COMMUNICATION COST IN BUS STRUCTURED DISTRIBUTED SYSTEMS

Brian E. Waldecker and Mario J. Gonzalez, Jr.
Computer Engineering Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas 78712

Abstract

As distributed systems become more common, a means to quantitatively compare candidate designs is needed. Gonzalez and Jordan have proposed a comparison framework which requires as input, a set of defined and quantified system attributes. In this paper, four bus connection schemes have been explored with respect to the attribute of system communication cost. Single and multiple buses were examined for two methods of bus protocol. One method involves holding the bus throughout the time needed to access and transfer the data while the other involves using tagged requests and replies. A means of quantifying communication cost for these distributed systems has been derived using elements of queueing theory. The expressions for communication cost were then compared to see how communication cost increases more rapidly for the bus-hold method than it does for the tagged request-reply method (as processor speed decreases). Lastly, the effect of varying (non-uniform) reference patterns on communication cost was analyzed for the request-reply bus system. This provides a quantitative measure of performance degradation as the distribution of references becomes less uniform.

1 Introduction

In the past few years, much attention has been focused on distributed systems in which a number of computers are connected. While various structures for distributed systems each have merits and advocates, the comparison of one distributed system with another is still largely a matter left to "experience" or intuition. Although experience will always play a part in the design process, it is nonetheless attractive to envision some framework which could be employed in the comparison of candidate systems. Such a framework has been presented in work done by M. J. Gonzalez and B. W. Jordan [1]. This paper examines communication cost for bus connected multiple processor systems and presents queuing models for use in quantifying this cost which can then be included in a comparison framework. Briefly, communication cost is the average amount of time that a processor spends referencing data in a non-local memory (e.g. another processor's memory or a common shared memory).

2 Background – The Framework of Gonzalez and Jordan

Let \( A = \{A_1, A_2, \ldots, A_n\} \) be a set of attributes which describe problem requirements and system characteristics. Also, let the distributed systems to be compared be represented by \( S = \{S_1, S_2, \ldots, S_N\} \). Let \( R = \{r_1, r_2, \ldots, r_n\} \) denote the problem requirements such that \( r_i \) represents the quantity of attribute \( A_i \) required for a given problem. Lastly, let \( Q \) be an \( N \times n \) matrix where element \( q_{ij} \) represents the extent to which system \( S_j \), \( 1 \leq j \leq N \), possesses attribute \( A_i, 1 \leq i \leq n \) [1]. In the framework put forth by Gonzalez and Jordan, weighted differences between the elements of \( R \) and \( Q \) are used to rank the candidate systems in order of preference. Fig. 1 illustrates the relationships between \( A, S, \) and the "weighted attribute differences" (denoted by lower case \( d's \)). The weighted attribute difference will be precisely defined shortly.

![Evaluation Framework](image)

To arrive at the weighted differences, the attribute set, \( A \), must be ordered according to importance. This is really just a permutation of the set \( A \) to \( A' = \{A'_1, A'_2, \ldots, A'_n\} \) such that \( A'_1 \) is the most important attribute and \( A'_n \) is least important. Each attribute must be assigned a weight from the set \( W = \{w_1, w_2, \ldots, w_n\} \). Here attribute \( A'_i \) is assigned weight \( w_i \) and this reflects the monotonic nonincreasing order of importance placed on the attributes in \( A' \). Gonzalez and Jordan discuss some ways the weights in set \( W \) may be chosen in [1]. To simplify notation, further reference to the attribute set \( A \) will actually be to the ordered set \( A' \). With this notation established, \( d_{ii} \) is defined for system \( S_i \) and attribute \( A_i \) as \( d_{ii} = w_i [(q_{ii} - r_i)/r_i] \).
Here, \( r_i \) is the problem requirement for attribute \( A_i \) and is placed in the denominator for normalization. Fig. 2 pictorially shows the weighted attribute differences for a single one of the candidate systems. The weight attached to an attribute corresponds to the bar's width along the x-axis.

For each system, the sum over all attributes \( A_i, 1 \leq i \leq n \), is taken of the \( d_i \)'s defined above. Thus, for each system, \( S_i \), a sum of the weighted attribute differences is calculated. This sum is given mathematically as \( D_i = \sum_{j=1}^{n} d_{ij} \).

This sum of weighted differences is computed for each member in the set, \( S \), of candidate systems, thereby producing the set \( D = \{ D_1, D_2, \ldots, D_n \} \) of sums. Finally, from this set, the most suitable candidate system can be selected. This system is given by \( S_k \in S \) where \( D_k = \max \{ D_1, D_2, \ldots, D_n \} \).

3 Systems to be Examined

Buses can be placed into two classes, buses with centralized control and buses with decentralized control. The bus structure to be considered here is the decentralized type. This is the class of interconnection structure which Anderson and Jensen have dubbed DSB, for “Direct Shared path global Bus” [2]. An illustration of the DSB structured system is shown in Fig. 3.

4 Bus Communication Cost

The amount of contention present in the system will vary as a function of how evenly the external references are distributed among the processors. Thus, the number and distribution of external references will be the basic variables in calculating system communication cost. Let \( a \) be a \( n \times n \) array where \( a_{ij} \) gives the number of references per unit time from \( PE_i \) to \( PE_j \). That is

\[
a = \begin{bmatrix}
d_{11} & \cdots & d_{1n} \\
\vdots & \ddots & \vdots \\
0_{n1} & \cdots & a_{nn}
\end{bmatrix}
\]  

It should be noted that if the array \( a \) has non-zero entries \( a_{ii} \) and zeros for \( a_{ij} \) entries \( i \neq j \), then the system represented has no external references and all \( PE \) memory requests are local. The matrix \( a \) can also represent a global memory system with no memory local to any \( PE \) if the row corresponding to the global memory contains all zeros and the column corresponding to the global memory is the only column that does not contain all zeros.

To model the shared bus system, locations in the system where contention exists must be identified. Contention will be present first for use of the bus and second, for access to the local memory of the remote \( PE \).

There are two ways of assuming how the bus behaves. First, it could be assumed that the bus is held throughout the access to the remote \( PE \). Alternatively, a system of tagged requests and replies could be assumed where the bus is released between a request's being sent and the reply's getting returned. With this second method, the bus is free more with the hope that better performance will result from allowing the bus to do useful work while the remote \( PE \) services the request for data. However, since every non-local memory access now involves the sending of two packets over the bus, the request rate for the bus will be double that of the case where the bus is held throughout the external reference.

4.1 Queueing Model for Bus-Hold System.

In this case, an external reference must contend for the bus, then transmit the memory addresses needed and wait until the information is returned, holding the bus throughout. Since the bus is held during the accessing time at the remote \( PE \) it seems reasonable to make the remote \( PE \) access time part of the total service time of the bus. Thus, a model is
defined with a source of external references that feeds a queue being served by the bus. The mean service time of the bus server must be calculated by including bus transfer time as well as the queuing times \(^1\) at remote PEs. The queuing times at each PE, weighted by the fraction of total bus traffic which addresses that PE, can be added to give a weighted sum that represents the average queuing time at any remote PE. Adding this weighted sum to the transfer times will give the mean bus service time. The mean bus service rate will be the reciprocal. Equation 2 expresses this mathematically.

\[
\mu_{B(hold)} \overset{\text{def}}{=} \frac{1}{2TB + \sum_{i=1}^{n} E_i Q_i(hold)}
\]

where, \(T_B\) = mean bus transfer time

\[
E_i = \frac{\text{number ext. refs. to } PE_i}{\text{total number of refs. on bus}}
\]

\[
Q_i(hold) = \frac{\text{average queuing time at } PE_i}{26}
\]

To determine the time until the data at the remote PE is ready to be sent back \((Q_i(hold))\), an additional single server queue must be employed. \(Q_i(hold)\) will equal the queuing time for PE\(_i\). The mean service rate for a PE\(_i\) is the number of requests per unit time which that PE can process. Similarly, the mean arrival rate at a PE is the number of requests per unit time that are actually arriving at that PE. Let these quantities be defined as follows:

\[
\mu_i \overset{\text{def}}{=} \text{mean service rate of } PE_i
\]

\[
\lambda_i(hold) \overset{\text{def}}{=} \text{mean arrival rate at } PE_i
\]

The arrivals into this queue are given by that PE\(_i\)’s local accesses plus all other external accesses to it from other PEs. Thus, \(\lambda_i(hold)\) is given by:

\[
\lambda_i(hold) = a_i + \sum_{j=1,j\neq i}^{n} a_{ji} = \sum_{j=1}^{n} a_{ji}
\]

Using Equations 3 and 4 for service and arrival rates with the result for queuing time of a M/M/1 queue, \(Q_i(hold)\) of Equation 2 is given by:

\[
Q_i(hold) = \frac{1}{\mu_i - \lambda_i(hold)}
\]

\[
= \frac{1}{\mu_i - \sum_{j=1}^{n} a_{ji}}
\]

\(Q_i(hold)\) is used in Equation 2 to find the mean bus service rate for the bus. Next, the arrival rate into the bus queue \((\lambda_B(hold))\) will be the sum of the external reference rates of all PEs. Mathematically this appears as Equation 5.

\[
\lambda_B(hold) = \sum_{i=1}^{n} \left( \sum_{j=1,j\neq i}^{n} a_{ji} \right)
\]

Having specified the mean arrival rate and service rate for the bus queue (Equations 5 and 2), the mean queuing time for the bus \((Q_B(hold))\) is calculated as:

\[
Q_B(hold) = \frac{1}{\mu_B + \sum_{i=1}^{n} E_i Q_i(hold)} - \sum_{i=1}^{n} \left( \sum_{j=1,j\neq i}^{n} a_{ji} \right)
\]

Thus, a hierarchical queueing model has been developed consisting of two levels with single server queues in each. The lower level’s mean queuing times are added in as a component to determine the mean service time of the higher level queue server. The hierarchical queueing model is illustrated in Fig. 5.

![Figure 5: Queueing Model for Bus-Hold DSB System.](image)

The probability distribution function for the lower level (PE) queues’ queuing times will affect the service time distribution function which can be assumed for the server in the higher level queue. If the lower level queue has an exponential distribution for its queuing time, the service time of the higher level queue server cannot have a constant probability distribution and most likely will be exponential. In such a case, the higher level queue must be analyzed as a \(-M/\lambda-\text{or } -G/\mu-\text{type queue}.\)

4.2 Deriving Communication Cost for Bus-Hold DSB System.

Having constructed a queuing model and analyzed it to find the mean queuing time, this queuing time will now be used to arrive at an expression for communication cost. This will be the average cost per external memory reference.

\(^1\) Queuing time refers to the time spent waiting in the queue plus the time spent being served.

\(^2\) "G" denotes a "General" distribution function. That is, a distribution function which is neither Markov nor constant.
per PE in the system. Given a PE, \(P_{Ek}\), will incur a communication cost of one bus queueing time per external reference. Let \(CC_{PE_k}(\text{hold})\) be the communication cost seen by \(P_{Ek}\) per external reference. Thus, \(CC_{PE_k}(\text{hold})\) will just be equal to \(Q_B(\text{hold})\).

\[
CC_{PE_k}(\text{hold}) = Q_B(\text{hold})
\]

(6)

The PE average communication cost will be the sum of the cost per reference incurred by the individual PEs divided by the number of PEs. This is just the same as \(CC_{PE_k}(\text{hold})\) which is the same as \(Q_B(\text{hold})\). The average communication cost for the system (\(CC_{avg}(\text{hold})\)) is given by the following expression.

\[
CC_{avg}(\text{hold}) = \frac{1}{2T + \sum_{i=1}^{n} \frac{1}{Q_{E_i}(\text{hold})}} - \sum_{i=1}^{n} \left( \frac{\sum_{j=1}^{m} a_{ij} \cdot q_{ij}}{\sum_{k=1}^{m} a_{ik}} \right)
\]

(7)

It is straightforward to handle the case of multiple buses. Each PE is connected to all buses and can use any available bus. The only effect of multiple buses on the communication cost of the system is in the calculation of \(Q_B(\text{hold})\). The queuing model of Fig. 5 will look the same; however, the bus queue will be a multiple server queue with the number of servers equal to the number of buses. \(Q_B(\text{hold})\) can now be calculated from known queuing theory results for multiple server queues. The same bus queue arrival rate and service rate as before will be used to calculate \(Q_B(\text{hold})\), now for multiple buses. The computation of \(CC_{avg}(\text{hold})\) follows in the same manner as for the single server case described above except the new \(Q_B(\text{hold})\) is used.

4.3 Queueing Model for Request-Reply DSB System.

Having developed a model for the DSB system where the bus is held while the remote PE accesses the desired information, development and analysis of a model for a DSB system with tagged requests and replies now follows. Here, the bus is released between the sending of a request for data from a remote PE and the return of a reply. The queuing model for this system will have a queue for the bus and a queue for each PE (corresponding to the contention points present). Fig. 6 illustrates the queuing model for the request-reply DSB system.

All requests and replies, one of each per external reference, will enter the bus queue. Upon being transmitted over the bus, the request will enter a queue in front of its destination PE. Replies will be given to the requesting PE after transmission over the bus.

In the case of a request, the referenced PE will take the request off its queue and access local memory. A reply will be packaged and enqueued to use the bus. There is no explicit path in the model from the PEs back to the bus queue since the reply will be accounted for by references emitted from the external reference source shown.

4.4 Communication Cost for Request-Reply DSB System.

The queuing times for the queues of the model (Fig. 6), will be as follows.

\[
Q_{B(i)} \overset{def}{=} \text{mean queuing time for bus (request-reply)}
\]

\[
= \frac{1}{\mu_{B(i)} - \lambda_{B(i)}}
\]

\[
= \frac{1}{T_B - 2 \sum_{i=1}^{n} \left( \frac{\sum_{j=1}^{m} a_{ij} \cdot q_{ij}}{\sum_{k=1}^{m} a_{ik}} \right)}
\]

(7)

\[
Q_{(i)} \overset{def}{=} \text{mean queuing time for PE, (request-reply)}
\]

\[
= \frac{1}{\mu_{i} - \lambda_{(i)}}
\]

\[
= \frac{1}{\mu_{i} - \sum_{j=1}^{m} a_{ij}}
\]

(8)

The factor of 2 appears in \(Q_{B(i)}\) because every reference now has both a request and a reply.

When \(P_{E_i}\) makes an external reference to \(P_{E_j}\), the communication cost observed by \(P_{E_i}\) is \(Q_{B(i)}\) for the request plus \(Q_{(i)}\) for the actual access by \(P_{E_i}\), plus \(Q_{B(j)}\) for the reply packet to be returned. Let \(E_{ij}\) be defined as the fraction of \(P_{E_i}\)'s external references per unit time that go to \(P_{E_j}\).

\[
E_{ij} = \frac{a_{ij}}{\sum_{k=1}^{m} a_{ik} \cdot q_{ik}}
\]

(9)

Using \(E_{ij}\) and the queuing times for the network given by \(Q_{B(i)}\) and \(Q_{(i)}\), the communication cost seen by a single
$PE$ can be expressed. Equation 10 gives the communication cost seen by $PE$.

$$CC_{PE_i} = E_i(2Q_{B_i} + Q_{const})$$

$$CC_{PE_i} = \sum_{j=1, j \neq i}^n E_j(2Q_{B_j} + Q_{const})$$ (10)

Having the communication cost as seen by each $PE$ allows the $PE$ average communication cost to be calculated by summing the communication costs seen by the individual PEs and dividing by $n$. Equation 11 gives the average communication cost for a $PE$ in a DSB system using the request-reply protocol for remote accesses.

$$CC_{avg[rr]} = \frac{\sum_{i=1}^n [E_i(2Q_{B_i} + Q_{const})]}{n}$$ (11)

Finding the communication cost for a multiple bus system is almost identical to the single bus system except the bus queue of Fig. 6 is analyzed as a multiserver queue with as many servers as there are buses. The same bus arrival and service rates ($\lambda_{B_i}$ and $\mu_{B_i}$) will be used with known multiserver queueing theory results to get a new value for $Q_{B_i}$.

### 4.5 Request-Reply DSB Communication Cost for a Uniform Distribution Case

Let all processors in a single bus DSB system be homogeneous so that the external reference rate for each $PE$ is the same. Further, assume a uniform distribution where every $PE$ sends an equal fraction of its external references to each of the other PEs. That is, the fraction of $PE_i$'s external references that go to $PE_j$ is $1/(n-1)$, for all $i$ and $j$ ($i \neq j$). The implication of these assumptions is that all non-diagonal elements of the reference array, $a$, will be equal $^4$. That is,

$$a_{ij} = a_{ji} = K_i$$

where, $K_i$ is a constant.

The average system communication cost for a system with such a uniform distribution will be given by:

$$CC_{avg[rr]} = 2Q_{B_i} + Q_{const}$$

Here, $Q_{const}$ denotes the $PE$ queuing times which turn out to be the same for all PEs. The result for average system communication cost shows it to be the $PE$ queuing time plus twice the bus queuing time. This result is consistent with intuition for this uniform case where:

1. All PEs have the same local reference rate.
2. All PEs have the same external reference rate.
3. The external references from each $PE$ are evenly distributed to the $n-1$ other PEs with each receiving the same fractional part.
4. The service rate for each $PE$ is the same.

$^4$was defined in Equation 1

### 5 Comparison of Request-Reply and Bus-Hold Protocols

Direct comparison of the expressions for average system communication cost under the two protocols is a non-trivial task. Intuition would indicate that the bus-hold method will be advantageous (or less disadvantageous) when the processors are fast compared to the bus queuing time. In such a case, it would be costly to relinquish the bus once it has been granted if the mean time for service at the remote $PE$ is short relative to the bus waiting time. On the other hand, the request-reply method will be advantageous when bus access is relatively cheap compared to the queuing time at the remote $PE$. That is, when the bus queuing time is short compared to the $PE$ queuing time. This would be the case if relatively slow PEs or PEs with slow memory were to be used in the system.

Examining the two communication cost expressions for the boundary condition of infinitely fast PEs ($PE$ queueing time goes to zero), yields the following:

$$CC_{avg[bus-hold]} = 1/\left(\frac{1}{2T_B} - \sum_{i=1}^n \left(\sum_{j=1, j \neq i}^n a_{ij}\right)\right)$$

$$CC_{avg[rr]} = 2Q_{B_i}$$

It is apparent that for $PE$s with zero service time (infinite service rate) the communication cost will be the same for the two protocols. As the processors slow down (service time becomes greater than zero), one would expect the request-reply method to be more beneficial. This is because request-reply allows other bus traffic to be sent during the queuing time at the remote $PE$, which is now increasing from zero in the boundary case.

The effect of progressively slower PEs (increasing $PE$ service time) on communication cost has been examined for a single bus system under each of the two protocols. A computer program was used to generate data points which are plotted in Fig. 7.

![Figure 7: Request-Reply vs. Bus-Hold Comparison.](image-url)
6 Request-Reply Communication

Cost as a Function of Reference Distribution

The effect of changing the distribution of references on communication cost has been examined for the single bus request-reply system. This was done by algebraic manipulation of the equations for communication cost and by means of a program that evaluated the models for different reference distributions. The two methods produced matching results thus providing some assurance that neither a coding nor algebra mistake had occurred. The results from the program run will be presented here. The input data was as follows:

- Number of PEs = 4
- Bus service rate = 240
- PE service rates: PE\(_1\) = 150, PE\(_2\) = 150
- PE\(_3\) = 150, PE\(_4\) = 150

Reference rates initial values:

\[
a = \begin{bmatrix}
78 & 12 & 12 & 0 \\
8 & 78 & 8 & 8 \\
8 & 8 & 78 & 8 \\
8 & 8 & 8 & 78
\end{bmatrix}
\]

In each iteration, PE\(_1\) will subtract one reference per unit time from the number of references it sends to PEs 2 and 3. Correspondingly, the total subtracted (2 in this case) in each iteration, will be added to the number of references per unit time from PE\(_1\) to PE\(_4\).

Plotting the communication cost as a function of reference distribution for the request-reply system gives the curve of Fig. 8. It is interesting to note that the minimum average communication cost occurs where the external references are uniformly distributed.

![Communication Cost vs. External Reference Distribution](image)

Figure 8: Request-Reply Communication Cost vs. External Reference Distribution.

7 Summary

The framework for comparison of distributed systems developed by Gonzalez and Jordan [1] provides a means of determining the most suitable of several candidate systems for a set of problem requirements. To do so however, requires that a set of attributes describing various system characteristics be defined and quantified.

The DSB ("Direct Shared path global Bus") system can be subdivided into four cases. The first distinction is single or multiple buses. The second distinction is the method of operation, which could either be bus-hold or request-reply protocols. The bus-hold protocol assumes that the bus is held throughout the entire access time at the remote processor. The request-reply protocol, on the other hand, assumes a request for non-local data is sent to the remote processor and the bus is then released to perform other work. Once the remote processor has accessed the requested data, a reply containing the data is returned to the requesting processor over the bus.

Communication cost is the average amount of time required by a processor in the system to make an external reference to another processor. Expressions are derived for communication cost of DSB bus-hold and DSB request-reply systems with single buses. These expressions are also extended to the respective multiple bus cases. To derive the communication cost expressions, queueing models were constructed and analyzed to determine mean queuing times.

The expression for the request-reply protocol communication cost has been reduced from the general case to a simpler case with uniform external references and equal memory access times for each processor. Such assumptions apply to homogeneous processors and algorithms which lend themselves well to parallelism.

The behavior of communication cost as a function of processor service time was also examined for the request-reply and bus-hold protocols. Results indicate that the two methods are identical for the limiting case of processors having zero service time. Both methods experience performance degradation as processor service time increases with the bus-hold method degrading faster than request-reply.

The degradation of performance as measured by communication cost has been quantitatively illustrated. The intuitive result that a uniform reference pattern provides minimum communication cost in a homogeneous system has also been observed.

Some areas for future research include further comparison of the bus-hold and request-reply protocols for the DSB system. This will consist of examining communication cost under the two protocols as a function of external reference rates, memory access time, and communication path transfer time. The operating point of the system in terms of external reference rates, memory access time, and path transfer time will then indicate the benefit of using one or the other of the two protocols.

References


