The Impacts of Timing Constraints on Virtual Channels Multiplexing in Interconnect Networks*

A. Khonsari [1,2], M. Ould-Khaoua [3], A. Nayebi [2], H. Sarbazi-azad [4,2]
[1] Department of ECE, University of Tehran
[2] IPM School of Computer Science, Tehran, Iran
[3] Department of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK
[4] Department of Computing Engineering, Sharif University of Technology, Tehran, Iran

Abstract

Interconnect networks employing wormhole-switching play a critical role in shared memory multiprocessor systems-on-chip (MPSoC) designs, Multicomputer systems and System Area Networks. Virtual channels greatly improve the performance of wormhole-switched networks because they reduce blocking by acting as "bypass" lanes for non-blocked messages. Capturing the effects of virtual channel multiplexing has always been a crucial issue for any analytical model proposed for wormhole-switched networks. Dally [8] has developed a model to investigate the behaviour of this multiplexing which have been widely employed in the subsequent analytical models of most routing algorithms suggested in the literature. It is indispensable to modify Dally's model in order to evaluate the performance of channel multiplexing in more general networks where restrictions such as timing constraints of input arrivals and finite buffer size of queues are common. In this paper we consider timing constraints of input arrivals to investigate the virtual channel multiplexing problem inherent in most current networks. The analysis that we propose is completely general and therefore can be used with any interconnect networks employing virtual channels. The validity of the proposed equations has been verified through simulation experiments under different working conditions.

1. Introduction

It is widely known that the critical component of a concurrent computer is its interconnect network [9]. Recently, SoC (Systems-on-Chips) design methodologies undergo revolutionary changes. According to recent publications [5], [10], [14], the emergence of SoC platforms consisting of a large set of embedded processors is imminent. A key component of these multiprocessor SoC (MPSoC) platforms [14] is the interconnect topology. Interconnect networks are composed of two types of recourses: buffers and channels (physical channels). Typically a single buffer is associated with each channel. To improve performance the buffer storage associated with each physical channel is divided to several small queues, *virtual channels*, rather than a single deep queue [8]. The virtual channels associated with one physical channel are allocated independently but compete with each other for physical bandwidth and thus virtual channels decouple buffer resources from transmission resources. This decoupling allows active messages to pass blocked messages using network bandwidth that would otherwise be left idle and thus greatly improves performance. Virtual channels have been also used in wormhole-switched deadlock avoidance routing algorithms particularly to avoid deadlock [7, 9]. Adaptive routing algorithms with deadlock recovery on the other hand do not dedicate a set of virtual channels in particular to avoid deadlocks. However, virtual channels are added to act as virtual lanes to provide "bypass" routes for non-blocked messages to improve network performance [9].

Analytical modelling is a versatile and cost-effective alternative to simulation for investigating system performance. Analytical models of routing algorithms with virtual channels in wormhole-switched networks have been widely reported in the literature [6, 11, 16, 17 and references there in]. A Markov chain proposed by Dally [8] has been used by almost all the models reported in the literature to capture the performance behaviour of virtual channels in the network. This Markov chain, however, cannot accurately model a system when message arrivals have deadline constraints. This type of arrivals frequently happens in internet, network systems and more specifically in multicomputers with deadlock recovery routing algorithms which often rely on time-out mechanism to detect potential deadlocks in the network [9]. In this paper, we use theoretical results from queueing systems to capture the effects of virtual channel multiplexing on network performance when arriving messages suffers time-out if do not receive service within a predefined time. Throughout the paper we use the terms "arrivals", "messages", "customers" interchangeably and also "missing deadline", "suffering timeout", "leaving

* This research was in part supported by a grant from I.P.M. (No. CS1384-3-01).

impatiently" liberally in connection with messages who leave the system due to their timing constraints before acquiring the service.

A queueing situation widely studied in the literature is a system in which customers wait for service for a limited time only and leave the system if not served during this time [2-4, 15, 18]. These queueing situations apply to many real-life systems, such as telephone systems and inventory systems with perishable goods. More importantly in high speed packet switching networks and internet individual packets may usually have some timing constraints within which they are to be received at their destinations. There have been many studies [2-4, 15, 18] and references thereafter that investigate different types of these queueing systems and seek to propose solutions for different measures of performance like the fraction of customers who are lost and the average delay in queue of a customer. In addition, there have been some efforts to propose closed-form solutions for the probability distribution functions of important random variables in these queues [3, 15]. In previous study [11] a channel has been considered as a queue with deterministic impatient time customers and the formula suggested by Tijms [18] has been used in the derivation of the model's equations. Using Tijms [18] formula greatly simplifies the derivation of the equations by ignoring the dependency of the events of time-out and the message blocking. These events are related to the number of messages in the system.

This paper presents an accurate model of virtual channel multiplexing in networks with messages having deadline constraints. A very important feature that we address in the paper is taking into considerations the dependency of the events of time-out and blocking in each channel which was assumed independent in the previous study [11]. To deal with this dependency firstly, Dally's Markov chain has been scrutinized and then modified to incorporate message arrivals with timing constraints. The modified Markov chain in particular reduces to Dally's model by setting the deadline of the arrivals to infinity. Secondly, we investigate the event of missing deadline by capturing the dependency of the events of timeout and blocking. In previous study [11] we have simply assumed that these events are independent and have used Tijms [18] equation to calculate the probability of the event of time-out and Dally's equation [8] to compute the probability of the event of blocking in order to determine the probability of the event of message missing deadline. It is worth mentioning that the probability of timeout proposed in [18] has been derived as a special case of our general solution. Thirdly, we calculate the average waiting time of customers in the system which again reduces to the equation of Tijms [18] as a particular case. The results obtained for deterministic impatience time has then been extended to capture the exponential impatience time as well.

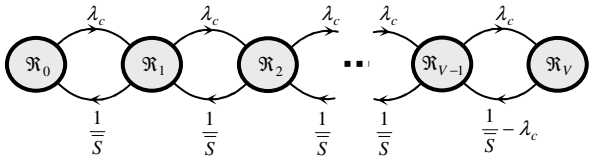


Fig. 1: Morkov model for virtual channel occupancy probabilities.

The remainder of the paper is organised as follows. Section 2 describes the analysis and derivations of the equations. Section 3 validates the model through simulation. Finally, Section 4 concludes this study.

2. Analysis

This section describes first the assumptions used in the analysis, and then presents the analytical model.

Calculation of the average degree of virtual Channels multiplexing ($\overline{V}$)

In virtual channel flow control, multiple virtual channels share the bandwidth of a physical channel in a time-multiplexed manner. To capture the effects of virtual channel multiplexing a Markov model have been proposed by Dally [8]. The model is based on assumptions that have been used in the literature [2-4, 6, 8, 15].

a) Messages inter-arrival time is exponentially distributed with average arrival time λ_c.

b) The message service time are exponentially distributed with average service time $\overline{S}$.

c) V $(V \geq 1)$ virtual channels are used per physical channel; V=1 corresponds to the case where no virtual channels are used. An arrival chooses randomly one of the available virtual channels at one of the physical channels.

In Dally's model [8] the probability, P_v that v virtual channels at a given physical channel are busy, determined by a Markovian model shown in Fig. 1. State $\Re_v$ corresponds to v virtual channels being busy. The transition rate out of state $\Re_v$ to $\Re_{v+1}$ is λ_c, where λ_c is the traffic rate on a given channel, while the rate out of $\Re_v$ to $\Re_{v-1}$ is $1/\overline{S}$. The transition rate out of the last state $\Re_v$ is reduced by λ_c to account for the arrival of messages while a channel is in this state. In the steady state, the model yields the following probabilities.

$$q_v = \begin{cases} 1 & v = 0 \\ q_{v-1}\lambda_c \overline{S} & 0 < v < V \\ q_{v-1}\dfrac{\lambda_c}{1/\overline{S} - \lambda_c} & v = V \end{cases} \qquad (1)$$

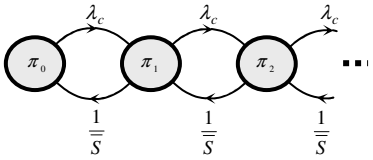


Fig. 2: Number of customers (requests) in an M/M/1 chain

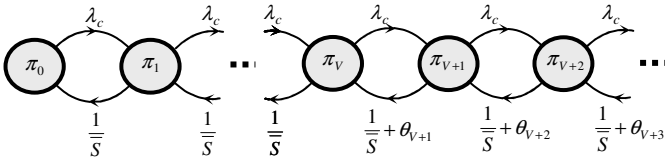


Fig. 3: Number of customers (requests) in a chain when customers are impatient

$$P_v = \begin{cases} 1 \Big/ \sum_{v=0}^{V} q_v & v = 0 \\ P_{v-1} \lambda_c \overline{S} & 0 < v < V \\ P_{v-1} \dfrac{\lambda_c}{1/\overline{S} - \lambda_c} & v = V \end{cases}$$

It is perceptible that this Markov chain is actually an M/M/1 chain [12] which calculates the number of requests in the system with the following condition. In M/M/1 chain shown in Fig. 2 State π_n corresponds to n virtual channels being requested and the probability that v virtual channels are busy, when $\pi_n; n \in [0, V-1]$, is the probability of being in state π_v, i.e. $P_v = \Pr[\pi_v]$. However, the probability that all virtual channels (i.e. V virtual channels) are busy is the summation of the probabilities of being in states $\pi_n; n \in [V, \infty)$ i.e. $P_V = \sum_{n=v}^{\infty} \Pr[\pi_n]$ which is the tail of the queue. The steady-state solution of the M/M/1 chain yields the probability P_v to be

$$P_v = \begin{cases} (1 - \lambda_c \overline{S})(\lambda_c \overline{S})^v, & 0 \le v < V \\ \sum_{n=v}^{\infty} \Pr[\pi_n] = (\lambda_c \overline{S})^V, & v \ge V \end{cases} \quad (2)$$

Rewriting equation (1) yields the following equations

$$P_0 = \frac{1}{\sum_{v=0}^{V} q_v} = \frac{1}{1 + \lambda_c \overline{S} + \ldots + (\lambda_c \overline{S})^{v-1} + \dfrac{(\lambda_c \overline{S})^v}{1 - \lambda_c \overline{S}}} = 1 - \lambda_c \overline{S}$$

$$P_v = \begin{cases} P_0 (\lambda_c \overline{S})^v = (1 - \lambda_c \overline{S})(\lambda_c \overline{S})^v & 0 < v < V \\ (1 - \lambda_c \overline{S})(\lambda_c \overline{S})^{v-1} \dfrac{\lambda_c}{1/\overline{S} - \lambda_c} = (\lambda_c \overline{S})^v & v = V \end{cases}$$

which exactly corresponds to the results obtained in equation (2) and shows that state $\Re_v = \pi_v$ when $v \in [0, V-1]$ and the last state of Dally's model is $\Re_V = \sum_{n=v}^{\infty} \Pr[\pi_n]\ \pi_n; n \in [V, \infty)$.

Now, let consider a situation that arriving messages have deadline constraints and experience timeout and become "lost" messages if do not acquire for service before expiration of their deadline. Only those messages, who acquired for service before their deadline, remain in the queue until served irrespective of whether or not their total waiting exceeds their timeout. The Markov chain in Fig. 3 illustrates the behaviour of virtual channel occupancy of messages taking into account the timing constraint of these messages. It is worth to mention that when number of messages $n \in [0, V-1]$, all arriving messages acquire a free virtual channel immediately. On the other hand, when $n \in [V, \infty)$, an arriving message finds all virtual channels busy and has to wait in the queue to acquire for service. In this case as virtual channels becomes free (by completing the servicing of a message), the queued message enters service with the provision that any message in the queue who has not been acquired service by its deadline time after entering the queue will leave the system impatiently. The balance equations of the length of the queue at time t, $\pi_v(t)$ are given by

$$\begin{aligned} \pi_0'(t) &= -\lambda_c \pi_0(t) + (1/\overline{S}) \pi_1(t), \\ \pi_n'(t) &= \lambda_c \pi_{n-1}(t) - (\lambda_c + 1/\overline{S}) \pi_n(t) \\ &\quad + 1/\overline{S}\, \pi_{n+1}(t), \qquad (n = 1, 2, \ldots, V-1) \quad (3) \\ \pi_n'(t) &= \lambda_c \pi_{n-1}(t) - (\lambda_c + 1/\overline{S} + \theta_n(t)) \pi_n(t) \\ &\quad + (1/\overline{S} + \theta_{n+1}(t)) \pi_{n+1}(t) \qquad (n \ge V) \end{aligned}$$

Where $\pi_n'(t)$ is the first derivative of $\pi_n(t)$ and $\theta_{n+1}(t)h + o(h)$ is the probability that a customer becomes lost during the time interval $(t, t+h)$ under the hypothesis that the queue length is n at time t. Under the assumption of statistical equilibrium, the system (1) becomes

$$\begin{aligned} & -\lambda_c \pi_0 + (1/\overline{S}) \pi_1 = 0, \\ & \lambda_c \pi_{n-1} - (\lambda_c + 1/\overline{S}) \pi_n \\ & + 1/\overline{S}\, \pi_{n+1} = 0, \qquad (n = 1, 2, \ldots, V-1) \qquad (4) \\ & \lambda_c \pi_{n-1} - (\lambda_c + 1/\overline{S} + \theta_n) \pi_n \\ & + (1/\overline{S} + \theta_{n+1}) \pi_{n+1} = 0 \qquad (n \ge V) \end{aligned}$$

In which $\pi_n = \lim_{t \to \infty} \pi_n(t)$ and $\theta_n = \lim_{t \to \infty} \theta_n(t)$.

System (4) has the general solution

$$\begin{aligned} \pi_n &= (\lambda_c \overline{S})^n \pi_0 \qquad (n \le V) \\ \pi_{V+i} &= (\lambda_c \overline{S})^V \pi_0 \lambda_c{}^i \Big/ \prod_{j=1}^{i} \left(\frac{1}{\overline{S}} + \theta_{V+j}\right) \quad (i \ge 1) \qquad (5) \end{aligned}$$

Deterministic Impatience

In this section, important performance measures for the model which are the fraction of customers who lost and the average waiting time in queue when impatience time of the customer is deterministic and is equal to fixed value τ are calculated. Using a similar technique as in [4] θ_{V+i} can be determined by

$$\theta_{V+i} = \tau^{i-1} e^{-\frac{1}{\bar{S}}\tau} \Big/ \int_0^{\tau} t^{i-1} e^{-\frac{1}{\bar{S}}t}\, dt \tag{6}$$

Substituting equation (6) into (5) and using normalization condition of probabilities yields

$$\pi_0 = = \frac{(1-\lambda_c \bar{S})}{1-(\lambda_c \bar{S})^{V+1} e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}} \tag{7}$$

The equilibrium fraction of lost customers which is the probability of time-out is given by

$$P_t = \frac{1}{\lambda_c} \sum_{i=1}^{\infty} \pi_{V+i} \theta_{V+i} = \frac{(1-\lambda_c \bar{S})(\lambda_c \bar{S})^V e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}}{1-(\lambda_c \bar{S})^{V+1} e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}} \tag{8}$$

To calculate steady-state mean waiting time of customers in the system with respect to both served and lost customers it is necessary first to compute average number of customers in the system and queue. Average number of customers in the system n_s can be determined as

$$\begin{aligned} n_s &= \sum_{n=1}^{\infty} n\pi_n = \sum_{n=1}^{V} n\pi_n + \sum_{i=1}^{\infty} (V+i)\pi_{V+i} = \\ &\frac{1-(\lambda_c \bar{S})^V - V(\lambda_c \bar{S})^V + V(\lambda_c \bar{S})^{V+1}}{(1-\lambda_c \bar{S})^2} (\lambda_c \bar{S})\pi_0 \\ &+ \frac{V - V\lambda_c \bar{S} - V\pi_0 + V\pi_0 (\lambda_c \bar{S})^{V+1}}{1-\lambda_c \bar{S}} + (\lambda_c \bar{S})^V \pi_0 \\ &\times \left[\frac{\lambda_c \bar{S}}{(1-\lambda_c \bar{S})^2} - \left(\frac{\lambda_c \bar{S}}{(1-\lambda_c \bar{S})^2} + \frac{(\lambda_c \bar{S})^2 (\tau/\bar{S})}{1-\lambda_c \bar{S}} \right) e^{-(1-\lambda_c \bar{S})\tau/\bar{S}} \right] \end{aligned} \tag{9}$$

Similarly, the average number in queue n_q is

$$\begin{aligned} n_q &= \sum_{n=1}^{\infty} n\pi_{n+V} = \\ &\left[\frac{\lambda_c \bar{S}}{(1-\lambda_c \bar{S})} - \left(\frac{\lambda_c \bar{S}}{(1-\lambda_c \bar{S})} + (\lambda_c \bar{S})^2 \frac{\tau}{\bar{S}} \right) e^{-(1-\lambda_c \bar{S})\tau/\bar{S}} \right] \\ &\times \frac{(\lambda_c \bar{S})^V}{1-(\lambda_c \bar{S})^{V+1} e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}} \end{aligned} \tag{10}$$

Using Little's formula [13] the steady-state average waiting in queue of a customer with respect to both served and lost customers, W_q can be stated

$$\begin{aligned} W_q = \frac{n_q}{\lambda_c} &= \left[\frac{\lambda_c \bar{S}^2}{(1-\lambda_c \bar{S})} - \left(\frac{\lambda_c \bar{S}^2}{(1-\lambda_c \bar{S})} + (\lambda_c \bar{S})^2 \tau \right) e^{\frac{(\lambda_c \bar{S}-1)\tau}{\bar{S}}} \right] \\ &\times \frac{(\lambda_c \bar{S})^{V-1}}{1-(\lambda_c \bar{S})^{V+1} e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}} \end{aligned} \tag{11}$$

For the special case and by setting the number of virtual channels in equations (8) and (11) to 1 give rise the following results that have been previously reported in Tijms [18].

$$P_t = \frac{1}{\lambda_c} \sum_{i=1}^{\infty} \pi_{1+i} \theta_{1+i} = \frac{(1-\lambda_c \bar{S})(\lambda_c \bar{S}) e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}}{1-(\lambda_c \bar{S})^2 e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}}$$

$$\bar{W}_c^{\exp} = \frac{\frac{\lambda_c \bar{S}^2}{(1-\lambda_c \bar{S})} - \left(\frac{\bar{S}}{(1-\lambda_c \bar{S})} + \lambda_c \bar{S}\tau \right)(\lambda_c \bar{S}) e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}}{1-\lambda_c^2 \bar{S}^2 e^{-(1-\lambda_c \bar{S})\tau/\bar{S}}} \tag{12}$$

By setting timeout to infinity π_0 given by equation (7) becomes $(1-\lambda_c \bar{S})$ and the Markov chain in Fig. 3 reduces to M/M/1 chain depicted in Fig. 2 and Dally's model is derived in a manner that mentioned in the equation (2). Finally, as an extreme case by setting timeout to infinity and number of virtual channels to 1 the performance measures of an M/M/1 queue is obtained, which agrees with the results already reported in the literature [12].

$$n_s = \frac{\lambda_c \bar{S}}{1-\lambda_c \bar{S}}$$

$$n_q = \sum_{n=1}^{\infty} n\pi_{n+1} = n_s - (1-\pi_0) = \frac{(\lambda_c \bar{S})^2}{1-\lambda_c \bar{S}}$$

$$W_q = \frac{n_q}{\lambda_c} = \frac{(\lambda_c \bar{S})\bar{S}}{1-\lambda_c \bar{S}}$$

Negative Exponential Impatience

In this section, expressions for the average number of customers in queue and system; the probabilities of waiting, acquiring service and the customer lost rate (timeout) are obtained. After joining the queue each customer will wait a certain length of time for service to begin. If it has not begin by then, he departs and becomes a lost customer. The impatience time is a random variable whose density function is given by a negative exponential density with parameter θ.

With V virtual channels and $(V+i)$ customers in the system any of i customers in the queue may renege and due to memory-less property of the exponential random variable, the minimum of $i : i \in [1, \infty)$ independent exponential random variable with parameter $-\theta$ is exponential with parameter $-i\theta$. By substitution in equation (5) we have

$$\pi_{V+i} = (\lambda_c \bar{S})^V \pi_0 (\frac{\lambda_c}{\theta})^i \Big/ \prod_{j=1}^{i} (\frac{1}{\bar{S}\theta} + j) \qquad (i \geq 1) \qquad (13)$$

using normalization condition of probabilities gives

$$\frac{1}{\pi_0} = \frac{1-(\lambda_c \bar{S})^{V+1}}{(1-\lambda_c \bar{S})} + (\lambda_c \bar{S})^V \left[-1 + \frac{1}{\bar{S}\theta} \left(\frac{\lambda_c}{\theta} \right)^{-1/\bar{S}\theta} e^{\frac{\lambda_c}{\theta}} \int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{\frac{1}{\bar{S}\theta}-1} dt \right] \qquad (14)$$

where $\int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{u-1} dt$ is the Incomplete Gamma function [1].

Analogous to the deterministic timeout case, the mean number of customers in system (n_s) and queue (n_q) and probability of timeout (P_t) for the negative exponential timeout are summarized below:

$$n_s = \sum_{n=1}^{\infty} n\pi_n = \sum_{n=1}^{V} n\pi_n + \sum_{i=1}^{\infty} (V+i)\pi_{V+i} \qquad (15)$$

$$= \frac{\lambda_c \bar{S} - (\lambda_c \bar{S})^{V+1} - V(\lambda_c \bar{S})^{V+1} + V(\lambda_c \bar{S})^{V+2}}{(1-\lambda_c \bar{S})^2} \pi_0 + V(\lambda_c \bar{S})^V \times \left[-1 + \frac{1}{\bar{S}\theta} \left(\frac{\lambda_c}{\theta} \right)^{-1/\bar{S}\theta} e^{\frac{\lambda_c}{\theta}} \int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{\frac{1}{\bar{S}\theta}-1} dt \right] \pi_0 + \frac{1}{\bar{S}\theta} (\lambda_c \bar{S})^V \times \left[\left(-\frac{1}{\bar{S}\theta} + \lambda_c / \theta \right) \left(\frac{\lambda_c}{\theta} \right)^{-1/\bar{S}\theta} e^{\frac{\lambda_c}{\theta}} \int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{\frac{1}{\bar{S}\theta}-1} dt + 1 \right] \pi_0$$

,

$$n_q = \left(1/(\bar{S}\theta)\right)(\lambda_c \bar{S})^V \pi_0 \times \left[\left(-1/(\bar{S}\theta) + \lambda_c / \theta \right) \left(\frac{\lambda_c}{\theta} \right)^{-1/\bar{S}\theta} e^{\frac{\lambda_c}{\theta}} \int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{\frac{1}{\bar{S}\theta}-1} dt + 1 \right] \qquad (16)$$

and

$$P_t = \frac{1}{\lambda_c} \sum_{i=1}^{\infty} \pi_{V+i} \theta_{V+i} = \frac{\theta}{\lambda_c} \sum_{i=1}^{\infty} i\pi_{V+i} = (\lambda_c \bar{S})^{V-1} \pi_0 \times \left[\left(-1/(\bar{S}\theta) + \lambda_c / \theta \right) \left(\frac{\lambda_c}{\theta} \right)^{-1/\bar{S}\theta} e^{\frac{\lambda_c}{\theta}} \int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{\frac{1}{\bar{S}\theta}-1} dt + 1 \right] \qquad (17)$$

Using a method similar to the deterministic case discussed in equation (11), mean waiting in the queue is calculated by

$$W_q = \frac{n_q}{\lambda_c} = \left(1/(\bar{S}\theta)\right)(\lambda_c \bar{S})^V \pi_0 \left(1/\lambda_c\right) \times \left[\left(-\frac{1}{\bar{S}\theta} + \lambda_c / \theta \right) \left(\frac{\lambda_c}{\theta} \right)^{-1/\bar{S}\theta} e^{\frac{\lambda_c}{\theta}} \int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{\frac{1}{\bar{S}\theta}-1} dt + 1 \right] \qquad (18)$$

Again, for the special case with one virtual channel equations (14) and (16) correspondingly reduce to

$$\pi_0 = \left[1 + \left(\frac{\lambda_c}{\theta} \right)^{1-(1/\bar{S}\theta)} e^{\frac{\lambda_c}{\theta}} \int_0^{\frac{\lambda_c}{\theta}} e^{-t} t^{\frac{1}{\bar{S}\theta}-1} dt + 1 \right]^{-1} \qquad (19)$$

$$n_q = \left(\lambda_c / \theta\right) - \frac{1}{\bar{S}\theta} (1 - \pi_0) \qquad (20)$$

which agrees with the results that have been derived in [2].

3. Validation

The proposed equations have been validated by means of a discrete-event simulation. Each simulation experiment was run until the system reached its steady state, that is until a further increase in simulated system does not change the collected statistics appreciably. Numerous validation experiments have been performed to validate the key measures of system performance such as mean waiting time in queue, probability of timeout and probability of blocking, and mean number of customers in queue for both deterministic and exponential impatience time. In all appropriate cases Dally's model and M/M/1 queue have been derived as particular cases of our general solution. However, for the sake of specific illustration, results are presented for some important cases. In what follows the horizontal axis in the figures represents the traffic generation rate.

The probability of timeout, mean waiting in the system and average number of customers in queue are shown in Fig. 4. In the figure average service times are set to $\bar{S} = 32$, $\bar{S} = 64$ and $\bar{S} = 128$. Moreover, timeout values are deterministic and are equal to fixed values $\tau = 0$, $\tau = \infty$ and $\tau = \bar{S}$ and the number of virtual channels per physical channel are set to *V*= 4. The above scenarios have been repeated for exponential timeout and some partial results have been illustrated in Fig. 5. As it is illustrated in figures the results obtained through simulations matches with mathematical equations with a high degree of precision which is less than 0.1 percent in almost all cases.

The equations presented in this paper are applicable in many practical cases that arriving customers have timing constraints. For instance, recently many routing algorithms proposed in the literature known as deadlock recovery routing algorithms employ a variation of timeout

mechanism to recover from deadlock [9]. These routing algorithms use the concept of virtual channels to achieve better performance. In order to compare the performance merits of these routing algorithms through analytical model and due to the lack of the necessary equations it was required to make the crude assumption of the independency of the probabilities of timeout and blocking [11]. This assumption was a major source of errors in the analytical models. Fig. 6 compares the important measures of performance in a virtual channel multiplexer using the crude independence assumption (depicted in the figure as "independent") and the equations proposed in this paper (shown as "new_equation"). As it is evident from the figure the discrepancies between the independent case and reality is near 40 percent in some traffic regions.

4. Conclusions

The performance of contemporary concurrent systems such as MPSoC, Multicomputers and System Area Networks are not only determined by the capacity of the node processors (e.g. CPU speed, cache size, etc.), but it is also limited by the interconnect network that connects the processors and memories in MPSoC and processors to processors in multicomputers. Design and optimization of such interconnect network are critical for these system performance. By introducing virtual channels in the input and output ports, we can increase channel utility of interconnect network considerably. Dealing with virtual channel multiplexing has always been a crucial issue for any analytical model proposed for wormhole-switched networks. Most existing analytical models proposed for evaluating the performance merits of different routing algorithms in multicomputers have used a method proposed by Dally [8] to investigate the effect of virtual channel multiplexing in the network. This method, however, loses accuracy as traffic increases especially for more complex networks.

In an effort to gain deep understanding of the issue of modelling virtual channel multiplexing, this paper is first to address a general solution for virtual channel multiplexing when arrival messages have timing constraints. Important measures of system performance such as mean number of messages in system and queue, probability of timeout and mean message latency has been derived for both deterministic an exponential impatience time for unbounded queues. Moreover, Dally's model [8], probability of timeout and mean waiting time given by Tijms [18] has been derived as particular case of our general solution. Finally, our boundary results agree with the results attained by Barrer [4] and Ancker [2] as well. Moreover, simulation experiments also have been conducted as a double confirmation. The results have revealed that important measures of performance obtained by developed equations are in agreement with those provided by the simulation with a high degree of accuracy. The next step of this work is to extend the above modelling approach to bounded queueing system and obtaining measures when service time of messages are generally distributed.

5. References

[1] Abramovitz-Stegun: *Handbook of Mathematical Functions,* (Dover Edition 1965).

[2] C.J. Ancker and A.V. Gafarian, Queueing with impatient customers who leave at random, *The Journal of Industrial Engineering,* XIII (2) (1962) 84-90.

[3] C.J. Ancker and A.V. Gafarian, Some queueing problems with balking and reneging, *Oper. Res.* (11) (1963) 88-100.

[4] D.Y. Barrer,Queuing with impatient customers and ordered service, Opr. Res. 5 (1957) 650-656.

[5] L. Benini and G. DeMicheli, "Networks on Chips: A New SoC Paradigm," Computer, 35(1) (2002) 70-78.

[6] Y. Boura, Design and analysis of routing schemes and routers for wormhole-routed mesh architectures, Ph.D. Dissertation, Department of Computer Science and Engineering, Penn State University (1995).

[7] W.J. Dally and C.L. Seitz, Deadlock-free message routing in multiprocessor interconnection networks, *IEEE Trans. Computers* 36(5) (1987) 547-553.

[8] W.J. Dally, Virtual channel flow control, *IEEE Trans. Parallel & Distributed Systems,* 3(2) (1992) 194-215.

[9] J. Duato, S. Yalamanchili and L. Ni, *Interconnection Networks: An Engineering Approach*, Morgan Kaufmann Publishers (2003).

[10] M. Horowitz and B. Dally, "How Scaling Will Change Processor Architecture," Proc. Int'l Solid State Circuits Conf. (ISSCC), (Feb. 2004), 132-133.

[11] A. Khonsari,, Performance Modelling and Analysis of Deadlock Recovery Routing Algorithms in Multicomputer Interconnection Networks, PhD Thesis, Compting Science Department, Glasgow University, (2003).

[12] L. Kleinrock, *Queueing Systems: Theory* (1) (John Wiley & Sons, New York, 1975).

[13] J. D. C. LITTLE, A proof for the Queueing Formula *Operations Research,* 9(3) (1961) 383-387.

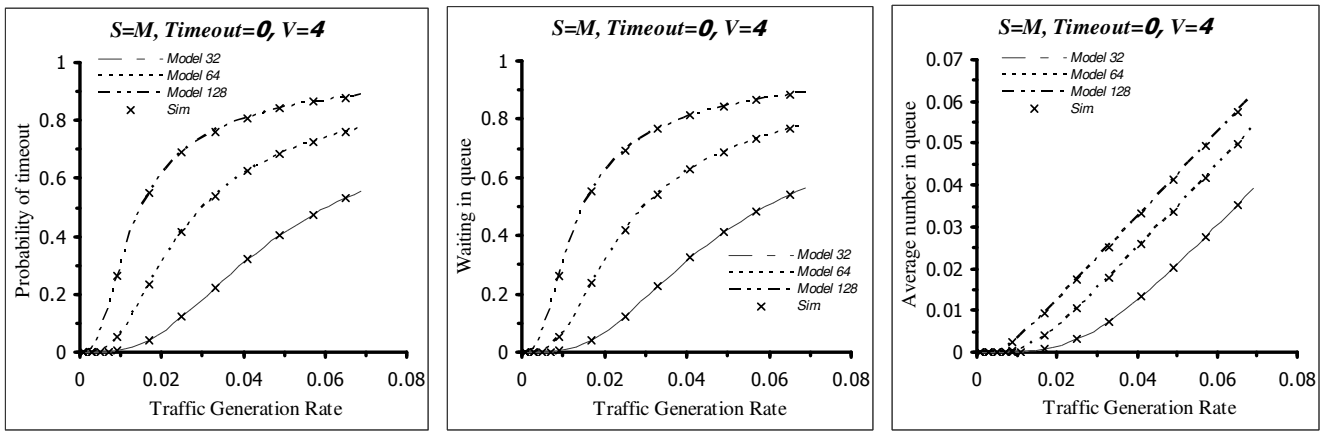
[14] P. Magarshack and P.G. Paulin, "System-on-Chip beyond the Nanometer Wall," Proc. Design Automation Conf. (DAC), (June 2003), 419-424.

[15] A. Movaghar, On queueing with customer impatience until the beginning of service, *Queueing Systems,* (29) (1998) 337-350.

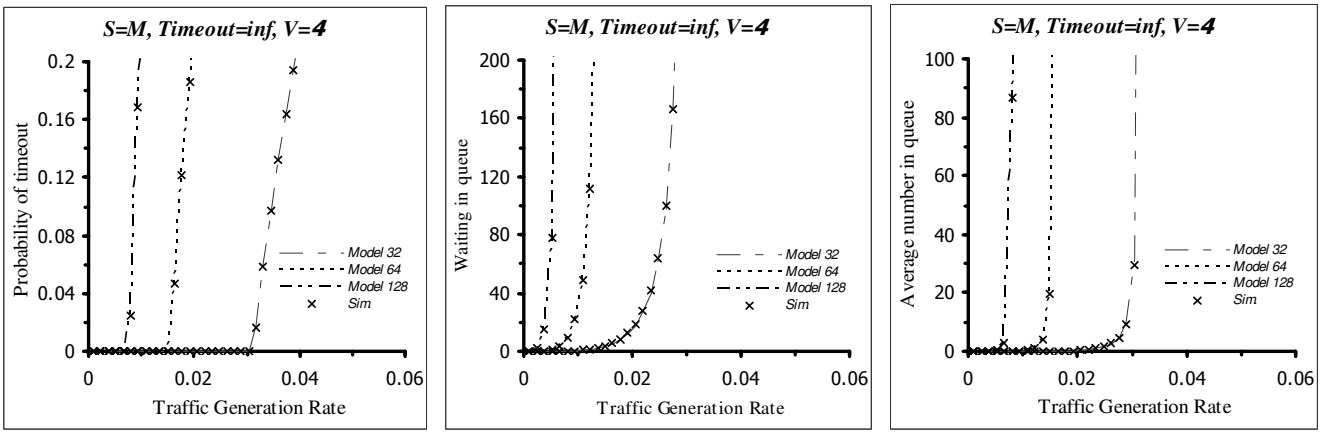
[16] M. Ould-Khaoua, A performance model for Duato's fully adaptive routing algorithm in *k*-ary *n*cubes, *IEEE Trans. Computers* 48(12) (1999) 1-8.

[17] H. Sarbazi-Azad, Performance Analysis of Wormhole Routing in Interconnection Networks, PhD Thesis, Compting Science Department, Glasgow University, (2001).

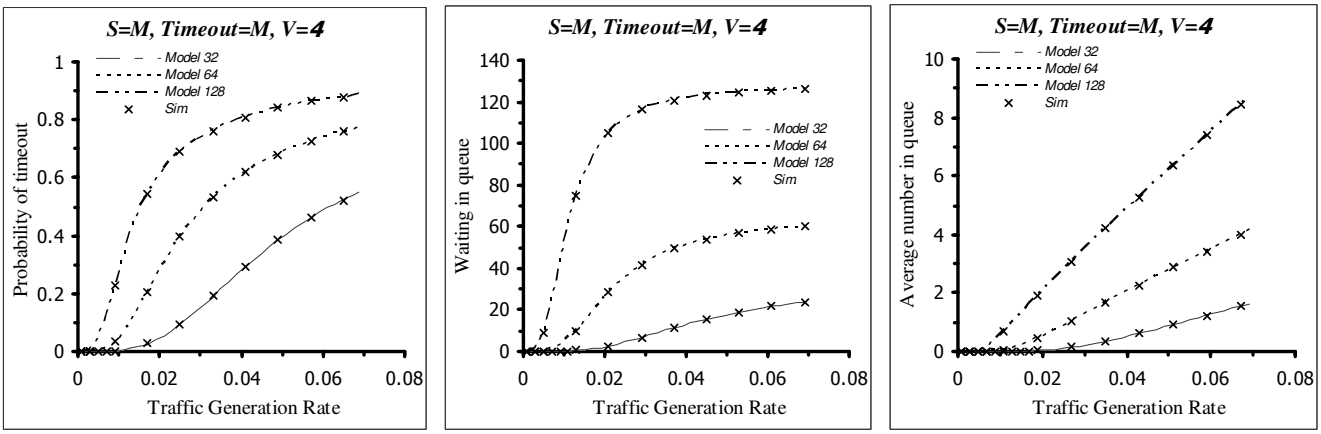
[18] H.C. Tijms, Stochastic modelling and analysis: A computational approach (John Wiley & Sons, 1986).



(a)



(b)



(c)

Fig. 4: The probability of timeout, mean message waiting in the queue and average number of messages in the queue versus traffic generation rate. The number of virtual channels $V = 4$ and mean service time $\overline{S} = 32, 64, 128$, timeout period is deterministic and equals (a) ($\tau = 0$), (b) ($\tau = \infty$) and (c) ($\tau = \overline{S}$).

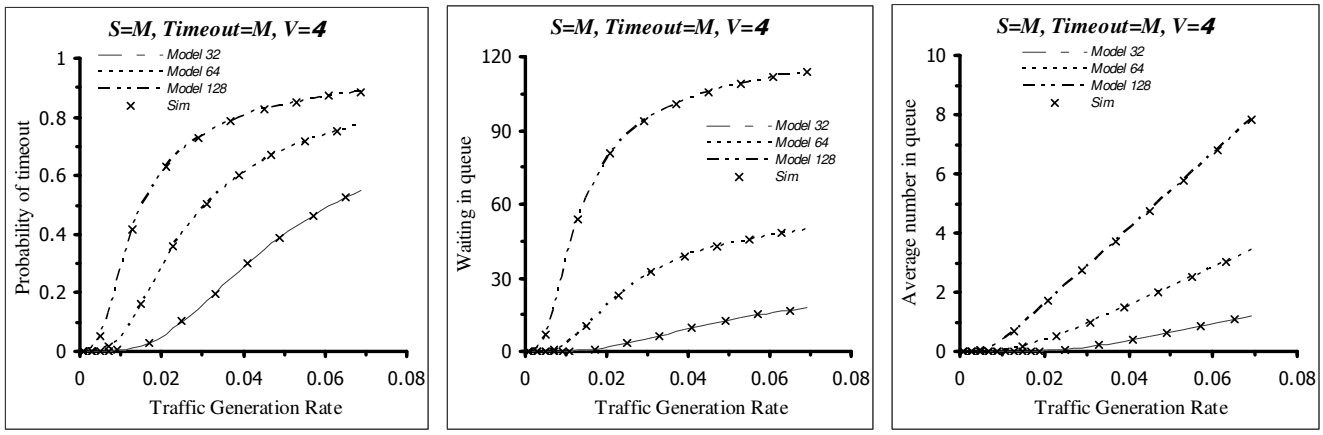


Fig. 5: The probability of timeout, mean message waiting in the queue and average number of messages in the queue versus traffic generation rate. The number of virtual channels $V = 4$ and service time $\overline{S} = 32, 64, 128$, timeout period is exponentially distributed with mean ($\tau = \overline{S}$).

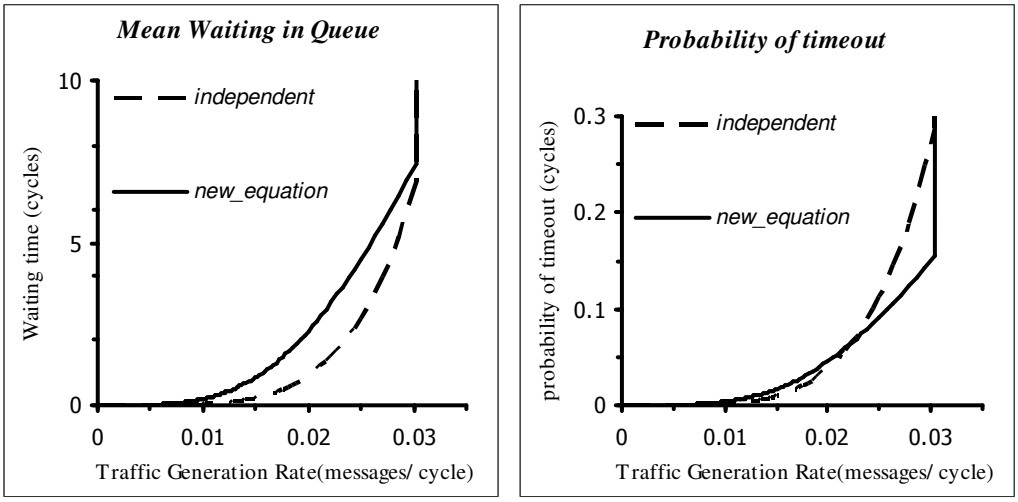


Fig. 6: comparing the important measures of performance in a virtual channel multiplexer using the crude independence assumption (depicted as "independent") and the equations proposed in this paper (illustrated as "new_equation").