On Queueing Systems with Finite Arrivals

Oualid Jouini
(Laboratoire Génie Industriel, Ecole Centrale Paris)

Joint work with **Rowan Wang** and **Saif Benjaafar**
(University of Minnesota)

YEQT IV, Eindhoven, November 26th 2010

Finite arrivals

- Queueing systems with a finite number of arrivals

- Arrivals occur over a period of time followed by few or no arrivals for an extended period thereafter

- Primary concern is the customer waiting time

<u>Examples</u>

- Boarding for scheduled flights
- Stadium checking for sporting event
- Concerts
- Restaurants during peak time
- Arrival of customers to a movie theater
- Arrivals of patients to a health care facility
- …

Finite arrivals

- Finite source of customers
- Heterogeneous inter-arrival times
- Heterogeneous service times

Research questions

- How does these characteristics affects performance?
- Is it efficient to simply use a standard queueing analysis (infinite number of customers)?

Related literature

- Hu and Benjaafer (2009)
  - Queueing system during rush hour, customers arrive all at once
  - Expected waiting time
- Hassin and Mendel (2008), Jouini and Benjaafar (2010)
  - Single server, appointment-driven arrivals with no-shows and non-punctuality
  - Optimal schedule to minimize the cost
- Parlar and Sharafali (2008)
  - Multiple servers, single queue
  - Optimal staffing level

- Our contribution:
  - Customer specific inter-arrival and service times
  - Multiserver setting
  - Accounting for these particular features is important

Single server model

- Single server
- single queue, FCFS discipline of service
- Finite population size: M customers
- Customer m (n=1..M)
  - arrives after a duration T_m after the arrival of customer m-1. The T_ms are independent
  - needs an exponential service time with rate $\mu_{n(m)}$

X: Waiting time in queue of an arbitrary customer ?

Analysis

- R_m : Number of customers found in the system by customer *m*
- We have a Markov chain at the arrival epoch of customer *m*
- System state probabilities
  - Probability $p_{m,i}$ = probability that customer *m* (*m*=1..*M*) finds *i* (*i*=0…*m*-1) customers at the epoch of her arrival
  - We recursively compute $p_{m,i}$ by relating it to $p_{m-1,j}$

$$p_{m,i} = \sum_{j=i-1}^{m-2} p_{m-1,j} \Pr\{R_m = i \mid R_{m-1} = j\}$$

Analysis

- The expected waiting time

$$E(X) = \frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_{n_l}}$$

- The cdf of the waiting time

$$\Pr\{X \le t\} = \frac{1}{M} + \sum_{m=2}^{M}\frac{1}{M}\left(p_{m,0} + \sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1} p_{m,i}\left(\prod_{r=m-i,\, r\neq l}^{m-1}\frac{\mu_{n_r}}{\mu_{n_r}-\mu_{n_l}}\right)\left(1-e^{-\mu_{n_l} t}\right)\right)$$

- Multiserver case (i.i.d. service times)

$$\Pr\{X < t\} = 1 - \frac{1}{M}\sum_{m=s+1}^{M}\sum_{i=s}^{m-1}\sum_{l=0}^{i-s} p_{m,i}\frac{(s\mu t)^l}{l!}e^{-s\mu t}$$

Special case: Exponential inter-arrival times

- A singe server and a multiple classes of customers

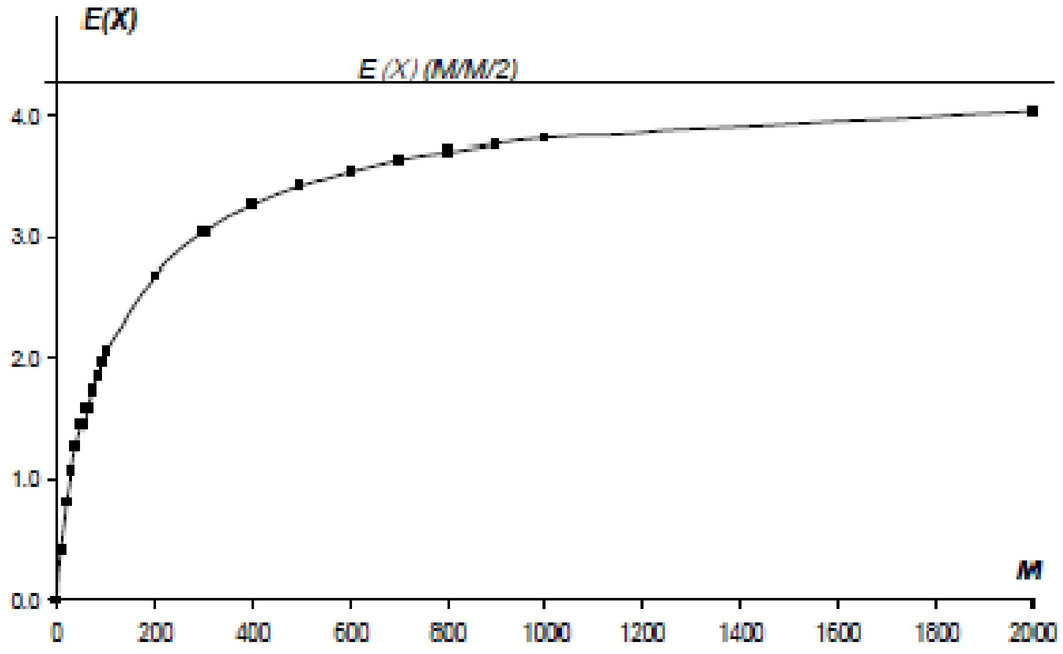
$$\Pr\{R_m = i \mid R_{m-1} = j\} = \left(\prod_{l=i+1}^{j+1} \frac{\mu_{n_{m-l}}}{\mu_{n_{m-l}} + \lambda_m}\right) \frac{\lambda_m}{\mu_{n_{m-i}} + \lambda_m}$$

- Multiple servers and a single class of customers

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \left(\prod_{l=i+1}^{j+1} \frac{\mu \min(l,s)}{\mu \min(l,s) + \lambda_m}\right) \frac{\lambda_m}{\mu \min(i,s) + \lambda_m}$$

Numerical experiments

- The effect of number of arrivals (homogeneous inter-arrival and service times)



Numerical experiments

- The effect of heterogeneous arrivals

  Model 4.8: the first and last 25% of the customers: $\lambda m = 3/4\lambda$. Otherwise it is $1/4\lambda$

  Model 3.1: $\lambda m = \lambda/2$ for all m (Approximation model)

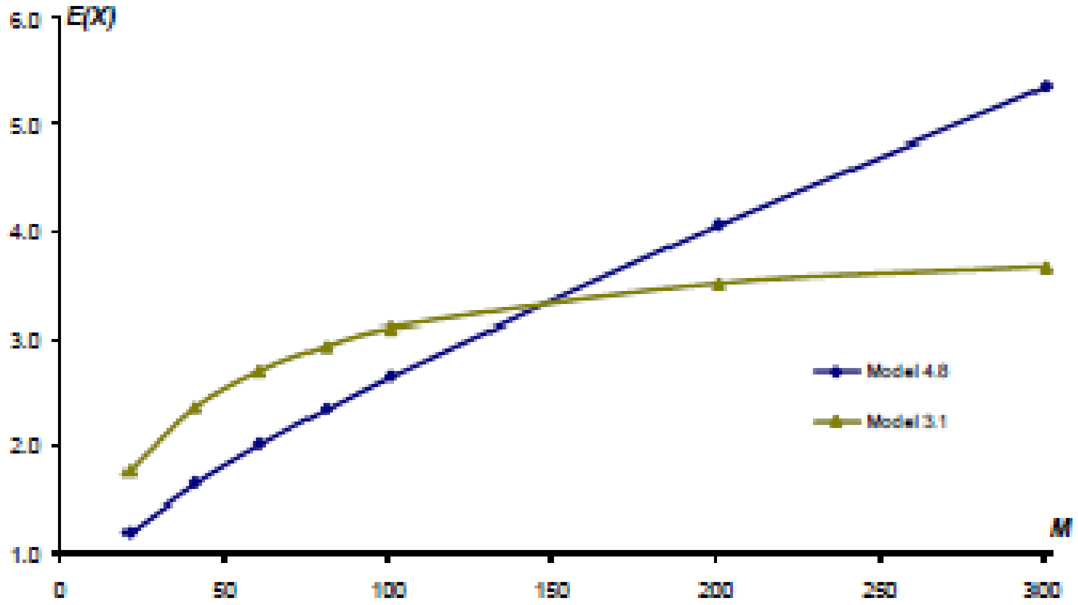


Figure 8: Heterogeneous Arrival, Single-server Comparison 4 ($\lambda = 1.6, \mu = 1$)

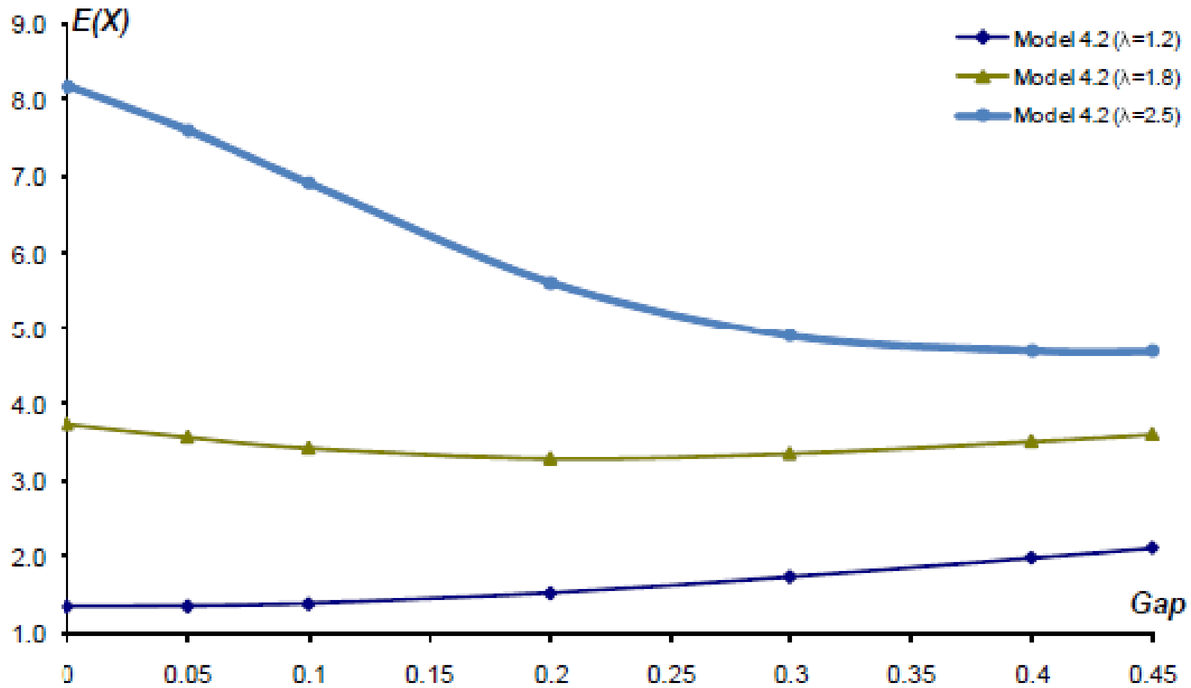
- Significant gap between the two models
- Variability does not always deteriorate perforamance

Numerical experiments

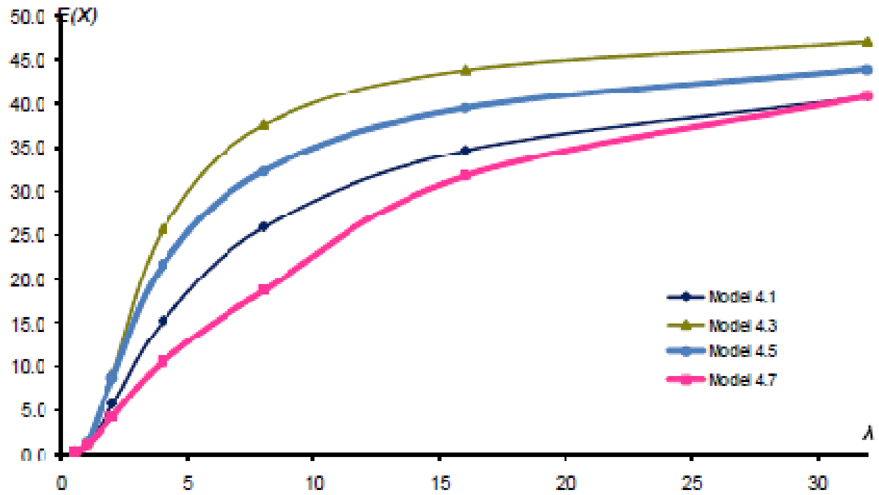
- The effect of heterogeneous arrivals

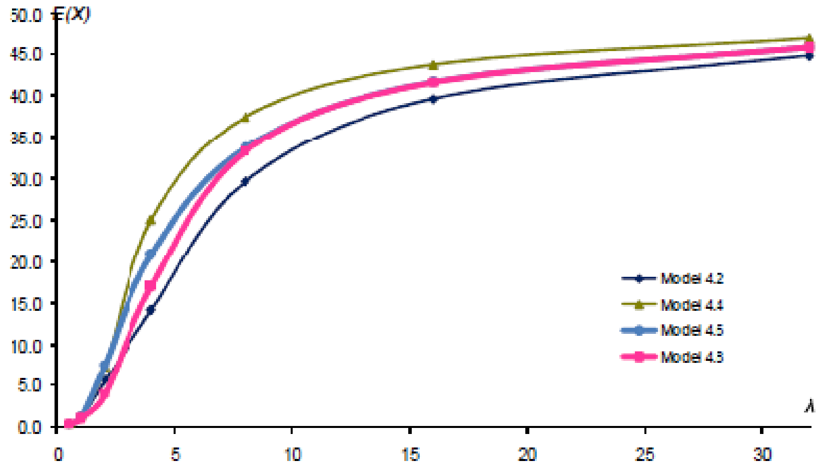
  Model 4.2: GAP= arrival rate of the first half of the customers – that of the second one

  Model 3.1: $\lambda m = \lambda/2$ for all m (Approximation model), GAP=0



Numerical experiments

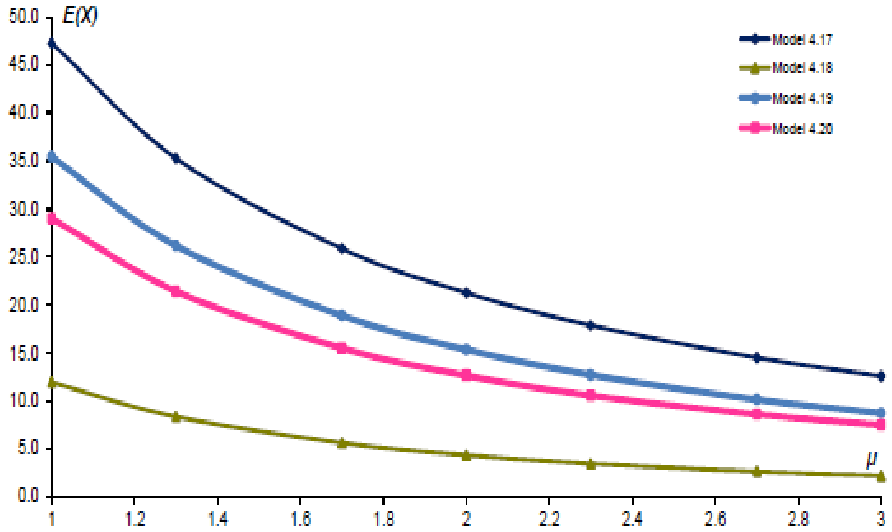


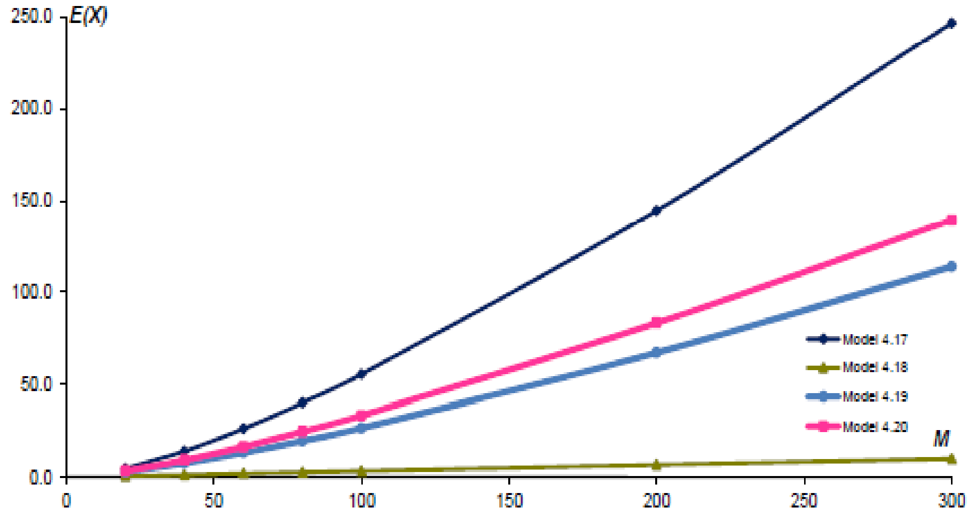


- Worst performance are for models with decreasing arrival patterns
- Scheduling of customers affects performance

Numerical experiments

- Effect of the service process





- It is better to schedule the fastest job first

Conclusions

- Difficulty (Difference from traditional queueing analysis)
  - The number of arrivals is finite
  - Non-homogeneous inter-arrival and service times
- Performance analysis
- Analysis of the impact of these particular features on performance

Thank you !