Multi-Cluster Computing Interconnection Network Performance Modeling and Analysis

#Bahman Javadi[1], Mohammad K. Akbari[1], Jemal H. Abawajy[2], Saeid Nahavandi[2]

[1]*Computer Engineering and Information Technology Department, Amirkabir University of Technology*
424 Hafez Ave., Tehran, Iran, {javadi,akbari}@ce.aut.ac.ir

[2]*School of Engineering and Information Technology, Deakin University*
Geelong, VIC. 3217, Australia, {jemal,nahavand}@deakin.edu.au

Abstract

The overall performance of a distributed system is often depends on the effectiveness of its interconnection network. Thus, the study of the communication networks for distributed systems is very important, which is the focus of this paper. In particular, we address the problem of fat-tree based interconnection networks performance modeling for multi-user heterogeneous multi-cluster computing systems. To this end, we present an analytical model and validate the model through comprehensive simulation. The results of the simulation demonstrated that the proposed model exhibits a good degree of accuracy for various system organizations and under different working conditions.

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 to the conventional supercomputing systems [3]. Advances in computational and communication technologies has made it economically feasible to conglomerate multiple independent clusters leading to the development of large-scale distributed systems commonly referred to as 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 interconnection networks performance modeling for multi-cluster computing systems. Although simulation has been used to investigate the performance of various components of multi-cluster computing systems [3], we are interested in analytical modeling of interconnection networks. We present an analytical performance model of fat-tree based interconnection networks for multi-cluster computing systems in a multi-user environment. The model is based on probabilistic analysis and queuing network to analytically evaluate the performance of interconnection networks for cluster of clusters system. The model takes into account stochastic quantities as well as cluster sizes heterogeneity among clusters. The model is validated through comprehensive simulation, which demonstrated that the proposed model exhibits a good degree of accuracy for various system sizes and under different operating conditions.

Several analytical performance models of multi-computer systems have been proposed in the literature for different interconnection networks and routing algorithms (e.g., [6,7,8]). Unfortunately, little attention has been given to the cluster computing systems. Most of the existing researches are based on homogenous cluster systems and the evaluations are confined to a single cluster [9,10,11] with the exception of [12, 19], which looked at processor heterogeneity. In contrast, we focus on heterogeneous multi-cluster computing systems. Moreover, unlike all previous studies, we consider multi-cluster computing systems in multi-user environment.

The rest of the paper is organized as follows. In Section 2, a brief overview of the multi-cluster the system architecture is discussed. In Section 3, detailed description of the proposed analytical model is presented. The model verification and validation is discussed in Section 4. We summarize our findings and conclude the paper in Section 5.

2. MULTI-CLUSTER ARCHITECTURE

The multi-cluster computing system architecture used in this paper is shown in Fig. 1. The system is made up of C non-dedicated (i.e., multi-user) clusters, each cluster with different number of computing nodes (i.e., cluster size). Each cluster i is composed of N_i computing nodes, $i \hat{I} \{0,1,...,C - 1\}$, each node comprising a processor with computational power (s_i) (i.e., processors may be heterogeneous) and its associated memory module.

Each cluster has two communication networks: an Intra-Communication Network (*ICN*1) and an intEr-Communication Network (*ECN*1). The ICN1 is used for the purpose of message passing between processors in the same cluster while the ECN1 is used to transmit messages between clusters as well as for the management of the entire system. It should be noted that, ECN1 can be accessed directly by the processors of each cluster without going through the ICN1 (see Fig. 2). ECN1 and ICN2 are connected by a set of Concentrators/Dispatchers [13], which combine message traffic from/to one cluster to/from other cluster.

1-4244-0716-8/06/$20.00 ©2006 IEEE.

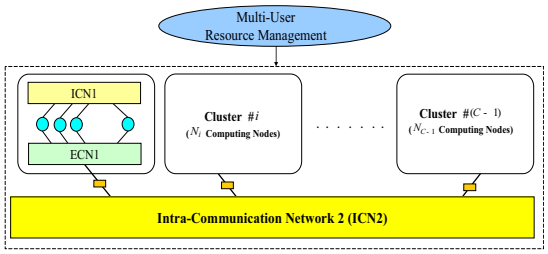

Fig. 1: The Heterogeneous Multi-Cluster Architecture

High performance computing clusters typically utilize *Constant Bisectional Bandwidth* (i.e., Fat-Tree) networks to construct large node count non-blocking switch configurations [4, 14]. In this paper we adopted *m*-port *n*-tree [15] as a fixed arity switches to construct the topology for each cluster in the system. An *m*-port *n*-tree topology consists of N processing nodes and N_{sw} network switches which can be calculated as follows:

$$N = 2 \times \left(\frac{m}{2}\right)^{n} \quad (1)$$

$$N_{sw} = (2n - 1) \times \left(\frac{m}{2}\right)^{n-1} \quad (2)$$

Flow control and *routing algorithms* are important components of a communication network. The flow control manages the allocation of resource to messages as they progress along their route. In this paper, we used the wormhole flow control, which is commonly used in cluster network technologies, e.g., Myrinet, Infiniband and QsNet [14]. *Routing algorithms* establish the path between the source and the destination of a message. Since the most of cluster network technologies adopted deterministic routing, we used a deterministic routing based on Up*/Down* routing [16] which is proposed in [17]. In this algorithm, each message experiences two phases, an *ascending phase* to get to a Nearest Common Ancestor (NCA), followed by a *descending phase*. Furthermore, since this algorithm performs a balanced traffic distribution, so the switch contention problem will be extinguished.

3. The Proposed Analytical Model

In this section, we develop an analytic model for the above mentioned cluster of clusters system. The proposed model is built on the basis of the following assumptions which are widely used in the similar studies [6-9]:

1. Nodes generate traffic independently of each other, and which follows a Poisson process with a mean rate of l_g messages per time unit.
2. The destination of each message would be any node in the system with uniform distribution.
3. The number of processors in each cluster is different (N_i) and the processing power of cluster's nodes are homogenous with the same computational power ($s_0 = s_1 = \ldots = s_{C-1}$).
4. The overhead of multi-user systems (e.g., context switch, etc.) is assumed to be t_i for all computing nodes in cluster *i*.
5. The network switches are input buffered and each channel is associated with a single flit buffer.
6. Message length is fixed (M flits).
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.

In this topology we have two types of connections, node to switch (or switch to node) and switch to switch. In the first and the last stage, we have node to switch and switch to node connection respectively. In the middle stages, the switch to switch connection is employed. Each type of connection has a service time which is approximated as follows:

$$t_{cn} = 0.5a_n + L_m b_n \quad (3)$$

$$t_{cs} = a_s + L_m b_n \quad (4)$$

Where t_{cn} and t_{cs} represent times to transmit from node to switch (or switch to node) and switch to switch connection, respectively. a_n and a_s are the network and switch latency, b_n is the transmission time of one byte (inverse of bandwidth) and L_m is the length of each flit in bytes.

The message flow model of the system is shown in Fig. 2, where the path of a flit through various communication networks is illustrated. As it is shown in this figure, we could find the mean message latency from cluster *i* point of view with the following equation:

$$\bar{l}^{(i)} = \left(1 - Q^{(i)}\right)\left(\bar{T}_{in-cluster}^{(i)}\right) + Q^{(i)}\left(\bar{T}_{out-cluster}^{(i)}\right) \quad (5)$$

where $Q^{(i)}$ is the probability of outgoing requests, $\bar{T}_{in-cluster}^{(i)}$ and $\bar{T}_{out-cluster}^{(i)}$ are the mean message latency in the intra-cluster network (i.e., ICN1) and the mean message latency in the inter-cluster networks (i.e., ECN1 and ICN2) respectively, from cluster *i* point of view.

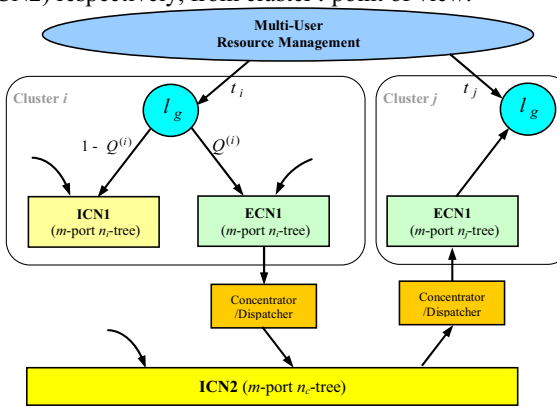

Fig. 2: Message flow model in the multi-cluster system

The probability $Q^{(i)}$ can be computed according to the assumption 3, by:

$$Q^{(i)} = 1 - \frac{N_i - 1}{N - 1} \tag{6}$$

In continue, to calculate the total mean of message latency in the system, we use a weighted arithmetic average as follows:

$$\bar{l} = \sum_{i=0}^{C-1} \left(\frac{N_i}{N} \times \bar{l}^{(i)} \right) \tag{7}$$

A Mean Message Latency of the Intra-Cluster Network

The mean message latency in each network contains three factors: the mean network latency, the average waiting time at the source node, and the average time for the last flit of a message to reach it destination. At first, we try to find the mean network latency of from cluster *i* point of view, $\bar{S}^{(i)}$. Since each message may cross different number of links to reach its destination, we consider the network latency of an *2h*-link message as $S_h^{(i)}$, and averaging over all the possible nodes destined made by a message yields the mean network latency as:

$$\bar{S}^{(i)} = \sum_{h=1}^{n_i} \left(P_h^{(i)} \times S_h^{(i)} \right) \tag{8}$$

where $P_h^{(i)}$ is the probability of a message which is originated from cluster *i* crossing 2*h*-link (*h*-link in ascending and *h*-link in descending phase) to reach its destination in a *m*-port n_i-tree topology. As it is mentioned in assumption 2, we take into account the uniform traffic pattern so, based on the *m*-port n_i-tree topology, we can define this probability as follows:

$$P_h^{(i)} = \begin{cases} \dfrac{\left(\frac{m}{2} - 1\right)\left(\frac{m}{2}\right)^{h-1}}{N_i - 1} & h = 1, 2, \ldots, n_i - 1 \\ \dfrac{(m-1)\left(\frac{m}{2}\right)^{h-1}}{N_i - 1} & h = n_i \end{cases} \tag{9}$$

As shown in the flow model, the processor requests will be directed to ICN1 and ECN1 by probabilities $1 - Q^{(i)}$ and $Q^{(i)}$ respectively, where $i \in \{0, 1, \ldots, C-1\}$. According to assumption 1, the request rate of a processor is λ_g per time unit, but assumption 5 implies the overhead due to the system is in the multi-user environment, so the effective request rate of a processor in cluster *i* would be approximately $(1 - t_i)^{-1}\lambda_g$. Therefore, the message rate received in the ICN1 can be obtained as follows:

$$\lambda_{I1}^{(i)} = \frac{N_i \left(1 - Q^{(i)}\right)}{(1 - t_i)} \lambda_g \tag{10}$$

Given that a newly generated message in cluster *i* makes *2h*-link to reach its destination with probability $P_h^{(i)}$, the average number of links that a message makes to reach its destination is given by:

$$d_{avg}^{(i)} = \sum_{h=1}^{n_i} \left(2h \times P_h^{(i)} \right) \tag{11}$$

By substituting of Eq.(9) in Eq.(11), the average message distance is obtained as,

$$d_{avg}^{(i)} = \frac{(mn_i - 2n_i - 1)\left(\frac{m}{2}\right)^{n_i} + 1}{\left[\left(\frac{m}{2}\right)^{n_i} - \frac{1}{2}\right]\left(\frac{m}{2} - 1\right)} \tag{12}$$

Consequently, we could derive the rate of received messages in each channel, which can be written as:

$$\lambda_{c(I1)}^{(i)} = \frac{\lambda_{I1}^{(i)} \times d_{avg(I1)}^{(i)}}{4 n_i N_i} \tag{13}$$

Our analysis begins at the last stage and continues backward to the first stage. The numbering starts from the stage next to the source node (stage 0) and goes up as we get closer to the destination node (stage $K - 1$). It is obvious that in *m*-port *n*-tree topology, the number of stages for *2h*-link journey is $K = 2h - 1$. The destination, stage $K - 1$, is always able to receive a message, so the service time given to a message at the final stage is t_{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 amount of time that a message waits to acquire a channel at stage *k* for cluster *i*, $W_{k,h}^{(i)}$, is given by the product of the channel blocking probability in stage *k*, $P_{B_{k,h}}^{(i)}$, and the mean service time of a channel at stage *k*, $S_{k,h}^{(i)}/2$ [13]:

$$W_{k,h}^{(i)} = \frac{1}{2} S_{k,h}^{(i)} P_{B_{k,h}}^{(i)} \tag{14}$$

The value of $P_{B_{k,h}}^{(i)}$ is determined using a birth-death Markov chain [18]. The rate of transition out and into the first state is $\lambda_{c(I1)}^{(i)}$ and $1/S_{k,h}^{(i)} - \lambda_{c(I1)}^{(i)}$ respectively. Solving this chain for the steady state probabilities gives:

$$P_{B_{k,h}}^{(i)} = \lambda_{c(I1)}^{(i)} S_{k,h}^{(i)} \tag{15}$$

The mean service time of a channel at stage *k* is equal to the message transfer time and waiting time at subsequent stages to acquire a channel, so:

$$S_{k,h}^{(i)} = \begin{cases} \sum_{a=k+1}^{K-1} \left(W_{a,h}^{(i)} \right) + Mt_{cs} & \text{otherwise} \\ Mt_{cn} & k = K - 1 \end{cases} \tag{16}$$

According to this equation, the network latency for a message with *2h*-link journey equals to $S_{0,h}^{(i)} (= S_h^{(i)})$.

A message originating from a given source node in cluster *i* sees a network latency of $\overline{S}^{(i)}$ (given by Eq.(8)). 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. The mean waiting time for an M/G/1 queue is given by [18]:

$$\overline{W}_s^{(i)} = \frac{l^{(i)} \left(\overline{x}^{2(i)} + s_{\overline{x}}^{2(i)} \right)}{2\left(1 - l^{(i)} \overline{x}^{(i)}\right)} \qquad (17)$$

Where $l^{(i)}$ is the mean arrival rate on the network, $\overline{x}^{(i)}$ is the mean service time, and $s_{\overline{x}}^{2(i)}$ is the variance of the service time distribution. Since the minimum service time of a message at the first stage is equal to Mt_{cn}, the variance of the service time distribution is approximated based on a method proposed by Draper and Ghosh [7] as follows:

$$s_{\overline{x}}^{2(i)} = \left(\overline{S}_{I1}^{(i)} - Mt_{cn} \right)^2 \qquad (18)$$

As a result, the mean waiting time in the source queue becomes,

$$\overline{W}_s^{(i)} = \frac{l_{I1}^{(i)} \left(\left(\overline{S}_{I1}^{(i)} \right)^2 + \left(\overline{S}_{I1}^{(i)} - Mt_{cn} \right)^2 \right)}{2\left(1 - l_{I1}^{(i)} \overline{S}_{I1}^{(i)}\right)} \qquad (19)$$

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

$$\overline{R}^{(i)} = \sum_{h=1}^{n_i} \left[P_h^{(i)} \cdot \left(\sum_{k=1}^{K-1} t_{cs} + t_{cn} \right) \right] \qquad (20)$$

Where $P_h^{(i)}$ can be computed by Eq.(9). The mean latency seen by the message, $\overline{T}^{(i)}$, crossing from source node from cluster *i* to destination, consists of three parts; the mean waiting time at the source queue ($\overline{W}_s^{(i)}$), the mean network latency ($\overline{S}_{I1}^{(i)}$), and the mean time for the tail to reach the destination ($\overline{R}^{(i)}$). Hence,

$$\overline{T}^{(i)} = \overline{W}_s^{(i)} + \overline{S}_{I1}^{(i)} + \overline{R}^{(i)} \qquad (21)$$

The mean message latency in the ICN1 from cluster *i* point of view, $\overline{T}_{I1}^{(i)}$, would be the intra-cluster message latency and is given by Eq.(21), that is $\overline{T}_{in-cluster}^{(i)} = \overline{T}_{I1}^{(i)}$.

B Mean Message Latency of the Inter-Cluster Networks

Similar to previous, we determine the same entity in inter-cluster networks. As mentioned before, external messages cross through both networks, ECN1 and ICN2, to get to the destination in other cluster. Since the flow control mechanism is wormhole, the latency of these networks should be calculated as a merge unit. Therefore based on the Eq.(8) we can write,

$$\overline{S}_{E1,I2}^{(i,j)} = \sum_{r=1}^{n_i} \sum_{v=1}^{n_j} \sum_{l=1}^{n_c} \left(P_{(r,v)+l}^{(i)} \cdot S_{(r,v)+l}^{(i,j)} \right) \qquad (22)$$

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

$$P_{(r,v)+l}^{(i)} = P_r^{(i)} \cdot P_v^{(i)} \cdot P_l^{(i)} \qquad (23)$$

Where all probabilities can be calculated by Eq.(9).

As before, we first find the message and channel rate in the networks. Based on the message flow model (Fig. 2), the external message (out of cluster) of cluster *i* leaves the ECN1 and crosses through the ICN2 and then goes to the ECN1 of the cluster *j* to reach its destination node. Therefore, the effective message rate received in each networks can be obtained as follows:

$$l_{E1}^{(i,j)} = \frac{N_i Q^{(i)}}{(1 - t_i)} l_g + \frac{N_j Q^{(j)}}{(1 - t_j)} l_g \qquad (24)$$

$$l_{I2}^{(i)} = \frac{N_i Q^{(i)}}{(1 - t_i)} l_g \qquad (25)$$

Consequently, the channel message rate can be written as:

$$l_{c(E1)}^{(i,j)} = \frac{l_{E1}^{(i,j)} \cdot d_{avg(E1)}^{(i)}}{4 n_i N_i} \qquad (26)$$

$$l_{c(I2)} = \frac{\sum_{i=0}^{C-1} l_{I2}^{(i)} \cdot d_{avg(I2)}}{4 C n_c} \qquad (27)$$

Where n_c is the number of trees in the ICN2 and would be computed such that $C = 2(m/2)^{n_c}$.

The message latency of inter-cluster networks from cluster *i* point of view can be found as the arithmetic average of all latencies which the message from cluster *i* to all other clusters, namely cluster *j*, might be seen as follows:

$$\overline{T}_{E1,I2}^{(i)} = \frac{1}{C - 1} \sum_{j=0, j \neq i}^{C-1} \left(\overline{T}_{E1,I2}^{(i,j)} \right) \qquad (28)$$

Where $\overline{T}_{E1,I2}^{(i,j)}$ would be determined with Eq.(21) by the following substitutions:

$$K = r + 2l + v - 1 \qquad (29)$$

$$l_c^{(i,j)} = \begin{cases} l_{c(I2)} & r \le k < r + 2l - 1 \\ l_{c(E1)}^{(i,j)} & \text{otherwise} \end{cases} \qquad (30)$$

$$l^{(i)} = l_{E1}^{(i,j)} \quad (31)$$

The mean waiting time at the concentrator/dispatcher is calculated in a similar manner to that for the source queue (Eq.(17)). The service time of the queue would be Mt_{cs} and there is no variance in the service time, since the messages length is fixed. By modeling of the injection channel in the concentrator/dispatcher as an M/G/1 queue, the mean waiting time is given by following equation:

$$\overline{W}_d^{(i)} = \frac{l_{I2}^{(i)} (Mt_{cs})^2}{2\left(1 - l_{I2}^{(i)} Mt_{cs}\right)} \quad (32)$$

Also, we model the ejection channel in the concentrator/dispatcher as an M/G/1 queue, with the same rate of injection channel. So, the total waiting time at the concentrator/dispatcher would be $2\overline{W}_d^{(i)}$.

The mean message latency in the inter-cluster networks from cluster *i* point of view can be found as:

$$\overline{T}_{out-cluster}^{(i)} = \overline{T}_{E1,I2}^{(i)} + 2\overline{W}_d^{(i)} \quad (33)$$

4. Model Validation

In order to validate the proposed model and justify the applied approximations, the model was simulated. Messages are generated at each node according to Poisson process with the mean inter-arrival rate of l_g. The destination node is determined by using a uniform random number generator. Each packet is time-stamped after its generation. For each simulation experiment, statistics were gathered for a total number of 100,000 messages. Statistic gathering was inhibited for the first 10,000 messages to avoid distortions due to the *warm-up* phase. Also, there is a *drain* phase at the end of simulation in which 10,000 generated messages were not in the statistic gathering to provide enough time for all packets to reach their destination. Extensive validation experiments have been performed for several combinations of clusters sizes, network sizes, network technologies, and message length. The general conclusions have been found to be consistent across all the cases considered. After all, to illustrate the result of some specific cases to show the validity of our model, the items which were examined carefully are presented in TABLE 1. Two different environments, a single-user environment (t_i = 0.0) and a multi-user environment with the t_i which are indicated in TABLE 1, are used for the model validation. Moreover, the two different message lengths, M=32 and 64 flits with different sizes, L_m=256 and 512 bytes are used. The network bandwidth is 500/time unit and network latency and switch latency are 0.02 and 0.01 time unit, respectively.

TABLE 1: SYSTEM ORGANIZATIONS FOR MODEL VALIDATION

N	*C*	*m*	*Cluster Organizations*		
1120	32	8	n_i=1 τ_i = 0.1 i∈[0,11]	n_i=2 τ_i = 0.15 i∈[12,27]	n_i=3 τ_i = 0.12 i∈[28,31]
544	16	4	n_i=3 τ_i = 0.1 i∈[0,7]	n_i=4 τ_i = 0.15 i∈[8,10]	n_i=5 τ_i = 0.12 i∈[11,15]

The results of simulation and analysis are shown in Fig. 3 and Fig. 4 in which the mean message latencies are plotted against the traffic generation rate for two different system organizations in two different environments.

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. 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 average waiting time at the source queue and concentrators/dispatchers. The approximation which is made to compute the variance of the service time received by a message at a given channel (Eq.(18)) is a factor of the model inaccuracy. 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.

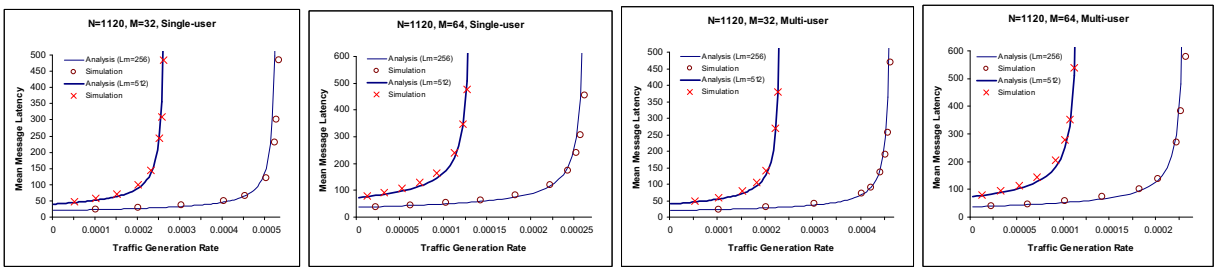

Fig. 3: Mean message latency in a single-user and multi-user system with N=1120, M=32, 64

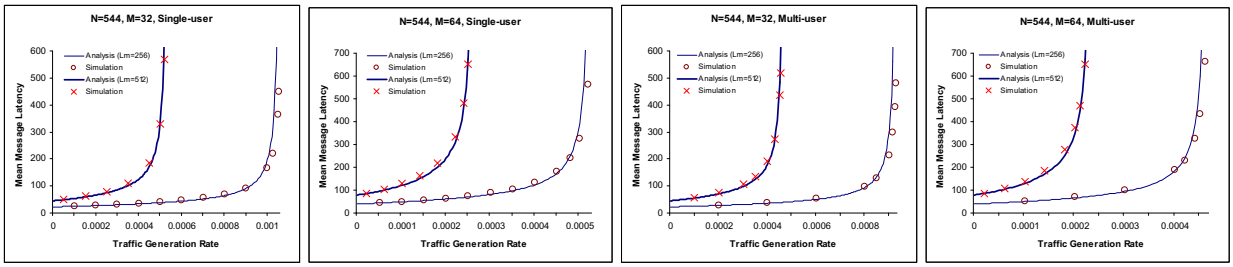

Fig. 4: Mean message latency in a single-user and multi-user system with N=544, M=32, 64

5. CONCLUSIONS

Analytical models play a crucial role in evaluation of a system under various design issues. In this paper, an analytical model of fat-tree based interconnection networks for heterogeneous multi-cluster computing systems is discussed. The proposed model has been validated with versatile configurations and design parameters. Simulation experiments have proved that the model predicts message latency with a reasonable accuracy. For future work, we intent to take the non-uniform traffic pattern into account, which is closer to the real traffic in such systems.

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] H. Sarbazi-Azad, A. Khonsari, and M. Ould-Khaoua. "Performance Analysis of Deterministic Routing in Wormhole *k*-ary *n*-cubes with Virtual Channels", *Journal of Interconnection Networks*, Vol. 3, Nos.1&2, 2002, pp.67-83.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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. 3483, Springer-Verlag, 2005, pp.1262-1271.

[12] A. Clematis and A. Corana, "Modeling Performance of Heterogeneous Parallel Computing Systems", *Journal of Parallel Computing*, Vol.25, No.9 (Sep), 1999, pp.1131-1145.

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

[14] "Building Scalable, High Performance Cluster/Grid Networks: The Role of Ethernet", White Paper, Force10 Networks Inc., Milpitas, CA, 2004.

[15] 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.

[16] 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.

[17] 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 Information Technology, Deakin University, Australia (Jan.). 2006.

[18] L. Kleinrock, *Queuing System: Computer Applications*, Vol.2, John Wily Publisher, New York. 1975.

[19] B. Javadi, J. H. Abawajy, and M. K. Akabri, "Analytical Modeling of Communication Latency in Multi-Cluster Systems", In *the Second IEEE International Workshop on Performance Modeling and Analysis of Communication in Wired and Wireless Networks*, Minneapolis, MN, USA, 2006, pp. 9-14.