Arctic Switch Fabric*

G. Andrew Boughton

Massachusetts Institute of Technology, Laboratory for Computer Science, Cambridge Mass 02139, USA

Abstract. The Arctic Switch Fabric technology is a scalable network technology based on the Arctic router chip. Switch Fabrics are fat tree networks that are capable of providing high performance even under a heavy load of large (96 byte) packets. They have a number of diagnostic features that make them well suited for experimental computer systems. Switch Fabrics will be used in the StarT-Jr, StarT-Voyager, and Xolas computer systems at MIT. This paper describes the overall characteristics of Switch Fabrics and describes the three layers of Fabric components. These are the 16-leaf network, the four-leaf board, and the Arctic router chip.

1 Introduction

The Arctic Switch Fabric technology is a scalable network technology based on the Arctic router chip [2]. An Arctic Switch Fabric is a fat tree network which supports 150 MB/sec bandwidth in each direction at each endpoint and has a bisection bandwidth equal to 150 MB/sec times the number of endpoints divided by two. A Switch Fabric is currently being tested on the PC-based StarT-Jr multiprocessor. In the next six months Switch Fabrics will be used in a number of other multiprocessor systems at MIT including the nine node (72 processor) Xolas and the 32 processor StarT-Voyager.

Switch Fabrics provide high performance and reliability for large systems. The fat tree structure allows traffic to be evenly distributed. The buffering and crossbar scheme of the Arctic chip allows high utilization of network links. A Switch Fabric is capable of accepting high bandwidth packet flows from all of its inputs simultaneously, even when a large fraction of the packets are big (96 bytes). A differential signaling scheme is used on the interchassis cables to provide reliable data transmission. Each network link has extensive error detection circuitry including a 16-bit CRC circuit which verifies the accurate transmission of each packet and Manchester code checking circuits which verify the accurate transmission of each control signal.

The Switch Fabric technology includes powerful diagnostic features. A Switch Fabric has a dedicated Ethernet-based maintenance system which accesses each router through its JTAG port and is thus independent of the status of the network link cables. The maintenance system has access to six 36-bit counters in

* The research described in this paper was supported in part by the Advanced Research Projects Agency under Office of Naval Research contract N00014-92-J-1310 and Ft. Huachuca contract DABT63-95-C-0150.

each router output port which are used to monitor traffic through the port. It can disable and enable links, flush packets, inject packets, and reconfigure the routing algorithm in each router. It can also run verification tests on each router.

2 16-Leaf Fat Tree Network

The basic building block of the Fabric technology is a 16-leaf fat tree network. As shown in Figure 1, the network provides 16 leaf ports and 16 root ports. The leaf ports can be attached by cables to either endpoints or other (logically lower) networks. The root ports can be attached by cables to other (logically higher) networks. Thus arbitrarily large systems can be constructed.

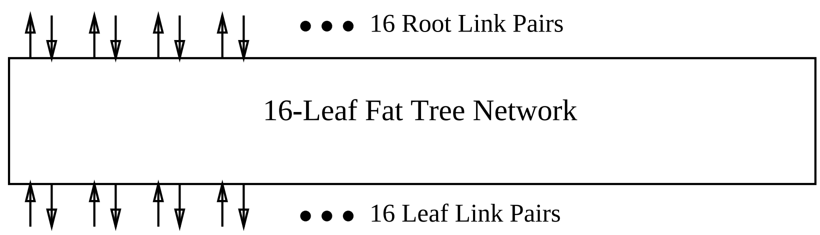

Fig. 1. 16-Leaf Fat Tree Network

The 16-leaf fat tree network is constructed from eight four-leaf boards as shown in Figure 2. Each four-leaf board provides four single ended GTL (low voltage swing CMOS) ports and four differential ECL ports. The GTL ports are used for the internal links of the network and the ECL ports are used for the external links.

Two 16-leaf fat tree networks fit in a single card cage as shown in Figure 3. The internal links of each network are implemented using backplane traces. A 32-leaf network (with no root ports) can be formed by cabling together the root ports of the two 16-leaf networks. Alternatively each 16-leaf network can be cabled to other 16-leaf networks in other cages as discussed above. Arbitrarily large fat tree networks can be constructed in this fashion.

Each card cage has an associated controller which is a small PC running the Linux operating system. The controller provides access to the maintenance systems of the cage's two networks.

3 Four-Leaf Board

Each board is a four-leaf fat tree network constructed from four Arctic router chips as shown in Figure 4. The board has four GTL ports attached to backplane

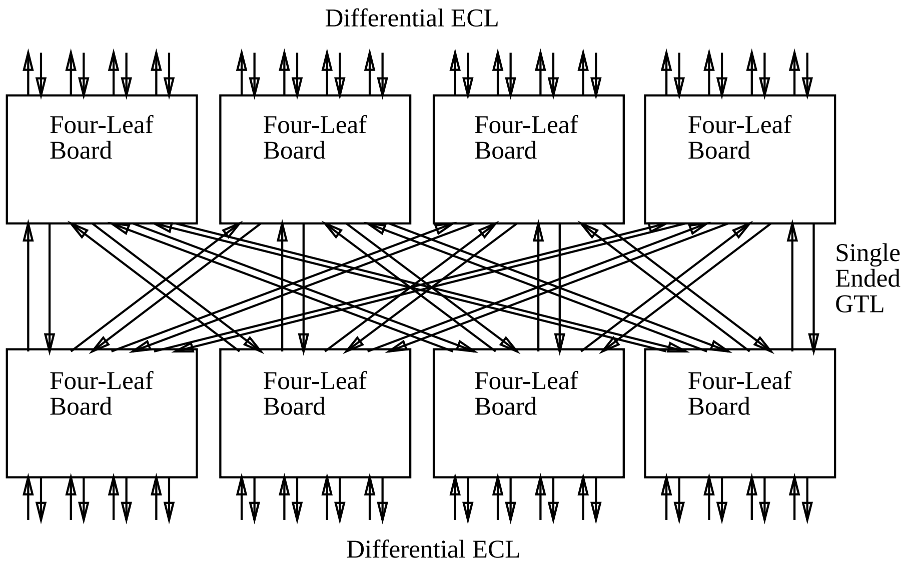

Fig. 2. Structure of the 16-Leaf Network

connectors and four ECL ports attached to cable connectors. Each port has both incoming and outgoing data lines.

Each differential ECL port is connected to a single ended GTL Arctic router port through an interface circuit. In addition to converting between logic families, the circuit registers both data coming out of Arctic and data going into Arctic. Arctic provides outgoing data at 77 Mbits/sec/line, and provides two complimentary 38.5 MHz clocks with the data. The interface circuit uses a PLL to generate a 77 MHz clock from these clocks. It uses the 77 MHz clock to register the data and outputs the clock on the ECL port with the registered data. Incoming data on the ECL port has an associated 77 MHz clock which the interface circuit uses to register the data and to generate, with a simple divide by two circuit, the appropriate 38.5 MHz input clocks for Arctic. The interface circuit sends these clocks with the registered data into Arctic.

The four-leaf board has a single maintenance port which is attached to a backplane connector. This port provides access to a scan chain which includes the JTAG ports of the four Arctics on the board.

4 Arctic Router Chip

Arctic [2] is a four input four output packet router implemented on a Motorola H4CP178 gate array chip in a 324 CBGA package. Arctic uses approximately 105,000 of the 178,000 available gates; roughly 44,000 are used for memories. Arctic supports two priority levels, packet sizes from 16 to 96 bytes, and exten-

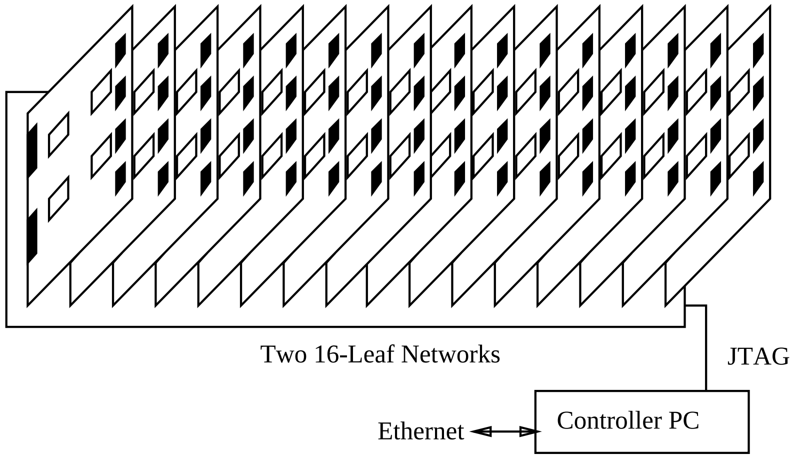

Fig. 3. Card Cage

sive error detection; it has limited error handling, keeps statistics, and provides significant testing support. The overall structure of Arctic is shown in Figure 5.

Arctic uses a sophisticated buffer management scheme that greatly increases the effectiveness of its buffers and the utilization of network links. This scheme is similar to that developed by Joerg for the PaRC routing chip [5, 6] of the Monsoon multiprocessor. All buffering in Arctic is located in the input sections. Each input section contains three buffers; each capable of storing a maximum size packet. Each buffer is capable of storing either a priority packet or a normal packet. The last empty buffer is reserved for a priority packet; this ensures that normal packets can not block priority packets. The head of a packet can be transferred out of a buffer before the end of the packet has been received (virtual cut-through with partial cut-throughs allowed). (The head of an unblocked packet will leave Arctic roughly 150 ns after it has entered Arctic.) The crossbar has 12 inputs and 4 outputs. Each of the packet buffers in the chip is directly connected to the crossbar. The scheduling scheme used in Arctic ensures that if any buffer in the chip contains a packet destined for a given output then a packet will be transferred to that output as soon as the output becomes available. Any four of the 12 input buffers can be simultaneously transferring to output sections. Packets received on a given input and destined for a particular output are transferred in the same order as they were received, thus in order delivery of packets is maintained along a given network path. Packets received on different inputs and destined for a particular output are scheduled using a round-robin scheme which guarantees that no packet is starved for service. All priority packets waiting for a particular output are transferred before any waiting normal packet is transferred to that output.

Each Arctic port has an incoming 16-bit data path and an outgoing 16-bit data path. Two clock lines, a frame line, and a (reverse direction) flow control line

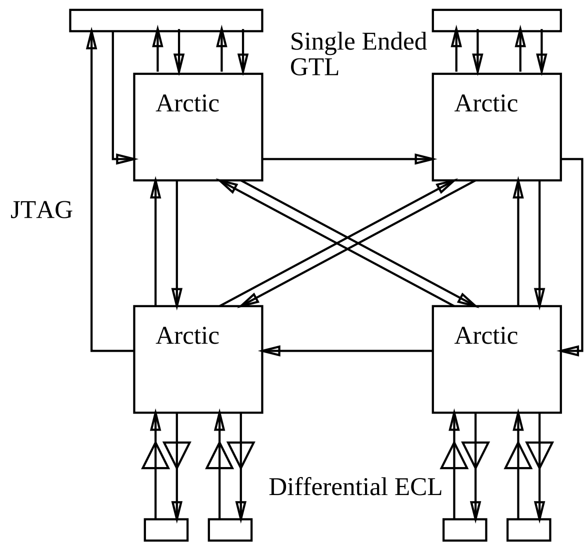

Fig. 4. Four-Leaf Board

are associated with each data path. Each data bit is transmitted using a single ended GTL driver at 77 Mbits/sec. This technology supports printed circuit board links that are one meter long. A 16-bit CRC is sent with each packet. The frame and flow control signals are Manchester encoded. Link errors are detected using a CRC check circuit for data and a Manchester check circuit for the frame and flow control signals. Since the two Switch Fabric link technologies (interchassis ECL and intrachassis GTL) have been designed for high reliability, Arctic does not provide link level packet retry. In the unlikely event of a CRC error, Arctic will record the error and will forward the packet, marked with a bad CRC, toward a Fabric endpoint.

Arctic uses source based routing; the course of a packet through a Switch Fabric is completely determined by routing information placed in the packet header by the source. Arctic can be configured to implement any of a wide variety of source based routing schemes. Each Arctic input port is configured with two masks, two reference values, and three bit extraction operations. For each incoming packet, Arctic uses the two masks to select two subfields of the packet's header and compares these subfields to the reference values. Based on the results of the comparison, Arctic selects an extraction operation and applies it to the header. The result of this operation determines the output port for the packet. Most Switch Fabrics will be configured for fat tree routing. For fat tree routing each packet header is divided into two fields. The source places a random bit string in the first and the destination address in the second. As the packet travels from the leaves to the root, each Arctic checks the appropriate prefix of the destination address to determine if the packet must travel farther from the

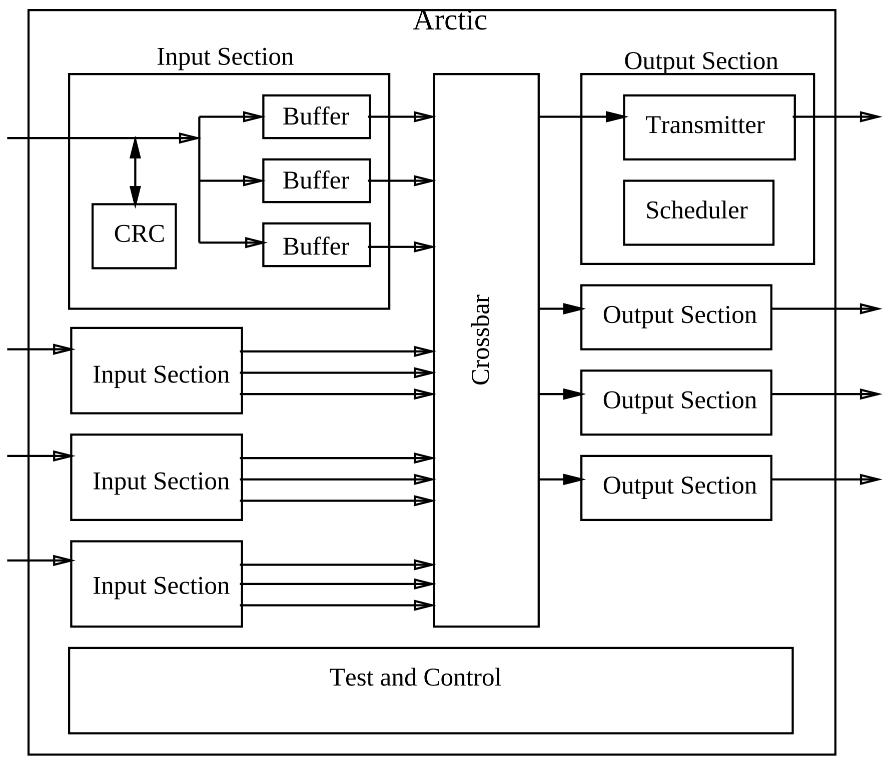

Fig. 5. Arctic Block Diagram

leaves. If this is the case then, since there are multiple acceptable paths away from the leaves, the Arctic uses a portion of the random bit string (the first header field) to choose an appropriate output port (away from the leaves). Otherwise the Arctic uses a portion of the destination address (the second header field) to determine the appropriate output port toward the leaves. The use of a random bit string generated at the source provides a more globally random distribution of traffic than would be produced if each Arctic dynamically routed packets based on local flow control information. Also since this string is generated at the source, a variety of modifications can easily be made to this fat tree routing algorithm.

Arctic has an extensive maintenance system which is accessed through its JTAG port. The maintenance system provides configuration data to Arctic including the routing algorithm for each input port. The maintenance system has real time control of Arctic. It can disable and enable each port, and can flush packets. It has access to the five error counters in each input port which record link errors, and to the six 36-bit counters in each output port which monitor packet traffic.

The maintenance system is capable of performing several off line test operations. It has access to Arctic's boundary scan ring. The boundary scan ring supports EXTEST operations which can be used to test the continuity of printed circuit board traces. The boundary scan ring also supports INTEST operations which can be used to run off line verification tests on Arctic through its I/O cells. The maintenance system has off line access to Arctic's manufacturing test rings which contain nearly all of Arctic's state elements. Thus the maintenance system allows ATPG generated manufacturing tests to be run on Arctic without a manufacturing tester. These tests verify the integrity of almost all of Arctic's internal circuitry.

Arctic's maintenance system allows test packets to be injected into the Switch Fabric. Packets are injected by taking Arctic off line, using the boundary scan ring to inject the packets into Arctic, and then placing Arctic back on line with the injected packets.

A more detailed description of the Arctic router chip can be found in the PCRCW 94 paper [2].

5 Status

A prototype Fabric circuit board has been constructed and successfully tested with StarT-Jr nodes. As shown in Figure 6, this prototype board includes one Arctic chip and one interface circuit. It has one ECL port cable connector, three GTL port cable connectors, and a maintenance port connector. Four copies of this board are being used in a four endpoint Switch Fabric. ECL links are being used to connect the prototype Fabric boards to StarT-Jr NIUs and GTL links are being used to interconnect the prototype Fabric boards.

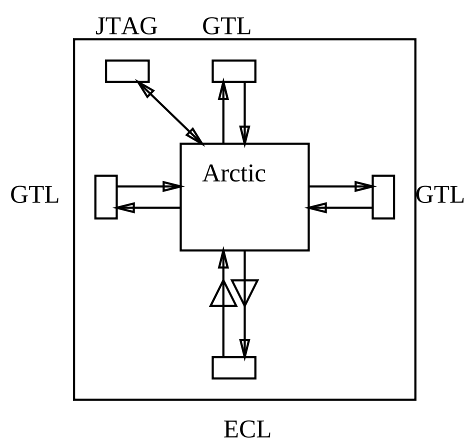

Fig. 6. Prototype Fabric Board

The four-leaf Fabric board is a simple extension of this prototype board. The design of the four-leaf board is currently being completed and fabricated four-leaf boards are expected in September. A Fabric card cage should be available in October.

Related work on network interfaces is being done by other members of our group. The StarT-Jr network interface [4] has been available for over a year. It is PCI-bus based and includes an embedded I960 microprocessor. The StarT-X network interface [3] is also PCI based but has a more direct hardware implementation and provides higher performance. While StarT-X has a more limited functionality than StarT-Jr, it provides the most important message passing functions including DMA. StarT-X boards will be available in August. The StarT-Voyager Network Endpoint System (NES) [1] is a much higher performance memory-bus based network interface. The NES contains an embedded PowerPC 604 service processor. The NES provides direct hardware support for four different types of messages and programmable support for cache coherent distributed shared memory. NES boards are expected in October.

6 Acknowledgments

I want to thank everyone who has contributed to the Arctic Switch Fabric project. Greg Papadopoulos, Bob Greiner, and Chris Joerg inspired many of the concepts used in the Arctic router chip. Many people have contributed to the detailed design of the Arctic router chip, the Fabric boards, and the support software. These include Jack Costanza, Doug Faust, Tom Durgavich, Ralph Tiberio, Richard Davis, Elth Ogston, Michael Sy, and Chen Chi Ming Hubert. Finally, I want to give special thanks to Arvind for providing his support and managerial guidance to the project.

References

1. Boon S. Ang, Derek Chiou, Larry Rudolph, and Arvind. Message Passing Support on StarT-Voyager. CSG Memo 387, Computation Structures Group, LCS, MIT, July 1996.
2. G. Andrew Boughton. Arctic Routing Chip. In *Proceedings of the 1994 University of Washington Parallel Computer Routing and Communication Workshop*, May 1994. Also CSG Memo 373, Computation Structures Group, LCS, MIT.
3. James Hoe. Functional Specification for a High-Performance Network Interface Unit on a Peripheral Bus. CSG Memo 389, Computation Structures Group, LCS, MIT, June 1997.
4. James C. Hoe and Mike Ehrlich. StarT-JR: A Parallel System from Commodity Technology. In *Proceedings of the 7th Transputer/Occam International Conference*, November 1996. Also CSG Memo 384, Computation Structures Group, LCS, MIT.
5. C. F. Joerg. Design and Implementation of a Packet Switched Routing Chip. TR 482, Laboratory for Computer Science, MIT, 1990.
6. C. F. Joerg and G. A. Boughton. The Monsoon Interconnection Network. In *Proceedings of the 1991 IEEE International Conference on Computer Design*, October 1991. Also CSG Memo 340, Computation Structures Group, LCS, MIT.