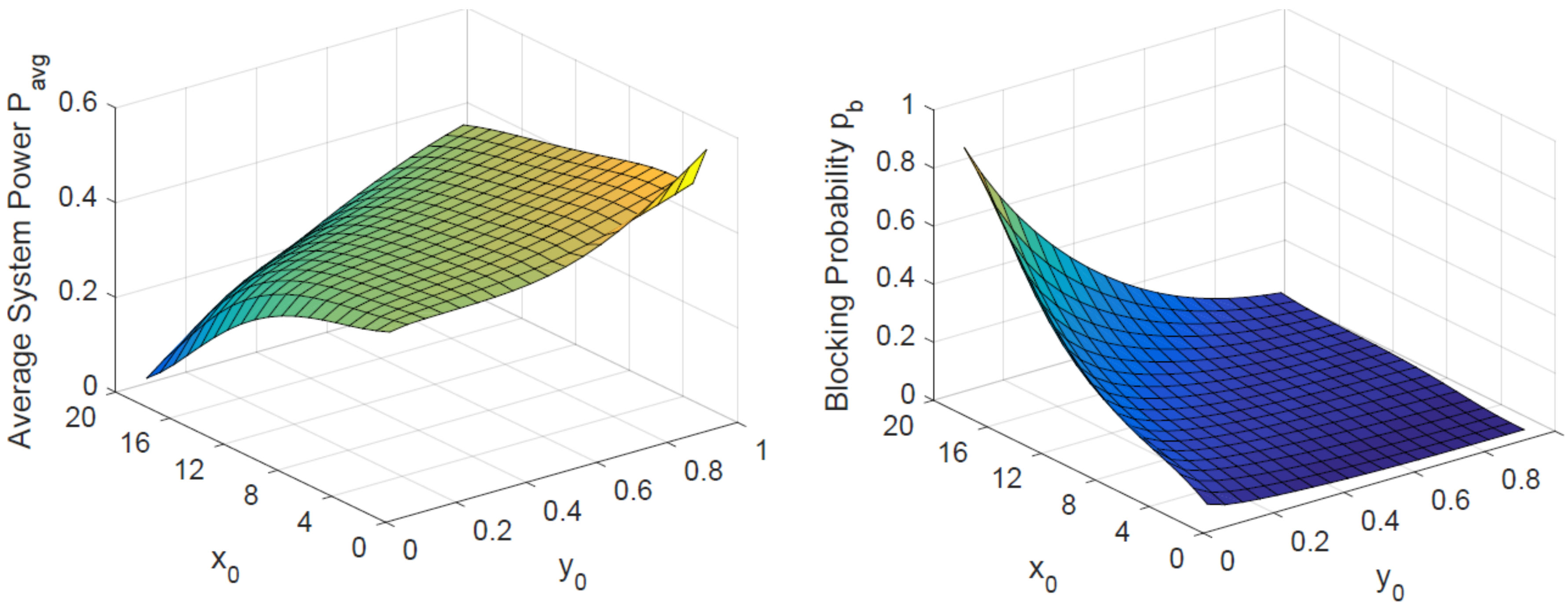


Figure 2: Average system power P_{avg} and blocking probability p_b as functions of parameters x_0 and y_0 for $K = 20$.

Example III – Comparison with Static Speed Scaling

- $K = 1, T^{(1)} = 20, \mu_1 = \mu_{max} = 1$
- M/M/1 queue with load $\rho = \lambda/\mu_{max} \rightarrow P_f = (1-\rho)P_I + \rho P_1$
- $G = 100\frac{(P_f - P_{avg})}{P_f}$
- Blocking probability should be less than 0.01

Example III – Comparison with Static Speed Scaling

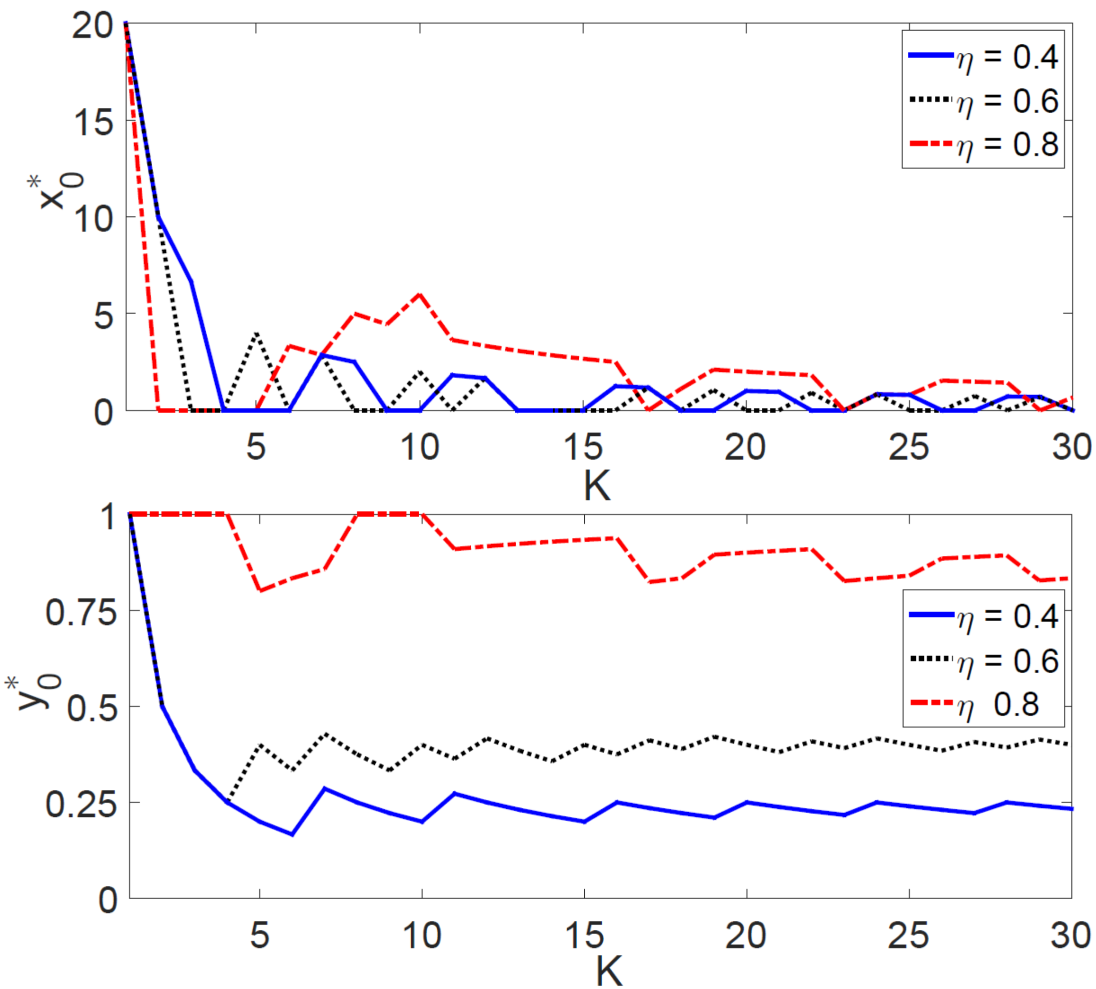


Figure 3: Optimal values of x_0 and y_0, denoted by x_0^* and y_0^*, as functions of K for $\eta = 0.4, 0.6, 0.8$.

Example III – Comparison with Static Speed Scaling

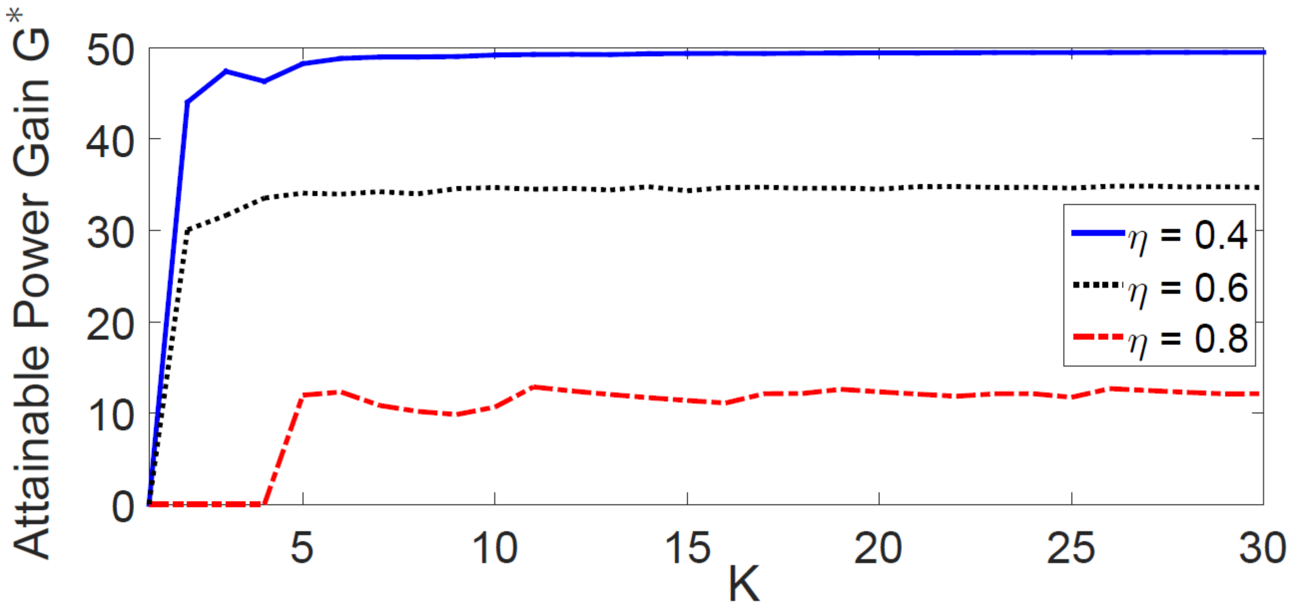


Figure 4: Attainable power gain, denoted by G^*, as a function of K for $\eta = 0.4, 0.6, 0.8$.

Example III – Comparison with Static Speed Scaling

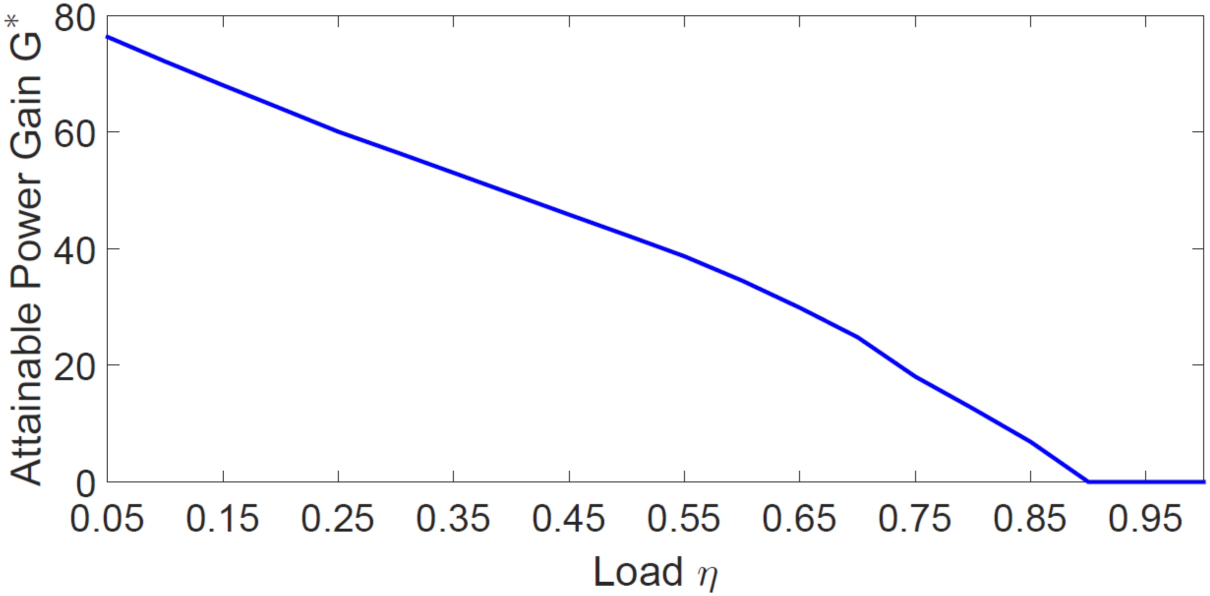


Figure 5: Attainable power gain, denoted by G^*, as a function of the load η for K=20.

Conclusion

- We propose an MRMFQ model of a dynamic speed scaling system, in which a service rate is decided according to the delay of the HOL job.
- Piecewise Linear Rate Adjustment Policy (PiLRAP) is proposed which minimizes the power consumption under job blocking probability constraints.

Future Work

- More general arrival process such as MAP
- Other service time distributions, such Phase-type distribution
- Detailed analysis of a real life application
- Zero-drift states to model abandonments to deal with the case $\mu_{K+1} \to \infty$
- Multi-server case

Acknowledgment

- This study is funded by TÜBİTAK (The Scientific and Technological Research Council of Turkey) with ARDEB 1001 (under the project number 115E360) program.

Thank you for your attention.
Any questions?

Markov Fluid Queues (MFQs)

- Single-Regime MFQ (SRMFQ)
  - Buffer considered as a single regime
  - Fixed infinitesimal generator and drift values
- Multi-Regime MFQ (MRMFQ)
  - Buffer is divided into a finite number of regimes
  - Each regime has own infinitesimal generator and drift values
- Continuous-Feedback MFQ (CFMFQ)
  - Infinitesimal generator and drift values as continuous functions of the buffer level

Steady-state Solution of MRMFQs

$$A^{(k)} = Q^{(k)}\left(R^{(k)}\right)^{-1} \rightarrow A^{(k)}Y^{(k)} = Y^{(k)}\begin{bmatrix} 0 & & \\ & A_{-}^{(k)} & \\ & & A_{+}^{(k)} \end{bmatrix},$$

$$\left(Y^{(k)}\right)^{-1} = \begin{bmatrix} L_0^{(k)} \\ L_{-}^{(k)} \\ L_{+}^{(k)} \end{bmatrix} \rightarrow f^{(k)}(x) = a^{(k)}\begin{bmatrix} L_0^{(k)} \\ e^{A_{-}^{(k)}\left(x-T^{(k-1)}\right)}L_{-}^{(k)} \\ e^{-A_{+}^{(k)}\left(T^{(k)}-x\right)}L_{+}^{(k)} \end{bmatrix},$$

$a^{(k)} = \left[a_0^{(k)}\ \ a_{-}^{(k)}\ \ a_{+}^{(k)}\right]$: vector of unknown coefficients

Stability Conditions

1. Mean drift in the last regime should be negative, i.e.,

$$\pi^{(K)} R^{(K)} \mathbf{1} < 0$$

2. $f^{(K)}(x)$ should be bounded, i.e.,

$$a_0^{(K)} = 0, a_+^{(K)} = \mathbf{0},$$

State Transitions