An Incremental Network Topology for Contention-free and Deadlock-free Routing

Pangfeng Liu
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan.
pangfeng@csie.ntu.edu.tw

Yi-Fang Lin Jan-Jan Wu
Institute of Information Science
Academia Sinica, Taipei, Taiwan.
wuj@iis.sinica.edu.tw

Abstract

Wormhole switching has become the most widely used switching technique for multicomputers. However, the main drawback of wormhole switching is that blocked messages remain in the network, prohibiting other messages from using the occupied links and buffers. To address the deadlock problem without compromising communication latency and the incremental expansion capability that irregular networks can offer, we propose a simple topology called *Extended Incremental Triangular Mesh* (EITM) for switch-based networks. EITM is an extension of a previous ITM (Incremental Triangular Mesh) topology with a more flexible structure. We also show that EITM is highly scalable, allows incremental expansion of systems, has guaranteed deadlock freedom, and can support contention-free multicast. First, we show that for an EITM, *any* shortest path routing method will not deadlock, therefore EITM networks are ideal for the escape paths in adaptive routing networks. Second, we show that it is possible to arrange the nodes of an EITM in a circular order so that two messages from independent parts of the circular order will not interfere with each other – this is extremely useful for implementing contention-free multicast and other collective communication operations. We also present the results on the relation between ITM/EITM, outer planar graphs and chordal graphs. We show that chordal graphs are strongly related to the freedom of deadlock for shortest path routing, and ITM in our previous paper is indeed maximum outer planar graph.

1 Introduction

Wormhole switching [4, 15] has become the most widely used switching technique for multicomputers. The availability of high-speed wormhole switches, such as Autonet [7], Myrinet [1], and Servernet [12], has also made network of workstations a promising alternative for cost-effective parallel computing. In earlier stored-and-forward routing method an entire message has to be stored in one node before it could be sent to the next. In contrast, wormhole routing uses a cut-through approach that divides the message into small flits that travel through the network in a pipelined fashion, therefore it eliminates the need to allocate large buffers in the intermediate nodes along the path [15]. This not only simplifies the switch design but also provides a distance insensitive routing methodology for sufficiently

large messages.

The main drawback of wormhole switching is that blocked messages remain in the network, prohibiting other messages from using the occupied links and buffers, therefore wasting channel bandwidth. We further classify this problem into two categories. First, a poorly designed routing algorithm might cause a *deadlock* on a wormhole routing network, in which messages are tangled together and no message can proceed. Secondly, for a particular communication pattern (e.g. multicast), a large number of messages may go through a common channel and cause significant delay. Although no deadlock occurs, the communication performance is degraded due to this *contention* problem.

Deadlock-free routing and contention minimization have been extensively studied for proprietary networks, in which the processing nodes are usually interconnected into a regular topology, such as mesh, torus or hypercubes [4, 5, 6, 9, 8, 10, 16]. On the other hand, switch-based interconnects have been a popular choice for building networks of workstations and PCs. Typically, these switches support networks with irregular topologies. Such irregularity allows easy design and wiring of scalable systems with incremental expansion capability (allows the addition of one or more switches at a time). However, the irregularity also makes routing and deadlock avoidance on such systems very complicated. Several deadlock-free routing algorithms have been proposed in the literature for irregular networks [1, 7, 12, 17]. These algorithms avoid deadlock by *restricting* routing to remove cyclic dependencies between channels. As a consequence, some messages may be routed through non-minimal paths, resulting in increased latency.

To address the deadlock-free routing problem without compromising communication latency and the incremental expansion capability that irregular networks can offer, in a previous paper [14] we proposed a simple topology called *Incremental Triangular Mesh* (ITM) for switch-based networks. ITM is highly scalable, allows incremental expansion of systems, has guaranteed deadlock freedom, and can support contention-free multicast [14]. This paper extends the idea of ITM into *Extended Incremental Triangular Mesh* (EITM). EITM has a more flexible topology, provides higher bandwidth, and most important of all, has all the routing properties as we showed for ITM. This paper also formally establishes the relation between ITM and outerplanar graph, and shows that ITM is maximum outerplanar graph.

For the nice routing properties of EITM, first, we show that on an EITM, *any* shortest path routing method will not deadlock. There are numerous deadlock-free routing algorithms in the literature that work in a similar fashion – messages must travel through the channels in a particular order to break the symmetry (e.g. dimensional ordering [11] or up-down routing in [7]). These approaches sacrifice certain throughput for deadlock free guarantee. In contrast we argue that in EITM a message can go through *any* shortest path without deadlock, therefore EITM can be used as dedicated virtual channels to avoid deadlock in many adaptive routing networks. Secondly, we show that it is possible to arrange the nodes of an EITM in a circular order so that two messages from independent parts of the circular order will not interfere with each other. It is shown in [18] that it is impossible to find a *linear order* for every irregular topology. Nevertheless, for every EITM we can define a *circular order* that has the contention-free property. This is extremely useful for implementing contention-free multicast and other collective communication operations.

The rest of the paper is organized as follows. Section 2 describes the related works. Section 3 describes the deadlock-free and contention-free property of EITM, and gives detailed proof, and Section 4 concludes.

2 Related Work

2.1 Deadlock-free Routing

Chien and Kim [3] describe a class of restricted adaptive routing algorithms suitable for packet-switched data transmission in multiprocessors. Planar-adaptive routing provides an effective compromise that sacrifices some routing freedom to reduce the possibility of deadlock. Restricting routing at each step to a specific hyperplane in k-ary n-cubes still leaves many alternative routes, but the restriction allows provably deadlock-free operation at a cost of only 3 virtual channels, regardless of the number of dimensions in the n-cube. The result is a much lower hardware cost for deadlock-free routers.

There are other general purpose deadlock-free routing algorithms for wormhole switches. Up-down routing [7] first constructs a breadth-first spanning tree on the switching network. A directed link is "up" if it goes from a node "upwards" towards the root, or if it goes from one node to another node in the same level, but with a higher processor id. A legal route has all the "up" links appearing *before* all the "down" links. Eulerian trail routing [17] assumes that the network topology is Eulerian, then routes the messages along this Eulerian path. Shortcut channels may be used to reduce the length of the route [17].

Tseng et. al. [20] focus on multicasting in wormhole-routed networks. A trip-based model is proposed to support adaptive, distributed, and deadlock-free multiple multicast on any network with arbitrary topology using at most two virtual channels per physical channel.

With the introduction of virtual channel, Duato et. al. [5, 19] suggested another approach for deadlock-free routing on any irregular networks. The network is split into two layers. An arbitrary routing algorithm is running on the first layer, while a deadlock-free routing algorithm is on the second layer. The key idea is to compromise between maximizing performance (on the first layer) and guaranteeing deadlock-free operation (on the second layer). If a message is blocked at the first layer, then it moves down to the second layer and stays there until it reaches its destination. The second layer network is used as escape paths to avoid deadlock.

2.2 Contention-free Routing

There are many contention-free multicast algorithms for regular switching topologies. For example, Esfahanian el. al. [16] suggested contention-free multicast on n-dimensional meshes and hypercubes, and provided good performance from implementation on n-Cube and Smalt 2-D mesh. Ho and Johnsson [11] suggested dimensional ordering algorithms for broadcast and personalized all-to-all communication on hypercubes. Other contention-free algorithms include [5, 10].

It is much more difficult to design contention-free routing algorithms for irregular network topologies. In many multicast algorithms processors are arranged as a linear list, with the property that if node a, b, c and d appear in the list in order, then the message between a and b will not interfere, or contend from any links with the message between c and d [18]. However, it is also shown in [18] that for some irregular topologies such an ordering simply does not exist.

3 Extended Incremental Triangular Mesh

This section defines EITM and describes its properties. First we define ITM and then extend the definition of ITM (by adding extra edges) into EITM. The first property of both ITM and EITM is that they guarantee freedom from deadlock for *any* shortest path routing. This property allows ITM and EITM to route messages through any shortest path without

the risk of deadlock. The second property is that we can partition an ITM or EITM so that the messages traveling in different sections will not interfere with one another. It is shown in [18] that for some irregular graphs this contention-free ordering simply does not exist. We show that both ITM and EITM, which can be very irregular, do provide this ordering.

3.1 Definition of ITM

We first define the ITM topology [14]. The concept of incremental triangular mesh is built on top of a series of *AddNode* operations. Let $G' = (V', E')$ be a undirected graph and $e' \in E$. To add a node v into G' at edge $e' = (x, y)$ means that we add v into V' and connect v to the two endpoints of e'. The edge e' is called the *corresponding edge* of v. We also assume that every added node has a *unique* corresponding edge. Formally we have the following definition: $AddNode(G', v, (x, y)) = (V' \cup v, E \cup \{(v, x), (v, y)\})$.

The *incremental triangular mesh* (ITM) is defined recursively as follows. First a clique of three nodes is an ITM. A graph G is an ITM if and only if there exists another ITM (denoted by G') of $n - 1$ nodes such that $G = AddNode(G', v, e')$, where $e' \in E'$ is the corresponding edge of the newly added node v.

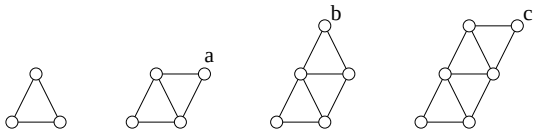

Figure 1: An incremental triangular mesh constructed by adding nodes a, b and c.

3.2 Definition of EITM

We now extend ITM to EITM for a more flexible structure and better bandwidth. After an ITM is constructed, we add additional edges into the graph. By the definition of ITM just before a node v is added to the corresponding edge (x, y), there exists a unique node z such that x, y and z form a triangle. We optionally add an additional edge (v, z), into the original graph so that v, x, y, and z form a tetrahedron. The original edges in G will be called *ITM-edges* or *i-edges*, and these added edges are *jump edges*, or *j-edges*. Formally we have the following definition: $AddNode(G', v, (x, y)) = (V' \cup v, E \cup \{(v, x), (v, y), (v, z)\})$

We define (x, y, z) to be the *corresponding triangle* for the added node v. A node can be added into an EITM by a corresponding edge (without a j-edge), or a corresponding triangle (with a j-edge). Also notice that after adding j-edges, the graph may no longer be planar. Figure 2 shows an EITM of 6 nodes.

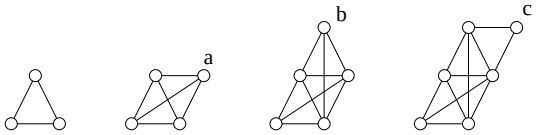

Figure 2: An EITM example.

3.3 Deadlock-free Routing

Most of the deadlock-free routing on a regular network requires certain "dimension ordering" in order to break the symmetry and guarantee freedom from deadlock. For example we may require that a message must traverse along the correct row of a mesh before traversing the column to make sure that deadlock will not happen. This restriction may limit the available bandwidth since some routes may be unnecessarily avoided just because of the possibility of a deadlock. In contrast, in an EITM network a message can choose any shortest path without risking a deadlock.

Before we show the deadlock-free properties for ITM and EITM, we establish the relation between chordal graph and the deadlock free property under shortest path routing.

Lemma 1 *A graph G is deadlock-free under every shortest-path routing method if and only if G is chordal.*

Theorem 2 *Any routing discipline that takes the shortest path is deadlock-free in an ITM or EITM.*

3.4 ITM and Outerplanar Graphs

ITM is closely related to outerplanar graph since the construction of ITM is the reverse process of the well-studied polygon decomposition problem [13, 2]. An outerplanar graph is a planar graph whose nodes are all on the boundary. An outerplanar graph is maximum if we cannot add another edge without destroying the outerplanarity. It is easy to see that ITM is an outerplanar graph. It is also true that a graph is an ITM if and only it is a maximum outerplanar graph.

Theorem 3 *[14] A graph $G = (V, E)$ has an ITM subgraph that contains all the nodes in V if and only if there exists a Hamiltonian cycle in G for which G is* totally triangulated. *A graph G is totally triangulated for a Hamiltonian cycle C if and only if when the vertices of G are around a circle according to the order they appear in C, no edge can be added without intersecting an edge of G.*

3.5 Contention-free Routing

This section describes the contention-free property of EITM. We assume that each link in the network is bi-directional and two messages are *contention-free* as long as they do not go through the *same* link in the *same* direction.

In switch-based network routing it is desirable to have an ordering among all the nodes in a network such that two messages involving processors from different sectors of this ordering do not interfere with each other. That is, suppose we can define a total order $<$ among processors such that when $w < x < y < z$, then any message-passing between w and x will not interfere with those between y and z. Using this property we can design simple contention-free recursive algorithms for broadcast and multicast, i.e. the source processor first sends the message to the processor in the *middle* of the list, and repeats the process on the two sub-lists. Despite that this property can be obtained for some regular graphs, it is shown in [18] that there exist irregular graphs where such orderings are not possible. Nevertheless, we show that for *EITM* we can define a similar order that has this nice "non-interfering" property, despite the irregularity of *EITM*.

Instead of a total order, we use a *circular order* to enumerate the nodes in an EITM $G = (V.E)$. It is easy to see that all the nodes in G form a simple cycle along the "boundary" of G, so we can define a circular order among the nodes by enumerating the nodes clockwise or counterclockwise. We then define that $w < x < y$ if and only if a node x appears between node w and y in the circular order.

Consider two messages m and m'. The message m goes from w to x, and m' goes from y to z. The two messages m and m' are *independent* if and only if $w < x < y < z$ in the circular order defined earlier. We will show that all the shortest paths of two independent messages will not share a communication link in the *same* direction. That is, a shortest path going from w to y will not share a directed edge with any shortest path going from y to z in an EITM.

3.5.1 Contention-free Multicast for EITM

We describe the contention-free property for EITM in this section. Notice that although we can define a circular order for EITM similar to ITM, we cannot use the results in [14] since EITM is not planar. As a result we establish new proof technique for the case of EITM.

Before going into the proof we establish a

key property of EITM that the introduction of a new node into an EITM G will not modify the shortest path between two nodes that were originally in G.

Lemma 4 *Let $G = (V, E)$ be an EITM and G' be the new EITM after adding a node v into G. A path P from x to y, where $x, y \in V$, is a shortest in G if and only if P is a shortest path in G'.*

This property of EITM is crucial since it allows the routing table on the nodes to be updated *incrementally*. After adding a new node all the previous shortest routes are not changed, and the system only needs to create the routing table entries for the newly added node.

Lemma 5 *Two independent messages will not travel through the same communication link in the same direction in an EITM under any shortest path routing discipline.*

Theorem 6 *Any multicast pattern can be completed within $O(\log D)$ non-interfering phases in an EITM where D is the number of destinations.*

4 Conclusions

This paper proposes a new interconnecting topology, *extended incremental triangular mesh*, for switch-based network of workstations. We have shown that EITM guarantees deadlock freedom for any shortest path routing and supports contention-free multicast.

The nice properties of EITM also make it an ideal candidate for supporting adaptive routing in many networks. Adaptive routing can be implemented by changing the routing tables and adding links in parallel with existing ones, or by splitting physical channels into virtual ones. Deadlock can be avoided either by restricting routing so that there are no cyclic dependencies between channels, or simply by providing some escape paths to avoid deadlock, without restricting routing. EITM's deadlock-free property and incremental expansion capability make it a suitable choice for building the escape paths.

We also establish the relation between ITM/EITM, chordal graphs, and maximum outer planar graphs. We show that the chordal graph criteria is both necessary and sufficient for a network topology to be deadlock free under any shortest path routing. We also show that ITM is equivalent to maximum outer planar graph.

The future work includes two directions – the investigation of general planar graph and more complex extensions of EITM. We show in a previous paper [14] that planar ITM can have contention-free multicast. That result combined with the outer planarity indicates that any outer planar graphs do support contention-free multicast. However, for general graph it is not clear how one could order the nodes so that the contention-free argument still works. On the other hand, the way we extend ITM is rather restricted – we are not able to form a tetrahedron from all the facets, but only those on the boundary. It is an interesting question how this restriction can be relaxed.

References

[1] N. J. Boden, D. Cohen, R. F. Felderman, A. E. Kulawik, C. L. Seitz, J. Seizovic, and W. Su. Myrinet - a gigabit per second local area network. *IEEE Micro*, pages 29–36, Feb. 1995.

[2] Bernard Chazelle. Triangulating a simple polygon in linear time. In *IEEE Symposium on Foundations of Computer Science*, pages 220–230, 1990.

[3] A. Chien and J. H. Kim. Planar-adaptive routing: low-cost adaptive networks for multiprocessors. *Journal of ACM*, 42(1):91–123, Jan. 1995.

[4] W. J. Dally and C. L. Seitz. Deadlock-free message routing in multiprocessor interconnection networks. *IEEE Transactions on Computers*, C-36(5):547–553, May 1987.

[5] J. Duato. On the design of deadlock-free adaptive routing algorithms for multicomputers. In *Proceedings of Parallel Architectures and Languages Europe 91*, June 1991.

[6] J. Duato. A necessary and sufficient condition for deadlock-free adaptive routing in wormhole networks. In *Proceedings of the 1994 International Conference on Parallel Proceeding*, August 1994.

[7] M. D. Schroeder et. al. Autonet: A high-speed, self-configuring local area network using point-to-point links. Technical Report SRC research report 59, DEC, April 1990.

[8] P. T. Gaughan and S. Yalamanchili. Adaptive routing protocols for hypercube interconnection networks. *IEEE Computer*, 26(5):12–23, May 1993.

[9] C.J. Glass and L.M. Ni. The turn model for adaptive routing. *J. ACM*, 41:847–902, Sept. 1994.

[10] G. Gravano, G. D. Pifarre, P. E. Berman, and J. L. C. Sanz. Adaptive deadlock- and livelock-free routing with all minimal paths in torus networks. *IEEE Trans. Parallel and Distributed Systems*, 5(12):1233–1251, Dec. 1994.

[11] C. Ho and S. Johnsson. Optimal broadcasting and personalized communcaiton in hypercudes. *IEEE Transaction on Computers*, 38:1249–1268, September 1989.

[12] R. Horst. Servernet deadlock avoidance and fractahedral topologies. In *Proceedings of the International Parallel Processing Symposium*, pages 274–280, April 1996.

[13] J. Mark Keil. Decomposing a polygon into simpler components. *SIAM J. Comput.*, 14(4):799–817, 1985.

[14] P. Liu, J. Wu, Y. Lin, and S. Yeh. A simple incremental network topology for wormhole switch-based networks. In *Proceedings of International Parallel and Distributed Processing Symposium*, May 2001.

[15] L.M. Ni and P.K. McKinley. A survey of wormhole routing techniques in direct networks. *IEEE Computer*, 26(2):62–76, February 1993.

[16] A.-H. Esfahanian P.K. McKinley, H. Xu and L.M. Ni. Unicast-based multicast communication in wormhole-routed networks. *IEEE Transactions on Parallel and Distributed Systems*, 5(12):1252–1265, December 1994.

[17] W. Qiao and L.M. Ni. Adaptive routing in irregular networks using cut-through switches. In *Proceedings of the 1996 International Conference on Parallel Proceeding*, pages I:52–60, August 1996.

[18] K. Bondalapati R. Kesavan and D.K. Panda. Multicast on irregular switch-based networks with wormhole routing. In *International Symposium on High Performance Computer Architecture*, Feb. 1997.

[19] F. Silla, M.P. Malumbres, A. Robles, P. Lopez, and J. Duato. Efficient adaptive routing in networks of workstations with irregular topology. In *Workshop on Communication and Architectural Support for Network-based Parallel Computing*, Feb. 1997.

[20] Y.-C. Tseng, D. K. Panda, and T.-H. Lai. A trip-based multicasting model in wormhole-routed networks with virtual channels. *IEEE Trans. Parallel and Distributed Systems*, 7(2), Feb. 1996.