I/O Placement Strategies and their Data Locality Sensitivities in $k \times k$ Toroidal Networks

Bader Almohammad*
Department of Computer Science
Oregon State University
Corvallis, OR 97331 - USA
almohaba@cs.orst.edu

Jong-Hyun Kim
Department of Computer Science
Yonsei University
Wonju, Kangwon-Do, 220-710, South Korea
jhkim@dragon.yonsei.ac.kr

Bella Bose
Department of Computer Science
Oregon State University
Corvallis, OR 97331 - USA
bose@cs.orst.edu

Abstract

Various I/O placement strategies have been proposed for a $k \times k$ toroidal distributed system. In previous studies, it has been shown that I/O placement strategy has a major impact on average network latency. In this study the sensitivity of different strategies toward I/O data locality is investigated. An intensive simulation was carried out for this purpose. The simulation considers both scientific and transaction computing environments. Simulation results show that the sensitivities of different strategies can vary significantly. In addition, it is noticed that the sensitivity of an I/O placement strategy is influenced by two other factors: network size, and network contention.

However, in previous studies [1,2], it was shown that I/O placement strategy has a major impact on average network latency. Furthermore, it was shown that Perfect/Quasi-perfect (PQ) I/O placement strategy, proposed in [3], is superior to Outer-Column (OC) in terms of average network latency. In a Perfect/Quasi-Perfect (PQ) placement strategy, I/O nodes are distributed uniformly in the interconnection network. More details about this strategy are discussed in Section 2.

1. Introduction

An Input/Output (I/O) placement strategy of a parallel system is a method according to which I/O nodes are distributed in the interconnection network. In practice, a simple strategy is usually employed. For instance, in Intel-Paragon system, which has a two dimensional mesh as the interconnection topology, the I/O nodes are placed at the boundary nodes. This we call as the Outer-Column (OC) I/O placement strategy. The reasoning behind using a simple I/O placement strategy is the claim that using more mature placement strategy will complicate building the system with no major gains especially if wormhole routing is used.

* Bader Almohammad is supported by Kuwait University (fellowship MATH/CS-3/93)

The sensitivity of an I/O placement strategy toward I/O data locality is the effect of varying data locality on the performance of a system that uses this strategy. Strategy $A$ is said to be less sensitive than strategy $B$ toward data locality if performance variations in a system that uses $A$ would be less than those of a system that uses $B$. It is assumed that the systems are identical except for using different I/O placement strategies. In this paper, "I/O data locality" and "data locality" are used interchangeably. Our aim is to investigate and compare sensitivities of Outer-Column (OC) and Perfect/Quasi-Perfect (PQ) strategies in $k \times k$ toroidal architectures. The measure of this sensitivity is chosen to be the average I/O or Processing Element (PE) communication latency. The rest of this paper is divided into four sections. Section 2 presents some background about Perfect/Quasi-Perfect (PQ) placement strategy. Section 3 describes workload characteristics and simulation environment. Simulation results are introduced in Section 4. Finally, conclusions of this study are drawn in Section 5.
2. Preliminaries

The Perfect/Quasi-Perfect (PQ) I/O placement strategy for toroidal networks was introduced in [4]. A partial solution to this problem has been provided in [3]. A placement is said to be perfect distance-\( t \) placement if and only if each node in the network is within a distance of \( t \) or less from exactly one I/O node. For example, Figure 1-(a) shows a perfect distance-1 placement for a \( 5 \times 5 \) torus. On the other hand, a distance-\( t \) placement is quasi-perfect if the following two conditions hold:

- Let \( a \) and \( b \) be any two I/O nodes and \( S_a \) and \( S_b \) be the sets of nodes at a distance \( t \) or less from \( a \) and \( b \) respectively. Then \( S_a \cap S_b = \phi \).
- No non-I/O node is at a distance of more than \( t + 1 \) from some I/O node.

These two conditions imply that the maximum number of non-I/O nodes are at a distance of \( t \) or less from the I/O nodes and the remaining nodes are at a distance of \( t + 1 \) from some I/O nodes. For example, Figure 1-(b) shows a quasi-perfect distance-1 placement for an \( 8 \times 8 \) torus. In this example each node is either at a distance of one from one I/O node or at a distance of two from two I/O nodes.

In [3], it was proved that for a \( k \times k \) torus, there exists a perfect distance-\( t \) placement if \( 2t^2 + 2t + 1 \) divides \( k \). For instance, if \( t = 2 \), then there exists a perfect distance-2 placement for a \( 13 \times 13 \) torus. In perfect cases, a generator matrix of the form \( [1 \ 2i t^2] \) can be used to locate the coordinates of the I/O nodes. Hence, a general formula of the I/O node locations in perfect cases is \( [i \ 2i t^2 mod k] \), where \( 0 \leq i \leq k - 1 \). Finding a general form generator matrix for quasi-perfect cases is still under investigation.

3. Workload and Simulation Environment

A modified version of Proteus [5] is used to simulate \( 5 \times 5, 8 \times 8, 13 \times 13 \), and \( 16 \times 16 \) tori. Figure 1 shows the I/O placements for these tori when Perfect/Quasi-perfect (PQ) is applied. The \( 5 \times 5 \) and \( 13 \times 13 \) tori represent perfect cases while \( 8 \times 8 \) and \( 16 \times 16 \) tori illustrate quasi-perfect placements. The simulation considers both transaction and scientific computing environments. The main distinction between them lies in the relation between I/O requests and PE-PE communication. The PE-PE communication and I/O requests overlap in transaction environment; however, in scientific environment they do not [6].

The simulation is designed to work as follows: the nodes generate requests at random instants. The average time between two sequential instants, Inter-Request time (IR), is assumed to be exponentially distributed. A specified percentage of these requests is considered to be I/O requests. The rest of the requests are set to be PE-PE communication requests (abbreviated as PE requests in the rest of this paper). The destination of a PE request is selected from other PE nodes with a uniform probability. A percentage of I/O requests that matches data locality factor is forwarded to the closest I/O node. The rest of I/O requests are uniformly distributed among the other I/O nodes. Each I/O node is assumed to satisfy all the requests it receives. The I/O requests are assumed to be single. More clearly, no I/O request makes an I/O node to broadcast to all other nodes or to accumulate from them.

In the transaction environment, the percentage of I/O requests out of the total requests usually varies from 5% to 20%. In addition, the size of PE and I/O requests are assumed to be 1 kbyte and 4 kbyte, respectively. In the scientific environment, I/O request percentage is set to be 100% since it is assumed that no overlap between PE and I/O requests exists as it is the case in SIMD model. Simulating
scientific environment was carried out for the cases in which I/O request sizes are 1 kbyte and 4 kbyte. Flit size is set to 1 byte in both environments. All the above choices of simulation parameters were based on the analysis provided in [6, 7]. Disk access time is not considered in this study. This is because the aim is to find out the effect of I/O placement strategy on variations of network latency.

4. Simulation Results

As stated earlier, the goal of this study is to compare data locality sensitivity of Outer-Column (OC) with Perfect/Quasi-Perfect (PQ). Average latency of I/O, or PE, communication is chosen to be the measure of this sensitivity. The following two sub-sections introduce simulation results in transaction and scientific environments.

4.1. Transaction Environment

Intensive simulation was executed to determine I/O request rates and IR values at which data locality sensitivity of the studied placement strategies is to be measured. The data locality rate indicates how much of I/O requests generated by a non-I/O node go to the nearest I/O node. For example, a rate of 0.4 means that 40% of the requests are forwarded to the nearest I/O node while the remaining requests are uniformly distributed among the other I/O nodes. Figure 2 shows the average I/O communication latency vs. IR with 10% I/O rate and 50% data locality in an 8 × 8 torus. In OC case, the network becomes saturated when IR = 1400 μsec. However in PQ case the network saturates at IR = 1000 μsec. Overall, it is clear that PQ results in a better performance, or shorter latency, than OC. At IR = 3000 μsec the performance gap between simulated strategies started to be stable. Figure 3 shows the average I/O communication latency vs. I/O rate at IR = 3000 μsec and data locality = 50% in an 8 × 8 torus. At I/O rate of 15% the gap in the performance started to be stable. Hence, sensitivity toward data locality in an 8 × 8 torus is simulated with IR = 3000 μsec and I/O rate of 15%. The same strategy is used with the other tori to determine values at which sensitivity toward data locality is to be measured. These values are shown in Table 1. Figure 4 shows data locality sensitivity of OC and PQ with respect to average PE-PE communication latency. Figure 5 shows the sensitivity of these strategies with respect to average I/O communication latency.

From the simulation results, the following observations can be made: (1) The sensitivity of PQ is, in general, less than or equal to the sensitivity of OC. (2) The sensitivity gap almost always increases as data locality decreases. (3) The sensitivity gap increases in larger tori. (4) The sensitivity gap increases as IR decreases. These observation can be explained according to what was shown in [2]. The PQ placement strategy reduces distances among non-I/O and I/O nodes. Therefore, the total network traffic decreases and the potential of hot spot growth is eliminated; consequently, the average network latency is reduced. One more observation is that the sensitivity gap with respect to average PE-PE communication latency is less than the one with respect to average network latency of I/O requests. The main reason behind this is the difference in the sizes of I/O and PE requests. I/O request size is 4 kbyte while it is 1 kbyte for PE request. However, this observation indicates that I/O placement strategy effects the traffic in the entire network.
Figure 4. Data locality sensitivities with respect to average PE-PE communication latency in transaction environment.

Figure 5. Data locality sensitivities with respect to average I/O latency in transaction environment.
Figure 6. I/O Inter-request vs. average network latency in 13x13 torus.

<table>
<thead>
<tr>
<th></th>
<th>5 x 5</th>
<th>8 x 8</th>
<th>13 x 13</th>
<th>16 x 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 kbyte</td>
<td>100</td>
<td>200</td>
<td>550</td>
<td>850</td>
</tr>
<tr>
<td>4 kbyte</td>
<td>1500</td>
<td>2500</td>
<td>3000</td>
<td>4500</td>
</tr>
</tbody>
</table>

Table 2. Values of IR [μsec] at which data locality sensitivity of OC and PQ are measured in scientific environment.

4.2. Scientific Environment

Simulating the scientific environment started with determining values of IR at which sensitivity toward data locality should be measured in simulated architectures. Figure 6 shows the average I/O communication latency in a 13 x 13 torus with 50% data locality and 4 kbyte I/O request size. Based on this graph, the sensitivity is decided to be measured at IR = 3000 μsec. This decision is based on the observation that the performance gap between OC and PQ started to be stable from that point on. Applying the same idea on the other dimensions resulted in the values given in Table 2.

Figures 7 and 8 show the simulation results when I/O requests are 1 kbyte and 4 kbyte, respectively. It can be noticed that the network traffic increases abruptly. This is due to the burst mode generation of I/O requests in this environment. All the observations, except the one related to PE requests, made in the case of transaction environment are also beheld here. Note that in the case of scientific environment all the network traffic is assumed to be I/O communications.

5. Conclusion

In this paper, data locality sensitivities of both Outer-Column (OC) and Perfect/Quasi-perfect (PQ) are investi-
gated. Simulation results show that the PQ is superior to OC in terms of being less sensitive towards I/O data locality. Consequently, a \( k \times k \) torus will, in general, have a smaller or equal average network latency if it uses PQ rather than OC placement. This will hold even if I/O data locality varies. Based on the simulation results, the following observations are made. First, in both of OC and PQ, the sensitivity toward I/O data locality is effected by two other factors: network size and network contention. Second, the sensitivity increases with the increase of network size or network contention. Third, the sensitivity growth in OC case is larger than in PQ case when the network size or contention is increased, or when data locality is decreased.

References


Figure 8. Sensitivity of simulated architectures in scientific environment when I/O request size is 4 byte.