Modelling QoS in IoT Applications

Irfan Awan
University of Bradford
Bradford, UK
e-mail: i.u.awan@bradford.ac.uk

Muhammad Younas
Oxford Brookes University
Oxford, UK
e-mail: m.younas@brookes.ac.uk

Wajia Naveed
University of Bradford
Bradford, UK
e-mail:
W.Naveed@student.bradford.ac.uk

Abstract — **Internet of Things (IoT) aims to enable the interconnection of a large number of smart devices (things) using a combination of networks and computing technologies. But an influx of interconnected things makes a greater demand on the underlying communication networks and affects the quality of service (QoS). This paper investigates into the QoS of delay sensitive things and the corresponding traffic they generate over the network. Things such as security alarms, cameras, etc, generate delay sensitive information that must be communicated in a real time. Such things have heterogeneous features with limited buffer capacity, storage and processing power. Thus the most commonly used Best Effort service model cannot be an attractive mechanism to treat delay sensitive traffic. This paper proposes a cost-effective analytical model for a finite capacity queueing system with pre-emptive resume service priority and push-out buffer management scheme. Based on the analytical model various simulation results are generated in order to analyse the mean queue length and the blocking probability of high and low priority traffic for system with various capacities.**

Keywords: Internet of Things, QoS, Modelling, Traffic, delay sensitive information.

I. INTRODUCTION

The Internet of Things (IoT) creates an Internet-based interconnected platform wherein smart devices (things) have interfaces and identities and can communicate with each other through standard and interoperable communication protocols [1, 2]. Such an interconnected platform of things has the potential to provide various benefits including efficient transportation, increased productivity, reduced energy and power consumption in various domains such as smart cities [3], supply chain management [4], and healthcare [5] to name a few [7].

In addition, IoT has various other benefits and applications. For instance, Yang et al [8] describe the motivation and benefits of the use of IoT technologies in emergency management operations and response times such as fires, floods, earthquakes and so on. Such emergency operations require real time information in order to aid situational awareness and to make appropriate decisions. IoT technology has the potential to meet the demands of emergency operations where information is highly delay-sensitive.

In IoT, things are required to be actively participating in various activities, exchanging information and making intelligent decision without much reliance on human involvement [6]. Smart homes are equipped with various intelligent devices to ensure safety, security and high performance of electrical appliances. These devices are usually connected with personal mobile devices through GSM or GPRS. Mobile devices are increasingly being used for Internet browsing, Internet telephony and electronic shopping in addition to its traditional use of voice calls. Mobile devices are equipped with small buffers but have to deal with IoT services that generate a huge volume of data. Thus an effective buffer management scheme together with an appropriate service scheduling mechanism is essential to ensure instant communication without facing any packet loss or queuing delay. A typical example can be an incident of a break into a smart home. An intelligent CCTV camera takes a picture of an intruder and instantly sends a priority packet together with the photograph of the intruder to the home owner's mobile device. In this perspective, it is necessary to design and develop service models that ensure appropriate level of QoS for delay sensitive applications in IoT.

Current approaches provide unsatisfactory solutions for delay-sensitive applications and ignore the key parameters such as timing, deadline and importance of delays. Such problems become more complicated in IoT where communication media and device resources are scarce. This paper, extending our previous work [14], focuses on modeling IoT applications that are involved in producing and consuming delay sensitive information.

The proposed model aims to analyse a finite capacity queue with push-out buffer management scheme and pre-emptive resume (PR) service priority. The cost-effective exact analytical solution for this queue is used to evaluate the performance of smart devices under varying traffic conditions to ensure preferential treatment of highest priority delay sensitive data. Based on the analytical model, we conduct various simulation experiments using Matlab. The experiments take into account some of important parameters of performance measures such as mean queue length, and blocking probability. The results obtained demonstrate how various parameters of the model influence

978-1-4799-4224-4/14 $31.00 © 2014 IEEE
DOI 10.1109/NBiS.2014.97

the performance measures of two classes of traffic; high priority and low priority

The rest of the paper is structured as follows. Section II reviews the related work. Section III presents the proposed analytical model. Section IV describes the performance evaluation model. Section V presents the implementation of the proposed model and the experimental results. Section VI concludes the paper and identifies future research work.

II. RELATED WORK

IoT has attracted significant attention from research community both in academia and industry. Existing research investigates into different aspects of IoT such as Web of Things [16], SOA-based IoT [15], and provides surveys of applications, issues and solutions related to IoT [7, 17].

IoT deals with all kind of traffic requiring various level of QoS constraints. For example, emergency messages are more sensitive to delay than other types of traffic. Various service scheduling have been proposed ranging from FCFS to Head of Line under different buffer management schemes ranging from dedicated buffer access for each class to the shared buffer with space priority. There is only a small number of studies for analyzing service priorities under push-out buffer management. Mostly such studies have been conducted for ATM networks with fixed size cells.

The authors in [7, 9] present surveys on IoT vision, IoT related projects, application and impact areas, enabling technologies and the research issues such as interoperability, privacy, trust, energy and resource scarcity of things in IoT. These are useful surveys and provide high level description of QoS issues but without proposing any model or solution.

Jin et al [3] present four types of network architectures in a smart city, which is one of the potential areas of IoT. These include autonomous network, ubiquitous network, application-layer overlay network and service-oriented network architectures. These network architectures are compared using factors such as design approach, connectivity model, Network hierarchy In-Network Processing, QoS complexity and progress in defining QoS. Though such comparison is very useful the analysis is very abstract and does not provide any in-depth analysis of the QoS issues. Nef et al [6] analyse WSNs integration approaches in the IoT, which are considered to be the main contributors to the IoT QoS. The paper also presents the feasibility and different (best) ways for integrating WSNs into the IoT. The above approaches do not look into the QoS of delay-sensitive applications in IoT.

Kleoec and Kos in [10] present the behaviour of packet transit times for a delay sensitive application with certain minimum bandwidth constraint. They propose a simple model with two queues under priority and FCFS service rules. Both queues work under a complete sharing scheme. Although this model is simple, the utilization of higher priority depends on the delay sensitive traffic intensity and can result in small queue length for high priority traffic at the cost of high data loss for low priority data under low high priority traffic load. The authors in [11] investigate into the impact of buffering under complete buffer sharing scheme on resource allocation in wireless local area networks under heterogeneous traffic loads. The results indicate that, in the presence of heterogeneous loads, 802.11 does not allocate transmission opportunities equally. They show that large buffers can help this inequality at the expense of significantly increased delay which is mainly due to absence of an effective buffering scheme and service priorities. Guo et al [12] introduced an Awareness Driven Schedule (ADS) scheme that enables sensors to provide differentiated data service by their awareness. The higher a sensor resource's awareness on the event is, the more detailed data service it should provide.

In summary the above approaches do not give particular attention to the performance analysis of delay sensitive applications in IoT.

III. THE PROPOSED MODEL

This paper proposes a $[M]^X/M/1/N$ queuing system with pre-emptive resume (PR) service priority and finite capacity queue with complete buffer sharing scheme by all classes of traffic under a push out mechanism. The proposed model can be used to evaluate the performance of smart devices to meet various QoS constraints under varying input parameterization. In the proposed system, it is assumed that the arriving traffic is classified into low priority (normal traffic) and high priority (emergency traffic). The inter-arrival times and service times for each arriving class of traffic are distributed according to exponential distribution. These are modeled as follows.

Let,

- λ_i be the arrival rate of class i traffic, for i=1,2, where class 1 has the highest priority
- μ_i be the service rate for class i traffic, for i=1,2,
- N be the total buffer capacity under complete buffer sharing with push-out mechanism
- $\mathbf{n} = (n_1, n_2, \ldots, n_R)$ be a joint queue state where $\sum_{i=1}^{R} n_i \leq N$, R=2
- $P(\mathbf{n})$ be the joint state probabilities

Finite capacity buffer is managed under a complete sharing scheme whilst giving highest priority to emergency traffic signals. Upon arrival to a full buffer, a highest priority class traffic packet will push-out the lower priority class packet. Service scheduling follows PR priority discipline in which higher priority traffic is serviced according to FCFS prior to lower priority traffic. Lower priority traffic is served only in absence of higher priority traffic and are immediately pre-empted upon arrival of emergency data packets and could also be pushed out if buffer is full to avoid data loss of the delay sensitive traffic.

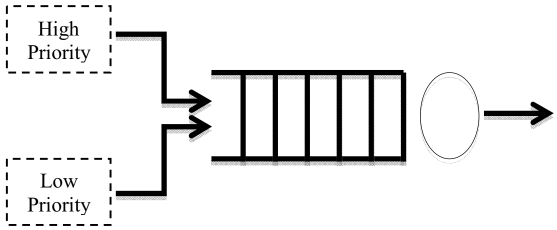

Fig. 1: $[M]^X/M/1/N$ queuing system with two classes of traffic and a shared buffer

Two dimensional Markov chain for $[M]^X/M/1/N$ with PR service priority and Push-out buffer management scheme can be constructed as shown in Fig. 2.

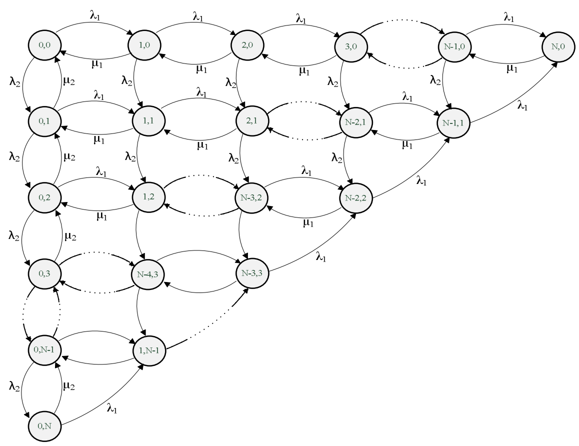

Fig. 2: Markov chain for $[M]^X/M/1/N$ queuing system with PR priority and push-out buffer

The flow balance equations, based on this Markov chain, can be derived as follows:

$$(\lambda_1 + \lambda_2)P_{0,0} = \mu_1 P_{1,0} + \mu_2 P_{0,1} \qquad (1)$$

$$(\lambda_1 + \lambda_2 + \mu_1)P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + \mu_1 P_{i+1,j}, \ \ \forall\, i \geq 1 = 0, \qquad (2)$$
$$where\ P_{i,-1} = 0$$

$$(\lambda_1 + \lambda_2 + \mu_2)P_{0,j} = \lambda_2 P_{0,j-1} + \mu_2 P_{0,j+1} + \mu_1 P_{1,j}, \forall\, j \geq 1, \qquad (3)$$

$$\mu_1 P_{N,0} = \lambda_1 P_{N-1,0} + \lambda_1 P_{N-1,1} \qquad (4)$$

$$(\lambda_1 + \mu_2)P_{0,N} = \lambda_2 P_{0,N-1} \qquad (5)$$

IV. PERFORMNACE EVALUATION

Exact analysis of a complex multi-class queue presented in Section 3 can be evaluated using Matrix Geometric Method [13]. In the proposed model the arrival process follows the Poisson distribution whereas the service times are exponentially distributed. The Matrix Geometric Method is used to solve the stationary state probability for vector state Markov process. The flow balance equations (1)-(5) can be easily solved using the following relation

$$PQ = 0 \qquad (6)$$

where P is a vector of joint state probabilities, represented as

$$P = (P_{0,0}, P_{1,0}, P_{2,0}, \ldots, P_{N,0}, P_{0,1}, \ldots, P_{0,N})$$

and Q is rate matrix represented as

$$Q = \begin{pmatrix} -(\lambda_1+\lambda_2) & \lambda_2 & \lambda_1 & & \\ \mu_1 & -(\lambda_1+\lambda_2+\mu_2) & & \lambda_2 & \vdots \\ \mu_2 & & -(\lambda_1+\lambda_2+\mu_1) & & \\ & \mu_2 & & -(\lambda_1+\lambda_2+\mu_2) & \\ & \mu_1 & 0 & & \vdots \\ & & \mu_1 & & \\ & & & \mu_2 & \\ & & & \mu_1 & \\ & & & & \vdots \end{pmatrix}$$

which can be simplified as

$$Q = \begin{pmatrix} A_0 & \Lambda_0 & & & & \\ M_1 & A_1 & \Lambda_1 & & & \\ & M_2 & A_2 & \Lambda_2 & & \\ & & M_3 & A_3 & & \\ & & & & & \Lambda_{N-1} \\ & & & & M_N & A_N \end{pmatrix} \qquad (7)$$

where

$$A_0 = -(\lambda_1 + \lambda_2)$$

$$A_1 = \begin{pmatrix} -(\lambda 1 + \lambda 2 + \mu 2) & \\ & -(\lambda_1 + \lambda_2 + \mu_1) \end{pmatrix}$$

$$A_2 = \begin{pmatrix} -(\lambda 1 + \lambda 2 + \mu 2) & & \\ & -(\lambda_1 + \lambda_2 + \mu_1) & \\ & & -(\lambda_1 + \lambda_2 + \mu_1) \end{pmatrix}$$

….

$$\Lambda_0 = (\lambda_2 \quad \lambda_1)$$

$$\Lambda_1 = \begin{pmatrix} \lambda_2 & \lambda_1 & \\ & \lambda_2 & \lambda_1 \end{pmatrix}$$

$$\Lambda_2 = \begin{pmatrix} \lambda_2 & \lambda_1 & \\ & \lambda_2 & \lambda_1 \\ & & \lambda_2 \end{pmatrix}$$

…

$$\mathrm{M}_1 = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\mathrm{M}_2 = \begin{pmatrix} \mu_2 & \\ \mu_1 & 0 \\ & \mu_1 \end{pmatrix}$$

$$\mathrm{M}_3 = \begin{pmatrix} \mu2 & & \\ \mu_1 & & \\ & \mu_1 & 0 \\ & & \mu_1 \end{pmatrix}$$

...

Using equations (6) and (7), we can derive the following expressions for steady state probabilities:

$$P_N A_N = P_{N-1}\Lambda_{N-1}$$

which can written as

$$P_N = P_{N-1}\Lambda_{N-1}A_N^{-1} \tag{8}$$

Similarly

$$P_{N-1}A_{N-1} = P_{N-2}\Lambda_{N-2} + P_N M_N$$

using equation (8), the above expressions can be written as

$$P_{N-1} = P_{N-2}\Lambda_{N-2}(A_{N-1} - \Lambda_{N-1}A_N^{-1}M_N)^{-1} \tag{9}$$

$$P_{N-2} = P_{N-3}\Lambda_{N-3}(A_{N-2} - \Lambda_{N-2}(A_{N-1} - \Lambda_{N-1}A_N^{-1}M_N)^{-1}M_{N-1})^{-1} \tag{10}$$

and so on.

Let

$$S_0 = \Lambda_{N-1}A_N^{-1}$$

Substituting above in equation (8), it yields

$$P_N = P_{N-1}S_0 \tag{11}$$

Similarly let

$$S_1 = \Lambda_{N-2}(A_{N-1} - S_0 M_N)^{-1}$$
$$S_2 = \Lambda_{N-3}(A_{N-2} - S_1 M_{N-1})^{-1}$$
$$S_3 = \Lambda_{N-3}(A_{N-3} - S_2 M_{N-2})^{-1}$$
…
…
$$S_n = \Lambda_{N-1-n}(A_{N-n} - S_{n-1}M_{N-n+1})^{-1}$$

Using these recursive relationships, state probabilities can be expressed as:

$$P_N = P_{N-1}S_0$$
$$P_{N-1} = P_{N-2}S_1$$
$$P_{N-2} = P_{N-3}S_2$$
…
…
$$P_1 = P_0 S_{N-1}$$

These can further be represented as recursive relationships as follows:

$$P_1 = P_0 S_{N-1}$$

$$P_2 = P_0 S_{N-1}S_{N-2}$$

$$P_3 = P_0 S_{N-1}S_{N-2}S_{N-3}$$
……
$$P_N = P_0 S_{N-1}S_{N-2}S_{N-3} \dots S_1$$

Finally the above expressions can be represented

$$P_N = P_0 \prod_{i=0}^{N-1} S_{N-1-i} \tag{11}$$

Using these steady state probabilities, various performance measures can be easily derived such as loss probability, mean waiting times and mean queue length occupancy

V. IMPLEMENTATION RESULTS

This section evaluates the performance of the proposed model presented in Section 4. The model is implemented in Matlab and some important parameters of performance measures such as mean queue length and blocking probability are taken into account. The graphs are presented for two different scenarios of traffic to demonstrate how various parameters of the models influence the performance measures of two classes of traffic; high priority and low priority.

The graphs are presented for system capacity, N=10, and N=15, where other input parameters include:

- λ_1 is the arrival rate of high priority traffic
- λ_2 is the arrival rate of low priority traffic

- μ_1 is the service rate of high priority traffic
- μ_2 is the service rate of low priority traffic.

In the experiments, the following values of input parameters are used:

- Arrival rate for high priority traffic, $\lambda 1$=0:0.5:23.5
- Arrival rate for low priority traffic, $\lambda 2$=5
- Service rate for high priority and low priority traffic, $\mu 1 = \mu 2$=10.

Set 1 Experiments: Mean Queue Length

The effect of increasing $\lambda 1$, on Mean Queue length of high and low priority traffic for system capacity N=10 and 15 is shown below.

For N=10

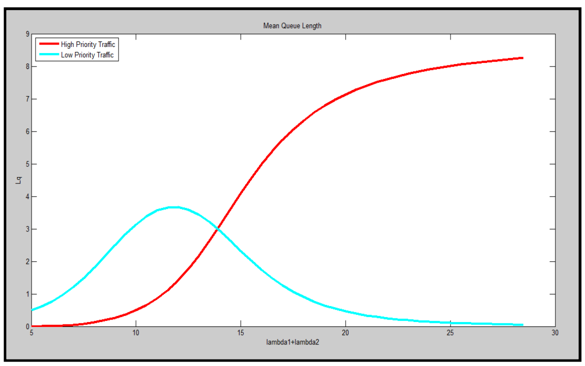

Fig. 3. Effect of increasing λ_1 on Mean Queue length of high and low priority traffic, N=10.

For N=15

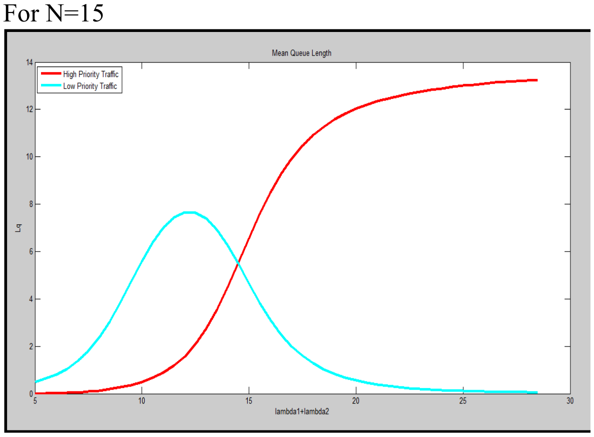

Fig. 4. Effect of increasing λ_1 on Mean Queue length of high and low priority traffic, N=15.

Observations:

Fig. 3 and Fig. 4 demonstrate the effect of increasing $\lambda 1$ on Mean Queue length for high and low priority traffic. These illustrate that: (i) Mean queue length for high priority traffic increases with an increase in $\lambda 1$. It continues to increase until it approaches the buffer capacity. After which it remains steady as it cannot exceed the buffer capacity. (ii) The mean queue length for low priority traffic rises in the beginning and then it begins to fall. This initial rise in mean queue length for the low priority traffic is due to the fact that lower priority traffic is served only in the absence of higher priority traffic and is immediately pre-empted upon arrival of emergency data packets. Whilst there is an increased volume of high priority traffic in the system, chances for lower priority traffic to be served diminishes. (iii) As the mean number in queue for high priority traffic increases, the graph of the mean queue length for the low priority traffic starts to decline. This is as a result of push-out mechanism where higher priority traffic could push-out the lower priority traffic upon arrival to a full buffer to avoid data loss of delay sensitive traffic.

Set 2 Experiments: Blocking Probability

The effect of increasing λ_1, on blocking probability of high and low priority traffic for system capacity N=10 and N=15 is shown below.

For N=10

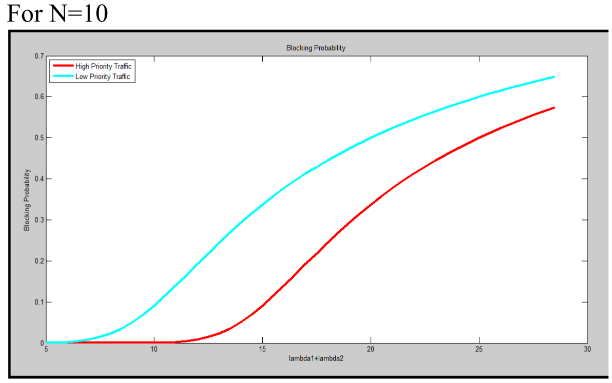

Fig. 5. Effect of increasing λ_1 on Blocking probability of high and low priority traffic, N=10.

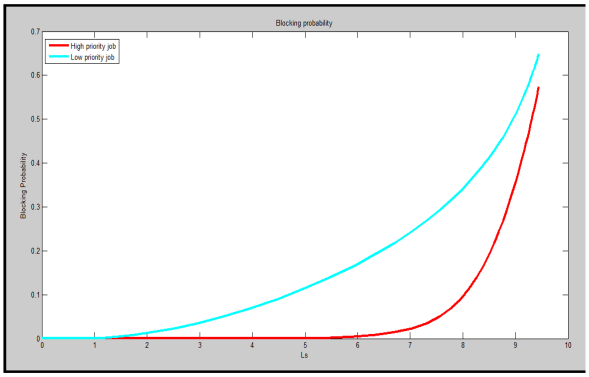

Fig. 6.: Graph between Ls (mean number in system) and Blocking Probability for high priority and low priority traffic.

For N=15

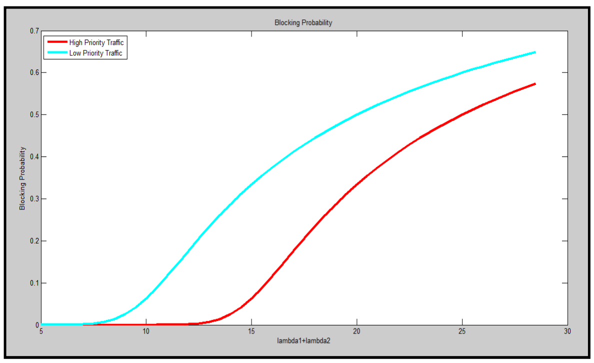

Fig. 7. Effect of increasing λ_1 on Blocking Probability of high priority and low priority traffic, N=15.

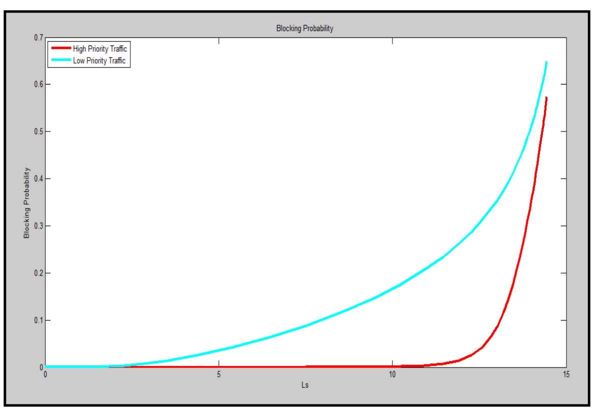

Fig.8. Graph between Ls (mean number in system) and Blocking Probability for high and low priority traffic.

Observations:

Fig. 5 and Fig. 6 illustrate the effect of an increasing arrival rate $\lambda 1$, on the blocking probability for two classes of traffic. It is observed that: (i) the blocking probability for low priority traffic starts to ascend earlier as compared to high priority traffic; this is due to the preferential treatment provided to the delay sensitive traffic. Where emergency data packets shall continue its arrival by pushing out low priority traffic when buffer is full. (ii) When the arrival rate for high priority traffic is less than the service rate, then the blocking probability is minimal for high priority traffic. (iii) The blocking probability for low priority traffic starts to increase as a sum of arrival rate (λ_1 +λ_2) approaches the service rate.

Fig. 7 and Fig. 8 present the results between the mean number in system and blocking probability. It could be noticed that: (i) The blocking probability of the high priority traffic is approximately zero in the beginning, but as average number of traffic in the system approaches the system capacity, blocking probability for high priority traffic increases at a very steep rate. As system capacity is increased, higher volume of traffic is absorbed by the large buffer capacity which reduces the probability of blocking high priority traffic in the beginning. It could cause an abrupt rise in the blocking probability for high priority traffic. (iii) The blocking probability for low priority traffic is always greater than high priority traffic. As high priority traffic will only be blocked if there are N (system capacity) high priority jobs in the system. Whereas low priority traffic will be blocked when there are N (system capacity) high or low priority jobs in the system. (iv) When a system has a large buffer capacity, it will take more time to occupy full buffer capacity. This will lead to reduced blocking probability.

VI. CONCLUSION

This paper investigated into the QoS for transmission of delay sensitive information in the IoT. It specifically examined that the most commonly used Best Effort service model cannot provide a proper solution for delay sensitive traffic. There exist a variety of devices in IoT; each having varying but limited capacity for storing, processing and exchanging information with other devices. In this paper we therefore take account of the characteristics of IoT devices and the importance of delay sensitive information. We proposed a cost-effective analytical model for a finite capacity queuing system with pre-emptive resume service priority and push-out buffer management scheme. The analytical model can be used to predict the performance of smart devices under various traffic conditions to meet the QoS constraints. The analytical model is implemented to evaluate the performance of smart devices under various traffic conditions to meet the QoS constraints.

The results show a clear improvement in the performance of high priority traffic over low priority traffic. As compared with low priority traffic, blocking probability for high priority traffic is very low. The results show the effectiveness of the buffer management scheme used and indicate how high priority traffic will continue its arrival by pushing out low priority traffic to avoid data loss of emergency data packets.

REFERENCES

[1] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, I.S. Jubert, M. Mazura, M. Harrison, M. Eisenhauer, P. Doody, "Internet of Things Strategic Research Roadmap", European Research Cluster on the Internet of Things, Cluster Strategic Research Agenda, 2011, http://www.internet-of-things-research.eu/pdf/IoT_Cluster_Strategic_Research_Agenda_2011.pdf

[2] A.K. Bandara, N. Smith, M. Richards, M. Petre, "Educating the Internet-of-Things Generation", Computer, vol. 46(2), 2013, pp. 53-61.

[3] J. Jin, J. Gubbi, T. Luo, M. Palaniswami, "Network Architecture and QoS Issues in the Internet of Things for a Smart City", Proc. 12th Int. Symp. on Communications and Information Technologies (ISCIT 2012), IEEE, October 2012, Gold Cost, Australia, pp. 974-979

[4] S. Konomi, G. Roussos, "Ubiquitous Computing in the Real World: Lessons Learnt from Large-Scale RFID Deployments", Personal and Ubiquitous Computing, vol. 11(7), 2007, pp. 507–521.

[5] A. Ghose, C. Bhaumik, D. Das, A.K. Agrawal "Mobile Healthcare Infrastructure For Home And Small Clinic" Proc. 2nd ACM International Workshop on Pervasive Wireless Healthcare (MobileHealth '12), 2012, ACM, New York, NY, USA, pp. 15-20

[6] M-A Nef, L. Perlepes, S. Karagiorgou, G.I. Stamoulis, P.K. Kikiras, "Enabling QoS in the Internet of Things", Proc. Fifth International Conf. on Communication Theory, Reliability, and Quality of Service (CTRQ 2012), Chamonix / Mont Blanc, France, 2012, pp.33-38

[7] L. Atzori, A. Iera, G. Morabito, "The Internet of Things: A Survey", Computer Networks, 2010, vol. 54 (15), pp. 2787-2805

[8] L. Yang, S.H. Yang, L. Plotnick, "How the Internet of Things Technology Enhances Emergency Response Operations" International Journal of Technological Forecasting & Social Change", vol. 80, no. 9, Nov. 2013, pp. 1854–1867

[9] D. Miorandi, S. Sicari, F.D. Pellegrini, I. Chlamtac, "Internet of Things: Vision, Applications and Research Challenges", Ad Hoc Networks, 2012, vol. 10 (7), pp. 1497-1516

[10]. B. Klepec, A. Kos, "Performance of VoIP Applications in a Simple Differentiated Services Network Architecture", Proc. International Conference on Trends in Communications, Vol.1, Bratislava, Slovakia, 2001, pp. 214-217

[11] K. Duffy, A.J. Ganesh, "Modeling the Impact of Buffering on 802.11" IEEE COMMUNICATIONS LETTERS, vol. 11(2), 2007, pp. 219-221

[12] H. Guo, S. Ma, F. Liang, "Enabling Awareness Driven Differentiated Data Service in IoT" Journal of Networks, 6(11), 2011, pp. 1572-1577

[13] M.F. Neuts, "Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach", Dover Publications, 1994.

[14] I. Awan, M. Younas, "Towards QoS in Internet of Things for Delay Sensitive Information" Trends in Mobile Web Information Systems, MobiWIS 2013 International Workshops, Paphos, Cyprus, August 26-28, 2013, CCIS, vol. 183, pp. 86-94

[15] P. Spiess, S. Karnouskos, D. Guinard, D. Savio, O. Baecker, L. Souza, V.Trifa, "SOA-based Integration of the Internet of Things in Enterprise Services" Proc. of IEEE ICWS July 2009, Los Angeles, Ca, USA

[16] D. Guinard, V. Trifa, "Towards the Web of Things: Web Mashups for Embedded Devices," In MEM 2009 in Proceedings of WWW 2009. ACM

[17] M-W Ryu, J. Kim, S-S Lee, M-H Song, "Survey on Internet of Things: Toward Case Study Smart", Computing Review, vol. 2, no. 3, June 2012, pp.195-202