Performance Analysis of Wormhole Switching with Adaptive Routing in a Two-Dimensional Torus

M. Colajanni[1] *, B. Ciciani[2] ** and F. Quaglia[2] ***

[1] University of Modena, Italy
[2] University of Roma "La Sapienza", Italy

Abstract. This paper presents an analytical evaluation of the performance of wormhole switching with adaptive routing in a two-dimensional torus. Our analysis focuses on minimal and fully adaptive routing that allows a message to use any shortest path between source and destination.

1 Introduction

In the wormhole switching technique [2], [11] any message is partitioned into a set of flits. The transmission works as follows: after passing through a channel, the *header flit* tries to get another channel while the *data flits* are transmitted through the already obtained channels. In the case of channel contention, the message flits are stored in the flit buffers of the nodes along the already established path. A channel is released only when the last flit (*tail flit*) of the message is transmitted through it.

If wormhole is combined with adaptive routing, then there is no predefined path from the source to the destination. Since already obtained channels are released only at the end of the message transmission, a mechanism to prevent deadlock must be considered. This is done by using multiple *virtual channels* multiplexed on each physical link and forcing a pre-defined order on the allocation of virtual channels to messages [2], [11]. An adaptive routing policy allowing messages to use only shortest paths is called *minimal*. Moreover, an adaptive routing policy is called *fully adaptive* if a message is allowed to follow any path of the minimal (or non-minimal) class.

In this paper we propose an analytical approach to evaluate the performance of wormhole switching with minimal and fully adaptive routing in a two-dimensional torus. Most previous papers presenting analytical models consider either different switching techniques (such as *circuit-switching* [4] and *virtual-cut-through* [10]) or deterministic routing [1], [3], [5], [6]. To the best of our knowledge, an analytical evaluation of wormhole with adaptive routing is done

* colajanni@unimo.it
** ciciani@dis.uniroma1.it
*** quaglia@dis.uniroma1.it

only in [8]. Such analysis differs from our work as it considers Duato's adaptive routing [7] instead of minimal and fully adaptive routing.

The remainder of the paper is organized as follows. In Section 2 we present the system model. In Section 3 we propose a solution for the estimation of the mean latency time. In Section 4 we validate the analysis with some simulation results.

2 System Model

Each node (i, j) of the two-dimensional torus consists of a *processing element* $P_{i,j}$ and a *router node* $N_{i,j}$ with a controller per each dimension. Nodes are connected through bi-directional full-duplex links[1]. The flit size (bits) is equal to the number of wires B of a link. Therefore, each flit is transmitted in a single cycle link time. This time represents the basic temporal unit of our analysis.

We assume that generation times of messages at each processor are independent and identically distributed in accordance with a Poisson process with rate $1/\tau_{bit}$ (bit/cycle). This assumption and the symmetry of torus topology guarantee that the network is *balanced* [9] that is, we can assume that channels are equally likely to be visited independently of the message destination distribution. As in Dally's approach [5], we assume that the network has no virtual channels.

We consider the transmission of a message as consisting of two consecutive and separate phases: *path-hole* and *data-trail*. During the path-hole phase, the *header flit* builds the path from the source to the destination. All delays due to channel contentions are taken into account in this phase. The data-trail phase models the transmission of data flits along the channels of the selected path. Since we are assuming uniformly chosen destinations, the average length of the path along each dimension is $K = k/4$, where k is the number of nodes of each dimension. We define *average message*, a message which must travel on exactly K channels in each dimension.

Due to network symmetry and bi-directional links, it is possible to partition the torus into four quadrants (north-east, north-west, south-west, south-east), with the same number of nodes, such that the path of an average message belongs entirely to one quadrant. Our analysis focuses on the south-east quadrant. This quadrant is shown in Figure 1, where router nodes (i.e., circles) and channels (i.e., rectangular boxes) are represented through a two-dimensional array notation that identifies their position with respect to the average message's view that is, $N_{1,1}$ is the first router that the average message crosses, $X_{1,1}$ and $Y_{1,1}$ are the first horizontal and vertical channels available to that message, respectively.

Since each message can choose among four directions, the flow considered in the analysis represents $1/4$ of the entire flow generated by each node. Let $1/\tau_E = 1/(L\tau_{bit})$ be the message generation rate (messages/cycle) per each node, where $1/\tau_{bit}$ is the emission rate in bit/cycle, and L is the message average length. The average flow considered in the analysis has generation time $\tau = 4\tau_E$.

[1] The terms *link* and *channel* are used interchangeably.

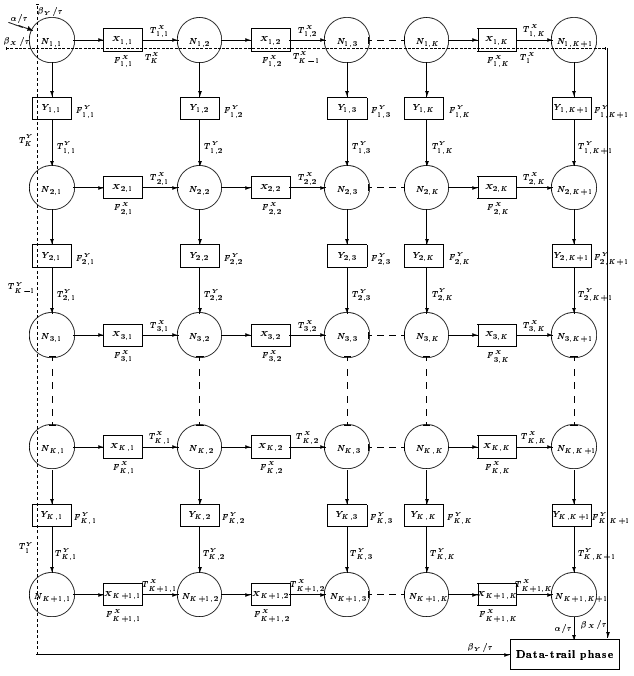

Fig. 1. Communication diagram for the *average message*

We can identify three flow streams: (i) the stream β^X denoting messages that use only channels of the dimension X; (ii) the stream β^Y denoting messages that use only channels of the dimension Y; (iii) the stream $\alpha = 1-(\beta^X+\beta^Y)$ denoting messages that can use adaptive path selection along both dimensions. As the traffic is uniform, the value of each stream can be estimated through the ratio between the number of nodes reachable by that stream and the total number of nodes: $\alpha = (k-1)^2/(k^2-1) = (k-1)/(k+1)$, $\beta^X = \beta^Y = (k-1)/(k^2-1) = 1/(k+1)$.

In Figure 1 plain lines denote the paths of the α stream, horizontal dotted lines refer to the β^X stream, and vertical dotted lines refer to the β^Y stream. $F^X_{i,j}$ represents the flow rate of the α stream that uses $X_{i,j}$. $T^X_{i,j}$ denotes the *residual transmission time* from $X_{i,j}$ that is, the mean time that the header flit takes to get from $X_{i,j}$ to the destination. Analogous notation is used for a vertical channel $Y_{i,j}$. A single parameter is sufficient for the identification of the residual transmission time of the streams β^X and β^Y, because they use channels along one dimension only. For example, T^X_{K-j+1} refers to the channel $X_{1,j}$ which is used by β^X, and T^Y_{K-i+1} refers to the channel $Y_{i,1}$ which is used by β^Y. The sub-index denotes the number of router nodes that the β stream has yet to cross.

3 The Analysis

When a message of the α stream reaches a node $N_{i,j}$, it attempts to continue along the dimension X. If the channel $X_{i,j}$ is not available (this happens with probability p_X), then the message tries to continue along the dimension Y. If the channel $Y_{i,j}$ is busy as well (this happens with probability p_Y), the message waits until at least one channel is released. The evaluation of p_X and p_Y is postponed. Let F be the adaptive flow reaching a node, F^X be the horizontal flow exiting that node, and F^Y be the vertical flow exiting that node. The following *bifurcation rule* holds: $F^X = F(1-p_X)/(1-p_Xp_Y) = Ff^X$ and $F^Y = Fp_X(1-p_Y)/(1-p_Xp_Y) = Ff^Y$ (due to space limitation the proof is omitted). By applying the bifurcation rule to the α stream, we obtain the equations for all the flows of the diagram in Figure 1. We partition the horizontal channels into six classes: the first channel $X_{1,1}$; the first row $X_{1,j}$ but the first channel; the first column $X_{i,1}$ but the first and the last channels; the intermediate channels $X_{i,j}$ but the first row, the first column and the last row; the last channel of the first column $X_{K+1,1}$; the last row $X_{K+1,j}$ but $X_{K+1,1}$. Using the bifurcation rule and Figure 1, we obtain the flow rates associated to each class of channels:

$$F^X_{1,1} = f^X\alpha/\tau \tag{1}$$

$$F^X_{1,j} = F^X_{1,j-1}f^X = (f^X)^j\alpha/\tau \qquad (j = 2,3,\ldots,K) \tag{2}$$

$$F^X_{i,1} = F^Y_{i-1,1}f^X = (f^Y)^{i-1}f^X\alpha/\tau \qquad (i = 2,3,\ldots,K) \tag{3}$$

$$F^X_{i,j} = F^X_{i,j-1}f^X + F^Y_{i-1,j}f^X \qquad (i,j = 2,3,\ldots,K) \tag{4}$$

$$F^X_{K+1,1} = F^Y_{K,1} = (f^Y)^K\alpha/\tau \tag{5}$$

$$F^X_{K+1,j} = F^X_{K+1,j-1} + F^Y_{K,j} \qquad (j = 2,3,\ldots,K) \tag{6}$$

Similarly, we partition the vertical channels into six classes, with the following flow rates associated to each class:

$$F^Y_{1,1} = f^Y\alpha/\tau \tag{7}$$

$$F^Y_{1,j} = F^X_{1,j-1}f^Y = (f^X)^{j-1}f^Y\alpha/\tau \qquad (j = 2,3,\ldots,K) \tag{8}$$

$$F^Y_{1,K+1} = F^X_{1,K} = (f^X)^K\alpha/\tau \tag{9}$$

$$F^Y_{i,1} = F^Y_{i-1,1}f^Y = (f^Y)^i\alpha/\tau \qquad (i = 2,3,\ldots,K) \tag{10}$$

$$F^Y_{i,j} = F^X_{i,j-1}f^Y + F^Y_{i-1,j}f^Y \qquad (i,j = 2,3,\ldots,K) \tag{11}$$

$$F^Y_{i,K+1} = F^X_{iK} + F^Y_{i-1,K+1} \qquad (i = 2,3,\ldots,K) \tag{12}$$

The mutual dependences existing among the previous equations can be solved analyzing the flow of each channel starting from $X_{1,1}$, and then following the west-east and north-south direction.

We model each link as an M/G/1 queue with multiple classes of flow indexed $m = 1,2,\ldots,M$. Assuming that messages of class m arrive with Poissonian rate λ_m and that their mean service time is $\overline{T}_m$, we obtain the following expression for the mean waiting time to get that channel [12, page 276]:

$$\overline{W} = \sum_{m=1}^{M}\frac{\rho_m}{\rho}E[W_m] = \frac{1}{2(1-\rho)}\sum_{m=1}^{M}\lambda_m\left(\overline{T}_m^2 + \sigma_m^2\right) \tag{13}$$

where σ_m^2 denotes the variance of $\overline{T}_m$, and $\rho = \sum_{m=1}^{M} \lambda_m \overline{T}_m$. The analysis focuses on messages that travel from $N_{1,1}$ to $N_{K+1,K+1}$ following the west-east and north-south directions that is, in the south-east quadrant. Hence, the waiting times of interest are W_{WE} and W_{NE} for the horizontal channels, and W_{NS} and W_{WS} for the vertical channels. We briefly present the steps for the estimation of W_{WE} for a generic horizontal channel $X_{i,j}$. The other waiting times are obtained through analogous considerations. The service time $\overline{T}_m$ (i.e., the link utilization times for the flow of messages of class m) will be evaluated later. To find W_{WE} we have to determine the M classes of flow that use $X_{i,j}$. Depending on the source and destination, that channel may represent the first horizontal channel requested by a message generated at $P_{i,j}$, or a channel of the last row of the diagram in Figure 1, or any horizontal channel in the middle of the same diagram. We have obtained for W_{WE} the following expression where $(S_{i,j}^X)^2 = [(\overline{T}_{i,j}^X)^2 + (\sigma_{i,j}^X)^2]$ (equations for the other waiting times are analogous):

$$W_{WE} = \frac{\sum_{j=1}^{K} F_{K,j}^Y (S_{K+1,j}^X)^2 + \sum_{i=2}^{K} \sum_{j=1}^{K} f^X F_{i-1,j}^Y (S_{i,j}^X)^2 + \frac{\beta^X (S_K^X)^2 + \alpha f^X (S_{1,1}^X)^2}{\tau}}{1 - \rho_{WE}} \quad (14)$$

A message holds a link for a period that we call *link utilization time*. Its value is equal to the residual transmission time of the message minus the number of flits already transmitted through that link. The following equations denote the link utilization time as a function of the link position in the average message transmission:

$$\overline{T}_j^X = T_j^X - j \qquad (j = 1, \ldots, K) \quad (15)$$

$$\overline{T}_i^Y = T_i^Y - i \qquad (i = 1, \ldots, K) \quad (16)$$

$$\overline{T}_{i,j}^X = T_{i,j}^X - (2K - i - j + 2) \qquad (i = 1, \ldots, K+1, j = 1, \ldots, K) \quad (17)$$

$$\overline{T}_{i,j}^Y = T_{i,j}^Y - (2K - i - j + 2) \qquad (i = 1, \ldots, K, j = 1, \ldots, K+1) \quad (18)$$

The general distribution assumed for the link utilization time would require also the specification of the variance of the service time $\overline{T}_{i,j}$, To this purpose, we observe that in wormhole switching the link utilization time is equal to the mean time to transmit the entire message plus the mean waiting time to obtain the remaining links of the path. The former term has a known distribution (that chosen for the message length), whereas the latter term and the covariance between these terms are unknown. We assume null covariance and exponential distribution for the mean waiting time.

The residual transmission time $T_{i,j}^X$ is the average time that the header of a message experiences while traveling from $X_{i,j}$ to the destination. Hence, $T_{latency}$ represents the residual transmission time of a message while traveling from the source node to the destination. To evaluate these residual transmission times, we adopt a backward analysis that moves from the *data trail* node of the diagram in Figure 1 back to the first line of horizontal and vertical channels.

The mean time T_{DT} to complete the *data trail* phase is a known parameter because it corresponds to the transmission time of all flits of an average message.

Therefore, we can write $T_{DT} = L/B$, where B denotes the number of wires per link, and L the average length of the message in flits.

For the messages following adaptive paths we distinguish four classes of horizontal channels: the last channel of the last row $X_{K+1,K}$; the last column $X_{i,K}$ but $X_{K+1,K}$; the last row $X_{K+1,j}$ but $X_{K+1,K}$; the other $X_{i,j}$. These classes have the following residual transmission time:

$$T^X_{K+1,K} = T_{DT} + 1 \tag{19}$$

$$T^X_{i,K} = W_{WS} + T^Y_{i,K+1} + 1 \qquad (i = 1, 2, \ldots, K) \tag{20}$$

$$T^X_{K+1,j} = W_{WE} + T^X_{K+1,j+1} + 1 \qquad (j = 1, 2, \ldots, K-1) \tag{21}$$

$$\begin{aligned} T^X_{i,j} = {} & (1 - p_X) T^X_{i,j+1} + p_X (1 - p_Y) T^Y_{i,j+1} + \\ & p_X p_Y \left[r(W_{WE} T^X_{i,j+1}) + (1 - r)(W_{WS} T^Y_{i,j+1}) \right] + 1 \\ & (i = 1, 2, \ldots, K; j = 1, 2, \ldots, K-1) \end{aligned} \tag{22}$$

The additional unit term denotes the time to transmit one data flit through a link. The residual transmission time in (22) for the intermediate links is obtained by adding the time values corresponding to the following events multiplied by the probability of their occurrence: the message continues along the dimension X; the message continues along the dimension Y; the message has found both channels busy, and continues when either one becomes free. The equations for the residual transmission time of the vertical channels are obtained in a similar way:

$$T^Y_{K,K+1} = T_{DT} + 1 \tag{23}$$

$$T^Y_{K,j} = W_{NE} + T^X_{K+1,j} + 1 \qquad (j = 1, 2, \ldots, K) \tag{24}$$

$$T^Y_{i,K+1} = W_{NS} + T^Y_{i+1,K+1} + 1 \qquad (i = 1, 2, \ldots, K-1) \tag{25}$$

$$\begin{aligned} T^Y_{i,j} = {} & (1 - p_X) T^X_{i+1,j} + p_X (1 - p_Y) T^Y_{i+1,j} + \\ & p_X p_Y \left[s(W_{NS} T^X_{i+1,j}) + (1 - s)(W_{NE} T^Y_{i+1,j}) \right] + 1 \\ & (i = 1, 2, \ldots, K-1; j = 1, 2, \ldots, K) \end{aligned} \tag{26}$$

The r and s terms in (22) and (26) are binary variables: if $W_{WS} < W_{WE}$, then $r = 0$, else $r = 1$; if $W_{NE} < W_{NS}$, then $s = 0$, else $s = 1$. The residual transmission times associated to the β^X and β^Y streams are given by:

$$T^X_1 = T^Y_1 = T_{DT} + 1 \tag{27}$$

$$T^X_j = W_{WE} + T^X_{j-1} + 1 \qquad (j = 2, \ldots, K) \tag{28}$$

$$T^Y_i = W_{NS} + T^Y_{i-1} + 1 \qquad (i = 2, \ldots, K) \tag{29}$$

Equations (27)–(29) are in recursive form. They can be solved through a backward flow analysis starting from T_{DT}, and then following the south-north and west-east direction, alternating horizontal and vertical channels.

The evaluation of the mean latency time requires an estimation of the probability of contention on the vertical and horizontal channels that is, p_X and p_Y.

These probabilities can be obtained as the sum of all flow rates multiplied by time values along horizontal and vertical channels, respectively. In addition to the flows that travel following the west-east direction in X and the north-south direction in Y (as shown by Figure 1), a horizontal channel is used also by the flows that follow the south-north and west-east direction. Analogously, a vertical channel is used also by the flows that follow the north-south and east-west direction. These contributions double the amount of messages on each link and motivate the multiplier terms two in the following equations:

$$p_X = 2 \sum_{i=1}^{K+1} \sum_{j=1}^{K} F_{i,j}^X \overline{T}_{i,j}^X + \frac{2\beta^X}{\tau} \sum_{j=1}^{K} \overline{T}_j^X \tag{30}$$

$$p_Y = 2 \sum_{i=1}^{K} \sum_{j=1}^{K+1} F_{i,j}^Y \overline{T}_{i,j}^Y + \frac{2\beta^Y}{\tau} \sum_{i=1}^{K} \overline{T}_i^Y \tag{31}$$

We evaluate the mean latency time as weighted sum of the residual transmission time values experienced by α, β^X, β^Y that is:

$$T_{latency}(\tau_E) = \alpha T^\alpha + \beta^X \left(T_K^X + W_{WE} + W_{NE} \right) + \beta^Y \left(T_K^Y + W_{NS} + W_{WS} \right) \tag{32}$$

where $T^\alpha = (1 - p_X) T_{1,1}^X + p_X (1 - p_Y) T_{1,1}^Y + p_X p_Y [v(W_{WE} + W_{NE} + T_{1,1}^X) + (1 - v)(W_{NS} + W_{WS} + T_{1,1}^Y)]$. The term v is a binary variable that is, if $(W_{WE} + W_{NE}) < (W_{NS} + W_{WS})$, then $v = 1$, else $v = 0$. The mutual dependency existing among some variables does not permit to achieve a closed formula for the mean latency time. However, a simple backward computation provides any performance value in few iterative steps.

4 Analysis Validation

In this section we validate the analytical model through a discrete event simulator. The Independent Replication Method was used to obtain confidence intervals at 95% level of confidence.

We report values related to four different network dimensions: 4×4, 8×8, 12×12 and 16×16. The average message length is 12 flit, hence we validate the model in the case of: average message length greater than the average path length, average message length equal to the average path length, and average message length smaller than the average path length. The message generation is modeled as a Poisson process with τ_E time cycles per node, and the message destination is an uniformly distributed random variable.

The tables below show the analytical and simulation latency times for the transmission of a message. The results show that the latency evaluated through the model is a good approximation of that obtained through simulation. In particular, the error is under 6% for low and medium arrival rates and under 12% for arrival rates close to the network saturation point.

Gen. rate per node	4 × 4			8 × 8		
	Simulation	**Model**	Error(%)	**Simulation**	**Model**	Error(%)
0.001	13.43	13.65	+1.6	15.55	15.73	+1.1
0.002	13.58	13.70	+0.9	15.96	15.92	-0.3
0.003	13.68	13.75	+0.5	16.27	16.11	-1.0
0.004	13.89	13.81	-0.6	16.81	16.31	-2.9
0.005	14.14	13.87	-1.9	17.10	16.51	-4.6
0.006	14.32	13.93	-2.7	17.66	16.72	-5.3
0.007	14.53	13.98	-3.8	18.15	16.94	-6.6
0.008	14.73	14.04	-4.7	18.65	17.18	-7.9
0.009	14.89	14.10	-5.3	19.14	17.42	-8.9
0.010	15.06	14.17	-5.9	19.52	17.69	-9.4
0.011	15.29	14.23	-6.9	20.12	17.97	-10.7
0.015	16.10	14.50	-9.9	22.18	19.39	-12.6

Gen. rate per node	12 × 12			16 × 16		
	Simulation	**Model**	Error(%)	**Simulation**	**Model**	Error(%)
0.001	17.79	17.90	+0.6	20.07	20.05	+0.1
0.002	18.43	18.31	-0.6	20.99	20.65	-1.6
0.003	19.09	18.76	-1.6	21.85	21.37	-2.2
0.004	19.88	19.27	-3.0	22.82	22.27	-2.5
0.005	20.73	19.87	-4.1	23.99	23.47	-2.2
0.006	21.33	20.58	-3.5	25.06	25.34	+1.1
0.007	22.15	21.40	-3.4	26.27	29.28	+11.4
0.008	22.65	22.52	-0.6	–	–	–
0.009	23.25	24.20	+4.0	–	–	–

References

1. V.S. Adve, and M.K. Vernon, "Performance analysis of mesh interconnection networks with deterministic routing", *IEEE Trans. on Parallel and Distributed Systems*, vol. 5, Mar. 1994, pp. 225–246.
2. K.M. Al-Tawil, M. Abd-El-Barr, and F. Ashraf, "A survey and comparison of wormhole routing techniques in mesh networks", *IEEE Network*, Mar.-Apr. 1997, pp. 38–45.
3. B. Ciciani, M. Colajanni, and C. Paolucci, "An accurate model for the performance analysis of deterministic wormhole routing", Proc. *IEEE 11th Int. Parallel Processing Symposium* (IPPS'97), Geneva, Switzerland, April 1997, pp. 353–359.
4. M. Colajanni, B. Ciciani, and S. Tucci, "Performance analysis of circuit-switching interconnection networks with deterministic and adaptive routing", *Performance Evaluation*, vol. 34, no. 1, Sept. 1998, pp. 1–26.
5. W.J. Dally, "Performance analysis of k-ary n-cube interconnection networks", *IEEE Trans. on Computers*, vol. 39, no. 6, June 1990, pp. 775–785.
6. J.T. Draper, and J. Gosh, "A comprehensive analytical model for wormhole routing in multicomputer systems", *J. Parallel and Distributed Computing*, vol. 23, no. 2, Nov. 1994, pp. 202–214.
7. J.Duato, "A new theory of deadlock free adaptive routing in wormhole routing networks", *IEEE Trans. on Parallel and Distributed Systems*, vol. 4, no. 12, 1993, pp. 1320–1331.
8. O. Khaoua, "An analytical model of Duato's fully-adaptive routing algorithm in k-ary n-cubes", Pro. *27th Int. Conference on Parallel Processing*, 1998.
9. P. Kermani, and L. Kleinrock, "Virtual cut-through: a new computer communication switching technique", *Computer Networks*, vol. 3, 1979, pp. 267–286.
10. A. Lagman, W.A. Najjar, S. Sur, and P.K. Srimani, "Evaluation of idealized adaptive routing on k-ary n-cubes", Proc. *IEEE Symp. on Parallel and Distributed Processing '94*, Dallas, TX, Dec. 1993, pp. 166–169.
11. L.M. Ni, and P.K. McKinley, "A survey of wormhole routing techniques in direct networks", *IEEE Computer*, vol. 26, no. 2, Feb. 1993, pp. 62–76.
12. H. Takagi, *Queueing Analysis - A Foundation of Performance Evaluation*, Elsevier-North-Holland, 1991.