Large Valency Serial Wormhole Routing Networks as a Scalable Multimedia Switching Infrastructure

Neil Davies and Peter Thompson
({neil, thompson}@pact.srf.ac.uk)

SRF/PACT
University Gate
Park Row
Bristol BS1 5UB
UK

Abstract

Multimedia data provides particular challenges for systems interconnect. Although high throughputs are required, e.g. for multiple video streams, low and predictable latency and round-trip times are also needed, e.g. for duplex voice communication. Most interconnects have high latency-bandwidth products, which forces a compromise. Interconnects based on IEEE 1355 have low inherent buffering, which translates to a low latency-bandwidth product, thus providing the basis for a variety of services for different categories of multimedia data.

Serial wormhole routing chips of valency 32 have been available for several years, providing paths with implicit flow control with very low per path buffering. An extensive study of networks of these devices has been performed to characterise their latency, throughput and delay variation over many network topologies. Networks of up to 1024 100Mbits/s terminal nodes have been simulated, constructed and measured, demonstrating properties that are beneficial for the construction of scalable infrastructures for high performance multimedia switches, including a rapid indication of internal contention that can be used to effectively manage the quality of service across the switch.

We will present the major results of the study and outline how they can be used to design low cost-per-port scalable switches with the potential for the modular addition of fault tolerance.

1 Introduction

1.1 Multimedia Data Streams

Until recently switching infrastructures were of two distinct types, one for telecommunications and one for computer data, typified by the telephone system on one hand and the Internet on the other. Telecommunications traffic has the characteristics of a large number of channels, constancy of bit-rate (e.g. 64kbits/s per channel) and intolerance both to high latency and to large packet delay variation. Dedicated fixed-rate channels have

traditionally handled such traffic. Computer data traffic, on the other hand, has the characteristics of highly variable bit-rate (from long periods of nothing to bursts of 10Mbits/s or more per channel), relatively high tolerance of latency (provided that protocol time-outs are not exceeded), and relatively low tolerance of errors. Such traffic is typically multiplexed onto shared channels with the assumption that the high-rate bursts are uncorrelated.

The introduction of digital video has brought a new set of requirements, with bit-rates in the order of 2-16Mbits/s per channel, with potential variability caused by compression. Because the corresponding uncompressed stream is constant-rate, any packet delay variation must be compensated by (costly) buffering, usually in the user terminal. Furthermore, the integration of multimedia facilities (audio and video) into computers, and of computers into the media infrastructure (such as digital TV and audio studio equipment) has blurred the distinction between the formerly separate categories. Consequently there is now a need to deal with "multimedia data" which combines the requirements of old-style constant bit-rate traffic, computer data and digital video. Multimedia data streams combine high bandwidth with low tolerance to latency, packet delay variation and errors. Moreover the potential number of streams in applications such as home video-on-demand (VoD) vastly exceeds the number of computers in a typical network. Such services will not be viable without a cost-effective means of handling large numbers of multimedia data streams since the cost that the individual consumer will bear for them is limited.

1.2 High Latency-bandwidth Product Interconnects

Interconnects designed primarily for long-distance connections, such as ATM, typically contain features that are oriented towards maximising the utilisation of the point-to-point bandwidth. If we consider a switch for such an interconnect, with data arriving on a number of inputs destined for a number of outputs with some degree of randomness, we can see that even if data arrives on every input in a given unit of time we can have no guarantee that this will include data destined for every output. Thus, if we take no action, a proportion (tending towards 36.8% as the number of outputs increases [1]) of the outputs will be idle, contradicting the intention to maximise the utilisation of the output links. The usual solution to this is to provide buffering inside the switch to hold data received from the inputs for retransmission at a later time. This improves the probability that any particular output can be provided with data to transmit in each unit of time, at the price of increasing the latency, not to mention the cost and complexity. This creates an awkward problem for switching multimedia data with its combination of high bandwidth and real-time requirements, the solution of which causes even more complexity to be added.

The alternative is to consider an interconnect in which the provision of raw point-to-point bandwidth is so inexpensive that heroic efforts to use it all are unnecessary, and for which low latency-bandwidth product switching is feasible. This is the subject of this paper.

1.3 Systems scenario

There are several potential multimedia system architectures that differ in where the multimedia traffic can be switched. We are focussing here on the easiest to envisage, where there is a central switching facility, with no additional switching occurring elsewhere. This simplifies the quality of service management issues, as they can all be solved centrally. The central switch consists of interfaces, which adapt to external protocols, be they long-distance to the end-users or short distance to service devices such as disks, interconnected

by a switching fabric. In this paper we are concentrating on the performance scalability of the switching fabric, since this is frequently the dominant cost in large systems.

In such a centrally switched environment there are several types of traffic that can be flowing through the switch fabric at any one time. These include VoD traffic from central repositories to end users, telephony and Internet access between end-users and the wider world, and even interactive games between end users. The end users are connected to the central switch only at the rate of their most 'greedy' service - thus minimising the cost of the transmission infrastructure.

The goal is to sustain the quality of service for the traffic that requires it (telephony, video), maximise the bandwidth available for other things while minimising the cost. This raises the question of what is a cost-effective way of switching to and from such a set of streams and interconnecting them with the video servers and other systems. Further questions are whether the solution is scalable, and whether it allows for high-availability, both while remaining cost-effective. These are the questions that we consider in this paper.

2 IEEE 1355

There has been an explosion of interest in serial interconnection techniques over the last five years, as the technology has become mature and the limitations of bus-based interconnects have become more severe. The IEEE 1355-1995 standard (ISO/IEC 14575) of scalable, heterogeneous interconnect [2] defines a set of lightweight serial protocols suitable for low-cost implementation. IEEE 1355 includes a range of different speeds and media to allow flexibility in the choice of cost-performance point, all of which support a common packet layer. At this layer, arbitrary length packets are transferred from point to point as if through fully handshaken FIFOs. All connections are full duplex and multiplex flow-control information with the data in order to produce the FIFO effect.

2.1 Serial Wormhole Routing

Wormhole (or “cut-through”) routing is a packet switching technique in which only the first part of a packet is required to make the routing decision, which can therefore be done before the whole packet has been received, thus minimising routing latency. It has been used for some time as the basis for interconnection between elements of large parallel processors [3]. However, in this context the connections between devices are generally wide, transmitting between 8 and 32 bits simultaneously, with additional signals for flow-control. Only with IEEE 1355 did the two concepts of serial communications and wormhole routing come together, as the protocols of 1355 were specifically designed to permit low-latency wormhole switching.

2.2 Large Valency Switches

The low implementation complexity of IEEE 1355 interfaces together with the minimal buffering requirement of wormhole routing allows single chip switching devices to be made with a large number of interfaces (a “high valency”). One such device can thus interconnect many other terminal devices equipped with 1355 interfaces. Examples are the STC104 developed by SGS-Thomson microelectronics with 32 100Mbit/s DS-SE-01 links and the R^3 router prototype with 8 1Gbit/s HS-SE-10 links developed by the University of Paris VI [5].

2.3 Wormhole Routing Networks

When the number of terminals is such that a single switching device is no longer adequate,

a number of such devices can be interconnected to make a network, since the protocols are symmetrical. As the individual components of the network are crossbar switches there are many ways of interconnecting them, each having a differing amount of internal bandwidth. The precise way in which the switches constituting the network are interconnected defines its topology, and determines its performance, in particular the data rate at which it saturates and the latency experienced by packets due to internal contention.

2.4 IEEE 1355 Network Simulations

In the adoption of any new technology, especially one where the behaviour of the whole system is not immediately obvious from the specification of the parts (however simple), there is a "bootstrap problem" to be overcome. On the one hand there is no incentive to investigate the application of the technology to an area where it may fail to deliver the right results, and on the other the technology cannot be shown to deliver such results until it is applied to a particular area. Determining whether a technology is suitable for an application can be costly, and so there is a strong motivation for a general study that will give good indications of which applications and technologies are well matched. An analogy would be a geological map that helps to determine the most profitable regions to explore for oil. In the case of 1355, the EU under the Macrame project (ESPRIT/OMI 8603) has funded this speculative study, and the results are now available.

The simulation study performed under the Macrame project investigated the performance of simulated networks of STC104 switch chips interconnected by 100Mbit/s DS-SE-01 links in a variety of topologies subjected to a variety of applied traffic loads. Topologies considered were various kinds of grids and k-ary n-cubes, together with multi-stage networks of the "folded Clos" type [1], interconnecting between 32 and 512 terminals. Traffic loads considered were uniform random traffic, random permuted traffic and permuted traffic exercising the longest path lengths in the network. The full results are tabulated in [6], but for the purposes of this paper we will concentrate only on those loads most similar to multimedia data streams, and those topologies most cost-effective in switching them.

Simulations were performed using a special-purpose simulator, TOPSIM, developed by SINTEF under the Macrame project, and typically used several hours of UltraSparc time to determine the stable network performance under each specific set of traffic parameters. The results were calibrated against measurements performed by CERN on the testbed constructed under the same project [7]. Having performed these simulations, we can observe the performance behaviour as the switching fabric scales in both traffic density and number of terminals.

3 Scalable Multimedia Switches

The system studied here through simulation represents the symmetric worst-case analysis of multimedia switching systems. To maximize the utility of the results we make very few assumptions on the pattern of the traffic, other than that the load per node is bounded at 20Mbits/s, corresponding to about 15Mbits/s of user data. We also assumed a worst case symmetric data-flow within the switching fabric, although in most applications data-flow would be asymmetric with a much lower flow from end-users to servers. We have chosen a packet size of 64 bytes (plus two byte routing header) since this represents a reasonable compromise between an ATM cell and an MPEG-2 transport packet. Because the switching technology is inherently packet-length independent, there are no dramatic changes in behaviour as the packet size is varied, and so an intermediate choice of packet size is valid.

The results in table 1 show the scaling of average packet latency and packet delay variation of a constant load from each terminal, as the number of terminals goes from 32 to 512. A variety of interconnection topologies are shown, and for each one the number of 32-valent switches needed to implement it and the minimum and maximum path lengths (number of switches between two terminals) is shown. The proximity to the saturation point of the network is also shown.

Table 1: Latency and Packet Delay Variation at a load of 20Mbits/s per terminal

No. Switches	Path length	Average Latency (microseconds)[1]			Packet Delay Variation (microseconds)			Saturation Bandwidth[3]		Topology[4]
	Min/ Max	Random	Permutation[2]		Random	Permutation[2]		Remaining	Percentage of	
			Avrg	Worst		Avrg	Worst			
32 Terminal Networks										
1	1/1	8.9	8.0		28.8	10.3		60	25	Single crossbar
64 Terminal Networks										
4	1/3	10.2	9.2	10.5	35.1	13.3	14.2	40	33	2D Hypercube
6	1/3	10.8	9.9	10.5	36.6	13.1	12.9	45	31	3 stage Clos
8	1/4	10.8	10.3	11.8	30.9	14.5	14.1	50	29	3D Hypercube
8	1/4	10.8	9.8	11.8	32.5	13.0	15.2	50	29	2D Torus
14	1/3	11.1	10.2	10.5	31.7	13.0	12.8	55	27	3 stage Clos
128 Terminal Networks										
8	1/5	11.4	10.4	39.4	37.9	23.9	74.0	20	50	2D Grid
8	1/4	10.8	9.9	12.3	41.3	19.1	24.5	20	50	3D Hypercube
12	1/3	11.2	10.2	10.5	37.2	15.9	13.0	40	33	3 stage Clos
16	1/6	11.7	11.0	13.1	33.5	16.6	21.9	30	40	3D Grid
16	1/5	11.4	10.9	13.0	32.5	15.4	15.6	45	31	4D Hypercube
20	1/5	12.9	12.1	13.0	37.9	15.6	15.4	45	31	5 stage Clos
28	1/3	11.4	10.4	10.5	17.8	43.3	12.8	55	27	3 stage Clos
256 Terminal Networks										
16	1/15	16.7	15.9	75.4	64.6	58.7	164.6	0	100	2D torus
16	1/5	11.9	10.9	16.0	37.6	24.0	43.0	20	50	4D Hypercube
24	1/3	11.5	10.8	10.8	35.5	22.0	13.2	35	36	3 stage Clos
32	1/7	12.3	11.3	15.8	43.2	14.5	37.8	10	67	3D Torus
32	1/6	12.1	12.0	15.0	32.5	20.8	26.9	40	33	5D Hypercube
40	1/5	13.4	12.7	13.5	34.6	18.5	16.0	35	36	5 stage Clos
56	1/7	15.5	15.1	16.2	39.3	18.7	18.7	40	33	7 stage Clos
512 Terminal Networks										
48	1/3	12.5	11.7	10.8	43.7	34.7	13.2	20	50	3 stage Clos
80	1/5	14.2	13.4	13.5	37.5	26.4	15.9	30	40	5 stage Clos
112	1/7	16.2	15.4	16.2	40.0	22.0	18.6	35	36	7 stage Clos

Notes:

1. The latency values are averaged over all packets, and have three components:
   - Time taken to transmit a 64 byte packet (fixed around 6.5 microseconds)
   - Time taken to make a switching decision (1 microseconds per switch traversed)
   - Time waiting due to contention for a transmission path (load dependant)

2. The “worst case” permutation is the one that maximises the path length, which in some situations (eg the 512 terminal folded Clos networks) can minimise the contention for internal resources.

3. The saturation value chosen was the minimum for the three traffic patterns - typically this is the random pattern for Clos networks and the worst case permutation for the others. This is because in the Clos case the traffic saturates due only to head of line blocking for the output port. In the other cases the contention that dominates is within the network.

4. All the Clos networks studied were 'fat' in the sense that internal bandwidth was sufficient to cope with a worst case permutation. 'Thinner' Clos networks would probably give reasonable performance for this traffic scenario, using fewer switches at the cost of a slight increase in the packet delay variation. Only Clos networks are included in the 512 case as the other topologies (grids, torii and hypercubes) are beyond saturation at 20Mbits/s per link.

3.1 Discussion

In the first instance we can see from these results that a scalable amount of multimedia traffic can be switched using wormhole routing networks constructed from large valency switches, providing an excellent quality of service, simply by choosing a topology and an operating point where the latency and packet delay variation are low. The important performance characteristics (sustainable throughput, latency and packet delay variation) remain excellent as the system grows. For instance, we see from table 1 that the difference in latencies from a minimum configuration of 32 nodes to 512 nodes is only 11.5 microseconds to 16.0 microseconds.

With the increasing interest in switching multimedia traffic and consequent consideration of what this entails, the significance of packet delay variation is beginning to emerge. Large variability in the transport time implies large buffers in the many end terminals, with corresponding impact on the overall cost of the system. In this light, the packet delay variation figures of table 1 are extremely promising, since they are an order of magnitude less than those encountered in even the best large latency-bandwidth product switching fabrics such as ATM.

Real-time applications have to accept that network information is lost and act accordingly. Delay variation rate smoothing buffers cannot be unduly large, even when cost is not an issue, as the additional delay introduced would be unacceptable. Given that a buffer size/additional delay choice has been made, those packets of information that arrive outside that delay will usually be treated as additional 'lost' packets. Given that optical and other communication techniques now deliver a 'true' loss rates in the 10^{-12} to 10^{-15}, loss due to other mechanisms such as out-of-time delivery dominate the application-perceived loss rate. The fraction of packets that will be delivered outside the acceptable envelope is a crucial design parameter in designing a system that is sensitive to these issues.

The packet delay variation in table 1 represents the period in which 99.9% of all the packets sent are delivered to the destination node. Although a real system would be designed around a delivery fraction one or two orders of magnitude smaller, the 10^{-3} level represents a figure that can be derived with reasonable accuracy from the simulation runs. One issue with simulation where the event of interest is rare is to get sufficient examples in a simulation run to ensure that the figure quoted has a reasonable statistical accuracy. As can be seen the packet delay variation is of the order of one or two packet transmission times, which is very low by store-and-forward switch standards.

We have studied the distribution of the 'tail' of this packet delay variation and discuss how significance levels of around 10^{-5} can be estimated from our simulation data in [6].

4 Further Optimisations

4.1 Improved quality of service control

With the transportation of multimedia data has come an increased interest in quality of the communication, as well as it quantity. This 'quality' has several distinct aspects, such as residual error rate and average packet latency. For real-time data an important property is variation in the delay that each packet experiences during its transportation. Simply, where this variation is large then large amounts of data buffering can be required by the end-host (with unfortunate consequences for the cost of the system) to ensure that the end-application can be delivered the data at the rate it requires. The buffering acts as rate smoothing that commutes the delay variation into latency (at least as far as the end-application is concerned). Although increasing latency does not impact uni-directional services (such as VoD) it is important in interactive applications - the very type of novel applications that it is hoped will act as the sales driver for the mass deployment of multimedia technology.

The traffic patterns and topologies that we have so far described posses their own uniformity. There is the assumption that the 'service' that is being utilised is the same for all the terminals. One important issue for the carrying of multimedia traffic is that, by definition, the traffic characteristics may be different for the different connections. In some of these cases the traffic will not require real-time delivery and it is to be hoped that this flexibility can be exploited. One possibility is to exploit the variety of paths available through a richly connected network to devote different paths to different traffic types.

Switching fabrics made from 1355 based switches have properties that make them well suited to building fabrics that can support many different traffic quality of service requirements. In traditional ATM switches there is a large computational load, or specialist hardware, to ensure that the real-time traffic is given the appropriate priorities within the fabric. One problem with these approaches is that once the end-host has launched its packets into the network there is little (if any) feedback as to their fate. The flow control and other feedback mechanisms happen over long periods, which is to be expected given the primary target market of *wide* area networking.

The minimal buffering and implicit low-level flow control of 1355, both in the point-to-point transmission and the cross bar switching has the effect of delivering rapid feedback to the end points. For example, given a maximum path length of 3, there can only be about two 64-byte packets in transit along a path. If the path becomes congested the feedback is rapid. This feedback is delivered to those end-points that are *contributing* to the congestion, unlike shared buffer switches where the feedback is to everyone, telling them that that the whole switch is congested. This certain knowledge that your traffic is contributing to the congestion, combined with the extremely rapid feedback (in the order of 15 microseconds) opens up new possibilities for control approaches, ones that can take the appropriate action for the particular traffic types.

4.2 Provision of Fault-tolerance

Multimedia data switching is becoming increasingly important for the provision of services, such as VoD, for which high availability as well as good performance scalability is a key requirement. Fortunately the use of IEEE 1355 switching fabrics permits a high degree of robustness in the system at moderate incremental cost. Using point-to-point connections rather than shared media enables any failed or misbehaving component to be isolated from the rest of the system. Further, using low-cost serial connections makes it straightforward to provide redundant connections on the interfaces, and using a fabric

constructed from (small numbers of) high valency switches makes it cost-effective to provide a 'hot standby' switching fabric, thereby avoiding single points of failure for the system. Of course, ensuring that the whole system is highly available requires more than this, but providing a robust switching infrastructure is important. Achieving this with either a bus-based system or with a complex switching system such as ATM would be a much more expensive proposition. A more detailed discussion of these issues can be found in [8].

5 Conclusions

Serial wormhole routing networks using lightweight IEEE 1355 protocols provide a scalable switching infrastructure for multimedia data. Our results have shown that a system with 512 terminals can be supported using only 48 switch devices, delivering constant streams of 15Mbits/s of user data to and from each terminal continuously under tight real-time constraints. Assuming a $200 unit cost of the switch chips, this translates to a cost per stream of less than $20 as far as the switching infrastructure is concerned. The authors know of no other technology which approaches this cost point, which is required for large-scale deployment of new services such as VoD. Furthermore, the use of symmetrical general-purpose switching elements gives considerable flexibility in the network topology, which can be exploited to further reduce costs when the constraints are looser or the traffic is more precisely characterised.

The use of point-to-point serial interconnect has considerable advantages, for robustness (leading to high availability), modularity and even the ability to upgrade and expand the switching fabric live in the field. Investment in the actual switching fabric is thus preserved, rather than just that in the line card interfaces as is typical in switching systems today.

Speculative simulations of such switching systems funded by the European Union (for which the authors wish to express their gratitude) have made it possible to identify promising application areas for this technology, thereby helping to overcome the inertia resisting its adoption.

6 References

[1] M.D. May *et al.*, Networks, Routers and Transputers: Components for Concurrent Machines. ISBN: 90 5199 129 0. IOS Press, Amsterdam, 1993.

[2] 1355-1995 IEEE Standard for Heterogeneous InterConnect (HIC) (Low-Cost, Low-Latency Scalable Serial Interconnect for Parallel System Construction). ISBN 1 55937 595 7. Institute of Electrical and Electronic Engineers, Inc, 1995.

[3] L. Ni and P. McKinley, A Survey of Wormhole Routing Techniques in Direct Networks, *IEEE Computer* **26** (2) (1993) 62-76

[4] P. Thompson and J.Lewis, The STC104 Packet Routing Chip, *Journal of VLSI Design* **2** (4) (1994) 305-314

[5] Z. Belkacem *et al.*, Rcube: Message Routing Device Using The OMI/HIC High Speed Link, Proceedings of the Third IEEE International Conference on Electronics Circuits and Systems, Rodos, Greece, October 1996.

[6] A. Jones *et al.*, The Network Designer's Guide. IOS Press, Amsterdam, 1997.

[7] S. Haas *et al.*, The Macrame 1024 Node Switching Network. In: B. Hertzberger and P. Sloot (eds.), High Performance Computing and Networking, LNCS, Springer 1997.

[8] P. Thompson, Globally Connected Fault-Tolerant Systems. In: J. Kerridge (ed.), Transputer and occam Research: New Directions. ISBN 90 5199 121 5. IOS Press, Amsterdam, 1993.