M. Jonsson, "Comments on interconnection networks for parallel radar signal processing systems," Research Report CCA - 9911, Centre for Computer Systems Architecture (CCA), Halmstad University, Sweden, May 1999.

Research Report CCA - 9911, Centre for Computer System Architecture (CCA), Halmstad University, Sweden, May 1999.

Comments on Interconnection Networks for Parallel Radar Signal Processing Systems

Magnus Jonsson

Centre for Computer Systems Architecture, Halmstad University, Box 823, S-301 18 Halmstad, Sweden. Email: magnus.jonsson@cca.hh.se. Phone: +46 35 16 71 00, Fax: +46 35 12 03 48

1. Introduction

In this report, some high-performance interconnection networks are briefly commented/evaluated against a specific radar signal processing system selected as a general representative system (see Figure 1). The computations in the shown signal processing chain are pipelined in three stages. All stages are therefore always kept busy and the transmissions of the dataflows between the stages are assumed to be spread over time to avoid burst traffic. Further on, we assume two PE-arrays are contained in one computational module sharing one network interface. If an odd number of PE-arrys is required in a stage, we assume that the spare PE-array also is used to get an equal load among the modules. We also assume that the special distribution module has several network interfaces to overcome a possible bottleneck in the network. The most demanding stage, from a node-bandwidth point of view, is then the first computation stage where the worst-case bandwidth demand is 500 MByte/s divided between two nodes (three PE-arrays are rounded up to four), i.e., 250 MByte/s (or 2 Gbit/s), both incoming and outgoing from a node. If the minimum number of PE-arrays required holds for systems with the maximum incoming bandwidth of 500 MByte/s is however uncertain. Anyway, one can spread the computations in the first stage (or even other stages too) over more modules than needed. This reduces the worst-case node bandwidth at the cost

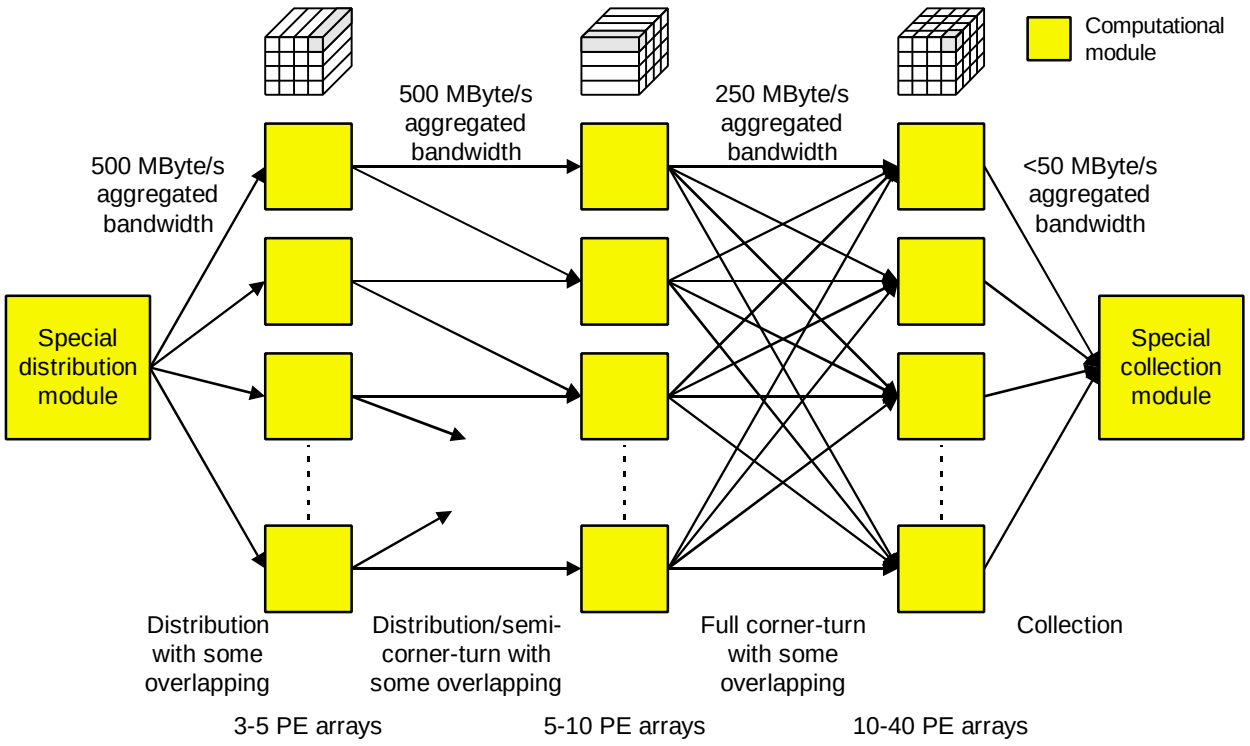

Figure 1: Sample signal processing chain with dataflows between computational modules shown.

of more computational modules. If the 500 MByte/s are spread over three modules in the first and second stages, the worst-case node bandwidth is reduced to 167 MByte/s (1.33 Gbit/s). Spreading over four modules instead gives a worst-case node bandwidth of 125 MByte/s (1 Gbit/s).

The bandwidth requirements mentioned above are given as efficient bandwidths. The bandwidth offered by the network must typically be much higher due to overhead depending on, for example, message sizes, offered low-level support for DMA, group communications etc, overhead in the protocol suit, how bursty the traffic is in the final system, and difficulties in overlapping computation and communication. When looking at peak-performance figures, like those given below for different networks, one must remember to account for these overheads which can be hard to estimate without a real implementation.

A more comprehensive evaluation and survey, but not focused on a specific signal processing chain, has recently been put together [1]. Parts of it were published internationally [2].

2. Mercury's RACEway and RACE++

RACEway from Mercury Computer Systems is a switch-based network especially developed for embedded systems [3] [4] [5] [6]. A RACEway system is built up with 6-port crossbar switches to get an active backplane. Several different topologies can be chosen but the typical topology is fat-tree of switches, where each switch has four children and two parents. Circuit-switching and source routing are used. Support of real-time traffic is obtained by using priorities, where a higher-priority transmission preempts a lower-priority transmission. The link bandwidth (or peak node-bandwidth) is 160 MByte/s.

Next generation of RACEway is called RACE++ and has a link bandwidth of 264 MByte/s. The crossbar size is extended to eight ports. According to Mercury, "Strided DMA" (DMA from/to regular but not consequent address space) is supported. This feature might be valuable when implementing, e.g., corner-turn.

We will now look at some suitable topologies that can be built using 8-port RACE++ switches. The fat tree (see Figure 2) is a fully scalable topology, i.e., the bi-section bandwidth scales linearly with the number of nodes. The number of switches needed is $N/4 \log_2(N/4)$, where N is the number of nodes. With a pipelined dataflow however, networks with fewer switches can be built. A ring of true crossbar switches (see Figure 3) can be used if the ring bandwidth is enough to carry the possible traffic between two stages (not always true as discussed below). Traffic between nodes connected to the same switch is not limited by the switch bandwidth. However, the influence of the actual communication pattern on the performance of the network must be considered as indicated above, e.g., overhead caused by the splitting into many small messages at a corner-turn. With one RACE++ link between each switch in a ring of switches, the bottleneck bandwidth is 264 MByte/s and $N/6$ switches are required. With two or four links in the ring, the number of switches required are $N/4$ and $N/2$, respectively. See Table 1 for the number of switches for some system sizes of the discussed topologies. As indicated, switches might be saved by using a ring-of-switches topology instead of the

Number of nodes:	**16**	**32**	**64**	**128**
Fat tree	8	24	64	160
One-link ring of switches (264 MByte/s)	2.67	5.33	10.7	21.3
Two-link ring of switches (528 MByte/s)	4	8	16	32
Four-link ring of switches (1056 MByte/s)	8	16	32	54

Table 1: Number of switches required for some topologies at different system sizes.

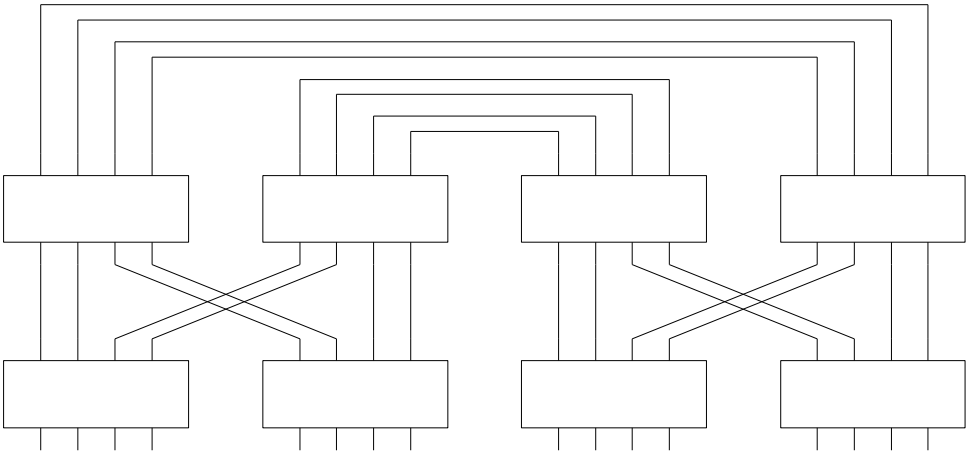

Figure 2: Fat tree of switches.

fully scalable fat-tree topology. One must however remember that multicast traffic might increase the required bandwidth in a ring where traffic else would be removed at a node before the node's outgoing traffic is added (see Figure 4).

3. Myrinet

Myrinet is a switch-based solution, similar to RACEway and RACE++, with support for arbitrary topologies [7]. In the current version (May 1999), full-duplex 1.28 + 1.28 Gbit/s links connect switches and nodes in the selected topology. Host interfaces for PCI and SBus are available, while switches with 4, 8, 12, and 16 ports exist. A parallel computer architecture with an hierarchy of Myrinet switches is reported in [8]. The lowest level in the hierarchy is a switch connecting several

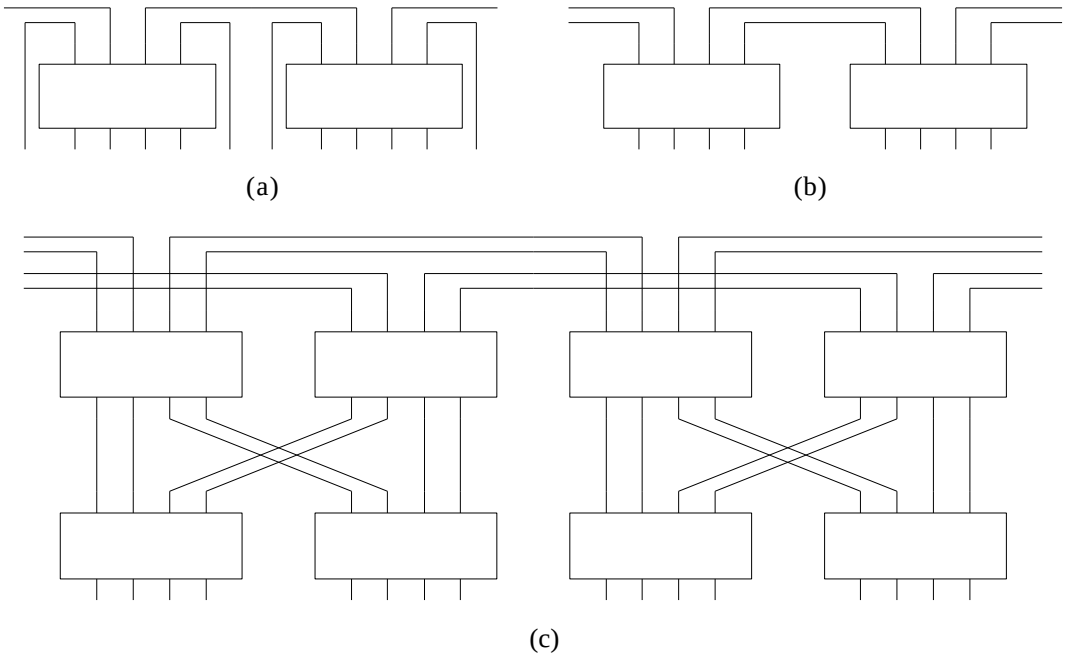

Figure 3: Rings of crossbars with (a) one link between each switch, (b) two links between each switch, and (c) four links between each cluster of switches. The links connecting the linear arrays into rings are not shown.

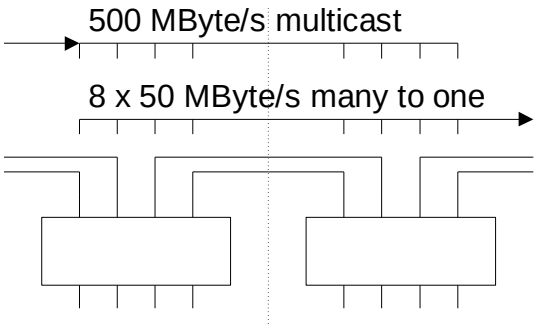

Figure 4: The two switches are assumed to connect all eight nodes of one pipeline stage to the ring of switches. Because the 4 ´ 50 MByte/s transmitted bandwidth by the four first nodes is added to the multicast bandwidth, the aggregated bandwidth crossing the dotted line is 700 MByte/s which is more than the dataflow between two stages (500 and 400 MByte/s, respectively).

processors on a single board. The same switch is an interface to the next level which connects several boards in a backplane. More references to reports on communication systems where Myrinet is used are found in [9].

Myrinet offers a full duplex link-bandwidth of 160 + 160 MByte/s (i.e., 320 MByte/s) compared to 264 MByte/s half duplex in RACE++. Myrinet also offers more freedom in the choice of topology, e.g., because of the variety of switch sizes. The disadvantage with Myrinet is its lack of support for real-time traffic.

4. Fiber-ribbon ring with packet and circuit switching

The fiber-ribbon ring network with two sub-networks, one for circuit switching and one for packet-switching, has been shown to be feasible in radar signal processing systems [10] [11]. The sample system used in the case study was, though, not exactly the same as the one shown above. The nodes in one stage are placed in a row in the ring, while the nodes in the next stage are placed in a row down-stream after the last node in the former stage. Because it is a pipeline ring network (spatial bandwidth-reuse is possible), the highest bandwidth needed over a single link in the ring equals the dataflow between two succeeding stages in the pipeline chain (see Figure 1), plus overhead.

Typically, only the circuit-switched sub-network is used for the dataflow in the signal processing chain. This should be considered when calculating the bandwidth demands. Fiber-ribbon cables with a bandwidth of, e.g., 10 × 800 Mbit/s exist. This translates to a bandwidth of 6.4 Gbit/s (800 MByte/s) if eight fibers are used for the circuit-switched sub-network. Fiber-ribbons with more than ten fibers have been reported in the research literature and are expected to appear commercially. An advantage with the fiber-ribbon ring over bit-serial fiber-optic communication is that a dedicated fiber can be used to clock the data. No clock-recovery circuits are therefore needed. Another advantage is the lack of active components other than the network interfaces (no switches are required for moderate-sized systems). A disadvantage, however, is the enforced link and network-interface bandwidth caused by the fact that the whole dataflow between two succeeding stages must traverse a single link. In a switched network, the topology can be chosen in a way so the dataflow is spread over several links and network-interface as discussed above.

5. Control-channel based fiber-ribbon ring

Most of the discussion for the circuit-/packet-switched ring network holds for the control-channel based network too [11] [12]. The control-channel based network is a little bit more complex but offer a larger range of services, e.g., real-time services and low-level support for process synchronization, group communication, reliable transmission etc. [13] [14]. These features are valuable both for the programmability of the system and to reduce the overhead at, e.g., corner-turns.

6. Fiber-channel etc

Assuming about 1 Gbit/s per node interface, it is not sure if even switched Fiber-Channel does not offer enough bandwidth. HIPPI is a related standard which was first designed for electrical cables [15] [16]. Switched HIPPI exist and the highest currently available link bandwidth known to the author is 1.6 Gbit/s full duplex.

Other networks similar to those discussed above, and Myrinet and RACEway, include:

- *SCI:* standardized network supporting cache-coherence in different topologies of the network, e.g., ring or switch-based [17] [18]. Used in, e.g., a system from Sequent [19]. Bandwidths between 1.25 and 8 Gbit/s was available 1996 according to the authors of [17].

- *Spider:* short-distance (few meters) switch-based network with 2 × 1 GByte/s full duplex links [20], used in SGI's Origin computer systems [21]

- *TNet:* switch-based wormhole-routing network with a link bandwidth of 50 MByte/s in the first implementation [22]

- *HAL's Mercury Interconnect Architecture:* network based on crossbars with six 1.6 + 1.6 Gbyte/s full duplex ports [23]

7. ATM

Experiments with ATM networks in parallel and distributed computing systems have been reported [24]. The small size of the cells (packets) might increase performance for many-to-many traffic where many small messages should be passed between the nodes. Also, ATM can be used with a large range of bit rates. It is, however, questionable if ATM is suitable for the system if *cost* and *complexity* are considered. A variant is to use the cell-switching of ATM and having own protocols above in the protocol suite.

8. WDM star network with the TD-TWDMA protocol

The WDM star network is a flexible network that, with the TD-TWDMA protocol, implements a kind of distributed crossbar with support for real-time services [25] [26]. It scales well up to the practical limit in number of wavelengths (16-32) because each node has its own channel (possibly with a capacity of several Gbit/s) to transmit on. The star-of-stars topology is used for larger networks where the backbone bandwidth, from each cluster, can be chosen similarly as for the ring of crossbars by having multiple channels from each cluster [27]. The WDM technique is promising but most commercially available components are still *expensive* ones made for long-distance telecommunications. Discussions on how to take advantage of the signal-chain properties to get cheaper/simpler networks when designing WDM star networks and similar networks are found in [2].

9. Conclusions

The two major candidates of network solutions seems to be the control-channel based ring and a switch-based solution. The ring offers low hardware complexity without external active hardware, but a higher link bandwidth is needed. The bandwidth requirements might, however, be reduced due to spatial reuse of bandwidth possible when the dataflow has a pipelined nature. The switch-based networks offer solutions where arbitrary topologies normally can be chosen, and where flexible communication patterns are supported due to the properties of true crossbars. Because the switches (depending on the chosen topology) separates the different dataflows, the network-interface bandwidth, and normally also the links, does not have to be capable of higher bandwidths than what can be produced from one node at worst.

Maybe, the only hardware needed for a ring (the network interface) fits on the same ASIC as the computational-module. A good solution would be to design the network interface for the ring, but such generic so the input and output ports can be used as general I/O interface to an external network interface or directly to a switch-based network. The I/O interface to the ring, or another topology, is assumed to be electrical on the ASIC but where an external optoelectronic converter can be used, e.g., OPTOBUS to have fiber-ribbon cables in the network when such are needed/preferred.

References

[1] M. Jonsson, "Optical interconnections in parallel radar signal processing systems," *Research Report CCA - 9909, Centre for Computer Systems Architecture (CCA), Halmstad University, Sweden*, Apr. 1999.

[2] M. Jonsson, "Fiber-optic interconnection networks for signal processing applications," *4th International Workshop on Embedded HPC Systems and Applications (EHPC'99) held in conjunction with the 13th International Parallel Processing Symposium & 10th Symposium on Parallel and Distributed Processing (IPPS/SPDP '99)*, San Juan, Puerto Rico, Apr. 16, 1999. Published in *Lecture Notes in Computer Science*. vol. 1586, Springer Verlag, pp. 1374-1385, 1999, ISBN 3-540-65831-9.

[3] B. C. Kuszmaul, "The RACE network architecture," *Proc. 9th International Parallel Processing Symposium (IPPS'95)*, Santa Barbara, CA, USA, Apr. 25-28, 1995, pp. 508-513.

[4] T. Einstein, "RACEway Interlink - a real-time multicomputing interconnect fabric for high-performance VMEbus systems," *VMEbus Systems*, Spring 1996.

[5] B. Isenstein, "Scaling I/O bandwidth with multiprocessors," *Electronic Design*, June 13, 1994.

[6] *RACE® Series RACEway Interlink Modules Data Sheet*. Mercury Computer Systems, Inc., 1998.

[7] N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic, and W.-K. Su, "Myrinet: a gigabit-per-second local area network," *IEEE Micro*, vol. 15, no. 1, pp. 29-36, Feb. 1995.

[8] T. Boggess and F. Shirley, "High-performance scalable computing for real-time applications," *Proc. of the Sixth International Conference on Computer Communications and Networks (IC³N 97)*, Las Vegas, NV, USA, Sept. 22-25, 1997, pp. 332-335.

[9] R. A. F. Bhoedjang, T. Rühl, and H. E. Bal, "User-level network interface protocols," *Computer*, vol. 31, no. 11, pp. 53-60, Nov. 1998.

[10] M. Jonsson, B. Svensson, M. Taveniku, and A. Åhlander, "Fiber-ribbon pipeline ring network for high-performance distributed computing systems," *Proc. International Symposium on Parallel*

Architectures, Algorithms and Networks (I-SPAN'97), Taipei, Taiwan, Dec. 18-20, 1997, pp. 138-143.

[11] M. Jonsson, "Two fiber-ribbon ring networks for parallel and distributed computing systems," *Optical Engineering*, vol. 37, no. 12, pp. 3196-3204, Dec. 1998.

[12] M. Jonsson, "Control-channel based fiber-ribbon pipeline ring network," *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'98)*, Las Vegas, NV, USA, June 15-17, 1998, pp. 158-165.

[13] M. Jonsson, C. Bergenhem, and J. Olsson, " Fiber-ribbon ring network with services for parallel processing and distributed real-time systems," *Submitted for reviewing*, Feb. 1999.

[14] C. Bergenhem and J. Olsson, "Protocol suite and demonstrator for a high performance real-time network," *Master thesis, Centre for Computer Architecture (CCA), Halmstad University, Sweden*, Jan. 1999.

[15] S. Saunders, *The McGraw-Hill High-Speed LANs Handbook*. McGraw-Hill, New York, NY, USA, 1996, ISBN 0-07-057199-6.

[16] D. Tolmie and J. Renwick, "HIPPI: simplicity yields success," *IEEE Network*, pp. 28-32, Jan. 1993.

[17] D. B. Gustavson and Q. Li, "The scalable coherent interface (SCI)," *IEEE Communications Magazine*, no. 8, pp. 52-63, Aug. 1996.

[18] *IEEE Standard for Scalable Coherent Interface (SCI)*. IEEE, New York, NY, USA, 1993, ISBN 1-55937-222-2.

[19] T. Lovett and R. Clapp, "STiNG: a CC-NUMA computer system for the commercial marketplace," *Proc. 23rd International Symposia on Computer Architecture (ISCA'96)*, May 1996.

[20]M. Galles, "Spider: a high-speed network interconnect," *IEEE Micro*, vol. 17, no. 1, pp. 34-39, Jan./Feb. 1997.

[21]J. Laudon and D. Lenoski, "The SGI Origin: a ccNUMA highly scalable server," *Proc. 24th International Symposia on Computer Architecture (ISCA'97)*, Denver, CO, USA, June 2-4, 1997.

[22]R. W. Horst, "TNet: a reliable system area network," *IEEE Micro*, vol. 15, no. 1, pp. 37-45, Feb. 1995.

[23]W.-D. Weber, S. Gold, P. Helland, T. Shimizu, T. Wicki, and W. Wilcke, "The Mercury interconnect architecture: a cost-effective infrastructure for high-performance servers," *Proc. 24th International Symposia on Computer Architecture (ISCA'97)*, Denver, CO, USA, June 2-4, 1997.

[24]T. von Eicken, A. Basu, and V. Buch, "Low-latency communication over ATM networks using active messages," *IEEE Micro*, vol. 15, no. 1, pp. 46-53, Feb. 1995.

[25] M. Jonsson, A. Åhlander, M. Taveniku, and B. Svensson, "Time-deterministic WDM star network for massively parallel computing in radar systems," *Proc. Massively Parallel Processing using Optical Interconnections, MPPOI 96*, Lahaina, HI, USA, Oct. 27-29, 1996, pp. 85-93.

[26] M. Jonsson, K. Börjesson, and M. Legardt, "Dynamic time-deterministic traffic in a fiber-optic WDM star network," *Proc. 9th Euromicro Workshop on Real Time Systems*, Toledo, Spain, June 11-13, 1997, pp. 25-33.

[27] M. Jonsson and B. Svensson, "On inter-cluster communication in a time-deterministic WDM star network," *Proc. 2nd Workshop on Optics and Computer Science (WOCS)*, Geneva, Switzerland, Apr. 1, 1997.