CORS 2016

Autoscaling Effects in Speed Scaling Systems

Carey Williamson
Department of Computer Science
(joint work with Maryam Elahi and Philipp Woelfel)

UNIVERSITY OF CALGARY

Introduction

- Dynamic CPU speed scaling systems
- Service rate adjusted based on offered load
- Classic tradeoff:
  - Faster speed → lower response time, higher energy usage
- Two key design choices:
  - Speed scaler: how fast to run? (static, coupled, decoupled)
  - Scheduler: which job to run? (FCFS, PS, FSP, SRPT, LRPT)
- Research questions:
  - What are the “autoscaling” properties of coupled (i.e., job-count based) speed scaling systems under heavy load?
  - In what ways are PS and SRPT similar or different?

UNIVERSITY OF CALGARY

Related Work

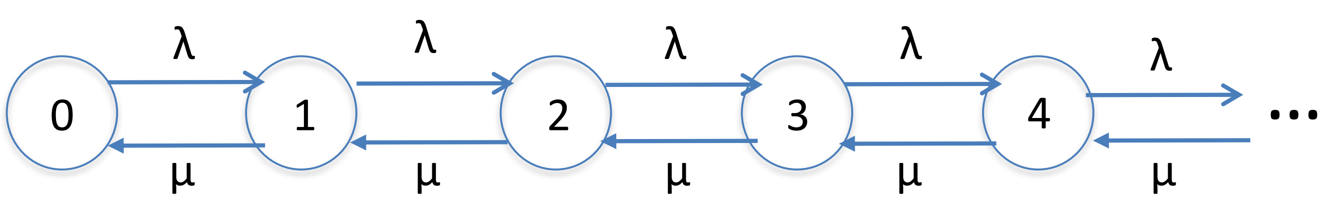
- [Albers 2010] “Energy-Efficient Algorthms”, CACM
- [Andrew et al. 2010] “Optimality, Fairness, and Robustness in Speed Scaling Designs”, ACM SIGMETRICS
- [Bansal et al. 2007] “Speed Scaling to Manage Energy and Temperature”, JACM
- [Elahi et al. 2012] “Decoupled Speed Scaling”, QEST, PEVA
- [Wierman et al. 2009] “Power-Aware Speed Scaling in Processor Sharing Systems”, IEEE INFOCOM, PEVA 2012
- [Weiser et al. 1994] “Scheduling for Reduced CPU Energy”, USENIX OSDI
- [Yao et al. 1995] “A Scheduling Model for Reduced CPU Engergy”, ACM FOCS

System Model (1 of 4)

Review: Birth-death Markov chain model of classic M/M/1 queue

Fixed arrival rate λ

Fixed service rate μ



Mean system occupancy: $N = \rho / (1 - \rho)$

Ergodicity requirement: $\rho = \lambda/\mu < 1$

$p_n = p_0 (\lambda/\mu)^n$

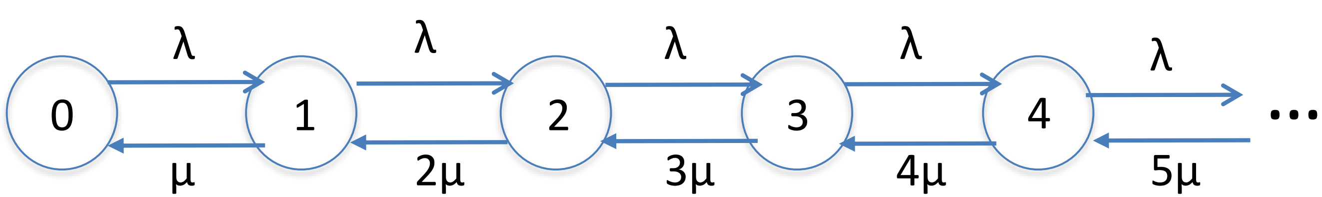
$U = 1 - p_0 = \rho$

System Model (2 of 4)

Birth-death Markov chain model of classic M/M/∞ queue

Fixed arrival rate λ

Service rate scales linearly with system occupancy (α = 1)



Mean system occupancy: $N = \rho = \lambda/\mu$

System occupancy has Poisson distribution

Ergodicity requirement: $\rho = \lambda/\mu < \infty$

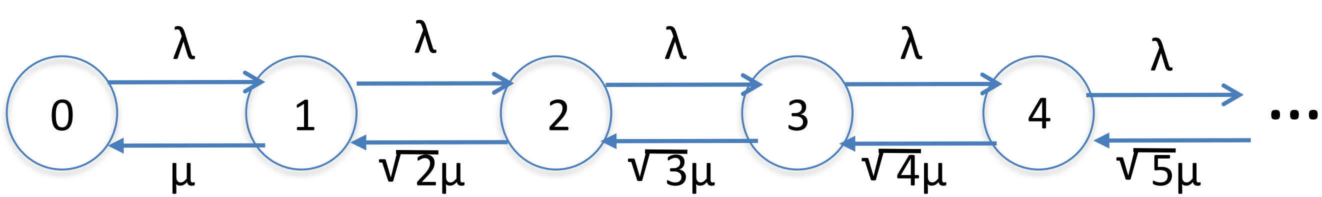
$$p_n = p_0 \prod_{i=0}^{n-1} (\lambda/(i+1)\mu)$$

$$U = 1 - p_0 \neq \rho$$

FCFS = PS ≠ SRPT

System Model (3 of 4)

Birth-death Markov chain model of dynamic speed scaling system
Fixed arrival rate λ
Service rate scales sub-linearly with system occupancy (α = 2)



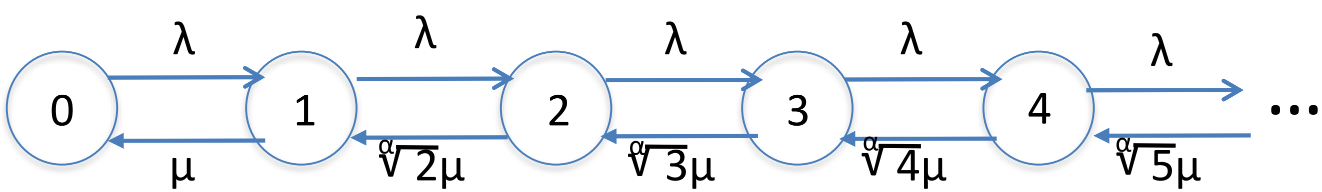
Mean system occupancy: $N = \rho^2 = (\lambda/\mu)^2$ $\quad p_n = p_0 \prod_{i=0}^{n-1} (\lambda/(\sqrt{i+1})\mu)$
System occupancy has higher variance than Poisson distribution
Ergodicity requirement: $\rho = \lambda/\mu < \infty$

System Model (4 of 4)

Birth-death Markov chain model of dynamic speed scaling system

Fixed arrival rate λ

Service rate scales sub-linearly with system occupancy ($\alpha > 1$)



Mean system occupancy: $N = \rho^{\alpha} = (\lambda/\mu)^{\alpha}$ $\quad p_n = p_0 \prod_{i=0}^{n-1} (\lambda/(\sqrt[\alpha]{i+1})\mu)$

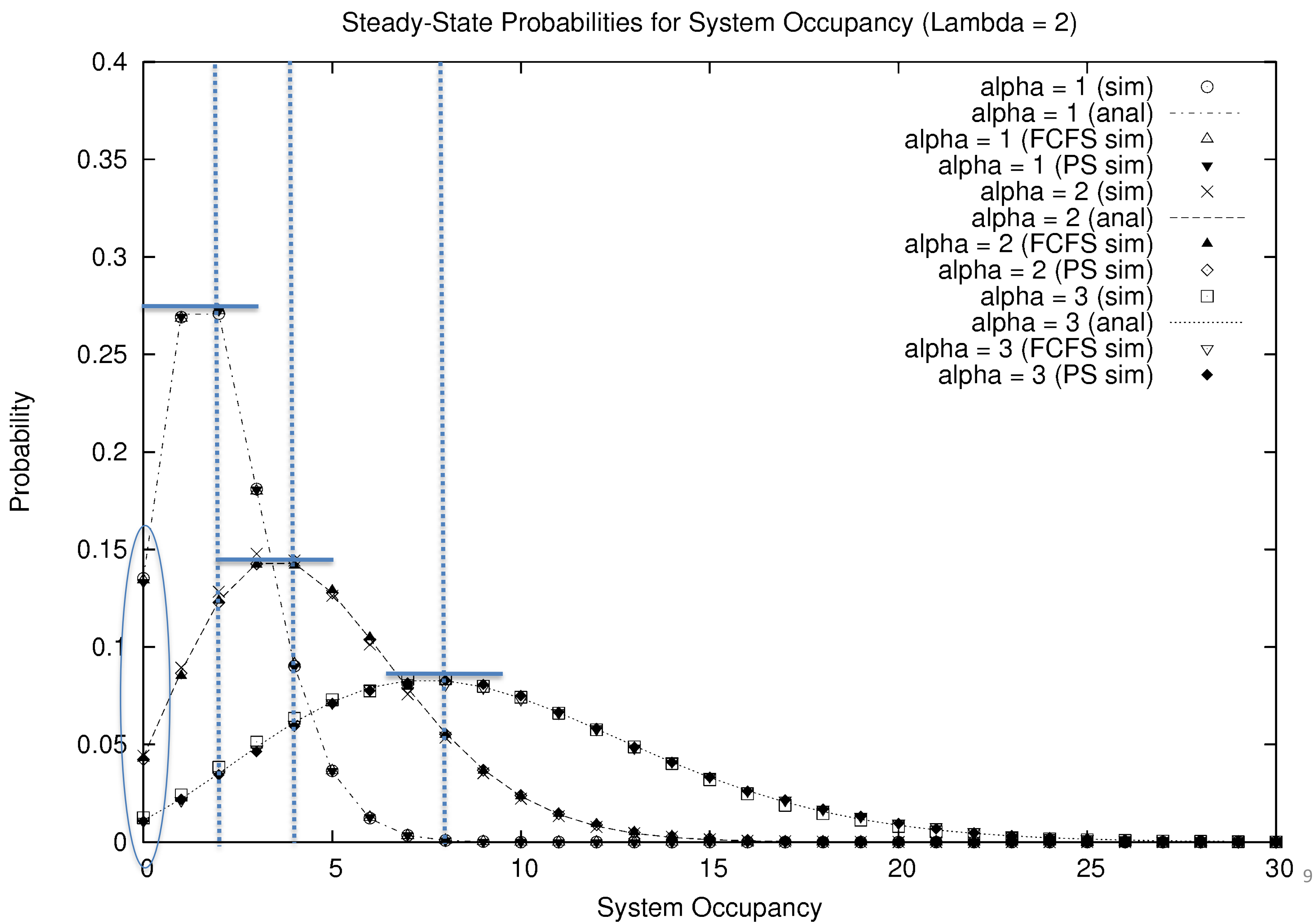
System occupancy has higher variance than Poisson distribution

Ergodicity requirement: $\rho = \lambda/\mu < \infty$

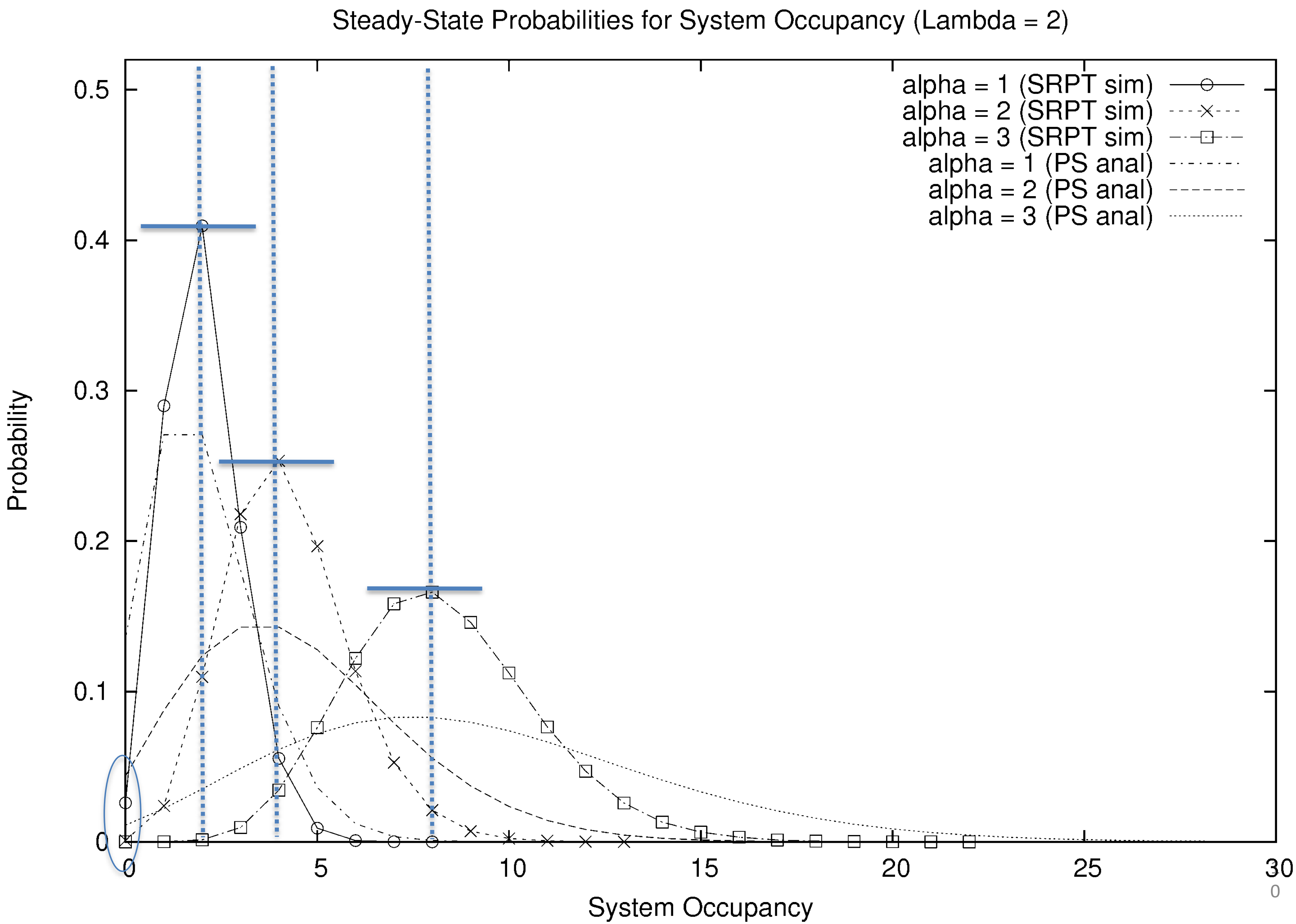
Analytical Insights and Observations

- In speed scaling systems, ρ and U differ
- Speed scaling systems stabilize even when $\rho > 1$
- In stable speed scaling systems, $s = \rho$ (an invariant)
- PS is amenable to analysis; SRPT is not
- PS with linear speed scaling behaves like $M/M/\infty$, which has Poisson distribution for system occupancy
- Increasing α changes the Poisson structure of PS
- At high load, $N \rightarrow \rho^{\alpha}$ (another invariant property)

PS Modeling Results



SRPT Simulation Results

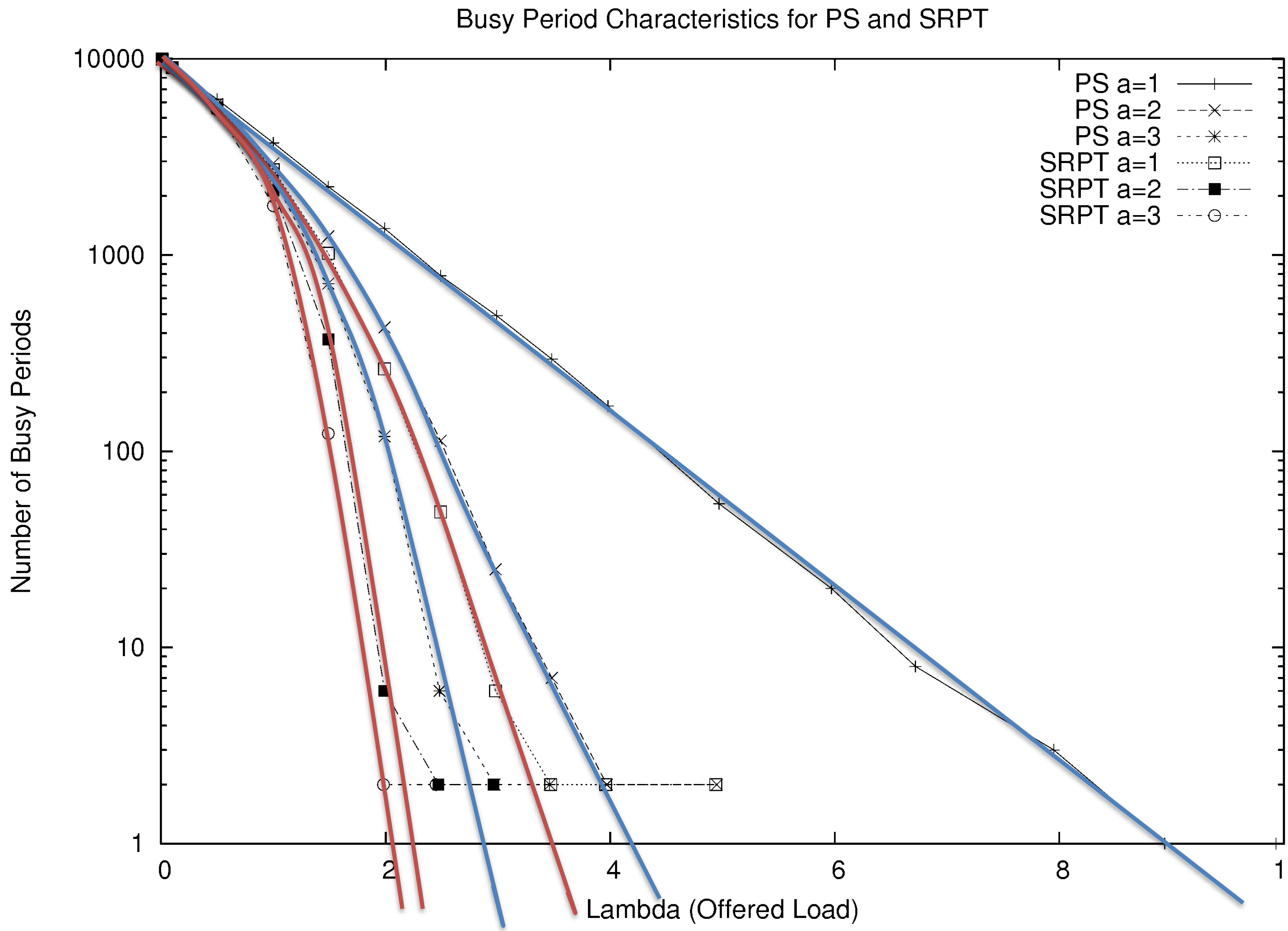


Comparing PS and SRPT

- Similarities:
  - Mean system speed (invariant property)
  - Mean system occupancy (invariant property)
  - Effect of α (i.e., the shift, the squish, and the squeeze)

- Differences:
  - Variance of system occupancy (SRPT is lower)
  - Mean response time (SRPT is lower)
  - Variance of response time (SRPT is higher)
  - PS is always fair; SRPT is unfair (esp. with speed scaling!)
  - Compensation effect in PS
  - Procrastination/starvation effect in SRPT

Busy Period Structure for PS and SRPT (simulation)



Simulation Insights and Observations

- Under heavy load, busy periods coalesce and U → 1
- Saturation points for PS and SRPT are <u>different</u>
  - Different "overload regimes" for PS and SRPT
  - Gap always exists between them
  - Gap shrinks as α increases
  - Limiting case ($\alpha = \infty$) requires $\rho < 1$ (i.e., fixed rate)
- SRPT suffers from starvation under very high load
- "Job count" stability and "work" stability differ

Conclusions

- The autoscaling properties of dynamic speed scaling systems are many, varied, and interesting!
  - Autoscaling effect: stable even at very high offered load ($s = \rho$)
  - Saturation effect: $U \rightarrow 1$ at heavy load, with $N \rightarrow \rho^{\alpha}$
  - The α effect: the shift, the squish, and the squeeze
- Invariant properties are helpful for analysis
- Differences exist between PS and SRPT
  - Variance of system occupancy; mean/variance of response time
  - Saturation points for PS and SRPT are different
  - SRPT suffers from starvation under very high load
- Our results suggest that PS becomes superior to SRPT for coupled speed scaling, if the load is high enough