This is the published version (version of record) of:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30009766

Copyright : ©2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Analytical Network Modeling of Heterogeneous Large-Scale Cluster Systems

Bahman Javadi\textsuperscript{1}, Jemal H. Abawajy\textsuperscript{2}, Mohammad K. Akbari\textsuperscript{1}, Saeid Nahavandi\textsuperscript{2}

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

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

Abstract

The study of the communication networks for distributed systems is very important, since the overall performance of these systems is often depends on the effectiveness of its communication network. In this paper, we address the problem of networks modeling for heterogeneous large-scale cluster systems. We consider the large-scale cluster systems as a typical cluster of clusters system. Since the heterogeneity is becoming common in such systems, we take into account network as well as cluster size heterogeneity to propose the model. To this end, we present an analytical network 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.

Keywords: Network Modeling, Analytical Model, Cluster of clusters, Heterogeneous Networks, Latency.

1. Introduction

The overall performance of a distributed system is often critically hinged on the effectiveness of the communication networks. This renders the study of interconnection networks very important. In this paper, we address the problem of communication networks analytical modeling for heterogeneous cluster of clusters computing systems. We are motivated to study this problem for a number of reasons. First, the interconnection network plays a central role in the performance of cluster of clusters computing systems. Second, due to the networks contention problems \cite{16}, having a fast communication network does not necessarily guarantee to obtain a good performance. The contention problems which adversely affect the overall performance would happen in host nodes, network links, and network switches \cite{16}. 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. Moreover, these problems may be more complicated in the presence of network heterogeneity.

Simulation has been used to investigate the performance of various components of cluster of clusters computing systems \cite{1, 2}, but we are interested in analytical modeling of communication networks. In the mean time, the heterogeneity is becoming common in such systems and heterogeneous cluster systems are using by more and more researchers \cite{3, 4}. Of this, the proposed analytical model takes into account network 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 working conditions.

Several analytical performance models of multi-computer systems have been proposed in the literature for different interconnection networks and routing algorithms (e.g., \cite{7-10}). However, interconnection network studying for the system of interest is generally rare. Most of the existing researches are based on homogenous cluster systems and the evaluations are confined to a single cluster \cite{11-14}, with the exception of \cite{15}, which looked at processor heterogeneity. A general model based on queuing networks was proposed for a single cluster computing system in \cite{11}. The model assumes that the system is homogenous. Also, extensive numerical calculation of the model renders it too complicated. Furthermore, the model cannot be used for cluster of cluster computing systems in the presence of network and cluster size heterogeneity. The authors recently proposed an analytical model for multi-cluster systems in the presence of processor heterogeneity in \cite{25}.

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

2. Large-Scale Cluster Computing Systems

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 cluster of clusters systems. Such 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 [1,2,5,6]. Examples of production-level cluster of clusters systems include the DAS-2 [5] and the LLNL cluster of clusters system [6].

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

Fig. 1. The Heterogeneous Cluster of clusters System

Each cluster has two communication networks: an Intra-Communication Network (ICN1) and an inter-Communication Network (ECN1). 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. All of these networks may have different characteristics, i.e., bandwidth and latency. 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 [22], 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 [6, 23, 24]. In this paper we adopted $m$-port $n$-tree [17] as a fixed arity switches to construct the topology for each cluster in the system. An $m$-port $n$-tree topology consists of $N = 2(m/2)^n$ processing nodes and $N_{sw} = (2n-1)(m/2)^{n-1}$ network switches. In addition, each network switch itself has $m$ communication ports $\{0,1,2,\ldots,m-1\}$ that are attached to other switches or processing nodes. Every switch except root switches uses ports in the range of $\{0,1,2,\ldots,(m/2)-1\}$ to have connection with its descendants or processing node, and using ports in the range of $\{(m/2),(m/2)+1,\ldots,m-1\}$ for connection with its ancestors.

Flow control and routing algorithms are other 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 [24]. Routing algorithms establish the path between the source and the destination of a message. Since the most of cluster network technologies adopted deterministic routing [18], we used a deterministic routing based on Up*/Down* routing [19] which is proposed in [20]. In this algorithm, each message experiences two phases, an ascending phase to get to a Nearest Common Ancestor (NCA), followed by a descending phase.

3. The Analytical Network Model

In this section, we develop an analytic network 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 similar studies [7-11, 25]:

1. Nodes generate traffic independently of each other, and which follows a Poisson process with a mean rate of $\lambda$ 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 \( (N_i = 2^{(m/2)^h}) \).

4. The processing power of cluster’s nodes is homogenous with the same computational power.

5. The intra-cluster networks and inter-cluster networks are heterogeneous with different characteristics.

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

7. Message length is fixed \( (M \) flits).

Since the \( m \)-port \( n \)-tree is not a node-symmetric topology, so it does not suffice to analyze the traffic situation at a single node. The mean waiting time at the source queue, the mean network latency, and the mean time for the tail flit to reach the destination. Hence,

\[
\bar{T}_{in}^{(i)} = \bar{W}_{in}^{(i)} + \bar{T}_{in}^{(i)} + \bar{E}_{in}^{(i)}
\]

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 \( 2h \)-link message as \( T_h^{(i)} \), and averaging over all the possible nodes destined made by a message yields the mean network latency as:

\[
\bar{T}_{in}^{(i)} = \sum_{h=1}^{n_i} (P_{h,n_i} T_h^{(i)})
\]

where \( P_{h,n_i} \) is the probability of a message which is originated from cluster \( i \) crossing \( 2h \)-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:

\[
P_{h,n_i} = \begin{cases} 
\left(\frac{m}{2}\right)^h \left(\frac{m}{2}\right)^{h-1} & h = 1,2,\ldots,n_i - 1 \\
\left(\frac{m}{2}\right)^n & h = n_i
\end{cases}
\]

As shown in the flow model, the processor requests will be directed to ICN1\(^{(i)}\) and ECN1\(^{(i)}\) by probabilities \( 1 - U^{(i)} \) and \( U^{(i)} \), respectively. Therefore, the message rate received in the ICN1\(^{(i)}\) can be obtained as follows:

\[
\lambda_{ICN1}^{(i)} = N_i (1 - U^{(i)}) \lambda_g
\]

3.1. Intra-Cluster Message Latency

The mean latency seen by the intra-cluster message, \( \bar{T}_{in}^{(i)} \), crossing from source node from cluster \( i \) to destination, consists of three parts; the mean waiting time at the source queue, the mean network latency, and the mean time for the tail flit to reach the destination. Hence,

\[
\bar{W}_{in}^{(i)} + \bar{T}_{in}^{(i)} + \bar{E}_{in}^{(i)}
\]

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

\[
Latency = \frac{\sum_{i=0}^{C-1} \left(\frac{N_i T_i^{(i)}}{N}\right)}{N}
\]
\[ D^{(i)} = \sum_{h=1}^{n_i} (2hP_{h,n_i}) \]  

(8)

By substituting of Eq.(6) in to Eq.(8), the average message distance is obtained as,

\[ D^{(i)} = \frac{(m_{i} - 2n_{i} - 1)\left(\frac{m}{2}\right)^{n_{i}} + 1}{\left(\frac{m}{2}\right)^{n_{i}} - \frac{1}{2}\left(\frac{m}{2} - 1\right)} \]  

(9)

Consequently, we could derive the rate of received messages in each channel in the ICN1, which can be written as:

\[ \eta_{i}^{(1)} = \frac{\lambda_{11}^{(i)} D_{11}^{(i)}}{4n_{i}N_{i}} \]  

(10)

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. In other words, 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 \( K = 2h - 1 \). It should be noted that, 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:

\[ t_{cs}^{(i)} = 0.5\alpha_{n}^{(i)} + d_{m}\beta_{n}^{(i)} \]  

(11)

\[ t_{es}^{(i)} = \alpha_{s}^{(i)} + d_{m}\beta_{n}^{(i)} \]  

(12)

Where \( t_{cs}^{(i)} \) and \( t_{es}^{(i)} \) represent times to transmit from node to switch (or switch to node) and switch to switch connection in the cluster \( i \), respectively. \( \alpha_{n}^{(i)} \) and \( \alpha_{s}^{(i)} \) are the network and switch latency, \( \beta_{n}^{(i)} \) is the transmission time of one byte (inverse of bandwidth) in the cluster \( i \) and \( d_{m} \) is the length of each flit in bytes. In the presence of network heterogeneity, for the intra-cluster networks the set of \( \{t_{cs_{1}}^{(i)}, t_{es_{1}}^{(i)}\} \) and for the inter-cluster networks the set of \( \{t_{cs_{1}}^{(i)}, t_{es_{1}}^{(i)}, t_{es_{2}}^{(i)}\} \) are adopted in the model.

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 \( t_{cs1}^{(i)} \). 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 \( k \), \( 0 \leq k \leq K - 1 \), for cluster \( i \), \( W_{k,h}^{(i)} \), is driven as follows [25]:

\[ W_{k,h}^{(i)} = \frac{1}{2} \eta_{i}^{(1)} (T_{k,h}^{(i)})^{2} \]  

(13)

Where \( T_{k,h}^{(i)} \) is the mean service time of a channel at stage \( k \) and it is equal to the message transfer time and waiting time at subsequent stages to acquire a channel, therefore we can write:

\[ T_{k,h}^{(i)} = \begin{cases} 
\sum_{s=k+1}^{K-1} (W_{s,h}^{(i)}) + Mt_{cst_{1}}^{(i)} & \text{otherwise} \\
Mt_{cst_{1}}^{(i)} & k = K - 1 
\end{cases} \]

(14)

According to this equation, the network latency for a message with \( 2h \)-link journey equals to the mean service time of a channel at stage 0 (i.e., \( T_{0,h}^{(i)} = T_{h}^{(i)} \)).

An intra-cluster message originating from a given source node in cluster \( i \) sees a network latency of \( T_{in}^{(i)} \) (given by Eq.(5)). 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 [21]:

\[ \bar{W}_{in}^{(i)} = \frac{\lambda^{(i)}(\bar{x}^{(i)} + \sigma_{x}^{(i)})}{2(1 - \rho^{(i)})} \]  

(15)

\[ \rho^{(i)} = \lambda^{(i)} \bar{x}^{(i)} \]  

(16)

Where \( \lambda^{(i)} \) is the mean arrival rate on the network, \( \bar{x}^{(i)} \) is the mean service time, and \( \sigma_{x}^{(i)} \) is the variance of the service time distribution. Since the minimum service time of a message at the first stage is equal to \( Mt_{cst_{1}}^{(i)} \), the variance of the service time distribution is approximated based on a method proposed in [9] as follows:

\[ \sigma_{x}^{(i)} = \left(\bar{W}_{in}^{(i)} - Mt_{cst_{1}}^{(i)}\right)^{2} \]  

(17)

As a result, the mean waiting time in the source queue becomes,
\[ W_m^{(i)} = \frac{\lambda_1^{(i)} \left( \left( T_m^{(i)} \right)^2 + \left( T_m^{(i)} - M_{t,m}^{(i)} \right)^2 \right)}{2 \left( 1 - \lambda_1^{(i)} T_m^{(i)} \right)} \] (18)

At last, the mean time for the tail flit to reach the destination can be written by the following equation:
\[ T_{m}^{(i)} = \sum_{h=1}^{n} \left[ P_{h,n_i} \left( 2(h-1)T_{m}^{(i)} + t_{cn,m}^{(i)} \right) \right] \] (19)

### 3.2. Inter-Cluster Message Latency

In this part, we determine the same entity in the inter-cluster networks. A typical inter-cluster message of cluster \( i \) leaves the ECN1(\( i \)) and crosses through the ICN2 and then goes to the ECN1(\( j \)) of the cluster \( j \) to reach its destination node. Since the flow control mechanism is wormhole, the latency of these networks should be calculated as a merge unit. So, based on the Eq. (5), we can write,
\[ T_{ex}^{(i,j)} = \sum_{r=1}^{n} \sum_{l=1}^{n} \left( P_{(r,v),+l,n} \times T_{(r,v) (+l,j)} \right) \] (20)

It means each message crosses \( r(v)-l \)-link through the ECN1 networks \( r(v)-l \) in the source cluster \( i \) and \( v \)-link in the destination cluster \( j \) and \( 2l \)-link in the ICN2 to reach its destination. The probability of such journey from cluster \( i \) point of view, \( P_{(r,v)+l,n} \) would be,
\[ P_{(r,v)+l,n} = P_{r,v,n_i} P_{v,n_j} P_{n_i} \] (21)

In what follow, we determine the mean network latency in the inter-cluster networks. 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. So, the message rate received in each networks can be obtained as follows:
\[ \lambda_{E1}^{(i,j)} = \left( N_i U_i^{(i)} + N_j U_j^{(j)} \right) \lambda_{l} \] (22)
\[ \lambda_{E2}^{(i,j)} = \left( N_i \left( N_i U_i^{(i)} \right) + N_j \left( N_j U_j^{(j)} \right) \right) \lambda_{l} \] (23)

Consequently, the mean message rate received by each channel in these networks can be written as:
\[ \eta_{E1}^{(i,j)} = \frac{\lambda_{E1}^{(i,j)} T_{E1}^{(i)}}{4n_i N_i} \] (24)
\[ \eta_{E2}^{(i,j)} = \frac{\lambda_{E2}^{(i,j)} T_{E2}^{(i)}}{4n_c} \] (25)

Where \( n_c \) would be computed such that \( C = 2(m/2)^N \). In the inter-cluster networks, the number of stages for each message journey is \( K = r + 2l + v - 1 \). Based on Eq. (13), the mean amount of time that a message waits to acquire a channel at stage \( k \), in the inter-cluster networks is as follows:
\[ W_{k,(r,v)+l,i,j} = \frac{1}{2} \eta_c^{(i,j)} \left( T_{k,(r,v)+l,i,j} \right)^2 \] (26)

Where the channel rate is driven by the following equation:
\[ \eta_c^{(i,j)} = \begin{cases} \eta_{E1}^{(i,j)} & \text{if } r \leq k < r + 2l - 1 \\ \eta_{E2}^{(i,j)} & \text{otherwise} \end{cases} \] (27)

Where \( \delta^{(i)} \) is the relaxing factor and could be define as follows:
\[ \delta^{(i)} = \frac{\beta_{n_c}^{(i)}}{\beta_{n_{E2}}} \] (28)

This is because of two networks (i.e., ECN1 and ICN2) have different bandwidth, so when the message flow comes into the ICN2 (with usually more bandwidth) the waiting time will be decreased proportional to the capacity of the ICN2 networks.

The mean service time of a channel in the inter-cluster networks from cluster \( i \) point of view can be found as follows:
\[ T_{k,(r,v)+l,i,j} = \begin{cases} M_{t,m}^{(i)} & \text{if } k = K - 1 \\ \sum_{s=k+1}^{K-1} \left( W_{s,(r,v)+l,i,j} \right) + M_{t,m}^{(i)} & \text{otherwise} \end{cases} \] (29)

Where \( t_m \) can be written based on the time to transmit of each flit in the correspondence network as:
\[ t_m = \begin{cases} \frac{t_m^{(i)}}{4n_i N_i} & \text{if } 0 \leq k < r \\ \frac{t_m^{(i)}}{4n_c} & \text{if } r \leq k < r + 2l - 1 \\ t_m^{(j)} & \text{if } r + 2l - 1 \leq k < K - 1 \end{cases} \] (30)

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.

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:
The message latency of inter-cluster networks from cluster \( i \) to cluster \( j \) can be found as follows:

\[
\bar{T}_{ex}^{(ij)} = \bar{W}_{ex}^{(ij)} + \bar{T}_{w}^{(ij)} + \bar{E}_{ex}^{(ij)}
\]  

(32)

Where \( \bar{E}_{ex}^{(ij)} \), the mean time for the tail flit to reach the destination, is given by the following equation:

\[
\bar{E}_{ex}^{(ij)} = \sum_{r=1}^{n_i} \sum_{e=1}^{n_j} \sum_{l=1}^{m} P(r,e)+t_{ex} E_{ex}^{(i,j)}
\]  

(33)

The \( E_{(r,e)+t_{ex}^{(i,j)}}^{(ij)} \) can be obtained as follows:

\[
E_{(r,e)+t_{ex}^{(i,j)}}^{(ij)} = (r-1)t_{ex1}^{(i)} + (e-1)t_{ex2}^{(j)} + (2)t_{ex2}^{(i)} + t_{ex1}^{(j)}
\]  

(34)

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

\[
\bar{T}_{ex}^{(i)} = \frac{1}{C-1} \sum_{j=0,j \neq i}^{C-1} \bar{T}_{ex}^{(ij)}
\]  

(35)

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.(15)). The service time of the queue would be \( M_{ex1} \) and although the messages length is fixed but there is a variance in service time because of different network characteristic. The variance of the service time distribution is approximated same as source queue as follows:

\[
\sigma_{2}^{(i)} = \left( M_{ex2} - M_{ex1}^{(i)} \right)^{2}
\]  

(36)

By modeling the concentrate buffers in the concentrator/dispatcher as an M/G/1 queue, the mean waiting time is given by the following equation:

\[
\bar{W}_{e}^{(i,j)} = \frac{\lambda_{e}^{(i,j)} \left( M_{ex1}^{2} + \left( M_{ex2} - M_{ex1}^{(i)} \right)^{2} \right)}{2 \left( 1 - \lambda_{e}^{(i,j)} M_{ex1} \right)}
\]  

(37)

The arithmetic average of sum of the two above mentioned waiting times gives mean waiting time at the concentrator/dispatcher as follows:

\[
\bar{W}_{d}^{(i,j)} = \frac{1}{C-1} \sum_{j=0,j \neq i}^{C-1} \left( 2\bar{W}_{e}^{(i,j)} \right)
\]  

(38)

At last, the mean message latency in the inter-cluster networks from cluster \( i \) point of view can be found as:

\[
\bar{T}_{lost}^{(i)} = \bar{T}_{ex}^{(i)} + \bar{W}_{d}^{(i)}
\]  

(39)

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 \( \lambda_{g} \). The destination node is determined by using a uniform random number generator. 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. 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 characteristics, 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. Also, the network characteristics which are used in the validation are shown in Table 2. The ICN1 and ICN2 networks used the Net.1 while the ENC1 networks used the Net.2 configuration. Moreover, the two different message lengths, \( M = 32 \) and 64 flits with different sizes, \( D_{in} = 256 \) and 512 bytes are used.

**Table 1. System Organizations for Model Validation**

<table>
<thead>
<tr>
<th>( N )</th>
<th>( C )</th>
<th>( m )</th>
<th>Node Organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1120</td>
<td>32</td>
<td>8</td>
<td>( n_{i}=1 ) ie [0,11]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( n_{j}=2 ) ie [12,27]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( n_{k}=3 ) ie [28,31]</td>
</tr>
<tr>
<td>544</td>
<td>16</td>
<td>4</td>
<td>( n_{i}=3 ) ie [0,7]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( n_{j}=4 ) ie [8,10]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( n_{k}=5 ) ie [11,15]</td>
</tr>
</tbody>
</table>
Table 2. Network Characteristics for Model Validation

<table>
<thead>
<tr>
<th>Network</th>
<th>Bandwidth</th>
<th>Network Latency</th>
<th>Switch Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net.1</td>
<td>500</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Net.2</td>
<td>250</td>
<td>0.05</td>
<td>0.01</td>
</tr>
</tbody>
</table>

The results of simulation and analysis are shown in Fig. 3 to Fig. 6 in which the mean message latencies are plotted against the traffic generation rate 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 terms 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.(17)) is a factor of the model inaccuracy. However, at light traffic the model differs from simulation by about 4 to 8 percent. 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.

To show the capabilities of the proposed model, we consider a typical analysis in the system under study. The results of simulation and analysis reveal that the inter-cluster networks, especially ICN2, are the bottlenecks of the system. To have a better analysis, we increase the bandwidth of ICN2 in amount of 20 percent in the two above-mentioned systems. The results of analysis for $M=128$ and $D_m=256$ are depicted in Fig. 7. The figure shows that this enhancement has a great impact on the system performance especially in the high traffic region. Also, the performance of the system with $N=544$ has better improvements than the system with $N=1120$, since it is strongly depended on the organization of the systems.

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 cluster of clusters 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 whereas at light traffic the model differs from simulation by less than about 4 to 8 percent. For future work, we intent to take the non-uniform traffic pattern into account, which is closer to the real traffic in such systems.

Acknowledgments

This work is funded in part by Iran Telecommunication Research Center (ITRC).
Fig. 4. Mean message latency in a system with $N=1120$, $M=64$

Fig. 5. Mean message latency in a system with $N=544$, $M=32$

Fig. 6. Mean message latency in a system with $N=544$, $M=64$

Fig. 7. Mean message latency in two systems with different bandwidth in the ICN2 network

References


