hardware implementation of the node.

References:

1. Louis Pouzin, "Virtual Circuits vs. Datagrams - Technical and Political Problems, "1976 Nat. Comput. Conf., AFIPS Conf. Prec., vol.45, 1976, pp.483-494.
2. W.L.Price, "Data Network Simulation, "Computer Networks, vol. 1.1977, pp. 199 - 210.
3. D.B.Johnson, "Efficient Algorithms for shortest Paths in Sparse Networks, "Journal of the Association for Computing Machinery, Vol.24, No.1, Jan, 1977,pp.1-13.
4. J.M.McQuillan, D.C.Walden "The ARPA Network Disign Decisions, "Computer Networks, vol.1, Aug. 1977, pp,243-289.
5. A.Giessler, J.Hanle, A.Konig, E.Pade,"Free Buffer Allocation - An Investigation by Simulation, "Computer Networks", vol.2, July 1978, pp.191-208.
6. D.E.Carision, "Bit-Oriented Data Link Control Procedures " , IEEE Transactions on Communications, vol.COM-28 No.4, Apr. 1980,pp.455-467.
7. Mischa Schwartz, T.E.Stern, "Routing Techniques Used in Computer Communication Networks", "IEEE Transaction on Communications, vol.COM-28
8. Mario Geria, Leonard Kleinrock, "Flow Control: A Comparative Survey, "IEEE Transactions on Communications" vol. COM-28 No.4, Apr. 1980,pp.553-574.
9. J.M.McQuillan, Ira Richer, E.C. Rosen, "The New Routing Algorithm for the ARPANET, "IEEE Transactions on Communications, vol. COM-28 No.5, May 1980,pp.711-719.
10. Louis Pouzin, "Methods, Tools and Observations on Flow Control in Packet-Switched Data Networks, IEEE Transactions on Communications, vol.COM-29 No.4, Apr. 1981,pp.413-426.
11. A.Giessler, J. Hanle, A. Konig, E. No.4,Apr. 1980,pp.539-552.
    Pade, "Flow Control Based on Buffer Classes, "IEEE Transactions on Communications, vol.COM-29 No.4, Apr.1981 pp.436-443.
12. Hassan Taheri, "An Adaptive Fault Tolerant Routing Algorithm for a Packet Switching System, "ph.D. thesis 1988.

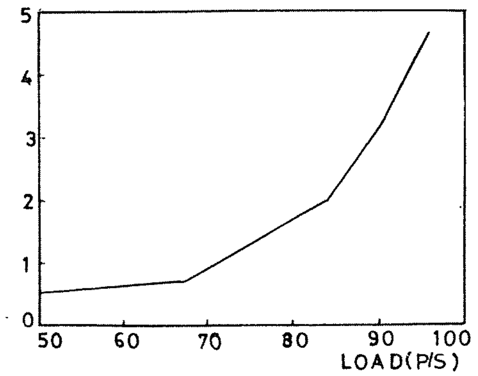


Fig. 11 Ring Delay Performance

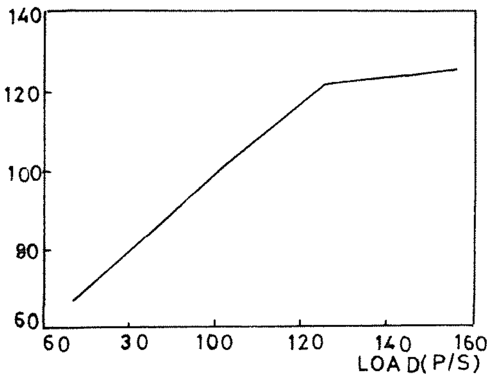


Fig. 12 Star Throughout Performance

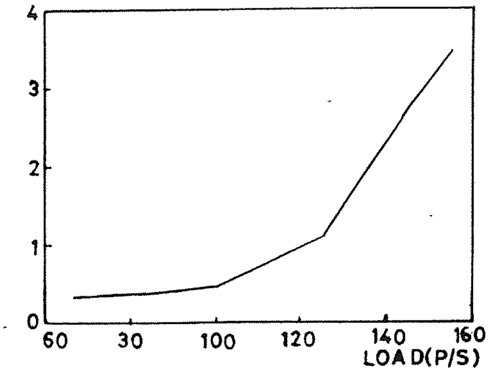


Fig. 13 Star Delay Performance

IX: CONCLUSIONS

The routing algorithm presented which is distributed, adaptive and easily reconfigurable, is capable of recovering from link failures without any user intervention.

This means the network can continue to work (albeit with reduced performance) in spite of occurrences of link failures as long as the network is not disjoint. In the case of the network becomes disjoint however, each part can still continue to work internally with only the packets destined to other part or parts discarded. This is a direct result of performing the routing algorithm in a totally distributed manner.

Insertion of a new line, or a recovered faulty line, is performed through operator who initiates generation of supervisory packets announcing cost value of the new line.

Extensive simulation runs indicate that the algorithm is deadlock free. Adaptivity, which acts upon the values of line costs, has shown to have severe effects on improving the network performance. This feature, which attempts to distribute traffic uniformly throughout the network, is specially useful after a change in topology (e.g. line failure).

Although the simulation results and the figures obtained indicate the performance of the routing algorithm under different network topologies, these performance figures depend largely upon the link speeds and also the transmission time from one sub unit to the other inside a node. This transmission time was 1 millisecond in one case and 15 milliseconds in all other cases. However, the precise value depends on the actual

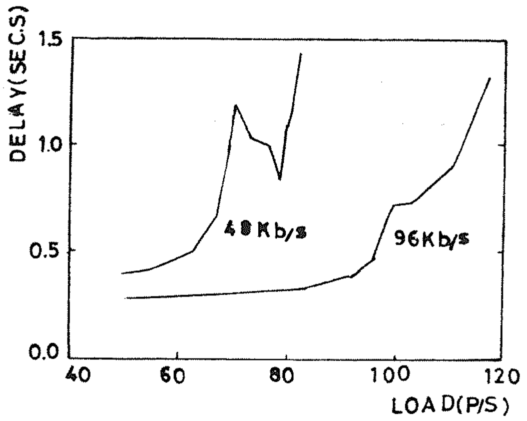


Fig. 7 NPL Delay Performance

The curves in figure 6 and 7 have been repeated for a line speed of 96 kbits per second. Comparison of the two sets of curves show that doubling the line speed has the effect of improving the performance by nearly 30% . This is because the delay from one sub unit to the other inside a node (15 milliseconds) does not change. In the case of 96 kbits per second the cost incrementing for the line connecting nodes 2 and 3 occurs at the load equal to 100 packets per second.

Figures 8 and 9 are the same as 6 and 7 except that the time required for transmission of a packet from one sub unit to the other inside a node is considered 1 millisecond instead of 15 milliseconds. Comparison of the two sets of figures shows an improvement by a factor of nearly 2. In both figures (8 and 9) dashed line shows the performance in the case of fixed line costs. Comparison of the solid and dashed lines in figures 8 and 9 shows the effect of adaptivity on network performance. Figures 10 to 13 show the same performance figures as 6 and 7 but for 10 node ring and star configurations respectively. Reference (12) provides more detailed information on simulation results.

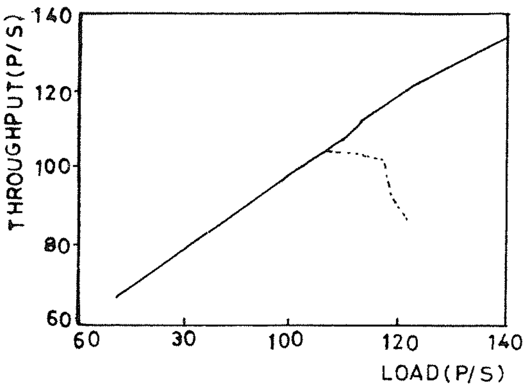


Fig. 8 NPL Throughout Performance

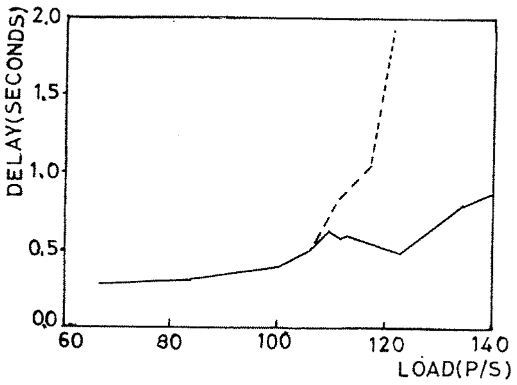


Fig. 9 NPL Delay Performance

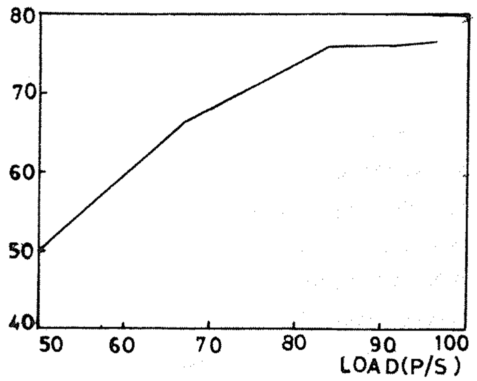


Fig. 10 Ring Throughout Performance

generate a test packet to send it to its neighbour sub unit at the other end of the line. If the line is O.K., the test packet will shortly be acknowledged (as the line is not busy). This acknowledgement packet will clear the timer.

In case the line is faulty, the test packet will not be acknowledged and the timer will continue to increment until it reaches a higher limit. At this time, the sub unit will declare the line as faulty by producing a supervisory packet which is sent to other sub units in the network. On receiving this packet, each sub unit will recompute its next node table assuming the faulty line as open.

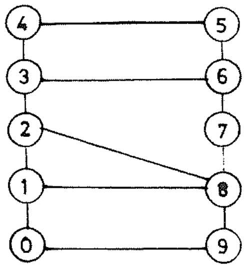


Fig. 5 NPL Configuration

VIII: SIMULATION RESULTS

Theroution and flow control procedures discussed has been tested on a number of simulated networks.

Performance figures obtained thruogh simulation experiments on a number of network configurations will be discussed.

The 10 node configuration of figure 5 was first used by the National Physical Laboratory (U.K.) and later used by the Institute for Teleprocessing of GMD (W.Germany) for some network simulation experiments. A number of simulation runs have been performed for this network configuration during which a cost value of 1 was considered for each link.

Figures 6 and 7 show respectively variations of network effective throughout and total average delay versus the load presented to this network configuration. The delay of a packet is measured as the time between it entering the end user sub unit of its source node and the sub unit receiving the packet's acknowledgement.

At loads above 70 packets per second (for line speed equal to 48 kbits/sec), there will be a cost increment for the line connecting nodes 2 and 3. Figures 6 and 7 show that increasing load has the effect of increasing packet delay; this is mainly due to accumulation of packets queuing to use heavily loaded lines. Increase in packet delay means reduction in network performance.

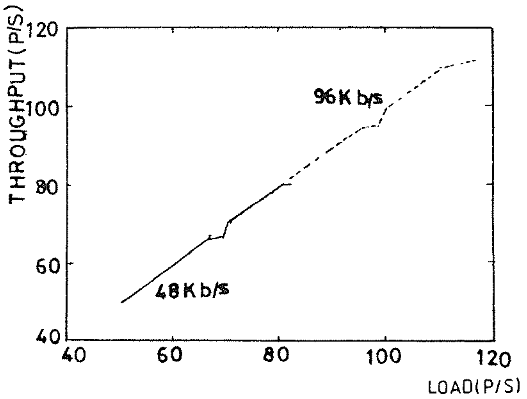


Fig. 6 NPL Throughout Performance

On the other hand if load continues to increase, at some stage, the network will react by increasing the cost of a heavily loaded line (or lines) which in turn will improve the performance.

In this technique of buffer allocation, occupancy level n indicaties the maximum number of buffer pages which can be available to packets of priority level n. This means that if the buffer is nearly empty, it can accommodate a relatively large amount of bursty traffic of mostly one priority level.

VI: PRIORITY DEADLOCK

Although the Structured Buffer pool method of flow control is designed to provide a deadlock free performance in packet routing (5), there is still the possibility of (at least) one type of deadlock. The following example shows how this type of deadlock which we call Priority Deadlock may occur.

Fig. 4 shows three nodes of some larger network. Assume buffer memories of sub units in the nodes 1,2, and 3 are all occupied above buffer level 2 (Fig. 3) mostly with packets of priority level 2 or greater.

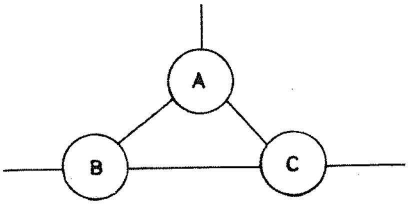


Fig. 4 Priority Deadlock

Therefore all sub units in figure 4 have masked receipt of packets of priority level 2 and greater. Also assume that all packet destinations are either of the nodes 1,2 or 3 and there is no alternative route through other parts of the network. As there is not any packet of priority level or 1 to pass through, the buffer levels can not come down. So the mask conditions persists and results in deadlock.

With this problem, a sub unit in trying to transmit a packet, finds that none of the three next node alternatives is successful in determining a receiving sub unit. To overcome, the sub unit will count the number of packets in the buffer memory with smaller priority level (higher priorities) than the current packet. If this number is smaller than the maximum gap between buffer occupancy levels, there exists some possibility of priority deadlock.

Therefore the priority level of the current packet is decremented by one to help the packet leave the sub unit.

This process will be repeated for similar packets until the occupancy level of the buffer memory is reduced sufficiently to change the mask condition. Therefore the sub unit will be able to accept more packets from the neighbour sub units so relieving the deadlock situation.

VII: LINE FAILURE DETECTION AND RECOVERY

Line failure detection is performed on the assumption that data link control layer detects and discards any packet which has been disrupted by the line. Therefore, a faulty line will not deliver any packet to either of its connected nodes.

A watchdog timer is provided in each transit sub unit (which is connected to a line). This timer is incremented each second and is cleared when a packet is delivered by the line to the sub unit. If the timer is not cleared and so reaches to some fixed limit, the sub unit will packet with priority level 4.

On the other hand, if the buffer occupancy comes down to level 5 (lowest level) the sub unit will remove limitation on receipt of priority level 4 packets.

the margin equal to the differece between level 4 and level 5 introduces a hysteresis effect and has the advantage of reducing the number of times the mask condition in any sub unit changes which in turn reduces the number of supervisory packets generated to announce them. In general if buffer occupancy rises to level n, the sub unit will limit receipt of any packet with priority level n (if already allowed). On the other hand if buffer occupancy falls to level n+1(level n+1<level n) the sub unit will remove limitation on receipt of priority level n packets (if any).

The only exception is on occupancy rising to buffer level 0. At this time receipt of priority level 0 will be limited. But supervisory packets can still be received. Occupancy levels must be chosen such that buffer memories may not become full under any load condition. Maximum allowable values of occupancy levels to satisfy this condition can be determined either in advance by simulation techniques or in actual network operation using statistical information.

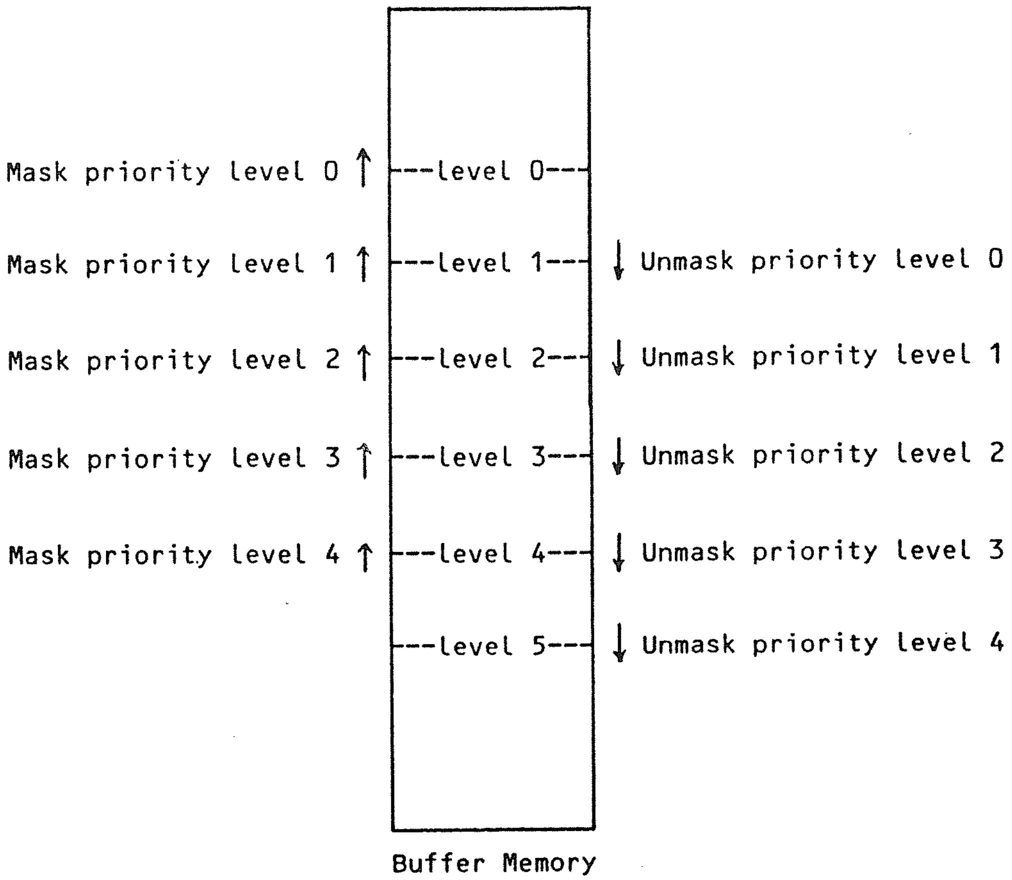


Fig. 3 Occupancy Levels in Buffer Memory

packet routing consists of a series of transmission between two neighbour sub unit.

The buffer memory in each sub unit is divided into a number of pages each of which can accommodate one packet of any type. When a packet arrives to a sub unit, it will occupy the first available page in the buffer memory. The buffer address will go at the end of one of the five priority queues according to the packets priority level.

Each sub unit on forwarding a packet towards its destination uses the first alternative in the next node table to find the appropriate neighbour sub unit. If the mask table shows that this sub unit cannot accept the packet, the second alternative in the next-node table will be tried. This cycle may be repeated once more for the third alternative in the next node table. If none of the three next nodes can satisfy the mask table, the packet address will go at the end of the same priority queue so that it can be tried later.

IV: ADAPTIVITY

Adaptivity to traffic fluctuations is performed by changing the line cost. A cost is assigned to each line the value of which depends on the line speed and the network topology. A cost value may change during the network operation in an attempt to redistribute traffic away from heavily loaded lines. This is performed by monitoring the lowest priority (priority level 4) packets which are delivered to each line. That is if packets of priority level 4 are available in a sub unit for delivering to a line but no such packets have been delivered for some fixed time interval, the sub unit will increment the cost of the line to make it less attractive. The sub unit will also announce this cost increment to the whole network using the flooding technique (9) so that every sub unit in the network can recompute its next-node table. This procedure is most useful after a change in network topology such as occurrence of a line failure.

V: FLOW CONTROL

Flow control is a version of the Structured Buffer Pool (SBP) method (11). Using this method, buffer memory in each sub unit is allocated to different packets according to their priority levels.

Five priority levels have been considered for different types of packets. On entering the network, data packets are assigned priority level 4 (lowest priority) while priority level 1 is assigned to acknowledgement packets. Supervisory packets will always have priority level 0 (highest priority).

A cost is assigned to any data or acknowledgement packet. The cost value of a packet at any time is equal to the total cost of the lines that the packet has traversed. As a packet's cost becomes larger, it will acquire higher priorities (lower priority levels). That is data packets may acquire priority levels 3 or 2 and acknowledgement packets may acquire priority level 0. Data packets will never get priority level 1 or 0 unless in a priority deadlock sitaution described later.

Figure 3 shows the buffer levels provided to allocate buffer utilization to packets of different priority levels. When the buffer is empty, the sub unit is free to receive packets of any priority level. When the buffer occupancy reaches level 4, the sub unit will limit receipt of any buffer memory. By ensuring enough buffers are reserved for highest priority packets a deadlock free performance can be achieved.

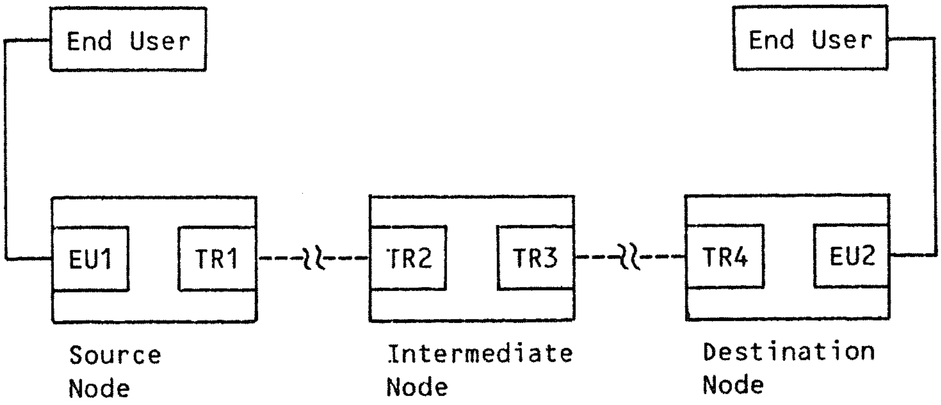
III. THE ROUTING ALGORITHM

At each sub unit two tables provide the necessary information required to perform the routing and flow control procedures. The first one is called the next node table and contains, for any destination, three directly connected nodes (next-nodes) which in turn lead to the three least cost paths to that destination node. This table is the same for all sub units inside a node and will be computed by the shortest path algorithm (12).

The second table, called mask table, provides flow control information and indicates priority levels of packets which may be sent to any neighbour sub unit. For the sub unit j inside node, i, by neighbour sub units we mean all other sub units inside node i. If the sub unit i is a transit sub unit which means it has been provided for connection to a line, the sub unit connected at the other end of the line is also considered a neighbour sub unit. In other words if two sub units can communicate directly to each other, either through the internal bus or through the external line, they are considered as neighbours.

Figure 2 illustrates how these tables are used for packet routing.

As the figure shows, each data packet is delivered by the source end user to the end user sub unit of the source node. This sub unit passes the packet through the internal bus (not shown in the figure) to the neighbour sub unit which provides the line connection to the next node. The later sub unit in turn passes the packet through the external line to its neighbour sub unit in the next node. This process continues until the packet reaches to the end user sub unit of the destination node and then to the destination end user. In summary, a



EU = End User sub unit. TR = Transit sub unit.

Fig. 2 A Packet Rout From Source to Destination

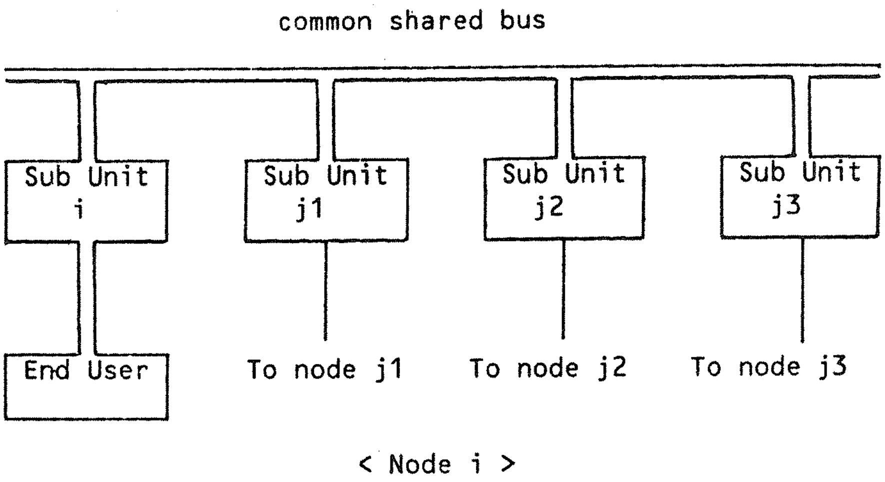


Fig. 1 Simplified Diagram of the Packet Switching Node

independently. Another disadvantage is the need for inter nodal communications which would introduce some delay.

II. DESIGN DECISIONS

The structure of the packet switching system necessitates a distributed (4) implementation of the routing algorithm. The major advantage of a distributed network is that it is less vulnerable to faults than centrally controlled networks.

To achieve efficiency and flexibility; the datagram technique has been used in contrast to the virtual circuit technique (1). The datagram technique relies on multiple route and is more flexible, more reliable and more efficient than virtual circuit technique. In the datagram technique, intermediate nodes do not have to keep track of the virtual circuit status, they only switch self contained packets towards their destination. In addition, error recovery and adaptive routing are much simplified.

Routing decisions are made on the basis of a shortest path algorithm (3). This algorithm routes packets from surce to destination over a path of least cost.

To reduce the communication overhead, destination - to - source acknowledgements are used in contrast to node-to-node acknowledgements.

To increase availibility of network resources to supervisory and control packets, five levels of packet priority have been assumed. Level O is assigned the highest priority pasckets and level (4) the lowest. Priority levels O and 1 are reserved for acknowledgements and control packets while levels 2 to 4 are used for data.

The flow control technique used is a modified version of the Structured Buffer Pool (SBP) method (8).

This method, which was developed by the GMD group in Germany for implementation in GMDNET, allows precise monitoring of each sub unit's

A Deadlock Free Routing Algorithm For A Mudular Type Packet Switching System

Hassan Taheri PhD.

Assistant Professor of the

Department of Electrical Engineering
Amirkabir University of Technology

ABSTRACT

This paper presents a routing and flow control strategy suitable for implementation on a modular type packet switching system. The routing algorithm, which is based on datagram techinque finds the shortest path between any source-destination pair as well as alternatives to take account of faulty or congested areas which are themselves discovered by the algorithm. Flow control which is based on buffer class schemes monitors the buffer occupancy in each module and controls traffic accordingly.

A quantitative performance evaluation of the algorithm found through simulation techniques has been presented.

1.THE SWITCH

This section provides the minimum system requirements of the packet switching node for which the routing and flow control procedures are developed.

Figure 1 shows the simplified diagram of such a switching node. The switch consists of an internal bus and a number of modules which we call sub units. Each sub unit is either connected to a link if it is a transit sub unit or to an end user if it is an end user sub-unit. Only one end user sub unit is assumed in each node. This results in reduced communication and processing overhead because there would be no need for inclusion of the source and destination sub units in the packet headers. The number of transit sub units in each node are equal to the number of links connected to that node which can be as many as n-1, where n is the number of nodes in the network.

An end user sub unit is responsible for communication with the end user as well as other sub units inside the node through the internal bus. A transit sub unit is responsible for communications with the neighbour node through the connected line as well as other sub units inside the node through the internal bus.

This type of structure as a packet switch has the advantages of simplicity, modularity and expandability. The main disadvantage is that each sub unit inside a node must maintain a database describing the complete network topology and must pereform all routing calculations