Dynamic Optical Switching: The Network is the Memory

M. Baldi[†‡] and Y. Ofek[†]

[†]Synchrodyne Networks, Inc.
New York, New York, USA
{baldi,ofek}@synchrodyne.com

[‡]Turin Polytechnic, Computer Engineering Dept.
Turin, Italy
mario.baldi@polito.it

Abstract: **This paper shows that the transition to dynamic all-optical packet switching, that fully exploits the benefits of fast dynamic optical components, will require changing the current asynchronous packet switching paradigm. This is shown by focusing on the relatively simple problem of optical memory, while assuming that the more complex problems associated with optical packet header processing and fast switching in the optical domain have been resolved–which is indeed not the case. Various comparative measures are used in the optical memory evaluation. For example, much more glass (Silicon) is needed for optical memory than for electronic memory; in fact the ratio is on the order of 1 million. As shown, a common time reference (CTR) is needed for realizing the optical random access memory (O-RAM) required for optical packet switching. Furthermore, a CTR is used for both minimizing the O-RAM requirements and distributing the O-RAM in the network realizing a new optical network architecture called Fractional Lambda Switching (Fl S).**

1. Introduction

The more optical transmission systems are deployed, the greater the bottlenecks in electronic processing and switching become. Consequently, a solution to these bottlenecks is sought in the optical domain.

Definition 1: *Dynamic optical switching* – Each data unit is transmitted end-to-end over an arbitrary topology network through optical fibers and switches with no conversion to electronics such that the network's switches may treat each data unit on a given optical channel differently.

Definition 2: *Static optical switching* – All data units that are transmitted end-to-end over an arbitrary topology network through optical fibers and switches with no conversion to electronics such that all data units on a given optical channel are treated the same way by the network's switches.

Clearly, *dynamic optical switching* is much closer to current packet and circuit switching than *static optical switching*. Furthermore, *static optical switching* requires the provisioning of a separate optical channel—a.k.a. wavelength or lambda (λ)—from each source to each possible destination. Consequently, if each source is also a destination then *n* sources require (n^2-n) optical channels. This *n* square requirement is a major limit to the scalability of *static optical switching* and hence a compelling reason for resorting to *dynamic optical switching*.

2. Comparative Evaluation of Optical Memory

In the context of this work the following two definitions are applied:

Definition 3: Bulk Optical Memory (BOM). An optical fiber operates as a delay line, and capable of storing a predefined amount of bits encoded within an optical signal. The access to such bulk optical memory is strictly sequential, or first-in-first-out (FIFO).

Definition 4: *Optical Random Access Memory (O-RAM).* An optical memory wherein each of the stored data units can be accessed at any predefined time, independent of the order in which they had entered the O-RAM. (In general in RAM *any data unit* can be accessed *at any time.*)

2.1 Device Level Analysis of Bulk Optical Memory

The physical dimension of optical memory is compared with that of electronic memory (i.e., a Dynamic RAM (DRAM)) with equivalent capacity. A synchronous DRAM chip capable of storing 256 Mbits is manufactured with state of the art technology on a $10\cdot10^{-3}\cdot10\cdot10^{-3}\cdot0.5\cdot10^{-3}=50\cdot10^{-9}$ meter^3 (or $50\cdot10^{-6}$ Liter) silicon chip. A 256 Mbit optical memory for an optical signal encoded at 10 Gb/s is realized with a $256\cdot10^6\cdot2\cdot10^{-2}$=5,120,000 meter fiber. Since the fiber diameter, core and cladding, is $125\cdot10^{-6}$ meter, the total volume is $\pi\cdot(125\cdot10^{-6}/2)^2\cdot5{,}120{,}000=62.8\cdot10^{-3}$ meter^3 (or 62.8 Liters). Hence, the step from DRAM to optical memory corresponds to an increase of more than 1,000,000 folds in the memory's physical size.

2.2 Sub-system Analysis of Bulk Optical Memory

The device level comparison between electronic memory and optical memory clearly indicates that in order to provide the required memory capabilities, optical packet switches are going to be many orders of magnitude larger than conventional electronic packet switches. Table 1 shows the amount of memory per optical channel on state-of-the-art terabit packet switches. Various physical dimensions, such as, length and weight, of the BOM required to implement the same amount of memory for a single 10 Gb/s optical channel are shown in Table 1 (using parameters from Corning).

When the switches in Table 1 have 1000 optical channels the numbers will increase 1000 folds. For example, the memory for an all-optical switch with 256 MB of optical memory per channel will be 3,686 tons. A network with

2,000 such switches would weigh as much as the Great Pyramid of Khufu in Giza, which weighs 6 million tons. Obviously, weight is not the only realization constraint; fiber length is another. Amplification is required to compensate the attenuation introduced by transmission over long fibers. Moreover, the signal degenerates due to the distortion introduced along the fibers and by the optical amplifiers; hence, after traveling a predefined distance it is necessary to regenerate the signal.

Table 1: O-RAM requirement with 10 Gb/s channel for current terabit switches

Product	Memory per channel [MB]	Fiber Length per channel [Km]	Weight per ch. [Kg]	Weight per 1000 chs. [Kg]
M160 (Juniper)	128	20,480	1,843	1,843,000
20000 Terabit Network Router (Pluris)	128	20,480	1,843	1,843,000
GSR12016 (Cisco)	256	40,960	3,686	3,686,000
Aranea-1 (Charlottes' Web)	1000	160,000	14,400	14,400,000

2.3 Sub-system Analysis of Linear Optical Random Access Memory – O-RAM

Random access implies that at any given time any part of the memory can be accessed. Since the stored optical data units travel along the optical fiber at the speed of light, as shown in Figure 1(A), at any given time the data units are in another position along the optical fiber. Consequently, random access has two basic requirements, as shown in Figure 1(B):

1. *Infinite number of taps:* a tap realized by a 1-by-2 switch enables the light that is stored in the fiber to either continue along the fiber or be switched out of the fiber.

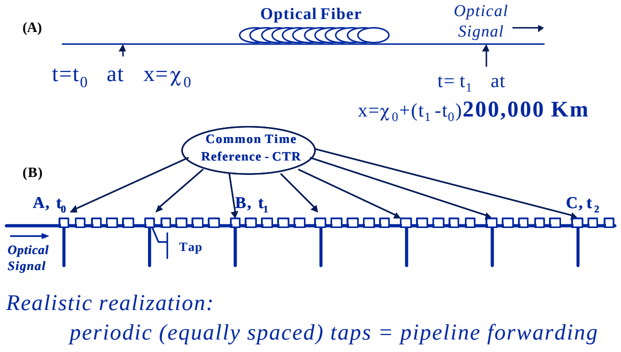

Figure 1: Linear O-RAM (B) – is time dependent (A)

2. *Common Time Reference (CTR)*: a precise knowledge of time is required in order to access a given data unit at a given position along the optical fiber. This knowledge of time should be based on the same time reference along the optical fiber, which is why; this timing requirement is called Common Time Reference or CTR.

Theorem 1: *Assuming that data units are stored in a linear O-RAM, then it is not possible to correctly switch out, at an arbitrary time, the stored data units without CTR.*

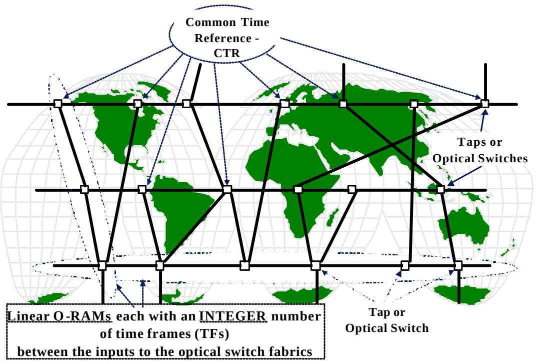
Common Time Reference - CTR
Taps or Optical Switches
Linear O-RAMs each with an INTEGER number of time frames (TFs) between the inputs to the optical switch fabrics
Tap or Optical Switch

Figure 2: Linear O-RAMs for FλS network

3. Time and Fractional λ Switching - FλS

This section studies the relationship between time measurements and scheduling in communications networks, in general, and a new optical network architecture called FλS, in particular. The broad approach is taken in order to provide the rationale for FλS whose principles of operation will be described in Section 4.

3.1 Why Time?

There are a number of answers; a simple one is that time minimizes the memory requirement, which is important for implementation in the optical domain. For example, the previously presented comparison of the amount of silicon needed for solid state memory versus the amount of silicon needed for optical memory (i.e., optical fiber) shows that asynchronous packet switching is not practical in the optical domain. Hence, the step from DRAM to optical memory is not practical without some changes that reduce the memory requirements. Using time and scheduling does this.

3.2 Time Measurement

Measuring time between two events in the same location is performed locally by counting periodic rotations of various sorts. In ancient era the time was measured by counting the earth rotations, or, as some argued, the sun rotations around the earth. Since then, the measurement of time has improved dramatically.

Today, a common time reference has been established by the *time-of-day* international standard that is called *Coordinated Universal Time* or *UTC* (a.k.a. *Greenwich Mean Time* or *GMT*). Specifically, time is measured by counting the oscillations of the cesium atom in multiple locations. In fact, 9,192,631,770 oscillations of the cesium atom define one UTC second. UTC is available everywhere around the globe from several distribution systems, such as GPS (USA satellites system) [1], GLONASS (Russian Federation satellites system) [2], and in the future by Galileo (European Union and Japanese satellites system) [3], or can be distributed via communications satellites through the Two-Way Satellite Time and frequency Transfer (TWTFT) method [4].

3.3 Scheduling

Scheduling requires the ability to measure time. We consider scheduling with two time measurement methods:

1. <u>Scheduling with local time based measurements.</u> The delay between nodes cannot be measured, and therefore, the scheduling is based on local time. This method is used in circuit switching (e.g., SONET).
2. <u>Scheduling with UTC-based measurements</u>. The delay between nodes can be measured by using UTC and scheduling can be based on UTC. Scheduling with UTC implies no clock slips or drifts, and consequently, very simple implementation. This method is used in FλS.

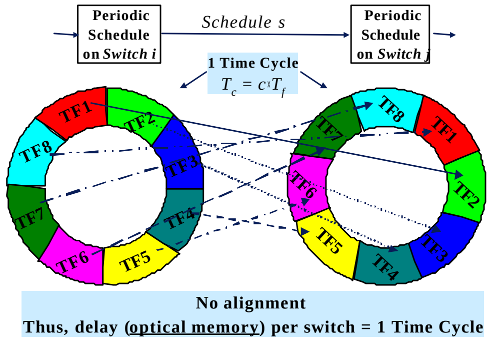
Periodic Schedule on Switch i
Schedule s
Periodic Schedule on Switch j
1 Time Cycle
$T_c = c \times T_f$
TF1
TF2
TF3
TF4
TF5
TF6
TF7
TF8
No alignment
Thus, delay (optical memory) per switch = 1 Time Cycle

*Figure 3: Local time based-scheduling - **SONET***

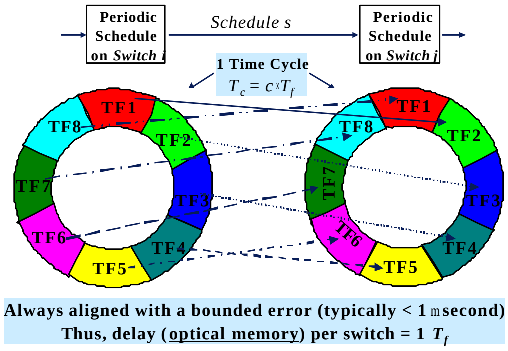
Periodic Schedule on Switch i
Schedule s
Periodic Schedule on Switch j
1 Time Cycle
$T_c = c \times T_f$
TF1
TF2
TF3
TF4
TF5
TF6
TF7
TF8
Always aligned with a bounded error (typically < 1 msecond)
Thus, delay (optical memory) per switch = 1 T_f

*Figure 4: UTC-based scheduling - **F l S***

Figure 3 and Figure 4 are examples of the above two scheduling methods[1]. In these examples, scheduling is periodic and time is divided into time frames (TF) of predefined duration T_f. For example, a time frame of 10 µseconds is obtained by dividing one UTC second by 100,000. For periodic scheduling time frames are grouped into *time cycles*; for example, 1,000 time frames of 10 µseconds create a 10 millisecond time cycle.

3.4 Per Switch Delay and Optical Memory

Let's assume that two neighboring switches, *Switch i* and *Switch j*, perform a given task — e.g., switching or transmitting data units — during predefined time frames according to a schedule, *Schedule s*. *Schedule s* repeats every time cycle, T_c, where $T_c = c \times T_f$. In the examples in Figure 3 and Figure 4, $T_c = 8 \times T_f$, and *Schedule s* on *Switch i* during time frame *k*, is scheduled on *Switch j* during time frame (*k+1*) mod *8*.

When the scheduling on *Switch i* and *Switch j* is based on local time, the delay between *Schedule s* on *Switch i* and on *Switch j* is not known, and consequently, the delay between TF *k* and TF (*k+1*) mod *c* is not known. Since the schedule repeats every time cycle, the maximum delay between a TF on *Switch i* and the corresponding TF on *Switch j* is one time cycle, T_c – where, $T_c = c \times T_f$.

When the scheduling on *Switch i* and *Switch j* is based on UTC, the delay between *Schedule s* on *Switch i* and on *Switch j*, is known, and consequently, the delay between TF *k* and TF (*k+1*) mod *c* is known. Consequently, the maximum time between the execution of the aforementioned task in *Switch i* and in *Switch j* is only one time frame – T_f (which results from the actual data unit propagation delay between the two switches not being an integer number of time frames – a.k.a. quantization delay). Since data units need to be stored while waiting for the task execution in *Switch j*, the time between the two task executions determines the amount of (optical) memory required within the switches.

3.5 SONET

SONET switches operate according to a reoccurring schedule that, as was mentioned before, is based on a local clock; consequently, data traversing a SONET switch are delayed up to a whole time cycle. Due to byte-by-byte channel multiplexing, the SONET time cycle is the time between the transmission of two successive bytes of the same channel. For example, the time cycle — hence the scheduling delay — of an STS-1 switch is 125/810 = 154 ns, independently of the line rate of its interfaces.

[1] Without loss of generality, the propagation delay between *Switch i* and *Switch j* was ignored.

Byte-by-byte de-multiplexing of STS-N frames into multiple STS-1 frames cannot be done in the optical domain. Consequently, in order to implement SONET-based dynamic optical switching, each incoming byte must be independently switched from input to output. This requires, for OC-192 line rates, optical switching and processing time well below 100 picoseconds, which is far beyond current technology.

In order to overcome the picosecond accuracy requirement, a SONET look-alike might be devised in which the multiplexed one byte slot size is increased by a factor of *x*. However, this will imply a factor of *x* increase in the time cycle, and, since SONET scheduling uses local time measurements, in the per switch delay and optical memory requirements. For example, if an STS-1 frame is used as multiplexed unit, i.e., *x*=810, the time cycle — hence the scheduling delay — of such a SONET switch is 125 µs.

Note that increasing the slot size of SONET by a factor of *x* will anyway not eliminate the need for optical processing of overhead information, such as, Synchronous Payload Environment pointers. These pointers are needed since local time measurements on different switches are continuously drifting from one another.

4. FRACTIONAL LAMBDA SWITCHING (Fl S): THE NETWORK IS THE MEMORY

In Section 2 it was shown that: (1) the physical size of optical memory is a million folds larger than electronic memory and (2) O-RAM is time dependent and requires a CTR. Consequently, there is a need to significantly reduce the optical memory size, which can be also achieved by using the CTR. In other words, both the realization of O-RAM and minimizing its size depends on the CTR. Fractional Lambda Switching (FλS) is a new optical network architecture basd on UTC as a CTR. There are several ways to explain the operation principles of FλS; in the context of this paper, the most natural way appears to be describing FλS as an optical memory that is distributed over the network, as shown in Figure 2. More specifically, such a network can be viewed as a mesh of linear O-RAMs where the taps are optical switches.

In fact, a linear O-RAM is implemented by connecting a sequence of optical switches via predefined length fibers– each switch providing access to a stored data unit at a given time. Propagation through a predefined length fiber requires a predefined time. The CTR is divided into time units called *time frames* (TFs), each with a predefined duration - T_f. Distributing the memory over a network requires the following O-RAM changes, as shown in Figure 2:

1. The 1-by-2 taps are replaced with N-by-N optical switches in order to enable the realization of a mesh topology.
2. The delay between the inputs of any two optical switches should be an integer number of time frames.

Providing a CTR in one box is simple, but doing it over a global network is less obvious. However, since UTC (Coordinated Universal Time, a.k.a. time-of-day) is available globally, it can be used as CTR for FλS networks. As said previously, UTC is distributed directly by various satellite constellations: GPS (USA satellites system) [1], GLONASS (Russian Federation satellites system) [2], and in the future Galileo (European Union and Japanese satellites system) [3]. There are other means for local distribution of UTC, such as, CDMA.

The operation of FλS is based on the following principles for realizing Pipeline Forwarding (PF) of time frames:

1. Switching of time frames: (*i*) each TF has a predefined duration and contains a *payload* with a predefined size (i.e., number of bytes), (*ii*) between two successive payloads there is a *safety margin* or *idle time* of a predefined duration, and (*iii*) the payload of a TF is switched as a whole from input to output, as shown in Figure 5.
2. **The Idle time or safety margin between two successive payloads** allows the switching matrix of the optical switch to be changed so that the payload of each TF of a given optical channel can be switched to a different output.
3. **CTR is UTC** and is coupled to all the (optical) switches. The UTC second is divided into a predefined number of equal duration time frames—T_f. Time frames are grouped into *time cycles* and time cycles are grouped into *super cycles*, wherein the super cycle is equal to one UTC second.
4. **Alignment of all received time frames to UTC**, such that the delay between inputs of adjacent *optical* switching *fabrics* (and consequently, between the inputs of any two optical switching fabrics) after alignment is an *integer number of time frames*, as shown in Figure 2. The alignment operation is performed before the optical switching, as shown in Figure 5.

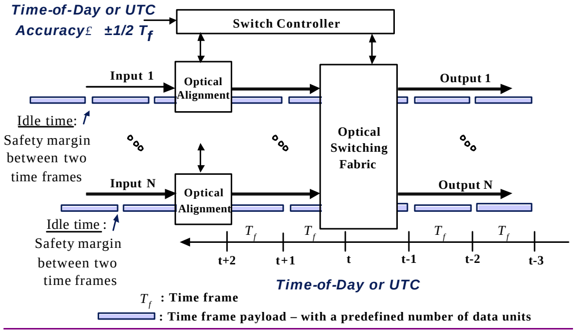

Figure 5: Incoming time frames are aligned with UTC before reaching the optical switching fabric

5. **Switching pattern in the optical switch fabric** repeats every time cycle or super cycle. In other words, during every time frame within a time cycle (or super cycle) the optical switching fabric has a predefined input/output configuration and the sequence of input/output configurations repeats every time cycle (or super cycle). This implies that for every time frame within a time cycle the time frame payload is switched to a predefined output.

6. **Scheduling of a fractional l pipe (FlP)**, *p*, is defined along a path of successive FλS switches: $S_p(1)$, $S_p(2)$, … , $S_p(k)$ such that, the forwarding of a time frame along *p* has a predefined schedule. More specifically, let (*i*) the delay, in time frames, between successive inputs of the optical switching fabric along *p* be $d_{1,2}$, $d_{2,3}$, … , $d_{k-1,k}$, (*ii*) the time cycle and super cycle duration be *c* and *s* time frames, respectively, and (*iii*) the scheduling per FλP repeat itself every *c* time frames, then, a time frame that is forwarded along path *p* from $S_p(1)$ at t_0 will be forwarded by $S_p(2)$ at $(t_0+d_{1,2}) \bmod c$, by $S_p(3)$ at $t_0+d_{1,2}+d_{2,3} \pmod c$, and so on; and will reach the last switch of *p*, $S_p(k)$, at $t_0+d_{1,2}+d_{2,3}+\ldots+d_{k-1,k} \pmod c$.

5. DISCUSSION

The evaluation of the optical memory required for optical packet switching has shown that, due to optical memory limitations, the currently deployed asynchronous packet switching paradigm cannot be realized in the optical domain. Fractional λ switching (FλS) was proposed as a switching paradigm that minimizes the need for optical memory. Moreover, FλS reduces the complexity of switching and eliminates the need for header processing that is the main unsolved problem in the optical domain. Thus, likewise static or whole λ switching, FλS can be realized, i.e., dynamic all-optical networking with FλS is viable with state of the art optical components. FλS enables bandwidth provisioning from fractional STS-1 to full channel capacity.

FλS uses UTC (Coordinated Universal Time) to implement *pipeline forwarding* (PF) of time frames, virtual containers of 5-20 Kbytes each. Pipeline forwarding, over a meshed FλS network, requires that the delay between any two switching fabric inputs be an integer number of time frames, which is realized with an *alignment to UTC* operation before each switching fabric input. Since the time frame boundaries are explicitly identified, a relaxed UTC accuracy of less than one half of a time frame suffices. Dynamic optical switching with FλS: (*i*) provides scalable switching with minimum complexity (i.e., a Banyan network can be deployed) – thereby solving the switching bottleneck, (*ii*) provides minimum complexity aggregation and grooming in the time domain – thereby solving the link bottleneck at the edges of the network, and (*iii*) is compatible with current public standards, such as IP/MPLS and related protocols.

REFERENCES

[1] National Institute of Standards and Technology (NIST), "GPS Data Archive," USA, `http://www.boulder.nist.gov/timefreq/service/gpstrace.htm`

[2] Russian Federation Ministry of Defense - Coordination Scientific Information Center, "Global Navigation Satellite System – GLONASS," Russian Federation, `http://www.rssi.ru/SFCSIC/english.html`

[3] European Union, "Transport-Satellite Navigation," Union Policies, Brussels, Belgium, June 2001, `http://europa.eu.int/scadplus/leg/en/lvb/l24205.htm`

[4] National Physical Laboratory, "Two-Way Satellite Time and frequency Transfer (TWTFT)," Teddington, Middlesex, UK, `http://www.npl.co.uk/npl/ctm/twstft.html`