Chapter 1 _____

Introduction

High-Speed Networks (HSNs) present two fundamental challenges for candidate network architectures. First, the network architecture should be scaleable with transmission speeds. Transmissions in the optical domain can use efficient coding techniques and very high bandwidth rates not available in current electronic technology. If the architecture relies on electronic frame processing, it will introduce two main limitations on the maximum end-to-end throughput. The first limitation is intrinsic in the fact that bit processing by electronic gates is currently less efficient than transmissions of optically coded information. If optical signals are converted to electronics for processing, the maximal end-to-end network efficiency will hit the natural electronic bound instead of the desirable optical one. The second limitation is due to the fact that networks must service multiplexed streams generated and consumed by electronic devices at the network periphery. If the multiplexed streams are switched using the very same electronic technology in internal network devices, a small number sources operating at peak rate may congest the network. The ideal architecture should be prepared to operate at traffic volumes much larger than what an individual source may generate.

Second, motions must happen with strict Quality of Service (QoS) guarantees. Applications need guarantees in the end-to-end delay, jitter, bandwidth, etc. to provide their

services. The ideal network architecture should strictly assure the negotiated QoS parameters. Many times such assurance imply trading optimal resource use for optimal service. For example, in some circumstances it is preferable to waste bandwidth on communication channels that require strict QoS in order to guarantee minimal delay and jitter.

The goal of this dissertation is to develop, build, and analyze the novel Isochronets switching architecture that solves both challenges. The reminder of this chapter overviews the architecture and summarizes main contributions of this work. It can be skipped by those interested in reading the work in more detail, starting from Chapter 2.

1.1 Isochronets and Route Division Multiple Access (RDMA)

Isochronets are a novel switching architecture that *coordinates* traffic motions in the network over time, similar to how green waves coordinate traffic motions in roads. In the latter scenario, coordination is accomplished through synchronized traffic lights to create routes for uninterrupted and collision-free motions from source to destinations. In Isochronets, coordination is accomplished by enabling a set of routes by means of periodic configuration of network switches. Switches set up routing trees, that is, routes from all nodes to a given destination. A frame is switched by following its path through an enabled tree. The coordination mechanism is called *Route Division Multiple Access (RDMA)* because it divides network bandwidth among routing trees.

In RDMA, the basic construct used to schedule traffic motion is a time-band (*green-band*) assigned to a routing tree, as depicted in Figure 1.1. The green-band is a time interval during which network switches are configured to route according to a particular routing tree. During the green-band (shaded), a frame transmitted by a source will propagate

down the routing tree to the destination root. If no other traffic contends for the tree, it will move uninterrupted, as depicted by the straight line. Additionally, intermediate nodes perform no header processing. Multiple simultaneous routing trees can schedule transmissions in parallel (have simultaneous green bands), as long as they do not share any link in the same direction.

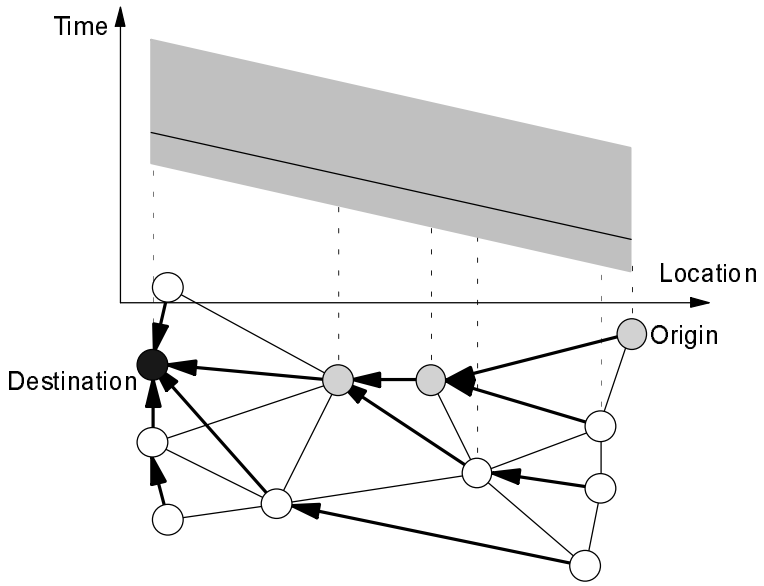

Figure 1.1: Green-band

Similar to Circuit Switching (CS), green-bands allocate reserved network resources. However, the units to which resources are allocated are neither point-to-point connections, nor traffic bursts, but routes. Routes represent long-lived entities and, thus, scheduling complexities can be resolved over time scales much longer than circuit slots. Additionally, routes are pre-allocated and not allocated on-demand like circuits.

Similar to Packet Switching (PS), traffic within a routing tree contends for outgoing links. However, contention is only among frames to the same final destination. Additionally, there is no contention for processing resources.

The allocation of synchronized time bands to routing trees and resolution of frame collisions are the primitive constructs used by RDMA to control traffic motions and QoS. *Contention bands* permit frame contention to happen for outgoing tree links. *Priority bands* are allocated to sources requiring absolute QoS guarantees, similar to a circuit service. Traffic from a priority source is given the right of way by switches on its path during its priority band. Unlike CS networks, however, priority sources do not own their bands. Contention traffic may access a priority band and utilize it whenever the priority source does not. During a *multicast band*, the routing tree is reversed and the root can broadcast to any subset of nodes.

Bands are allocated as portions of a fixed *clock cycle*. Every time the clock turns, the bands become available in sequence according to their position in the cycle.

The collision resolution during a contention band is designed in terms of signs "-", "+", and "++". In RDMA- only one frame proceeds, while the other colliding frames are discarded. In RDMA+, only one frame proceeds, but the others are buffered. Nevertheless, when a band finishes, all buffered frames are discarded. RDMA++ uses the same collision resolution technique as RDMA+, but stores frames beyond band termination, rescheduling them during the next band.

Isochronets present a few distinguishing characteristics:

- The band periodicity may vary with the type of traffic served. For example, traffic such as file transfers may use periods of long duration, whereas interactive voice or video traffic may use much shorter periods.
- All stack layers above the Medium Access Control (MAC) layer are delegated to

interfaces at the network periphery. Multiple protocol frames may coexist in the same network without adaptation.

- Interconnection of Isochronets can be accomplished via MAC layer bridges using extensions of current well-understood technologies.

It is interesting to compare RDMA with respect to PS and CS. If all the bands associated with a routing tree are priority bands, RDMA is operating on an optimized CS mode when the tree is enabled. That is, each source is allocated a circuit (priority band) to the tree root. The form of CS supported by Isochronets is superior to traditional CS in a few ways. First, circuits only get priority over band usage but do not own it. In a situation where the band has been allocated to a source serving real-time isochronous traffic, non-real-time data traffic may take advantage of underutilized parts of the band. Second, CS requires strict synchronization, usually per byte when implemented using Time Division Multiple Access (TDMA) [Halsall 92]. RDMA can offer the same services requiring synchronization per band that is usually much larger than a byte. Third, CS requires fast switching of bytes that can be difficult to accomplish when link speeds are gigabits or even terabits per second. RDMA needs only to switch routes independently of transmission speeds. Finally, CS requires a connection setup phase prior to communication that potentially can take longer than the transmissions in HSNs. RDMA can forward frames when the proper band starts, which depends on the clock cycle and not on the propagation delay in the network.

Consider now an Isochronet operating in RDMA++ contention resolution mode. If the entire band is allocated to contention traffic, frames moving down the tree will be stored

and forwarded as in an ordinary PS network. The form of PS supported by Isochronets is superior to traditional packet switching in a few ways. First, Isochronets support virtual cut-through mechanisms [Kermani 78] as frames arriving to a free switch will continue without store-and-forward delay. Second, there is no header processing in Isochronet switches. Third, contention happens only among frames to the same destination (and not among non-correlated traffic). Finally, RDMA can offer QoS to demanding sources.

In summary, at the two extremes of band allocations, Isochronets compare favorably with CS and PS networks. In-between, Isochronets can allocate combinations of priority and contention bands to produce a spectrum of switching techniques.

1.2 Protocols in Isochronets

1.2.1 The Problem

Isochronets need to solve three essential operational problems:

- How to allocate trees in the networks?
- How to allocate bands to trees?
- How to synchronize the bands at different nodes?

This section overviews each of these problems.

Tree allocation. In this problem, the objective is to allocate routing trees in the network so as to satisfy some optimization criteria. The exact criterion is dependent on the particular goals that the network is intended to satisfy. For example, trees can be allocated to minimize the distance between sources and destinations. This is particularly useful when links are operating at high speeds because the propagation delay is a major component in

the total end-to-end delay. Another example particular to Isochronets is to allocate trees such that the number of bands necessary to cover all possible destinations is minimized. This goal can be accomplished by allocating as many non-interfering trees as possible in the same band. In this case, the tree allocation mechanism should be geared toward choosing as many trees as possible that do not share links in the same direction.

Band synchronization. Given the band allocations in each node, the objective is to synchronize the bands, that is, assure that frames transmitted from a given node during band B_i reach the next node in the tree within the same band B_i. Synchronization mismatches in Isochronets may result in frames being misrouted. For example, if a frame transmitted during band B_i reaches an intermediate node during another band B_j, frames are forwarded from that point on toward B_j's destination. Another concern is to avoid bands intersecting each other. For example, if the intersection between bands B_i and B_j at some node is not empty, routing during the overlap period is ambiguous because incoming frames could potentially follow two separate paths.

Band allocation. Given a set of band transmission requirements, the goal is to allocate bands that will fit the requirements while being synchronized and not interfering with each other. The band allocation is thus concerned in finding two parameters for each band: initiation time within the clock cycle and size. The allocation changes over time given new demands computed from past history.

1.2.2 Main Results

Tree allocation. The proposed initial solution is an exhaustive search algorithm. It

begins by generating all spanning trees in the network. Then it builds all possible combinations of generated trees into maximal sets of non-interfering trees. That is, sets that contain only non-interfering trees such that no tree can be added without violating the non-interference constraint. Trees in the same maximal set can be assigned to the same band. Finally, the protocol finds the minimal number of maximal sets that covers all possible destinations. This number is also the number of bands in the cycle.

This exhaustive search solution is acceptable because the tree allocation protocol is executed only when the network configuration changes. The tree allocation protocol can be implemented off-line and execute at slow speed relative to transmissions.

Band synchronization. This problem is solved by manipulating the link propagation delays in the network. There is a device in each link that can be set to increase the link delay. The protocol computes the real link delay and then increases it using the device to make the final delay a multiple of the clock cycle period. Because of this manipulation, if something is sent at time *t* within the cycle from a given node, it reaches the next node still at time *t* but many cycles later. This fact trivializes synchronization because all that is necessary after link delays are set is to start and end each band exactly at the same moment within each cycle.

This protocol can operate off-line with transmissions. It is invoked periodically, depending on the accuracy desired.

Band allocation. For the purpose of this protocol, bands are always synchronized using the band synchronization protocol. The main concern of the band allocation protocol is then to size the bands. The task is pursued using a dynamic algorithm based on the Least

Recently Used (LRU) memory allocation policy. The protocol operates by shrinking band sizes for destinations that are not accessed frequently and then increasing band sizes of destinations that need higher demands.

This protocol may also operate off-line with transmissions. It may operate in a central location responsible for computing the allocation and then distributing it.

1.3 Protocols for Loosely-Synchronous Stacks

1.3.1 The Problem

Existing protocol stacks must be extended to support real-time traffic required by current applications. New real-time applications, such as multimedia exchanges, demand synchronous or isochronous services from the protocol stack. One important question is how to extend existing stacks in this direction while still providing current functionality. Ideally, the new services must be provided as extensions of current protocols to maintain compatibility with existing systems.

1.3.2 Main Results

This thesis presents two main contributions in the direction of providing real-time services. First, a novel *Loosely-synchronous Transfer Mode (LTM)* provides the functionality needed to support synchronous services to upper layers. Second, any protocol stack can be extended with synchronization signals from the LTM and become a *Synchronous Protocol Stack (SPS)*.

LTM. LTM is a novel transfer mode that issues *loosely-synchronous signals* to periphery nodes about current network status, that is, destinations that are reachable and as-

sociated QoS. An LTM network enables transmissions to some destinations during certain periodic time intervals or *bands*. Each band has associated QoS in terms of the bandwidth, end-to-end delay, jitter, and loss experienced in the communication.

LTM has many advantages:

- Sources and destinations are synchronized with the network through loosely-synchronous signals. That is, sources are synchronized with network status and not with frames or slots within frames as in Synchronous Transfer Mode (STM). Consequently, the synchronization does not need to be as accurate as in STM but still can attain the same sort of guaranteed service.
- QoS can be controlled and guaranteed because the loosely-synchronous signals coordinate how sources and the network must interact. Guaranteed performance is essential for time-sensitive applications such as multimedia conference.
- Multiple protocol frames can be transferred without fragmentation and reassembly because there is no pre-defined frame structure in the communication. This makes the extension of existing protocol stacks with the loosely-synchronous signals simple. Additionally, interconnection of multiple stacks is simplified.

RDMA in Isochronets is an example of an LTM.

SPS. The SPS is any stack extended with the loosely synchronized signals provided by the LTM. An SPS forwards LTM synchronization signals through its layers up to the applications. In the SPS:

- Existing protocol stacks are supported unchanged. For example, the Internet Protocol (IP) stack can be accommodated as follows. For sending, sources collect

frames and forward them when bands to the desired destinations are signaled. For receiving, the protocol works as in IP. The Service Access Points (SAPs) available at each IP layer can be supported unchanged.

- Existing stacks are extended with novel source and destination synchronization. For example, the IP stack can be enhanced with real-time service support. Real-time demanding frames are forwarded only when the network signals the required QoS to their destinations. The SAPs are extended with these novel services, but otherwise kept unchanged.

1.4 An Isochronet Switch Design and Implementation

1.4.1 The Problem

One of the problems addressed in this thesis is how to build an Isochronet switch (Isoswitch). The switch must operate at giga bit per second rates and ideally use only simple (that is, off-the-shelf) components.

1.4.2 Main Results

Electronic switch design and implementation. An electronic Isoswitch has been designed and implemented. It has four input and four output ports each operating at 1 Gb/s. The overall organization is depicted in Figure 1.2. The Host machine computes band allocation information off-line with respect to data transmission and reception. Sporadically, it stores the new configurations in a table inside the Control Unit. The Control Unit contains the arbitration logic that, based on the current configuration, sets up the switching fabric to provide the desired input and output connectivity. The configurations are changed over

time according to the band allocations. The Input Line Card receives bit-serial data from the network trunks and converts them into bit-parallel words that are routed to the proper output ports by the switching fabric. Similarly, the Output Line Cards converts bit-parallel words into bit-serial streams to be transmitted through the outgoing network trunks.

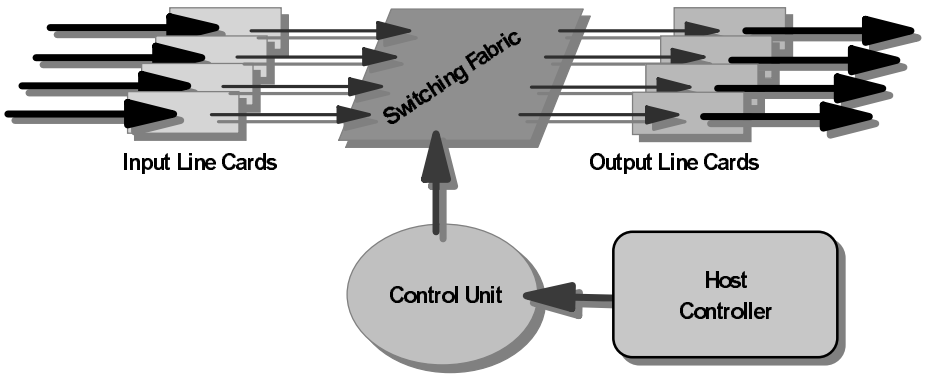

Figure 1.2: Isochronet switch organization

The design addressed the following challenges (where n is the number of input lines and m is the number of output lines):

- The design is scaleable with respect to number of channels and link speeds. The overall arbitration complexity is $O(nm)$ and its latency is $O(n)$.
- The implementation is easy to integrate with existing hardware. There is no adaptation among different protocol structures which makes the interface to other network devices simple.
- The Isoswitch can provide novel services. The first new service is the provision of synchronization signals that can be used at periphery nodes to schedule transmissions with required QoS. The second new service is direct frame forwarding without adaptation among different frame structures. The third new service is guaranteed (as opposed to statistical) QoS provision.

Interface card to a SPARC 1 machine. An interface card between the Isoswitch and a Sun SPARC 1 machine has been built. The card has a nominal throughput of 22 Mb/s. The card signals the beginning of band to the SPARC processor and can thus be directly used to implement LTM and SPS.

Optical switch design. A preliminary design of an all-optical switch implementation is proposed. An all-optical realization of Isochronets must avoid buffering at intermediate switches. The design proposed uses Wavelength Division Multiplexing (WDM) [Acampora and Karol 89, Brackett 91, Dono et al. 90] to allocate one wavelength for each band.

The implementation has a few advantages:

- No conversion of en-route signals to the electronic domain, which makes possible implementations potentially at hundreds of terabits per second.
- All routing trees are open at all times, thus contention band synchronization is not necessary.
- Fewer wavelengths are necessary than in pure WDM, namely one per band that, in the worst case, is the number of destinations in the network.

1.5 Performance Characterization

1.5.1 The Problem

The goal of the performance study is to evaluate the performance of RDMA when compared with more traditional switching techniques: PS and CS. The study is carried using both analytical and simulation models.

1.5.2 Main Results

Analytical approximations. The time-dependent behavior of RDMA complicates the performance study. Usual analytical performance analysis techniques ignore time-dependent (or transient) behaviors when simplifying models in search for tractable solutions. This study derived approximate closed-form formulas for end-to-end delay and loss in RDMA-, RDMA+, and RDMA++ for a single node and Poisson arrivals with mean rate λ. The analytical approximations are very accurate when checked with the simulation results, even for a network of nodes.

Under RDMA-, the whole routing tree can be seen as a server that transmits frames with sizes distributed according to some generic distribution $B(x)$ with mean rate μ. Whenever the server is transmitting, arriving frames are lost. If λ is the mean rate at which frames are arriving to the tree and $\tilde{z}$ is the inter-departure time for successfully transmitted frames. Then, the distribution of $\tilde{z}$ is:

$$\Pr\{\tilde{z} > z\} = e^{-\lambda z} - \int_0^z \overline{B}(z-x)\lambda e^{-\lambda x} dx \tag{1.1}$$

From this expression the mean loss rate can be computed as (where $\rho = \lambda/\mu$ is the tree utilization):

$$L = \frac{\rho}{\rho + 1} \tag{1.2}$$

One can see that as the load reaches 1, the maximum loss in the system is 50%.

Additionally, there is no queueing delay in RDMA- and the mean delay is equal to the mean transmission time ($1/\mu$):

$$D = \frac{1}{\mu} \tag{1.3}$$

Under RDMA++, the whole tree can again be seen as a server. It is active during the band (of size *U*) to the tree root and then goes on vacation (of size *V*) during other bands. From the point of view of an arriving frame when the system load is not very high, it needs to wait for the following events before being serviced. First, it needs to wait for the service of all frames in the queue upon its arrival plus the residual transmission time (or the time left to complete transmission) of the frame in service. Second, it needs to wait for the residual vacation time upon arrival. Finally, it needs to wait for the frame transmission. The formula that approximates the mean delay of a frame in RDMA++ is the sum of three expressions, where $\overline{x^2}$ is the second moment of the frame size distribution:

$$D = \frac{1}{2}\frac{\lambda \overline{x^2}}{(1-\rho)} + \frac{1}{2}\frac{V^2}{(U+V)}\frac{1}{(1-\rho)} + \frac{1}{\mu} \tag{1.4}$$

One can identify the first expression in the sum in the right hand size as the mean queueing delay for an M/G/1 system. The third expression is the mean transmission time by definition. The second may be interpreted as the probability of arriving during the vacation period ($V/(U+V)$) times the mean delay due to vacation ($V/2$) and the penalty due to the system load ($1/(1-\rho)$).

To be exact, the formula should additionally take into account vacations that occurred while servicing frames in the queue. The computation of such additional delay is not trivial because the expressions are not linear. Nonetheless, when the load in the system is low or medium, such added delay does not happen most of the time because all frames in the queue are serviced before the tree goes on vacation. The formula in (1.4) is thus a good

approximation for low to medium loads.

Finally, there is no loss in RDMA++ and:

$$L = 0 \tag{1.5}$$

RDMA+ mean delay can be approximated as the mean delay in M/G/1 if the band size is large compared to the mean frame size:

$$D = \frac{\lambda \overline{x^2}}{2(1-\rho)} + \frac{1}{\mu} \tag{1.6}$$

Finally, the loss in RDMA+ can be approximated using the mean number of frames in an M/G/1 system. This is also the mean number of losses per band that can then be approximated as:

$$L = \frac{\lambda \overline{x^2}}{2U(1-\rho)} + \frac{1}{\mu U} \tag{1.7}$$

When the band size is large, the loss tends to 0.

Simulation studies. The main goal of the simulation study is to compare RDMA with respect to PS and CS for Poisson arrivals and bursty on-off arrivals. Figure 1.3 depicts a typical behavior when the frame size is fixed at 53 bytes, the cycle size is 125 μs, the band size is 25 μs, and transmissions happen at 2.4 Gb/s. The load axis is the mean input frame rate and the delay axis indicates the mean frame delay in the system.

For low input loads, PS outperforms RDMA and CS because there is no admission delay waiting for a band or circuit. When PS needs to guarantee QoS, the performance advantage may vanish due to admission control mechanisms. When the load increases, the frame header processor becomes saturated and then RDMA outperforms PS. CS always performs worst than RDMA because the penalty incurred waiting for a circuit is larger

than the one waiting for a band.

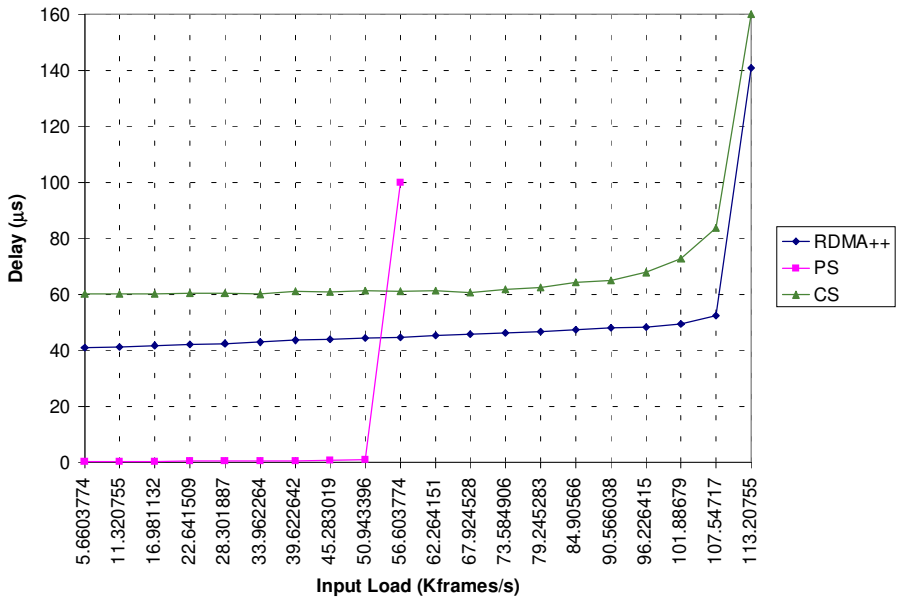

Figure 1.3: Mean frame delay comparison (in µs) for Poisson arrivals and deterministic service (the frame size is 53 bytes, the link speed 2.4 Gb/s, the band size 25 µs, the cycle period 125 µs)

1.6 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 analyzes other HSN architecture proposals with respect to a set of desirable characteristics for HSNs. Chapter 3 is a detailed account on the concepts in Isochronets and RDMA. Chapter 4 elaborates further on Isochronet protocols for tree allocation, band allocation, and band synchronization. Chapter 5 details LTM and SPS and overviews how existing stacks can be enhanced to provide real-time services. Chapter 6 explains the Isoswitch electronic design and implementation as well as the optical design. Chapter 7 uses analytical and simulation tools to study the performance of Isochronets and compare it with PS and CS. Chapter 8 suggests future work in Isochronets. Finally, Chapter 9 concludes.