Optics Communications 89 (1992) 178–182
North-Holland

OPTICS
COMMUNICATIONS

Organization of information flow in computation for efficient utilization of high information flux communication media

Haldun M. Ozaktas [1] and Joseph W. Goodman

Information Systems Laboratory, Durand Building, Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA

Received 30 May 1991; revised manuscript received 16 December 1991

We discuss the importance of organizing information flow in computation in a manner enabling multiplexing of signal paths with distinct source and destination localities. This allows efficient use of high bandwidth interconnection media, leading to a decrease in system size and propagation delay for communication limited layouts. Among the three methods we consider towards this end, the fat-tree architecture is found to be near optimal.

1. Introduction

A major advantage of optical and superconducting interconnections is their ability to transfer large amounts of information per unit cross section over long distances. Let the maximum information flux a given communication medium can support be denoted by $\mathscr{I}$ and be measured in bits/m^{2}s. For the length scales involved in a computing system (<10 m), it is possible to reduce the effects of dispersion and attenuation to the extent that $\mathscr{I}$ may be assumed to be independent of length for optical and superconducting interconnections [1–3]. On the other hand, $\mathscr{I}$ is a decreasing function of communication length for resistive interconnections, making them disadvantageous over longer distances. However, for distances less than about the order of a centimeter, they can provide greater information flux optical or superconducting interconnections.

Let T denote the minimum pulse repetition interval for a single physical optical communication channel (i.e. corresponding to a single spatial degree of freedom). Since we are ignoring dispersion, T will probably be set by the speed of the switching devices or electrooptic transducers. If wavelength division multiplexing is employed, an appropriate effective value of T should be used.

We assume that we would like to establish a prespecified pattern of $n/2$ pairwise connections among a collection of $n \gg 1$ points. For simplicity the extension to fan-out and fan-in is not considered. Although we restrict ourselves to a fixed connection pattern, the extension to reconfigurable or message routing systems is possible. We also limit ourselves to single layer two-dimensional layouts, the extension to multi-layer and three-dimensional layouts being straightforward. B will denote the rate at which binary digital pulses are emitted into each connection. Our purpose is to implement the given pattern of connections in a manner that results in smallest possible system area, which we assume is dominated by the space required for establishing communication.

The number of binary pulses in transit at any given time in an optical communication network occupying area A may not exceed $\sim A/(f\lambda cT)$, where c and λ denote the speed of light and wavelength of radiation respectively [1]. f is a dimensionless constant factor which in principle can approach the order of unity, but may be larger in practice. The number of pulses in transit in our system is given by [#1]

[1] Present address: Department of Electrical Engineering, Bilkent University, 06533 Bilkent, Ankara, Turkey.

[#1] Since order of magnitude accuracy is sufficient for the purpose of this paper, we ignore constants such as 2, $\sqrt{2}$ etc. for simplicity.

$(n/2)B\tau_{ave} \sim nB\tau_{ave}$ where $\tau_{ave}=l_{ave}/c$ is the average propagation delay and l_{ave} is the average interconnection length. l_{ave} is often expressed in the literature in terms of the system linear extent $L=A^{1/2}$ as $l_{ave}=\kappa n^{q-1}L$ where κ is a constant coefficient and $1/2 \leqslant q \leqslant 1$ is a measure of the connectivity of the system [4]. Using these relations, we find an approximate lower bound on the area and linear extent of our system

$$nB\tau_{ave} \leqslant A/f\lambda cT\,, \tag{1}$$

$$A \geqslant n(BT)c\tau_{ave}f\lambda = n(BT)l_{ave}f\lambda\,, \tag{2}$$

$$L=A^{1/2} \geqslant \kappa n^{q}(BT)f\lambda\,. \tag{3}$$

The above bounds represent the intrinsic information carrying capacity of optical wavefields and apply to any architecture or implementation. Notice the tradeoff between system size and B.

One way of implementing the desired pattern of connections is simply to allocate $[BT] \simeq \max(BT, 1)$ parallel channels between every pair of points to be connected. When $BT \geqslant 1$, such an implementation is as efficient as any other in terms of making maximum usefulness of the available capacity of the optical channels. In this case, the above lower bounds may be approached, for instance, by use of waveguides with effective [#2] line to line spacing of $\sim f\lambda$. To see this, notice that the total area required for communication is $(n/2)(BT)l_{ave}f\lambda \sim n(BT)l_{ave}f\lambda$ since $BT \geqslant 1$ physical channels, each occupying an average area of $l_{ave}f\lambda$ is allocated per connection. Thus the total communication area $A = n(BT)l_{ave}f\lambda = n(BT)\kappa n^{q-1}A^{1/2}f\lambda$, leading to eq. (3). However, if B is less than $1/T$, the channels are underutilized and the bound of eq. (3) cannot be approached, since no matter how small B is, a channel with capacity $1/T$ is allocated for every pairwise connection. Thus when $b < 1/T$, the layout area is not any less than when $B=1/T$, so that L can at best approach the bound

$$L \geqslant \kappa n^{q} f\lambda\,. \tag{4}$$

In this paper we concern ourselves with methods of restoring the broken tradeoff between system size and B when $BT<1$. If B is independent of n, such methods may lead to reduction of the system linear extent by a constant factor of BT, compared to the direct implementation just described (eq. (4)). In some case, B may decrease with increasing n, since the computational processes become bottlenecked by the increasing propagation delays so that it is not useful to employ high bit repetition rates. When this is the case, restoring the mentioned tradeoff allows one to slow down the growth rate of system size as a function of n, as evident from eq. (3). In fact, if B decreases at the same rate as $1/\tau_{ave}$, linear growth of A as a function of n can be achieved regardless of the value of q, as evident from eq. (2).

To achieve our objective, we would like to multiplex $1/BT > 1$ independent signal paths into the same physical channel, so as to saturate its capacity. However, this is not straightforward when the many signal paths have distinct source and destination localities. In the next sections we describe three architectures which enable information flow to be organized in a manner enabling overlap between such signal paths, allowing them to be multiplexed. The reduction in the number of physical channels thus possible results in a decrease in system size and propagation delay for communication limited layouts.

2. The multiplexed grid architecture

The multiplexed grid architecture is based on the family of k-ary m-dimensional meshes (grids) of $k^m=n$ nodes [5]. The hypercube is a special case with $k=2$ and $m=\log_2 n$. For sake of illustration, we consider the case $m=2$ and $k=n^{1/2}$, which corresponds to the familiar planar mesh with $n^{1/2}$ nodes on an edge. An arbitrary connection is established in several nearest neighbor (in m-space) "hops", and multiplexed together with other connections with which it overlaps, as illustrated in fig. 1. Notice that this procedure enables us to break down in independent signal paths into overlapping segments which may then be multiplexed. If at least $1/BT$ connections can be overlapped along each edge of the mesh, then complete utilization of the available capacity $1/T$ of the physical channels may be achieved [#3]. Fi-

[#2] That is, including all inefficiency factors due to routing etc.

[#3] It is assumed that switching devices with response time at least as fast as T are available.

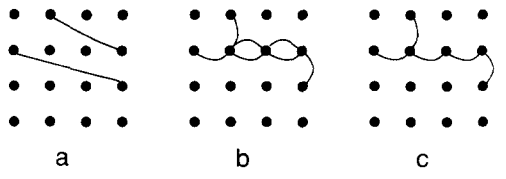

Fig. 1. The multiplexed grid architecture with $m=2$, $k=4$ and $n=16$. (a) Two of many to-be-established connections. (b) Each connection established in several hops. (c) Overlapping portions of these connections multiplexed into high capacity channels, reducing the total number of physical channels and thus layout area.

nally, the multiplexed m-dimensional mesh is laid out in two dimensions, as described in ref. [5]. Of course, this is a trivial task when $m=2$.

The price that must be paid in return for efficient utilization of the high capacity optical channels is the additional area and delays associated with routing of independent signal paths. Low dimensional meshes allow a larger number of connections to be overlapped, but increase the number of hops, and hence the number of device delays a signal must go through. High dimensional meshes decrease the number of hops but do not enable as many signal paths to be overlapped and multiplexed, possibly resulting in less than complete utilization of the capacity of the channels and thus larger layout area and propagation delays. The optimal of m minimizing overall signal delay (propagation plus device) is found to decrease with increasing n and asymptotically approaches 2 for two-dimensional layouts. In this case $\propto n^{1/2}$ node delays are suffered in the worst case. This is the basis of the *delay balanced* architecture of Hartmann and Ullman [6].

The problem of finding the optimal dimension of a multidimensional mesh organization was previously discussed by Dally [5] in the context of a message passing multiprocessor, based on somewhat different considerations.

3. The multiplexed global interconnection architecture

We now turn our attention to another architecture, illustrated in fig. 2. The n points among which connections are to be established are partitioned into n/n_1 "modules" of n_1 points each. All connections

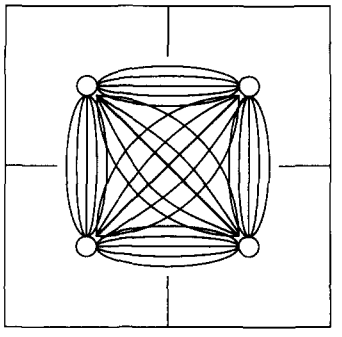

Fig. 2. The multiplexed global interconnection architecture with $n/n_1=4$. Connections internal to a module (not shown) are made directly, probably with conductive wiring, A connection to a destination in another module is first wired to a common locality with other connections destined to the same target module and multiplexed together. Demultiplexing takes place at the destination module, followed by wiring to the individual destinations. Thus two node delays are involved for global connections.

between points in one particular module to another particular module are bundled together and multiplexed into the smallest possible number of physical channels. The relatively short connections between points in the same module are made directly and would probably be implemented with conductive wiring, because of the greater density they offer over short distances.

The larger the value of n_1, the larger the number of connections between each module pair, so that a greater number of independent signal paths may be bundled (overlapped) and multiplexed together, resulting in a reduction of the area consumed by global communication channels. On the other hand, increasing n_1 also increases the area required by the internal connections. Thus there is an optimal value of n_1 resulting in minimum system area.

The multiplexed global interconnection architecture is not very useful for connection patterns exhibiting a great degree of locality. In such systems there will not be enough connections between distant module pairs to saturate the capacity $1/T$ of a single physical channel. It may be useful, for instance, for the implementation of fine grain parallel random access machines [6] or connectionist systems.

As a more specific example, this architecture can be used to implement a complete graph on n nodes with $BT \ll 1$. This might roughly model a multiprocessor interconnection scenario where each element

wants to be able to talk to every other, but only at a relatively low term average data rate.

4. The multiplexed fat-tree architecture

The fat-tree architecture, illustrated in fig. 3, was first advocated by Leiserson [7] in a multiprocessor interconnection context. We define the fat-tree to have $[n'^q] \simeq n'^q$ connections emanating from sub-trees containing n' points, where q is the same measure of connectivity introduced in the first section. That is, this rate of growth of capacity as we climb the tree is consistent with a layout with average connection length $\propto n^{q-1}A^{1/2}$, as discussed in detail by Donath [8,9] and Feuer [4]. Our definition is somewhat different than that originally given by Leiserson.

For concreteness, let us assume that waveguides of effective line to line spacing of $f\lambda$ are used. Assuming that the area required for routing functions [#4] is not the limiting factor, the linear extent $L(n')$ of a sub-tree containing n' points can be seen to satisfy the following recursion

$$L(n') = 2L(n'/4) + \max(n'^q BT, 1) f\lambda \qquad (5)$$

since n'^q connections emanating from a sub-tree of n' points can be multiplexed into $\max(n'^q BT, 1)$ physical channels. $L(1)$ corresponds to the linear extent of a single "point" and will be assumed to be negligibly small. From this recursion we may show that $L(n)$ approximately satisfies

$$\max[n^q(BT), n^{1/2}] f\lambda \leqslant L(n) \leqslant \max[n^q(\log_2 n^{1/2})(BT), n^{1/2}] f\lambda . \qquad (6)$$

The second term $\propto n^{1/2}$ is unavoidable for *any* two-dimensional layout. The first term corresponds to the communication area and is what we are interested in. Upon comparison with eq. (3), we observe that the multiplexed fat-tree allows the smallest possible system size to be approached within a logarithmic factor [#5]. (Of course, if BT is not small enough to satisfy $BT < \kappa/\log_4 n$, the use of a fat-tree may not prove advantageous.) What essentially happens is that the total communication area is dominated by the longer higher level connections, in the same sense that a geometric series is dominated by its leading terms. We succeed in multiplexing these to the greatest possible extent so that we can reduce the layout area near to that predicted by eq. (3). The fat-tree architecture allows a greater number of signals to be multiplexed than the multiplexed grid architecture at a cost of fewer node delays.

Once again the price paid is the area and delays associated with routing functions. It would probably be preferable to partition the n points into n/n_1 modules with direct internal connections (implemented with conductive wires) and then use a fat-tree organization to connect the n/n_1 modules. This would enable reduction of the number of device delays incurred and the routing circuitry. Larger values of n_1 would result in fewer node delays. Smaller values of n_1 would enable multiplexing at deeper levels

[#4] I.e. the areas of the switching networks to be contained in the internal nodes.

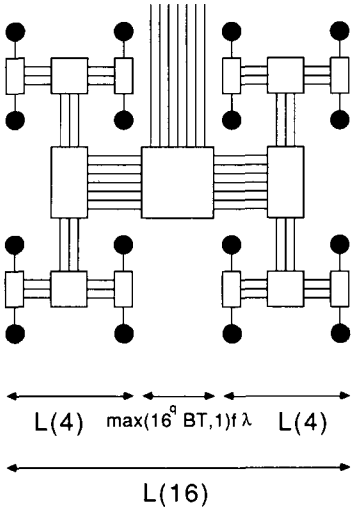

Fig. 3. The multiplexed fat-tree architecture. The points to be connected are located at the leaves, and the internal nodes provide routing functions. Each connection is established in several hops, $2\log_2 n$ in the worst case. The number of connections emanating from the sub-trees increase as we go up the tree. The overlapping portions of the connections are multiplexed into the smallest possible number of physical channels.

[#5] A similar optimality result can be shown to hold also for three-dimensional layouts if, ignoring constructional difficulties, we assume the optical communication channels can be freely routed in three dimensions.

of the system. A detailed simulation would reveal the optimal value of n_1.

5. Conclusion

We have discussed the importance of organizing information flow in a manner enabling maximum multiplexing of independent signal paths, leading to a reduction in the number of physical interconnections, which results in smaller communication area and propagation delays. Among the architectures discussed, the fat-tree is near optimal in this respect.

The latter two of the presented architectures provide a natural environment for the joint use of optical and conducting interconnections so as to bring out the best in both and may prove more promising than simple replacement of individual long wires with optics. Optical interconnections are used to provide high density/bandwidth multiplexed information transfer over long distances. Submicron scaled normal conductors are used to provide communication at a density unachievable with optics over shorter distances. This is also consistent with the energetic properties of the interconnection media. Optical interconnections consume less energy per transmitted bit over longer distances compared to normal conductors [10–12].

Both the multiplexed global interconnection architecture and the fat-tree architecture are especially suited for high density (i.e. f close to unity) free-space optical implementations because of the regular pattern of interconnections. Of the two, the fat-tree architecture is useful even for relatively local connection patterns [#6].

[#6] However, cases of extreme locality may not benefit from the use of such schemes in the first place.

Needless to say, a multitude of issues must be considered in contemplating the construction of a high performance computing system. We have focused our attention on the limitations imposed by the area consuming long distance connections and discussed how these limitations can be alleviated by exploiting the high bandwidth potential offered by optical and superconducting interconnections.

Acknowledgements

This work was supported by the Air Force Office of Scientific Research under Grant No. AFOSR-88-0024.

References

[1] H.M. Ozaktas and J.W. Goodman, J. Opt. Soc. Am. A 7 (1990) 2100.
[2] H. Kroger, C. Hilbert, U. Ghoshal, D. Gibson and L. Smith, IEEE Circuits and Devices magazine (May 1989) p. 16.
[3] R.C. Frye, IEEE Circuits and Devices Magazine (May 1989) p. 27.
[4] M. Feuer, IEEE Transactions on Computers 31 (1982) 29.
[5] W.J. Dally, A VLSI Architecture for concurrent data structures (Kluwer Academic Publishers, Dordrecht, 1987).
[6] A.C. Hartmann and J.D. Ullman, in: Model categories for theories of parallel systems, eds. G.J. Lipovski and M. Malek, Parallel computing: theory and experience (Wiley, New York, 1986).
[7] C.E Leiserson, IEEE Transactions on Computers 34 (1985) 892.
[8] W.E. Donath, IBM Journal of Research and Development 25 (1981) 152.
[9] W.E. Donath, IEEE Transactions on Circuits and Systems 26 (1979) 272.
[10] R.K. Kostuk, J.W. Goodman and L. Hesselink, Appl. Optics 24 (1985) 2851.
[11] M.R. Feldman, S.C. Esener, C.C. Guest and S.H. Lee, Appl. Optics 27 (1988) 1742.
[12] D.A.B. Miller, Optics Lett. 14 (1989) 146.