Optical Burst Switching: The Next IT Revolutoin Worth Multiple Billions Dollars?

Liwu Liu,Peng-Jun Wan, Ophir Frieder
Department of CS,Illinois Institute of Technology
10W 31ST Chicago, IL 60616
{liwuliu,wan,ophir}@cs.iit.edu

June 30, 2000

Abstract

All-optical Internet may one day come true, due to the explosive bandwidth requirement, and advances of enabling optical communication techniques. And exactly NOW we are at the turning point. We study one possible optical packet switching technique-optical burst switching. We present the overall network protocol, enabling techniques, architectures, switching cores and performance evaluations.

1 Introduction

While Internet and on-line e-business are booming, data networks grow rapidly, and direct video distribution is in expectation, it is proposed to use all-optical network (AON) to support next generation BISDN networks, to avoid electronic bottlenecks and decrease the cost by simplify the multiple layer architecture using now (such as ATM/SONET/WDM, IP/ATM/SONET/WDM). Since not faraway from nowadays SONET facilities, AON is expected to provide smoother network immigration, easier capacity expansion, cost reduction, signal transparency, more reliable networking and better qualities of service.

For packet switching networks such as Internet, on one hand, there are intrisical limitations to build huge-volume electronic switches with multiple ports using VLSI/CMOS techniques, such as limitation of electronic RAM accessing speed and limited size of VLSI fabrication. On the other hand, simply using an arrayed-waveguide grating router, a set of wavelength with speed like OC-192 (10Gbps) can be switched from input ports to output ports. We refer readers to [1]-[5] for researcher's efforts to build optical packet switches. Also companies such as Alcatel, Corvis, Telllabs, Nortel, Cisco, Qtera, Lucent, Ciena etc focus on commerical products and some of them have all-optical solutions.

2 Optical Burst Switching Technique

[5] then proposed *Optical Burst Switching (OBS)* network as one solution for future IP/DWDM networking(nowadays IP routers are directly running either over ATM or SONET). In OBS networks, a small routing header is sent before a burst of packets are sent. This header attempts to reserve the wavelength channel along its routing (also computed by intermediate switches/nodes) and it is processed by intermediate nodes in eletronic domain. Proper timing relation ensures that exactly when a burst enters an intermediate node, this node has switched the corresponding incoming wavelength to proper outgoing wavelength in the proper outgoing fiber. After a period of time, the last bit of the burst has gone out of this node, it then can re-use the incoming and outgoing wavelengths to setup different switching relationships for other bursts that will arrive late. Thus a burst arrives at the destine without any intermediate E/O or O/E conversion. And the time that the burst needs the switching relationship is the length of the burst divided by the channel line speed.

2.1 One-way Delayed Reservation Protocol

One specialty of DWDM optical networks is that, each wavelength channel provides transportation pipes with huge capacity, such as OC-48 (2.5Gbps) and OC-192 (10Gbps), and the in-test OC-768 (40Gbps). But compared with classical copper networks, the end-to-end delay does not vary too much, since there is not a significant difference between the light speed in fiber glass and the speed of electronic-magnetic wave in copper. Thus for packet switching network basing on DWDM technique, two-way reservation scheme, such as ATM Q.2931, IETF RSVP, can not leads to high channel utilization since reserved channels remain idle during the round trip time the reservation signal travels when reserving sources along routing.

Another reason is that, if using two-way reservation protocol, channels are held during the total round trip time, the blocking probability of other burst increase very rapidly when their resource requirement has common set with those held channels. Thus totally the network has very lower utilization and very high burst dropping probability.

So one-way tell-and-go delayed reservation protocol is proposed to support all-optical burst switching networks. Under this scenario, a burst header is sent out to pre-

0-7803-6512-6/$10.00 (C) 2000 IEEE

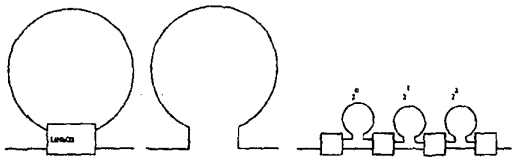

Figure 1: loop, line, and multistage delay device

reserve channel resources, namely a time interval over channel and storage in which it passes through, before the burst is transmitted. This protocol can results in higher channel utilization and lower blocking probability.

2.2 Enabling Techniques

Theoretical analysis and simulations showed that the multiple parallel interchangeable channels provided by D-WDM technique and fast-tunable wavelength conversion technique is critical to the practicality of optical burst switched networks. Given traffic pattern, the burst dropping probability is mainly decided by channel interchangeability. Partial interchangeability (from using partial conversion) is often not enough to result in low burst loss rate. When channel interchangeability is given, optical buffering utilizing fiber delay lines/loops then helps to decrease the burst loss rate, especially the channel contention caused by traffic time correlation. And, for optical packet networks, fast MEMS (Micro Electro Mechanical Systems) switches and fast converters with wide conversion range are needed. Combined with conversion capacity, cheap and experienced switching technology such as arrayed-waveguide grating, or mirror matrix can be used to build the core wavelength space division switch. Till now these dynamically reconfigurable optical enabling techniques mostly are basing on Lithium Niobate technologies and semiconductor optical amplifiers.

2.3 Global Timing

The strict global timing in fact has been implemented by SONET. For example, in the Sprint OC-48X4 SONET networks, primiary atomic clocks generate timing information that is transmitted to big central offices via Satellites and GPS (to make sure the strict uniformly timing). Small centeral offices can extract the timing information from SONET itself andthen drive local clocks. Similar global timing can be used in OBS networks to provide necessary synchronization.

2.4 Optical Buffering

In real world, the Internet is bursty, according to analysis on logged traffic from several big communication companies. We need use buffering to statistically multiplex those bursty traffic into channels so as to improve channel utilization and network performance.

Delay line can only provides one single delay and under certain conditions delay loop performances much better than delay line since it can provides a set of (discrete) delays, when channel contention has to be solved. If the traffic time correlation is in long term, long delay line is needed but small sized delay loop works well. But delay loop has its own limitations: (1) the length of delay loop must be larger than any burst it delays, which is not true for delay line; (2) delay too long in loops attenuate its power and amplifier such as EDFA is needed. (3) Small switches or SOA gates are needed to control whether the delayed burst is accessed or continues to circulate in the loop.

A burst can not circulate too many times in a loop and a loop thus provides a set of delays as $\delta, 2\delta, ..., n\delta$ where n is the maximal number of circulating times. A multistage switched delay device also provides delay $0, 1, 2, ..., n = 2^k - 1$, where k is the number of stages. For a loop, if it provides $k\delta$ delays for some burst, then a time interval with length $k\delta$ is used by this burst. For a multistage switched delay device, no matter whatever delay it provides for a burst with length l, a time interval with length l is used by the burst since when the last bit of the burst has entered the delay device, it is ready for delaying the next burst.

2.5 Optical Channel and Optical Buffer Scheduling

In electronic packet switch, we can use memory management unit(MMU) in CPU, (address/data)bus, bus arbiter logic, etc to control electronic memory. To address optical memory consisting of fiber delay line/loops in parallel, wavelength space division multiplexing switch is used, such as boradcast-and-select, crossbar, arrayed-waveguide grating. Electronic signals then are generated by optical memory scheduler to control SOA gates, filters, crosspoints, configuration of AWG routers, point of tunable converters, and other active components. Considering the discrete time attribute of optical memory and channels and bursts, the optical memory scheduler then should maintain (1)information about the time intervals on which memory and channels are idle and available to efficiently support IP/DWDM; (2)information about control signals that should be generated at specific time poistions.

Several buffering such as output buffering, shared buffering, recirculation buffering are possible since each has its own merits. For each buffering mechanism we have two basic optical implementations : (1) fiber delay line, looped or unlooped (2)fiber delay loop. (Till now we do not have optical memory such as proposed holography memory.) Accordingly fast algorithms to allocate channel and buffer

3 An Output Buffered Switch: Detail Case Study for Optical Bursting Switching Mechanism

Assume after demultiplexing wavelengths in each incoming fiber, the space division switch accepts NW incoming wavelength links as input ports. It has $N(W+F)$ outputs, in which NW ports address the outgoing NW wavelengths and NF outports address NF buffers, F buffers per outgoing fiber. Each buffer is implemented as a delay line/loop. Each outgoing fiber uses a passive coupler to multiple W channels among W channels and F buffers. We do not focus on how the switch is built and assume it is strictly nonblocking.

In this section, we assume the buffer is implemented by single delay lines, to make our presentation simple. At the end of this section, other buffer implementations then are discussed. Normally the length of delay lines is selected to match the traffic burstiness.

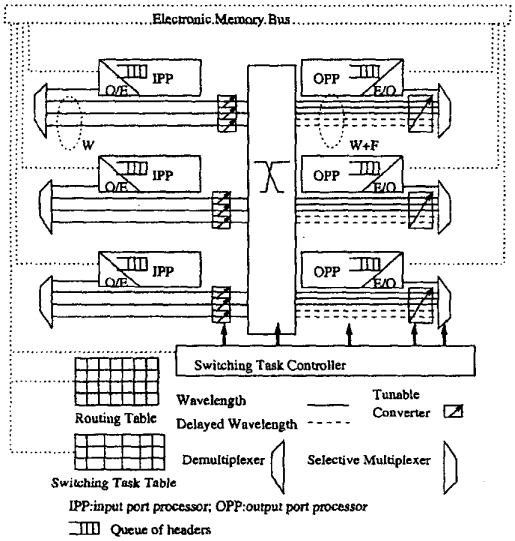

Figure 2: An Output-Buffered OBS Switch

In Figure 2, the control core then consists of (1)N input port processors, each of which listens to the control network to accept burst headers, looks up the routing table to decide the routing, and passes burst headers to corresponding output processors. (2) N output multiplexing processors, each maintains the resource utilization information related to its corresponding channel and storage resources and manages traffic oriented to this output. OPP processes each burst header and assign time intervals in buffer and wavelength and then write this information into a common switching task table. (3) A switch controller periodically read the switching task table, reconfigure the space division switch and converters and passive couplers according those switching task information.

Such an architecture permits parallel processing of routings and resource schedulings and it is necessary for fast burst switching. Assume we have $K = W + F$ outgoing wavelengths from the switch per outport. Thus each output can accept K bursts at the same time from the switch and F is the capacity of buffering to handle traffic burstiness (normally the outgoing link accept at most W bursts each time and at the worst case, NW bursts from NW incoming wavelengths are targeted to the same output port).

The scheduling of buffer and incoming/outgoing channel are tightly couple in time dimension: On one hand, exactly when the first data bit of a burst enters the switch, it starts to use the pre-assigned buffer. Exactly after the delay time provided by optical memory, one pre-assigned outgoing channel is used for this burst.

Each node maintains a scheduling table of *switching tasks*. A switching task entry is a tuple (w_i, w_o, s, e, id, f). It records the time intervals at which electronic control signals are to send to converters, switch and selective couplers. Here when fiber delay line $f = 0$, (w_i, w_o) is the pair of incoming wavelength and outgoing wavelength for which at time interval $[s, e]$ a switching relation needs. It is a transmitting task from w_i to w_o and corresponds to a switch control signal, for example, a signal to enable a specific crosspoint if crossbar is used. When incoming wavelength $w_i = 0$, then (f, w_o) are the pair to be set up switching relation for, which also corresponds to a transmitting task, but from the buffer to the outgoing wavelength-it corresponds to a control signal sent to the selective coupler. When $w_o = 0$, the pair (w_i, f) will be set up with their switching relationship during time interval $[s, e]$, which is a buffering task-it corresponds also to a control signal sent to the switch. $[s, e]$ is the corresponding time interval reserved for this task. id is the sequence number of the burst assigned by the source node. Here for a specific pair (w_i, w_o) or (w_i, f) in some task entry, it also defines the necessary converter tuning position at that time.

Periodically a switching controller reads entries in the scheduling table and finds the switching tasks whose starting time is exactly the *current time*, and drives the control logic to switch corresponding entries. So each switch task (w_i, w_o, s, e, f, id) is executed just upon time s. Here the time needed to setup switching relation is ignored.

Assume at time t, one OPP receives some burst header $h(w_i, t_b, l_b, t_h, s, d, id)$ forwarded by one IPP·(it means IPP has completed routing-it forwards the header to the outgoing link the OPP is managing.), where w_i is the required wavelength and t_b is the time when the first bit of the burst is transmitted by the previous node, l_b is the required time interval length, t_h is the time the previous node sends the header and s, d are source and destine which help to find routing and id is the burst sequence number.

The OPP then knows at time $t_a = t_b + t - t_h$, or $t_a - t$ later (from *now, time t*), the first bit of the burst will arrives and the burst lasts for time l_b. It then attempts to find one outgoing wavelength w_o.Here one wavelength w_o is legal when no switching task $(*, w_o, s, e, *, *, *)$ is in the scheduling table that satisfies interval $[s, e]$ overlaps with

interval $[t_a, t_a+l_b]$. If one legal wavelength w_o is found, then (1) a switching task entry $(w_i, w_o, t_a, t_a + l_b, 0, id)$ is added into the scheduling table. (2) a new header $h(w_o, t_a, l_b, t + \delta, s, d, id)$ is generated and sent to the next hop at time $t + \delta$, while δ is the electronic processing time and $t + \delta$ is the time to send out the new header.

Otherwise, we have to consider buffering. If there is a FDL, assuming the f-th FDL, is idling at time interval $[t_a, t_a + l_b]$ [1] (we select the legal one with minimal delay when there are many legel FDLs), and one "legal" compatible wavelength w_o, then we select f and w_o. Here w_o is legal when no switching task $(*, w_o, s, e, *, *)$ in the scheduling table satisfies that $[s, e]$ overlaps with $[t_a + d_f, t_a + d_f + l_b]$. Under this situation, a transmission task and a buffering task will be inserted into the schedule table and a new header $h(w_o, t_a + d_f, l_b, t + \delta, s, d, id)$ is generated and sent out to the next hop. Otherwise we can not buffer this burst and it is dropped.

One possible channel/buffer scheduling is so-called Horizon Based Scheduling: For each channel, simply its latest time at which the channel is currently scheduled to be used is recorded as its "scheduling horizon". When a burst header arrives, simply select a channel with horizon less then the burst arrival time. If there are multiple channels whose scheduling horizon precedes the burst arrival time, we select the one with latest scheduling horizon.

A scheduling algorithm that utilizes the total channel utilization information (states), rather than its scheduling horizon, then is expected to be more efficient, though it introduces the requirement of more data storage and more time complexity. Also, for a group of wavelengths outgoing to the same neighbor, we might permit to swap wavelengths for those switching tasks that are assigned over this group, if their time intervals do not overlap on every wavelength. Here the assumption is that after swapping, we still have enough time to set up new switching relation for those will-be up-to-date tasks.

More efficient approaches exploiting the resource utlization information (but with more time complexity) is also possible. Assume the group of outgoing wavelengths to one neighbor is $\{1, 2, ..., W\}$. The set of FDLs (FDLs are shared by all outgoing wavelengths in this group.) is $F = \{0, 1, 2, ..., F\}$ with delay $D = \{d_0 = 0, d_1, d_2, ..., d_{|F|}\}$, where 0-delay FDL ($d_0 = 0$) denotes the special case when we do not use delay, but directly use a channel. Assume the switching task scheduling table is $T = \{(w_i, w_o, s, e, f, id)\}$. For a coming burst header $h(w_i, t_b, l_b, t_h, s, d, id)$, we need to find a pair (f, w_o), where $1 \leq w_o \leq W, 0 \leq f \leq F$. The solution (f, w_o) need satisfy the following constraints to avoid any channel/FDL buffer conflict:

Channel constraints:

$$\forall (*, w_o, s, e, *, *) \in T, [s, e] \cap [t_a + d_f, t_a + d_f + l_b] = \Phi$$

[1] $t_a, t_a + l_b$ are the time when the first and last bit of the burst enters the FDL, respectively. Upon the last bit of the burst enters the FDL, another burst can start to enter the FDL.

FDL constraints:

$$\forall (*, 0, s, e, f, *) \in T, f \neq 0, [s, e] \cap [t_a, t_a + l_b] = \Phi$$

Those two constrains simply say that when two tasks are assigned the same channel/FDL, then their associated time intervals can not overlap. For example, consider the FDL constraint. Here $[s, e]$ is the time span on which the switching task is using the $f$the coming burst uses the f-th FDL since the burst comes at time t_a and $t_a + l_b$ is the time when the last bit of it enters the FDL. Thus we require the two spans do not overlap. Under this situation, at time t_a, the first bit of the burst arrives and starts to buffer at the f-th FDL, at time $t_a + d_f$ the first bit of the burst is switched and transmitted to the outgoing wavelength and at time $t_a + d_f + l_b$ the last bit of the burst is switched and transmitted to the outgoing wavelength. According to the above requirement, we proposed state based resource scheduling approach, which decrease the burst discard probability to one tenth of that from horizon based scheduling, for reasonable traffic intensity. (Detail omitted)

If delay loops or multistage switched lines are used, OPP has more options to select one among delays the loops/switched lines provide. For a burst arrives at time t, an idle time interval $[a \leq t, b]$ of a loop then can support delay within range $[1, b - t]$, provided that the burst length is no greater than the loop length. An idle time interval $[a \leq t, b]$ of a switched delay line can provide delay in range $[0, 2^k - 1]$, provided that $b - t$ is no less than the burst length. Thus we can select from multiple solutions of pairs (w, f) the one with minimal delay. similar management can be implemented.

We also can have other architectures such as input-buffering, share-buffering.

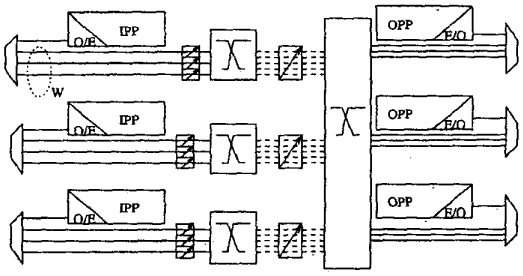

Figure 3: Input-Buffered Switch

4 Efficiency Evaluation for Case Study

Idealy, when we assume we can access the burst buffered somewhere at any time, OPP multiplexing performance can be modeled by a M/M/W/W+F queue. For M/M/W/W+F queue, we just assume the Poission arrival and exponentially distributed burst length. But there are several issues when adapt to this queue model. One is

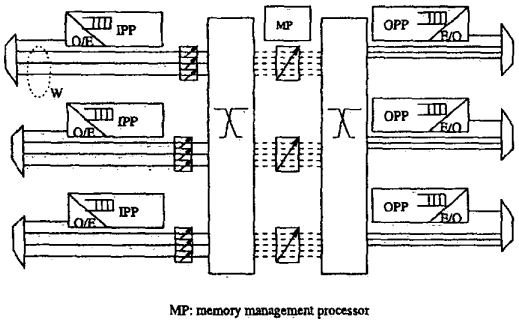

Figure 4: Shared Memory Buffering

the traffic burstiness of realworld traffic and Poission input might not model traffic very well. Another issue is that we can not access the buffered burst at any time, unlike electronic memory.

Our simulation results shows that the burst discard probability curves are close to that derived by using M/M/W/W+F queueing model. Figure 5 draws the OPP multiplexing performance under certain sceanrioes derived by simulation.

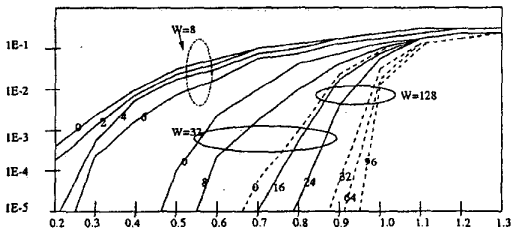

Figure 5: OPP Multiplexing Performance Using State Based Scheduling.

5 Conclusion

The rapidly growing bandwidth requirement makes optical IP router a potential solution as future Internet infrastructure. One key to a successful OBS switch is a fast and efficient resource scheduling implementation.

We did not consider the time uncertainty in this paper. Several factors cause uncertainty, such as light speed variance in different frequencies, temparory, etc. On one hand, high-quality wave guide should be used and uncertainty considered, on the other hand, we can assign a guard time before and after the time interval of a burst.

References

[1] Chlamtac, I.; Fumagalli, et al. CORD: contention resolution by delay lines. Selected Areas in Communications, IEEE Journal on. Volume: 145 , June 1996 , Page(s): 1014 -1029

[2] S.L Danielsen, P.B Hanse, K.E. Stubkjar. Wavelength conversion in optical packet switching. J lightwave. Dec. 1998 pp 2095-

[3] Transparent optical packet switching: network architecture and demonstrators in the KEOPS project. Gambini, P.; et al. Selected Areas in Communications, IEEE Journal on Volume: 16 7 , Sept. 1998 , Page(s): 1245 -1259

[4] C. Guillemot et al. Transparent Optical packet switching: the europe ACTS KEOPS project approach. J. lightwave. Dec. 1998, pp. 2117-

[5] C. Qiao et al. Optical Burst Switching (OBS) - A New Paradigm for an Optical Internet. J. High Speed Networks (JHSN) Vol. 8, No. 1, pp. 69-84.