Non-blocking Conditions in Scalable ATM Switches Using Path-Switching Scheme

Mai Jin*, Tony T. Lee, Soung C. Liew, Soung Y. Liew and Franklin F. K. Tong

* Center for Wireless Communications
National University of Singapore
20 Science Park Road, Singapore 117674

Department of Information Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong

Abstract- **The three-stage Clos network has been studied extensively as a framework for the implementation of large-scale ATM switches. Recently, a quasi-static routing scheme for three-stage Clos networks, called the path-swtiching scheme, has been proposed by Lee and Lam to achieve low switch complexity and high throughput. We derive in this paper the nonblocking conditions when the bandwidth requirements of traffic are rounded up to simplify switch operation. We discuss the implications of these results for switch implementation. In particular, we show how use the results to build a semi-optical network to reduce switch complexity.**

I. Introduction

The three-stage Clos network has been extensively studied as a framework for large-scale ATM switches. For a symmetric three-stage Clos network $C(n, m, k)$, there are k modules in the first stage, each with n inlets and m outlets. The middle stage has m k-by-k switch modules. The third stage has k modules, each with m inlets and n outlets. One link connects each input module to each central module, and one link connects each central module to each output module.

There are mainly two different routing schemes for routing traffic in the switch. One is dynamic routing. In dynamic routing systems, each input module will evenly distribute the arriving cells to different central modules to balance the loading. The central modules and the output modules further route cells to their destinations independently. As the cells may get out of order as they pass through the switch, additional buffers are needed at the output ports to re-sequence the cells. Since it is unpredictable which central module a cell will be routed to, the bandwidth or switch resources can not be reserved and scheduled for each individual Virtual Circuit (VC). When the total bandwidth of 2nd stage is same with that of 1st stage or 3rd stage, the Clos network can achieve non-blocking switching by using the 2nd stage which can forward all traffic from the 1st stage to the 3rd stage. Therefore, the number of central modules needed in dynamic routing systems is

$$m = n, \tag{1}$$

Another scheme is to use static routing [1, 2]. In static-routed networks, all cells belong to the same VC are constrained to follow the same path from the input module all the way to the output module. When a new connection request is arrived, the controller of the static-routed switching system must find a path from its input module to its destined output module with sufficient unused bandwidth on each of its links to accommodate the new VC. In this routing scheme, the bandwidth and switch resources are scheduled and reserved for each individual VC to guarantee the QoS. However, in order to guarantee a path with sufficient bandwidth for any new call arrival, the number of central modules must be very large. For a three-stage Clos network $C(n, m, k)$, minimum allowed VC bandwidth b and maximum allowed VC bandwidth B, the number of central modules needed to avoid call blocking have to satisfy the following inequality [3]

$$m \geq 2 \max_{b \leq \omega \leq B} \lfloor \frac{n - \omega}{\max\{1 - \omega, b\}} \rfloor + 1 \tag{2}$$

This number can be very large with either large B or small b. For example, with input/output module size

0-7803-6283-7/00/$10.00 © 2000 IEEE

of $n = 32$, $B = 0.3$ and $b = 0$, the number of central modules m is about three times the input/output module size n [1]. Analyses for VC blocking for this class of systems can be found in [3, 4]. A tighter result that improves the bound by two in some cases, and some possible call admission algorithms to reduce the number of central modules needed, are discussed in [5].

The path-switching scheme [6] is a compromise between these two routing schemes. Path switching is based on the concept of virtual paths. As shown in

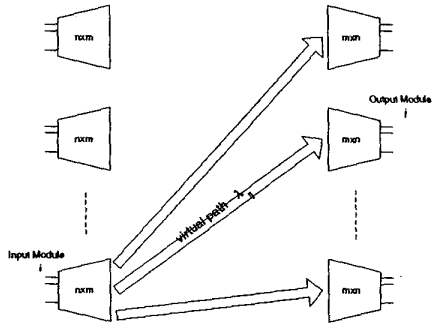

Figure 1: Virtual paths between pair of input-output modules

Fig.1, in a three-stage Clos network model, the virtual path from input module i and output module j must carry the aggregate traffic from i to j, λ_{ij}. For virtual path scheduling, the virtual paths are calculated and scheduled for each pair of input-output modules by setting up necessary paths through central modules. There are two features with this switching scheme. First, the bandwidth of each VC is scheduled and reserved to guarantee the QoS. Second, different VC flows between the same pair of input-output modules can share the bandwidth and switch resources within the same virtual path.

As we can expect, in the path-switching scheme, the number of central modules needed to avoid call blocking will be somehow between the two requirements (1) and (2). In this paper, we derive the necessary and sufficient non-blocking conditions of the whole switch under several different assumptions in the path-switching scheme. Furthermore, we discuss the WDM implementation of our switching scheme. Compared with other WDM ATM switches, our implementation is highly scalable, because the optical cross-connects in our implementation is not based on slot-by-slot routing assignment or contention resolution algorithms. Various schemes are also proposed in this paper to reduce the tuning speed requirements of optical transmitters or receivers.

II. Preliminaries

Our switch model is a three-stage Clos network with k input/output modules, m central modules, and n inlet/outlet for each input/output module. Let λ_{ij} be the total traffic between input module i and output module j. The unit of λ_{ij} is cells per time slot, where a time slot is defined to be the duration of a cell with respect to the speed of an inlet/outlet. $\Lambda = (\lambda_{ij})_{k \times k}$ is the traffic loading matrix.

$$\Lambda = \begin{pmatrix} \lambda_{00} & \lambda_{01} & \cdots & \lambda_{0,k-1} \\ \lambda_{10} & \lambda_{11} & \cdots & \lambda_{1,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{k-1,0} & \lambda_{k-1,1} & \cdots & \lambda_{k-1,k-1} \end{pmatrix} \quad (3)$$

The main issue is how to provide the necessary bandwidth λ_{ij} between any pair of input-output modules by setting up paths through central modules. In other words, to assign the central stage routings such that our three-stage Clos network is non-blocking.

Definition 1 A three-stage Clos network is said to be **non-blocking** for a traffic loading matrix Λ, if there exists a path assignment scheme such that the assigned bandwidth between each pair of input-output modules are larger than or equal to the corresponding entry λ_{ij}.

First, let us frame the routing problem in a three-stage Clos network as a bipartite graph problem. With

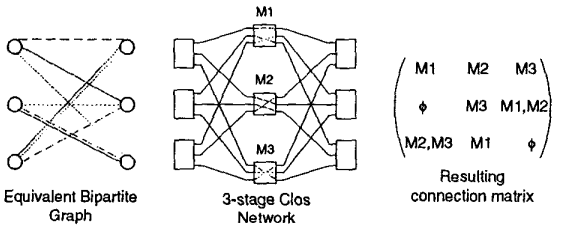

Figure 2: Correspondence between the bipartite graph and the 3-stage Clos network

reference to Fig. 2, let each left/right side node of the bipartite graph denote each input/output module of the 3-stage Clos network. Then, the number of edges linking each pair of node represents the number

of paths connecting the corresponding pair of input-output modules, which equals to the number of cells that can be transferred between the modules within one time slot. Since there are at most m paths connecting to each input/output module, where m is the number of central modules, the vertex degree of the corresponding bipartite graph is less then or equal to m.

Conversely, given a bipartite graph with vertex degree less then or equal to m, it is well known that the graph is m-colorable. Therefore, let each different color represents a different central module. We can map the coloring solution into the routing solution in the Clos network. We then have the connection matrix $C = (C_{ij})_{k\times k}$. The entry C_{ij} is the set of central modules through which the paths are set up between input module i and output module j. Here, we must have

$$|C_{ij}| \geq \lambda_{ij}, \forall i, j \tag{4}$$

in order to provide the necessary bandwidth λ_{ij} between input module i and output module j.

III. Non-blocking conditions

In order to set up paths for the traffic load of unicast calls, we consider the following capacity constraints that the traffic-loading matrix must satisfy. If the link capacity is normalized to one, then the total traffic load leaving each input module can not exceed n. That is, we have the following input-module capacity constraints:

$$\Sigma_j \lambda_{ij} \leq n, \forall 0 \leq i \leq k-1 \tag{5}$$

Similarly, we have the output-module capacity constraints:

$$\Sigma_i \lambda_{ij} \leq n, \forall 0 \leq j \leq k-1 \tag{6}$$

The value of λ_{ij} in general is not integral. However, $|C_{ij}|$ must be integral. Therefore, we consider a scheme called *round-up scheme* to modify the traffic matrix Λ such that each entry in the modified traffic matrix $P = (P_{ij})_{k\times k}$ is given by

$$P_{ij} = \lceil \lambda_{ij} \rceil \tag{7}$$

In other words, we round up the value of λ_{ij} and aim to set up P_{ij} paths between input module i and output module j.

Theorem 1 *Incorporated with round-up scheme, a three-stage Clos network is non-blocking with respect to any possible traffic matrix $(\lambda_{ij})_{k\times k}$ satisfying (5) and (6) if and only if the number of central modules $m \geq n + k - 1$.*

Proof: The matrix $(P_{ij})_{k\times k}$ satisfies

$$\Sigma_j P_{ij} = \Sigma_j \lceil \lambda_{ij} \rceil \leq n + k - 1, \forall 0 \leq i \leq k-1 \tag{8}$$

$$\Sigma_i P_{ij} = \Sigma_i \lceil \lambda_{ij} \rceil \leq n + k - 1, \forall 0 \leq j \leq k-1 \tag{9}$$

for any matrix $(\lambda_{ij})_{k\times k}$ satisfying (5) and (6).

Let P_{ij} be the number of edges of the corresponding bipartite graph, then the bipartite graph has vertex degree less then or equal to $n + k - 1$. Therefore, it is $(n + k - 1)$-colorable. In other words, $n + k - 1$ central modules is necessary and sufficient to assign paths corresponding to any possible traffic matrix $(\lambda_{ij})_{k\times k}$ satisfying (5) and (6). □

If λ_{ij} were integral, it is a well known result that $m = n$ is sufficient to guarantee the non-blocking operation of the switch. The above theorem states that using the round-up scheme the overhead of having non-integral λ_{ij} is $(k-1)$, which may be acceptable if k is small. For large k and n, two ways of scaling the system will be discussed later.

Compared with multirate static routing in three-stage Clos network, the number of central modules needed here with round-up procedure can be much smaller and is independent of the maximum or minimum bandwidth of calls. Compared with dynamic routing in three-stage Clos network, all calls between the same pair of input/output modules are allowed to share the same virtual path in which bandwidth has been reserved. At the same time, cells belonging to different pairs of input/output modules will not affect each other by using different virtual paths.

The predetermined and relatively static routing pattern suggests that the virtual path scheme can best be implemented in optical domain using wavelength division multiplexed (WDM) technique. One such implementation is based on broadcast-and-select (Figure 3) configuration, where each of the m central switch modules is replaced by a $k \times k$ optical star coupler. The transmission from each input module is carried out by an array of m fixed-tuned laser transmitters of identical wavelength, but different wavelengths will be assigned for different input modules. Each of the k output modules consists of an array of m tunable receivers capable of selecting any of the n incoming wavelengths from the star coupler as shown. In this

way, both uni- and multicasting operations can be supported. The fixed-tuned transmitters and tunable receiver (FTTR) configurations is not unique. It can be replaced by a symmetrical tunable transmitter and fixed-tuned receiver (TTFR) configuration. There are, however, variations in the implementation (e.g. protocol) and different technology issues (e.g. tunable transmitter) presented in these configurations.

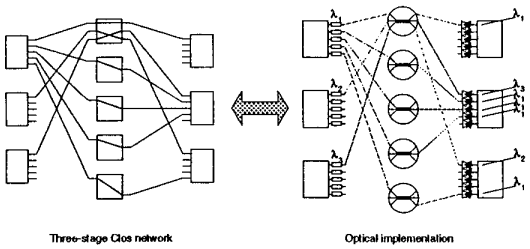

Figure 3: The WDM implementation of the central stage routing in a 3-stage Clos network

One of the benefits of using WDM is the enormous bandwidth available in the optical fibers and waveguides that allow speed-up operation of the central modules. Let s be the speed-up factor relative to the transmission speed of the external link. By speeding up the central modules, the number of central modules needed in the three-stage Clos network to achieve the non-blocking property can be reduced.

Theorem 2 *With central-module speedup factor s, a three-stage Clos network is non-blocking with respect to any possible traffic matrix $(\lambda_{ij})_{k\times k}$ satisfying Eq. (5) and (6) if and only if the number of central modules $m \geq \lceil \frac{n}{s} \rceil + k - 1$.*

Proof: With speedup factor s, the number of virtual paths needed between input module i and output module j in a time slot becomes

$$P_{ij} = \lceil \frac{\lambda_{ij}}{s} \rceil, \tag{10}$$

and P_{ij} must satisfy the following row sum, column sum constraints:

$$\Sigma_j P_{ij} = \Sigma_j \lceil \frac{\lambda_{ij}}{s} \rceil \leq \lceil \frac{n}{s} \rceil + k - 1, \forall 0 \leq i \leq k-1 \tag{11}$$

$$\Sigma_i P_{ij} = \Sigma_i \lceil \frac{\lambda_{ij}}{s} \rceil \leq \lceil \frac{n}{s} \rceil + k - 1, \forall 0 \leq j \leq k-1 \tag{12}$$

Therefore, the corresponding bipartite graph has vertex degree less then or equal to $\lceil \frac{n}{s} \rceil + k - 1$, hence $(\lceil \frac{n}{s} \rceil + k - 1)$-colorable. Thus it is also the minimum number of central modules required. □

Theorem 2 implies that for $s = n$, the number of central modules needed is $1 + k - 1 = k$ which is the number of input/output modules. The result is quite intuitive because when $s = n$ and $m = k$, each input-output modules pair can be connected by a fixed link that transmission rate is n times faster than the external link without any blocking.

The previous description is restricted to cases where the connection pattern of each central module is static. Suppose we allow the connection pattern of each central module to change from slot to slot and introduce the concept of a frame, which composes of f time slots. The central-module connection pattern is scheduled and varied from slot to slot in each frame and the group of connection patterns are repeated from frame to frame. The round-up scheme is modified as follows:

We multiply each entry λ_{ij} of the traffic loading matrix by an integer f. Then $f\lambda_{ij}$ is the total traffic between input module i and output module j at each frame. After round up, we need $P_{ij} = \lceil f\lambda_{ij} \rceil$ paths to be set up between input module i and output module j in f time slots. These virtual paths satisfy constraints:

$$\Sigma_j P_{ij} = \Sigma_j \lceil f\lambda_{ij} \rceil \leq fn + k - 1, \forall 0 \leq i \leq k-1 \tag{13}$$

$$\Sigma_i P_{ij} = \Sigma_i \lceil f\lambda_{ij} \rceil \leq fn + k - 1, \forall 0 \leq j \leq k-1 \tag{14}$$

Therefore, the corresponding bipartite graph, called capacity graph, has vertex degree less then or equal to $fn + k - 1$, and it is $(fn + k - 1)$-colorable.

Theorem 3 *With introduction of frames, a three-stage Clos network is non-blocking with respect to any possible traffic matrix $(\lambda_{ij})_{k\times k}$ satisfying (5) and (6) if and only if the number of central modules $m \geq n + \lceil \frac{k-1}{f} \rceil$.*

Proof: We will use the principle of time-space interleaving, which is proposed in [6], to prove the theorem. Consider any particular color assignment $a \in \{0, 1, \cdots, fn + k - 2\}$ of an edge between input node I_i and output node O_j of the capacity graph, we try to decompose it to either different time slot or different central module. If the number of central modules m satisfies the inequality

$$fm \geq fn + k - 1, \tag{15}$$

then we can always find a pair of integers $r \in \{0, 1, \cdots, m-1\}$ and $t \in \{0, 1, \cdots, f-1\}$ such that

$$a = r \cdot f + t. \tag{16}$$

That is, the color assignment a, or equivalently the assignment pair (t, r), of the edge between I_i and O_j indicates that the central module r has been assigned to a route from I_i and O_j in the t^{th} time-slot of every frame. As illustrated by the example shown in Fig. 4, where $m = 3$ and frame size $f = 2$, the decomposition of the edge-coloring into assignment pairs guarantees that route assignments are either space interleaved or time interleaved.

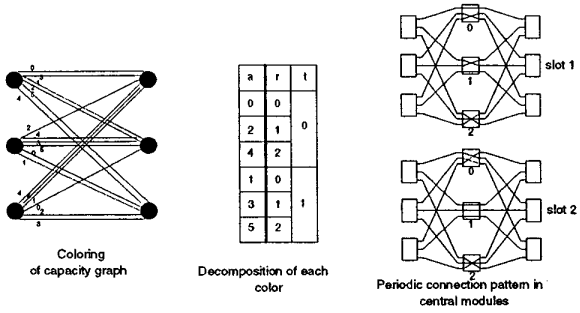

Figure 4: Illustration of Time-Space Interleaving Principle

On the other hand, we can see that if $fm < fn+k-1$, then we cannot decompose all colors to different central modules or different time slots. Therefore, $m \geq n + \frac{k-1}{f}$ is the necessary and sufficient condition for successful setting up the virtual paths. □

For WDM implementation, whether FTTR or TTFR is a preferred configuration depends on the performance requirements (e.g. call setup time and the protocol design). The high data rates will impose stringent performance requirements on both tunable transmitters and tunable receivers if tuning has to be completed from slot to slot over the entire wavelength range of operation. One way to get around this requirement is to introduce a "sustained period" u in the WDM system during which no tuning is performed. Instead of changing the connection pattern slot by slot for every f slots, we maintain each connection pattern for u time slots. In a frame, there are f connection patterns, with each connection pattern lasts for u continuous time slots. There is a guard time of g time slots between each of the connection pattern for wavelength tuning (Fig. 5)

The guard time is needed in practice not only to account for the finite tuning speed of optical components, but also to allow for synchronizing to take place after tuning. That is, the input modules may not synchronize perfectly the frame boundaries with sustained period boundaries. Some guard time is needed for the detection of new sustained period boundary upon tuning. Now, in total $fu+fg$ time slots in a frame, there are total $(fu+fg)\lambda_{ij}$ traffic between the input module i and output module j for the traffic loading matrix Λ. In other words, $\lceil \frac{(fu+fg)\lambda_{ij}}{u} \rceil$ paths need to be setup between them. Considering again the speedup factor s, the constraints become:

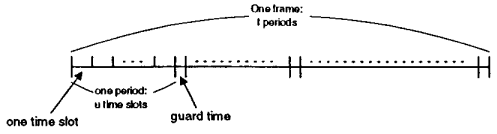

Figure 5: The Composition of a Frame

$$\Sigma_j P_{ij} = \Sigma_j \lceil \frac{(fu+fg)\lambda_{ij}}{us} \rceil \leq \lceil \frac{(fu+fg)n}{us} \rceil + k - 1, \quad \forall 0 \leq i \leq k-1; \tag{17}$$

$$\Sigma_i P_{ij} = \Sigma_i \lceil \frac{(fu+fg)\lambda_{ij}}{us} \rceil \leq \lceil \frac{(fu+fg)n}{us} \rceil + k - 1, \quad \forall 0 \leq j \leq k-1. \tag{18}$$

Similarly, for the number of central modules equals to m, there are total fm possible paths leaving and entering each input/output module. Therefore, the following inequality must be satisfied for successful path assignment:

$$fm \geq \lceil \frac{(fu+fg)n}{us} \rceil + k - 1 \tag{19}$$

In other words, we have the following non-blocking condition in this case:

Theorem 4 *Given frame size f, the central-module speedup factor s, the sustained period u and the guard time g between any two successive sustained period, a three-stage Clos network is non-blocking with respect to any possible traffic matrix $(\lambda_{ij})_{k\times k}$ satisfying Eq. (5) and (6) if and only if the number of central modules $m \geq \lceil (1+\frac{g}{u})\frac{n}{s} + \frac{k-1}{f} \rceil$.* □

So far, we have introduced the speedup factor s, frame size f, guard time g, sustain period u, as the control parameters of the number of central modules m to satisfy the non-blocking conditions. Obviously, increasing s or f reduces m, but the implementation complexity is generally high. Furthermore, when f goes up, the quasi-static routing scheme is more likely

to be a dynamic routing scheme. It is difficult to provide quality of service (QoS) guarantee in this case. On the other hand, given fixed g, smaller u causes lower utilization which implies that larger m is needed. However, larger u causes unwanted delay jitter which could adversely affect QoS, due to the virtual paths leaving an input module can only change their destinations after $g+u$ time slots.

By theorem 4, we can choose appropriate speedup factor s and frame size f to minimize the number of central modules needed. For example, let $s = f = \lceil(1+\frac{g}{u})n+k-1\rceil$, then by theorem 4,

$$m \geq \lceil(1+\frac{g}{u})\frac{n}{s}+\frac{k-1}{f}\rceil = \frac{\lceil(1+\frac{g}{u})n+k-1\rceil}{\lceil(1+\frac{g}{u})n+k-1\rceil} = 1 \quad (20)$$

This implies that one central module will be enough to make the overall switch non-blocking. It should be noted that the required tuning speed and tuning range of the combined WDM-TDM scheme is then more stringent. The recent advent in selectable WDM receivers [7] suggests that these devices could be useful in implementing such scheme. Such receiver is based on a passive optical demultiplexer followed by photodetector array and amplifier array circuits. The number of supported wavelengths is scalable, and the switching can be very swift ($\sim ns$).

IV. Conclusions

In this paper, we have investigated how the non-blocking property can be achieved in a three-stage clos network when a path switching scheme is used to route cells of multirate connections. Techniques for performance enhancement, such as speedup operation of the central modules and frame-size setting, have been studied in detail. In particular, their quantitative effects on the reduction of the central modules required are derived. Qualitatively, it is found that the speeding up of the central modules can substantially reduce the number of modules required, with the implication that optical implementation is worth serious consideration as far as the central stage is concerned.

A particular interesting result is the use of the so-called round-up scheme to round up the bandwidth requirements of connections when assigning bandwidths. This has the effect of simplifying the overall switch operation with minimal additional switch complexity. For instance, it is found that simple static routing is perfectly acceptable when the round-up scheme is used. This has an important implication if WDM optical technology is used for the central stage: static routing implies no wavelength tuning is required.

As an extension, this paper has also studied what if tuning is allowed and to what extent further decrease in switch complexity can be achieved. In particular, we have introduced the concept of sustained non-tuned period and guard time to lower the tuning-speed requirements.

References

[1] Yoshito Sakurai, Nobuhiko Ido, Shinobu Gohara, Noboru Endo, "Large-Scale ATM Multistage Switching Network with Shared Buffer Memory Switches," *IEEE Communications Magazine*, Jan. 1991.

[2] H. Kuwahara, N. Endo, M. Ogino and T. Kozaki, "Shared Buffer Memory Switch for an ATM Exchange," *Proc. Int. Conf. on Communications*, Boston, MA, June 1989, pp. 4.4.1-4.4.5.

[3] Riccardo Melen and Jonathan Turner, " Nonblocking Networks for Fast Packet Switching," *IEEE INFOCOM'89*, pp. 548-557, April 1989.

[4] Riccardo Melen, Jonathan Turner. "Nonblocking Multirate Distribution Networks," *IEEE Trans. on Commun.*, Vol. 41, no. 2, pp. 362-269, Feb. 1993.

[5] Soung C. Liew, Ming-Hung Ng, and Cathy W. Chan, "Blocking and Nonblocking Multirate Clos Switching Networks," *IEEE/ACM Trans. on Networking*, pp. 307-318, vol. 6, no. 3, June 1998.

[6] Tony T. Lee and Cheuk H. Lam, "Path Switching - A Quasi-static Routing Scheme for Large-Scale ATM Packet Switches," *IEEE JSAC*, pp. 914-924, vol. 15, June 1997.

[7] F. Tong, " Multiwavelength receivers for WDM systems," *IEEE Communication Magazine*, Dec. 1998.

[8] Charles A. Brackett, "Dense Wavelength Division Multiplexing Networks: Principles and Applications," *IEEE JSAC* Vol. 8, No. 6, August 1990.

[9] Youngbok Choi, Hideki Tode, Hiromi Okada, and Hiromasa Ikeda, "A Large Capacity Photonic ATM Switch Based on Wavelength Division Multiplexing Technology," *IEICE Trans. Commun.*, Vol. E79-B, No. 4, April 1996.

[10] Arturo Cisneros and Charles A. Brackett, "A Large ATM Switch Based on Memory Switches and Optical Star Couplers," *IEEE JSAC* Vol. 9, No. 8, October 1991.