# Analysis of Interconnection Networks in Heterogeneous Multi-Cluster Systems

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

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

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

## Abstract

*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. In the mean time, the heterogeneity is one of the most important factors of such systems. This paper addresses the problem of interconnection networks performance modeling of large-scale distributed systems with emphases on heterogeneous multi-cluster computing systems. So, we present an analytical model to predict message latency in multi-cluster systems in the presence of cluster size heterogeneity. The model is validated through comprehensive simulation, which demonstrates 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. Due to advances in computational and communication technologies, it is economically feasible to conglomerate multiple clusters to development of large-scale distributed systems known as multi-cluster systems. These systems are gaining momentum both in academic and commercial sectors and a wide variety of parallel applications are being hosted on such systems as well [3-5].

In this paper, we focus on the interconnection networks 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. Also, the interconnection network design plays a central role in the design and development of multi-cluster computing systems. Simulation has been used to investigate the performance of various components of multi-cluster computing systems [3]. Instead, we focus on analytic model.

Several analytical performance models of multi-computer systems have been proposed in the literature for different interconnection networks and routing algorithms (e.g., [6-9]). 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 [10-12]. A general model based on queuing networks was proposed for a single cluster computing in [10]. The model assumes that the processors are homogenous. Also, extensive numerical calculation of the model renders it too complicated. Furthermore, the model cannot be used for multi-cluster computing systems in the presence of heterogeneity. Also, a performance model for Network of Workstations with processor heterogeneity is discussed in [13]. The authors recently proposed an analytical model for multi-cluster systems in the presence of processor heterogeneity in [24, 25].

To this end, we present an analytical performance model of interconnection networks for multi-cluster computing systems. The model takes into account stochastic quantities as well as cluster sizes heterogeneity. 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.

The rest of the paper is organized as follows. In Section 2, a brief background is discussed. In Section 3, we give detailed description of the proposed analytical model. We present the model validation experiments in Section 4. We summarize our findings and conclude the paper in Section 5.

## 2. Background

The system under study in this paper is a multi-cluster computing systems which is made up of $C$ clusters, each cluster $i$ is composed of $N_i$ computing nodes, $i \in \{0,1,...,C-1\}$, each comprising a processor with processing power $\tau_i$ and its associated memory module as is depicted in Fig. 1. Also, each cluster has two communication networks, an Intra-Communication Network (*ICN*), which is used for the purpose of message passing between processors, and an intEr-Communication Network (*ECN*), which is used to transmit messages between clusters, management of the system, and also for the scalability of the system.

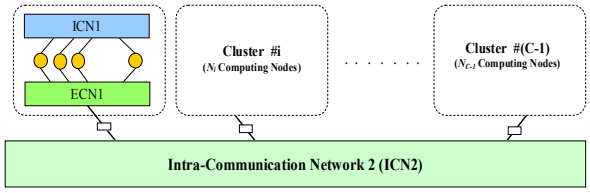


**Fig. 1. Heterogeneous multi-cluster system**

It should be noted that, ECN1 can be accessed directly by the processors of each cluster without going through the ICN1 (see Fig. 2). To interconnect ECN1 and ICN2, a set of Concentrators/Dispatchers [20] are used, which combine message traffic from/to one cluster to/from other cluster.

As mentioned before, the interconnection networks in such systems are crucial in gaining a desirable speedup. However, having a rapid network does not necessarily guarantee to obtain a good performance, due to contention problems. The contention problems which adversely affect the overall performance would happen in *host nodes*, *network link*s, and *network switches* [14]. Node contention happens when multiple data packets compete to contain a receive channel of a node, but link contention occurs when two or more packets share a communication link. The switch contention is due to unbalanced traffic flow through the switch, which would result in overflow of the switch buffer. The main factors which have impact on contention of an interconnection network and determine its performance are *Topology, Routing algorithm* and *Flow control mechanism.*

*Topology* defines how the network is physically connected together. High performance computing clusters typically utilize *Clos* networks, more commonly knows as "Fat-Tree" or *Constant Bisectional Bandwidth* networks to construct large node count non-blocking switch configurations [21, 22]. In this paper we adopted *m*-port *n*-tree [15] as a fixed arity switches to construct the topology for each cluster system. An *m*-port *n*-tree topology consists of $N$ processing nodes and $N_{sw}$ network switches which can be calculated as follows:

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

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

In addition, each network switch itself has $m$ communication ports $\{0,1,2,...,m-1\}$ that are attached to other switches or processing nodes. Every switch except root switches uses ports in the range of $\{0,1,2,...,(m/2)-1\}$ to have connection with its descendants or processing node, and using ports in the range of $\{(m/2),(m/2)+1,...,m-1\}$ for connection with its ancestors. It can be shown that the *m*-port *n*-tree is a full bisection bandwidth topology [18], so the link contention doesn't occur in such network.

*Routing algorithms* establish the path between the source and the destination of a message. The most commercial cluster networks (e.g. Myrinet, InfiniBand and QsNet) adopt deterministic routings [16]. The simplest deterministic routing used in such networks is Up*/Down* routing [17] which can be used in networks with both source and distributed routing. Of this, we used a deterministic routing based on Up*/Down* routing which is proposed in [18]. 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.

*Flow control* manages the allocation of resource to messages as they progress along their route. Two most famous flow control mechanisms are *store-and-forward* and *wormhole flow control* which are widely used in the commercial switches. Since the most dedicated cluster network technologies are using wormhole flow control, we adopt this mechanism to outline the analytical model.

In regards of existing heterogeneity in such systems and based on the discussions of [23], we categorized possible heterogeneity in multi-cluster systems as follows:

- Communication networks
- Processors computational power
- System organizations (e.g., cluster size)

It is obvious that develop an analytical model to cover all types of heterogeneity would be quite complicated.

Hence, in this paper we consider the last category (i.e., system organization) as it is illustrated in the following sections.

## 3. The Analytical Model

In this section, we develop an analytic model for the above mentioned multi-cluster system. The proposed model is built on the basis of the following assumptions which are widely used in similar studies [6-10]:

1. Nodes generate traffic independently of each other, and which follows a Poisson process with a mean rate of $\lambda_g$ messages per time unit. Moreover, the arrival process at a given channel of each network is approximated by an independent Poisson process.
2. The destination of each request 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 processing power ($\tau_0 = \tau_1 = ... = \tau_{C-1}$).
4. The network switches are input buffered and each channel is associated with a single flit buffer.
5. Message length is fixed (*M* flits).
6. 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.

### 3.1. Mean Network Latency

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

$$\overline{S}^{(i)} = \sum_{j=1}^{n_i} \left( P_{j,n_i} \times S_j^{(i)} \right) \tag{3}$$

Where $P_{j,n_i}$ is the probability of a message which is originated from cluster *i* crossing 2*j*-link (*j*-link in the ascending and *j*-link in the 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:

$$P_{j,n_i} = \begin{cases} \dfrac{\left(\frac{m}{2}-1\right)\left(\frac{m}{2}\right)^{j-1}}{N_i - 1} & j = 1,2,...,n_i - 1 \\ \dfrac{(m-1)\left(\frac{m}{2}\right)^{j-1}}{N_i - 1} & j = n_i \end{cases} \tag{4}$$

#### *3.1.1. Channel Message Rate*

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 shown in the model, the processor requests will be directed to ICN1 and ECN1 by probabilities $1 - P_o^{(i)}$ and $P_o^{(i)}$ respectively, where $i \in \{0,1,...,C-1\}$. The external message of cluster *i* leaves the ECN1 at the end of ascending phase and crosses through the ICN2 and then start the descending phase in the ECN1 of the cluster *v* to reach its destination node. Hence, it is like that a complete journey in the ECN1. Therefore, the message rate received in each networks can be obtained as follows:

$$\lambda_{I1}^{(i)} = N_i (1 - P_o^{(i)}) \lambda_g \tag{5}$$

$$\lambda_{E1}^{(i,v)} = N_i P_o^{(i)} \lambda_g + N_v P_o^{(v)} \lambda_g \tag{6}$$

$$\lambda_{I2}^{(i,v)} = \frac{\left( \left( N_i P_o^{(i)} \right) N_i + \left( N_v P_o^{(v)} \right) N_v \right) \lambda_g}{N_i + N_v} \tag{7}$$

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

$$d_{avg}^{(i)} = \sum_{j=1}^{n_i} \left( 2j \times P_{j,n_i} \right) \tag{8}$$

With substituting of Eq.(4) in Eq.(8), the average message distance is obtained as,

$$d_{avg}^{(i)} = \frac{(mn_i - 2n_i - 1)\left(\frac{m}{2}\right)^{n_i} + 1}{\left[\left(\frac{m}{2}\right)^{n_i} - \frac{1}{2}\right]\left(\frac{m}{2} - 1\right)} \tag{9}$$

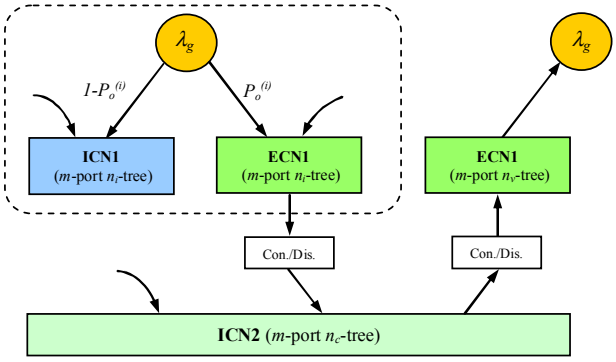


**Fig. 2. Message flow model in the system**

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

$$\eta_{I1}^{(i)} = \frac{\lambda_{I1}^{(i)} \times d_{avg(I1)}^{(i)}}{4n_i N_i} \tag{10}$$

$$\eta_{E1}^{(i,v)} = \frac{\lambda_{E1}^{(i,v)} \times d_{avg(E1)}^{(i)}}{4n_i N_i} \tag{11}$$

$$\eta_{I2}^{(i,v)} = \frac{\lambda_{I2}^{(i,v)} \times d_{avg(I2)}}{4n_c} \tag{12}$$

where $n_c$, the number of trees in the ICN2 compute such that $C = 2\times(m/2)^{n_c}$. As it is depicted in Fig. 2, the probability $P_o^{(i)}$ has been used as the probability of outgoing requests within cluster *i*. According to assumption 2, this parameter is computed by the following equation:

$$P_o^{(i)} = \frac{\sum_{j=0, j\neq i}^{C-1} N_j}{N-1} \tag{13}$$

#### *3.1.2. Mean Channel Service Time*

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:

$$t_{cn} = \frac{1}{2}\alpha_{net} + L_m\beta_{net} \tag{14}$$

$$t_{cs} = \alpha_{sw} + L_m\beta_{net} \tag{15}$$

where $t_{cn}$ and $t_{cs}$ represent times to transmit from node to switch (or switch to node) and switch to switch connection, respectively. $\alpha_{net}$ and $\alpha_{sw}$ are the network and switch latency, $\beta_{net}$ is the transmission time of one byte (inverse of bandwidth) and $L_m$ is the length of each flit in bytes.

Our analysis begins at the last stage and continues backward to the first stage. The number of stage for a message with *2j*-link journey is $K = 2j-1$. 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*, $W_{k,j}^{(i)}$, is given by the product of the channel blocking probability in stage *k*, $P_{B_{k,j}}^{(i)}$, and the mean service time of a channel at stage *k*, $S_{k,j}^{(i)}/2$ [25]:

$$W_{k,j}^{(i)} = \frac{1}{2} S_{k,j}^{(i)} P_{B_{k,j}}^{(i)} \tag{16}$$

The value of $P_{B_{k,j}}^{(i)}$ is determined using a birth-death Markov chain [25]. Solving this chain for the steady state probabilities gives:

$$P_{B_{k,j}}^{(i)} = \eta_k^{(i)} S_{k,j}^{(i)} \tag{17}$$

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:

$$S_{k,j}^{(i)} = \begin{cases} \sum_{s=k+1}^{K-1} \left(W_{s,j}^{(i)}\right) + Mt_{cs} & \text{otherwise} \\ Mt_{cn} & k = K-1 \end{cases} \tag{18}$$

According to this equation, the mean network latency is equal to $S_{0,j}^{(i)} \left(= S_j^{(i)}\right)$.

### 3.2. Mean Message Latency

A message originating from a given source node in cluster *i* sees a network latency of $\overline{S}^{(i)}$ (given by Eq.(3)). 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 [19]:

$$\overline{W}^{(i)} = \frac{\rho^{(i)}\overline{x}^{(i)}\left(1+C_{\overline{x}}^{2(i)}\right)}{2\left(1-\rho^{(i)}\right)} \tag{19}$$

$$\rho^{(i)} = \lambda^{(i)}\overline{x}^{(i)} \tag{20}$$

$$C_{\overline{x}}^{2(i)} = \frac{\sigma_{\overline{x}}^{2(i)}}{\overline{x}^{2(i)}} \tag{21}$$

where $\lambda^{(i)}$ is the mean arrival rate on the network, $\overline{x}^{(i)}$ is the mean service time, and $\sigma_{\overline{x}}^{2(i)}$ 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 [8] as follows:

$$\sigma_{\overline{x}}^{2(i)} = \left(\overline{S}^{(i)} - Mt_{cn}\right)^2 \tag{22}$$

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

$$\overline{W}^{(i)} = \frac{\lambda^{(i)}\left(\overline{S}^{(i)}\right)^2\left(1+\frac{\left(\overline{S}^{(i)}-Mt_{cn}\right)^2}{\left(\overline{S}^{(i)}\right)^2}\right)}{2\left(1-\lambda^{(i)}\overline{S}^{(i)}\right)} \tag{23}$$

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

$$\overline{R}^{(i)} = \sum_{j=1}^{n_i}\left[P_{j,n_i}\times\left(\sum_{k=1}^{K-1}t_{cs}+t_{cn}\right)\right] \tag{24}$$

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 ($\overline{W}^{(i)}$), the mean service time at the first stage ($\overline{S}^{(i)}$), and the mean time for the tail flit to reach the destination ($\overline{R}^{(i)}$). Hence,

$$\overline{T}^{(i)} = \overline{W}^{(i)} + \overline{S}^{(i)} + \overline{R}^{(i)} \tag{25}$$

The mean message latency in the ICN1 from cluster $i$ point of view, $\overline{T}_{I1}{}^{(i)}$, would be found by the above equation with substitution of $\eta_k{}^{(i)} = \eta_{I1}{}^{(i)}$ and $\lambda^{(i)} = \lambda_{I1}{}^{(i)}$.

### 3.3. Mean Message Latency for 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 one. Therefore based on the Eq.(3), we can write,

$$\overline{S}_{E1\&I2}{}^{(i,v)} = \sum_{j=1}^{n_i}\sum_{l=1}^{n_v}\sum_{h=1}^{n_c}\left(P_{(j,l)+h}\times S_{(j,l)+h}{}^{(i)}\right) \tag{26}$$

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

$$P_{(j,l)+h} = P_{j,n_i}\times P_{l,n_v}\times P_{h,n_c} \tag{27}$$

In the inter-cluster networks, the number of stages for each message journey is $K = j+2h+l-1$. Based on Eqs.(16) and (17), the mean amount of time that a message waits to acquire a channel at stage $k$, in the inter-cluster networks is as follows:

$$W_{k,(j,l)+h}{}^{(i,v)} = \frac{1}{2}\eta_k{}^{(i,v)}\left(S_{k,(j,l)+h}{}^{(i,v)}\right)^2 \tag{28}$$

Where the channel rate is driven by the following equation:

$$\eta_k{}^{(i,v)} = \begin{cases}\eta_{I2}{}^{(i,v)} & j \le k < j+2h-1 \\ \eta_{E1}{}^{(i,v)} & \text{otherwise}\end{cases} \tag{29}$$

Similar to the intra-cluster network, the network latency for an inter-cluster message equals to the mean service time of a channel at stage 0 and can be found by Eq.(18).

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 as:

$$\overline{W}_{E1\&I2}{}^{(i,v)} = \frac{\lambda_{E1}{}^{(i,v)}\left(\overline{S}_{E1\&I2}{}^{(i,v)}\right)^2\left(1+\frac{\left(\overline{S}_{E1\&I2}{}^{(i,v)}-Mt_{cn}\right)^2}{\left(\overline{S}_{E1\&I2}{}^{(i,v)}\right)^2}\right)}{2\left(1-\lambda_{E1}{}^{(i,v)}\overline{S}_{E1\&I2}{}^{(i,v)}\right)} \tag{30}$$

Finally, the arithmetic average is used to compute the mean message latency in the inter-cluster networks from cluster $i$ point of view, as follows:

$$\overline{T}_{E1\&I2}^{(i)} = \frac{\sum_{v=0,v\neq i}^{C-1}\left(\overline{W}_{E1\&I2}^{(i,v)} + \overline{S}_{E1\&I2}^{(i,v)} + \overline{R}_{E1\&I2}^{(i,v)}\right)}{C-1} \quad (31)$$

Where the mean time for the tail flit to reach the destination can be obtained as follows:

$$\overline{R}_{E1\&I2}^{(i,v)} = \sum_{j=1}^{n_i}\sum_{l=1}^{n_v}\sum_{h=1}^{n_c}\left[P_{(j,l)+h}\times\left(\sum_{k=1}^{K-1}t_{cs}+t_{cn}\right)\right] \quad (32)$$

The concentrator/dispatcher is working as simple bi-directional buffers to interface two external networks (i.e., ECN1 and ICN2). The mean waiting time at the concentrator/dispatcher is calculated in a similar manner to that for the source queue (Eq.(19)). 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. So, the mean waiting time are given by following equations:

$$\overline{W}_s^{(i,v)} = \frac{\lambda_{I2}^{(i,v)}\left(Mt_{cs}\right)^2}{2\left(1-\lambda_{I2}^{(i,v)}Mt_{cs}\right)} \quad (33)$$

Also, we model the dispatch buffers in the concentrator/dispatcher as an M/G/1 queue, with the same rate of concentrate buffers. So the mean waiting time is given similarly by Eq.(33). The arithmetic average of sum of the two above mentioned waiting times gives mean waiting time at the concentrator/dispatcher as follows:

$$\overline{W}_d^{(i)} = \frac{1}{C-1}\sum_{v=0,v\neq i}^{C-1}\left(2\overline{W}_s^{(i,v)}\right) \quad (34)$$

Putting all together, we could find the mean message latency from cluster $i$ point of view (based on Fig. 2) with the following equation:

$$\overline{\ell}^{(i)} = \left(1-P_o^{(i)}\right)\left(\overline{T}_{I1}^{(i)}\right) + P_o^{(i)}\left(\overline{T}_{E1\&I2}^{(i)} + \overline{W}_d^{(i)}\right) \quad (35)$$

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

$$\overline{\ell} = \sum_{i=0}^{C-1}\left(\frac{N_i}{N}\times\overline{\ell}^{(i)}\right) \quad (36)$$

## 4. Validation of the Model

In order to validate the proposed model and justify the applied approximations, the model was simulated. 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 $\lambda_g$. The destination node is determined by using a uniform random number generator. For each simulation experiment, statistics were gathered for a total number of 100,000 messages. Statistic gathering was discarded 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. 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 validation**

| *N* | *C* | *m* | *Node Organizations* | | |
|---|---|---|---|---|---|
| 1120 | 32 | 8 | $n_i$=1 i∈[0,11] | $n_i$=2 i∈[12,27] | $n_i$=3 i∈[28,31] |
| 544 | 16 | 4 | $n_i$=3 i∈[0,7] | $n_i$=4 i∈[8,10] | $n_i$=5 i∈[11,15] |

The results of simulation and analysis for a system with above mentioned parameters are depicted in Fig. 3 and Fig. 4 in which the mean message latencies are plotted against the offered traffic for two different system organizations.

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.(22)) 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.

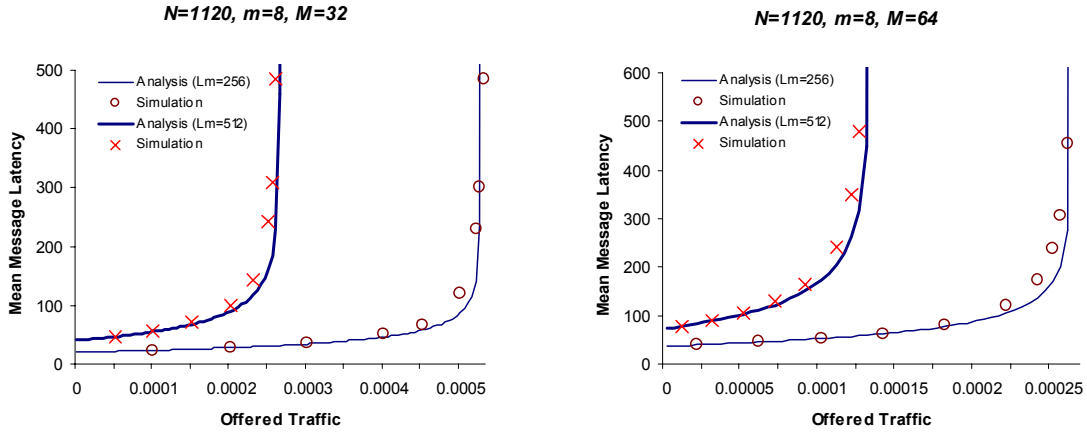


**Fig. 3. Mean message latency in a system with N=1120, M=32 and 64**

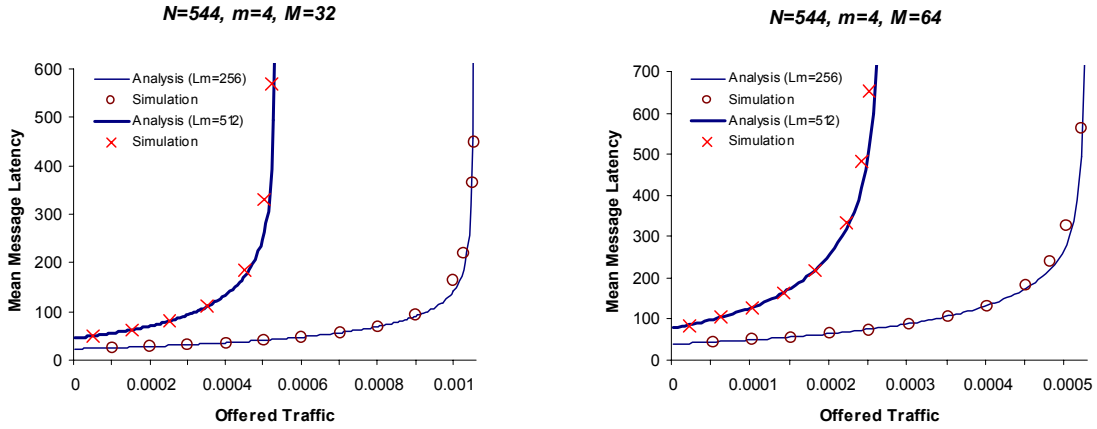


**Fig. 4. Mean message latency in a system with N=544, M=32 and 64**

## 5. Conclusions

Analytical models play a crucial role in evaluation of a system under various design issues. In this paper, an analytical model of interconnection networks for multi-cluster computing systems in the presence of cluster sizes heterogeneity 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 develop and extend the model to cover other categories of heterogeneity and non-uniform traffic pattern as well.

## Acknowledgments

Special thanks to Mr. D. Sgro and R. Ruge regarding the system setup for the simulations. The help of Dr. A. Khonsari from Tehran University for his worthwhile comments is appreciated. Financial support from Intelligent System Research Laboratory is also greatly appreciated.

## 6. 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 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. 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. 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] M. Ould-Khaoua "A Performance Model for Duato's Fully-adaptive Routing Algorithm in *k*-ary *n*-cubes", *IEEE Transaction on Computers*, Vol.42, No.12, 1999, pp. 1-8.

[8] 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.

[9] 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.

[10] 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. 1995, pp. 584-593.

[11] 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.

[12] 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.

[13] 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.

[14] 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. 2000, pp. 57-64.

[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. 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. 2003, pp. 395-402.

[17] 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.

[18] 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.

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

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

[21] "InfiniBand Clustering, Delivering Better Price/Performance than Ethernet", White Paper, Mellanox Technologies Inc., Santa Clara, CA, 2005.

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

[23] J. Dongarra and A. Lastovetsky, "An Overview of Heterogeneous High Performance and Grid Computing", In Engineering the Grid: Status and Perspective, Eds B. DiMartino, J. Dongarra, A. Hoisie, L. Yang, and H. Zima, American Scientific Publishers, Feb. 2006.

[24] B. Javadi, M. K. Akbari, and J. H. Abawajy, "Analysis of Multi-Cluster Computing Systems with Processor Heterogeneity", In *Proceedings of the 11th International Computer Society of Iran Computer Conference*, Tehran, Iran, Jan. 2006, pp. 369-376.

[25] B. Javadi, J. H. Abawajy, and M. K. Akabri, "Analytical Interconnection Networks Model for Multi-Cluster Computing Systems", In *13th International Conference on Analytical and Stochastic Modeling Techniques and Applications*, Bonn, Germany, to be appeared.