The Mesh Superceded?

L. V. Kalé
Department of Computer Science
University of Illinois at Urbana–Champaign

Abstract

Two dimensional interconnection schemes have some inherent advantages because of their linear area and constant wire–lengths. The nearest–neighbor mesh is such a topology that has enjoyed a widespread acceptance. We investigate a family of bus–based topologies called the double–lattice–meshes, and propose a variation to improve their properties. We show that the bus–based topologies perform better than the mesh for a variety of communication structures. In particular, when global communication is needed, they provide larger effective bandwidth, and when localized communication is permissible, they provide largest neighborhoods for a given communication capacity.

1 Introduction

Despite the emergence of high–dimensional interconnection schemes such as the hypercube the two–dimensional topologies retain their attraction. This is mainly because of one useful property: the length of wires (communication channels) does not increase with increase in the number of PEs. This is a consequence of the *linear area* property [10]. Also, routing messages on such topologies is easier, and can be implemented efficiently in hardware [3]. The most popular two–dimensional topology is the nearest–neighbor mesh (called just the mesh in the rest of the paper). The researchers at CalTech who were instrumental in developing hypercube multi–processors have recently advocated [2] mesh–connected multiprocessors, and predict that two–dimensional meshes will be standard topologies in second generation multiprocessors. We have proposed a 2–dimensional bus–based topology called the double–lattice–mesh (DLM) [7]. It is a generalizations of the mesh. In an earlier paper [7] we have shown that the DLM is superior to a mesh when communication is to be localized to a neighborhood (this result is briefly summarized in Section 4). In this paper we show that the DLM, and a new variant of it called the laddered DLM, are superior to the mesh for almost all communication structures needed for different applications. Both the message latencies and network throughput are better for the these bus topologies compared those for the mesh.

In the next section we briefly describe the DLM, and propose an improved variant, the laddered DLM. We also derive the average number of hops between all pairs of PEs for these topologies. In the succeeding sections, we compare the performance of these topologies for applications with different communication requirements.

2 The double lattice mesh

The *lattice–mesh* is described in [5]. We repeat its definition for ease of reference: Assume that each communication channel (bus) connects S PEs. S is said to be the span of a bus. N PEs are laid out in a $\sqrt{N} \times \sqrt{N}$ matrix where $\sqrt{N}$ is a multiple of S. Each PE is connected to 2 buses. We associate a label = (1 + (x+y) mod S) with a PE (x,y). All the buses parallel to the X–axis start at a PE with label X_s, and all the buses parallel to the Y–axis start at a PE with a label Y_s, where $X_s \neq Y_s$.

The *double–lattice–mesh* (DLM) can be thought as two overlapped lattice–meshes. Here, each PE is connected to 4 buses. The simple DLM is characterized by 4 parameters, X_{s1}, Y_{s1}, X_{s2}, and Y_{s2}, similar to the two parameters of the lattice–mesh. Figure 1 shows a 12x12 DLM using buses that span 3 PEs. We assume wrap around connections in this as

well as all other topologies considered in this paper. A DLM with S=2 is identical to the nearest neighbor mesh (with wrap around connections). Thus the DLM generalizes the mesh.

How many hops does a message has to traverse to go from one PE to another? Because the topology is homogeneous, (With minor local variations due to the buses) we can assume that the source of a message is a PE at (0,0). Let the co–ordinates of the destination be (X,Y). Because of the wrap–around connections, we can assume $X, Y \leq \sqrt{N}/2$. We will derive an upper bound on the number of hops:[1] It is always possible to travel parallel to one axis covering a distance of S in 2 hops. (Referring to Figure 1, one may start from a PE labeled 1, travel along one horizontal bus as much as possible to the right, then take the other horizontal bus to travel to the next label 1. Traveling along the second bus as much to the right as possible is tempting, but cannot be sustained over many hops, as is easy to verify.) So, it is clearly possible to reach (X,Y) in $\frac{(X+Y)}{S/2} = \frac{2(X+Y)}{S}$ hops. This calculation is asymptotic, in that it is valid only if the distance, X+Y, is sufficiently larger than S. As the PE at $(\sqrt{N}/2, \sqrt{N}/2)$ is the farthest, the maximum hops $= 2\sqrt{N}/S$.

We now derive the average of the number of hops needed between the PE at (0,0) and all the other PEs. Because of the wrap around connections, it suffices to consider the PEs in the quadrant where X and Y are at most $\sqrt{N}/2$, as shown in Figure 2.

Let $\overline{hops}_{DLM}(N,S)$ denote the average number of hops for a DLM with N PEs connected using buses that span S PEs each. As PEs at a distance of i can be reached in 2i/S hops,

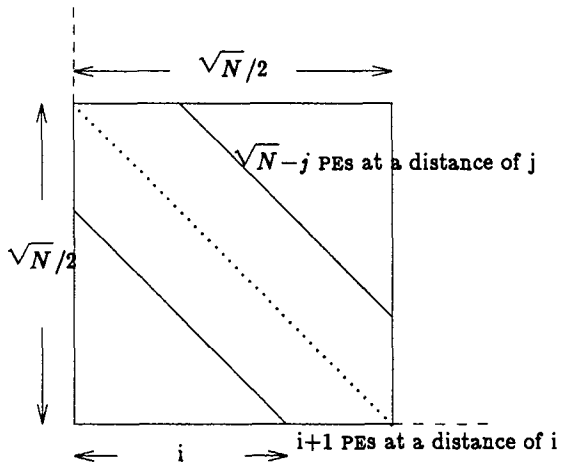
$$\overline{hops}_{DLM}(N,S) = \frac{1}{(\sqrt{N}/2)(\sqrt{N}/2)} \sum_{i=1}^{\sqrt{N}} (\textit{number of PEs at distance } i)(2i/S).$$

Referring to Figure 2, it is clear that this sum can be broken into 2 parts as:

$$= \frac{4}{NS}\left[\sum_{i=1}^{\sqrt{N}/2} (i+1).(2i) + \sum_{j=\sqrt{N}/2+1}^{\sqrt{N}} (\sqrt{N}-j).(2j)\right].$$

which, after some symbolic manipulation, simplifies to

[1] Throughout this paper we use *distance* to refer to the manhattan distance between two PEs. E.g. the distance between a PEs at (0,0) and (X,Y) is X+Y. We use the word *hops* to refer to the number of communication links visited on the shortest path between two PEs. We avoid the word *internode distance*, which is commonly used for the latter, to avoid confusion between them.



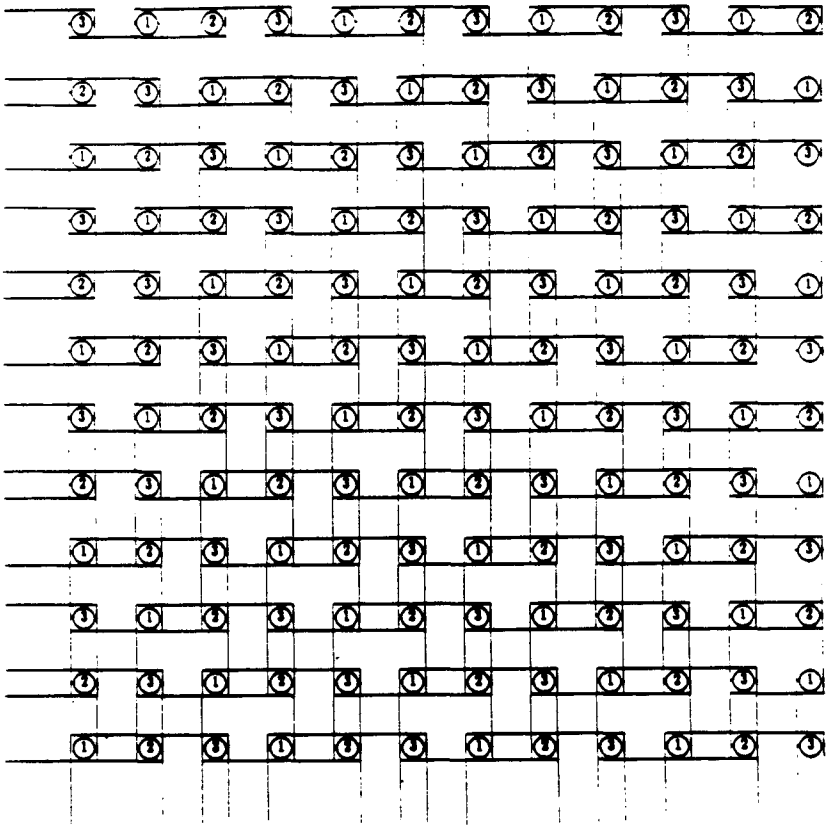
Average hops for DLM
Figure 2

$$\overline{hops}_{DLM}(N,S) = \frac{\sqrt{N}}{S} \qquad 2.1$$

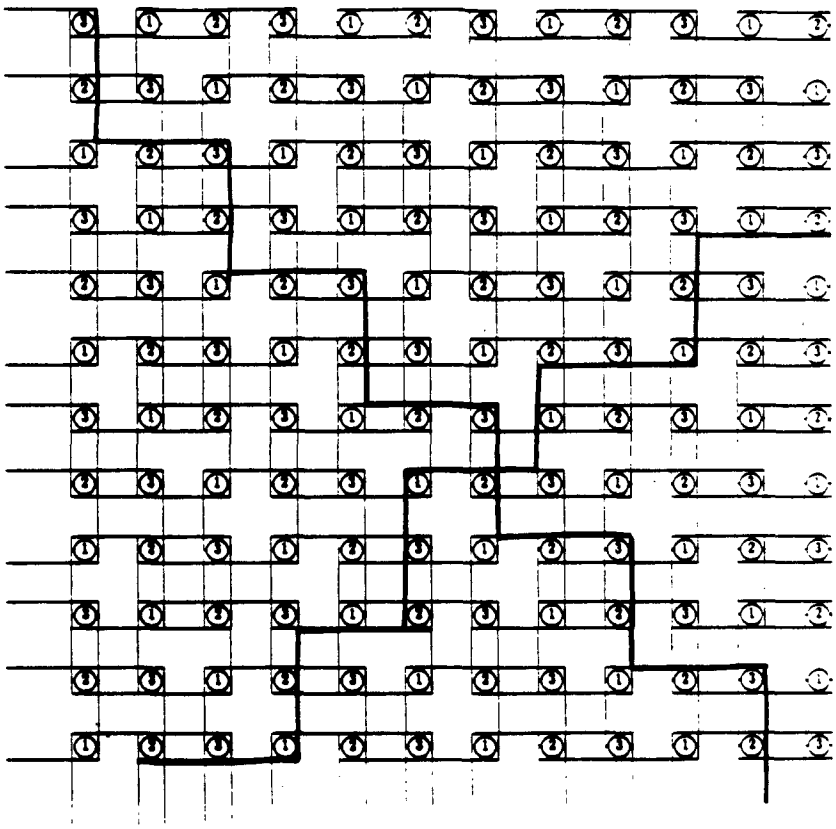
(This result can also be obtained by a pairing argument: the average of the hops for a PE at (X,Y) and its reflection about the dotted line, at $(\sqrt{N}/2-Y, \sqrt{N}/2-X)$ is always $\sqrt{N}/S$).

This is a conservative calculation. We assumed the message traverses along one axis as much as possible before traversing along the other axis. By interspersing the X and Y movements, fewer hops may suffice. Unfortunately, such interspersing cannot be done systematically enough to lead to a derivable and better formula.

The largest distance that can be traversed in one hop is $S-1$. As we saw above, such a distance cannot be sustained over long hauls. The reader should verify that even by interspersing X and Y movements, it is not possible to cover a distance of S–1 per hop in a sustained manner. Is it possible to design a bus topology such that this rate can be sustained? This question led us to the topology we describe next, *viz.* the laddered DLM (abbreviated LDLM). It can be thought as an inverted (mirror image) lattice–mesh overlapped on another normal lattice–mesh. More specifically: The layout and labeling of PEs is as before. There are 2 sets of horizontal (vertical) buses. All the buses in the first set start at PEs labeled 1, as in a DLM. The starting point of a bus in the second set depends on which row it is in (i.e. its Y co–ordinate). In a row that is distance R away from 0'th row, all the X buses in the second set start at PEs labeled $(X_s+2R-1) \bmod S + 1$. Also, in rows where the two sets of X–buses would overlap with this formula, buses in one set



A 12x12 Standard Double Lattice Mesh
Bus Span=3, $X_{s1}=Y_{s1}=1$, $X_{s2}=3$, $Y_{s2}=2$
Figure 1



A 12x12 Laddered Double Lattice Mesh
Bus-span=3, $X_{s1}=Y_{s1}=1$,
X_{s2} on the first X bus=3,
Y_{s2} on the first Y bus=2
Figure 2

are shifted right. See Figure 3 for an example of a LDLM. Thus, as we go up through the rows, the starting point of buses in the first set moves left while that of buses in the second set moves right. The vertical buses are analogously defined. X_s and Y_s are the two parameters of this topology, in addition to S.

The important feature of the new topology is the presense of *ladders*. A ladder is a sequence of alternate X and Y buses such that endpoint of one bus connects to the start point of the next bus. A ladder thus allows a sustained rate of $S-1$ physical distance per hop, as long as one is traveling along it. The worst case hops for LDLM with N PEs is $\sqrt{N}/(S-1)$.

There are ladders of both positive and negative slopes in a LDLM. It is possible to design a simple DLM (by choosing $X_{s1}=Y_{s2}$, say) that has negatively sloped ladders. However, no choice of parameters can produce a DLM with both kinds of ladders. We do need the power and 'irregularity' introduced by inverted lattice-mesh used in the LDLM. Also, notice that one can get on to a ladder, positive or negative, in one hop from any PE.

Calculation of the average number of hops is somewhat involved. Let us first calculate the hops to go from (0,0) to (X,Y). We assume $X>Y$ by symmetry. The routing strategy is shown in Figure 4. To traverse the diagonal of the square using a ladder, one needs $2Y/(S-1)$ hops. The rest of the distance is covered, as before, in $(X-Y)/(S/2)=2(X-Y)/S$ hops.

Hops between (0,0) and (X,Y) *in a LDLM* =

$$\frac{2Y}{(S-1)} + \frac{2(X-Y)}{S} \approx \frac{2X}{S} \qquad 2.2$$

The approximation is valid whenever S is sufficiently larger than 1. Also, X and Y must be larger than S.

To calculate the average number of hops, notice that (see Figure 5) all the PEs on the perimeter of a (2i)x(2i) square are at $\frac{i}{(S/2)}$ hops away from the source PE at center of the square. As there are 8i PEs on the perimeter, we can write the average number of hops as:

$$\overline{hops}_{LDLM}(N,S) = \frac{1}{N}\sum_{i=1}^{\sqrt{N}/2} 8i \cdot 2i/S$$

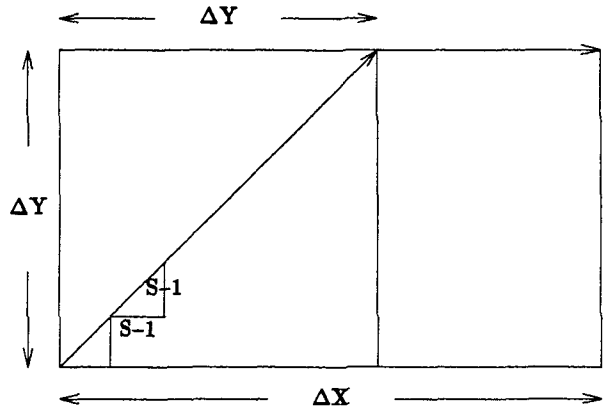
$$= \frac{16}{SN}\sum_{i=1}^{\sqrt{N}/2} i^2 = \frac{16}{SN}\frac{K(K+1)(2K+1)}{6} \text{.sp} \quad 0.4 \quad \text{where}$$

$K=\sqrt{N}/2$. Simplifying, we get Equation 2.3:

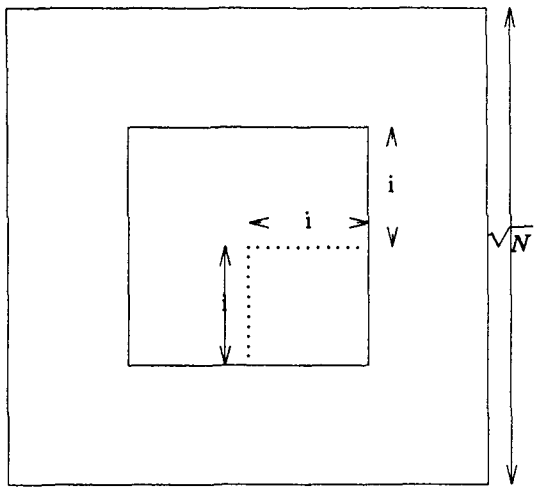
$$\overline{hops}_{LDLM}(N,S) = \frac{4}{6SN}\sqrt{N}(\sqrt{N}+1)(\sqrt{N}+2) \approx \frac{2\sqrt{N}}{3S}$$

3 Global Communication

Now let us consider a specific communication structure: Assume that each PE is equally likely to send a message to any other PE. Here, the performance of a topology is measured in terms of how much data it can deliver per unit time. We define the *delivered bandwidth* as the number of bytes of messages that can be delivered to their destinations



Routing on the LDLM
Figure 4



PEs on the perimeter of a square in a LDLM
(All the PEs are at 2i/S hops from the center)
Figure 5

per unit of time, in the steady state.

If a topology has B buses, and if a message traverses H hops on the average, clearly the delivered bandwidth is at most B/H times the bandwidth of an individual bus. We now compare this upper-bound for the different topologies.

A DLM has 4N/S buses; so does a LDLM. A mesh is special case of DLM, with S=2. So it has 2N buses, as expected. For H, we use the asymptotic expressions we computed in the previous section. The results are shown in the table below:

	Mesh	DLM	LDLM
Number of buses	2N	4N/S	4N/S
Average Hops	$\frac{\sqrt{N}}{2}$	$\frac{\sqrt{N}}{S}$	$\frac{2\sqrt{N}}{3S}$
Delivered bandwidth	$4\sqrt{N}$	$4\sqrt{N}$	$6\sqrt{N}$

Table I: Delivered bandwidth with global communication

Thus, the LDLM provides 1.5 times the bandwidth of a mesh of the same size. The DLM at least provides as much bandwidth as the mesh. As the calculation of hops for the DLM was conservative, we can expect the actual bandwidth to be better than the mesh. An added advantage of bus topologies is that they require fewer hops than the mesh. When the communication load is low, the net time required to transfer an individual message is smaller in a bus topology than on a mesh. Thus, the recommendation is clear: for connecting a 2-dimensional matrix of PEs for an application involving global communication, use a LDLM (or a DLM if its regularity is preferred) as opposed to a mesh. What span of buses (S) should be used? To minimize the maximum hops and to justify our approximation of $S \gg 1$, it seems tempting to say 'larger the better'. That would lead us to $S = \sqrt{N}$. However, at such large values of S the boundary effects become significant enough to invalidate the results obtained by asymptotic analysis above. (For example, the average hops for a DLM with $S = \sqrt{N}$ are close to 2, but the formula predicts 0.67.) Hardware considerations also dictate a limit on S. We present the results of some empirical experiments to obtain realistic comparisons between specific topologies, and to select values of S to optimize the delivered bandwidth.

3.1 Empirical experiments

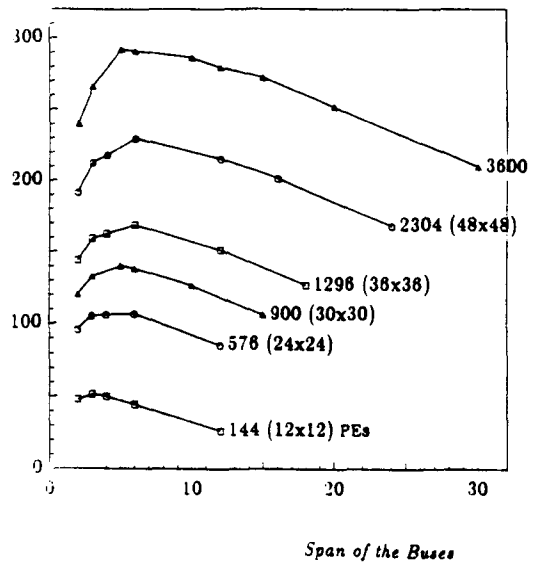
A set of programs available under the ORACLE simulation system at University of Illinois were used to produce the adjacency data for the topologies concerned, and to compute the average number of hops for each specific topology. These numbers were used to obtain the delivered bandwidth as before. The results are shown in Figure 6 and Figure 7. The mesh appears as a special case, with $S=2$, on both the graphs. Each curve refers to a fixed number of PEs. For example, the top curve for the LDLMs shows the delivered bandwidths of 3600 PE systems arranged as 60x60 matrix, with various bus-spans. The leftmost point (S=2) corresponds to a mesh, and yields a net delivered bandwidth equal to 240 times that of an individual bus. Using buses that connect 6 PEs each (S=6), the effective bandwidth rises by more than 35% to 320, which turns out to be optimal for 3600 PEs.

The graphs clearly demonstrate the superiority of bus-based design (i.e. S>2). Also, it can be seen that relatively short buses are enough to obtain the optimal performance. For the topologies considered, for both sets of curves, the maximum is obtained for bus spans of 6 or less.

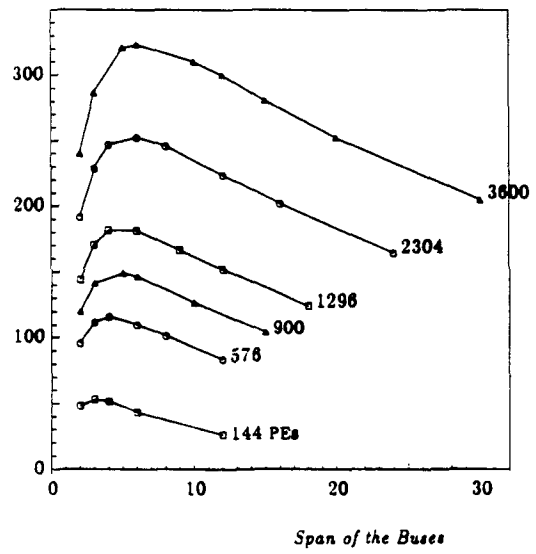
4 Localised communication

As we saw above, when uniformly distributed global communication is required, the 2-dimensional topologies we looked at provide only $O(\sqrt{N})$ delivered bandwidth although there are N PEs waiting for these messages. Thus, global communication leads to a communication bound system. Fortunately, for a significant class of applications, global communication can be avoided. Consider the parallel execution of functional programs (or logic programs, or any algorithm with divide-and-conquer flavor). Every *task* typically spawns smaller sub-tasks, collects and combines results from the sub-tasks, and returns them. The sub-tasks can be executed *on any PE*; in particular, it is possible to spawn the sub-tasks on a PE within a pre-specified neighborhood around the PE that is executing the parent task. The neighborhood is specified in terms of a bound on the number of hops. Smaller the neighborhood, smaller is the number of hops each message has to travel (each message is either a *create-sub-task* message or a *response* from a sub-task), and smaller is the required bandwidth. However, for uniform distribution of work, it is preferable to have access to a larger pool of PEs for allocating a sub-task. This is true of many dynamic load balancing strategies such as those described in [6] and [7]. The problem then is to maximize the number of PEs in the neighborhood while satisfying the bandwidth requirements.

We have shown elsewhere [7] that the DLM provides opportunities for optimizing the neighborhood, and that DLMs with S>2 provide significantly larger neighborhoods than the meshes of same size. We simply cite this result



Delivered Bandwidth for Various sizes of topologies: DLM
Figure 6



Delivered Bandwidth for Various sizes of topologies: LDLM
Figure 7

here to point to another communication structure for which DLM performs better than the mesh.

5 Arbitrary communication structures

Now we consider arbitrary communication structures. Let the application require that messages be sent to PEs at distance of i with probability q_i. (Recall that by *distance* we mean the simple Manhattan distance which does not depend on the span of the buses). To compute the net bandwidth that is avaliable to an application, we compute average number of hops a message in the application travels, for each of the three topologies. The formula for the number hops that we have to use now is:

$$\overline{hops}_{\tau}(N,S) = \sum_{i=1}^{\sqrt{N}} q_i \; \overline{hops}(\tau,S,i) \qquad 5.1$$

where $\overline{hops}(\tau,S,i)$ is the average number of hops needed to reach PEs at distance i, in a topology τ, with a bus-span of S.

For τ = mesh, the number of hops is identical to the distance, and we get:

$$\overline{hops}_{MESH}(N,S) = \sum_{i=1}^{\sqrt{N}} q_i \; i \qquad 5.2$$

For a DLM, $\overline{hops}(DLM,S,i) = 2i/S$. (Recall: $i = \Delta X + \Delta Y$). So,

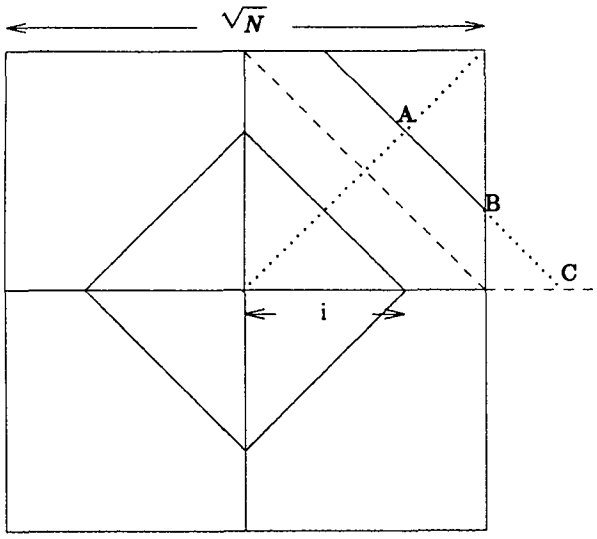
$$\overline{hops}_{DLM}(N,S) = \sum_{i=1}^{\sqrt{N}} q_i \frac{2i}{S} = \frac{2}{S} \sum_{i=1}^{\sqrt{N}} q_i \; i \qquad 5.3$$

For the LDLM, the number of hops are not the same for all the PEs at distance i, because they depend on max(ΔX, ΔY). Figure 8 shows all the PEs at a distance of i from the source PE at (0,0). By the various symmetries involved, it suffices to compute the average hops for the strip of i/2 PEs darkened in the figure. For all these PEs, the X co-ordinate is larger than the Y co-ordinate. The number of hops needed to reach a PE at (K,_) is then 2K/S. As K varies from i/2 to i, we get

$$\overline{hops}(LDLM,S,i) = \frac{1}{i/2} \sum_{k=i/2}^{i} \frac{2K}{S} =$$

$$\frac{4}{iS} \sum_{k=i/2}^{i} K = \frac{4}{iS} \cdot \frac{3i^2+2i}{8} = \frac{3i+2}{2S} \approx \frac{3i}{2S} \qquad 5.4$$

Examining Figure 8 carefully, it can be seen that the formula is not exact for PEs beyond a distance of $\sqrt{N}/2$. It computes the average for all the PEs (some of them hypothetical) along the line AC, while we are interested in



Average hops for the LDLM:
PEs at the distance i from the source
Figure 8

the strip AB. However, the average for the strip AC is clearly larger than that for AB. So, the formula 5.4 really gives an upper bound on the average number of hops. I.e.

$$\overline{hops}(LDLM,S,i) \leq \frac{3i}{2S}. \qquad 5.5$$

Substituting equation 5.5 in 5.1, we get:

$$\overline{hops}_{LDLM}(N,S) \leq \sum_{i=1}^{\sqrt{N}} q_i \frac{3i}{2S} = \frac{3}{2S} \sum_{i=1}^{\sqrt{N}} q_i \cdot i \qquad 5.6$$

The results of these calculations are shown in the table below.

	Mesh	DLM	LDLM
Number of buses	2N	4N/S	4N/S
Average Hops	$\sum q_i\, i$	$\frac{2}{S} \sum q_i\, i$	$\leq \frac{3}{2S} \sum q_i\, i$
Delivered bandwidth	$\frac{2N}{\sum q_i\, i}$	$\frac{2N}{\sum q_i\, i}$	$\geq \frac{8N}{3 \sum q_i\, i}$
Ratios	1	1	≥ 1.333

Table II: Arbitrary communication structures
(all the sums in the table range from i=1 to $\sqrt{N}$)

The results are similar to those for 'global communication': The mesh and the DLM perform equally well (again, the DLM may be better because of the conservative calculation of hops), and the LDLM performs provably better than the mesh. The advantages of fewer hops also argue in favor of bus topologies, as before.

Notice that the results apply as long as the approximations used to derive them are valid. In particular, the expressions used for $\overline{hops}(r,S,i)$ are valid only for $i \gg S$. For most communication structures (i.e. distributions of q_i) this is a valid approximation. However, in a communication structure known as 'sphere of locality' [9], where q_i=0 for all $i>R$, it is valid *only if* the radius of the 'sphere', R, is sufficiently larger than the span of buses, S. An interesting and important case is that of communication within a distance of 1 only. Obviously the equations do not apply, as R=1. Intuitively, the buses provide connections to PEs beyond a distance of 1, which are of no use for the communication required in the application. Thus for purely systolic algorithms, and other algorithms that do need only the near-neighbor communication (such as many low-level vision algorithms) the mesh provides better bandwidth than the bus topologies. Again, if the application needs a mix of near-neighbor and global communication, a bus topology may be worth investigating.

6 Discussion

The use of buses may seem to present a problem: resolving the contention for a bus. However, a mesh is not immune from that problem either, unless one has dedicated uni-directional channels, which doubles the cost of ports for each PE. More important, it is relatively simple to resolve the contention by using a round robin strategy with an appropriate protocol. No expensive hardware such as that used for ethernets is needed. Many real multi-processors, such as ELXI and S/Net [1] have been designed around a single bus. For the practical importance of keeping the length of buses short, the reader is referred to [1]

Prasanna kumar [8] has proposed the use of row and column buses in addition to the regular mesh connections. He demonstrates many algorithms that run efficiently on such a topology. We believe that a bus spanning $\sqrt{N}$ PEs will be impractical beyond systems with a few hundred PEs. It might be useful, then, to consider the topologies discussed in this paper for such applications. A group at MITRE has also proposed [4] a different variety of planar bus topologies that are particularly suitable for wafer scale implementation.

We considered only the topologies with a wrap around. It seems reasonable to assume that similar relationships will hold between the versions of the corresponding topologies without the wrap-around connections. We plan

to conduct more empirical calculations to confirm that.

7 Conclusion

We derived the average 'internode distance', in terms of the number of hops, for a bus topology called the Double Lattice–Mesh. We proposed a variant of this topology, the laddered DLM, that minimizes the average and the worst–case number of hops. We showed that for a variety of communication requirements, the DLM and the LDLM perform better than the mesh. For the case of global communication, we also confirmed the results by empirical calculations for specific topologies. Although all the topologies considered have the same asymptotic *order* of performance (e.g. for global communication, all provide $O(\sqrt{N})$ bandwidth), the bus topologies are better by a multiplicative constant. Therefore we can state that for most applications where a mesh can be used, a bus topology such as a DLM or LDLM will perform better. The one important exception involves applications that need communication within a very narrow neighborhood, such as purely systolic algorithms. Here, the large number of channels of a mesh proves advantageous.

8 References

1. S. Ahuja, "S/Net: A High Speed Interconnect for Multiple Computers", *IEEE Journal on Selected Areas in Communications*, SAC–1, 5 (November 1983), .
2. W. C. Athas and C. Seitz, "Multicomputers: Message–Passing Concurrent Computers", *Computer*, August 1988, 9–24.
3. W. J. Dally and C. L. Seitz, "Deadlock–free Message Routing in Multiprocessor Interconnection Networks", *IEEE Trans. Computers*, **36**, 5 (May 1987), 547–553.
4. J. D. Harris and H. E. T. Connell, "An Interconnection Scheme for a tightly coupled massively parallel computer network", Proceedings of ICCD, PortChester, N.Y., 1985.
5. L. V. Kale, "Lattice–Mesh: a Multi–Bus Topology", Proc. of ICPP, St. Charles, Illinois, August 1985.
6. L. V. Kale, "Parallel Architectures for Problem Solving", Doctoral Thesis, Dept. of Computer Science, SUNY, Stony Brook, Dec. 1985.
7. L. V. Kale, "Optimal Communication neighborhoods", Proc. of ICPP, St. Charles, Illinois, August 1986.
8. V. K. P. Kumar, "Array Processor with Multiple Broadcasting", Intl. Conference on Computer Architecture, 1985.
9. D. A. Reed, *Performance Based Design and Analysis of Multimicrocomputer Networks*, Ph.D. Thesis, Purdue Univ., May 1983.
10. J. D. Ullman, "Flux, Sorting and Supercomputer Organization for AI Applications", *Journal of Parallel and Distributed Computing*, 1, (1984), 133–151.