An Analytical Model of Communication Networks in Multi-Cluster Systems in the Presence of Non-Uniform Traffic

Hojjat Sharifi, Mohammad K. Akbari, Bahman Javadi

Computer Engineering and Information Technology Department
Amirkabir University of Technology
P.O.Box 19514, Tehran, Iran
{hojjatsharifi,akbarif,javadi}@aut.ac.ir

Abstract

Several analytical models of interconnection networks of multi-cluster systems under uniform traffic pattern have been proposed in the literature. However, there has been hardly any work reported yet that deals with other important non-uniform traffic patterns in parallel applications. In this paper we propose a new analytical model based on fat-tree interconnection networks in the presence of traffic pattern generated by matrix-transpose permutation, which is an important communication operation in parallel applications such as matrix computation problems. The model is validated through comprehensive simulations, which demonstrated that the proposed model exhibit a good degree of accuracy for various system organizations and under different working conditions.

Keywords: Analytical Modeling, Multi-Cluster, Heterogeneity, Matrix-transpose traffic pattern, Latency.

1. Introduction

An increasing trend in the high performance computing (HPC) development is towards the networked distributed systems such as commodity-based cluster computing [1] and grid computing [2] systems. These network-based systems have proven to be cost-effective parallel processing tools for solving many complex scientific, engineering and commercial applications as compared with conventional supercomputing systems [3]. Advances in computational and communication technologies have made it economically feasible to conglomerate multiple independent clusters towards development of large-scale distributed systems, commonly referred to multi-cluster systems. Examples of production-level multi-cluster systems include the DAS-2 [4] and the LLNL multi-cluster system [5].

In this paper, we address the problem of communication networks performance modeling for multi-cluster computing systems. The study of interconnection networks is important because the overall performance of a distributed system is often critically hinged on the effectiveness of its interconnection network [6].

Although many works on network analysis employ an uniform reference model, it is not always appropriated in practice because there are many real-world applications that exhibit non-uniform traffic behavior. For instance, computing multi-dimensional FFTs, matrix problems and divide and conquer strategies exhibit regular communication patterns. Traffic patterns such as matrix-transpose, bit-reversal, shuffle, exchange and butterfly are examples of non-uniform traffic patterns [7].

Several analytical performance models of multi-computer systems have been proposed in the literature for different interconnection networks and routing algorithms (e.g., [9,10,11]). However, research activities regarding interconnection network for the system of interest is rare and most of the existing researches use homogenous cluster systems and the evaluations are confined to a single cluster system [12,13,14]. In contrast to these researches, our model: (1) considers multi-cluster computing systems in the presence of cluster size and network heterogeneity, (2) take into account variable message length, (3) exhibit non-uniform traffic pattern.

The rest of the paper is organized as follows. In Section 2, a brief overview of the multi-cluster system architecture and its communication issues are presented. In Section 3, detailed description of the proposed analytical model is discussed while section 4 validates the model using simulation results. We summarize our findings and conclude the paper in Section 5.

2. System Description

The multi-cluster computing system architecture used in this paper is shown in Figure 1. The system is made up of C clusters, each cluster is composed of N_i computing nodes. Moreover, each node comprising a processor and its associated memory module, $\{Pn_0, Pn_1, ..., Pn_{N_i-1}\}$.

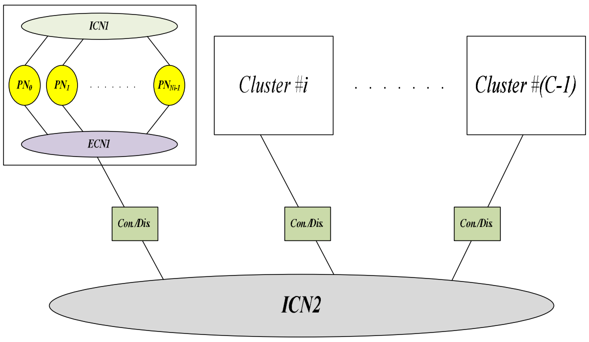

Figure 1. The Heterogeneous Multi-Cluster Architecture

Each cluster has two communication networks: an Intra-Cluster Network (ICN1) and an intEr-Cluster Network in level 1 (ECN1). The ICN1 is used for the purpose of message passing between nodes in the same cluster while the ECN1 is used to transmit messages between clusters as well as for the management of the entire system. All clusters interconnect to each other through Intra-Cluster Network in level 2. It should be noted that, ECN1 can be accessed directly by the nodes of each cluster without going through the ICN1 (see Figure 2). Also, the ECN1 and ICN2 are connected by a set of Concentrators/Distributors [15], which combine message traffic from/to one cluster to/from other cluster.

High performance computing clusters typically utilize *Constant Bisectional Bandwidth* (i.e., Fat-Tree) networks to construct large node count non-blocking switch configurations [5, 16]. In this paper we adopted *m*-port *n*-tree [17], which is a *Full Bisection Bandwidth* member of such networks to construct the topology for each cluster in the system. An *m*-port *n*-tree topology consists of $2(m/2)^n$ processing nodes and $(2n-1)(m/2)^{n-1}$ network switches. In this network the processing node is labeled as n-tuple $A = (A_0 A_1 ... A_{n-1})$ where $A \in \{0,1,...,m-1\} \times \{0,1,...,(m/2)-1\}^{n-1}$. In addition, each network switch itself has m communication ports $\{0,1,2,...,m-1\}$ that are attached to other switches or processing nodes.

Flow control and *routing algorithms* are other important components of a communication network. Routing algorithms establish the path between the source and the destination of a message. Since most of commercial network technologies adopted deterministic routing [18], we used a deterministic routing based on well-know Up*/Down* routing [19] which is proposed in [20]. In this algorithm, each message experiences two phases, *an ascending phase* to get a nearest common ancestor (NCA), followed by a *descending phase.*

3. The Proposed Analytical Model

In this section, we develop an analytic model for the multi-cluster system described in the pervious section. The proposed model is built on the basis of the following assumptions which are widely used in similar studies [9-14]:

1. There are two types of traffic in the network: “matrix-transpose” and “uniform”. When a message is generated it has a finite probability θ of being an external message and probability $1-\theta$ of being internal message. The external messages are destined to any other clusters in the system with equal probability. The internal messages are destined to a node within the cluster based on the matrix-transpose permutation.

2. Nodes generate traffic independent of each other, and which follows a Poisson process with a mean rate of λ messages per time unit.

3. The number of nodes in each cluster is different (N_i).

4. The network switches are input buffered and each channel is associated with a single flit buffer.

5. The network heterogeneity is presence between inter-cluster and Intra-cluster communication networks.

6. The message length is variable. Based on the reported measurements in [8], the most application programs have only two or the three distinct message sizes that processors send, so we adapted a weighted arithmetic mean as the average message length ($\overline{M}$ flits).

$$M = \sum_{i=1}^{f} M_i F_i \qquad (1)$$

Where M_i and F_i are message sizes and its probability, respectively. f is the number of distinct message lengths.

7. The source queue at the injection channel in the source node has infinite capacity. Moreover, messages are transferred to the node once they arrive at their destinations.

We have two types of connections in this topology, node to switch (or switch to node) and switch to switch. In the presence of network heterogeneity, we have two values for times to transmit. For intra-cluster networks the set of $\left(t_{cn}^{I1}, t_{cs}^{I1}\right)$ and for inter-cluster networks the set of $\left(t_{cn}^{E1}, t_{cs}^{E1}\right)$ and are adopted in the model. [16]

3.1. Traffic Pattern Analysis

There are two types of traffic in the system: matrix-transpose and uniform and it mainly affects the average message distance which is the expected number of links that a message traverses to reach its destination. For a newly generated message, the average number of links that the message traverses to reach its destination $\overline{d}^{(i)}$ is given by the following equation:

$$\overline{d}^{(i)} = \sum_{j=1}^{n_i} \left(2jP_{n_i,j}\right) \qquad (2)$$

Where $P_{n_i,j}$ is the probability of a message crossing $2j$-link (j-link in ascending and j-link in descending phase) to reach its destination. Different choices of $P_{n_i,j}$ lead to different distribution for message destination, and consequently different average message distance. As it is mentioned in assumption 1, the probability θ is defined as the number of external messages to the total number of messages. Note that, in the inter-cluster traffic pattern, an external message is destined to any other nodes in the system with equal probability.

In the *m*-port *n*-tree, with recalling that a node can not send a message to itself, the probability that a newly generated message makes $2j$-link with uniform distribution, $P_{n_i,j}^{u}$ can be defined as: [16]

$$P_{n_i,j}^{u} = \begin{cases} \dfrac{\left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{j-1}}{N_i - 1} & j = 1,2,\ldots,n_i - 1 \\ \dfrac{(m-1)\left(\dfrac{m}{2}\right)^{j-1}}{N_i - 1} & j = n_i \end{cases} \qquad (3)$$

To describe the matrix-transpose traffic pattern, let each node $PN(A_0, A_1, \ldots, A_{n_i-1})$ can also be labeled with a number such as Q which can be obtained as follows:

$$Q = A_0\left(\frac{m}{2}\right)^{n_i-1} + A_1\left(\frac{m}{2}\right)^{n_i-2} + \ldots + A_{n_i-2}\left(\frac{m}{2}\right) + A_{n_i-1} \qquad (4)$$

The binary representation of Q is $b_0 b_1 \ldots b_{n_i \log_2^{m/2}}$. For internal messages, we consider the matrix-transpose traffic pattern. In the traffic pattern generated according to the matrix-transpose permutation, a message generated in the source node $B = b_0 b_1 \ldots b_{n_i \log_2^{m/2}}$ is transferred to the destination node $D(B)$ as follows,

$$D(B) = \begin{cases} b_k b_{k+1} \ldots b_{2k-1} b_1 b_2 \ldots b_{k-1} & if \;\; n_i \log_2^{m/2} = 2k-1 \\ b_k b_{k+1} \ldots b_{2k} b_1 b_2 \ldots b_{k-1} & if \;\; n_i \log_2^{m/2} = 2k \end{cases} \qquad (5)$$

With this type of permutation the number of possible combinations that a message in cluster i crosses $2j$-link to reach its destination is as:

If n_i is even:

$$Count_j^{(i)} = \begin{cases} \left(\dfrac{m}{2}\right)^{j} - 2 & j = 1 \\ \left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{j-1} & j = 2,\ldots,(n_i+2)/2-1 \\ \left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{n_i-j-1} & j = (n_i+2)/2 \\ \left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{j-1} & j = (n_i+4)/2,\ldots,n_i-1 \\ (m-1)\left(\dfrac{m}{2}\right)^{j-1} & j = n_i \end{cases} \qquad (6)$$

If n_i odd and if z is odd:

$$Count_j^{(i)} = \begin{cases} \left(\dfrac{m}{2}\right)^{j} - 2 & j = 1 \\ \left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{j-1} & j = 2,\ldots,(n_i+1)/2-1 \\ \left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{n_i-j} & j = (n_i+1)/2 \\ \left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{j-1} & j = (n_i+3)/2,\ldots,n_i-1 \\ (m-1)\left(\dfrac{m}{2}\right)^{j-1} & j = n_i \end{cases} \qquad (7)$$

If n_i odd and if z is even:

$$Count_j^{(i)} = \begin{cases} 0 & j = 1,2,\ldots,(n_i+1)/2-1 \\ \left(\dfrac{m}{2}-2^{\frac{z}{2}}\right)\left(\dfrac{m}{2}\right)^{j-1} & j = (n_i+1)/2 \\ \left(\dfrac{m}{2}-1\right)\left(\dfrac{m}{2}\right)^{j-1} & j = (n_i+3)/2,\ldots,n_i-1 \\ (m-1)\left(\dfrac{m}{2}\right)^{j-1} & j = n_i \end{cases} \qquad (8)$$

Where $z = \log_2^{m}$, also, the probability of an internal message with matrix-transpose traffic pattern crossing $2j$-link to reach its destination in cluster i can be defined as:

$$P^m_{n_i,j} = \frac{Count^{(i)}_j}{N_i - Count^{(i)}_0} \tag{9}$$

Where $Count^{(i)}_0$ is the number of nodes that send the message to itself in cluster i and can be determined as follows,

$$Count^{(i)}_0 = 2^{\left\lceil \frac{1}{2}\log_2^{N_i} \right\rceil} \tag{10}$$

Since the *m*-port *n*-tree is not a node-symmetric topology, so it does not sufficient to analyze the traffic situation at a single node. Moreover in the presence of cluster size heterogeneity, this asymmetric problem must be solved for inter-cluster messages efficiently. The message flow model of the system is shown in Figure 2, where the path of a flit through various communication networks is illustrated. A processor in cluster i, which is shown as a circle in this figure, sends its request to ICN1$^{(i)}$ and ECN1$^{(i)}$ with probabilities $1-\theta$ and θ respectively. Where $i \in \{0,1,...,C-1\}$. The message path is depicted by arrows. Since the effective message rate of a processor in each cluster would be λ, so the rate of message received by each channel in the ICN1$^{(i)}$ can be obtained as follows:

$$\varphi^{(i)}_{I1} = \frac{(1-\theta)\lambda \overline{d}^{(i)}_{I1}}{4n_i} \tag{11}$$

The external message (uniform message) of cluster i leaves the ECN1$^{(i)}$ and crosses through the ICN2 and then goes to the ECN1$^{(v)}$ of the cluster v to reach its destination node. A simple way to deal with the asymmetric problem in the inter-cluster networks is compute the message rate from each cluster point of view and then averaging over all clusters. Therefore, the rate of message rate received by each channel in the inter-cluster networks can be driven as follows:

$$\varphi^{(i,v)}_{E1} = \left(1 + \frac{N_v}{N_i}\right) \times \frac{\theta\lambda_g \overline{d}^{(i)}_{E1}}{4n_i} \tag{12}$$

$$\varphi_{I2} = \frac{\theta \sum_{i=0}^{C-1} N_i \lambda \overline{d}_{I2}}{4n_c C} \tag{13}$$

Where $\overline{d}^{(i)}_{I1}$ is the average distance in the ICN1$^{(i)}$ and is given by Eq.(9). Also, $\overline{d}^{(i)}_{E1}$ and $\overline{d}_{I2}$ are the average distance in the ECN1$^{(i)}$ and ICN2, respectively and are given by Eq.(2). The n_c is the number of trees in the ICN2 and would be computed such that $C = 2(m/2)^{n_c}$.

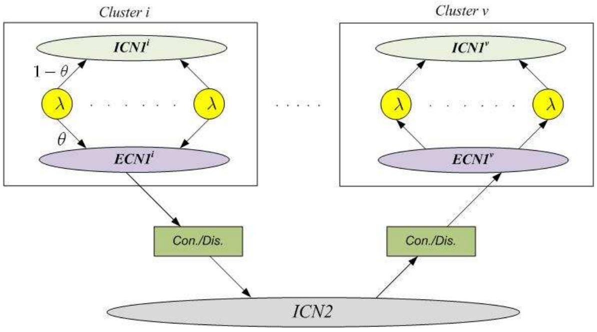

Figure. 2. Message flow model in the multi-cluster system between two typical clusters

3.2. Mean Message Latency for Intra-cluster Network

The mean latency seen by the matrix-transpose message, $T^{(i)}_m$, crossing from source node from cluster i to destination, consists of three parts; the mean waiting time at the source queue ($W^{(i)}_m$), the mean network latency ($S^{(i)}_m$), and the mean time for the tail flit to reach the destination ($R^{(i)}_b$). Hence,

$$T^{(i)}_m = W^{(i)}_m + S^{(i)}_m + R^{(i)}_m \tag{14}$$

At first, we find the mean network latency of intra-cluster network from cluster i point of view. Since each message may cross different number of links to reach its destination, we consider the network latency of an $2j$-link message as $S^{(i)}_j$, and averaging over all the possible nodes destined made by a message yields the mean network latency as:

$$S^{(i)}_m = \sum_{j=1}^{n_i} \left(P^m_{n_i,j} S^{(i)}_j \right) \tag{15}$$

Our analysis begins at the last stage and continues backward to the first stage. The network stage numbering is based on location of switches between the source and the destination nodes. It is obvious that in *m*-port *n*-tree topology, the number of stages for $2j$-link journey is $Y = 2j-1$. The destination, stage $Y-1$, is always able to receive a message, so the service time given to a message at the final stage is t^{I1}_{cn}. The service time at internal stages might be more because a channel would be idled when the channel of subsequent stage is busy. The mean service time of a channel at stage l is equal to the message transfer time and waiting time at subsequent stages to acquire a channel, so:

$$S_{l,j}^{(i)} = \begin{cases} Mt_{cn}^{I1} & l = Y-1 \\ \sum\limits_{h=l+1}^{Y-1} \left(W_{h,j}^{(i)}\right) + Mt_{cs}^{I1} & \text{otherwise} \end{cases} \tag{16}$$

According to this equation, the network latency for a message with $2j$-link journey equals to mean service time of a channel at stage 0. In this equation, $W_{h,j}^{(i)}$ is the mean waiting time seen by a $2j$-link message to acquire a channel at stage h from cluster i point of view. The mean waiting time depends on the probability of blocking at a given channel and on the mean service time of the channel. Consider a message that has to cross $2j$-link to reach its destination, suppose that this message reached in the stage h along its path. Let $P_{B_{h,j}}^{(i)}$ and $S_{h,j}^{(i)}$ denote the blocking probability of a $2j$-link message in stage h and the mean service time of a channel at stage h of the network from cluster i point of view. The mean waiting time is given by:

$$W_{h,j}^{(i)} = 1/2(S_{h,j}^{(i)} P_{B_{h,j}}^{(i)}) \tag{17}$$

The probability of channel blocking is determined using a birth-death Markov chain that is described in [25] and is as follows:

$$P_{B_{h,j}}^{(i)} = \varphi_{I1}^{(i)} S_{h,j}^{(i)} \tag{18}$$

An intra-cluster message originating from a given source node in cluster i sees a network latency of $S_m^{(i)}$ (given by Eq.(15)). Due to blocking situation that takes place in the network, the distribution function of message latency becomes general. Therefore, a channel at source node is modeled as an M/G/1 queue. So, as it has been shown in [21] the mean waiting time in the source queue becomes as follows, Where ξ_{I1} is the mean arrival rate on the network.

$$W_m^{(i)} = \frac{\xi_{I1}\left(S_m^{(i)}\right)^2 \left(1 + \frac{\left(S_m^{(i)} - Mt_{cn}^{I1}\right)^2}{\left(S_m^{(i)}\right)^2}\right)}{2\left(1 - \xi_{I1} S_m^{(i)}\right)} \tag{19}$$

$$\xi_{I1} = (1-\theta)\lambda \tag{20}$$

At last, the mean time for the tail to reach the destination can be written by the following equation:

$$R_m^{(i)} = \sum_{j=1}^{n_i} \left(P_{j,n_i} \left(\sum_{a=1}^{Y-1} t_{cs}^{I1} + t_{cn}^{I1} \right) \right) \tag{21}$$

3.3. Mean Message Latency for Inter-cluster Networks

External messages cross through both networks, ECN1$^{(i)}$ and ICN2, to get to their destination. Since the flow control mechanism is wormhole, the latency of these networks should be calculated as a merge one. Therefore, based on the Eq.(15) we can write [21],

$$S_{ex}^{(i,v)} = \sum_{j=1}^{n_i}\sum_{r=1}^{n_v}\sum_{l=1}^{n_c} \left(P_{(j,r,l)} S_{(j,r,l)}^{(i,v)} \right) \tag{22}$$

It means each external message cross $(j+r)$-link through the ECN1 networks (j-link in the source cluster i and r-link in the destination cluster v) and $2l$-link in the ICN2 to reach its destination. Also the probability of $P_{(j,r,l)}$ would be,

$$P_{(j,r,l)} = P_{i,n_i} P_{r,n_v} P_{l,n_c} \tag{23}$$

Based on the Eq.(16), we can drive the mean service time of a channel at stage h, where $0 \le h \le Y-1$, for inter-cluster networks as follows:

$$S_{h,(j,r,l)}^{(i,v)} = \begin{cases} Mt_{cn}^{E1} & h = Y-1 \\ \sum\limits_{k=h+1}^{Y-1} \left(W_{k,(j,r,l)}^{(i,v)}\right) + Mt_{cs} & \text{otherwise} \end{cases} \tag{24}$$

Where t_{cs} can be written based on the time to transmit of each flit in the correspondence network as fallows,

$$t_{cs} = \begin{cases} t_{cs}^{I2} & j \le h < j + 2l - 1 \\ t_{cs}^{E2} & \textit{otherwise} \end{cases} \tag{25}$$

Similar to the intra-cluster network, the latency for an external message equals to the mean service time of a channel at the first stage, i.e., $S_{0,(j,r,l)}^{(i,v)}$.

The uniform messages in the inter-cluster networks traverse ECN1$^{(i)}$ and then ICN2 . the uniform messages cross $(j+l)-1$ stages in the ascending and $(r+l)$ in the descending phase. So, based on Eq.(17) the mean amount of time that a message waits to acquire a channel at stage h, in the inter-cluster networks, is as follows:

$$W_{h,(j,r,h)}^{(i,v)} = 1/2(S_{h,(j,r,l)}^{(i,v)} P_{B_{h,(j,r,l)}}^{(i,v)}) \tag{26}$$

Where $P_{B_{h,(j,r,l)}}^{(i,v)}$ denote the blocking probability of a uniform message from cluster i to cluster v in stage h . This probability can be found similar to intra-cluster network with slightly modification by the following equation:

$$P^{(i,v)}_{B_{h,(j,r,l)}} = \varphi_h^{(i,v)} S^{(i,v)}_{h,(j,r,l)} \tag{27}$$

Where the channel rate is driven based on the current position of a message in each network by the following equation:

$$\varphi_h^{(i,v)} = \begin{cases} \varphi_{\mathrm{I2}} & j \le h < j + 2l - 1 \\ \varphi_{E1}^{(i,v)} & \text{otherwise} \end{cases} \tag{28}$$

As before, the source queue is modeled as an M/G/1 queue and the same method is used to approximate the variance of service time. Thus, the mean waiting time of the source queue in the inter-cluster networks can be calculated by Eq.(29). Where ξ_{E1} is the mean arrival rate on the network :

$$W_{ex}^{(i,v)} = \frac{\xi_{E1}\left(S_{ex}^{(i,v)}\right)^2 \left(1 + \frac{\left(S_{ex}^{(i,v)} - Mt_{cn}^{E2}\right)^2}{\left(S_{ex}^{(i,v)}\right)^2}\right)}{2\left(1 - \xi_{E1} S_{ex}^{(i,v)}\right)} \tag{29}$$

$$\xi_{E1} = \theta\lambda \tag{30}$$

As the last part, the mean time for the tail flit to reach the destination $R_{ex}^{(i,v)}$ is given by the following equation:

$$R_{ex}^{(i,v)} = \sum_{j=1}^{n_i}\sum_{r=1}^{n_v}\sum_{l=1}^{n_c}\left(P_{(j,r,l)}\left(\sum_{a=1}^{j+r-2} t_{cs}^{E1} + \sum_{a=0}^{2l-1} t_{cs}^{I2} + t_{cn}^{E1}\right)\right) \tag{31}$$

Finally, the arithmetic average of all latencies which the message from cluster i to all other clusters, namely cluster v, might be seen gives the message latency of inter-cluster networks as follows:

$$T_{ex}^{(i)} = \frac{1}{C-1}\sum_{v=0, v\neq i}^{C-1}\left(W_{ex}^{(i,v)} + S_{ex}^{(i,v)} + R_{ex}^{(i,v)}\right) \tag{32}$$

The mean waiting time at the concentrator/distributor is calculated in a similar manner to that for the source queue (Eq.(19)). By modeling the concentrate buffers in the concentrator/distributor as an M/G/1 queue, the mean waiting time is given by following equation where ${\xi_{I2}}^{(i)}$ is the message rate received in ICN2 through cluster i:

$$\overline{W}^{(i)}_{con./dis.} = \frac{\xi_{I2}^{(i)}\left(Mt_{cs}^{I2}\right)^2}{2\left(1 - \xi_{I2}^{(i)} Mt_{cs}^{I2}\right)} \tag{33}$$

$$\xi_{I2}^{(i)} = N_i\theta\lambda \tag{34}$$

Also, we model the distribute buffers in the concentrator/distributor as an M/G/1 queue, with the same rate of concentrate buffers. So the mean waiting time is given similarly by the above equation and consequently the mean waiting time at the concentrator/distributor would be $2\overline{W}^{(i)}_{con/dis}$. Finally, the mean message latency from cluster i point of view can be found as:

$$\overline{T}^{(i)} = (1-\theta)\left(T_m^{(i)}\right) + \theta\left(T_{ex}^{(i)} + 2\overline{W}^{(i)}_{con./dis.}\right) \tag{35}$$

To calculate the total mean of message latency, we use a weighted arithmetic average as follows:

$$\overline{T} = \sum_{i=0}^{C-1}\left(\frac{N_i}{\sum_{l=0}^{C-1} N_l} \times \overline{T}^{(i)}\right) \tag{36}$$

4. Validation of the model

In order to validate the proposed model and justify the applied approximations, the model was simulated. We have developed a discrete-event simulator based on the OMNET++ simulation environment [22], the simulator uses the same assumptions as the analysis. Messages are generated at each node according to Poisson process with the mean inter-arrival rate of λ. The destination address for an external message is determined by using a uniform random number generator while internal messages are sending to nodes which their address patterns are the matrix-transpose permutation of the source nodes address patterns. Each packet is time-stamped after its generation. The request completion time is checked in every "*sink*" module at each node to compute the message latency. Each simulation experiment was run until the network reached its steady state, that is, until a further increase in simulation network cycles does not change the collected statistics appreciably. Extensive validation experiments have been performed for several combinations of cluster sizes, network sizes, message length and its probability, and network heterogeneity. The general conclusions have been found to be consistent across all the cases considered. However, for the sake of specific illustration, we provide results for the cases that are presented in Table 1. Also, two different sets of networks are used in validation experiments are shown in Table 2 and Table 3. For all cases the intra-cluster network used Net.1 and inter-cluster networks adopt Net.2 configuration.

4.1. Results and Discussions

The results of simulation and analysis are shown in Figure 6 to Figure 8 in which the mean message latencies are plotted against the offered traffic rate for three different system organizations with different message weights. The figures reveal that the analytical model predicts the mean message latency with a good degree of accuracy when the system is in the steady state region, that is, when it has not reached the *saturation* point. It is assumed that network enters the saturation region when the system utilization becomes greater or equal to one. However, there are discrepancies in the results provided by the model and the simulation when the system is under heavy traffic and approaches the saturation point. This is due to the approximations that have been made in the analysis to ease the model development. For instance, in this region the traffic on the links is not completely independent, as we assume in our analytical model. Also, one of the most significant term in the model under heavily loaded system, is the mean waiting time at the source queue. The approximation which is made to compute the variance of the service time received by a message at a given channel is a factor of the model inaccuracy. However, at light traffic the model differs in average from simulation by less than about 8 percent for different systems. Since, the most evaluation studies focus on network performance in the steady state regions, so we can conclude that the proposed model can be a practical evaluation tool that can help system designer to explore the design space and examine various design parameters.

Table 1. System organization for model validation

N	C	m	*Cluster organization*		
1144	32	8	$n_i = 1$ $i \in [0,14]$	$n_i = 2$ $i \in [15,26]$	$n_i = 3$ $i \in [27,31]$
592	16	4	$n_i = 3$ $i \in [0,6]$	$n_i = 4$ $i \in [7,9]$	$n_i = 5$ $i \in [10,15]$

Table 2. Network configuration for model validation

	Net.1		Net.2	
t_{cn}	0.375	0.736	0.517	1.029
t_{cs}	0.375	0.736	0.522	1.034
L_m	256	512	256	512

Table 3. Network configuration for model validation

	Net.1		Net.2	
t_{cn}	0.436	0.863	0.645	1.285
t_{cs}	0.436	0.863	0.65	1.29
L_m	256	512	256	512

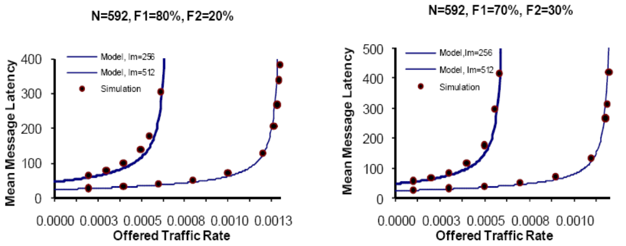

Figure 6. Mean message latency in a system, Net Table 2, $\theta = 0.6$, $M_1 = 32$ and $M_2 = 64$

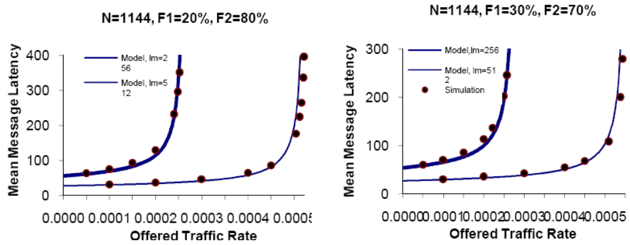

Figure 7. Mean message latency in a system, Net Table 2, $\theta = 0.5$, $M_1 = 32$ and $M_2 = 64$

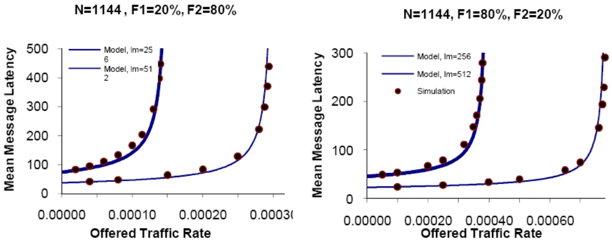

Figure 8. Mean message latency in a system, Net Table 3, left $\theta = 0.7$, right $\theta = 0.4$, $M_1 = 32$ and $M_2 = 64$

5. Conclusions

This paper has presented a new analytical model to compute message latency in the presence of non-uniform traffic based on fat-tree interconnection networks for heterogeneous multi-cluster systems. The model takes into account message and cluster size as well as network heterogeneity among clusters. The proposed model has been validated with versatile configurations and design parameters. Simulation experiments have proved that the model predicts message latency with a high degree of accuracy. It should be noted that at light traffic the model differs from simulation by less than about 6 percent. The simplicity and reasonable accuracy of the model make it an attractive tool for prediction the performance behavior of typical cluster and multi-cluster systems under different working conditions.

ACKNOWLEDGMENT

This project was partially supported by Iran Telecommunication Research Center (ITRC).

References

[1] M.Q. Xu, "Effective Meta-Computing using LSF Multi-Cluster". In *Proceedings of the IEEE International Conference on Cluster and Grid* (Brisbane, Australia, May 15-18), 2001, pp.100-106.

[2] I. Foster, "The Grid: A New Infrastructure for 21st Century Science". *Physics Today*, Vol.55, No.2 (Feb), 2002, pp.42-48.

[3] J. H. Abawajy, and S. P. Dandamudi. "Parallel Job Scheduling on Multi-Cluster Computing Systems". In *Proceedings of the IEEE International Conference on Cluster Computing* (Hong Kong, Dec. 1-4). 2003, pp.11-18.

[4] DAS-2 2002. The DAS-2 Supercomputer. http://www.cs.vu.nl/das2

[5] B. Boas, "Storage on the Lunatic Fringe". Lawrence Livermore National Laboratory, *Panel at Supercomputing Conference 2003* (Phoenix, AZ, Nov.15-21). 2003.

[6] A.T.T. Chun and C.L. Wang. "Contention-free Complete Exchange Algorithm on Clusters", In *Proceedings of the IEEE International Conference on Cluster Computing*, (Saxony, Germany, Nov. 28- Dec. 1), 2000, pp.57-64.

[7] M. Grammatikakis, D.F. Hsu, M. Kratzel Sibeyn, J. F Sibeyn, "Packet routing in fixed-connection networks": A survey, *Journal of Parallel and Distributed Computing*, Vol. 54, 1998, pp.77-132.

[8] J. Kim and D. J. Lilja, "Characterization of Communication Patterns in Message-Passing Parallel Scientific Application Programs", *Lecture Notes in Computer Science*, Vol. 1362, Springer-Verlag, 1998, pp.202-216.

[9] J.T. Draper and J. Ghosh, "A Comprehensive Analytical Model for Wormhole Routing in Multi-computer Systems", *Journal of Parallel and Distributed Computing*, Vol. 23, No.2, 1994, pp.202-214.

[10] Y.M. Boura and C.R. Das, "Performance Analysis of Buffering Schemes in Wormhole Routers", *IEEE Transactions on Computers*, Vol. 46, No. 6 (Jun). 1997, pp.687-694.

[11] A. Khonsari, H. Sarbazi-Azad, M. Ould-Khaoua, "An Analytical Model of Adaptive Wormhole Routing with Time-out", *Journal of Future Generation Computer Systems*, Vol. 19, No. 1, 2003, pp.1-12.

[12] X. Du, X. Zhang, and Z. Zhu, "Memory Hierarchy Consideration for Cost-Effective Cluster Computing", *IEEE Transaction on Computers*, Vol.49, No.5 (Sep), 2000, pp.915-933.

[13] B. Javadi, S. Khorsandi, and M. K. Akbari, "Study of Cluster-based Parallel Systems using Analytical Modeling and Simulation", *Lecture Notes in Computer Science*, Vol. 1911, Springer-Verlag, 2005, pp.1262-1271.

[14] P.C. Hu and L. Kleinrock, "A Queuing Model for Wormhole Routing with Timeout". In *Proceedings of the 4th International Conference on Computer Communications and Networks* (Nevada, LV, Sep.20-23). 1995, pp.584-593.

[15] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, Morgan Kaufmann Publisher, San Francisco. 2004.

[16] B. Javadi, M. K. Akbari, and J. H. Abawajy, "A performance model for analysis of heterogenous multi-cluster systems". *Journal of Parallel Computing*, Vol. 32, 2006, pp.831-851.

[17] X. Lin, An Efficient Communication Scheme for Fat-Tree Topology on Infiniband Networks, M.Sc Thesis, Department of Information Engineering and Computer Science, Feng Chia University, Taiwan. 2003.

[18] M. Koibuchi, K. Watanae, K. Kono, A. Jouraku, and H. Amano, "Performance Evaluation of Routing Algorithm in RHiNET-2 Cluster", In *Proceedings of the IEEE International Conference on Cluster Computing*, (Hong Kong, Dec.1-4), 2003, pp.395-402.

[19] M. D. Schroeder et. al. "Autonet: A High-Speed, Self Configuring Local Area Network Using Point-to-Point Links". SRC research report 59, Digital Equipment Corporation (Apr). 1990.

[20] B. Javadi, J. H. Abawajy, and M. K. Akbari , "Modeling and Analysis of Heterogeneous Loosely-Coupled Distributed Systems", Technical Report TR C06/1, School of Technology, Deakin University, Austra, Jan. 2006.

[21] B. Javadi, J. H. Abawajy, and M. K. Akbari, "Analytical modeling of interconnection networks in heterogeneous multi-cluster systems". *The Journal of super Computing*, Vol. 40, No. 1, 2006, pp.29-47.

[22] A. Varga, The OMNET++ discrete event simulation system, in: Proceedings of the European Simulation Multiconference, 2001.