Communication in parallel and distributed systems

Rohan Patil
Network Engineer,
Mphasis an HP Company.
www.mphasis.com
Pune, India
Email:rohan.anand.patil@gmail.com
Email:Rohan.P@mphasis.com

Pallavee Patil
Lecturer,T.K.I.E.T Shivaji University,
www.tkietwarana.org
Kolhapur, India.
Email: pallavee_patil@tkietwarana.org

Abstract

Communication is an essential part of distributed computing and parallel computing. There exist many topologies, network architectures and routing schemas for such type of communication. In this paper, we will review a few selected network architectures and routing schemes, which are continually going through evolution process.

Key Words

Alpha 21364, ASCI Q Machine, Network of Workstations, Virtual Cut through Switching, Wormhole switching.

I. Introduction

Parallel and distributed computing is an emerging field and it provides a fast and reliable means of efficient computing on low cost. By parallel computing we mean that multiple instructions of same program executed at the same time on multiple processors whereas in distributed computing, multiple computers executes a single task. These computers could be on different geophysical locations. Thus, to achieve fast computing goal it is very important that all we should have a very efficient communication mechanism. If communication mechanism is not efficient, we will lose much of our useful computational time just in communication overhead, and that is focus point of our article. In this article, we will discus different network architectures that include ASCI Q, Alpha 21364 and then we will evaluate efficient routing schemes for different networks such as virtual cut through switching and wormhole switching. We will discuss main building block of quadrics networks i.e. Élan (which is programmable network interface) and its communication switch i.e. Elite. We will also discuss how virtual cut through switching is efficient in network of workstation and then we will put some light on Alpha 21364 network architecture.

A. Parallel and Distributed Computing

Parallel computing is simultaneous execution of same task on multiple processors. V

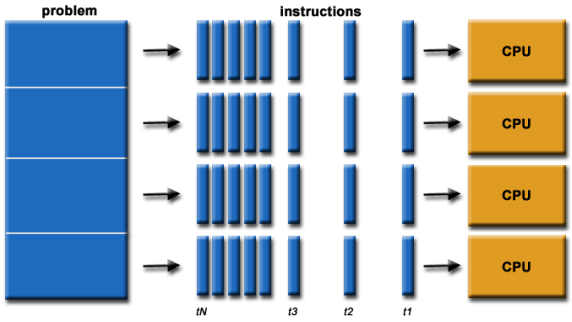

Figure 1 Parallel Computing (Taken from [1])

The motivation behind parallel computing is fast computation and also the fact the amount of memory available on single processor is limited.

.

Figure 2 Distributed Computing (Courtesy University of Surrey)

In distributed computing the task is subdivided into number of smaller task and the executed on number of different computers, they may be separated by large distance and combining the parts to obtain the final result. The main motivation behind distributed computing is to solve the large problem by distributing it over number of processors which is impossible for single processor.

SEEK DIGITAL LIBRARY

The report is arranged as follows in first part we will describe the ASCI Q architecture topology and features, in second part we will describe VCT and Wormhole switching in NOW environment, in third part we will describe Alpha 21364 architecture, network topology and deadlock avoidance schemes.

II. ASCI Q Architecture Topology and Features

First we will take the quick overview of quadrics network backbone of ASCI Q machines, its relevant features. We will describe the network architecture and topology of ASCI Q machine and main mechanism that are at the base of several collective communication patterns.

A. Quadrics Networks

The main building blocks for quadrics networks [3] are programmable network interface Elan [6] and communication switch Elite [7].

Figure 3 The ASCI Q machine at Los Alamos National Laboratory

B. Network Topology

In ASIC Q architecture, the switches are generally interconnected in quaternary fat tree topology. The board consists of 8 switches interconnected in 2 level fat tree.

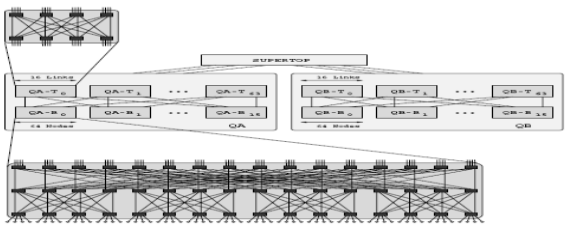

Figure 4 Network Topology of ASCI Q Machines. (Taken from [17])

1) Elan Network Interface

The Elan network interface links multistage quadrics network processing nodes containing one or more CPUs.it accepts packets to and from network; in addition it provides local processing power to implement various communication protocols.

2) Elite Switch

Elite switch has following features, 8 bidirectional links providing two virtual channels in each direction, an internal full crossbar switch, transmission bandwidth of 400 MB/S on each link with latency of 35 ns, CRC protected and it also supports adaptive routing[4]

Elite switches are typically interconnected in quaternary fat tree topology. The main building block ASCI Q machine is board with 8 Elite switches that implement two level fat tree with 16 up and 16 down connections. Four of these boards with backplane are used as switch providing 64 up and 64 down connections. The ASIC Q machine is divided into two clusters QA and QB each with 1024nodes/4096 processors each for the total of 2048 nodes/8192 processors forming two levels of switches with each segment is fat tree of dimension five. The bottom level has 16 switches (labeled as QA/B-B [0-15]) with 64 downlinks to nodes and 64 uplinks to top level switches. The top level has 64 switches (labeled as QA/B-T [0-63]) with 16 up and 16 down connections. The down connections are connected to distinct bottom switch. The set of switches at the top are called supertop.Due to the low latency of Elite switches the latency between any two nodes is almost constant independent of the distance.

C. Hardware and software support for collective communication

Quadrics network provides hardware support in both network interface and the switches for collective communication. Multicast packets can be sent either by hardware multicast capability of the network or the software tree.

1) Hardware Multicast

The process in the node injects packet in the node. Packet can only follow predetermined ascending path. By default, the top leftmost switch is chosen as a root node every packet must pass trough this switch in order to avoid the deadlock. Every packet must reach the root node before it can be multicasted.If another collective communication is issued while the first one is still in progress it is serialized at or before root switch. The second communication can start only when all circuits of first communication are cleared. For the successful transaction, single positive acknowledgement must be received from all recipients of multicast group pioneered by NYU U1 tracomputer [5]. Therefore, all members of multicast group will receive either the massage or none.

SEEK DIGITAL LIBRARY

2) Software tree

In hardware multicast all the recipients of multicast massage must be contiguous if this is not the case multicast can still be possible by using software trees. This is achieved by using thread processors of Elan, which receives the massage and sends multiple copies of the massage to the appropriate destinations within few microseconds without any interaction with the main processor.

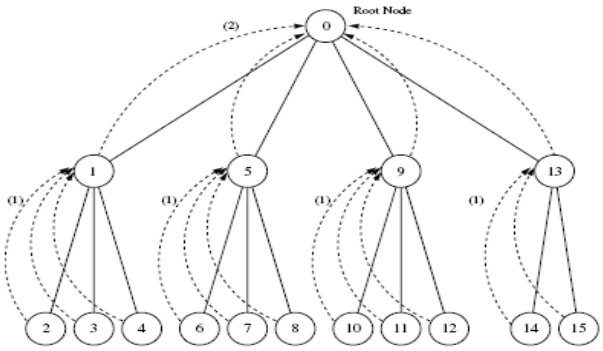

Figure 5 Software tree implemented with élan and network interface (Taken from [17])

a) Barrier Synchronization

Barrier synchronization is a logical point in the flow control of parallel program where all processes in multicast group must arrive before they are allowed to proceed. If all the nodes in barrier synchronization are contiguous, it is possible to use hardware multicast. When barrier is performed all the processes in the group write a sequence number to the memory location and wait for global go signal from the master process(one with the lowest ID).The master process sends special test and send multicast packet using Elan threads. This packet spans all the processes in the group and checks the barrier sequence numbers of all the processes. All the replies are combined into a single massage by Elite switches on the way back so the master process in the root node receives the single acknowledgement.

The software algorithm based on point to point massaging uses balanced tree to send the ready signal to the master process. Each process in the tree waits for the ready signal from its children and then sends its own ready signal to the parent process. When parent process receives ready signal from all of its children nodes it broadcasts a go signal using the same tree structure.

b) Broadcast

In broadcast root node sends chunk of data to multiple destinations. All the recipients of broadcast massage have the same virtual address. The hardware multicast can be used for user level broadcast. If all the recipients of broadcast massage are on the same node then original massage is copied into shared memory from where all recipients copy the massage using double buffering technique.

c) Allreduce

The common pattern used in collective communication is allreduce in which it computes arithmetic operations on set of processes collects the information from one vector and updates another vector of same size with the result based on arithmetic operations. The allreduce operation can be logically divided into two phases collection phase and distribution phase. Point to point messaging can used for collection phase while broadcast algorithm can be used for distribution phase.

d) Hot Spots

All the nodes send massage to single node, which becomes hot spot. This pattern represents system network traffic rather than user level communication.

III. WORMHOLE SWITCHING VS. VIRTUAL CUT THROUGH SWITCHING IN A "NOW" ENVIRONMENT

Most of the NOW networks traditionally use wormhole switching for fast computing but in "NOW" environment (Network of Workstations) virtual cut through switching gives higher throughput than wormhole switching. The long distance between workstations requires use of long wires and as a result large buffers to avoid overflow in wormhole switching and also traditional disadvantages of Virtual cut through switching like buffer size, packetizing overheads disappear in NOW's.

Wormhole switching was introduced as a suitable switching technique for fast and compact routers. Wormhole switching pipelines the massage across multiple routers and requires very small buffers. If the channel requested by the massage is busy it is blocked in a place. Flow control signal are exchanged between routers so buffers need to store packet for few flow control digits or flits. [8, 12]

Virtual cut through switching also pipelines massage across multiple routers as wormhole switching but flow control is performed at the level of packets requiring the buffers with capacity to store one or more packets. When the packet is blocked it is removed from the network and stored in the buffer reduces contention considerably.

The main reason for using wormhole switching instead of virtual cut through switching is due to large buffers required in virtual cut through switching. In wormhole, switching buffer size is limited by wire length. In multicomputer system as computers are closer to each other the wire length and as a result buffer size required in wormhole switching is small but in NOW's distance between workstations is long so wire length and buffer size is large in wormhole switching. So virtual cut through switching is a better option in NOW environment.

Most of the latency in a communication network is due to overheads in messaging layer and interface card. If massage is not packetized communication cannot start until whole massage is stored in the memory. In virtual cut through

IEEE COMPUTER SOCIETY DIGITAL LIBRARY

switching massage is divided into packets of fixed size, performance is increased[10] because packets are pipelined through network interface card, concurrently transmitted from main memory to network interface card and from there to the network.

A. Network of Workstations

Routing decisions in irregular networks can be based on source routing or distributed routing. In both cases network mapping should be performed in order to fill the routing table. The routing algorithm must determine the path before routing can take place. Several deadlock free algorithms and two general methodologies for adaptive routing have been proposed-

1) Up/Down* Routing*

In up*/down* routing [11] the spanning tree is computed for each operational link. The up end is defined as 1) the end whose switch is closer to the root in the spanning tree, 2) the end whose switch has lower ID if both ends are at switches at same tree level.
To avoid deadlocks following rule is used: the legal route must transverse through zero or more links in up direction followed by zero or more links in down direction. This routing avoids deadlocks but does not always provide minimal path between workstations [13].
(Minimal path: The path which will take shortest time to reach destination, it may not be the shortest path.)

2) Adaptive Routing

In adaptive routing physical links are divided into two virtual channels 1) original and 2) new channels. Newly injected packet can only take new channels belonging to minimal path. If all the new channels are busy then up*/down* routing is used to select original channel providing minimal path if any. If no minimal path is available then shortest path is selected. If packet reserves original channel it is only routed through original channel until delivered using up*/down* routing [13, 14].

B. Design Considerations

Normally wormhole router consist of flit buffer, switch and link controllers associated with input and output links and routing and arbitration unit [9]. Virtual channels are implemented by using separate flit buffers for each virtual channel. For efficiency switch should have as many ports as virtual channels. Multiplexed crossbar switch or non-multiplexed crossbar switch is used; the percentage of silicon area occupied by switch becomes very large when non-multiplexed switch is used but multiplexed switch increases complexity of the design and propagation delay.
The architecture of VCT router is quite similar to wormhole switch but the only mandatory difference is the buffers in VCT must be able to store one or more packets.

The main difference between VCT and Wormhole router is the type of crossbar used.VCT uses simpler crossbar with less number of ports but it introduces some contention as packets ask for same input port. The throughput of VCT is quite same as wormhole router but it is simpler to implement, it uses the less number of switches, occupies less silicon area and propagation delay is less than wormhole routing [8].

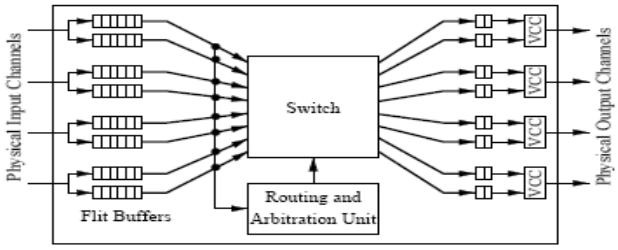

Figure 6 Organization of wormhole router with virtual channels.(Taken from [16])

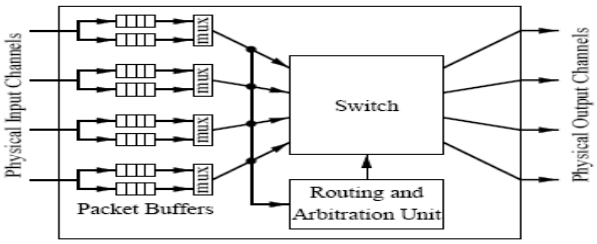

Figure 7 Organization of VCT router with multiplexed crossbar (Taken from [16])

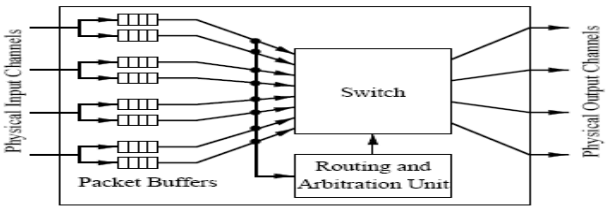

Figure 8 VCT router with partially multiplexed crossbar. (Taken from [16])

C. A New Routing Algorithm

The new routing algorithm is implemented on VCT router. This has two queues associated with each physical link namely new and original respectively. Each queue can store one or more packets. The main difference between new algorithm and previous one is that new one allows new queues to be reserved after reserving the original ones, which is forbidden in previous one. Once the packet is routed through new queue it can be stored in any queue either original or new associated with incoming link of next switch. New routing scheme allows packets to follow minimal path with higher probability than the previous one. This routing algorithm is fully adaptive as it allows packets

SEEK DIGITAL LIBRARY

to pass from new queues to original queues so the packets can follow the minimal path. This routing algorithm can also be implemented on wormhole router if the buffer size is increased to store entire packet. In this case, buffers associated with virtual channels replace queues.

D. Performance Evaluation

For the performance evaluation of VCT router using the new algorithm, we assume that the queue can store one packet.

We will compare it with performance of wormhole router using the previous algorithm. We assume that the buffer can store entire packet

1) Switch and Network Model

In a switch model, we assume that it takes one clock cycle to compute routing algorithm also I takes one clock cycle to transmit a flit over internal crossbar. The data is injected at one flit per cycle and the fly time for each flit is 4 cycles. For wormhole switch, we assume non-multiplexed crossbar switch with input buffer size if 27 flits and output buffer size of 4 flits.The VCT switch each queue can store single packet with partially multiplexed switch.

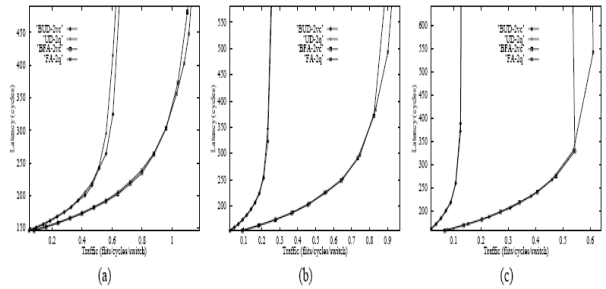

Figure 9 Performance Evaluation of VCT and wormhole Switching using new algorithm (Taken from [16])

2) Simulation Results

Figure 9 shows that the new algorithm reduces the latency and increases the throughput as compared to old one. This improvement in the performance is due to the increase in buffer size and full adaptability of routing algorithm. As the network size increases, the new routing algorithm is more efficient than the old one. It can be seen from the result that for the small networks VCT router performs better than wormholes but as the network size increases the distance between the switches increases and the routing algorithm that provides higher adaptively and routing through minimal path will help greatly in improving the performance.

IV. ALPHA 21364 NETWORK ARCHITECTURE.

Alpha 21364 processor provides high performance reliable network architecture with the router that runs at 1.2 GHz with peak bandwidth of 22.4 GB/second. This architecture is well suited for intensive server applications. [15] Alpha 21364 microprocessor uses 152 million transistors for Alpha 21364 processor core, 1.75 Mbyte second level cache, two memory controllers, cache coherence hardware and a multiprocessor router on a single die. It runs at 1.2 GHz provides 12.8 GB/s local memory bandwidth 22.4 GB/s router bandwidth supporting configuration up to 128 processors. A fully configured 128 processors provide 4 terabyte of Rambus memory and hundreds of terabytes of disk storage. This architecture is well suited for Web server applications, database servers. This configuration can be easily modified to support even larger configuration. This architecture offers very low latency, enormous bandwidth. The router offers very low latency because it runs at same clock speed as main processor i.e. 1.2 GHz. The pin to pin latency in router is 10.ns.The pin to pin latency for ASIC based router is 40 ns.The architecture also provides large bandwidth and sustains 70 to 90 percentage of it because of carefully designed algorithms, large amount of on chip buffering and fully pipelined router implementation.

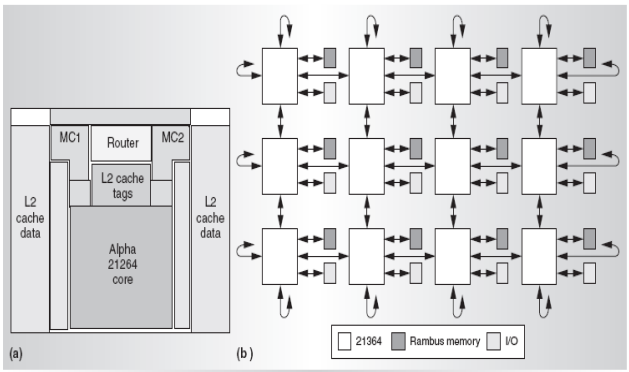

Figure 10 Alpha 21364 Architecture (Taken from [15])

A. Network Packet Classes

The packet is the portion of massage transported from one router to the other across the network. The flit is the portion of the packet transported in Parallel on a single clock edge.

The flit has 39 bits 32-bit payload and 7 bit error correction code. The 21364-processor support packets of 2,3,18 or 19 Flits. Packets first one to three flits contain packet Header.18 or 19 flit packet consist of 16 flit cache blocks or 16 flit I/O data in addition to header. There are 7 packet classes supported by 21364 processor.

Request (3 flits): A processor or I/O uses request packet to obtain 64 byte (16 flit) data or cache block.

SEEK DIGITAL LIBRARY

Forward (3 flits): The memory controller MC1 or MC2 uses forward packet to forward request packet to the current owner of cache block (processor or I/O device)

Block Response (18 to 19 flits): The processor or I/O device uses block response packet return data requested by request class packet.

Nonblack Response (2 or 3 flits): The processor or memory controller uses nonblock response packet to acknowledge requests.

Write I/O (19 flits): The processor or I/O device uses write I/O packet when it stores data to I/O memory.

Read I/O (3 flits): This packet is used to read data from I/O memory.

The packet header contains the information regarding packet class and function routing information for the packets. It can also contain physical address of I/O or cache block and flow control information.

B. Network Architecture

1) Virtual Cut through Routing

21364 is a 2D torus network 21364 network uses virtual cut through routing in which if there is the congestion in the network flits of packet are buffered until congestion clears. The network has buffer space to store 316 packets.

2) Adaptive Routing

In 21364 architecture packets adaptively route within minimum rectangle.(two points on rectangle i.e. current router and destination processor as opposite vertices of the diagonal of a rectangle). The minimum rectangle is the rectangle with the minimum diagonal distance between current router and destination processor.

The adaptive routing algorithm picks one of two possible output port through which the packet at the input of current router can be routed. Either it can continue in same dimension or it can change the dimension this is because manhattan distance to the destination reduced at every hop. If both the output ports are free then packet is routed via the port in the same dimension.

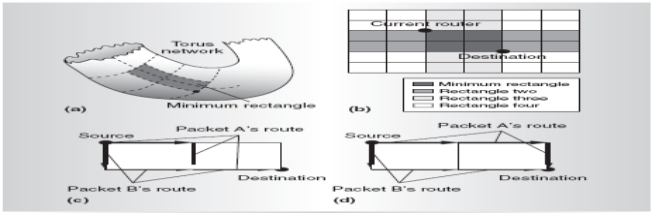

Figure 10 Adaptive routing in torus network (Taken from [15])

C. Deadlock Avoidance Rules

In this section, we will discuss deadlocks in different protocols of Alpha 21364 architecture and the mechanisms implied to overcome these problems.

1) Avoiding Deadlock in Coherence Protocol

The 21364 avoids the deadlock in coherence protocol that is the cyclic dependencies, request packets fill the network and block response packets reaching the destination, by providing virtual channels for all packet classes so the request packet can never block the block resonance packet. In addition, it assigns priorities to packet classes. Thus request can generate block response but block response cannot generate request.

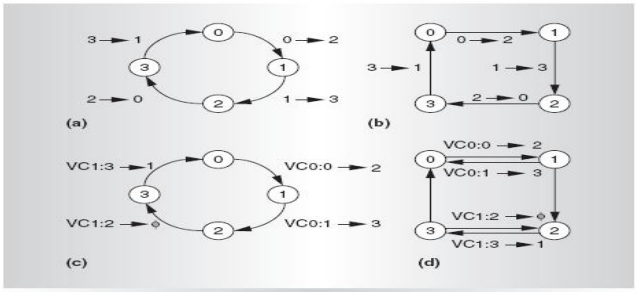

Figure 11 Deadlock avoidance by dividing buffers into virtual circuits (Taken from [15])

2) Avoiding Deadlock in Adaptive Routing

Adaptive routing can generate two types of deadlock in 21364-network intra and inter dimension deadlock. This is caused due to packets are waiting for buffers in forward direction to free up causing the deadlock within the processors of different dimensions.

The 21364 breaks this deadlock by dividing the buffers into two virtual channels VC0 and VC1.In virtual channel each packet is first routed along the primary (horizontal) and then along secondary axis (vertical). The packets along the primary axis depend on packets along secondary axis but the packets along secondary axis do not depend on the packets along the secondary axis.

D. Router Architecture

The 21364 has nine pipeline types based on input and output ports. There are three types of input and output ports: local (cache and memory controllers) interprocessor (off-chip network) and I/O. Any type of input port can route packet to any type of output port. In addition to pipeline latency, there are synchronization delay, pad, receiver and driver delay and transport delay from pins to router and form router back to the pins. The packet header entering the router is of 16 bit these bit contain information about the destination virtual channel number, packet type (I/O) and reserved bits. Each configuration table entry contains

SEEK DIGITAL LIBRARY

bits containing headers routing information, access control and parity bit.

E. Error Correction Code

Each 32-bit flit is protected by 7-bit error correcting code. The router cheeks ECC for every flit of a packet. For every packet arriving trough network the router calculates ECC and compares it with received ECC, if there is single bit error router corrects it and reports it to operating system via interrupt; it does not correct double bit errors.

F. Input Buffering

The 21364 provides buffering only at the input port. Each port can store single packet of any class except local ports. For local ports, packet payload resides in internal buffers.

G. Arbitration

One of the most demanding components of 21364 routers is arbitration mechanism. This schedules the dispatch of the packets arriving at the input ports. The 21364 breaks the arbitration logic into local and global arbitration. There are 16 local arbiters one for each input port and 7 global arbiters 1 for each output port. In each cycle local arbiter schedules packet for dispatch to an output port. One cycle after the completion of local arbitration global arbiter selects one out of up to seven scheduled packets for the dispatch through the output port. Once global arbiter makes such decision all flits in the crossbar follow the input port to the output port.

V. CONCLUSION

By above report we conclude that the collective communication in large networks is possible in a scalable way with the latency comparable to single point-to-point communication.

We compared the performance of Virtual cut through switching and wormhole switching in “NOW” Network of Workstations. In Wormhole switching buffer size is limited by wire length but as distance between workstations is large, so VCT is more suitable in such environment as it can store whole packet. In VCT buffer must be large enough to store the whole packet so the packets are divided into number of fixed size packets which induces packetization delay.VCT allows more routing flexibility preventing deadlock because blocked packets are buffered and removed from the network. The performance of the network not only depends on the network architecture but also the algorithm in use. As we have seen, the new routing algorithm provides fully adaptive routing the probability of routing the packet trough minimal path thus reducing the latency.

With development in System on Chip technology, it is expected that on chip networks will become more common. The router pipeline will face same problem as microprocessor pipeline so the data and control pats must be carefully designed to avoid delays. In addition, the routing algorithms must be carefully implemented to match the technology. Thanks to latest technology, we are able to develop very small transistors and even nanotechnology is not far away from so the chip size is decreasing day by day but we have to find a mean between network size, scalability, performance and cost to meet our requirements.

VI. REFERENCES

[1] http://distributedcomputing.info/

[2]https://computing.llnl.gov/tutorials/parallel_comp/#Whats

[3] F.Petrini, W.Feng, A.Hoisie, S.Coll, and E.Frachtnberg The Quadrics Network: High-Performance Clustering Technology. IEEEMicro, 22(1):46-57, Jan./Feb.2002.Availablefromhttp://www.c3.lanl.gov/~abrizio/Papers/ieeemicro.pdf.

[4]F.Petrini and M.Vanneschi.”Performance Analysis of k-ary n trees”. International Journal Wormhole Routed on FoundationsofComputerScience,9(2):157-177,June 1998.

[5] G.F.Pfister and V.A.Norton.”HotSpot”Contentionand Combining in Multistage Interconnection Networks. IEEETransactions on Computers, C-34(10):943-948, Oct.1985.

[6] Quadrics Supercomputers World Ltd. Elan Reference Manual, Jan.1999.

[7] Quadrics Supercomputers World Ltd. Elite Reference Manual, Nov.1999.

[8] W.J.Dally, “Virtual-channel powcontrol,” IEEE Transactions on ParallelandDistributedSystems, vol.3, no.2, Pp.194–205, March1992.

[9] J.Duato, et al., Interconnection Networks: An Engineering Approach.IEEEComputerSocietyPress, 1997

[10] J.RexfordandK.G.Shin, “Support for multiple classes of traffic in multicomputer routers “Proceedings of the PCRCW94, pp.116–130, May1994.

[11] M.D.Schroederetal., “Autonet: A high-speed, self- Configuring local area network using point-to-point links,”TechnicalReportSRCresearchreport59, DEC, April1990.

[12]S.L.ScottandG.Thorson, “TheCrayT3Enetwork: adaptiveroutinginahighperformance3Dtorus,”ProceedingsofHot Interconnects Symposium IV, August1996.

[13]F.Silla, et al., “Efficient adaptive routing in networksof workstations with irregular topology,” Proceedings of CANPC’97, pp.46–60, February1997.

[14]F.Silla and J.Duato, “Improving the efficiency of adaptive routing in networks with irregular topology,” Proc. of the 1997 Conf.on High Performance Computing, Dec.1997.

[15] Mukherjee, S. S., P. Bannon, S. Lang, A. Spink, and D.Webb, "The Alpha 21364 Network Architecture," *IEEEMicro*, vol. 22, no. 1, pp. 26-35, Jan.-Feb. 2002.

[16] Duato J., A. Robles, F. Silla, R. Beivide,"A Comparison of Router Architectures for Virtual Cut Through and Wormhole Switching in a NOW Environment," 13th International Parallel Processing Symposium & 10th Symposium on Parallel and Distributed Processing (IPPS/SPDP '99), San Juan, Puerto Rico, Apr. 12-16, 1999, pp. 240-247.

[17]Petrini, F., J. Fernandez, E. Frachtenberg, S. Coll, "Scalable Collective Communication on the ASCI Q Machine," Proceedings of the 11 th Symposium on High Performance Interconnects (HOTI’03), Aug. 20 22, 2003, pp. 54-59.

SEEK DIGITAL LIBRARY