On the Probability Distribution of Busy Virtual Channels

N. Alzeidi [1], A. Khonsari [2,3], M. Ould-Khaoua [1], L. M. Mackenzie [1]

[1] University of Glasgow
Computing Science Department.
Glasgow, UK
{zeidi, mohamed, lewis}@dcs.gla.ac.uk

[2] University of Tehran, Dept. of ECE
[3] IPM, School of Computer Science
Tehran, Iran
ak@ipm.ir

Abstract

A major issue in modelling the performance merits of interconnection network is dealing with virtual channels. Some analytical models chose not to deal with this issue at all i.e. one virtual channel per physical channel. More sophisticated models, however, relayed on a method proposed by Dally to capture the effect of arranging the physical channel into many virtual channels. In this study, we investigate the accuracy of Dally's method and propose an alternative approach to deal with virtual channels in analytical performance modelling. The new method is validated via simulation experiments and results reveal its accuracy under different traffic conditions.

1. Introduction

Wormhole switching [7] has become the dominating switching technique used in contemporary multicomputers and more recently in clusters [8] and system area networks [14]. This is because it requires minimum buffer space and it makes message latency almost independent of the message distance in absence of blocking. In wormhole switching, a message is broken into flits (few bytes each) for transmission and flow control. The header flit (contains the routing information) governs the route and the remaining data flits follow in a pipelined fashion. If the header flit is blocked, the data flits are blocked in situ.

As network traffic increase, messages may experience large delay to cross the network due to chains of blocked channels. To reduce the blocking delay, the flit buffers associated with a given physical channel are organized into several virtual channels [3], each representing a "logical" channel with its own buffer and flow control logic. Virtual channels are allocated independently to different messages and compete with each other for the physical bandwidth in a time multiplexed manner. This de-coupling allows messages to bypass each other in the event of blocking, using network bandwidth that would otherwise be wasted. Adding virtual channels to wormhole-switched networks greatly improves the performance because they reduce blocking by acting as "bypass" lanes for non-blocked messages. Fig. 1 illustrates the use of virtual channels as bypass lanes.

The concept of virtual channels has also been exploited to develop deadlock-free routing algorithms [4]. A routing algorithm specifies how a message selects its network path. Dealing with deadlock situations, that is when no message can advance towards its destination due to blocked channels, is a critical requirement for any routing algorithm. Deterministic routing [7] has been widely deployed in existing multicomputers because it is simple to implement and requires minimum number of virtual channels [9, 16, 17]. However, messages with the same source and destination addresses always follow the same path. As a result, they can not take advantage of the alternative paths that a topology may provide to reduce latency and avoid faulty links. In contrast, many adaptive

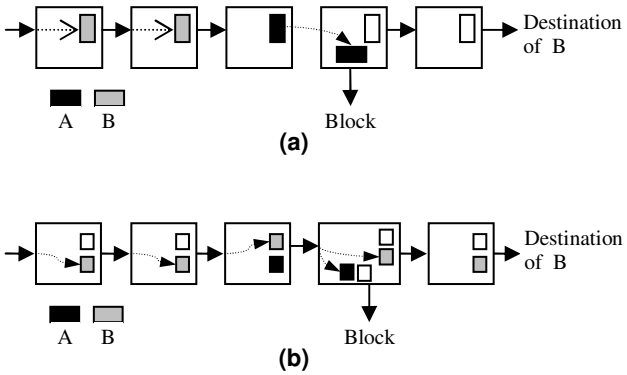


Fig. 1: (a) Message B is blocked behind message A, while physical channels remain idle. (b) Virtual channels provide additional buffers allowing message B to pass blocked message A.

1-4244-0054-6/06/$20.00

routing algorithms have been proposed where messages can use any of the available alternative paths between a given pair of nods to advance towards their destinations [5, 6].

Finding the probability distribution of the number of busy virtual channels per physical channel has always been a critical issue in all analytical models that were developed to assess the performance of interconnection networks that employ several virtual channels per physical channel. Previous studies reveal that adding virtual channels greatly improves the performance and reduces the blocking delay. Hence, it is of a significant importance for any analytical model to accurately calculate the probability distribution of the number of busy virtual channels per physical channel. Many analytical models have been proposed to evaluate the performance of wormhole networks [2, 10, 11, 12, 18, 19, 20]. Almost all of these models have used a method proposed by Dally [3] to calculate the probability distribution of the number of busy virtual channels. The method proposed by Dally is based on a Markov process and although it is useful in some cases, it loses its accuracy as network traffic increases.

In this paper we reinvestigate Dally's method and study its accuracy under deferent traffic conditions. We also propose a new general method for calculating the probability distribution of the number of busy virtual channels based on an M/G/1 queuing system. The new method can easily be tailored for different traffic conditions by simply using different service time distributions for the M/G/1 queue. Moreover, we showed that Dally's method can be derived as a special case from our new general method. To conclude this section we now define some variables that are necessary for the rest of the paper.

1- Each physical channel is divided into V virtual channels.
2- The mean traffic received by each physical channel is λ_c messages per cycle.
3- The mean service time for each message is $\overline{S}$ cycles.
4- We will refer to Dally's probability distribution of the number of busy virtual channels per physical channel as $\{ P_v^{Dally} ; 0 \le v \le V \}$. Similarly we will refer to our new general version of the probability distribution as $\{ P_v^{General} ; 0 \le v \le V \}$.
5- The probability distribution of the number of customers in an M/G/1 queuing system at an arbitrary time will be referred to as $\{ \pi_i^{M/G/1} ; i = 0,1,2,\dots \}$.

The rest of this paper is organised as follows. Section 2 briefly explains Dally's method of calculating the probability distribution of the number of busy virtual channels per physical channel and shows its relation to M/M/1 queuing system. Section 3 is devoted to the development of the new general method. In Section 4 we validated the new method and compared it to Dally's method. Finally, we conclude the study in Section 5 and present some future directions.

2. Dally's Method

Dally determined $\{ P_v^{Dally} ; 0 \le v \le V \}$ using a Markov process [3], shown in fig. 2. State V_v corresponds to v virtual channels being busy. The transition rate out of state V_v to state V_{v+1} is λ_c, while the rate out of state V_{v+1} to state V_v is $1/\overline{S}$. The transition rate out of the last state is reduced by λ_c to account for the arrival of messages while a channel is in this state.

Solving this model for the steady state probabilities gives [3]

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

$$P_v^{Dally} = \begin{cases} \dfrac{1}{\sum_{v=0}^{V} q_v} & v = 0 \\ P_{v-1}^{Dally} \lambda_c \overline{S} & 0 < v < V \\ P_{v-1}^{Dally} \dfrac{\lambda_c}{1/\overline{S} - \lambda_c} & v = V \end{cases} \quad (2)$$

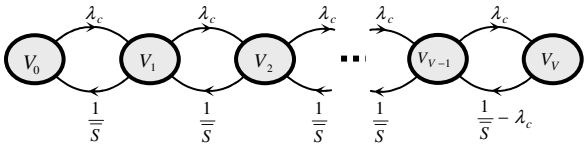


Fig. 2: Markov process for virtual channel occupancy

We now present the following lemma

Lemma:

The probability distribution $\{ P_v^{Dally} ; 0 \le v \le V \}$ of the number of busy virtual channels per physical channel as calculated by Dally in [3] is exactly the same as the probability distribution of the number of customers in an M/M/1 queuing system.

Proof of lemma :

Rewriting equation (2) yields the following

$$P_0 = \frac{1}{\sum_{v=0}^{V} q_v}$$

$$= \frac{1}{q_0 + \lambda_c \bar{S} q_0 + \ldots + (\lambda_c \bar{S})^{V-1} q_0 + (\lambda_c \bar{S})^{V-1} q_0 \frac{\lambda_c}{1/\bar{S} - \lambda_c}}$$

$$= \frac{1}{1 + \lambda_c \bar{S} + \ldots + (\lambda_c \bar{S})^{V-1} + \frac{(\lambda_c \bar{S})^V}{1 - \lambda_c \bar{S}}}$$

$$= \frac{1}{\frac{1 - (\lambda_c \bar{S})^V}{1 - \lambda_c \bar{S}} + \frac{(\lambda_c \bar{S})^V}{1 - \lambda_c \bar{S}}}$$

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

We then can write

$$P_v^{Dally} = \begin{cases} P_0 (\lambda_c \bar{S})^v & 0 \le v < V \\ (1 - \lambda_c \bar{S})(\lambda_c \bar{S})^{v-1} \dfrac{\lambda_c}{1/\bar{S} - \lambda_c} & v = V \end{cases} \quad (3)$$

After some manipulation of the above equations, $\{ P_v^{Dally} ; 0 \le v \le V \}$ can be rewritten as

$$P_v^{Dally} = \begin{cases} (1 - \lambda_c \bar{S}) (\lambda_c \bar{S})^v & 0 \le v < V \\ (\lambda_c \bar{S})^V & v = V \end{cases} \quad (4)$$

This is exactly the probability distribution of the number of customers in an M/M/1 queuing system; see for example[13, 15]. This proofs the lemma ■

By virtue of the above lemma, Dally's method of calculating the probability distribution of the number of busy virtual channels per physical channel, $\{ P_v^{Dally} ; 0 \le v \le V \}$ is now reduced to the calculation of the probability distribution of the number of customers in an M/M/1 queuing system. This approach is accurate under low traffic where the performance measures of an M/M/1 queue do not deviate too much compared to other queues and the service time is nearly exponential. However, as the traffic increases the blocking nature of the wormhole-switched networks interrupts the service time at each switch and the service becomes more general rather than exponential as in the M/M/1 queuing system. This explains the degradation of the accuracy of the analytical models that are based on Dally's method under moderate and high traffic. In other words, it is the assumption of exponential service times that contributes to pure accuracy of Dally's method under moderate and high traffic.

3. A New General Method

In this section we propose a new general method for calculating $\{ P_v^{General} ; 0 \le v \le V \}$. Here, we calculate $\{ P_v^{General} ; 0 \le v \le V \}$ as the probability distribution $\{ \pi_i^{M/G/1} ; i = 0, 1, 2, \ldots \}$ of the number of customers in an M/G/1 queuing system at arbitrary times. The mean arrival rate is exponentially distributed with mean λ_c. The fact that the service time is general gives us the freedom to either assume well-known service time distributions or to approximate it using approximation methods. This gives the model the flexibility to be adapted to different network and traffic conditions and hence more accurate results.

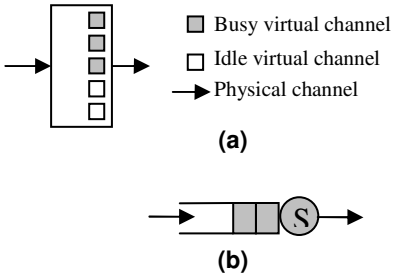


Fig. 3. (a) Three busy virtual channels corresponds to (b) Three customers in the M/G/1 queuing system: two in the queue and one being serviced

As illustrated in fig. 3, when there are *v* customers in the system this corresponds to *v* virtual channels being requested. The probability that *v* virtual channels are busy, when $0 \le v \le V-1$, is the probability of *v* customers in the system, i.e. $P_v^{General} = \pi_v^{M/G/1}$. However, the probability that all virtual channels are busy is the summation of the probabilities of v customers being in the system where $V \le v < \infty$ i.e. $P_V^{General} = \sum_{i=V}^{\infty} \pi_i$. In other words, $P_v^{General}$ is equal to the tail of the probability distribution $\{ \pi_i^{M/G/1} ; i = 0, 1, 2, \ldots \}$ of the number of customers in the M/G/1 queuing system at arbitrary time. This is to account for all the new requests for virtual channels when all the virtual channels are occupied. To summarize we can write

$$P_v^{General} = \begin{cases} \pi_v^{M/G/1} & 0 \le v \le V-1 \\ \sum_{i=V}^{\infty} \pi_i^{M/G/1} & v = V \end{cases} \quad (5)$$

The probability distribution $\{ \pi_i^{M/G/1} ; i = 0, 1, 2, \ldots \}$ of the number of customers in an M/G/1 queuing system is

obtained by inverting the probability generation function of the queue length of the M/G/1 queue which is given by [13, 15]

$$G_N(z) = \frac{(1-\rho)(1-z)}{S^*(\lambda_c(1-z)) - z} S^*(\lambda_c(1-z)) \qquad (6)$$

Where $S^*(s)$ is the Laplace transform of the service time distribution and $\rho = \lambda_c \overline{S}$ is the server utilization.

In the first instance some important observations from the above equation can be realized. First, the initial value property of generating functions [13, 15] and by setting z=0 shows that the probability that the system is idle is given by $1-\rho$. Moreover, the complete summation property of the probability generation functions implies that $G_N(1) = 1$. However, since the numerator of equation (6) has the factor of (1-z), this implies that the denominator should have this factor to avoid ambiguity of this equation at z=1. The important thing about equation (6) is that inverting it yields an expression for $\{\pi_i^{M/G/1}; i = 0,1,2,\ldots\}$ which consequently, by using equation (5), gives the number of virtual channels being requested.

We now show how Dally's method can be derived from our general method. This will give an example of the methods used to invert the z-transform presented in equation (6). As we have proven in lemma 1, Dallys method is actually identical to finding the probability distribution of the number of customers in an M/M/1 queue. In M/M/1 queue the Laplace transform of the service time, which is exponentially distributed with parameter, $1/\overline{S}$ is given by

$$S^*(s) = \frac{1/\overline{S}}{s + 1/\overline{S}} \qquad (7)$$

And hence

$$S^*(\lambda_c(1-z)) = \frac{1/\overline{S}}{\lambda_c(1-z) + 1/\overline{S}} = \frac{1}{\lambda_c \overline{S}(1-z) + 1} \qquad (8)$$

Substituting this into equation (6) and simplifying yields

$$\begin{aligned} G_N(z) &= \frac{1-\lambda_c \overline{S}}{1-\lambda_c \overline{S} z} \\ &= (1-\lambda_c \overline{S})(1 + (\lambda_c \overline{S} z) + (\lambda_c \overline{S} z)^2 + \ldots) \\ &= (1-\lambda_c \overline{S}) \sum_{i=0}^{\infty} (\lambda_c \overline{S})^i z^i \end{aligned} \qquad (9)$$

It is noteworthy that the common factor (1-z) in the numerator and the denominator of the above equation has been removed in order to avoid undefined situation in the equation when z=1. By inverting the above equation the probability distribution $\{\pi_i^{M/M/1}; i = 0, 1, 2, \ldots\}$ of the number of customers in the M/M/1 queuing system can be written as

$$\pi_i^{M/M/1} = (1-\lambda_c \overline{S})(\lambda_c \overline{S})^i \quad i = 0, 1, 2, \ldots \qquad (10)$$

Now the probability distribution $\{P_v^{General}; 0 \le v \le V\}$ of the number of busy virtual channels is obtained by substituting equation (10) into equation (5). Hence, after simplification, we can write

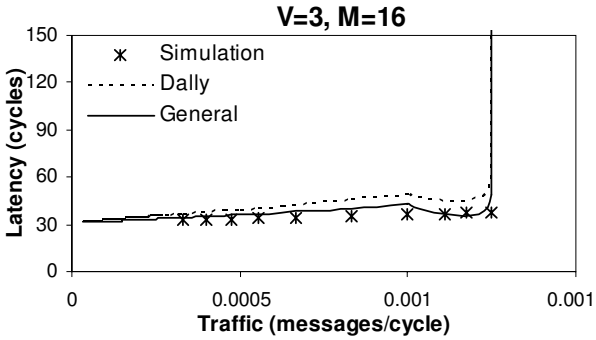
$$P_v^{Dally} = \begin{cases} (1-\lambda_c \overline{S})(\lambda_c \overline{S})^v & 0 \le v < V \\ (\lambda_c \overline{S})^V & v = V \end{cases} \qquad (11)$$

This is exactly the equation that has been derived by Dally in [3] and we derived it from equation (5).

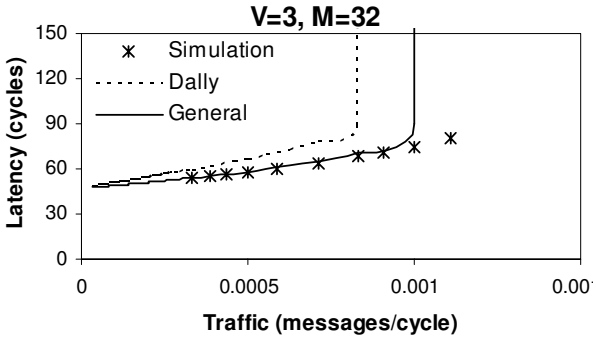
It is important to mention that as long as the Laplace transform of the service time distribution is available we can use any method of inverting the z-transform such as Complex Variable, Long Division, Partial Fraction or Power Series Expansion Methods [1] to invert equation (6). However, if the resulting z-transform is hard to invert, then we can use algorithmic approach to numerically invert the transform. Also, when the Laplace transform of the service time distribution is not available, it can be approximated using two-moment matching approximation [13]. The numerical inversion of the Laplace transform and the two moment approximation of the service time are two complementary issues that should be further investigated in order to make the general method more comprehensive and self-contained. These two issues will be the focus of our future research.

4. Validation and Comparison

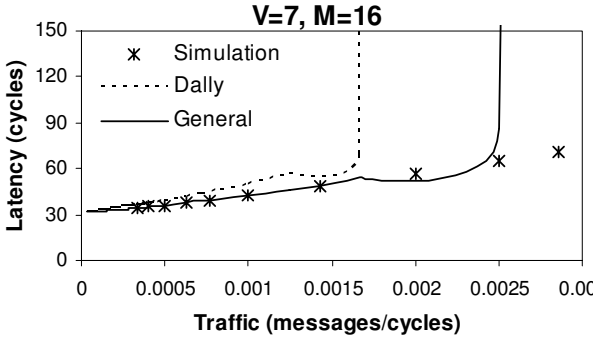
There have been many analytical models in the literature that employed Dally's method to calculate the distribution of the number of busy virtual channels. In [18], Ould-Khaoua presented an accurate analytical model for Duato's fully adaptive routing algorithm [5] and employed the Dally's method [3] for virtual channel multiplexing. For the purpose of verification and comparison, in this section, we will amend the model presented in [18] with our new approach and compare it with Dally's method. The implementations of both models are identical for both methods except for the calculation of the probability of busy virtual channels. This is to make sure that the differences are due to the different methods of calculating the distribution of the number of busy virtual channels. Furthermore, we will keep the same assumptions as in[18], which are outlined bellow for the purpose of completeness:



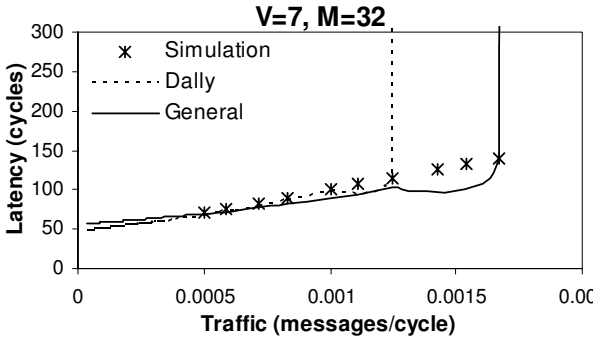
(a)



(b)

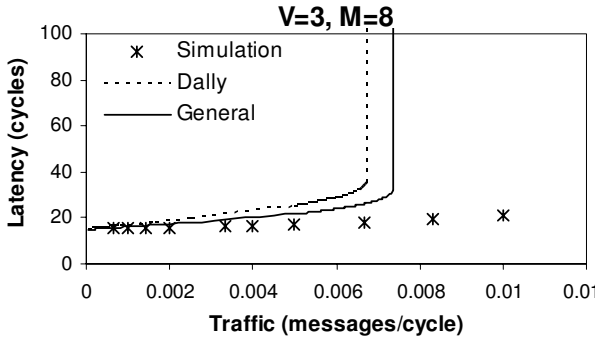


(c)

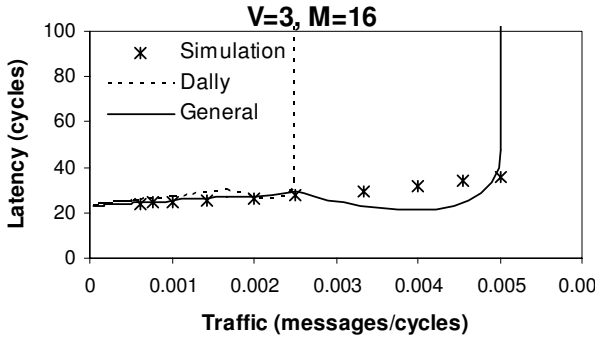


(d)

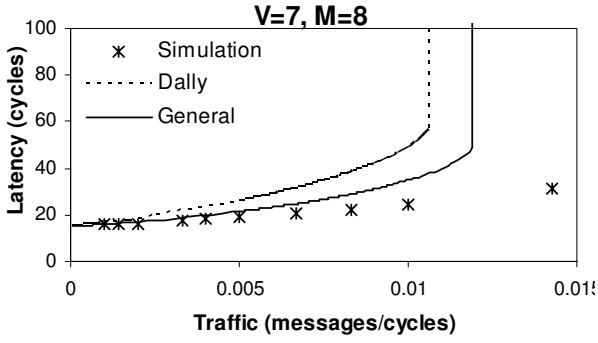
Fig. 4: Validation of the model and comparison with Dally's method for 16-ary 2-cube: (a) M=16, V=3, (b) M=32, V=3, (c) M=16, V=7, (d) M=32,V=7.



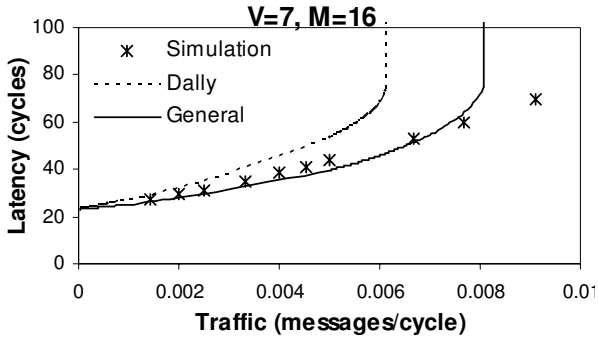
(a)



(b)



(c)



(d)

Fig. 5: Validation of the model and comparison with Dally's method for 8-ary 2-cube: (a) M=8, V=3, (b) M=16, V=3, (c) M=8, V=7, (d) M=16, V=7.

- Message destinations are uniformly distributed across the network nodes.
- Nodes generate traffic independently of each other and according to a Poisson process with mean rate of λ_g messages per node per cycle.
- Message length is fixed at M flits, each of which requires one-cycle to be transmitted from one router to another. Moreover, a message is long enough so that its data flits span from the source to the destination.
- The local queue at the injection channel in the source node has infinite capacity. Messages at the destination node are instantly transferred to the local PE as soon as they arrive.
- V (V>2) virtual channel are used per physical channel. In Duato's routing algorithm [5], class a contains (V-2) virtual channels, which are crossed adaptively, and class b contains two virtual channel, which are crossed deterministically (e.g. in an increasing order of dimensions). Let the channels in class a and b called adaptive and deterministic virtual channels respectively. When there is more than one adaptive virtual channel available, a message chooses one at random. Even though there are two deterministic virtual channels, a message can use only one at a time.

As we mentioned earlier, we kept the same implementation for both models except for the calculation of the probability of number of busy virtual channels. The model developed in [18] calculates the probability of busy virtual channels based on Dally's method by using equations (1 and 2) or alternatively, as we showed, by using equation (4). In this paper we used our new general method that is based on the M/G/1 queuing system. We experimented with different service time distributions. Our new model and Dally's model are both plotted against results obtained from an event-driven simulator that mimics the behaviour of a unidirectional wormhole-switched k-ary n-cube interconnection network. Each simulation experiment was run until the network reaches its steady state, that is, until a further increase in simulated job does not change the collected statistics appreciably. Numerous validation experiments have been performed for several combinations of network sizes, message lengths, and virtual channels

Figs 4 and 5 depict the mean message latency results predicted by both models plotted against those provided by the simulator as a function of the traffic injected in the network. The figures reveal that both models that are based on Dally's method and our new general method are in close agreement with simulation results under low traffic. However, as the injection rate increases, the model based on Dally's method starts to deviate from the simulation results. Meanwhile, the model that is based on our new general method continues to match the simulation results as the network approaches the saturation point. However, some discrepancies are still apparent due to the approximations made to ease the derivation of the model. Namely, the approximation we made to determine the variance and the approximation of the service time. Nevertheless, it can be concluded that our new general method of calculating the probability of busy virtual channels is, in one hand, more accurate than Dally's method, and on the other hand, it can be customize to adopt with different traffic conditions by using different distributions for the service time.

5. Conclusions

Almost all previous analytical models developed to assess the performance of wormhole-switched networks, employed a method presented by Dally to calculate the distribution of the number of busy virtual channels. Dally's method is accurate only under light traffic and degrades significantly as the traffic increases. In this study, we proposed a new general model to calculate the distribution of the number of busy virtual channels. Our model is based on an M/G/1 queue instead of a Markov process as opposed to Dally's method. We showed that Dally's method can be derived from our new general method and is equivalent to finding the queue size probability distribution of an M/M/1 queuing system. This explains the accuracy degradation of analytical models based on Dally's approach especially under moderate and high traffic. Beside the accuracy that it achieves under low, moderate and high traffic, a main advantage for our new approach is also the simplicity of adapting it to work with different traffic conditions and network setups by using different service time distribution.

The validity and the accuracy of the model has been demonstrated by comparing it to Dally's method as will as to a discrete event simulator. Results show that our model is more accurate than Dally's especially under moderate and high traffic. Furthermore, the design of the model is general so that it can be tailored to fit different traffic conditions and network topologies. Future research will focus on setting criteria to define what type of service time distributions to use under different traffic conditions and network setups. Numerical inversion of equation (6) and two moment approximation of the service time will also be studied in detail.

References

[1] W. Bolton, *Laplace and Z-Transforms*: Longman Publishing Group, 1997.

[2] Y. Boura, C. R. Das, and T. M. Jacob, "A performance model for adaptive routing in hypercubes," Proceedings of International Workshop Parallel Processing, 1994.

[3] W. J. Dally, "Virtual channel flow control," *IEEE Transactions on Parallel and Distributed Systems*, vol. 3, no. 2, pp. 194-205, 1992.

[4] W. J. Dally and C. L. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks," *IEEE Transactions on Computers*, vol. 36, no. 5, pp. 547-553, 1987.

[5] J. Duato, "A New Theory Of Deadlock-Free Adaptive Routing In Wormhole Networks," *IEEE Transactions On Parallel And Distributed Systems*, vol. 4, no. 12, pp. 1320-1331, 1993.

[6] J. Duato and P. Lopez, "Performance Evaluation of Adaptive Routing Algorithms for k-ary-n-cubes," Proceedings of First International Workshop on Parallel Computer Routing and Communication, 1994.

[7] J. Duato, S. Yalamanchili, and L. M. Ni, *Interconnection Networks: An Engineering Approach*. Los Alamitos: Morgan Kaufmann Publishers Inc., 2002.

[8] V. Halwan, F. Ozguner, and A. Dogan, "Routing in wormhole-switched clustered networks with applications to fault tolerance," *IEEE Transactions On Parallel And Distributed Systems*, vol. 10, no. 10, pp. 1001-1011, 1999.

[9] R. E. Kessler and J. L. Schwarzmeier, "CRAY T3D: a new dimension for Cray research," *1993 IEEE Compcon Spring*, vol., no., pp. 176, 1993.

[10] A. Khonsari and M. Ould-Khaoua, "Compressionless wormhole routing: An analysis for hypercube with virtual channels," *Computers and Electrical Engineering*, vol. 30, no. 1, pp. 45, 2004.

[11] A. Khonsari, H. Sarbazi-Azad, and M. Ould-Khaoua, "A Performance Model of Software-Based Deadlock Recovery Routing Algorithm in Hypercubes," *Parallel Processing Letters*, vol. 15, no. 1-2, pp. 153, 2005.

[12] A. Khonsari, A. Shahrabi, M. Ould-Khaoua, and H. Sarbazi-Azad, "Performance comparison of deadlock recovery and deadlock avoidance routing algorithms in wormhole-switched networks," *IEE Proceedings: Computers and Digital Techniques*, vol. 150, no. 2, pp. 97, 2003.

[13] L. Kleinrock, *Queueing Systems*, vol. 1. New York: John Wiley, 1975.

[14] S. Lee, "Real-time wormhole channels," *Journal Of Parallel And Distributed Computing*, vol. 63, no. 3, pp. 299-311, 2003.

[15] R. Nelson, *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling*. New York: Springer-Verlag, 1995.

[16] L. M. Ni and P. K. McKinley, "A Survey Of Wormhole Routing Techniques In Direct Networks," *Computer*, vol. 26, no. 2, pp. 62-76, 1993.

[17] M. D. Noakes, D. A. Wallach, and W. J. Dally, "J-Machine multicomputer. An architectural evaluation," *Conference Proceedings - Annual Symposium on Computer Architecture*, vol., no., pp. 224, 1993.

[18] M. Ould-Khaoua, "A performance model for Duato's fully adaptive routing algorithm in k-ary n-cubes," *IEEE Transactions On Computers*, vol. 48, no. 12, pp. 1297-1304, 1999.

[19] H. Sarbazi-Azad, M. Ould-Khaoua, and L. M. Mackenzie, "An accurate analytical model of adaptive wormhole routing in k-ary n-cubes interconnection networks," *Performance Evaluation*, vol. 43, no. 2-3, pp. 165-179, 2001.

[20] H. Sarbazi-Azad, M. Ould-Khaoua, and A. Y. Zomaya, "Design and performance of networks for super-, cluster-, and grid-computing: Part I," *Journal of Parallel and Distributed Computing*, vol. 65, no. 10, pp. 1119, 2005.