Performance Modelling of A Hybrid Scheduling Scheme under Self-Similar Network Traffic

Xiaolong Jin and Geyong Min

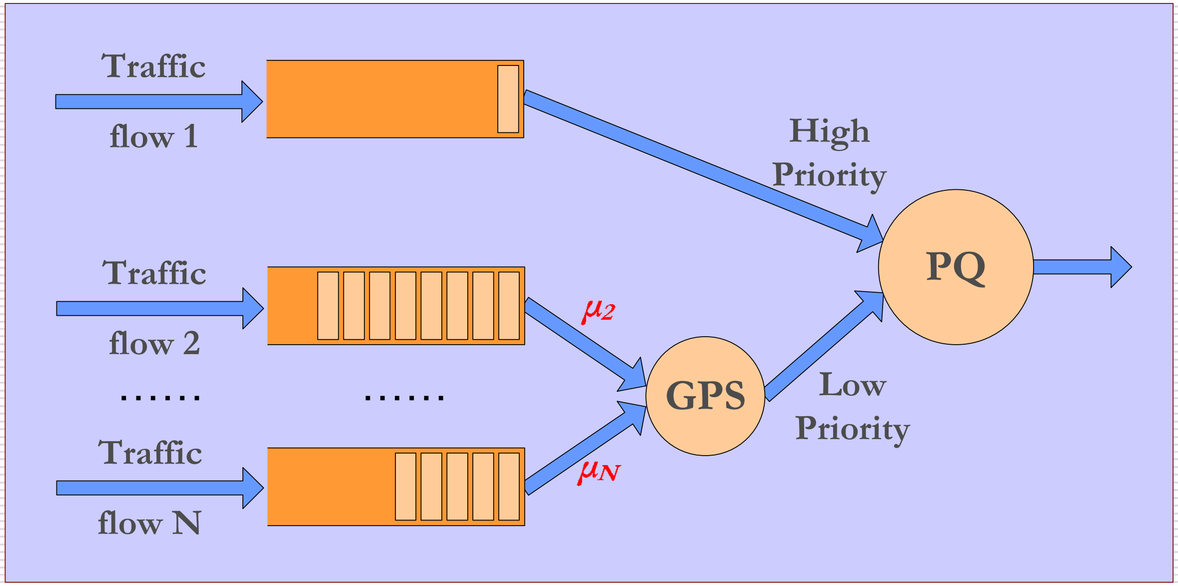
Department of Computing
University of Bradford
U.K.
{x.jin, g.min}@brad.ac.uk

Outline

- Introduction
  - The hybrid PQ-GPS scheduling mechanism; Self-similar traffic; Our motivations
- Queueing systems with self-similar traffic
  - Self-similar traffic modelling
  - Total queue length distribution
- Flow-decomposition of the PQ-GPS system
- Model validation and performance analysis
- Conclusions

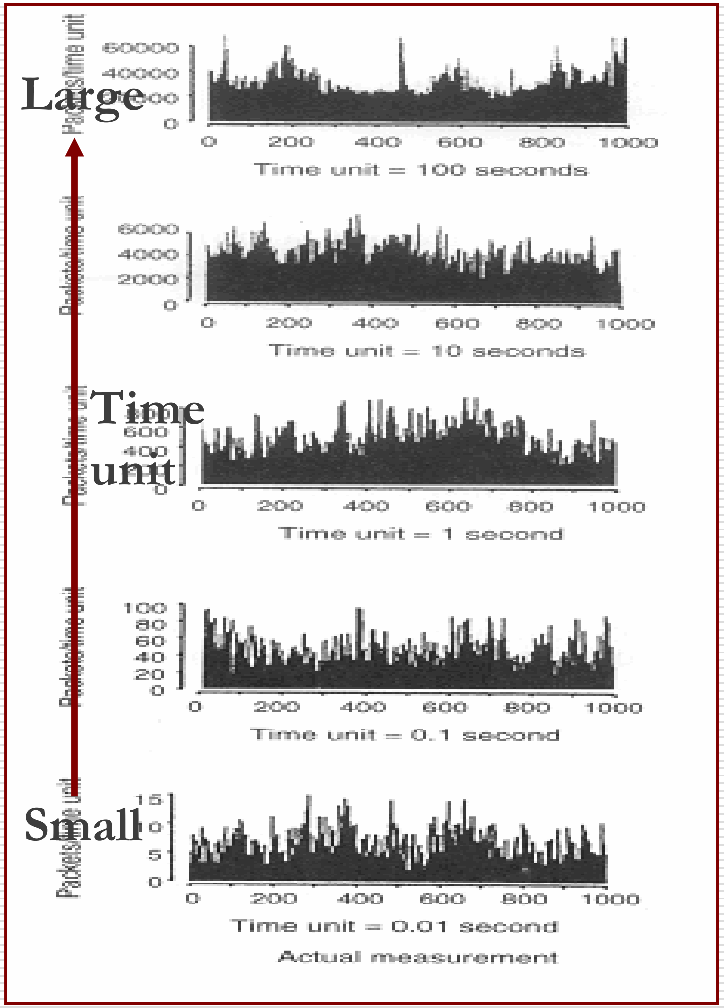
Introduction

- Traditional best-effort service can no longer meet the QoS requirements of advanced multimedia applications
- The provisioning and implementation of differentiated QoS has emerged as a pressing demand
- Hybrid PQ-GPS Scheduling mechanism: A promising scheduling policy for supporting QoS differentiation
  - Priority Queueing (PQ)
  - Generalized Processor Sharing (GPS)



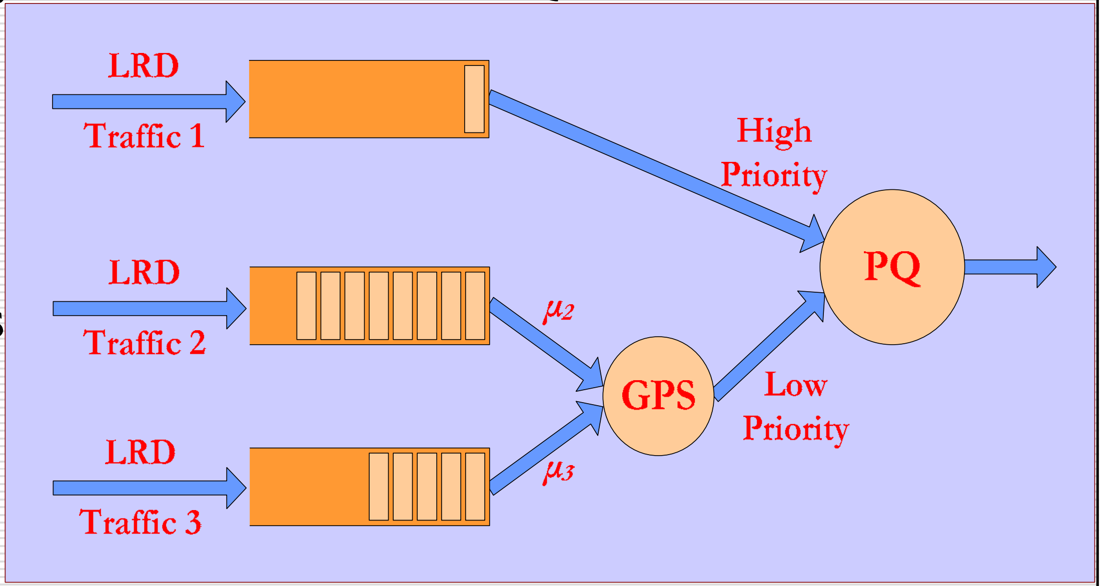
Introduction (cont.)

- Self-similar traffic
  - Being ubiquitous
    - LAN, WLAN, WWW traffic
    - TCP, FTP, and TELNET traffic
    - VBR video
    - Ad-Hoc network traffic
  - Features
    - Scale-invariant burstiness
    - Large-lag correlation



Introduction (cont.)

- Research problem
  - How to analytically address the hybrid PQ-GPS scheduling mechanism under self-similar traffic?
- Motivations
  - To develop a novel analytical model for the PQ-GPS scheduling mechanism subject to multi-class self-similar traffic
  - To derive the analytical upper and lower bounds for the queue length distributions of individual traffic flows



Modelling self-similar traffic

- fractional Brownian motion (fBm): An efficient way for modelling and generating self-similar traffic
- Formulation: $A_i(t) = m_i t + \sqrt{a_i m_i}\bar{Z}_i(t)$
  - $A_i(t)$: The cumulative amount of fBm traffic flow i arriving up to time t
  - m_i: Mean arrival rate; a_i: variance coefficient
- Variance: $v_i(t) = a_i m_i t^{2H}$, H: Hurst parameter
- fBm traffic is Gaussian in essence

Total queue length distribution

- Queueing systems with self-similar traffic
  - Total queue: $Q(t) = \sup_{s \le t} \left\{ \sum_{i=1}^{N} A_i(s,t) - C(t-s) \right\}$
  - An existing approach*: Based on the large deviation principle; For Gaussian traffic
  - Upper and lower bounds:

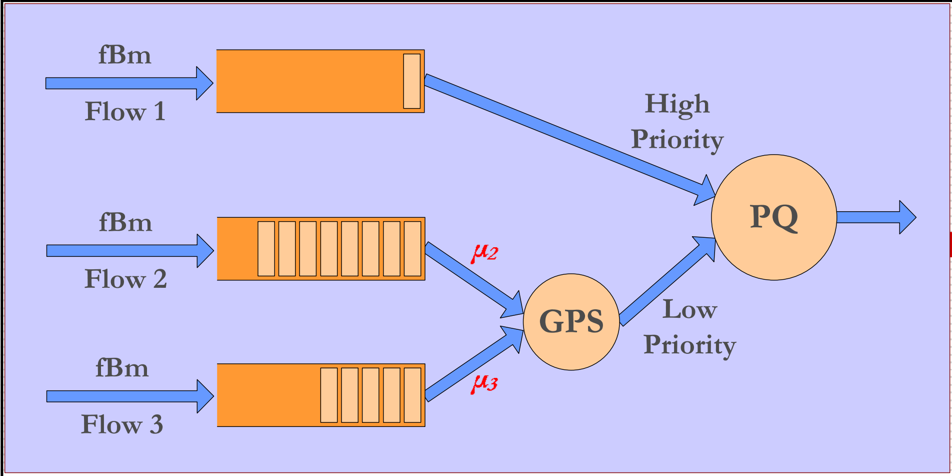
$$\frac{\exp(-\frac{1}{2}U(t_x))}{\sqrt{2\pi(1+\sqrt{U(t_x)})^2}} \le P(Q > x) \le \exp\left(-\frac{1}{2}U(t_x)\right)$$

where $U(t) = \dfrac{\left(-x + (C - \sum_{i=1}^{N} m_i)t\right)^2}{\sum_{i=1}^{N} a_i m_i t^{2H}}$

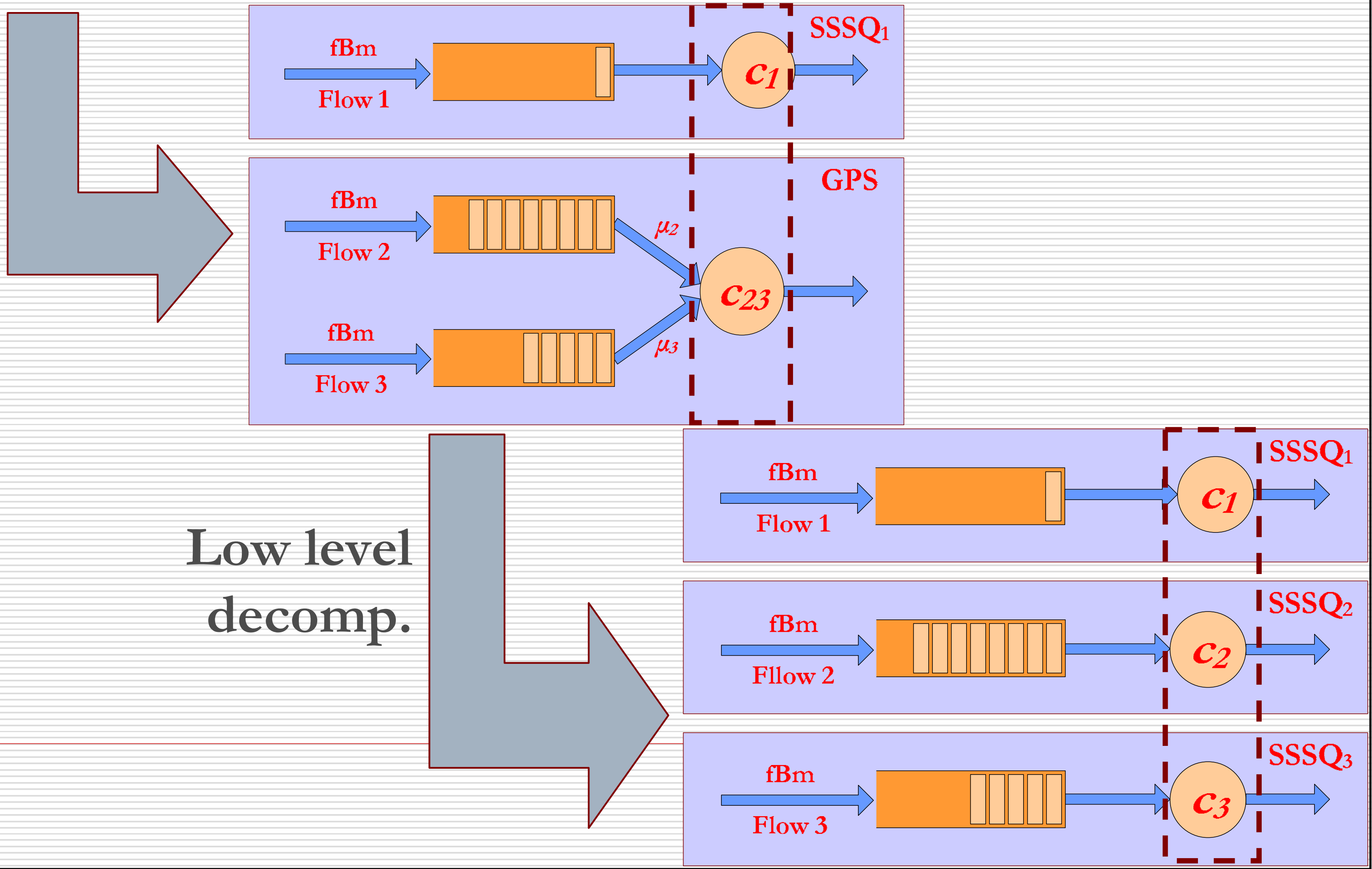
is the determinative function and $t_x = \arg\min_t U(t)$

* P. Mannersalo and I. Norros. A most probable path approach to queueing systems with general Gaussian input. Computer Networks, 40(3):399–412, 2002.

Flow decomposition

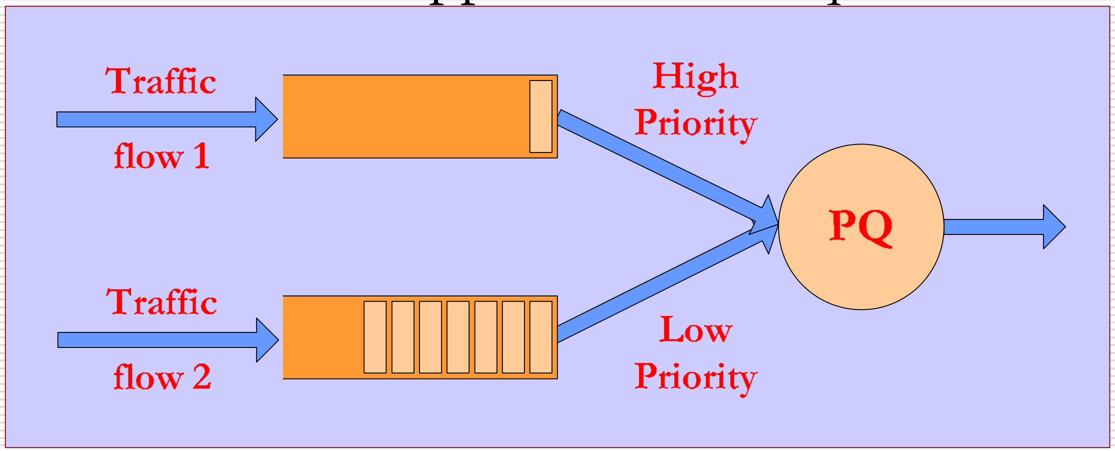


High level decomp.



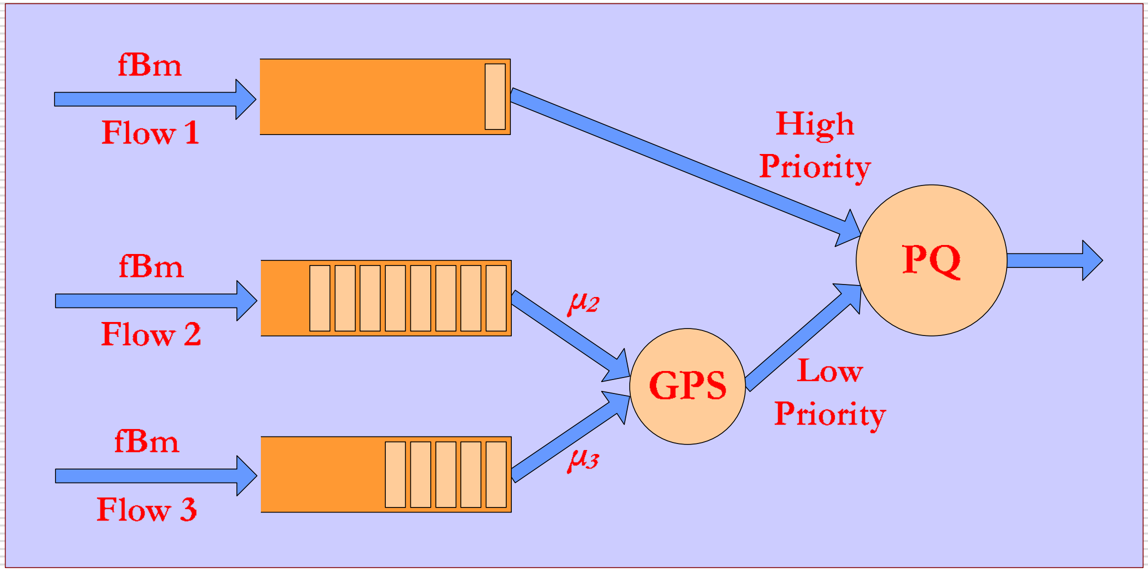
Flow decomposition at the high level

- Service capacity c_1 of $SSSQ_1$
  - fBm 1 is served with the strict high priority. Therefore, the service capacity of $SSSQ_1$ as $c_1 = C$
- Empty Buffer Approximation (EBA)
  - The total queue in a two-class priority system is almost exclusively composed of the low priority traffic
  - Its total queue length can be used to approximate the queue length of the low priority traffic



Flow decomposition at the high level (Cont.)

- The PQ-GPS and GPS systems
  - The aggregate traffic flows (fBm 2 and fBm 3) of the GPS system are served with the low priority. Therefore,
  - The total queue length of the PQ-GPS system can be utilised to approximate that of the GPS system



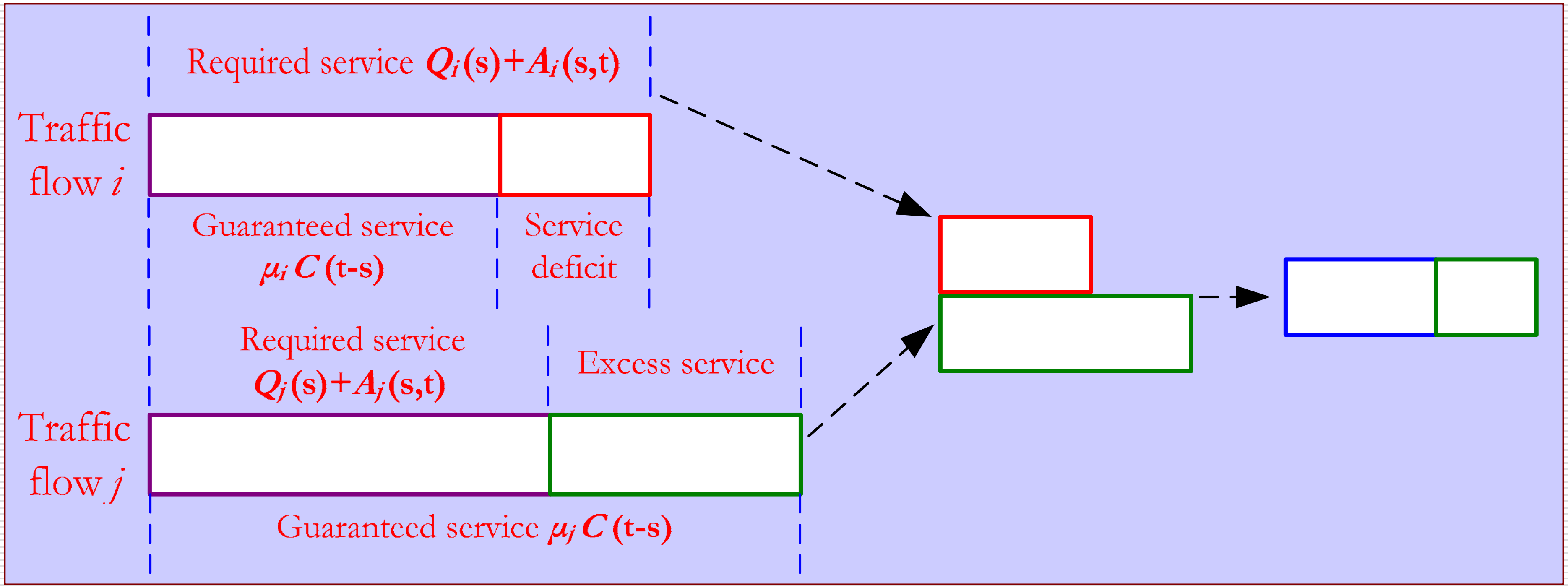
Flow decomposition at the high level (Cont.)

- The service capacity of the GPS system
  - Idea: The minimum values of the determinative functions corresponding to the the PQ-GPS system and the GPS system should be close to each other
  - PQ-GPS: $U(t) = \frac{\left(-x+\left(C-\sum_{i=1}^{3} m_i\right)t\right)^2}{\sum_{i=1}^{3} a_i m_i t^{2H}}$
  - GPS: $U_{gps}(g) = \frac{(-x+(c_{23}-m_2-m_3)g)^2}{(a_2 m_2+a_3 m_3)g^{2H}}$
  - By setting $\min_t U(t) = \min_g U_{gps}(g)$, we obtain

$$c_{23} = (m_2 + m_3) + \left(C - \sum_{i=1}^{3} m_i\right)\left(\frac{a_2 m_2 + a_3 m_3}{\sum_{i=1}^{3} a_i m_i}\right)^{\frac{1}{2H}}$$

Flow decomposition at the low level

- Excess service sharing in time interval (s, t]
  - One flow is guaranteed excess service, while the guaranteed service of the other is less than the mean arrival rate (Situation II)
  - Both flows are guaranteed an excess service rate (Situation I)



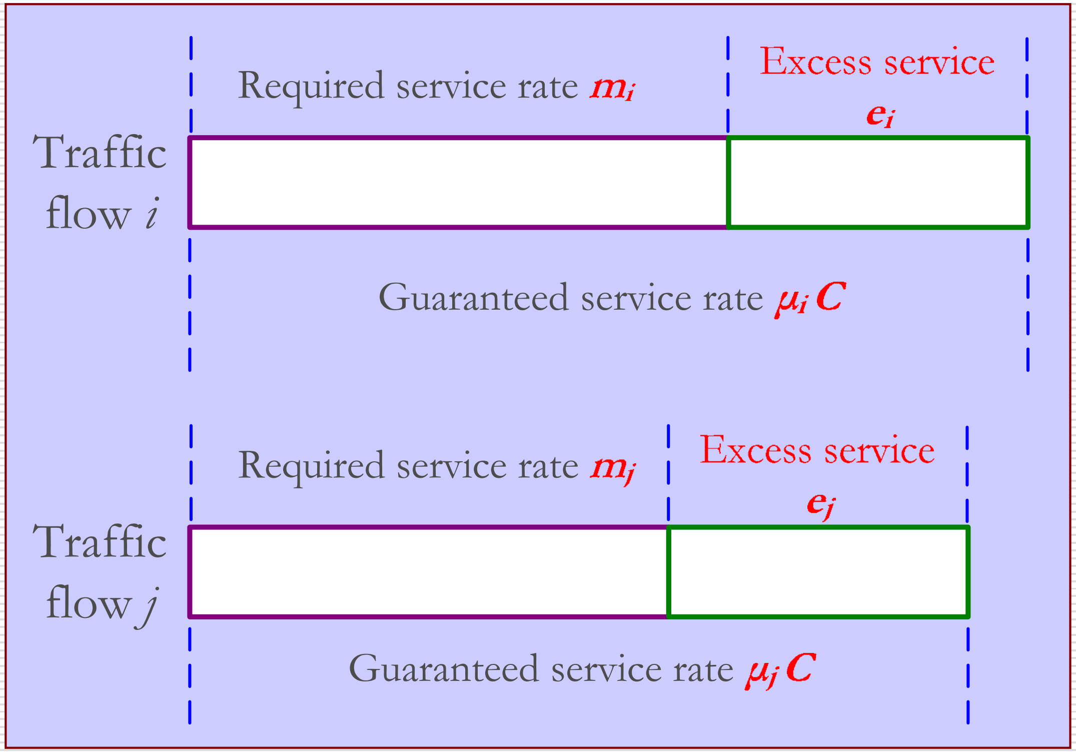
Flow decomposition at the low level (Cont.)

- Let $e_i = \mu_i C - m_i,\ i \in \{2,3\}$
  - If $e_i \geq 0$, indicating an excess service rate
  - If $e_i < 0$, indicating a service deficit
- Service capacities of SSQ_2 and SSQ_3 ?

  $c_i = \mu_i C + \hat{c}_i$ and $c_j = \mu_j C + \hat{c}_j;\quad i,j \in \{2,3\}$
  - $\hat{c}_i$ and $\hat{c}_j$: service rates taken by flows *i* and *j* from each other

Flow decomposition at the low level (Cont.)

- Situation I: $e_i \geq 0$ and $e_j \geq 0$

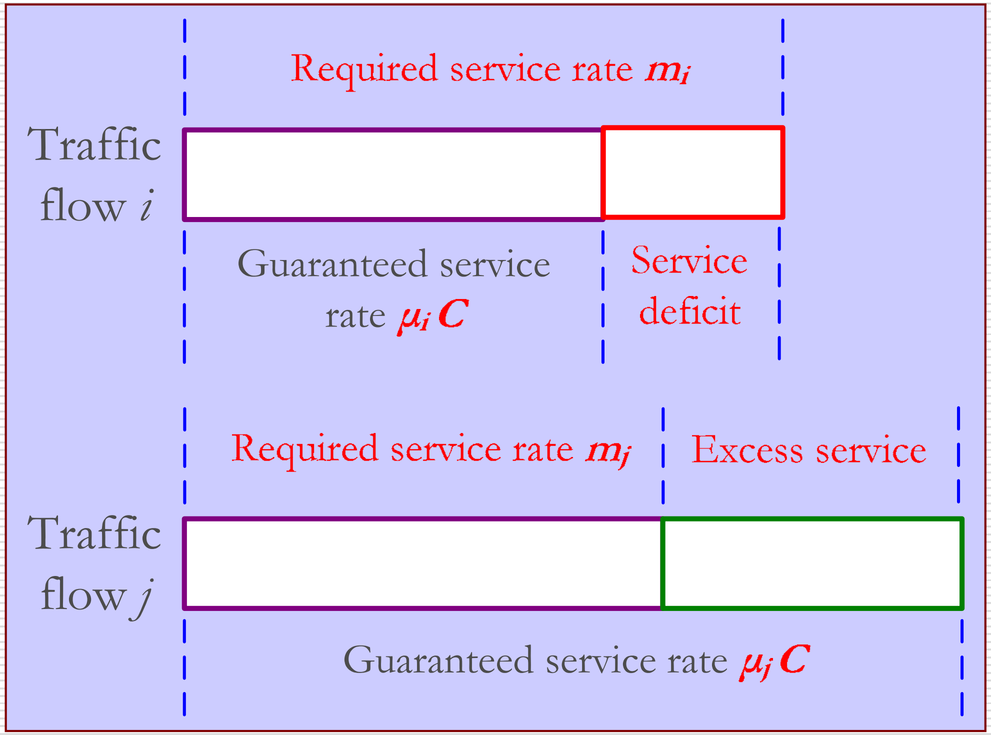


$$\hat{c}_i = e_j \left(\frac{a_i m_i}{a_i m_i + a_j m_j} \right)^{\frac{1}{2H}}$$

$$\hat{c}_j = e_i \left(\frac{a_j m_j}{a_i m_i + a_j m_j} \right)^{\frac{1}{2H}}$$

Flow decomposition at the low level (Cont.)

- Situation II: $e_i < 0$ and $e_j > 0$



$$\hat{c}_i = -e_i + (e_j + e_i) \times \left(\frac{a_i m_i}{a_i m_i + a_j m_j} \right)^{\frac{1}{2H}}$$

$$\hat{c}_j = 0$$

Flow decomposition at the low level (Cont.)

- Service capacities of SSQ_2 and SSQ_3

  $c_i = \mu_i C + \hat{c}_i$ and $c_j = \mu_j C + \hat{c}_j$

  - Situation I: $e_i \geq 0$ and $e_j \geq 0$

$$c_i = \mu_i C + (\mu_j C - m_j)\left(\frac{a_i m_i}{a_i m_i + a_j m_j}\right)^{\frac{1}{2H}}$$

$$c_j = \mu_j C + (\mu_i C - m_i)\left(\frac{a_j m_j}{a_i m_i + a_j m_j}\right)^{\frac{1}{2H}}$$

  - Situation II: $e_i < 0$ and $e_j > 0$

$$c_i = m_i + (C - m_i - m_j)\left(\frac{a_i m_i}{a_i m_i + a_j m_j}\right)^{\frac{1}{2H}}$$

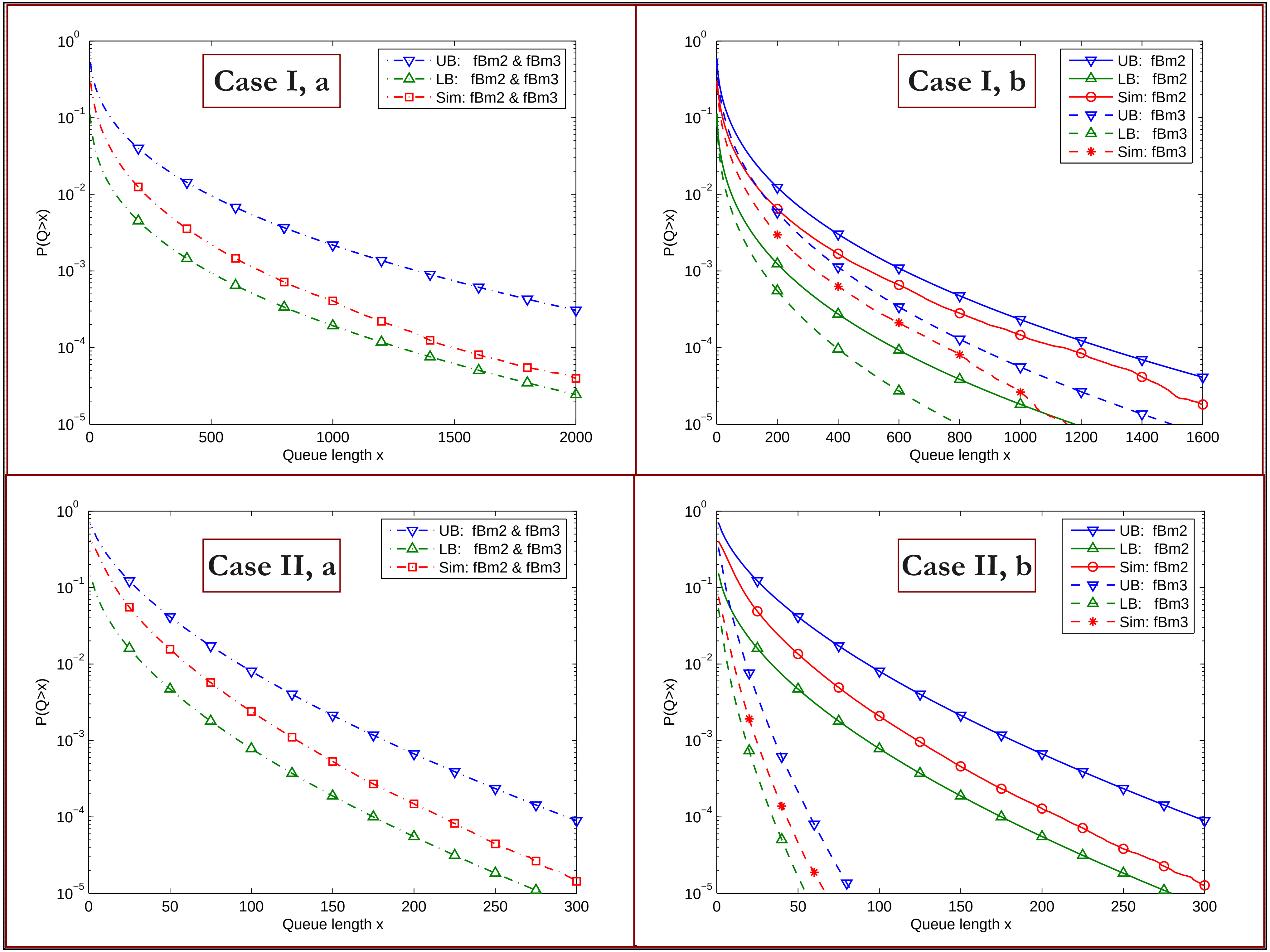
$$c_j = \mu_j C$$

Performance results and model validation

- Methodology
  - Comparing analytical and simulation results of queue length distributions
- Two typical scenarios

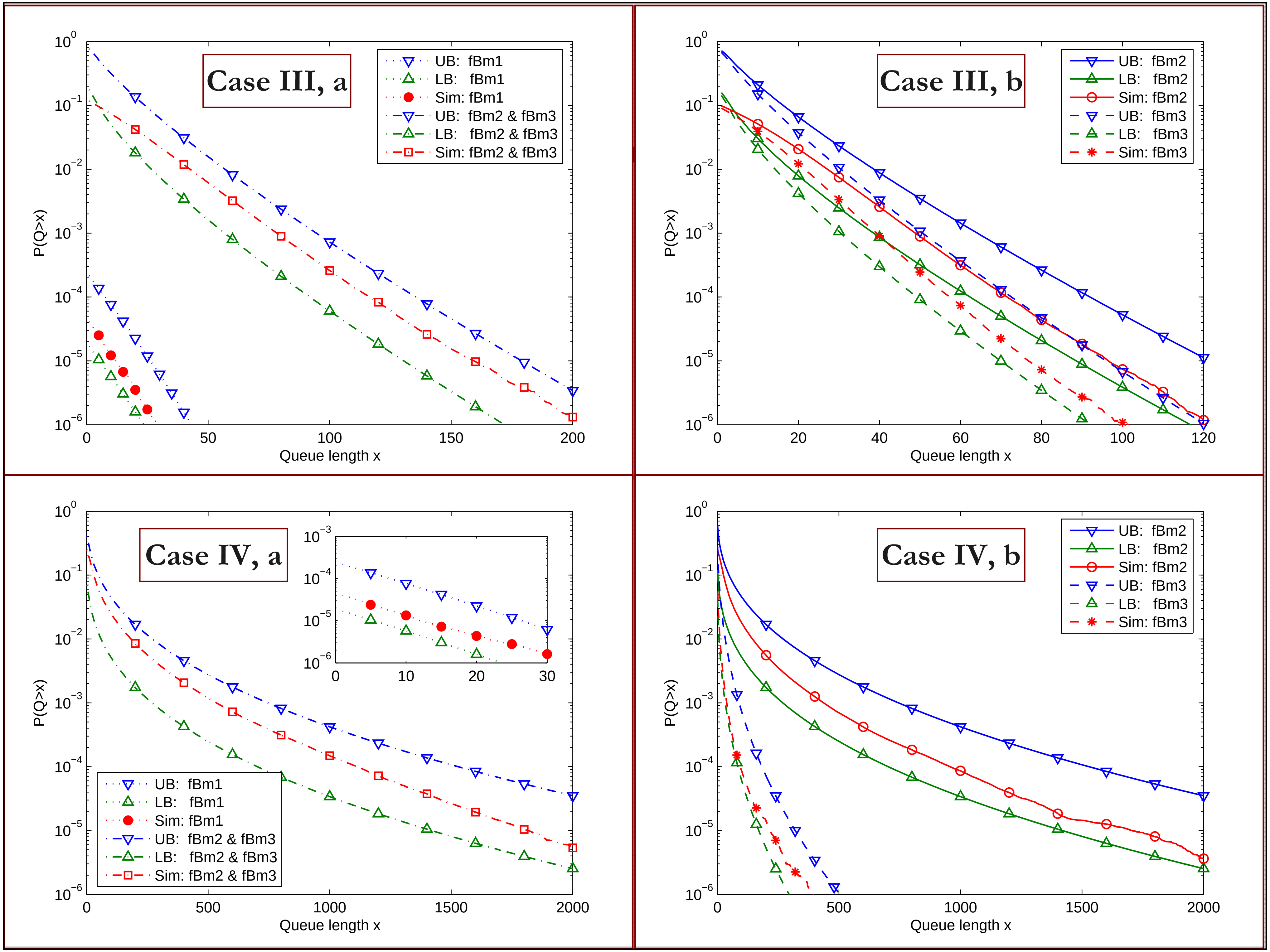
Scenario	Case	C	H	fBm_1	fBm_2		fBm_3	
				m_1	m_2	μ_2	m_3	μ_3
A	I	150	0.8	50	50	0.55	40	0.45
	II	150	0.7	40	55	0.50	45	0.50
B	III	1500	0.6	1350	50	0.55	40	0.45
	IV	1500	0.8	1350	55	0.50	45	0.50

  - Cases I and III: Situation I; Cases II and IV: Situation II
  - Scenario A: fBm 2 and fBm 3 dominate the input of the system
  - Scenario B: fBm 1 dominates the input


Case I, a
UB: fBm2 & fBm3
LB: fBm2 & fBm3
Sim: fBm2 & fBm3
P(Q>x)
Queue length x
Case I, b
UB: fBm2
LB: fBm2
Sim: fBm2
UB: fBm3
LB: fBm3
Sim: fBm3
P(Q>x)
Queue length x
Case II, a
UB: fBm2 & fBm3
LB: fBm2 & fBm3
Sim: fBm2 & fBm3
P(Q>x)
Queue length x
Case II, b
UB: fBm2
LB: fBm2
Sim: fBm2
UB: fBm3
LB: fBm3
Sim: fBm3
P(Q>x)
Queue length x

Observations on Scenario A

- The simulation results of fBm traffic flows as well as the GPS system are well situated within the scopes between their corresponding analytical bounds
- No curves corresponding to fBm 1 are presented
  - Small arrival rate → empty buffer of fBm 1 → no empirical curves
- In Case I, the scope between the fBm 2 bounds overlaps that of fBm 3. In Case II the two scopes are clearly separated
  - The difference on the ratio of the mean arrival rate of fBm 2 to its guaranteed service rate is close to the ratio of fBm 3



Case III, a
UB: fBm1
LB: fBm1
Sim: fBm1
UB: fBm2 & fBm3
LB: fBm2 & fBm3
Sim: fBm2 & fBm3
P(Q>x)
Queue length x
Case III, b
UB: fBm2
LB: fBm2
Sim: fBm2
UB: fBm3
LB: fBm3
Sim: fBm3
P(Q>x)
Queue length x
Case IV, a
UB: fBm1
LB: fBm1
Sim: fBm1
UB: fBm2 & fBm3
LB: fBm2 & fBm3
Sim: fBm2 & fBm3
P(Q>x)
Queue length x
Case IV, b
UB: fBm2
LB: fBm2
Sim: fBm2
UB: fBm3
LB: fBm3
Sim: fBm3
P(Q>x)
Queue length x

Observations on Scenario B

- The simulation results of fBm traffic flows are also well situated in the scopes between their analytical bounds
- The bounds and simulation results of fBm 1 are presented due to its relatively large arrival rate
- The performance results of fBm 1 are under those of fBm 2 and fBm3
  - Reason: it is served with the high priority, although its mean arrival rate is much larger than those of fBm 2 and fBm 3

Conclusions

- Proposed a novel and cost-efficient flow-decomposition approach to analytically modelling the PQ-GPS system under self-similar traffic
  - Equivalently decomposed the PQ-GPS system into individual SSSQ systems and derived their service capacities
- Derived the analytical upper and lower bounds of the queue length distributions for individual traffic flows
- Validated the effectiveness of the proposed flow-decomposition approach and the accuracy of the analytical model

The End!

Q. & A.