

Queuing Theory Systems Analysis in Wireless Networks

Mobile Stations with Non-Preemptive Priority

Bakary Sylla

**Senior Systems Design Engineer
Radio Access Network
T-Mobile Inc. USA**

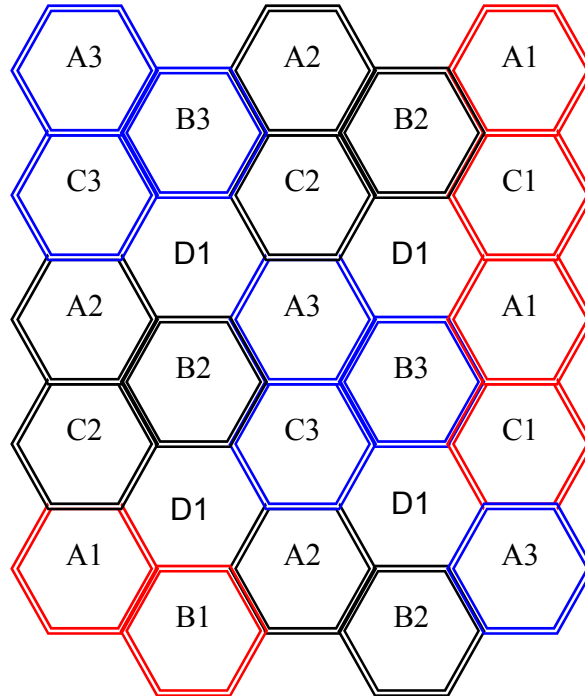
&

**Southern Methodist University
Department of Operations Research**
<http://engr.smu.edu/emis>

INTRODUCTION

When the mobile network is congested new mobile call attempts are discarded. Due to regulations in some countries it is not possible to pre-empt an existing call. Letting the subscriber redial randomly does not guarantee any success in a busy network. Thus the need to implement a queuing mechanism in a cellular network arises. The queuing mechanism is intended to provide service to selected classes of subscribers using a queue management scheme during congestion.

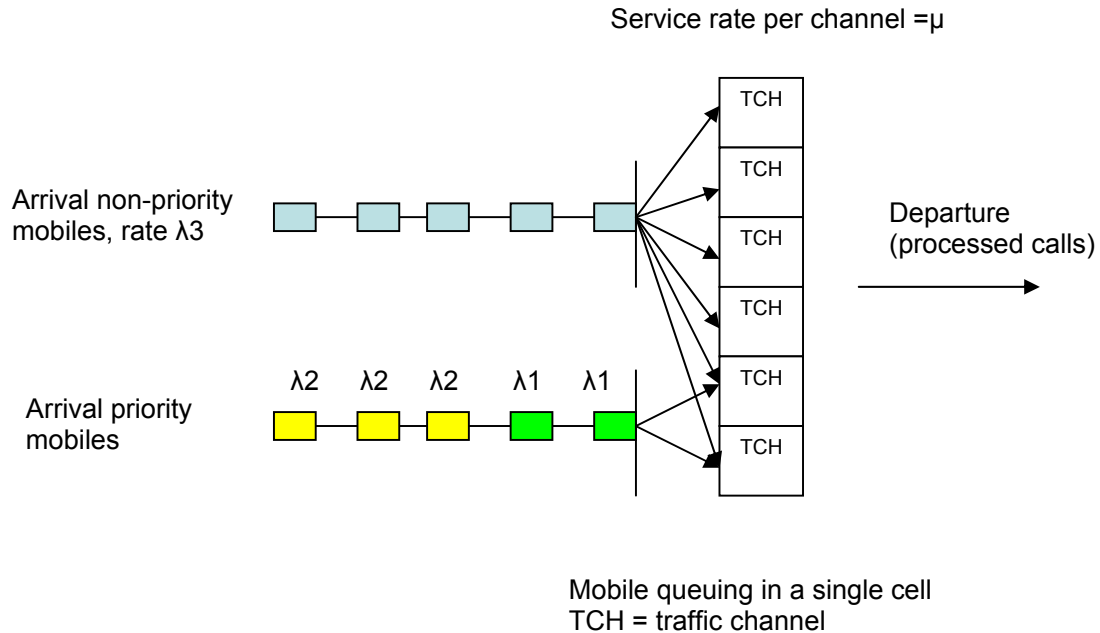
Queuing System Description



Example cellular network
configuration
Ai, Bi, Ci, Di = cell names

A mobile network divides a geographic area into cells. Each cell has a dedicated specific transmitter/receiver frequency. Adjacent cells have different frequency assigned to them. In a digital mobile network the frequency is divided into channels for traffic and signaling. A service is identified by allocating a traffic channel to an incoming subscriber's call. The service time is the duration of the call. In order to increase the probability that a given subscriber will get service it is queued in the system when the call attempt is received. Different priority levels are assigned to different categories of subscribers.

Definition:



- **Arrival:** An arrival is an incoming call. Only priority calls are considered. A non-priority can only access channels reserved for priority calls when they are idle.
- **Server:** A server is an allocated traffic channel. At congestion, traffic channels will be occupied by other prior calls with priority as well as with non-priority. Several servers can exist in the same cell. Several cells can exist in a given geographic area.
- **Departure:** Processed calls are pushed out of the system.

Queue Management:

- A fixed percentage of available traffic channels in the cell is reserved for priority calls.
- The length of the queue is set for each cell
- Calls are queued based on the priority level
- When a queue is full new call attempts are rejected

- A high priority call will remove a lower priority call from the queue
- For the same priority level a FIFO selection is adopted
- When a call queue time expires it is removed from the queue.

Queuing System Assumptions

A simplified queuing systems is analysed based on the following assumptions

1. Only two classes of nonpreemptive priority are considered in this project
 2. The number of servers or traffic channels allocated for priority calls is set to 2
 3. The two servers operate independently
 3. The call processing time or service rate μ is identical for both servers
 2. The arrival rate for each class of priority is a Poisson process
 3. The service time in each traffic channel has an exponential distribution
 4. Service time are identicals for each traffic channel or server
 6. Let m be the number of class 1 priorities mobiles in the system (highest priority)
 7. Let n be the number of class 2 priorities mobiles in the system (lowest priority)
 8. Let i be the number of class 1 priority mobiles in service ($i= 0,1,2$)
 9. Let j be the number of class 2 priority mobiles in service ($j= 0,1,2$)
- $p_{mni j} = \text{Pr}(\text{in steady state, } m \text{ mobiles of class 1 in the system, } n \text{ mobiles units of class 2 in the system, } i \text{ mobiles of class 1 in service, } j \text{ mobiles of class 2 in service})$

Transition States

The possible probability transition states of class 1 and class 2 mobiles in the system is given by the following 3 tables:

Note: $p_{mni j} =$ in steady state, m mobiles of class 1 in the system, n mobiles units of class 2 in the system, i mobiles of class 1 in service, j mobiles of class 2 in service

m	n	i	j		m	n	i	J		m	n	i	J
1	0	0	0		0	1	0	0		0	2	0	0
1	0	0	1		0	1	0	1		0	2	0	1
1	0	0	2		0	1	0	2		0	2	0	2
1	0	1	0		0	1	1	0		0	2	1	0
1	0	1	1		0	1	1	1		0	2	1	1
1	0	1	2		0	1	1	2		0	2	1	2
1	0	2	0		0	1	2	0		0	2	2	0
1	0	2	1		0	1	2	1		0	2	2	1
1	0	2	2		0	1	2	2		0	2	2	2

m	n	i	j		m	n	i	j		m	n	i	J
1	0	0	0		1	1	0	0		1	2	0	0
1	0	0	1		1	1	0	1		1	2	0	1
1	0	0	2		1	1	0	2		1	2	0	2
1	0	1	0		1	1	1	0		1	2	1	0
1	0	1	1		1	1	1	1		1	2	1	1
1	0	1	2		1	1	1	2		1	2	1	2
1	0	2	0		1	1	2	0		1	2	2	0
1	0	2	1		1	1	2	1		1	2	2	1
1	0	2	2		1	1	2	2		1	2	2	2

m	n	i	j		m	n	i	j		m	n	i	j
2	0	0	0		2	1	0	0		2	2	0	0
2	0	0	1		2	1	0	1		2	2	0	1
2	0	0	2		2	1	0	2		2	2	0	2
2	0	1	0		2	1	1	0		2	2	1	0
2	0	1	1		2	1	1	1		2	2	1	1
2	0	1	2		2	1	1	2		2	2	1	2
2	0	2	0		2	1	2	0		2	2	2	0
2	0	2	1		2	1	2	1		2	2	2	1
2	0	2	2		2	1	2	2		2	2	2	2

In order to minimize the number of stationary equations the following assumptions are valid:

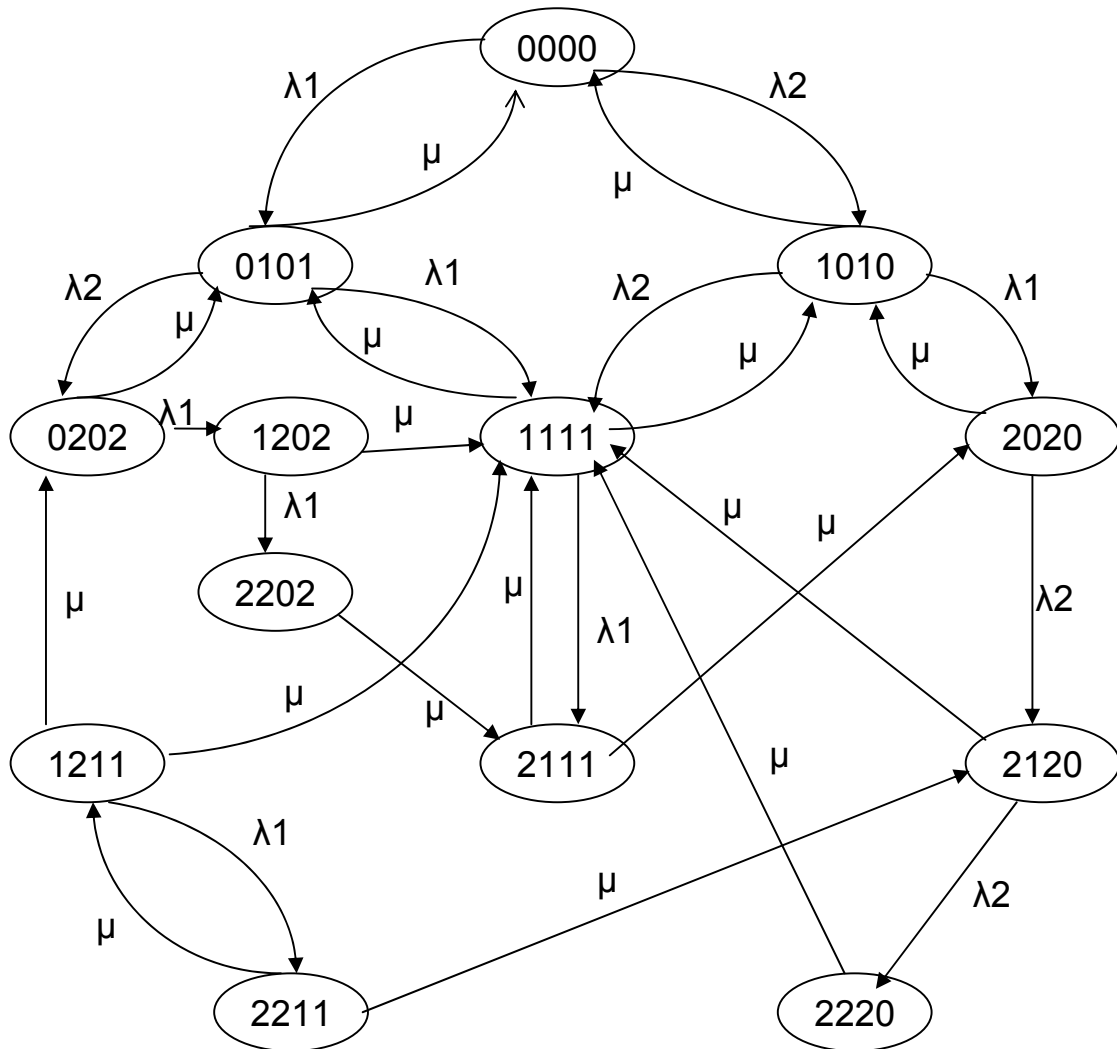
1. The maximum number of allowed mobiles in the system at any given time is 3
2. The maximum number of class 1 priority mobiles in the system at any given time is 2 ($m= 0, 1, 2$)
3. The maximum number of class 2 priority mobiles in the system at any given time is 2 ($n= 0, 1, 2$)
4. Due to traffic channel limitation, the maximum number of mobiles in service is 2 ($i,j = 0, 1, 2$)
5. Class 1 mobiles have higher priority than class 2 mobiles
6. FCFS priority is applied within the same class of priority
7. Channels are idle only if there is no incoming call.
8. Service rate is μ
9. Arrival rate for class 1 priority is λ_1
10. Arrival rate for class 2 priority is λ_2

These assumptions lead to the simplified transition states below:

m n i j = in steady state, m mobiles of class 1 in the system, n mobiles units of class 2 in the system, i mobiles of class 1 in service, j mobiles of class 2 in service

m	n	i	j
0	0	0	0
0	1	0	1
0	2	0	2
1	0	1	0
1	1	1	1
1	2	1	1
2	0	2	0
2	1	1	1
2	1	2	0
2	2	0	2
2	2	1	1
2	2	2	0

State Diagram



State diagram

λ_1 = arrival rate class 1 priority mobiles
 λ_2 = arrival rate class 2 priority mobiles
 μ = call duration time (service rate)

Difference Equations

To assure that the queue will not grow forever a steady state condition is assumed. Thus server utilization is strictly less than 1 ($\rho = \lambda/2\mu < 1$). The following difference equations are derived:

- (1) $\lambda p_{0000} = \mu(p_{0101} + p_{1010})$
- (2) $(\lambda + \mu)p_{0101} = \lambda p_{0000} + \mu(p_{0202} + p_{1111})$
- (3) $(\lambda + \mu)p_{1010} = \lambda p_{0000} + \mu(p_{1111} + p_{2020})$
- (4) $(\lambda + \mu)p_{0202} = \lambda p_{0101} + \mu p_{1211}$
- (5) $(\lambda + 2\mu)p_{1111} = \lambda p_{0101} + \lambda p_{1010} + \mu(p_{1202} + p_{1211} + p_{2111} + p_{2220} + p_{2120})$
- (6) $(\lambda + \mu)p_{2020} = \lambda p_{1010} + \mu p_{2111}$
- (7) $(\lambda + \mu)p_{1202} = \lambda p_{0202}$
- (8) $\mu p_{2202} = \lambda p_{1202}$
- (9) $(\lambda + 2\mu)p_{1211} = \mu p_{2211}$
- (10) $2\mu p_{2111} = \lambda p_{1111} + \mu p_{2202}$
- (11) $(\lambda + \mu)p_{2120} = \lambda p_{2020} + \mu p_{2211}$
- (12) $2\mu p_{2211} = \lambda p_{1211}$
- (13) $\mu p_{2220} = \lambda p_{2120}$
- (14) $\sum p_{mnij} = 1$

Numerical Analysis

Solving the system of difference equations in the previous section is quite tedious. For the purpose of a simplified numerical analysis the earlier assumptions are dropped. New assumptions are made which allow direct use of formulas for non-preemptive priority multiple channels queuing systems to derive system characteristics. The following new assumptions apply:

1. There is no limit on the number of allowed mobiles in the system
2. No restriction on the number of class 1 priority mobiles in the system
3. No restriction on the number of class 2 priority mobiles in the system
4. There are 2 servers (traffic channels) available for priority
5. Class 1 mobiles have higher priority than class 2 mobiles

6. FCFS priority is applied within the same class of priority
7. Channels are idle only if there is no incoming call.
8. Service time is 4 min, service rate $\mu = 1/4 \text{ min} = (1/15) \text{ hour}$
9. Arrival rate for class 1 priority is $\lambda_1 = 15/\text{hour}$
10. Arrival rate for class 2 priority is $\lambda_2 = 10/\text{hour}$

This is a 2-server Markov model with 2 priority classes. The model is solved using QTS software:

System characteristics are given in the table below:

INPUT VARIABLES:

lam1	15.	
lam2	10.	
st	0.066667	Mean time to complete service
c	2	Number of servers in the system

OUTPUT VARIABLES:

lambda	25.0	Overall arrival rate (#/time)
iat	0.04	Mean interarrival time
mu	15.0	Service rate (# served/unit of time)
r	1.666667	Average # arrivals during mean service time
rho	0.833333	Fraction of time each server is busy [MUST BE < 1]
p0	0.090909	Fraction of time the server is idle
Lq	3.787879	Expected queue size
L	5.454545	Expected system size
Wq	0.151515	Expected waiting time in the queue
W	0.218182	Expected waiting time in the system

PRIORITY CLASS 1

W1	0.117172	Expected time in the system
Wq1	0.050505	Expected waiting time in the queue
L1	1.757576	Expected number in the system
Lq1	0.757576	Expected number in the queue

PRIORITY CLASS 2

W2	0.369697	Expected time in the system
Wq2	0.30303	Expected waiting time in the queue
L2	3.69697	Expected number in the system
Lq2	3.030303	Expected number in the queue

References

1. Donald Gross and Carl M. Harris
Fundamentals of Queuing Theory
2. Xinyu Chen and Michael R. Lyu
Message Queuing Analysis in Wireless Networks with Mobile Station Failures and Handoffs
3. MS Queuing Implementation proposal (Ericsson Internal document)
4. Wireless Priority Service Industry Requirements (Ericsson Internal document)