ANALYTICAL INTERCONNECTION NETWORKS MODEL FOR MULTI-CLUSTER COMPUTING SYSTEMS

Bahman Javadi  
Computer Eng. and IT Department  
Amirkabir University of Technology  
424 Hafez Ave., Tehran, Iran  
E-mail: javadi@ce.aut.ac.ir

Jemal H. Abawajy  
School of Information Technology  
Deakin University  
Geelong, VIC. 3217, Australia  
E-mail: jemal@deakin.edu.au

Mohammad K. Akbari  
Computer Eng. and IT Department  
Amirkabir University of Technology  
424 Hafez Ave., Tehran, Iran  
E-mail: akbari@ce.aut.ac.ir

KEYWORDS  

ABSTRACT  
This paper addresses the problem of interconnection networks performance modeling of large-scale distributed systems with emphases on 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. We present an analytical model that considers stochastic quantities as well as processor heterogeneity of the target system. The model is validated through comprehensive simulation, which demonstrates that the proposed model exhibits a good degree of accuracy for various system sizes and under different operating 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 and grid computing 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 (Abawajy and Dandamudi 2003).

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 performances of various components of multi-cluster computing systems (Abawajy and Dandamudi 2003). Instead, we focus on analytical model. An accurate analytical model can provide quick performance estimates and will be a valuable design tool. The significant advantage of analytical models over simulation is that they can be used to obtain performance results for large systems which may not be feasible to study using simulation due to the excessive computation demands.

Several analytical performance models of multi-computer systems have been proposed in the literature for different interconnection networks and routing algorithms (e.g., Sarbazi-Azad et al. 2002; Boura and Das 1997; Drapper and Ghosh 1994). Unfortunately, little attention has been given to cluster computing systems. Most of the existing researches are based on homogeneous cluster systems and the evaluations are confined to a single cluster (Du et al. 2000; Hu and Kleinrock 1995). In contrast, we focus on heterogeneous multi-cluster computing environment.

To this end, we present an analytical performance model of interconnection networks for multi-cluster computing systems. The model is based on probabilistic analysis and queuing network to analytically evaluate the performance of interconnection networks for multi-cluster systems. The model takes into account stochastic quantities as well as processor 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.

The rest of the paper is organized as follows. In Section 2, we give a brief overview of the multi-cluster systems. 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. SYSTEM OVERVIEW

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 Ni processors of type τ, i ∈ {0, 1, ..., C − 1}. Also, each cluster has two communication networks, an Intra-Communication Network (ICN1), which is used for the purpose of message passing between processors, and an intEr-
Communication Network (ECN1), which is used to transmit messages between clusters, management of the system, and also for the scalability of the system.

It should be noted that, ECN1 can be accessed directly by the processors of each cluster without going through the ICN1 (see Fig. 1). To interconnect ECN1 and ICN2, a set of Concentrators/Dispatchers (Dally and Towles 2004) are used, which combine message traffic from/to one cluster to/from other cluster.

Cluster #i
Cluster #(C·1)

Fig. 1. The Multi-Cluster Computing System

The main factors which have impact on the performance of a multi-cluster system are Topology, Flow control mechanism, and Routing algorithm. The connection pattern of nodes defines the network’s topology. Most current cluster systems (Boas 2003) employ fat-tree for scalability and high bandwidth inter-processor communication. In this paper we adopted m-port n-tree (Lin 2003) 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 Nsw communication switches which can be calculated with Eqs.(1) and (2), respectively.

\[ N = 2 \times \left( \frac{m}{2} \right)^n \]  \hspace{1cm} (1)
\[ N_{sw} = (2n-1) \times \left( \frac{m}{2} \right)^{n-1} \]  \hspace{1cm} (2)

In addition, each communication switch itself has m communication ports{0,1,2,...,m -1} that are attached to other switches or processing nodes.

In regards of flow control mechanism, since the dedicated cluster network technologies, e.g., Myrinet, Infiniband and QsNet are using wormhole flow control, so we adopt this mechanism to outline the analytical model. Also, these commercial networks adopt deterministic routing algorithms. Of this, we used a deterministic routing based on Up*/Down* routing (Schroeder 1990) which is proposed in (Javadi et al. 2006). In this algorithm, each message experiences two phases, an ascending phase to get to a Nearest Common Ancestor (NCA), followed by a descending phase.

Unlike most works on heterogeneous parallel systems, we express the speeds of various nodes in each cluster relatively to a fixed reference machine (Clematis and Corana 1999), and not relatively to the fastest node. Although the latter choice may appear more natural since it makes it possible to obtain the speed-up by comparing performance of the parallel system with that of the fastest single node available, we think that choosing a fixed reference allows clearer performance analysis, especially if we vary the number and/or the power of nodes. Since we consider the processor heterogeneity between each cluster, the total relative speed and the average relative speed of the C clusters in the system is as follows, respectively:

\[ S = \sum_{i=0}^{C-1} s_i \]  \hspace{1cm} (3)
\[ \bar{s} = \frac{S}{C} \]  \hspace{1cm} (4)

Where \( s_i \) is the relative speed of a processor in the cluster \( i \).

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 the similar studies (Javadi et al. 2005, Sarbazi-Azad et al. 2002; Boubas and Das 1997; Hu and Kleinrock 1995):

1. Each processor in cluster \( i \) generates packets independently, which follows a Poisson process with a mean rate of \( \lambda_i \).

2. The arrival process at a given communication network is approximated by an independent Poisson process. The destination of each request would be any node in the system with uniform distribution.

3. The number of processors in all clusters are equal \( (N_0 = N_1 = \ldots = N_{C-1}) \) and the clusters' nodes are heterogeneous in their speed \( (\tau_i = s_i) \).

4. The communication switches are input buffered and each channel is associated with a single flit buffer. Message length is fixed \( (M \text{ flits}) \).

3.1. Traffic Analysis

The traffic pattern affects mainly the average message distance, \( d_{\text{avg}} \), which is expected number of links that a message makes to reach its destination. The average message distance is generally given by

\[ d_{\text{avg}} = \sum_{j=1}^{N_{sw}} 2j \times P_j \]  \hspace{1cm} (5)

Where \( P_j \) is the probability of a message crossing 2j-link (j-link in ascending and j-links in descending phase) to reach its destination in a m-port n-tree topology. As it is mentioned in assumption 3, we take into account the uniform traffic pattern so, based on the
m-port n-tree topology, we can define this probability as follows:

\[ P_j = \begin{cases} \frac{(m-1)^j}{N_0-1} & j = 1, 2, \ldots, n-1 \\ \frac{(m-1)^{n-j}}{N_0-1} & j = n \end{cases} \]  

(6)

With substituting of Eq.(6) in Eq.(5), the average message distance is obtained as,

\[ d_{av} = \frac{(nm - 2n - 1)\left(\frac{m}{2}\right)^n + 1}{\left(\frac{m}{2}\right)^n \cdot \frac{1}{2} \left(\frac{m}{2}\right)^{-1}} \]  

(7)

### 3.2. Arrival 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. The processor requests will be directed to the ICNI and the ECNI by probabilities \(1 - P_o\) and \(P_o\) respectively.

Fig. 2. Message Flow Model in each Communication Network

The external request (out of cluster) of cluster \(i\) goes through the ECNI with probability \(P_o\) and then ICN2. In the return path, it again accesses the ECNI in cluster \(v\) to get to the destination node. So, the message rate received by ICNI and ECNI in cluster \(i\) (to cluster \(v\)) can be calculated as follows:

\[ \lambda_{71}^{(i)} = (1 - P_o) \lambda_{x}^{(i)} \]  

(8)

\[ \lambda_{71}^{(v)} = P_o \lambda_{x}^{(i)} + P_o \lambda_{x}^{(v)} \quad \forall \neq i \]  

(9)

In the second stage, the input message rate of ICN2 can be computed by following equation:

\[ \lambda_{72}^{(i)} = N_o P_o \lambda_{x}^{(i)} \]  

(10)

We now derive the rate of received messages in each channel, which can be written as:

\[ \eta_{71}^{(i)} = \frac{(1 - P_o) \lambda_{x}^{(i)} \times d_{av(I)}(1)}{4n} \]  

(11)

\[ \eta_{72}^{(i)} = \frac{N_o P_o \lambda_{x}^{(i)} \times d_{av(2)}}{4n} \]  

(12)

Where \(n_o\) is the number of tree in ICN2 and can be computed by Eq.(1). In the Eq.(8) to Eq.(13), the probability \(P_o\) has been used as the probability of outgoing request within a cluster. According to assumption 3, this parameter is obtained by the following equation:

\[ P_o = \frac{\sum_{i=1}^{C-I} \lambda_{71}^{(i)}}{N_o} = \frac{(C-1) \times N_o}{C \times N_0 - 1} \]  

(14)

### 3.3. Average Network Latency

In this section, we find the average latency of each communication network from cluster \(i\) point of view, \(T_j^{(i)}\). Since each message may cross different number of hops to reach its destination, we consider the network latency of an \(2j\)-hop message as \(T_j^{(i)}\), and averaging over all the possible nodes destined made by a message yields the average message latency as:

\[ T^{(i)} = \sum_{j=1}^{\infty} (P_j \times T_j^{(i)}) \]  

(15)

Where \(P_j\) can be calculated from Eq.(6). 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 \(2j\)-hop journey is \(K = 2j - 1\). 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. Each type of connection has a service time which is approximated as follows:

\[ t_{ce} = \frac{1}{2} \alpha_{sw} + L_{sw} \beta_{sw} \]  

(16)

\[ t_{cz} = \alpha_{sw} + L_{sw} \beta_{sw} \]  

(17)

Where \(t_{ce}\) and \(t_{cz}\) represent times to transmit from node to switch (or switch to node) and switch to switch connection, respectively. \(\alpha_{sw}\) and \(\alpha_{sw}\) are the network and switch latency, \(\beta_{sw}\) is the transmission time of one byte (inverse of bandwidth) and \(L_{sw}\) is the length of each flit in bytes.

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 \(t_{sx}\). At stage \(k\), the average amount of time that a message waits to acquire a channel for cluster \(i\), \(W_{k,i}^{(i)}\), is given by the product of ...
the channel blocking probability, $P_{b,i}$, and the average service time, $S_{k,j}(i)/2$ (Dally and Towles 2004):

$$W_{k,j}(i) = \frac{1}{2}S_{k,j}(i)P_{b,i}$$  (18)

The value of $P_{b,i}$ is determined using a birth-death Markov chain (Kleinrock 1975). As it can be seen in Fig. 3, the rate of transition out and into the first state is $\eta(i)$ and $\frac{1}{S_{k,j}(i)} - \eta(i)$ respectively.

![Fig. 3. Markov Chain to Calculate Blocking Probabilities](image)

Solving this chain for the steady state probabilities gives:

$$P_{b,i} = \eta(i)S_{k,j}(i)$$  (19)

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

$$S_{k,j}(i) = \left\{ \begin{array}{ll}
M_{m} & k = K - 1 \\
\sum_{j=k+1}^{K-1} (W_{j,j}(i)) + M_{m} & \text{otherwise}
\end{array} \right.$$

(20)

According to this equation, the average service time of a message, not being in the first stage, is equal to average network latency. Hence, $T_{j}(i) = S_{k,j}(i)$.

A message originating from a given source node in cluster $i$ sees a network latency of $T^{(i)}$. 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 average waiting time at an $M/G/1$ queue is given by (Kleinrock 1975):

$$\bar{W}^{(i)} = \rho^{(i)}x^{(i)}\left(1 + \frac{\sigma^{2}(i)}{x}\right) \frac{1}{2(1 - \rho^{(i)})}$$

(21)

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

(22)

Where $\lambda^{(i)}$ is the average arrival rate on the network, $x^{(i)}$ is the average service time, and $\sigma^{2}(i)$ is the variance of the service time distribution. Since the minimum service time of a message at the first stage is equal to $M_{m}$, the variance of the service time distribution is approximated based on a method proposed in (Draper and Ghosh 1994),

$$\sigma_{x}^{2}(i) = \left(\bar{T}^{(i)} - M_{m}\right)^{2}$$

(23)

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

$$\bar{W}^{(i)} = \frac{\lambda^{(i)}\left(\bar{T}^{(i)}\right)^{2}}{2\left(1 - \lambda^{(i)}\bar{T}^{(i)}\right)^{2}}$$

(24)

Finally, the average message latency, $\bar{L}^{(i)}$, seen by the message crossing from source node from cluster $i$ to its destination, consists of three parts; the average waiting time at the source queue ($\bar{W}^{(i)}$), the average network latency ($\bar{T}^{(i)}$), and the average time for the tail to reach the destination. Therefore,

$$\bar{L}^{(i)} = \bar{W}^{(i)} + \bar{T}^{(i)} + \sum_{k=1}^{2(j+h) - 1} t_{es} + t_{en}$$

(25)

The average message latency in the ICN1 from cluster $i$ point of view would be found by Eq.(25) with substitution of $\eta(i) = \eta_{1}(i)$, $\lambda(i) = \lambda_{1}(i)$, and $d_{es} = d_{es}(i)$.

### 3.3.1. Average Latency in the Inter-Cluster Networks

As mentioned before, external messages cross through both networks, ECNI 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. Of this and based on the Eq.(15) we can write,

$$\bar{T}^{(i)} = \sum_{i=1}^{3j} \left(\sum_{j=1}^{K} \left(\sum_{k=1}^{2(i) - 1} \left(P_{j,k} \times T_{j,k}\right)\right)\right)$$

(26)

Where the probability $P_{j,k}$ is,

$$P_{j,k} = P_{j} \times P_{k}$$

(27)

Where $P_{j}$ and $P_{k}$ can be calculated from Eq.(6). The average network latency of inter-cluster networks can be founded with the equations which are presented in the previous section by following substitutions:

$$K = 2(j + h) - 1$$

(28)

$$\eta_{i}(j) = \left\{ \begin{array}{ll}
\eta_{1}(j) & j \leq k < j + 2h - 1 \\
\eta_{2}(j) & \text{otherwise}
\end{array} \right.$$  (29)

$$\lambda^{(i)} = \lambda_{1}^{(i)}$$

(30)

$$d_{es} = d_{es}(1) + d_{es}(2)$$

(31)
3.3.2. Average Waiting Time at the Concentrator/Dispatcher

The average waiting time at the concentrator/dispatcher is calculated in a similar manner to that for the source queue (Eq.(21)). By modeling the injection channel in the concentrator/dispatcher as an M/G/1 queue, the average arrival rate and average waiting time are given by following equations:

\[
W_{\text{con}}^{(i)} = \frac{\lambda_{12}^{(i)} \overline{T}_{12}^{(i)}}{2(1 - \lambda_{12}^{(i)} \overline{T}_{12}^{(i)})}
\]

where \(\overline{T}_{12}^{(i)}\) is the average network latency of the ICN2 from cluster \(i\) point of view. 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 service time of the queue would be \(M_{tcs}\) and there is no variance in the service time, since the messages length is fixed. Hence,

\[
W_{\text{dis}}^{(i)} = \frac{\lambda_{21}^{(i)} (M_{tcs})^2}{2(1 - \lambda_{21}^{(i)} M_{tcs})}
\]

The sum of the two above mentioned waiting times gives average waiting time at the concentrators/dispatchers as follows:

\[
W^{(i)} = W_{\text{con}}^{(i)} + W_{\text{dis}}^{(i)}
\]

Putting all together, we could find the average message latency of cluster \(i\) based on Fig. 2 with the following equation:

\[
\overline{\tau}^{(i)} = (1 - P_a) \overline{L}_{f1}^{(i)} + P_a \overline{L}_{f2}^{(i)} + \overline{W}_{e}^{(i)}
\]

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

\[
\overline{\tau} = \frac{1}{C} \sum_{i=1}^{C} \frac{S^{(i)}}{S} \times \overline{\tau}^{(i)}
\]

At last, to perform our analysis we chose to express the degree of heterogeneity of the system through a single parameter, i.e., the standard deviation of relative speeds as follows:

\[
H = \sqrt{\frac{1}{C} \sum_{i=1}^{C} (S^{(i)} - \overline{S})^2}
\]

4. VALIDATION OF THE MODEL

In order to validate the proposed model and justify the applied approximations, the model was simulated. Requests are generated randomly by each processor with an exponential distribution of inter-arrival time with a rate of \(\lambda_{a}^{(i)}\). 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 inhibited for the first 10,000 messages to avoid distortions due to the warm-up phase.

Extensive validation experiments have been performed for several combinations of clusters sizes, network sizes, message length, and degree of heterogeneity. 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 as follows:

- System size: \(N=2^9\) and \(N=2^{10}\)
- Cluster size: \(C=2^4\) and \(C=2^5\)
- Total relative speed: \(S=C\)
- Switch size: \(m=4\) and \(m=8\) ports
- Message length: \(M=64\) flits
- Flit length: \(L_{m}=256\) and \(512\) bytes
- Network technology bandwidth: \(500/\text{time unit}\)
- Network latency: \(0.02\) time unit
- Switch latency: \(0.01\) time unit

The results of simulation and analysis for the systems with above mentioned parameters are depicted in Fig. 4 to Fig. 7 in which the average message latencies are plotted against the offered traffic with different values for degree of heterogeneity.

The figures reveal that the analytical model predicts the average 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. 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.

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 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.
REFERENCES


