A new DWDM core switch design

Minjie Dai

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Recommended Citation

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
A new DWDM core switch design

by

Minjie Dai

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Engineering

Program of Study Committee:
Steve Russell, Major Professor
James Davis
Clifford Bergman

Iowa State University

Ames, Iowa

2001
Graduate College
Iowa State University

This is to certify that the master’s thesis of

Minjie Dai

has met the thesis requirements of Iowa State University

Signatures have been redacted for privacy
To my lovely wife for all the support

To my parents for always believing in me
Acknowledgements

It has been like an adventure for me for the past two years. What I have gone through gives me strength and confidence in my future career.

Dr. Steve Russell first opened the door of opportunity for me. He guided me through the only proven path to success in life – hardworking, with trust and support. He has been helping me through teaching and advices on how to strengthen the weakness of my knowledge. His constant encouragement has made it possible to reach my professional goal. Dr. Ahmed Kamal opened my eyes to the fascinating world of network technologies. His deep understanding of various technologies motivated me to keep learning in this area. My committee members, Dr. James Davis and Dr. Clifford Bergman have been real supportive of my career. Their valuable advices helped me to always set a higher standard for myself.
Table of contents

Abstract vi

Chapter 1 Introduction 1

Chapter 2 Optical components in DWDM technology 5

Chapter 3 Local area WDM network design 9

Chapter 4 Regional network design 13

Chapter 5 DWDM WAN architecture and IP internetworking 15

Chapter 6 My new core optical switch design 30

Chapter 7 Conclusion 50

Chapter 8 Future research 51

Reference 52
Abstract

WDM (wavelength division multiplexing) network has been envisioned as the next generation network carrier technology. IP (internet protocol) is currently the dominant network layer technology and is believed to continue to be in the foreseen future. Although ATM and other protocols challenged furiously with IP in the past, IP networking has managed to make intelligent modification to accommodate various demands yet stay flexible and relative simple to implement. It has been strengthening its dominance in most area of competition. The seamless combination of upper level protocols with underlying WDM facility is the goal of future network design. My focus of research was on the core switch design for backbone networks. Its unique combination of cost structure and performance demand is the focus of design for real world applications. Trade offs were balanced among various options; discussion of choices of selections was also presented. To accommodate real time packet level switching, fast control and switching algorithms have to be deployed and hardware has to be able to reconfigure fast enough to service requests from control module. Therefore, fixed filters and wavelength converters were used instead of tunable ones to make it possible for faster reconfiguration for packet switching. There was no O/E/O conversion for the data; only control headers and routing information need to be converted to electronic form for analysis and computation. A new space division optical switch was also proposed to facilitate the exploitation of those components and to support desired functionalities. Multicast was supported since I believe it will be the vehicle for majority of the network traffic in the future. Synchronized network is also assumed to support more efficient use of bandwidth. The core router has to use control and switching algorithms that can overcome the bottleneck of computation to achieve high throughput, low delay and less storage, thus it can be cost effective yet support desired features. Decisions were made based on my knowledge of present and future network applications, and they intended to solve believed more important aspects of the problems. The success of this design can be tested by simulation in future effort.
Chapter 1 Introduction

The fast expanding internet-based services are driving the need for even more bandwidth from backbone and regional network. Most of today's data networks are supported using the IP and it is widely believed that IP will remain the dominant network layer technology. Practically all forms of end user communications today make use of the ubiquitous TCP/IP protocol. Furthermore, many new services and applications being offered are also based on IP protocols. Although many other protocols are also seeing strong demands for some of their unique design advantages over IP protocols, changes are made constantly to improve the ability of IP.

Recent advances in the WDM (wavelength division multiplexing) technology open the door to an almost unlimited source of bandwidth. By dividing wideband light spectrum into tens even hundreds of individual channels, each carry 10Gbs or higher bit rate, we can foresee the huge bandwidth carried by a single fiber, roughly 50 THz, waiting to be tapped into. By dividing the whole bandwidth into sub-channels we can overcome the limit of current electronic processing speed and transmission rate, namely 10Gbs or higher right now for single channel, also improve scalability and ease of upgrade.

There are currently a number of issues keeping IP and other upper layer network protocols from taking advantage of the huge capacity of the underlying transmission systems. For example, in today's internet, IP routers are connected using additional layers of electronic multiplexing (e.g. ATM, SONET) that were designed for supporting voice communication and are a part of the existing network infrastructure. These layers add complexity and produce unnecessary bandwidth inefficiencies. Furthermore, today's IP routers are difficult to scale to higher data rates due to their significant electronic processing requirements. While current developments in IP, such as hardware routing and IP switching, promise improvements for the future, they will still be limited by electronic processing from taking full advantage of the capacity of WDM transmission systems. One vision of future network is an all-optical network where all of the switching is done in the optical domain. At present, WDM deployment is mostly for point-to-point connection, statically assign wavelength channels to the upper applications and uses SONET/SDH as the standard layer for interfacing to the higher layers of the protocol stack. However, large-scale efforts are
underway to develop standards and products that will eliminate one or more of these intermediate layers and run IP directly over the WDM layer and **dynamically assign** wavelength channels to the upper level applications.

It is still not clear yet what putting IP directly over WDM really means. Figure 1 and 2 illustrate current implementations and proposed solutions for carrying IP traffic via WDM systems. Although the multitude of electronic layer services present in today's networks represent a significant overkill, they still serve certain functions that are presently not part of the IP or the WDM layers. For example, since IP transmits packets asynchronously, at the very least some framing mechanism is needed in order to transport these packets over WDM channels. Also, IP offers a connectionless service and is not designed to provide any service guarantees. While future versions of IP are likely to provide some form of service guarantees, today customers in need of such service guarantees typically use an ATM or Frame Relay connection with guaranteed service, on top of which they set up their IP network. Clearly, in designing an efficient and streamlined IP-over-WDM network it will be necessary to somehow preserve certain functions that are presently provided by the various electronic layers.

IP over WDM has been envisioned as the winning combination due to the ability of the IP to be the common revenue-generating convergence sub-layer and WDM as a bandwidth-rich transport sub-layer. Unfortunately, however, optical packet switching technology is not yet mature. Hence, a near term vision is to use electronic switches to connect between the different WDM channels. In that context, the goal is to jointly design the electronic and optical layers in order to optimize the network's performance, so duplication of the network functions at multiple layers can be eliminated as much as possible [2]-[18]. **One of the main bottlenecks in the present Internet is routing at the IP layer. Several methods have been proposed to alleviate this bottleneck by switching long duration flows at lower layers. One most promising example is MPLambdaS.**

There are also many issues such as quality of services, differentiated services and multicasting need to be considered in the new context of WDM infrastructure.
This thesis will first survey current research effort on combining IP service and WDM system to serve as the future network infrastructure. It briefly goes through the components, technology and design considerations of WDM system, then discusses the unique issues applying WDM to serve IP services and some of the new considerations in the new context. The discuss is not limited to WDM serving IP only, the underlying technology of WDM should also be able to adapted to new higher level of service stacks if necessary in the future. The WDM is still a not so mature system, a lot of issues need to be addressed in the context of given types services and network structure. Various important concerns still need to be addressed regarding IP/WDM integration. These include light-path routing coupled with tighter networking with IP routing and resource management protocols,
survivability provisioning, framing/monitoring solutions, and equally important, cost. Last part will present the proposed design and system consideration for the core optical switch in detail. The proposed design targets nature support of multicasting and all optical switching at packet level. The design takes into considerations of component characteristics imposed on system performance and by moving functionalities to other parts of the system to work around the bottlenecks.
Chapter 2  Optical components in DWDM technology

A. Components

System designer has to take into considerations of limits and properties of components [1] and not to over-simplify his protocols or algorithm and make unrealistic assumptions. In order to take advantage of full potential of different components to realize desired system performance, deep understanding of limits and capabilities of components are very important. Advances in the component technology which keep breaking the limits of current technology to further the freedom of system design. Here is the simple list of the common components in WDM system:

Optical Transmitters: semiconductor diode laser, tunable laser (mechanically tuned, acousto-optically tuned, electro-optically tuned, injection current tuned), fixed laser. Table 1 shows the characteristics of some common types of tunable lasers.

<table>
<thead>
<tr>
<th>Tunable Transmitter</th>
<th>Approx. Tuning Range (nm)</th>
<th>Tuning Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical (external cavity)</td>
<td>500</td>
<td>1-10 ms</td>
</tr>
<tr>
<td>Acousto-optic</td>
<td>83</td>
<td>10 μs</td>
</tr>
<tr>
<td>Electro-optic</td>
<td>7</td>
<td>1-10 ns (estimated)</td>
</tr>
<tr>
<td>Injection-Current (DFB and DBR)</td>
<td>10</td>
<td>1-10 ns</td>
</tr>
</tbody>
</table>

Table 1. Tunable Optical Transmitters and Their Associated Tuning Ranges and Times [1].

Fibers: multimode fiber, single mode fiber. In the most recent applications, single mode fiber is almost exclusively used for long haul networks for its much better properties. In the system design, the complications of attenuation, dispersion, and all of the nonlinear effects in fiber (nonlinear refraction, stimulated Raman scattering, stimulated Brillouin scattering, four wave mixing) will have to be considered to avoid pitfalls.

Couplers: splitter, PSC (passive star coupler).


Filters and Receivers: photo-detectors, tunable optical filters, fixed optical filters. Table 2 shows the characteristics of some common types of tunable filters.
Switching elements: two types: Relational, in which content of the inputs will not affect pre-decided route, has the property of data transparency and can accommodate different data rates and high data rate. Logic, in which contents of the inputs will decide the route, has the flexibility to change with the content of input but will have limited data rate imposed by speed of transition of the control logic states. Fiber cross-connect elements, non-reconfigurable wavelength router, reconfigurable wavelength routing switch and photonic packet switches are commonly used for different systems.

Wavelength converters: In order to eliminate the restriction of wavelength continuity constraint, wavelength converters are used to use wavelength channels more efficiently. Types include using coherent effects of FWM (four wavelength mixing), DGF (difference frequency generation), also using cross modulation, such as SOA in XGM (cross gain modulation) and XPM mode and semiconductor lasers as shown in figure 3 and figure 4.

Table 2. Tunable Optical Filters and Their Associated Tuning Ranges and Times [1].

<table>
<thead>
<tr>
<th>Tunable Receiver</th>
<th>Approx. Tuning Range (nm)</th>
<th>Tuning Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fabry-Perot</td>
<td>500</td>
<td>1-10 ms</td>
</tr>
<tr>
<td>Acousto-optic</td>
<td>250</td>
<td>10 µs</td>
</tr>
<tr>
<td>Electro-optic</td>
<td>16</td>
<td>1-10 ns</td>
</tr>
<tr>
<td>LC Fabry-Perot</td>
<td>30</td>
<td>0.5-10 µs</td>
</tr>
</tbody>
</table>

Figure 3. A Wavelength converter using co-propagation based on XGM in an SOA [1].
B. System considerations

A basic WDM system is considered to be composed of N light sources sending information simultaneously to N receivers over an interconnecting fiber system. Each source is assigned a separate wavelength band within a specified tuning range. Tuned lasers are used as the optical source for each transmitter, and then data is modulated within its assigned band for transmission. The receiver isolates a transmitter by using some type of de-multiplexer to separate bands into individual channel to be detected or switched. This can be accomplished by the use of tunable optical filters or prism followed by photon detectors and decoders or optical space switch. The major disadvantage in optical WDM is that it requires fairly expensive and sophisticated hardware. Depends on the system design, wavelength-controlled tunable lasers and high-quality narrow band tunable filters may be required for each channel. The tuned filter at the receivers must be able to tune over the entire range to select a specific band. The most popular types of tunable filters being used are Fabry-Perot and Mach-Zehnder interferometer filters. Other types of optical tuning filters use acoustic-optic and electro-optic mode coupling, or can be obtained from narrow band laser amplifiers.

It has been realized that WDM has network applications beyond the simple increase of link capacity. When bit rate is getting higher and higher, it's harder and harder to increase the rate with time-division multiplexing and it's easier to increase capacity with WDM. The key feature of dense WDM is that the discrete wavelengths form an orthogonal set of carriers which can be separated, routed, and switched without interfering with each other, as long as the total light intensity is kept sufficiently low. It is this use of wavelength and processing in passive network elements that distinguish optical networks from other network technology.
The two general architectural forms that have been most commonly used in WDM networks are wavelength routing networks and broadcast-and-select networks, depends on the cost and applications of the system. Wavelength routing networks are composed of one or more wavelength selective elements and have the property that the path that the signal takes through the network is uniquely determined by the wavelength of the signal and the port through which it enters the network. In broad-and-cast network, all inputs are combined in a star coupler and broadcast to all outputs. It is more suitable to low cost environment, such as LAN. Several different schemes exist, depending on whether the input lasers, the output receivers, or both are made tunable.

There are also two types of wavelength switching. First one dynamically switches signals from one path to another by changing the WDM routing in the network. This type of switches can select an arbitrary subset of the wavelengths on one fiber and redirect them to another fiber. The selection is rearrangeable, so the subset selected can be changed whenever desired, and the larger the subset which can be selected simultaneously, the better statistical multiplexing it can achieve. It is also important to have the switching action fast enough to meet the application requirements. The second one is really wavelength conversion, where the information on a signal is effectively transferred from an optical carrier at one wavelength to another wavelength. This can be easily done by detecting the first signal and using the detected current to modulate a laser at the desired second wavelength. By using a tunable laser for the second wavelength, it is possible to switch information from one wavelength to any of the set of output wavelengths.
Chapter 3  Local area WDM network design

A. Overview of LAN

The challenge in LAN system design is that it has to minimize the cost for individual user yet provides enough bandwidth for them. That is one of the reasons that most of the LAN designs now choose broadcast and select architecture, in the form of passive star or passive bus shown in figure 5. Realizing that the maximum rate at which each user can access the network is limited by electronic speed (to a few gigabits per second), the key in designing light-wave networks in order to exploit the huge bandwidth is to introduce concurrency among multiple user transmissions into the network architectures and protocols. In an all-optical network, concurrency may be provided according to wavelength or frequency (wavelength division multiple access--WDMA), time slots (time division multiple access--TDMA), or wave shape (spread spectrum or code division multiple access--CDMA). However, the basic needs to have nodes synchronized to within one time slot (for TDMA) and one chip time (CDMA) make them less attractive than WDMA. By adopting the WDM, end users only need to operate at the bit rate of certain WDM channel.

In WDM LANs, the physical topology is the one seen by the optical layer. It consists of passive or configurable optical nodes interconnect to each other via optical fiber. The most common physical topology is passive star because of its logarithmic splitting loss in the coupler although other topology such as linear bus or a tree can also be used. LANs can be built based on single-hop approach or multi-hop approach. From the performance perspective, both approaches are currently equally attractive.

B. Single-hop WDM system

In a single-hop system, there are no intermediate nodes; each node must communicate with one another directly, so a significant amount of dynamic coordination between nodes is required. For a packet transmission to occur, one of the transmitters of the sending node and one of the receivers of the destination node must be tuned to the same wavelength for the duration of the packet's transmission. In a single-hop environment, it is also important that transmitters and receivers be able to tune to different channels quickly, so those packets may be sent or received in quick succession. For a single-hop system to be efficient, the bandwidth allocation among the contending nodes must be dynamically managed.
C. Multi-hop DWDM system

In multi-hop system, the channel to which a node's transmitter or receiver is to be tuned is relatively static, and this assignment normally is not expected to change except when a new global assignment of all transceivers is deemed to be more beneficial. It is unlikely that there will be a direct path between every node pair so that, in general, a packet from a source to a destination may have to hop through zero or more intermediate nodes. Different virtual structures will have different operational features and different performance characteristics.

![Figure 5 a) A passive star topology and (b) A passive bus topology.](image)

D. Network survivability issues

For designed network to be robust, establishment of network protection strategy and operation administration and management technologies is essential. There have been many papers addressing the problem and provided several different solutions.

E. Comparisons of proposed resource reservation protocols

There are a number of reservation protocols have been proposed to address resource reservation issue under single-hop WDM network. Access to the various channels of a single-hop network is usually based on reservation schemes that require the use of control channels. Existing protocols require that control information be transmitted on the control channel for each packet sent on the data channels. Typically, TDMA is employed in the
control channel with a control slot consisting of N mini-slots, one for each of the N nodes in the network. Now we compare some of the protocols.

In **tell-and-go** protocols, the data packet is sent on the node's home channel immediately after the transmission of the corresponding control information. Thus, receiver collisions may arise and explicit acknowledgments are needed.

In **tell-and-wait** protocols [22], nodes send the control information and wait for the control slot to reach all receivers. Then they process the information in the control slot to determine if a data slot has been reserved for them. In the event of a successful reservation, the packet is transmitted in the corresponding slot and channel. Individual node uses the control slot information in **tell-and-wait** protocols to build a picture of the packet queues at all other nodes in network.

The above protocols suffer from two problems. First, the control channel represents an electronic processing bottleneck as control information for N packets must be received and processed for each packet transmission and reception. Secondly, all these control channel protocols operate by scheduling a single packet from each transmitter at a time. This packet is scheduled independently of other packets waiting for transmission at the same node. Hence, depending on the protocol, one transmitter or receiver tuning time is incurred for each packet transmission/reception.

The PROTON protocol tried to overcome the processing bottleneck introduce more than one control channels. The drawback is its lack of scalability, as they require N+k wavelength (k is the number of control channels). It can operate with any number of wavelengths, and its design explicitly considers tuning and processing times. However, it schedules one packet at a time, and the high processing and tuning time have a significant effect on delay and throughput. The MaTPi protocol also considers tuning times, and uses pipelining to mask their effect.

DQMW (distributed queue multiple wavelengths) protocol [23] can also operate with any number of wavelengths and considers tuning times when scheduling packets. DQMW attempts to overcome the head-of-line blocking of other media access schemes by considering multiple packets for transmission by a given node. But these packets are still scheduled independently of each other; thus a tuning overhead is incurred for each. In
addition, this protocol has higher processing requirements compared to other protocols, as two control packets must be sent for each data packet: one before its transmission and one after the end of its transmission.

FatMAC is a reservation protocol that does not require a separate control channel. Instead, all channels operate in cycles, with each cycle consisting of a control and data phase. Reservations are transmitted in the control phase and the corresponding data packets are sent in the following data phase. As in other protocols, reservations are made only for the head-of-line packets, thus a control and tuning overhead is incurred for each packet.

HiPeR-I protocol [23] claim it can operate with any number of channels less than $N$ and can operate without the control channel, thus all channels are available for data transmission and no extra hardware is needed to monitor and access a control channel. It requires tuning at one end to make sure the packet transmission is free of channel and receiver collisions. It schedules multiple packets for transmission by a node on a given channel to mask the tuning latency and to keep the control requirements low. It also uses pipelining to overlap processing with packet transmissions and hide the effects of propagation delay.

This area of research is heavily dependent on the higher lever service requirements and technology feasibility to support them and need to have different schemes to exploit the trade off among many factors to be considered.
Chapter 4 Regional network design

While the growth of the backbone communication capacity has been tremendous, end user access to this capacity is still expensive and limited to data rates of kilobits and megabits per second. The WDM technology in MAN can be a powerful approach and is currently being actively explored.

The challenge in developing the physical architecture of the WDM network is to provide intelligent network functionality at the optical layer, as well as to allow optical and electronic switching layers to operate in synergy. One of proposed MAN architecture is shown in figure 6.

![Physical architecture of a proposed WDM optical metropolitan access network](image)

Figure 6. Physical architecture of a proposed WDM optical metropolitan access network [36].

The access network [36] is divided hierarchically into a feeder network and multiple distribution networks. End users are locally connected to the distribution networks, which in turn are connected to access nodes in the feeder network. In addition to connecting access nodes to one another, the feeder network has one or more connections to the communication backbone via backbone access nodes. A ring physical topology is proposed as the feeder
network although other topologies are also possible. The WDM feeder ring carries multiple wavelength channels, and different access nodes are connected by different wavelength. In this optical WDM, both optical and electronic intelligent network functionality reside in the access nodes, which interconnect data flows between the feeder and distribution networks. The function for the feeder access node is to route full optical wavelength channels and individual IP data packets inside wavelength channels toward their destinations. There have been great effort put into the research of cost effective traffic grooming in this context to balance the cost yet deliver satisfying bandwidth to access network.
Chapter 5  DWDM WAN Architecture and IP internetworking integration

A. Overview of current application and research direction

In the days of low traffic volume between IP routers, bandwidth partitions over a common interface make it attractive to carry IP over a frame relay and/or an ATM backbone. The fast expanding internet-based services are driving the need for even more bandwidth from backbone and regional network.

The traditional technique for increasing capacity has been to deploy more fibers and replace SONET TDM (time-division multiplexing) systems with new higher rate TDM systems. Since deploying new fibers can be extremely expensive and deploying higher rate TDM systems requires an inflexible replacement of the whole existing operational system, many network planners are turning to WDM transport systems as their mechanism for cost-effective flexible capacity expansion. In TDM system such as SONET, capacity scaling is achieved by increasing the rate of transmission. Capacity scaling can also be done by transmitting multiple TDM signals, each over different wavelengths, on the same fiber with WDM. For existing Internet network providers, WDM maximizes reuse and minimizes lifecycle cost of existing fiber facilities. Fiber layout between major switches hundreds miles away is the most expensive resource for the network carriers and should be used as efficient as possible. With WDM transport, capacity growth between regional core router sites can be easily be matched to actual demand while with TDM transport interfaces, router link capacity upgrades must take place in inflexible multiples. Using WDM to interconnect core router sites also has the added advantage that multiple interfaces mapping technologies can be transported on same fiber. It will be important because it allow the embedded base to still be used and at the same time allow flexibility in the evolution of the backbone network evolution in supporting a variety of options of data transport, over HDLC, over ATM, or some future yet-to-be-determined link protocol.

It is claimed that as technology involves, we will have more and cheaper bandwidth, which will make dynamic switching to use the bandwidth more efficiently out of favor. According to one of the study [14], the cost of computation to improve the bandwidth usage will still be much less than deploying more bandwidth for at least ten years assuming the current trend holds. Figure 7 shows the current trend and future prediction. So for current
developments, dynamic packet switching will still be in the center stage until we have cheaper enough bandwidth compare to the cost of computation and we can go back to optical circuit switching. Besides, the limit of Shannon limit will probably put a stop to the trend some time around 2008 and the drop of cost for deploying optical capacity will slow and we might not ever see the cross of the two cost lines. As we have learned in old electronic circuit switching system, typical data interaction has a peak rate of 15 times of their average rate and current packet network is approaching this ratio.

![Figure 7. Trends in packet switching computer cost and communications cost [14].](image)

*Increasingly, real-time services such as voice and video are likely to be transported on IP networks. In such cases, both QoS and fast restoration under failure will emerge as central considerations in the operation of future IP networks. Multicasting capability is also a desired feature for many applications to be able to scale up to more users yet conserve the bandwidth efficiently. As optical networking elements such as multiplexers and crossconnects are developed and deployed, sub-second restoration at the optical layer may*
become feasible. This will allow the IP routers to concentrate on QoS differentiation and multi-service integration issues.

Current applications of WDM focus on the static use of individual WDM channels, which may not support IP traffic efficiently. The challenge now is to combine the advantages of WDM with emerging all-optical switching capabilities to build high-throughput optical routers. *Systems have been developed so far are trying to bridge the gap of fast transmission and slow processing.*

Problems facing WDM now are all-optical synchronization, buffering, and node cascadability. To effectively develop an IP router capable of providing a full optical data path at speeds approaching the limits of current electronic devices, namely 10 Gb/s and higher, a transfer mode has to be adopted robust and efficient enough to overcome these technological limitations.

Figure 8 shows the foreseen general trend in future development of IP over WDM and its possible improvement [3].

Figure 8. Possible future trends in optical network technology [3].
B. Optical IP packet switching in high-speed backbone

The design considerations for the backbone switches are different from LAN or MAN system design. Although WDM bring almost unlimited bandwidth to the network, the deployment of multiple fibers to provide static circuit like channels are and will be more expensive than use statistically multiplexed packet switching network, although it needs extensive and complex control computation.

One of the main bottlenecks in present Internet is routing at the IP level. Several methods have been proposed to alleviate this bottleneck by switching long duration flows at lower layers. What really boil down to are the tradeoffs of circuit switching, currently represented by MPLambdaS, and traditional packet switching, in which decisions are made at each router. Tag switching schemes use routing protocols to predefine routes within the network and assign tags to the routes as the virtual-circuit. Packets are then switched based on these tags avoiding the need for routing table lookups. IP switching dynamically sets up layer-2 virtual circuits for those connections that are perceived to be long. The challenge now is to combine the advantages of WDM techniques focused on the static usage of individual wavelength channels with emerging all-optical switching capabilities to yield a high throughput optical platform directly underpinning next generation data networks. Shifting most of the switching burden into the optical domain permits the successful scaling of IP router throughput to WDM transmission. In doing so, an effective decoupling between the bandwidth and routing issues is also achieved.

Currently, many schemes have been developed to tap into the huge bandwidth WDM offer, yet to overcome the limit of electronic processing bottleneck. These schemes use combinations of many available technologies to find the best tradeoff to best serve the future development of WDM backbone services.

There are generally two methods for transmitting the control header in current development. One uses in-band header, in which data packet follows control header in the same wavelength channel, the other uses out-band control in which data packet and control header are transmitted in separate channels. They both have pros and cons, and I generalized some of them as follow:

Using a separate channel for carrying headers has a number of advantages:
1. More transparent of data rate. When data rate in the data channel increase, it will not affect the control channel bit rate. It will be scaled according to the processing speed of switching node, and as long as this is not changed, the edge router can upgrade the transmission rate at any time and not affecting the switching node at all. So the continuous system upgrading is easy. Although in-band header can have different bit rate than the data part, it nevertheless demand more complex controls.

2. Less challenge for system design point of view. When coding in different rate for header and data in the same channel, laser sources either have to be switched between different speeds frequently and the time between switches will not be used for transmission and wasted. Although separate lasers can be used for different rate, it will invoke higher system cost.

3. More efficient usage of the bandwidth. In the in-band control scheme, Packets have to be delayed for the time of processing the header, which will involve delay line which is costly since the only practical option now is using long fiber delay line. The bandwidth waste will be aggravated when the speed of data rate increase at a faster rate than the processing speed increase. While by using out-band header, headers can be sent out in a controlled time ahead of the transmission of data part to compensate the processing of header so data packets can cut though switch without being delayed. By using out-band scheme, excessive delay in the intermediate node can be avoided. Delay is realized at edge router instead of core switches, which can lead to using of inexpensive electronic storage instead of much more expensive fiber delay line in core switch, so simpler design for the core switches. DiffServ (differentiated services) can also be deployed by using different delay time between header and data packet. Control headers will be sent out to intermediate router or switches for processing ahead of the data packets. If we adopt the simple first come first serve scheduling at the switches, then start sending the header for the high priority packets at a much earlier time has the effect of giving them higher priority since they will take up the spots earlier and prevent low priority packet from taking the positions. The amount of time for different priority header to be sent out can be adjusted to realize the degree of differentiation [50].
4. Optical synchronization can be simpler, since in the in-band scheme, synchronization has to be done both at the input to recover header and output to put header back to the packet.

But there are also a number of drawbacks limiting its applications.

1. Since control header has to be sent out ahead of the real data packet, it has to include the exact time and the corresponding channel data packet will be sent out. This will limit the options control module can do since rescheduling after the control header is sent out is not an option. The implication is that the edge router will have most of intelligence and do most the scheduling and core router only need to exchange route definition information and based on the control header to do predefined routing.

2. If FDL buffers are used to resolve conflict, the control algorithm will be required to be able to know if packets will be switched to an output link or switched to a buffer and it will be scheduled for future transmission. It adds difficulties for the scheduling algorithm, since it adds the uncertainty when the data packet can actually be sent out which is needed for sending out the control header. Decide the offset time can be challenging, considering at edge router, all possible events have to be taken into account to find the balance, If offset time is too long, it will increase delay and require more storage buffer at edge router, if offset time is too short, packets might be dropped due to not enough time for processing before data packet’s arrival, or intermediate routers have to delay the packet to compensate it.

Although in-band control method will likely incur more processing delay and other unwanted system complications, it nevertheless holds some of the advantages.

1. More options of control scheduling can be deployed. Since the control headers are stripped off when the packets come and then processed, control system know exactly which channels are in use and can use any algorithm to schedule transmission.

2. There is no early arrival dropping of packets, and since inlet delay line can be used, it can be scaled according to the processing speed of the switch.

3. Packet synchronization can be done more naturally since detection of the header has to be done for headers to be processed.
Overall I think using separate channels for control purpose seems to be more promising and appropriate in the future although both schemes can be used.

Because of the cost and bandwidth it can provide, optical routers should be used more of core routers in Internet backbones. The inlets and outlets of a core router are optical fibers, which operate in DWDM mode with a large number of wavelengths per fiber depending on the technology. More wavelengths per fiber could yield a higher degree of statistical multiplexing and lower blocking probability if the router design takes advantage of it. The transmission speed per wavelength can be very high depending on the technology available. The functions of the core router are the same as the normal routers, namely forwarding and switching packets and running routing protocols.

The core routers are supposed to exploit optical technology to reach a throughput on the order of multiple terabits per second and beyond. However, the processing of predominantly short IP packets becomes a problem at such high speeds. Therefore, it is necessary to introduce some lower limit on the granularity of switched data units in the backbone. To this end, the burst switching [48] transfer mode paradigm has been proposed. Under this technique, the Internet backbone architecture can be viewed as figure 9. Incoming packets are collected into bursts in the edge routers and sent through the backbone network where core routers will perform forwarding according to the egress destination in the burst switching cloud. There are various stages may be implemented by means of different technologies and with different architectures, but the basic functions to be performed are those shown. In this optical backbone, a burst travels through an all-optic path between source and destination edge routers. During the travel in the core network, no optical to electrical conversion of data burst is performed, this task is achieved by means of an all optical switching matrix based on the use of wavelength converters and semiconductor optical space switches. However, the control is still performed electronically (figure 9). A burst control packet (BCP), corresponding to an optical data burst on a data channel group (DCP), is sent on a separate control channel, or sometime before the data packet on the same channel, terminated electronically at each hop. In the first case, a set of DCP on each fiber is reserved exclusively for this purpose. Depending on the opto-electronic conversion complexity, the control and data channel bit rates may differ. In the control path, BCPs are
processed at each core and ingress edge router to reserve a time slot on the data and control channels for the transmission of both the data burst and the BCP in a synchronized manner. There are many different proposed protocols for control and reservation scheme. There are many design issues still need to be addressed in this scheme in terms of performance such as burst delay and burst loss probability. Compare to the conventional packet switching, the optical burst packet switching system has two different characters. First one is that the buffer is realized by means of fiber delay lines, therefore, the storage capacity and holding time in the buffer are related. The second one is that each burst must pass through the optical switching matrix via a pre-established optical path. Therefore, processing on the BCP and related forwarding decisions must be performed before the burst enters the matrix.

![Illustration of burst packets traveling in core network](image)

Figure 9. Illustration of burst packets traveling in core network [48].

It is also conceivable that separate network for routing information should be used or at least separate data channel should be dedicated to that to avoid the complexities of dealing with data and routing information at the same time.

C. Virtual topology and routing and wavelength assignment issues

Routing and wavelength assignment (RWA) issues is intended to efficiently manage WDM network resource utilization, many different constraints serve to complicate this
process, examples being wavelength-continuity requirements, wavelength conversion limitations, analog attenuation effects, etc. Many authors have proposed algorithms to provide optimum solution based on variable constraints. For those schemes [31-34], they usually require global knowledge of the network and more amenable to centralized implementation. The computation time is significant and delay will be long for decision to propagate and take effect, so they may not be feasible for the practical applications. Some of the dynamic and distributed schemes are proposed to reduce the computation burden, but will inevitably incur higher blocking rates. Currently, there is no standard yet for RWA problems. It is very likely that some hybrid schemes will be used that computation intensive centralized schemes used for long-term planning and dynamic distributed schemes used for short-term decision to adjust to the fast changes in the network.

The topology design, which is usually done at the higher protocol level, currently has no clear strategies for WDM network as they are tied directly to the whole network design goal.

D. Network survivability

Network survivability can be classified into two forms, protection and restoration, like in any LAN or MAN. The differences in the designs of WAN from other scale of network require different considerations and schemes to provide the survivability in the framework of the integrated system design.

Protection schemes are referred to those pre-provisioned failure recovery methods and usually faster than restoration schemes. Backup channels are provided and to reduce the redundancy, they can be shared. Shared channels design is the trade-off of reduced channel recovery possibility and lower cost of system. Most current networks used modified SONET like schemes to suit specialized topology of the given networks.

Restoration schemes, based on signaled recoveries, do not require pre-routed backup channels. They will dynamically recomputed and select new routes to solve the network problems. It is considered to be self-healing process and provides a more “medium-level” backup capability. It is no surprise that the computation involved to dynamically decide the routes will be much slower than protection schemes and also when distributed schemes are used, need time for all the nodes to converge and routes might not be optimum.
I will not consider above issues in my design and instead will focus on the internal node design issue to provide enough modification capability to meet many possible strategies exploitation in the future integration. I will assume that topology design, routing and optimization are done through exchanging network information over the control channels in some way, same as network survivability issues, and switching decisions will be made based on that information by the control module in the switch. Many strategies [27-30] have been proposed to provide better survivability for the future networks.

E. Contention resolution for optical switches

Various contention resolution techniques can be used, optical buffering, wavelength conversion, multi-path routing. The performance, measured in terms of packet loss probability, has been used to compare different schemes.

It is well known that Internet traffic shows self-similar property, so leads to fluctuation and burst-rate condition. This makes circuit switching much less attractive than cell/packet switching, because statistical multiplexing in cell/packet switching can use bandwidth more efficiently. But, the trade off for the performance is the lack of guarantee of service and best effort nature, which is compounded in very bursty conditions. Numerous methods have been proposed and used in current network to add more quality of service to the network yet maintain the benefit of the statistical multiplexing. It looks right now although optical buffer might be expensive to use, it is probably will be necessary to a certain degree considering the bursty nature of the traffic. Traffic balancing and policing can provide relief to the burden when we have deeper understanding of the traffic characters. With the adoption of wavelength dimension multiplexing, we can use much fewer and shorter depth of buffers which is the goal of my design.

F. QoS for IP over WDM networks

Quality of service (QoS) is measured in many factors, such as packet loss rate, end to end delay, delay jitters and so on. Scalable end-to-end QoS in DWDM networks with large numbers of wavelengths penetrate the data centric backbone networks, Specific IP service requirements such as priority restoration, scalability, dynamic provisioning of capacity and
routes, and support for coarse-grain QoS capabilities will have to be addressed in the optical domain in order to achieve end-to-end QoS over a DWDM network. IP QoS model that has been studied extensively in the literature is integrated services (Intserv). It is based on the definition of several traffic classes that, if supported by the network routers and switches traversed by IP flows, can deliver packets to destination with certain QoS commitments. Intserv requires a signaling mechanism, such as the RSVP (resource reservation protocol) to reserve network resources along the flow path. With Inserv, each packet in the network is processed by the router to determine its service class and scheduled accordingly. In large IP networks, the processing and policing of individual packets to ensure QoS impose a computation burden on the packet-forwarding engine that limits the scalability of Inserv. The DiffServ (differentiated services) model was introduced to deal with the scalability issues in the Inserv service model. In DiffServ, scalability is achieved by aggregating packets with the same QoS requirements into fewer but coarser-grained flows. This significantly reduces the computational burden since the packet forwarding engine has to police fewer coarse-grained flows. DiffServ flows are enforced locally on per-hop basis, further simplifying the complexity of end-to-end QoS policing mechanisms. A further aggregation of DiffServ flows into coarser-grained flows can be directly mapped into optical channels at optical channels. The mapping may be performed in such a way that the optical quality of the lightpath matches that of the aggregate DiffServ flows. The major design issue of the DWDM optical QoS model is scalability, especially at gigabit data rates. This implies that a heavy computation burden must be overcome if the DWDM edge device has to process, track, and maintain the QoS state of each flow of aggregated traffic in the optical domain.

One approach proposed [38] to apply DiffSer approach to the DWDM environment is shown as the figures 11 and 12 below. Classification is based on the quality of wavelength channels and packets are carried in corresponding channels. This model bears the similarity of DiffServ model and develops efficient support of IP differentiated services in DWDM network.
G. Real time service on WDM networks

One of the issues IP service needs to address and support in the future is the real time service. Although some schemes have been developed for IP, this has to be readdressed in the WDM optical domain. ATM, FDDI and other networks provide real-time service to messages with time constraint. If WDM wants to serve different upper level services, it also need to take into the consideration of implementing real time service in this optical network which will put some restrain on the protocol design.

H. WDM multicasting in IP over WDM networks

a) Overview of IP multicasting over WDM

<table>
<thead>
<tr>
<th>Classification Criteria</th>
<th>Class 1 $(\lambda_1, \lambda_2)$</th>
<th>Class 2 $(\lambda_3, \lambda_4)$</th>
<th>Class 3 $(\lambda_5, \lambda_6)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BER</td>
<td>$10^{-9}$</td>
<td>$10^{-7}$</td>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>Survivability</td>
<td>90%</td>
<td>70%</td>
<td>20%</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Intrusive</td>
<td>Intrusive</td>
<td>Nonintrusive</td>
</tr>
<tr>
<td>Security</td>
<td>Secure</td>
<td>Unsecure</td>
<td>Unsecure</td>
</tr>
<tr>
<td>Processing</td>
<td>1R</td>
<td>2R</td>
<td>3R</td>
</tr>
</tbody>
</table>
Supporting WDM multicasting in and IP over WDM network poses interesting problems. The number of wavelengths available in a WDM network continues to be a major constraint, even in DWDM. As a result, such a network has to use and reuse wavelengths efficiently in order to support a large number of end users. The design of multicast algorithms has to take this unique situation into the consideration. Effort has been given to derive the asymptotic upper bounds on the number of wavelengths needed to support multicasting capability in all-optical network and this can serve as the limit on scalability of all-optical networks that support multicasting. The result is based on the assumption that the switches are capable of multicast so a signal coming to an input port of a switch can be copied to more than one output port of the switch, also the switch can be independently configured for different wavelength channels.

Multicast sessions may run IP over WDM networks by letting each router on a multicast tree make copies of a data packet and transmit a copy to each downstream router. This requires Optical-Electrical-Optical (O/E/O) conversions of every data packet at all the routers on the tree, which may be inefficient and undesirable. Such O/E/O conversion may be avoided by using a virtual topology consisting of light-paths from the multicast source to each destination of the multicast session. However, for large multicast groups, the network bandwidth consumed by such a scheme may become unacceptable because of the unicasting nature of the light-paths. The true multicasting at the WDM layer by taking advantage of the power splitting capability of the WDM switches is desirable. WDM multicasting has several potential advantages. First, with the knowledge of the physical topology (WDM layer), which may differ from what is seen at the upper electronic layer (IP), more efficient multicast trees can be constructed at the WDM layer. Secondly, some WDM switches use power-splitting components and power splitting is more efficient than copying (by IP layer) for multicasting purposes. Finally, multicasting at the WDM layer provides a higher degree of data transparency in terms of bit rate and coding format.

b) WDM switch architectures for multicasting

There are generally two types of WDM switch architectures in the view of multicasting. One is multicasting capable (MC), the other is multicast incapable (MI). In general, a space division switch in a WDM network, such as those made of electro-optical
directional couplers, are MI. An MC switch needs to use splitters or passive couplers as a component.

In one example of MI switches, each input WDM signal is de-multiplexed first, and each channel may then be converted to different wavelength using a wavelength converter in order to avoid contention. Each channel is routed to desired output port by a NM by NM space division switch.

An example of MC switch architecture is shown in figure 12. Each input WDM signal is split into NM signals through two stages of splitters (first 1: N then 1: M), one WDM signal for each N by 1 space division switch at the third stage. Each space division switch then selects one of the N input WDM signals, out of which one wavelength is extracted using a tunable filter. The extracted wavelength may then be converted into a different wavelength to avoid contention. To support multicasting, the same input WDM signal is selected by multiple space division switches, which are normally connected to different outputs. This architecture can also send multiple copies to the same output using different wavelengths by letting multiple space division switches to be connected to the same output select the same input WDM signal.

Another approach proposed is to use optical burst/label switching under the framework of MPLS (multi-protocol label switching) and modifies a multicast tree constructed by distance vector multicast routing protocol into a multicast forest based on the local information only.

C) Some of the solutions proposed for including multicasting incapable routers

Because some WDM switches may be incapable of switching an incoming signal to more than one output interface, and keeping multicasting data in the optical domain may not be easy, the multicast tree constructed by IP must be modified to ensure that the multicast data can reach all the receivers. One approach to WDM multicasting is based on wavelength routing, which constructs a multicast forest for each multicast session so that multicast-incapable WDM switches do not need to multicast and router can use different wavelength for each outgoing output.

In the environment of only part of the WDM switches is capable of multicast, which is called sparse splitting, one way is to use a new or modified IP multicast protocol that
constructs an appropriate IP multicast tree, where an IP router can serve as a branching node only if the router has an underlying WDM switch which is multicast capable. This tree can be used by the WDM layer to switch multicast data.

Another possible solution is to keep the semantics of existing IP multicast protocol intact but let the WDM layer to modify the IP multicast tree constructed by MOSPF or DVMRP. This approach requires global knowledge of the physical (WDM) network, including the multicast capability of each WDM switch.

1. Buffered or Buffer-less switch

There has been real divergence in regard to whether buffers should be used or not in the optical switch. Currently, the only practical optical buffer are fiber delay lines as shown in figure 13, which can be very costly and space demanding.

![Figure 11. Optical buffer implementations [8]:](image)

(a) Fiber delay line. (b) Programmable fiber delay line. (c) Feed-forward time slot interchanger. (d) Active switched re-circulating delay line
Chapter 6  My new core switch design

A. Overview of my design

Three of the most commonly used schemes of routers are shown in figure 12. Scheme has less involvement of electronics interpretation and fewer O/E and E/O conversions from top to bottom.

*Electrically-switched router/transponder

*Optically-switched router/transponder

*All-optical wavelength router

Figure 12. Tunable wavelength router/transponder implementations [36].

Rx: optoelectronic receiver
Tx: optoelectronic transmitter
My system design selections are based on comparisons of many options available and tradeoffs for overall system performance. It can be categorized into the third type in the design trend. Also considered is the system stack, by adding a separate optical control layer (such as MPLS like control and burst switching) to glue the optical physical layer and upper control layer. I will fuse the functionalities of these together by putting functionalities into the structure of these two layers and make the system stack work as a whole. The design follows the same philosophy as in proposed MPLS-lambda switching network.

Some of the assumptions are used as ground of arguments in my design. For OC-192, the speed of one channel is about 10G/s, which is close to the limit of processing speed right now, to store 1 bit of data, FDL needs to be about 3cm. An ATM cell will occupy 11.5m FDL and a 1500 byte IP packet will take 360m FDL. For a 600mi link, if we assume 10G/s and 1MB per burst in the burst-switching scheme, only 4 bursts can be in the link, which means if we want to buffer one, it will be 150mi for each FDL. The burst switching technology currently under study has to take into the account the physical dimension problems also. If long bursts are used, the less likely many or if any FDL will be used, And other mechanisms have to be used to improve packet drops due to burst-rate nature of traffic. If short bursts even down to single packets are used as data unit, the more likely we can afford many FDL to resolve the contentions and conflicts. Reuse of the fiber will be attractive in either case if FDL is used. To achieve true packet switching, system has to be able to reconfigure fast enough to keep up with the data rate, which means system design has to take into account of limits of components. Any time needed for reconfiguring the system will be overhead and it should be a small percentage of transmission. Long overhead has to be compensated by longer stream of data transmission with the same configuration which either achieved by only change for perceived long stream or long burst duration of single packet. As we have shown in chapter 1, tunable laser or tunable filter or tunable wavelength converters have certain range of operation for wavelength versus tuning time. To achieve the true packet level switching, I will avoid using those components if tuning time is long. Currently only electronically tuned components are in the order of ns to keep up with average size packets, but their tuning range is very limited as to achieve high degree of multiplexing of wavelength channels. As we know, most of Internet traffic now is small
sized packets. A packet of 125 bytes is 1000 bits, to limit the overhead within 10%, switch should be able to reconfigure within 100 bits transmission time. At current highest rate of 10Gb/s, that is 10ns. Electronic buffer can still be used as long as the delay will not exceed the overall time limit. This will complicate the system of the switch and it is considered to be out of the scope of my all-optical switch design discussion.

Since control module will have information about all the output links in its control, it is possible to use deflecting algorithm to switch packets to an optional output link instead of default link since packets are broadcasted to every output link. Switch decisions can be made by the control module are as below:

A. Switch to output
B. Switch to buffer
C. Switch to another output (deflecting)
D. Discard (packet dropping)

Some new designs are introduced in various parts to complete and improve the functionalities needed. The switch consists of four major parts. A: input links which consist of fibers with multiple wavelength channels. Every input link will split the light evenly to feed into every output link’s input interface. B: control module for processing control header and other network exchange information. C: space switching concentrator for every output link, which concentrates all input channels into smaller number of output channels and FDLs (fiber delay line) if included. Each switching fabric can be different and is dependent on characteristics of specific link. D: output links with wavelength converters for every output wavelength channel and FDLs are shared among all the channels as buffer pool.

An example of a 5X5 router is shown below.
My design will assume that the core switches will be used as cut through routers which will not examine the content of packet data for transmission errors, checksum errors and so forth. It is edge router’s responsibility to verify those before sending packets into the core network cloud and after packets exit from it. It will be equipped with limited number of or even no FDL buffers and edge router will have to store, multiplex and forward packets. If burst packets are used to transfer packets in the core network, Edge routers will fulfill the job of assembling burst packets for core switches and disassembly bursts into packets or cells for transmission into regional network. The edge routers will also play the role of traffic policing so as to restrain outgoing traffic when packet drops are detected and balance the traffic among available links and channels for outgoing traffic to core switch. Buffer is used for conflict resolution only if used, not for storing packet at intermediate node. As being believed, if intermittent delay can be mostly eliminated by not queuing packet for long time in the intermediate routers, real time requirement for packets will not be a concern. With the cost of adding capacity getting lower, fewer and fewer buffers should be used which leads to shorter queue and delay. In the control header for the data burst, a TTL (time to live) or
similar time stamp should be used so that packets will be dropped when delay is detected to be too long.

The reasons to choose broadcasting for all input links to output links are as follow: A: the number of links a core switch will connect is likely to be small, most likely in the range of three to below ten, so the splitting to every output link will not complicate the whole design and hardware cost will not be major burden. Only a few splitters and possibly optical amplifiers are needed for this purpose. B: it will automatically support multicasting without the need of complicated control algorithm or hardware structure. It is strongly believed that applications foreseen for multicast will lead to high possibility that most of the output links will be targeted if a multicast is initiated. The future growth driven applications such as video on demand or virtual private network are intrinsically multicast oriented and the majority of the traffic will be multicast packets. C: natural implementation of output queuing to reduce blocking since every input will be available to all output link concentrators. It is well known that output queuing has the best system performance to avoid blocking. The new design’s many desired features are listed as below.

a. Built in multicasting support.
b. Support deflecting algorithm if choose to.
c. Support possible different QoS (Quality of Service) schemes.
d. Simple structure and control algorithm. Many variations can be made for the exact system requirements.
e. Reduced hardware requirements, especially for expensive optical components
f. Good scalability, work in the framework of current limits of hardware capability and easy future upgrading.
g. Modular design to simplify design and manufacturing
h. Output queuing system reducing blocking possibility
i. Support fast packet switching among all grouped wavelength channels.
j. Limited wavelength conversions, only one converter for one output channel.
k. Small space-division switching fabric to reduce cost and help fabrication.
1. Totally shared buffers for all output channels in a given output link to reduce hardware and *improve packet dropping loss*. Do not require faster buffer access than link speed as in traditional output queuing system.

The concentrator is at the center of the switch, which will be controlled by center control module. The new design reduces the number of elements for the space division switching fabric needed to realize the output queuing system.

The logic diagram for the most commonly used building elements for the space division switches is $2 \times 2$ and $2 \times 1$, which are shown as figure 18. The actual implementation can be different as shown in figure 19 and figure 20.

![Logic diagrams of cross-connect elements.](image)

*Figure 14. Logic diagrams of cross-connect elements.*

![Implementations of cross-connect elements with amplifier gate switch.](image)

*Figure 15. Implementations of cross-connect elements with amplifier gate switch.*

![Schematic of optical cross-point elements [1].](image)

*Figure 16. Schematic of optical cross-point elements [1].*
What we need in the packet switching network is a dynamically reconfigurable switching matrix. A static switching matrix will not be considered here for the core switch design. There are many dynamically reconfigurable space switching matrix designs, some of them have strictly non-blocking property which usually will require more than $N \times N$ switching elements for a $N$ inputs $N$ outputs matrix, some of them are blocking switches, which will have high possibility of internal blocking. I will not consider blocking switch in my design because it is under the erroneous assumption of low packet drops because internal blocking is small. It has been shown that between the two extremes, there can be designs categorized as wide-sense non-blocking switches. In this category, the switching fabric will switch the packets non-blocked under certain restrictions. The number of switching elements can be reduced because of some combinations of input to output are not possible under certain system restrictions. Some types of switches are shown below.

![Switching Matrix Diagrams](image-url)

Figure 17. Basic schemes optical cross-connect architecture [6].
(a) Rigid; (b) Rearrange able; (c) Strictly non-blocking; (d) Strictly non-blocking with lower-dimension space-switches
<table>
<thead>
<tr>
<th>Architecture</th>
<th>Number of active elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional AS/AC</td>
<td>2N(N-1)</td>
</tr>
<tr>
<td>Conventional PS/AC, AS/PC</td>
<td>N(N-1)</td>
</tr>
<tr>
<td>Enhanced PS/AC</td>
<td>N(2N-1)</td>
</tr>
<tr>
<td>Enhanced PS/PC</td>
<td>N^2</td>
</tr>
<tr>
<td>Simplified AS/AC</td>
<td>N(5N/4-2)</td>
</tr>
<tr>
<td>Simplified PS/AC, AS/PC</td>
<td>N(3N/4-1)</td>
</tr>
<tr>
<td>Two-active-stage (point-to-point)</td>
<td>3N^2/4</td>
</tr>
<tr>
<td>Two-active-stage (multicast)</td>
<td>N^2</td>
</tr>
</tbody>
</table>

Table 4. Number of active elements in tree-type architectures [80].

Some of the typical space switches and their total number of elements and other properties are listed in the table 5.

Most of these designs do not take into account of real world network needs, they assume N input channels and N output channels and they assume same number of inputs and outputs, this might not be the case, and those designs do lack of the flexibility to accommodate this possibility. The Internet traffic is never symmetrical. The bursty nature of traffic and multicast requirements often lead designers to add extra modules to accommodate the needs. It is also clear that it is not possible to implement multicast at optical level in those design of switches. It is intrinsically one to one, not one to many capable.

In the effort to solve some of these problems and take into account of the need of simple control scheme to keep up with the need of fast switching for future optical network, I propose a new structure and its corresponding space switch.

There was one popular design called knockout switch for electronic packet switching design, which has the similar functionality of my concentrator. Knockout switch has been proposed for electronic switches and routers. Its diagram is shown as figure 22 and 23.
Figure 18. Schemes of original knockout switch [67].

Figure 19. Internal structure of knock-out switch [71].

My design also takes advantage of non-symmetric inputs and outputs structure and central control philosophy to make the structure much simpler. Figure 24 illustrates the internal structure of a shift-and-drop concentrator for an output link and figure 25 shows the diagram for the 4 by 2 concentrator using 2x1 elements.
Since we only need strictly non-blocking switching for M outputs from N inputs under the control of the control module, the number of elements in the concentrator can be reduced dramatically. Since through the reservation and computation, only M packets can be switched to the M available output channels (in the case of no FDL buffer) or M + L channels (in the case of having L FDL buffer), we can predetermine the inputs which will be switched out and block those which will not be. As shown in figure 24, all the winners from
input link 1 and 2 will travel to the bottom of the outputs of concentrator and all the winners from input 3 and input 4 will be shifted to the leftmost of the outputs of concentrator. Since the sum of the survivors from the two concentrators will never exceed M, there will be no conflict in the 2M x M selector. The structure can be cascade to higher number of nodes. For example, in the case of 8 input links, two intermediate M outputs will pass through another 2M x M concentrator. Table 6 lists the correspondence of the number of input links and number of concentrators.

<table>
<thead>
<tr>
<th>Input Links</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentrator</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>11</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 5. Number of concentrators versus number of input links.

As we can see, the new switch has (N-M) stages of shifting for the N input M output concentrator with N, N-1, ..., M switching elements at each stage. The total number of 2 x 1 elements will be \((N-N-M-N+M)/2\), considerably fewer than most of switch designs. It is especially attractive when N and M are close in value, especially in the design of having many parallel FDL at outputs. The number of elements will be dramatically reduced compared to the knockout switch design also. In the original knockout switch, every input has to compete with every other input to decide if it can be a winner at the output or not. Since in my optical switch design, winners will be decided by the central control module, which can simply block all the losers and only switch all the winners to the output. Another advantage for my design is that when switching fabric gets bigger and bigger, it will be harder to fabricate, it is desirable to build the switch in a hierarchy. In my design, the number of elements will only increase linearly with total number of inputs. Because of this, we can opt to add more elements to reduce cross talking among channels when they travel through the switching matrix to improve S/N, as people have shown improved S/N in designs of dilated-Benes switch over Benes switch.

Because of the structure of the concentrator, simple 2 X 1 SOA switch can be used as the basic building element. As an example, assume there are 5 links each with 20 channels,
most of traditional NM by NM non-blocking switch will require $100 \times 100 = 10000$ $2 \times 2$ elements while not supporting multicast. This will put a limit on the size of the switch and limit the number of wavelength channels can be multiplexed. In my design, for each link, it will require 2 much smaller switching fabric, each with two 600 concentrators. Total $2 \times 1$ elements are 1200 for each link and 6000 total for the whole switch with multicast support. Smaller modules also have the advantage of facilitating faster control modules. If we use SOA to build the switching elements, $2 \times 2$ will need 4 SOA and $2 \times 1$ only needs 2 SOA. This will make the difference even bigger. My design uses considerably fewer elements compare to the traditional switching matrix yet has more capability.

Space division switch might be too large to exploit when each link has around a hundred wavelength channels as envisioned in the future. One possible solution is to divide channels into groups, basically treat each group as a separate layer. Switching will be limited within same group, logically several vertical networks running at the same time. This will lose some of the advantage of statistical multiplexing compare to using all channels for statistical multiplexing, but can meet the system limits and leads to simpler design and smaller components. It also has the advantage of being less demanding for all wavelength-related components, such as optic filters and wavelength converter where wavelengths in a group can be very close to each other. Also upgrading capacity will be easy too, without throwing away all the switching components. Load balancing can be done at the edge router to dispatch traffic evenly into groups.

In this design, each output channel has one fixed wavelength converter. Signal will only need to be converted into the new wavelength once in the switch. It is designed to reduce the number of costly wavelength converters and avoid using slow tunable wavelength converters. It will be possible to use much smaller size of burst packet since reconfiguration time can be very small, in the order of tens of ns.

Buffer can be used in the switch if contention resolution is needed when capacity is limited. What is needed from the buffer is to have large enough number of interface to the concentrator to reduce the conflicts, also deep enough to hold enough packets to smooth out the bursty nature of the Internet traffic. Figure 26 shows normal setup of buffer in packet switches. Buffer pool can either be recycled through the input of concentrator again thus can
be shared by all output or can be switched to output directly thus can only be used by one output link. In my design, a FDL pool will be used and shared by all the channels in an output link, so the buffers will be dynamically available to all the links and statistically deeper than if they are divided statically among all the channels. It can also be designed to support priority-based content replacing policy. Any priority based QOS control can be done in the FDL array and controlled electronically. High priority incoming packets can replace the low priority ones in the buffer if we choose to do so.

![Diagram](image)

**Figure 22.** Possible buffer locations for (a) electronic packet switch and (b) all-optic photonic packet switch [8].

Since FDL buffers are expensive, they should be used as efficiently as possible. Total shared buffer offer best system performance as shown in figure 23, the complexity of SHWP
will lead to fewer and shallower FDL buffers used at each output link. I will choose to use shared output buffer for each output link if buffer needs to be used, but that will increase the dimension of the switching matrix dramatically. All the output wavelength channels for the given output link will share all FDL buffers, and they are used to reduce packet drop and resolve contention. As studies have shown, dynamic switching among multiplexed wavelength channels provide much better system behavior than just switching within same wavelength channels and higher degree of multiplexing and better the system can handle contention (figure 24).

![Graph](image)

Figure 23. Comparison of required buffer depth for shared and unshared schemes [12]. An idealized comparison of buffer depth per wavelength for both SCWP and SHWP at different loads for a 32-input/output switch to achieve a packet loss of $10^{-9}$, with fixed-length packets.
Figure 24. Discard probabilities versus degree of channel multiplexing [48].

Compared with a conventional electronic packet switching system, the optical packet switching system will have to enter the switching matrix at the pre-established route, which will be set up by the central control module after receiving control headers for the data packets prior to their coming or data packets will have to stay in delay line waiting control headers to be processed. There can be several schemes exploited for the QoS. In one algorithm, we can assign channels with different priorities, and input channels are aligned according to their priorities. Only channels with higher priority and have packet to transmit will be able to take output channels if available. The channels in every link will be aligned in the sequence according to their priorities. Control headers for the corresponding channels can be sent out in that sequence for the potential benefit of taking early position. This does lead to the problem of priority reverse in some situations. For example, in the situation that in some links low priority packets get transferred because lack of high priority packets, so they are sent in higher priority channels and if no other information is carried in control headers, they will be considered to have higher priority than packets from other input links which actually have higher priority. If priority information can be extracted, packets can be
rearranged according to their priority, which can be achieved by adding M x M matrix at the end of the concentrator to reshuffle outputs according to their priorities in the intermediate switches. We can prioritize packets at the edge router and let packets interact at intermediate switches. Since priorities are kept at node only, certain degree of priority-based differentiation will be kept but not strictly, as in the traditional electronic router. The degree of differentiation can be a good research topic to study. To further the concept of using offset time between control header and its corresponding data packet as a mechanism to achieve priority differentiation. Even with the huge bandwidth WDM brings to alleviate the burden of Internet, the challenges facing the system designer to integrate QoS remains. Now ATM is the major candidate for providing the solution for integrating the QoS on top of WDM service, but the small cell size and huge buffer size required by the ATM system are not efficient to fully utilize the bandwidth. Burst Switching has been proposed to reduce the overhead and system complexity in core switches. MPLS is proposed as the model to be modified to suit the need of new context to facilitate the need of fast switching decision. As we have shown in the calculation, although two-way reservation protocols can guarantee resources, it is not desired because of the long round trip delay before actual transmission can happen. In packet switching, which dose not make reservations, it utilizes bandwidth more efficiently in the bursty nature of Internet traffic, but lacks QoS support. To include the QoS support, usually the payload needs to be stored and header processed to decide if and how it will be serviced. There are numerous schemes proposed for IP service and most of them require electronic storage of data packet and queuing method. Also the delay time needed for control header to be processed before switching decision can be made adds very high overhead in the context of high-speed network.

<table>
<thead>
<tr>
<th>Optical switching paradigm</th>
<th>Bandwidth utilization</th>
<th>Latency</th>
<th>Implementation difficulty</th>
<th>Adaptivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelength</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Packet/cell</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>OBS</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
</tbody>
</table>

Table 6. Comparisons between three optical switching paradigms [46].
Optical burst switching (OBS) has been proposed to achieve the balance between the coarse-grained circuit switching and fine grained packet/cell switching. By aggregating short packets into relatively long bursts at the edge router before entering the switching matrix of WAN, we will be able to reduce the ratio of the overhead and use bandwidth more efficiently.

Figure 25. Illustrations of (a) optical packet switching and (b) optical burst switching [46].

Generally, there are two methods used in OBS. In the first one, packets are sent out at the same time as the control headers, in-band or out-band, and intermediate router/switch buffer or delay the data payload while processing the control header. In the other one, control headers are sent out at certain offset time before the actual payload is sent out, the time gap is
calculated to compensate the processing delays at all the intermediate nodes. Since the delay is realized at the edge router, it is possible to use buffer-less schemes in the core routers.

![Diagram](image)

Figure 26. Using offset time to realize reservation [46].

Algorithms have been proposed to provide some extent of priority-based differentiation to support QoS in the core switches. At the edge router, the reservation control headers can be sent out before packet assembly is over. Based on the service classes, the starting times can be different. For example, the highest class can start reservation once the queue is started, since the queuing time will be defined by the algorithm, and also it is the first class, it will be able to reserve bandwidth over other classes. Mechanisms can be provided so that certain bandwidth is allocated for lower classes to prevent starvation. Service class will decide the waiting time to assembly a burst. The offset times to send the corresponding control headers relative to the start of the assembly are set according to the degrees of differentiations of classes. Different classes can be merged into upper class if space is open, but the complexity of implementation will decide if the scheme is feasible. Separate channels will be used for control headers and depends on the header/burst ratio, one control channel is used for many data channels. Since switches are connected point to point,
two switches can statically setup certain channels for control purpose only. Also since optical/electronic conversion will be done to the control headers, control channels can also carry other network control packets such as synchronization, flow control, negotiation and such. The control/protection/restoration messages for immediate switches should be carried in the control channel since they will not be converted into the electronic form for interpretation if they are carried in the data channels.

It has been proven that when arriving packets are synchronized and aligned, the packet drop rate is two to three orders lower than random pattern. This cost of synchronization is to decide a fixed time slot for all switches to follow and using delay lines or other mechanism to align incoming packets from all links. The mechanism borrowed from SONET and other proven systems can be modified to suite the needs of optical switches. If variable duration packets are allowed, variable length FDLs have to be used to accommodate them if FDLs are used. This will present even more complexity for the system design. The fixed duration packet scheme will be used in my design. Using fixed duration at edge router to decide the packet size instead of actual number of bits will solve the problems of buffer size and data rate difference among different wavelength channels more gracefully, the proposed design will adopt this philosophy. I will assume transmission rates are the same for all data channels, so the fixed duration will mean fixed packet size for ease of explanation.

Assume all the links are synchronized, the central control module can decide if a higher priority channel will have input or empty once the control header for that channel comes and do not have to wait for all the control headers to make a decision. To make the mechanism fair to all links, the channels from all links will be interleaved so as to form a round robin fashion. To make the situation even more fairly, we can add a 4X4 space switch and rotate the sequence among links every certain period of time. Another case is there is no predefined priority for each channel, rather the packet priority is indicated in the control header, so that we need to dynamically decide which channels will be the winners to be switched out. We have to wait until all control headers have been processed to decide the winners. This means that there will be a minimum offset time between a control header and the corresponding datagram and which might need to be negotiable between routers or set for requirement of routers. The control process in the first case can be much faster, since the
control headers can be sent in the order of their priorities. Once one header comes, it can be processed and control bits for its path can be set. As soon as the control module receives the last header, all control bits can be set and the concentrator is ready to switch packets. In the second case, the switching decision can be more flexible, but the control module will have to be more complicated, it has to wait till last header before it can make any decision. So depends on the principles of system design, we can adopt different schemes to better serve the requirements. In this thesis, I will assume the second design.

If the short packet scheme is adopted, the fast decision will be desirable to reduce the processing overhead. If the long burst scheme is adopted, then longer control process and more sophisticated control can be used since the decision time can be longer.

If we don’t want the data packet to be delayed at the intermediate node other than normal travel and switching latencies, the control header needs to be sent out ahead of its corresponding data packet for at least the sum of all processing time which the header will experience in the switching cloud if no extra delays are set up to compensate for the processing time. Accurate time to live should be used as a guidance to decide how early the header should head out which will be different than the current use of it, in which hop counts are used. If switch does not have time to process the control header before the arrival of the data, it will have to drop the data packet or use buffer to delay it if available.

If we do not want to estimate the processing delay since it might not be possible to be decided ahead of time, we can use inlet FDL at every input of the switch to delay the data packet for the time to process the control header provided processing time is relative small.

In the in-band control header scheme, inlet FDL has to be used to delay the data packet part while the control header is being processed.
Chapter 7 Conclusion

This new design emphasizes lowing system cost while not sacrificing performance. This is realized by implementing better strategies from results of current research effort by numerous researchers. Strategies used include: using light splitting to support multicast, using wide sense non-blocking space division switch to lower switch dimension and eliminate usage of tunable components, using large degree of channel multiplexing to reduce buffer depth and reduce packet drop rate, using synchronous mode of transmission with fixed duration to simplify buffer design, increase link utilization and reduce packet drop rate, separating control header with data body to avoid buffering in the intermediate switches, using offset time as an effective way to realize differentiations among classes to provide certain extent of QoS.
Chapter 8  Future research

The WDM technology is still under active research. The property of different topologies, algorithms and protocols investigated will not only decide the implementations for future application, but also affect upper level services. The deeper understanding of the capability of WDM will definitely be the key in future network designs and application adaptations.

Most of the design decisions are still based on theoretical models and they can only be proven by running real traffic throw them. Since it is expensive to build actual hardware for testing, tools such as VHDL or SystemC which can incorporate accurate timing information can be used to simulate the system behavior to demonstrate some degree of confidence of successfuleness of the model and that is the next step of my research.
References


