Contribución a la Conmutación Óptica de Paquetes. Arquitecturas, Evaluación de Prestaciones y Análisis Comparativo

Universidad Politécnica de Cartagena

Departamento de Tecnologías de la Información y las Comunicaciones

Pablo Pavón Mariño

Director

Dr. Juan García Haro

2004

A mi familia

Agradecimientos

Quiero expresar mi gratitud a todas las personas que me han acompañado a lo largo de estos años.

He sido afortunado por contar con el consejo y la orientación de Joan García Haro, Director de esta Tesis Doctoral. Más allá de los contenidos desarrollados, están los valores y las actitudes. Mi relación con él marcará sin duda mi vida como investigador. Su honestidad en el trabajo, es mi referente. Con toda sinceridad, gracias Joan.

Quisiera agradecer también su ayuda y colaboración a todos los compañeros del Área de Ingeniería Telemática, y del Departamento de Tecnologías de la Información y las Comunicaciones. En estos años, con el ritmo de trabajo que exige una Universidad en creación como la nuestra, ha habido momentos duros para todos. Para mí, tiene por eso mayor valor vuestro apoyo, y sabéis que siempre contáis con el mío.

De todo corazón, gracias a todos los buenos amigos, por los momentos que hemos pasado, y por los que pasaremos. En especial, y por orden alfabético, Cristina, Ginés, Juan y Lola, que más me han aguantado en estos años, ¡y lo que les queda!.

Por encima de todo, gracias a mi familia, por el cariño, por el ánimo, por la comprensión. Sin vosotros, habría sido imposible. Gracias a ti, Susana, que le das sentido a todo esto.

ÍNDICE

Capítulo 1. Introducción

1.1 Multiplexación por División en Longitud de Onda

El impresionante crecimiento de Internet ha estimulado el avance en tecnologías que permitan soportar la demanda de tráfico. La transmisión por fibra óptica ha dado enormes progresos en este sentido, ofreciendo un canal de baja atenuación, baja distorsión de señal, bajo coste, y enorme ancho de banda. Como muestra la figura 1-1 el ancho de la ventana de transmisión para fibras monomodo se encuentra alrededor de los *50* THz, lo que podría permitir velocidades binarias en el orden de decenas de *Tbps*.

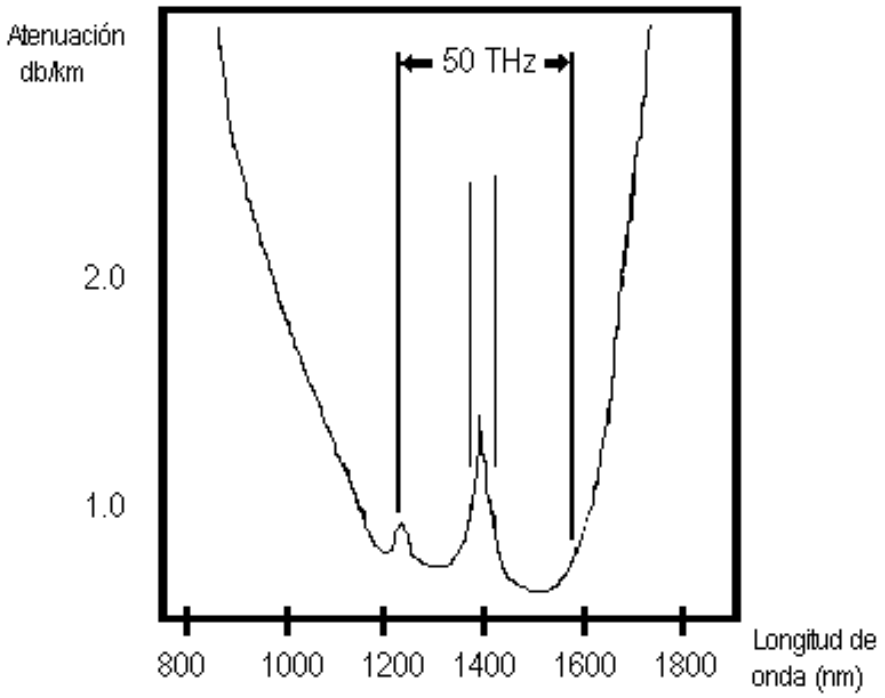


Figura 1-1. Atenuación en fibra monomodo

Por otro lado, la capacidad de procesamiento electrónico no ha crecido de manera similar, situándose en el orden de decenas de *Gbps*. Para intentar atacar este desajuste entre capacidad de transmisión óptica y capacidad de procesamiento electrónico, se requiere por tanto un sistema de multiplexación adecuado, que combine la transmisión de múltiples fuentes de tráfico, en el orden de los *Gbps*, en un agregado en el orden de (potencialmente) *Tbps*. De manera natural, recogiendo las estrategias de multiplexación aplicadas en otros canales, las tres tendencias bajo estudio en el entorno óptico son [Siv00][Ram01][Mur02]:

- Multiplexación por división en frecuencia, llamada a frecuencias ópticas multiplexación por división en longitud de onda (en adelante WDM, *Wavelength Division Multiplexing*).
- Multiplexación por división en tiempo (en adelante TDM, *Time Division Multiplexing*).
- Multiplexación por división en código (en adelante CDM, *Code Division Multiplexing*).

En una multiplexación óptica TDM, la operación de *mux/demux* necesaria para transmitir y extraer información de cada canal exige ser capaz de sincronizarse con la señal binaria multiplex, en el orden de *Tbps*, fuera del ámbito de operación de los circuitos electrónicos. En la multiplexación CDM, la operación de *mux/demux* debe trabajar con una señal de *chirp*, de nuevo a una velocidad mayor que la velocidad binaria de cada canal multiplexado, y mayor que la velocidad de procesamiento electrónico. Sin embargo, en la multiplexación WDM, la operación de *mux/demux* se realiza mediante filtros pasivos en longitud de onda (de coste relativamente bajo), que combinan/separan cada canal de usuario. De esta manera se elimina la necesidad de sincronización global, provocando que el límite en cuanto a la velocidad de procesamiento electrónico afecte únicamente a la velocidad binaria de la señal de cada canal individual.

Por estas razones, asociadas al escenario de desajuste entre velocidad de procesamiento electrónico vs. velocidad de transmisión óptica en el que nos encontramos, la opción que ha recibido mayor interés en el ámbito investigador y comercial es la multiplexación WDM.

1.2 Conmutación en redes WDM

Una red troncal WDM de interconexión arbitraria (*mesh*), como la que muestra la figura 1-2, se encuentra formada por un conjunto de nodos de conmutación interconectados por enlaces WDM. Los **nodos frontera** son el origen y destino de conexiones de tráfico que atraviesan la red troncal. Actúan como interfaz con las redes metropolitanas, y se encuentran por tanto situados cerca de grandes centros de tráfico (zonas urbanas de población más o menos elevada). Los **nodos de interconexión (nodos de tránsito, *cross-connect*)** realizan estrictamente la función de conmutación del tráfico entre sus enlaces con otros nodos. El número de enlaces que conectan un nodo (grado del nodo), es raramente superior a *5* en las topologías troncales actuales. El número de longitudes de onda, y la capacidad de cada canal de un enlace, son factores que se dimensionan en función de las necesidades de tráfico.

La multiplicación en el ancho de banda de transmisión en los enlaces que ha permitido la multiplexación WDM, ha creado un cuello de botella en la capacidad de conmutación de los nodos, que deben repartir esta enorme capacidad entre los usuarios de la red. Para atacar este cuello de botella, se ha intensificado la investigación en tecnologías de conmutación para la red troncal WDM. Los objetivos fundamentales que deben satisfacer las alternativas bajo estudio, vienen marcados por las necesidades de las empresas operadoras de este tipo de redes:

- **Eficiencia de reparto del ancho de banda.** Las demandas de conexiones a la red troncal pueden variar desde el orden de los *Mbps* hasta los *Gbps*. Se requiere, por tanto, un mecanismo que permita el control de la partición de la

capacidad de los enlaces (*granularización*) de una manera eficiente, dando servicio simultáneo a demandas de distintos anchos de banda.

- **Integración de distintos tipos de tráfico.** La red troncal debe transportar distintos tipos de tráfico (p.e. tráfico telefónico y tráfico IP) sobre los enlaces WDM. De esta manera, y como muestra la figura 1-2, conmutadores ATM, conmutadores telefónicos o encaminadores IP son fuentes de tráfico, potenciales *clientes* de la red troncal. La eficiencia en el mecanismo de integración viene determinada por la complejidad de la pila de protocolos necesaria, en la transmisión del tráfico relativo a cada dispositivo.

Relacionándolo con las claves indicadas anteriormente, en esta sección se hará un breve resumen de las tecnologías aplicadas y bajo estudio para la operación de las redes troncales WDM, siguiendo una aproximación cronológica.

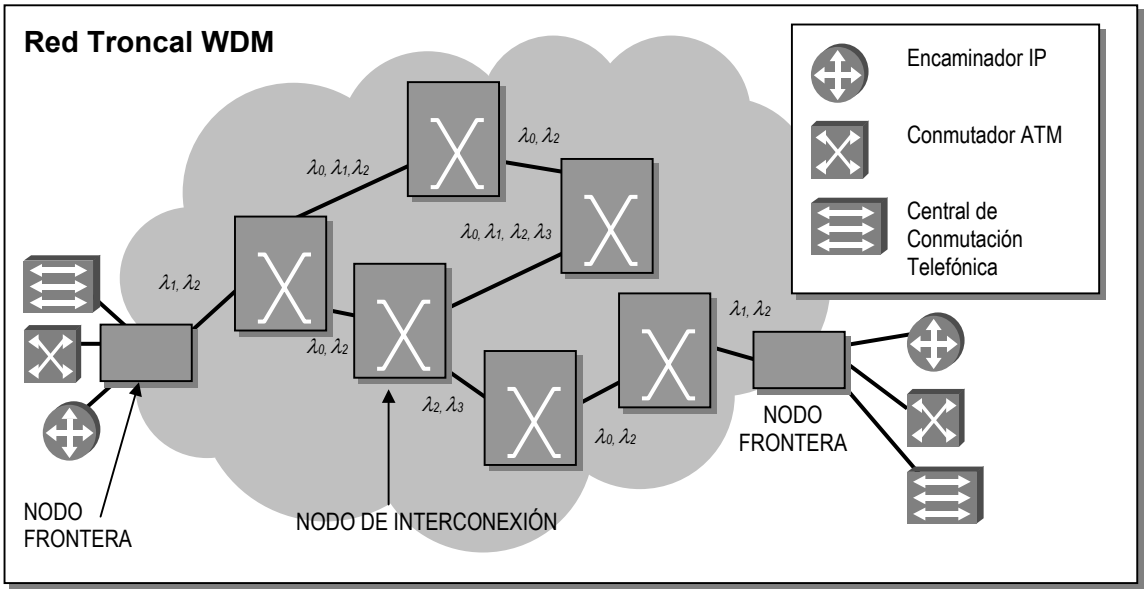


Figura 1-2 Red troncal WDM

1.2.1 SONET/SDH

Hasta la década de los 80, los sistemas de transmisión basados en fibra empleados en las redes troncales eran sistemas propietarios. Esto suponía que aspectos como el equipamiento, la codificación de línea, el mantenimiento, la administración o el provisionamiento de conexiones fueran específicos, y difícilmente interoperables. La situación de dependencia de los fabricantes de equipamiento, con los altos costes que esto generaba, provocó una fuerte presión de las empresas operadoras para la estandarización de una tecnología que pusiera fin a este escenario. En 1984 se establecieron los necesarios comités de trabajo dentro del ANSI (*American National Standards Institute*) produciendo como resultado el estándar llamado SONET (*Synchronous Optical Network*). Mientras los comités norteamericanos trabajaban en el desarrollo de SONET, en Europa, los comités de estandarización lo hacían en una tecnología para el reemplazo de los sistemas PDH (*Plesiochronous Digital Hierarchy*), por sus limitaciones debidas al funcionamiento no-síncrono. El resultado de estos comités fue una jerarquía de multiplexación semejante a SONET (y altamente interoperable con SONET), llamada SDH (*Synchronous Digital Hierarchy*),

estandarizada hoy en día por la ITU (*International Telecommunications Union*), y empleada en Europa.

La velocidad de línea de los enlaces SONET/SDH ha ido creciendo hasta los *40 Gbps* en estado comercial y *160 Gbps* en estado de prueba, según una determinada jerarquía [Sex97]. El reparto de este ancho de banda, se basa en la reserva de porciones de tamaño **fijo** adecuado en las tramas SONET/SDH (de duración *125* μs), de manera que tributarias de jerarquía inferior puedan agregarse en una transmisión de jerarquía superior. Los conmutadores SDH que implementan esta funcionalidad de interconexión y granularización de tráfico se conocen como *Digital Cross-Connect Systems* (DCS).

La introducción de la tecnología de multiplexación WDM en la transmisión/recepción, ha evolucionado de manera natural hacia la transmisión de una señal SONET/SDH independiente por cada longitud de onda, procesada por el dispositivo DCS del nodo como un canal independiente. Esta estrategia tiene tres desventajas fundamentales:

- La operación O-E-O de este tipo de redes se basa en la recepción de una señal óptica (O), que sufre una conversión opto-electrónica para su procesamiento electrónico (E) SONET/SDH, y conversión electro-óptica para el tráfico saliente (O). El elevado número de dispositivos de conversión opto-electrónica y electro-óptica provocan problemas de disipación y consumo de potencia.
- La necesidad de realizar procesamiento electrónico SONET/SDH en *todos* los nodos de la red, cuando la mayor parte del tráfico debe *atravesar* esos nodos, supone un mayor coste total en equipamiento. Asimismo, el coste unitario de los dispositivos DCS es superior, debido al aumento en el número y velocidad de los enlaces, que hace más costosa la función de granularización de la capacidad.
- Como red de transporte diseñada para una mejor adaptación al tráfico telefónico, su eficiencia en el transporte de tráfico de datagramas es mucho menor. Sin embargo, la tendencia en el crecimiento del tráfico de paquetes (en especial tráfico IP) ha generado un creciente interés por simplificar la pila de protocolos involucrada en la transmisión de datagramas IP, en la que la eliminación de SDH se baraja como una posibilidad. Una completa discusión en este punto puede encontrarse en [Bon01].

1.2.2 Conmutación O-O-O

Para intentar evitar los problemas asociados al cuello de botella electrónico, y gracias a la mejora constante de la tecnología de dispositivos ópticos, se ha planteado una serie de alternativas que se basan en el tratamiento óptico del tráfico que debe atravesar un nodo de conmutación de la red troncal (es decir, tráfico no originado ni destinado al mismo). Este tipo de alternativas se engloban dentro lo que se ha llamado conmutación O-O-O. Desde la década de los 90, el estudio de este tipo de conmutación en redes WDM ha seguido tres estrategias distintas [Siv00][Ram01][Mur02] Conmutación Óptica de Longitudes de Onda (*Wavelength Routing,* WR), Conmutación Óptica de Ráfagas (*Optical Burst Switching,* OBS) y Conmutación Óptica de Paquetes (*Optical Packet Switching,* OPS).

Las redes WDM que utilizan la técnica de **Conmutación Óptica de Longitudes de Onda** (*Wavelength Routing, WR*) [Muk97][Ram98][Ste99] son la opción más

avanzada que se puede encontrar en fase comercial en el actual estado del arte de la tecnología fotónica. La arquitectura de este tipo de redes está basada en una topología de interconexión arbitraria de nodos WXC (*Wavelength Crossconnect Nodes*), también llamados encaminadores WDM. Una conexión de tráfico entre dos nodos frontera implica la reserva en modo conmutación de circuitos de un canal (longitud de onda) a lo largo de una serie de fibras atravesando la red troncal, sufriendo o no conversiones de longitud de onda en su camino. El circuito establecido actúa como un medio de transmisión transparente, ahorrando costes respecto a las redes troncales convencionales, ya que no es necesario realizar un procesado electrónico de la señal en cada salto, sino únicamente en los nodos inicio y destino de la conexión de tráfico (que en general siguen siendo SDH). Sin embargo, la ineficiencia inherente de los mecanismos de conmutación de circuitos enfrentados a patrones de tráfico como el de Internet, hace que sea necesario sobredimensionar los nodos de conmutación y el número de canales por enlace para tener una probabilidad de bloqueo aceptable. Asimismo, el reparto entre los usuarios del ancho de banda de un camino óptico exige aplicar técnicas de aglomeración de tráfico (*traffic grooming*) en los nodos frontera, con el gasto de gestión y procesamiento electrónico que conllevan. Este problema de la granularización de tráfico empeora a medida que la evolución tecnológica eleva las velocidades de transmisión en cada canal (hasta *40 Gbps*, como se ha indicado anteriormente). Por todos estos motivos, la utilización del ancho de banda en este tipo de redes para las fluctuaciones de tráfico en redes como Internet es muy pobre.

La **Conmutación Óptica de Ráfagas**, *Optical Burst Switching* (OBS) [Tur97][Qia99][Qia00] permite un control más fino del ancho de banda de transmisión. Esta técnica de conmutación establece que cuando una fuente desea transmitir una cierta cantidad de datos a un mismo nodo destino, debe encapsularlo en una ráfaga de tamaño variable (en el orden de milisegundos de duración), que es conmutada como un todo a lo largo de la red. La configuración adecuada en los nodos es controlada por una cabecera, transmitida por un canal separado, y que precede la llegada de la ráfaga de datos. La Conmutación Óptica de Ráfagas mejora la utilización del canal respecto a las redes de Conmutación de Longitudes de Onda, a costa de una mayor complejidad *hardware* de los dispositivos. Para una visión global actualizada del estado de esta tecnología, puede consultarse [Bat03].

Sin embargo, el proceso de reparto de capacidad que permite la utilización más alta del canal, debido a la multiplexación estadística del tráfico, y el mecanismo más directo para adaptarse a patrones de tráfico de las redes de datos multimedia actuales se alcanza con la tercera alternativa: **Conmutación Óptica de Paquetes** (*Optical Packet Switching, OPS)* [Hun00][Xu01] [OMa01][Jou01][Elb02][Yao02]. En el dominio óptico, OPS es similar a las técnicas tradicionales de conmutación de paquetes bajo tecnología electrónica, excepto que la carga de datos (*payload*) del paquete permanece en estado óptico en los nodos centrales de la red (lo que supone un almacenamiento óptico en caso de ser necesario), mientras su cabecera es procesada electrónicamente. Este paradigma proporciona la mayor eficiencia en el uso del canal, al operar con la granularidad de un paquete. Asimismo, también proporciona ventajas en cuanto a los mecanismos de protección y restauración de la red, comparado con los sistemas de protección en las redes *Wavelength Routing*, como ha sido argumentado en [Dit03]. Sin embargo, la necesidad de realizar la función de conmutación paquete por paquete a las velocidades existentes en la red troncal impone los requisitos tecnológicos más estrictos. Por ello, a pesar de ser una opción tecnológica investigada desde la década de los 90, en la que se han construido prototipos con éxito, la Conmutación Óptica de Paquetes se considera como la situación final a medio/largo plazo en la evolución de la red troncal, pasando o no por la solución intermedia de Conmutación Óptica a Ráfagas [IETF03-2]. En la siguiente

sección, enfocamos el estudio en las redes troncales OPS, donde se enmarca esta tesis doctoral.

1.3 Redes Troncales de Conmutación Óptica de Paquetes

1.3.1 Arquitectura de Red

Conceptualmente, la arquitectura de una red troncal OPS está basada en un conjunto de nodos centrales de interconexión conocidos como *Optical Cross-Connect Nodes (OXC)*, y un conjunto de nodos frontera con las redes de acceso de conmutación electrónica, interconectados por enlaces WDM, de manera semejante a como se muestra en la Figura 1-2.

Los **nodos frontera** de la red OPS son el origen y destino de conexiones de tráfico que atraviesan la red, siguiendo el modelo de Circuito Virtual, habitual en las redes troncales de Conmutación de Paquetes. Son por tanto estos nodos los encargados de la conformación y extracción de tráfico en y desde paquetes ópticos. Debido a su función de interfaz con las redes de conmutación electrónicas, las características deseables de este tipo de dispositivos son un funcionamiento en el orden de los *Tbps*, con una gran capacidad de almacenamiento de paquetes (almacenamiento electrónico), y un funcionamiento multiprotocolo que permita la interacción con redes electrónicas de paquetes de distinto tipo [Yao02][Oma01]. Los **nodos de interconexión**, son los encargados de la conmutación de los paquetes ópticos, desde y hacia los nodos frontera.

Las redes de Conmutación Óptica de Paquetes pueden ser clasificadas en función de dos parámetros [Dit03][Bre03]:

- **Tamaño de paquete (tiempo de transmisión de paquete).** Discriminando entre redes que operan con tamaño fijo y con tamaño variable de paquete.
- **Sincronización de la red de transporte.** Distinguiendo entre redes síncronas o ranuradas (*slotted*), y asíncronas o no ranuradas (*unslotted*). Según el modo ranurado, los paquetes ópticos recibidos por los nodos de conmutación deben ser alineados en los puertos de entrada. En el modo no ranurado, los paquetes pueden llegar y ser procesados por el nodo de conmutación en cualquier instante temporal.

A pesar de que el funcionamiento no ranurado simplifica la necesidad de una etapa de sincronización en los nodos de conmutación, tiene como desventaja mayores necesidades de almacenamiento en los mismos para compensar una mayor probabilidad de contención. Por ello, los mayores esfuerzos de investigación en la Conmutación Óptica de Paquetes se han enfocado hasta el momento en las redes de tamaño fijo de paquete, con nodos de conmutación trabajando en modo ranurado, y tamaño de la ranura temporal igual al tamaño de paquete. Como ejemplo, la selección de esta alternativa ha sido una conclusión dentro del proyecto DAVID (*Data And Voice Integration over DWDM*) [Dit03] financiado por la Unión Europea. Aspectos como la transmisión de paquetes de tamaño variable (p.e. datagramas IP) sobre este tipo de paquetes ópticos, o la longitud de los mismos [Cal97] no han sido definidos todavía. Las mejores prestaciones se obtendrían con una segmentación en paquetes de pequeño tamaño en los nodos frontera. Sin embargo, este aspecto está limitado por la pérdida de eficiencia debido a la relación tamaño de cabecera y tiempo de guarda respecto a tamaño de paquete, y por el tiempo de procesamiento de paquete. Un tamaño de paquete realista [Dit03] de *1* μ*s* transporta *5000* bytes en a un canal a *40*

Gbps, lo cual es un orden de magnitud mayor que el tamaño de los datagramas IP generados por la mayoría de las aplicaciones. Por ello, parece necesario algún tipo de agregación en el conformado de paquetes en los nodos frontera, de manera semejante a lo planteado para la Conmutación Óptica de Ráfagas.

La Conmutación Óptica de Paquetes síncrona (ranurada), para paquetes de tamaño fijo igual a la longitud temporal de la ranura, será el marco de trabajo en esta tesis doctoral. Algunas arquitecturas de conmutación evaluadas para tamaño variable de paquete, pueden consultarse en [Cal00-1][Cal00-2][Cal02][Tan00][Bre03] y dentro de la literatura referente a la Conmutación Óptica a Ráfagas. Una aproximación a arquitecturas de conmutación diseñadas para paquetes de tamaño fijo, en modo no-ranurado puede encontrarse en [Han98].

1.3.2 Arquitectura de conmutador

En esta sección se describirán los bloques funcionales de un nodo de Conmutación Óptica de Paquetes ranurado, mostrado en la Figura 1-3.

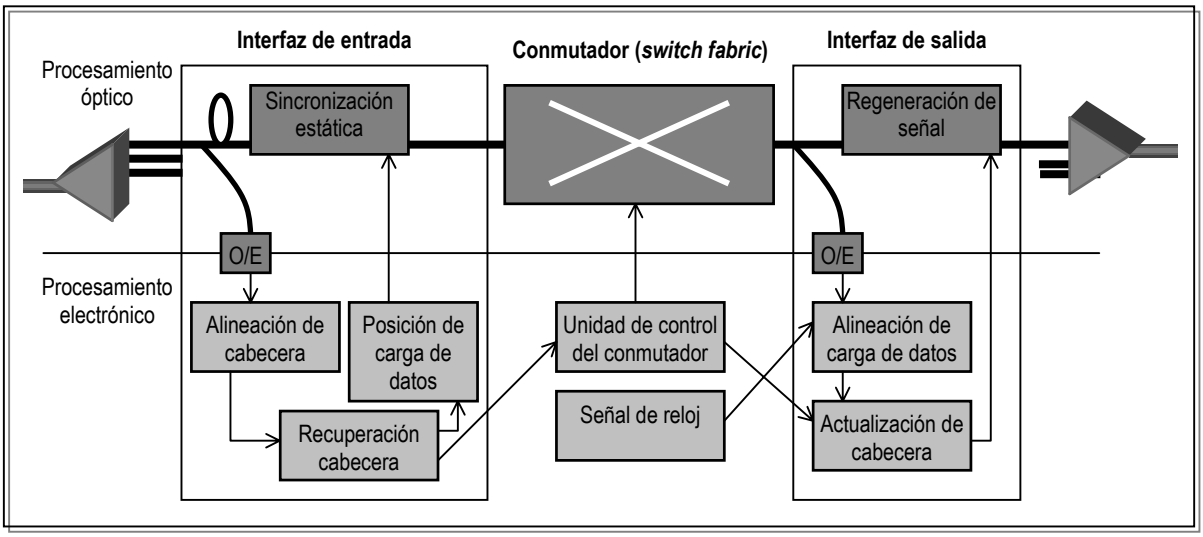


Figura 1-3 Arquitectura de conmutador OPS síncrono

1.3.2.1 Interfaz de entrada

Como corresponde a un entorno de actuación WDM, las fibras de entrada y salida al nodo de conmutación son multiplexadas y demultiplexadas mediante los adecuados dispositivos pasivos. Como resultado, el número de puertos de entrada (salida) del conmutador es igual al número de fibras de entrada (salida) por el número de longitudes de onda en cada fibra.

De manera similar a lo que sucede en los nodos de conmutación puramente electrónicos, existe una interfaz para cada uno de los puertos de entrada, que en el caso de los nodos de conmutación OPS ranurados realiza las funciones de (1) sincronización estática de la carga de datos, y (2) detección de cabecera, que se explican a continuación.

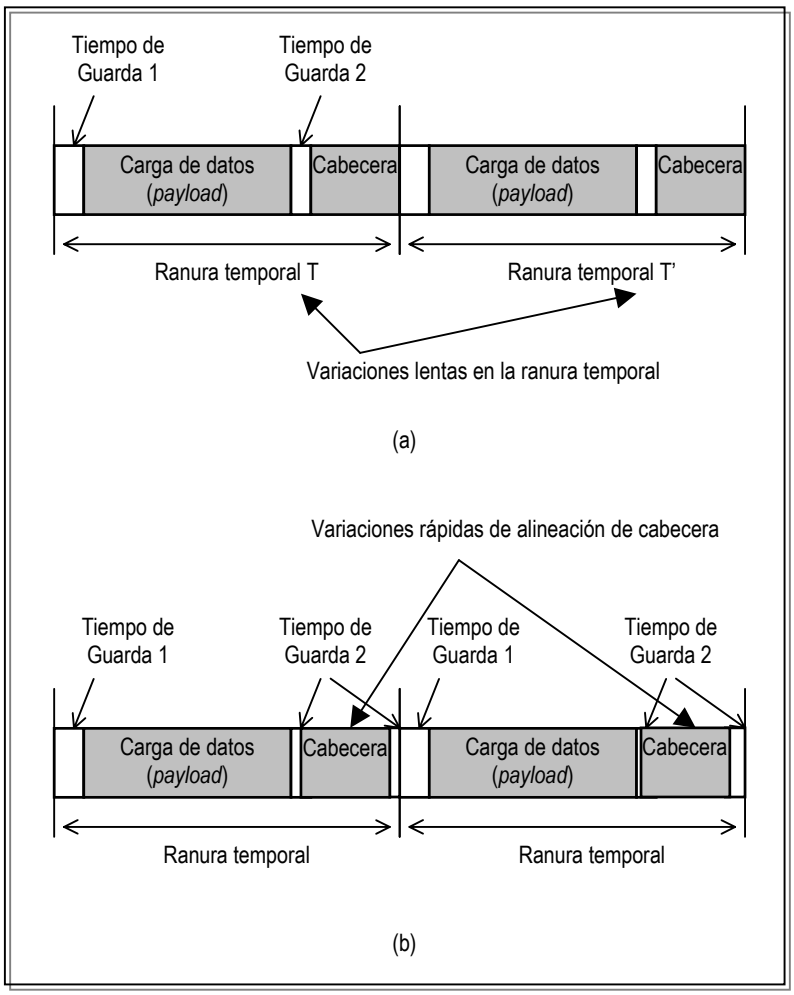


Figura 1-4 Sincronización en redes ranuradas, (a) sincronización estática, (b) alineación de cabecera

1.3.2.1.1 Sincronización estática

La problemática asociada a la sincronización de la carga de datos se muestra en la Figura 1-4-(a). El funcionamiento ranurado requiere una alineación gruesa de la carga de datos respecto al reloj de sincronización del nodo. Sin embargo, la dispersión cromática dependiente de la temperatura en los enlaces entre nodos, contribuye a una desincronización de los paquetes provenientes de cada uno de los puertos de entrada, sobre la que no se tiene control. La alineación que se requiere es de variación lenta (y por ello también llamada "estática" [Yao00]), debido a que no son necesarias correcciones paquete a paquete, sino una alineación durante la inicialización del nodo, y una monitorización a gran escala temporal posterior.

En [Yao00] y [Gui98] se describen dos estrategias de sincronización. La primera, mostrada en la Figura 1-5-(a), consiste en un conjunto de conmutadores ópticos *2x2* que dirigen la señal luminosa de entrada a través de una serie de retardos ópticos, basados en fibras de longitud apropiada. La elección del camino a través estos retardos se decide con el objetivo de alineamiento a la salida de la etapa de sincronización. El segundo mecanismo descrito en [Yao00] consiste en la utilización de una sección de fibra altamente dispersiva, y un convertidor de longitud de onda sintonizable (Figura 1-5-(b)). La elección de la longitud de onda se realiza de manera que su retardo produzca un alineamiento del paquete a la salida.

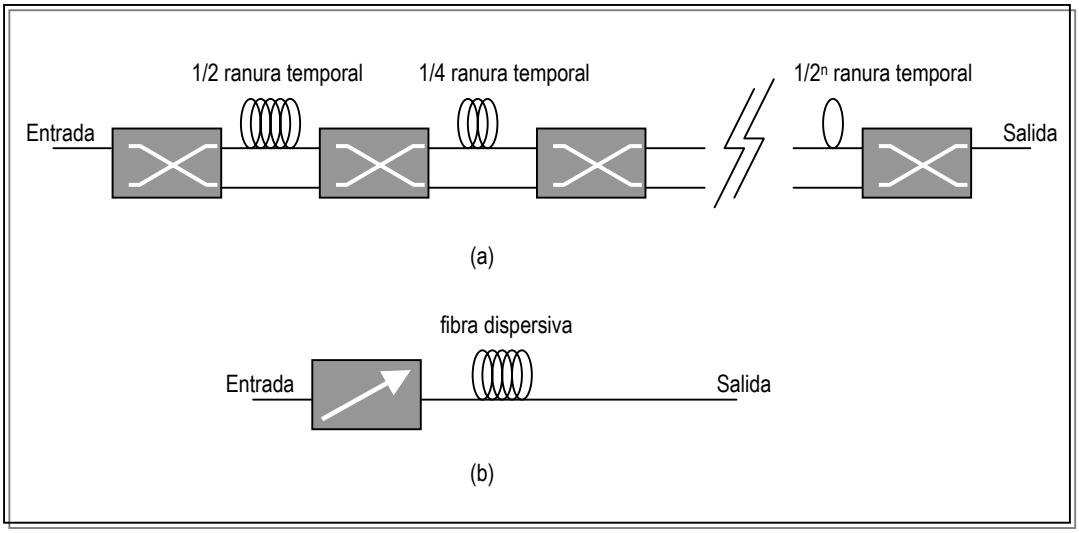


Figura 1-5 Sincronización estática (a) mediante retardos en cascada, (b) mediante fibras dispersivas

1.3.2.1.2 Detección de cabecera

La detección de cabecera es el proceso por el cual la información binaria de cabecera es extraída de la señal óptica, y entregada a la unidad de control para su procesamiento electrónico. Una alineación precisa de la señal óptica de cabecera (figura 1-4-(b)) es un requisito primordial para que este proceso pueda llevarse a cabo en los nodos de conmutación. En los nodos frontera de la red, donde la carga de datos también debe ser extraída, se aplica, asimismo, la necesidad de alineación [Bos97].

Durante el proceso de alineamiento, los bits a ser leídos deben ser sincronizados con la señal interna de reloj. Algunos formatos de paquete, requieren una sincronización a nivel de bit, que debe realizarse paquete a paquete. Esto supone que el proceso de sincronización debe conseguirse en pocos bits. Las soluciones basadas en circuitos PLL (*Phase Locked Loop*) no son utilizables en este escenario, por el elevado número de bits que se requieren para la sincronización. Se han desarrollado y probado sistemas de sincronización que cumplen con los requisitos planteados, cuyo estudio está fuera del ámbito de esta tesis doctoral [Zuc96][Ker97][Gui97].

1.3.2.2 Unidad de control

La unidad de control se encarga de la generación de las señales que gobiernan el funcionamiento de los componentes del nodo de conmutación. En la actualidad, este procesamiento es realizado por circuitos electrónicos, y no se espera que esto cambie en el medio/largo plazo. Las decisiones involucradas abarcan desde las referentes a la planificación de la arquitectura de conmutación, hasta las operaciones de gestión y mantenimiento del nodo.

1.3.2.3 Reloj del dispositivo

Debido al funcionamiento ranurado de los nodos de conmutación, la utilización de una red de sincronización plesiócrona, podría provocar errores de sincronización y por tanto la pérdida de paquetes. Por ello, se presupone la utilización de una red de sincronización propia, o derivada de otro equipamiento (p.e. SDH) que gobierne la ranuración temporal de una manera síncrona entre nodos.

1.3.2.4 Interfaz de salida

Existe una interfaz para cada puerto de salida, que acondiciona la señal de salida a los requisitos de transmisión, incluyendo niveles de potencia, o compensación de (pequeños) desalineamientos de paquetes a la salida del conmutador, debidos a diferencias de camino óptico que los paquetes pueden sufrir. Las necesarias actualizaciones de cabecera de paquete son realizadas también en esta interfaz de salida. Algunas de estas funcionalidades pueden ser innecesarias, en función de la arquitectura de conmutación empleada [Yao00].

1.3.2.5 Arquitectura de conmutación

La arquitectura de conmutación (*switch fabric*) es el componente del nodo que se encarga de la función de transferencia de los paquetes ópticos desde sus puertos de entrada hacia sus correspondientes puertos de salida. Para resolver la posible contención entre paquetes destinados hacia el mismo puerto de salida, en un mismo ciclo de conmutación, un esquema de almacenamiento (óptico) debe ser necesariamente implementado.

La arquitectura de conmutación es un componente crítico dentro del nodo de conmutación, por su impacto en el coste y las prestaciones. La evaluación de distintas alternativas de diseño en este campo es el objetivo principal de esta tesis doctoral.

1.4 Componentes fotónicos de las arquitecturas de conmutación OPS

En esta sección se describirá un conjunto de dispositivos fotónicos involucrados en la implementación de arquitecturas de conmutación OPS. La visión que se proporcionará está enfocada hacia los aspectos de interés para un proceso de evaluación comparativa de arquitecturas, objeto de esta tesis doctoral. Por ello, después de una síntesis del funcionamiento de cada componente, no nos centraremos en los detalles físicos del mismo, sino en parámetros como su escalabilidad (en número de puertos o en número de longitudes de onda), o aspectos que nos puedan orientar sobre el coste relativo de los dispositivos. Es esta información la que se requiere en un proceso de evaluación, para proporcionar respuestas que balanceen prestaciones y coste de las distintas opciones.

Sin embargo, debemos tener en cuenta que: (1) el estado del arte de componentes fotónicos, es un campo de gran dinamismo, donde la mejora de prestaciones y la bajada de costes es constante, (2) teniendo como objetivo el medio plazo (para la aplicación comercial de la Conmutación Óptica de Paquetes), el coste relativo de los componentes puede variar enormemente, variando el interés en unas arquitecturas frente a otras.

Por ello, los datos expuestos en esta sección, que serán empleados en distintas justificaciones y conclusiones dentro de esta tesis doctoral, tienen que ser valorados en conjunción con el escenario tecnológico. En el caso concreto de las comparativas incluidas, se harán desglosando el número de dispositivos de cada tipo que requiere cada arquitectura de conmutación. Se pretende aportar una información que pueda ser utilizable tras posibles cambios en el coste relativo de los componentes.

1.4.1 Función de conmutación

Los dispositivos de conmutación utilizados en las redes *Wavelength Routing*, basados en las propiedades electroópticas de los cristales líquidos [Nog98], en generación de burbujas de aire en guías planares [Ven01], o en orientación de micro espejos MEMS (*Micro-Electro-Mechanical-Systems*) [Neu01] proporcionan tiempos de conmutación del orden de milisegundos. Estos tiempos son suficientes para las

necesidades de este tipo de redes, donde el tiempo crítico de la función de conmutación viene determinado por el tiempo máximo de recuperación ante un fallo en un nodo o en un enlace. Sin embargo, en las redes OPS, la función de conmutación debe realizarse paquete a paquete, en un tiempo limitado por la suma del tiempo de guarda más el tiempo de cabecera, en el orden de nanosegundos [Bre03]. Esta característica impone los mayores requisitos, dejando un estrecho abanico de dispositivos ópticos viables, que forman la base de los conmutadores OPS:

- **Puertas ópticas.** Las puertas ópticas basadas en Amplificadores Ópticos Semiconductores (*Semiconductor Optical Amplifier, SOA*) [Kal92][Chi01-2] [Stu00][Ren98], permiten el bloqueo (estado OFF) o transmisión (estado ON) de una señal óptica, mediante la variación de la corriente de inyección al amplificador SOA. Este funcionamiento permite una velocidad de conmutación en el orden del nanosegundo con un ratio ON/OFF de unos *50* dB [Gui98]. El ancho de banda de las puertas es amplio (en 1995 fue descrito un sistema de *40* nm de ancho de banda [Mur02]), con lo que son capaces de bloquear/dejar pasar señales combinando un número elevado de canales WDM. Su mayor limitación se encuentra en la adición de ruido debido al efecto de emisión espontánea ASE (*Amplified Spontaneous Emission*), lo que debe tenerse en cuenta en el caso de que una señal deba atravesar una cascada de estos dispositivos. En cuanto a sus costes de fabricación, son todavía elevados, debido al insuficiente nivel de integración conseguido: en [Sha00] se puede encontrar un ejemplo de una matriz de *32* puertas ópticas fabricada por Alcatel, sobre una placa de unos *40x30* cm (estimado a partir de ilustraciones, al no haber sido posible obtener las medidas exactas). Asimismo, el consumo de potencia es alto en este tipo de componentes, y se hacen necesarios mecanismos de control de temperatura de los dispositivos.
- **Conversores de Longitud de Onda Sintonizables** (*Tunable Wavelength Converter, TWC*). Estos dispositivos permiten convertir la longitud de onda de un paquete de información, en otra seleccionable dentro de un rango. Existen distintas tecnologías probadas de dispositivos TWC para redes OPS, descritos en la literatura (ver [Dur96] y [Whi02] como referencia inicial). Su funcionamiento se basa en la combinación de un láser sintonizable [Bru02] con las no linealidades de los dispositivos SOA. Cumplido el requisito de velocidad de sintonización (en el orden del nanosegundo), el parámetro de interés a tener en cuenta para la evaluación de arquitecturas es el rango de longitudes de onda sintonizables, habiéndose probado con éxito dispositivos con rangos del orden de *30* nm [Gui98], y siendo éste un valor en constante mejora.
- **Conversores de Longitud de Onda Fijos** (*Fixed Wavelength Converter, FWC*). Estos dispositivos convierten la longitud de onda de un paquete de información, en otra fija. Su funcionamiento se basa en la combinación de un láser a frecuencia fija [Bru02] con las no linealidades de los dispositivos SOA. Se trata de dispositivos, lógicamente, más económicos que los TWC.

1.4.2 Encaminamiento de la señal óptica

Los dispositivos AWG (*Arrayed-Waveguide-Gratings*) son componentes pasivos que permiten encaminar las señales ópticas en los puertos de entrada hacia los puertos de salida, en función de su longitud de onda [Tak90]. En un dispositivo AWG de *K* puertos de entrada y *K* puertos de salida (figura 1-6-(a)), las reglas que gobiernan la transmisión son las que se muestran en la figura 1-6-(b).

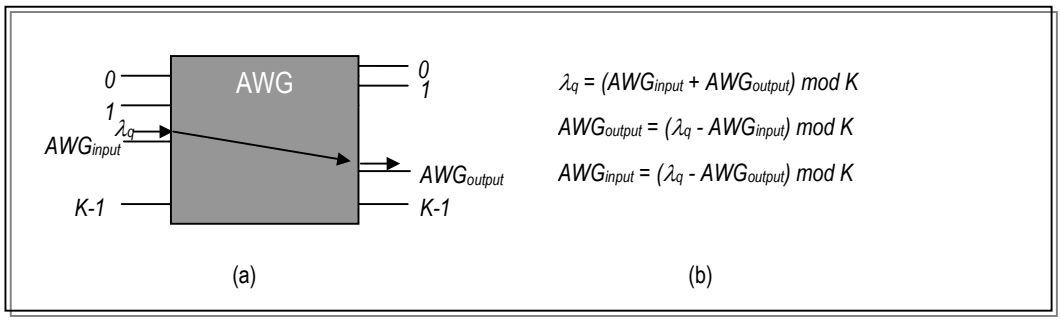


Figura 1-6 (a) Dispositivo AWG, (b) relaciones en el encaminamiento de señales ópticas

El diseño de los AWGs se basa en la transmisión de las señales de entrada sobre un alto número de guías, cuya longitud tiene una diferencia constante. En esta transmisión se produce un fenómeno de dispersión electromagnética, de tal manera que la señal en un puerto de entrada aparece con mayor potencia en un puerto de salida, dependiente de la longitud de onda de la señal. El alto grado en el que se produce la dispersión permite una separación de longitudes de onda del orden de *1* nm (y por tanto homologable a la separación de canales WDM).

El progreso en la fabricación de los dispositivos AWG mediante tecnología PLC (*Planar-Lightwave Circuit*) impulsó su utilización en combinación con dispositivos TWC, en determinadas arquitecturas de Conmutación Óptica de Paquetes, como veremos en esta tesis doctoral, donde la conversión de longitud de onda selecciona el puerto de salida del AWG. La capacidad de integración alcanzada, y la posibilidad de interconexión de dispositivos AWG [Bar93] ha permitido construir módulos de elevado número de puertos a un coste comparativamente bajo (en [NTT00] se describe un dispositivo de *256x256* puertos, con una separación entre canales de *25* GHz, sobre un chip de *55x75* mm). Como desventaja, destacar que se trata de dispositivos que requieren un control de temperatura estricto en su funcionamiento, aunque perfectamente alcanzable en un nodo de interconexión de red troncal.

1.4.3 Almacenamiento

Un aspecto crucial para el desarrollo de las redes OPS es el problema del almacenamiento (*buffering*) de los paquetes ópticos. La funcionalidad de lectura-escritura habitual en las memorias electrónicas, no es implementable para señales a frecuencias ópticas. Por ello, las memorias en las arquitecturas de conmutadores ópticos de paquetes están basadas casi únicamente en líneas de retardo, que retrasan la señal un tiempo prefijado por la longitud física de la fibra y la velocidad de propagación de la señal (ver [Hun98-2] como referencia inicial). En el caso más sencillo, una cola de *B* posiciones de memoria puede ser emulada mediante *B* líneas de retardo de longitudes ópticas de *0* hasta *B-1*.

Las memorias basadas en lazos recirculantes, donde la señal óptica circula por un bucle cerrado hasta el momento de la lectura, no son todavía viables, aunque son consideradas como un punto de máximo interés de investigación. Esta alternativa está directamente relacionada con la necesidad de regeneración de la señal óptica en cada recirculación. Los distintos niveles de regeneración de señal se clasifican en: 1R (sólo amplificación de señal), 2R (amplificación y reconformado del pulso) y 3R (amplificación, reconformado de pulso y resincronización temporal) [Whi02]. Sin embargo, la construcción de regeneradores 3R de alta velocidad en el dominio óptico, todo integrado en un mismo dispositivo, se encuentra en un estado inicial de investigación en el laboratorio.

Al contrario de lo que sucede con las memorias electrónicas, las líneas de retardo son inicialmente, componentes de bajo coste. Debe sin embargo hacerse una puntualización en cuanto a la magnitud de la longitud de las fibras manejadas. La longitud de una línea de retardo viene dada por el número de ranuras temporales que debe retrasar la señal, y la duración de dicha ranura. Con un retardo de propagación en fibra de *0,5* μs cada *100* m, esto supone *200* m de fibra para un retardo de una ranura temporal de *1* μs. De esta manera, un dispositivo con líneas de retardo desde *1* a *50* ranuras temporales requeriría un total de más de *250* km de fibra. Esto hace muy interesante aquellas alternativas que permiten una reducción del número de retardos necesario, como las que serán estudiadas en esta tesis doctoral.

1.4.4 Simbología en figuras

A continuación, se muestran los símbolos de cada uno de los componentes, empleados en los diagramas de las arquitecturas de conmutación en este documento.

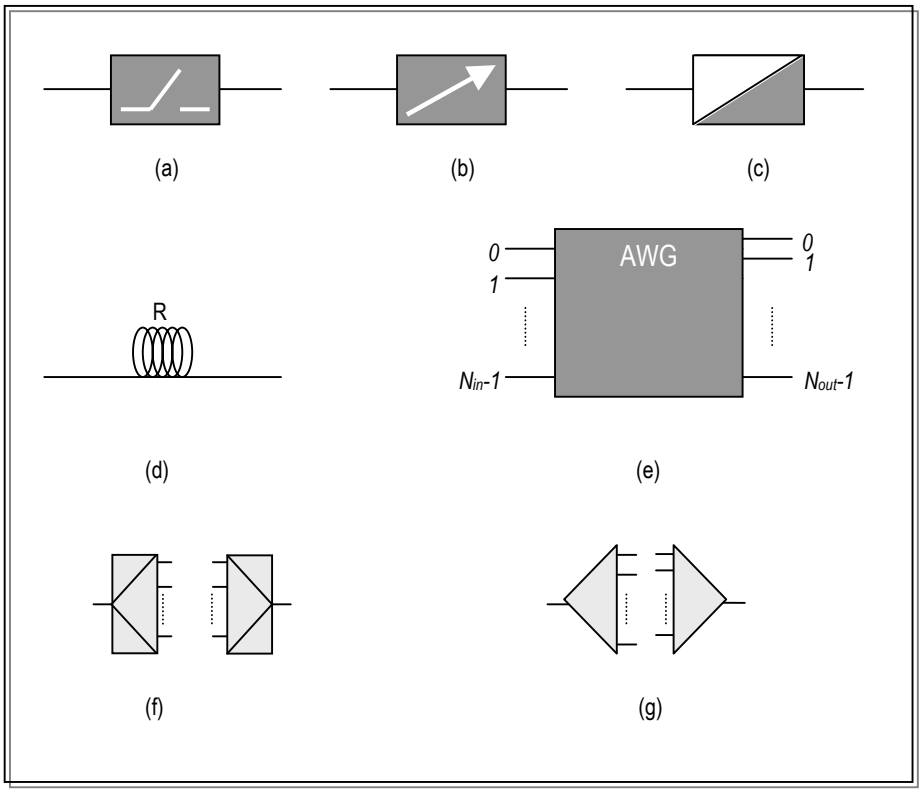


Figura 1-7 (a) Puerta óptica, (b) Conversor de Longitud de Onda Sintonizable, (c) Conversor de Longitud de Onda Fijo, (d) Línea de retardo de *R* ranuras temporales, (e) Dispositivo AWG de N_{in} puertos de entrada y N_{out} puertos de salida, (f) acoplador/combinador óptico, (g) demultiplexor/multiplexor óptico

1.5 Modos de operación de las redes OPS

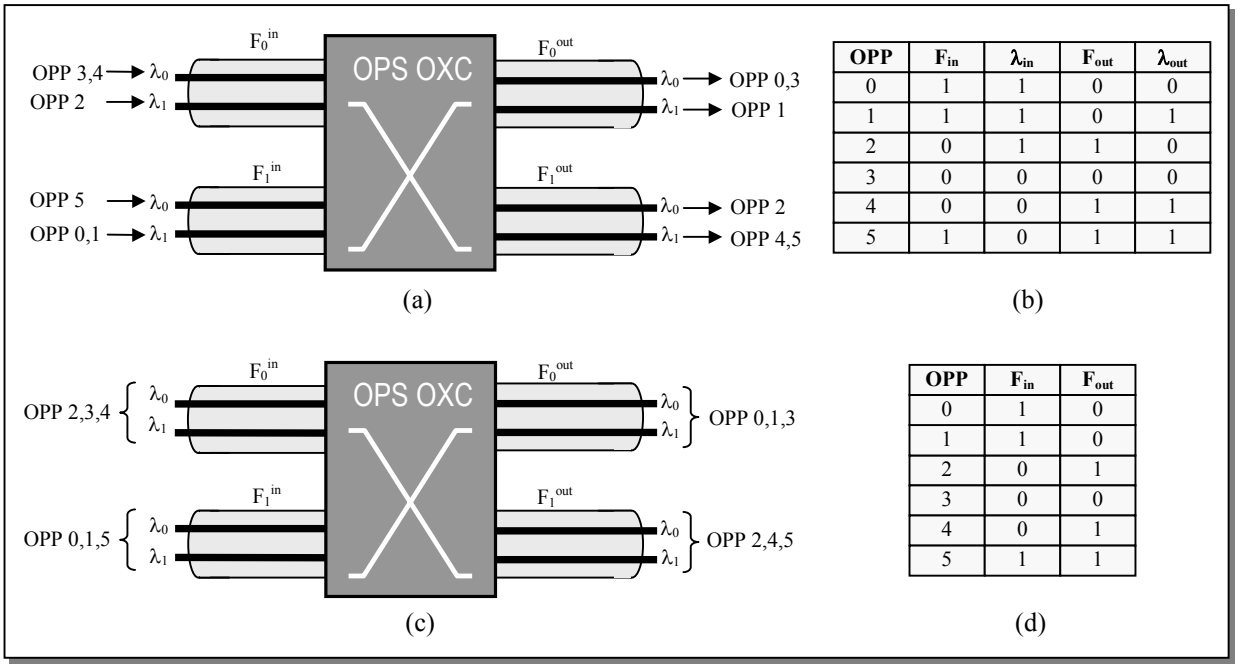
Como se ha comentado en una sección anterior, está generalmente aceptado que las conexiones de tráfico entre extremos de una red troncal OPS, se establecerán mediante un tipo de circuito virtual permanente, que llamaremos *Optical Packet Path, (OPP)*. Siguiendo el concepto de circuito virtual, los paquetes pertenecientes al mismo

OPP siguen una secuencia de saltos desde el nodo frontera de entrada (*ingress*) hasta el nodo frontera de salida (*egress*). Esta secuencia es prefijada durante el establecimiento (provisionamiento) del circuito, siguiendo parámetros de ingeniería de tráfico, de tal manera que el número de conexiones de tráfico encaminadas a través de un mismo enlace entre dos nodos, puede ser potencialmente mucho mayor al número de longitudes de onda de ese enlace. Es éste el escenario de aplicación de OPS, donde sus ventajas en cuanto a la eficiencia de reparto granularizado del ancho de banda la hacen rentable frente a otras alternativas como la Conmutación Óptica de Longitudes de Onda.

En esta situación, el **modo de operación de la red** define el mecanismo por el cual el tráfico de las conexiones que atraviesan un enlace, es repartido entre las longitudes de onda del mismo.

Las propuestas en este campo se han dado dentro del proyecto WASPNET [Niz98][Hun99][Chi99][Chi01-1], que plantea dos alternativas: redes SHWP (*Shared Wavelength Path*) y redes SCWP (*Scattered Wavelength Path*). En las redes SHWP, los paquetes pertenecientes al mismo OPP siguen una secuencia fija de saltos donde la fibra y la longitud de onda de transmisión se encuentran fijadas en cada uno de los saltos. Estos valores se almacenan durante el provisionamiento (*provisioning*) de la conexión en una tabla de consulta en cada nodo atravesado (figura 1-8-(a,b)). Por otro lado, en las redes SCWP, las conexiones OPP determinan la fibra de transmisión en cada salto, pero no la longitud de onda de salida (figura 1-8-(c,d)). Por lo tanto, el sistema planificador de cada nodo puede seleccionar la longitud de onda de transmisión dinámicamente de cada paquete, según determinados criterios de eficiencia, como la ocupación de memorias. SHWP y SCWP son los competidores previsibles para el futuro despliegue de una red OPS sobre tecnología WDM. La adopción de una u otra alternativa involucra distintos factores [Niz98]:

- SCWP ofrece unas prestaciones de utilización de enlace mucho mejores, debido a la multiplexación estadística que se puede obtener mediante un algoritmo adecuado de selección de canal de salida. Esto permitiría disminuir los requisitos de almacenamiento, y reducir los retardos.
- Por otro lado, las redes SCWP conllevan una mayor complejidad en el control del nodo, por la necesaria decisión de selección de longitud de onda de salida paquete a paquete.
- Según ha sido descrito en [Niz98], el modo de operación provoca una mayor complejidad en los algoritmos de recuperación ante fallo de la red. En opinión del autor de esta tesis doctoral, este es un aspecto que requiere un estudio más profundo.
- Las redes SHWP aseguran una entrega ordenada extremo-a-extremo de paquetes pertenecientes al mismo OPP, mientras los nodos intermedios no añadan desorden interno. Esta condición tiene una matización dentro de las redes SCWP, ya que dos paquetes del mismo OPP con una determinada precedencia temporal, pueden ser transmitidos *simultáneamente* en longitudes de onda distintas de la misma fibra. Esto plantea la necesidad de que cada nodo sea capaz de conocer cuál es el orden entre paquetes recibidos simultáneamente, con el fin de poder mantener ese orden en el siguiente salto. A lo largo de este documento se prestará también especial atención al problema del desorden de paquetes en las distintas arquitecturas bajo estudio.



OPP	F_{in}	λ_{in}	F_{out}	λ_{out}
0	1	1	0	0
1	1	1	0	1
2	0	1	1	0
3	0	0	0	0
4	0	0	1	1
5	1	0	1	1

OPP	F_{in}	F_{out}
0	1	0
1	1	0
2	0	1
3	0	0
4	0	1
5	1	1

Figura 1-8. Ejemplo de aplicación de los modos de operación SHWP/SCWP, (a) Provisión de OPPs SHWP, (b) Tabla de consulta SHWP del nodo, (c) Provisión de OPPs SCWP, (b) Tabla de consulta SCWP del nodo

1.6 Motivación, objetivo y desarrollo de esta tesis

La visión de la Conmutación Óptica de Paquetes como una alternativa elegante y prometedora, pero no viable hasta el medio plazo, ha sido un pensamiento acentuado por el frenazo tecnológico del que ahora comenzamos a recuperarnos. Esto ha tenido su repercusión en los distintos foros estandarizadores, que han centrado sus esfuerzos en la tecnología de la red troncal *Wavelegth Routing*, y más recientemente *Optical Burst Switching*. Como ejemplo, los documentos del grupo de trabajo IPO (*IP over Optical*) del IETF, no han realizado (hasta el conocimiento del autor) ninguna mención a la transmisión de datagramas IP en redes OPS. Tampoco existe ningún grupo de trabajo IRTF (*Internet Research Task Force*) que trate OPS como estrategia de evolución de la red troncal WDM. Las menciones encontradas recientemente en documentos del IETF ahondan en esta línea:

- Dentro del grupo de trabajo GSMP (*General Switch Management Protocol*), en Junio de 2003 se ha publicado un documento con status de *Internet Draft* y fecha de expiración Diciembre de 2003, en el que se valora la aplicación del protocolo GSMPv3 para la gestión y control de nodos de conmutación óptica [IETF03-1]. La arquitectura de los nodos de conmutación se considera de manera abstracta como un dispositivo capaz de conmutar información en función de etiquetas ópticas, independientemente de la red de transporte subyacente (todo óptica o no). Además de la descripción de los mensajes de control y gestión del nodo aplicados a redes *Wavelength Routing*, y redes *Optical Burst Switching*, se menciona (sin describirla) la posible aplicación en redes de Conmutación Óptica de Paquetes.

- Dentro del grupo de trabajo CCAMP (*Common Control And Measure Plane*), en Junio de 2003 se ha publicado un documento con status de *Internet Draft*, y fecha de expiración Diciembre de 2003, en el que se valoran ciertos requisitos para la aplicación de políticas de control y señalización en redes de Conmutación Óptica de Ráfagas [IETF03-2]. Dentro de este documento, se menciona la Conmutación Óptica de Paquetes como la solución definitiva para la transmisión y conmutación en redes ópticas, pero se argumenta su inviabilidad en el corto plazo, y se plantea OBS como una opción en el plazo intermedio.

En definitiva, después de los casi *15* años de investigación heterogénea en el campo de la Conmutación Óptica de Paquetes, nos encontramos con un marco prematuro, difuso y poco uniformizado. En este sentido, uno de los puntos abiertos más relevantes, es la manera en la cual la Conmutación Óptica de Paquetes puede ser aplicada sobre redes WDM. Hasta las propuestas realizadas dentro del proyecto WASPNET, sobre el modo de operación SHWP/SCWP de la red, este tema no había sido tenido en cuenta en la mayor parte de las propuestas de arquitecturas de conmutación. Sin embargo, el modo en el que una arquitectura opera, es un factor de gran impacto sobre las prestaciones de los conmutadores. Esto nos empujó a tomar este *modo de operación* como base de una clasificación de arquitecturas: debíamos determinar primero el modo de operación (SHWP/SCWP) en el que una arquitectura de conmutación debía trabajar, para posteriormente realizar la necesaria evaluación de prestaciones. No es de interés comparar arquitecturas que trabajan en distintos modos de operación. Sí es de interés evaluar qué arquitectura se adapta mejor a cada modo de operación (relación coste/prestaciones). Así como, sí es de interés conocer qué modo de operación proporciona mejores prestaciones en las distintas arquitecturas.

Se trata de una clasificación sencilla, que ha estado en la base de todos los trabajos publicados como parte de esta tesis doctoral. El argumento a favor de su utilización es subjetiva: la opinión del autor de que el concepto de *modo de operación* estará sin duda presente en el momento en que la madurez tecnológica centre el interés de los comités de estandarización en la Conmutación Óptica de Paquetes. La elección de uno u otro modo, valorará aspectos diversos, como el impacto en los mecanismos de protección y recuperación, de provisionamiento, de gestión de red, y también en el coste de las arquitecturas de conmutación. Es en este último aspecto, de gran influencia en el coste final de los nodos de conmutación, en el que se centra el trabajo de investigación emprendido.

Sin embargo, la aplicación de esta clasificación presenta el inconveniente de que para gran parte de las arquitecturas de conmutación, propuestas fuera del proyecto WASPNET, (1) no han sido descritos sus mecanismos de integración en una red WDM, (2) no han sido especificados sus algoritmos de planificación bajo los modos de operación SHWP y SCWP. Por ello se ha requerido, previamente, un proceso de adaptación o de normalización de los diseños. En [Pav03-2], fue presentado un conjunto de *3* puntos sencillos, pero necesarios como proceso de adaptación:

(1) Las arquitecturas no-WDM (es decir, aquellas cuyas puertos de entrada y salida transmiten una única longitud de onda), deben adaptarse al entorno multi-canal, para poder aplicar los modos de operación SHWP/SCWP. Las modificaciones *hardware* que necesarias, deben ser tenidas en cuenta en las comparativas de costes del conmutador.

(2) La planificación del conmutador para el modo de operación SHWP debe ser especificada. Los objetivos de este proceso de planificación serán optimizar las prestaciones del conmutador, atendiendo o no al posible desorden dentro del nodo.

(3) La planificación del conmutador para el modo de operación SCWP debe ser especificada. Esta planificación incluye el mecanismo de decisión sobre la longitud de onda de salida para cada paquete.

Como veremos a lo largo de esta tesis doctoral, este proceso de adaptación es seguido mecánicamente para todas las arquitecturas comparadas, como paso previo a su posterior evaluación. De nuevo, esto nos permitirá responder a preguntas como qué arquitectura es preferible dado un modo de operación, o cuantificar las consecuencias en coste que supone la aplicación de un modo de operación en cada arquitectura.

Sin pretender realizar un trabajo enciclopédico, nos centraremos en la evaluación de aquellas arquitecturas de conmutación propuestas en la literatura que han tenido una mayor relevancia, o mejores prestaciones fruto de comparativas anteriores. Para consultar otras propuestas el lector puede dirigirse a [Haa93][Cho95][Chl96][Hun98-1][Gui99][Sas97].

El desarrollo de la tesis doctoral continúa en los siguientes capítulos de la siguiente manera:

- **Capítulo 2**: Se enfoca en un subconjunto de arquitecturas de conmutación OPS, con capacidad de emular el comportamiento de un conmutador con colas a la salida [Hlu88].
  - Conmutador KEOPS, propuesto dentro del Proyecto Europeo ACTS (*Advanced Communications Technologies and Services*) KEOPS (*KEys to Optical Packet Switching*) [Gui98].
  - Conmutador *Output Buffered Wavelength-Routed Optical Packet Switch,* propuesto dentro del proyecto A49702803 del ARC (*Australian Research Council*) [Zho98].
  - Conmutador *Space Switch* [Dan98], asociado también al proyecto KEOPS.

  En este capítulo, los *3* puntos anteriores del proceso de adaptación serán especificados para cada una de las arquitecturas. Para la planificación SCWP, serán descritos dos algoritmos de selección de longitud de onda y de retardo, propuestos dentro de esta tesis doctoral. Se demostrará su optimalidad cuando es aplicado a arquitecturas con capacidad de emular colas a la salida. La evaluación de prestaciones y costes, se realizará de manera comparativa para las tres arquitecturas. El trabajo expuesto en este capítulo contiene lo publicado en [Pav03-2][Pav03-3] y [Pav03-5].

- **Capítulo 3**: En este capítulo nos centraremos en la evaluación de la arquitectura *Input Buffered Wavelength-routed Switch* propuesta dentro del proyecto A49702803 del ARC (*Australian Research Council*) [Zho98]. Su interés se fundamenta en su menor complejidad *hardware* respecto a otras arquitecturas. Por contra, presenta una mayor dificultad de planificación, en cuanto a la decisión de asignación de retardo. Nuestro estudio se basa en la

formalización de este problema de planificación, y la caracterización del mismo como un problema de Emparejamiento Máximo en Grafos Bipartitos [Bon76]. Este capítulo contiene lo publicado en [Pav03-1], junto con la propuesta y evaluación del algoritmo de planificación SCWP PDBM (*Parallel Desynchronized Block Matching*).

- **Capítulo 4**: Este capítulo se enfoca en el estudio de las arquitecturas de conmutación OPS de gran escala (alto número de puertos). Después de una descripción del estado del arte en este campo (fruto del trabajo publicado en [Pav02]), se describirá la propuesta realizada de arquitecturas *knock-out*, donde se han obtenido prometedores resultados en el modo de operación SCWP [Pav03-4]. El capítulo finaliza con una comparativa de costes de distintas alternativas en este campo.

- **Capítulo 5**: Este capítulo conluye esta tesis doctoral. Se destacarán las conclusiones obtenidas, y se propondrán un conjunto de líneas de investigación que aparecen como prometedoras en el ámbito de la Conmutación Óptica de Paquetes.

Capítulo 2. Arquitecturas OPS de colas a la salida

2.1 Introducción

En este capítulo nos centraremos en el estudio de las arquitecturas de conmutación OPS con capacidad de emular colas a la salida. Se describirán algoritmos de planificación SHWP y SCWP que ofrecen prestaciones óptimas. Posteriormente, estas prestaciones serán evaluadas, y los costes comparados para tres arquitecturas OPS, que han destacado por la atención recibida en la literatura. El trabajo expuesto en este capítulo contiene lo publicado en [Pav03-2][Pav03-3] y [Pav03-5]. El algoritmo de planificación SCWP uniforme presentado será empleado en el capítulo *4*, en la evaluación de arquitecturas *knock-out* de gran escala.

2.2 Descripción de las arquitecturas

2.2.1 Conmutador KEOPS

A mediados de la década de los 90, el proyecto europeo ACTS KEOPS (*Keys to Optical Packet Switching*) enfocó sus esfuerzos en la viabilidad de la Conmutación Óptica de Paquetes como base de la red OTP-N (*Optical Transparent Packet Network*) [Gui98]. Fruto de este esfuerzo, se propuso el conmutador KEOPS, de tipo difusión-selección (*broadcast-and-select*). Asimismo, se construyó un prototipo de *16x16* puertos, operando a *10 Gbps*, con un tamaño de ranura de paquete de *1,646* μ*s* (*1680* bytes de datos a *10 Gbps*). El diseño original del conmutador KEOPS, mostrado en la figura 2-1-(a), fue un desarrollo del conmutador OASIS, del proyecto RACE R2039 ATMOS (*Asynchronous Transfer Mode Optical Switching*).

Según el diseño original, un conmutador KEOPS de *N* puertos de entrada y *N* puertos de salida, opera del siguiente modo:

1) Los paquetes a la entrada se convierten a una longitud de onda fija, y distinta para cada puerto. Para ello se requieren *N* convertidores de longitud de onda fija (FWC, *Fixed Wavelength Converter*), cada uno de ellos a una longitud de onda distinta $\lambda_0,\ldots,\lambda_{N-1}$.
2) A continuación, las *N* señales provenientes de los *N* puertos de entrada, son combinadas y difundidas por las *M* líneas de retardo. Las longitudes de las líneas de retardo, retrasan el paquete un número entero *0...M-1* de ranuras temporales.
3) Cada puerto de salida dispone de dos etapas de puertas ópticas que permiten elegir el paquete a transmitir por ese puerto en cada ranura temporal. La primera etapa de puertas ópticas selecciona la línea de retardo de la que se recoge el paquete, y por tanto el tiempo de entrada del mismo. La segunda etapa de puertas ópticas selecciona la longitud de onda del paquete, y por lo tanto el puerto de entrada original del mismo.

La arquitectura KEOPS no requiere de ningún componente sintonizable. Soporta tráfico *multicast*, y es capaz de emular el comportamiento de un conmutador con colas a la salida, asignando de manera incremental los retardos a los paquetes entrantes destinados al mismo puerto de salida. Asimismo, permite un cierto grado de priorización de tráfico, que no se encuentra en otras arquitecturas basadas en líneas de retardo. Esto es así, ya que un paquete se encuentra disponible para ser seleccionado por cualquier puerto de salida durante *M* ranuras temporales consecutivas. La decisión de planificación sobre los paquetes salientes se podría tomar por tanto en el momento de la salida, y no en el momento de entrada del paquete en el conmutador, como sucede en otras arquitecturas. Esto permitiría el diseño de un sistema de planificación del conmutador en el que algunos paquetes pudiesen adelantar a otros llegados anteriormente, en función de una marca de prioridad. Hasta el conocimiento del autor, esta funcionalidad aún no ha sido explorada.

La desventaja fundamental de la arquitectura KEOPS es su deficiente escalabilidad debido al crecimiento del número de puertas ópticas necesarias en su implementación ($NM+N^2$), y a las pérdidas de potencia de señal debidas a su funcionamiento *broadcast*, proporcionales a NM^2. Siendo un dato fuertemente dependiente de la evolución de la tecnología, el límite barajado por los diseñadores es de arquitecturas de hasta *32x32* puertos y pocas decenas de retardos [Raf00-1] [Raf00-2].

El conmutador KEOPS, tal y como se ha mostrado en la figura 2-1-(a), no fue diseñado para trabajar con fibras de entrada y salida WDM. Por ello, la integración de este conmutador en una red OPS WDM requiere la adaptación a este escenario (paso *1* del proceso de normalización de arquitecturas). Las modificaciones de la arquitectura aplicadas se ilustran en la figura 2-1-(b).

- Se ha añadido una etapa de demultiplexación WDM en las fibras de entrada, que separan los paquetes entrantes por su longitud de onda, en *n* fibras distintas.
- Las *nN* fibras resultantes son conectadas a un conmutador KEOPS de *nNxnN* puertos de entrada y salida.
- En el conmutador original, no hay control sobre la longitud de onda de salida de un paquete, que está únicamente determinada por su puerto de entrada (longitudes de onda internas al *switch* $\lambda^s_0,\ldots,\lambda^s_{nN-1}$). Por ello, para fijar la longitud de onda de transmisión $\lambda^t_0,\ldots,\lambda^t_{n-1}$, cada uno de los *nN* puertos de salida del conmutador están conectados a un convertidor a longitud de onda fija, y posteriormente a un multiplexor para cada fibra de salida. De esta forma, la decisión sobre la fibra y longitud de onda de transmisión de un paquete se traduce en una decisión sobre su puerto de salida del conmutador KEOPS subyacente.

  - En el modo de operación SHWP, el OPP al que pertenece un paquete establece su fibra y longitud de onda de transmisión, y por tanto el puerto de salida del conmutador KEOPS subyacente.

  - Para el modo SCWP, la libertad en la selección de longitud de onda de salida se traduce en una libertad para conmutar el paquete hacia uno de los *n* puertos del conmutador KEOPS, asociados a la fibra de salida de destino.

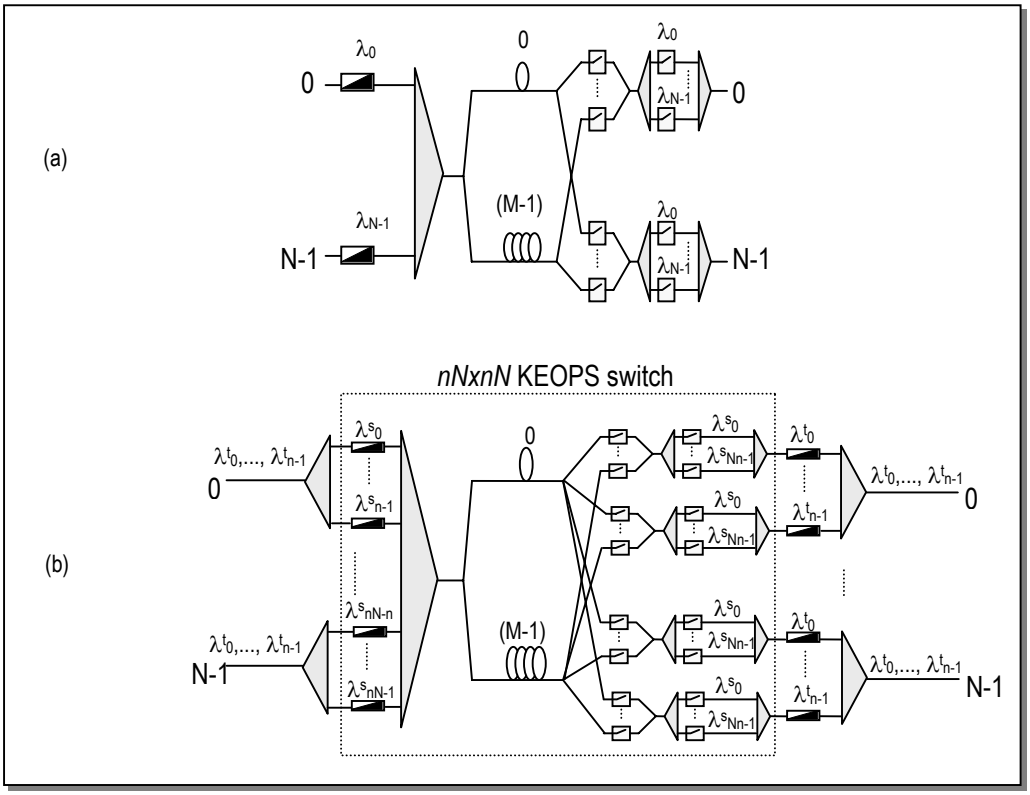


Figura 2-1. Arquitectura de conmutación *broadcast-and-select* KEOPS, (a) arquitectura original, (b) adaptación WDM

2.2.2 Conmutador OB-WR

Los progresos en la fabricación de dispositivos AWG (*Arrayed-Waveguide-Grating*) mediante tecnología planar [Tak90] potenció el estudio de conmutadores ópticos de paquetes que se beneficiasen de sus características de enrutamiento de la señal óptica. Una de las ventajas de este tipo de dispositivos es la posibilidad de ser combinados con líneas de retardo en el diseño de módulos de almacenamiento, como los mostrados en la figura 2-2-(a). Un módulo de almacenamiento de tamaño *NxN* puertos de entrada y salida, con retardos de *0* a *M-1* ranuras temporales, puede ser implementado con *N* convertidores de longitud de onda sintonizables (TWC, *Tunable Wavelength Converter*) de rango de sintonización $\lambda_0 \ldots \lambda_{K-1}$, y dos dispositivos AWG de tamaño *KxK* interconectados a través de las *M* líneas de retardo, donde *K=max (N,M)*.

El modo de operación cíclico respecto a la longitud de onda de transmisión de los dispositivos AWG, descrita en la figura 1-5, provoca que un paquete entrante por el puerto de entrada *i*-ésimo, salga por el puerto de salida *i*-ésimo del módulo, independientemente de la longitud de onda a la que es convertido por el dispositivo TWC. La longitud de onda es empleada para seleccionar el puerto de salida del primer AWG, y por tanto la línea de retardo que será atravesada. El retardo sufrido por cada paquete vendrá especificado por la regla λ*=(puerto E/S + retardo) mod K*.

Un conmutador OPS con capacidad de emular colas a la salida, construido mediante este tipo de módulo de almacenamiento fue presentado en [Zho98]. El conmutador original de tamaño *NxN*, denominado *Output-Buffered Wavelength-Routed (OB-WR) switch* se muestra en la figura 2-2-(a). Consiste en un conjunto de *N* convertidores TWC, un conmutador espacial sin bloqueo de tamaño *NxN*, construido

mediante interconexión de N^2 puertas ópticas, y un módulo de almacenamiento de *NxN* puertos.

Según el diseño original, un conmutador OB-WR de *N* puertos de entrada y *N* puertos de salida, opera del siguiente modo:

1) Los paquetes a la entrada se convierten a una longitud de onda determinada por el retardo que deben sufrir en el módulo de almacenamiento.
2) La etapa de conmutación espacial *NxN* (sin memoria) envía cada paquete hacia el puerto de entrada del módulo de almacenamiento, que es el mismo que el puerto de salida final del paquete. Paquetes simultáneos encaminados hacia el mismo puerto de salida, llegan al módulo de almacenamiento siempre en longitudes de onda distintas, en función del retardo que les ha sido planificado.
3) La etapa de almacenamiento encamina los paquetes entrantes por líneas de retardo distintas en función de su longitud de onda.

Como podemos observar, el conmutador OB-WR original no está diseñado para escenarios con fibras de entrada y salida WDM. Es de nuevo necesario realizar el paso *1* del proceso de adaptación, para poder aplicar los modos de operación SHWP y SCWP. La figura 2-2-(b) muestra los cambios propuestos a la arquitectura, similares a los realizados con la arquitectura KEOPS:

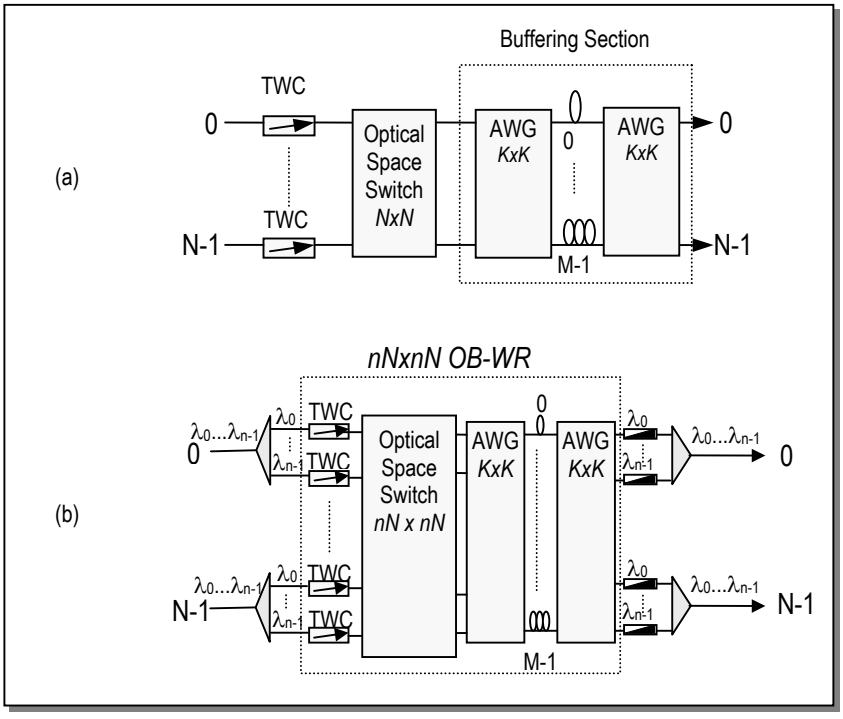


Figura 2-2. Arquitectura de conmutación OB-WR, (a) arquitectura original, (b) adaptación WDM

- Se ha añadido una etapa de demultiplexación WDM en las fibras de entrada, que separan los paquetes entrantes por su longitud de onda, en *n* fibras distintas.

- Las *nN* fibras resultantes son conectadas a un conmutador OB-WR de tamaño *nNxnN*.
- En el conmutador original, la longitud de onda de salida de un paquete viene determinada por el retardo que el paquete debe sufrir (longitudes de onda de uso interno al conmutador). Para fijar la longitud de onda de transmisión $\lambda_0,...,\lambda_{n-1}$, cada uno de los *nN* puertos de salida del conmutador están conectados a un convertidor a longitud de onda fija, y posteriormente a un multiplexor para cada fibra de salida. De esta forma, la decisión sobre la fibra y longitud de onda de transmisión de un paquete se traduce en una decisión sobre su puerto de salida del conmutador OB-WR subyacente, de manera similar a lo descrito para el conmutador KEOPS:

  - En el modo de operación SHWP, el OPP al que pertenece un paquete establece su fibra y longitud de onda de transmisión, y por tanto el puerto de salida del conmutador OB-WR subyacente.

  - Para el modo SCWP, la libertad en la selección de longitud de onda de salida se traduce en una libertad para conmutar el paquete hacia uno de los *n* puertos del conmutador OB-WR, asociados a la fibra de salida de destino.

2.2.3 Conmutador espacial (*space switch*)

Un conmutador OPS basado enteramente en conmutación espacial fue presentado en [Dan98]. El conmutador original descrito (figura 2-3), trabaja en un entorno WDM, de *N* fibras de entrada y *N* fibras de salida, con *n* longitudes de onda por fibra. Consiste en un conjunto de *nN* convertidores TWC, una etapa de conmutación espacial sin bloqueo de tamaño *nNxMN* (donde *M* es el número de retardos del conmutador), y una etapa final de almacenamiento formada por *N* grupos de *M* líneas de retardo, un grupo por cada fibra de salida.

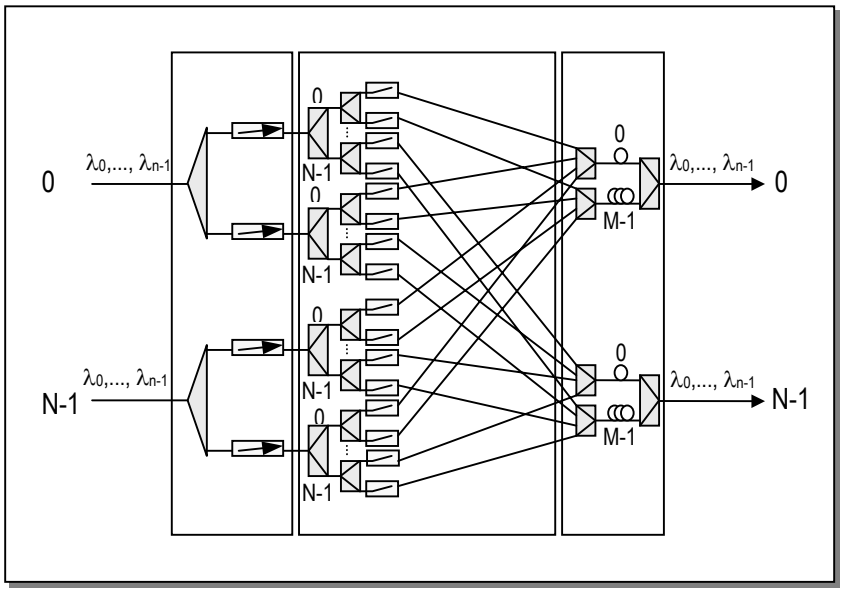


Figura 2-3. Arquitectura de conmutación *space switch*

El conmutador espacial así descrito opera del siguiente modo:

1) Los paquetes a la entrada se convierten a la longitud de onda de salida $\lambda_0 \ldots \lambda_{n-1}$ final del paquete (fijada por el identificador de OPP del paquete -SHWP-, o por un algoritmo de selección de longitud de onda –SCWP-).
2) La etapa de conmutación espacial encamina cada paquete al puerto de salida determinado por el retardo y la fibra de salida que corresponde al paquete. Distintos paquetes pueden llegar simultáneamente al mismo retardo de salida, en distintas longitudes de onda.
3) Después de atravesar la línea de retardo correspondiente, los paquetes son multiplexados hacia la fibra de salida asociada.

Como podemos observar, este conmutador ha sido diseñado para ser integrado en un entorno de puertos de entrada y salida WDM, con lo que no es necesaria ninguna adaptación *hardware* para la aplicación de los modos de operación SHWP y SCWP.

2.3 Planificación del conmutador

En esta sección describiremos los algoritmos de planificación SHWP y SCWP, que corresponden a los pasos *2* y *3* del proceso de adaptación seguido. Estos algoritmos serán posteriormente aplicados a las arquitecturas descritas en la sección anterior. Podrían asimismo aplicarse a otras arquitecturas basadas en líneas de retardo con colas a la salida, manteniendo las mismas propiedades.

2.3.1 Planificación SHWP

En el modo de operación SHWP, cada paquete entrante tiene determinado el puerto de salida por el que debe ser transmitido (obtenido a partir de su identificador de OPP). Por ello, la única decisión que se requiere es respecto al retardo que se debe asignar a cada paquete, o el posible descarte del mismo.

Los objetivos de esta decisión de planificación son:

- Optimizar el caudal de salida: un puerto de salida se encuentra inactivo en una ranura temporal únicamente si no existen paquetes almacenados destinados a ese puerto.
- Optimizar el retardo medio: un paquete es asignado el retardo mínimo posible, acotado por la contención a la salida.
- Mantenimiento del orden extremo a extremo: en las redes SHWP, el orden extremo a extremo se mantiene asegurando que cada nodo de conmutación no genere desorden interno. Dados p_i, p_j, $i<j$, paquetes asociados al mismo OPP, el método requiere que en cada nodo p_i sea transmitido antes de p_j.

La solución óptima obvia, que cumple todos nuestros requisitos, es la asignación de retardos emulando un comportamiento de colas a la salida con disciplina FIFO (*first-in first-out*). En los conmutadores OPS basados en líneas de retardo como los descritos en este capítulo, esto se consigue mediante la utilización de un contador para cada puerto de salida, que asigne retardos crecientes a los paquetes destinados al puerto asociado, hasta el número de retardos que dispone el conmutador. Después de cada ranura temporal, el contador de un puerto de salida debe ser decrementado en el caso de que un paquete haya sido transmitido. La figura 2-4 muestra el algoritmo de planificación SHWP.

```
/* N = n° de fibras entrada/salida */
/* n = n° de long. de onda por fibra */
/* M = n° de retardos */
/* p0 = primer puerto de entrada visitado */
/* delay [f,λ] = próximo retardo a asignar al puerto de salida
                  asociado a f, λ  */

for input i = 0 to nN-1 do
  if (packet p in input (i + p0) mod nN) then
    f,λ = output fiber and wavelength of p /* ( dependiente de opp (p) ) */
    if (delay [f,λ] < M) then
       associate delay [f,λ] to p
       delay [f,λ] ++
     else
       packet p is lost
     endif
  endif
endfor

/* Rotar p0 tras cada ranura temporal, para ser justo
con las fuentes de tráfico */
p0 = (p0 + 1) mod nN

/* decrementar delay [f,λ], f=0..N-1, λ=0..n-1 tras cada ranura temporal */

for output fiber f=0 to N-1 do
  for output wavelength λ=0 to n-1 do
    delay [f,λ] = max (0, delay [f,λ]-1)
  endfor
endfor
```

Figura 2-4. Algoritmo de planificación SHWP (pseudocódigo)

2.3.2 Planificación SCWP

En arquitecturas con colas a la salida, el modo de operación SCWP supone tomar dos decisiones para cada paquete entrante: la longitud de onda de salida, y el retardo a asignar (o el posible descarte del paquete). Las decisiones de retardo y longitud de onda para paquetes destinados a fibras de salida distintas son independientes. El modelo del sistema bajo estudio se reduce por tanto a las posibilidades de distribución de paquetes entre *n* colas de *M* posiciones cada una, correspondientes a las *n* longitudes de onda de una fibra de salida bajo estudio (figura 2-5-(a)). El objetivo es de nuevo encontrar un algoritmo que maximice el *throughput*, minimice el retardo, y permita mantener el orden entre paquetes.

El orden extremo a extremo entre paquetes del mismo OPP en redes SCWP ha sido tratado en profundidad en [Niz98]. La diferencia respecto a las redes SHWP, estriba en que más de un paquete perteneciente al mismo OPP puede ser transmitido simultáneamente, en longitudes de onda distintas. Por ello, es necesario un criterio que permita a un nodo conocer cuál es el orden entre paquetes llegados simultáneamente, con el objetivo de mantener ese orden hacia el siguiente salto.

En [Niz98] fue descrito un algoritmo de planificación SCWP que permite mantener la secuenciación de paquetes sin necesidad de un contador por OPP en la cabecera del paquete, para la versión retroalimentada (*feedback*) del conmutador WASPNET [Hun99]. La necesidad de un contador en la cabecera del paquete es una solución a evitar ya que aumenta el tamaño de la misma y añade complejidad al procedimiento. El método especificado en [Niz98] se basa en la transmisión de los paquetes simultáneos del mismo OPP, ordenados por su longitud de onda, de tal forma que los paquetes de orden más bajo deben transmitirse en longitudes de onda

menores. Dados p_i, p_j, $i<j$, paquetes asociados al mismo OPP, el método requiere que: (1) p_i sea transmitido antes de p_j, o (2) p_i sea transmitido durante la misma ranura temporal que p_j, y $\lambda_i<\lambda_j$. El orden extremo a extremo dentro del mismo OPP se mantiene, si todos los nodos atravesados cumplen las condiciones anteriores.

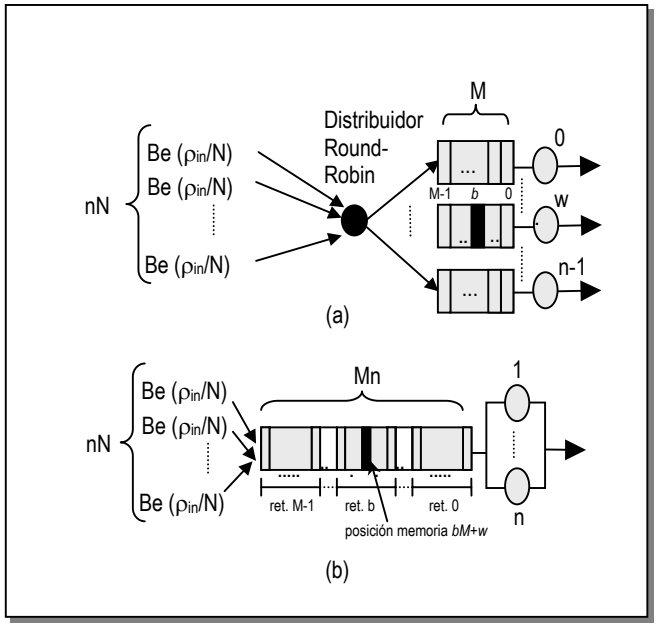


Figura 2-5. (a) Modelo de colas para un conmutador de colas a la salida de *N* fibras de entrada y salida, *n* longitudes de onda por fibra, *M* retardos, (b) Modelo de colas equivalente por la aplicación del algoritmo de planificación SCWP propuesto

En [Pav03-2] se propone una modificación del algoritmo presentado en [Niz98], adaptado para la aplicación a conmutadores que emulen colas a la salida, para los cuales presenta la característica de proporcionar unas prestaciones óptimas, permitiendo mantener el orden extremo a extremo, de nuevo sin la necesidad de un número de orden dentro de la cabecera del paquete.

El algoritmo propuesto en [Pav03-2] se muestra en la figura 2-6. Su funcionamiento se resume en los siguientes puntos:

- Asignación cíclica (*round-robin*) de las longitudes de onda λ_0 a λ_{n-1}, para paquetes destinados a la misma fibra de salida. Para ello se requiere un puntero *round-robin* independiente para cada fibra de salida.
- Un puntero *round-robin* es puesto a *0* cuando tras una ranura temporal, no existen paquetes destinados a su fibra de salida.
- Para todas las fibras de entrada, los paquetes en los puertos correspondientes a las longitudes de onda menores son asignados retardo y longitud de onda de salida antes.

Esto proporciona el comportamiento óptimo para arquitecturas con capacidad de emular colas a la salida:

- Un paquete siempre sufre el mínimo retardo disponible: se asigna a un paquete un retardo *d* y una longitud de onda *w* si y sólo si las colas *0...w-1* tienen *d* *(0≤d≤M-1)* paquetes almacenados, y las colas *w...n-1* tienen *max(0,d-1)* paquetes almacenados. La puesta a cero del puntero *round-robin* asegura el cumplimiento de esta condición también tras el vaciado de las colas asociadas a la fibra.
- El *throughput* se maximiza, ya que un paquete es descartado únicamente cuando las *n* colas correspondientes a las *n* longitudes de onda de su fibra de salida están ocupadas, y *Mn* paquetes están almacenados.
- Se preserva el orden de los paquetes en los mismos términos que en los indicados en [Niz98]: (1) para las fibras de salida, los paquetes de menor orden se transmiten en longitudes de onda menores, (2) para las fibras de entrada, los puertos correspondientes a longitudes de onda menores son visitados primero.
- Debido a la operación con *paquetes de tamaño fijo*, existe una identificación biunívoca entre la posición M_b, $0 \le b \le M-1$ en la cola de longitud de onda de salida λ_w, $0 \le w \le n-1$, y la posición de búffer M_{bn+w} de una cola equivalente con *n* servidores y *Mn* posiciones de memoria (ver figura 2-5-(b)). Ambos sistemas son indistinguibles, y por lo tanto presentan las mismas prestaciones para todos los tráficos de entrada.

```
/* N = n° de fibras entrada/salida */
/* n = n° de long. de onda por fibra */
/* M = n° de retardos */
/* delay [f] = próximo retardo a asignar a paquetes destinados a fibra f */

for input i = 0 to nN-1 do
  if (packet p present on input i) then
    f = output fiber demanded by p /* ( dependiente de opp (p) ) */
    if (delay [f] < M) then
      associate delay [f] to p
      associate wav. λ [f] to p
      /* λ [f] es un puntero RR */
      λ [f] ++
      if (λ [f] == n)
        λ [f] = 0
        delay [f] ++
      endif
    endif
  endif
endfor

/* decrementar delay [f,λ], f=0..N-1, λ=0..n-1 tras cada ranura temporal */

for output fiber i=0 to N-1 do
  if (delay [f] == 0)
    λ [f] = 0 /* reset puntero RR */
  else
    delay [f] --
  endif
endfor
```

Figura 2-6. Algoritmo de planificación SCWP [Pav03-2] (pseudocódigo)

2.3.2.1 Análisis de la distribución de la selección de longitud de onda

El algoritmo de planificación SCWP no selecciona con la misma probabilidad todas las longitudes de onda de una fibra de salida. Esto es debido a que la asignación no se realiza por un proceso *round-robin* puro, sino que un puntero es reiniciado si tras una ranura temporal las *n* colas correspondientes a su fibra de salida están vacías (o lo que es lo mismo, se han transmitido menos de *n* paquetes en la ranura temporal anterior). Como consecuencia, para cualquier distribución de tráfico de entrada se cumple

$$P\left[\lambda_i\right] \geq P\left[\lambda_j\right], \forall j > i \qquad \text{(Ec. 2.1)}$$

Donde *P[λ]* indica la probabilidad de que, *si* un paquete se transmite (no es descartado), se seleccione la longitud de onda λ.

$$P\left[\lambda_j\right] = E\left(paq.\ transmitido\ \ por\ \lambda_j \middle|\ \ paq.\ transmitido\right), j = 0,...,n-1 \qquad \text{(Ec. 2.2)}$$

Para conocer la probabilidad de selección $P[\lambda_j]$, $j=0...n-1$, debemos enfocar nuestro estudio hacia el número de paquetes que se transmiten entre dos reinicios (*reset*) de puntero de una cola multiservidor. Un puntero es reiniciado cuando en la ranura temporal anterior se han transmitido menos de *n* paquetes. Por ello, el numero de paquetes transmitidos entre dos *reset* del puntero es igual al tamaño (en número de paquetes) del periodo ocupado de la cola multiservidor. Supongamos que conocemos la distribución de esta variable aleatoria *B* (de *busy period*).

$B_i = P[B=i],\ i=1,2,...$ (Ec. 2.3)

Esto nos indica que:

- Con probabilidad B_1, λ_0 será elegida *1* vez, y $\lambda_1...\lambda_{n-1}$ *0* veces.
- Con probabilidad B_2, λ_0, λ_1 serán elegidas *1* vez, y $\lambda_2...\lambda_{n-1}$ *0* veces.
- ...
- Con probabilidad B_n, $\lambda_0...\lambda_{n-1}$ serán elegidas *1* vez.
- Con probabilidad B_{n+1}, λ_0 será elegida *2* veces, y $\lambda_1...\lambda_{n-1}$ *1* vez.
- ...
- En general se cumple que, con probabilidad B_i, λ_j será elegida $\left\lceil \frac{i-j}{n} \right\rceil$ veces.

Sea *T* el periodo de observación, medido en número de periodos ocupados consecutivos.

$$P\left[\lambda_j\right] = \lim_{T\to\infty} \frac{paquetes\ tx.\ \lambda_j}{paquetes\ tx.} = \lim_{T\to\infty} \frac{\sum_{i=1}^{\infty} T\cdot B_i \cdot \left\lceil \frac{i-j}{n} \right\rceil}{\sum_{i=1}^{\infty} T\cdot B_i \cdot i} =$$

$$= \lim_{T\to\infty} \frac{\sum_{i=1}^{\infty} B_i \cdot \left\lceil \frac{i-j}{n} \right\rceil}{\sum_{i=1}^{\infty} B_i \cdot i}, j = 0...n-1 \qquad \text{(Ec. 2.4)}$$

Y por lo tanto

$$P\left[\lambda_j\right] = \frac{\sum_{i=1}^{\infty} B_i \cdot \left\lceil \frac{i-j}{n} \right\rceil}{E(B)}, j = 0...n-1 \qquad \text{(Ec. 2.5)}$$

Que es la expresión que relaciona las probabilidades de selección de longitud de onda con la distribución de probabilidad de periodo ocupado.

2.3.2.2 Cálculo de periodo ocupado de una cola multiservidor

En esta tesis doctoral, la distribución de periodo ocupado se ha resuelto siguiendo una aproximación sencilla, válida para colas multiservidor finitas, con tráfico de entrada independiente idénticamente distribuido (IID).

La cola en estudio es la mostrada en la figura 2-5-(b), correspondiente a una fibra de salida de un conmutador SCWP de *N* fibras de entrada y salida, *n* longitudes de onda por fibra, y *M* retardos por puerto de salida (y por tanto *nM* retardos de la cola multiservidor equivalente). La evolución de la cola multiservidor viene expresada por una cadena discreta de Markov. Denotando Q_m como el número de paquetes en la cola al final de la ranura temporal *m*, y A_m como el número de llegadas durante la ranura temporal *m*, se obtiene:

$$Q_{m+1} = min\{max\{0, Q_m - n\} + A_m, nM\} \qquad \text{(Ec. 2.6)}$$

La variable aleatoria A_m se asume independiente e idénticamente distribuida para todas las ranuras temporales *m*.

$$P[A_m = k] = P[A = k] = a_k, k = 0...nN \qquad \text{(Ec. 2.7)}$$

La figura 2-7 muestra las variables en juego para el análisis de la distribución del periodo ocupado:

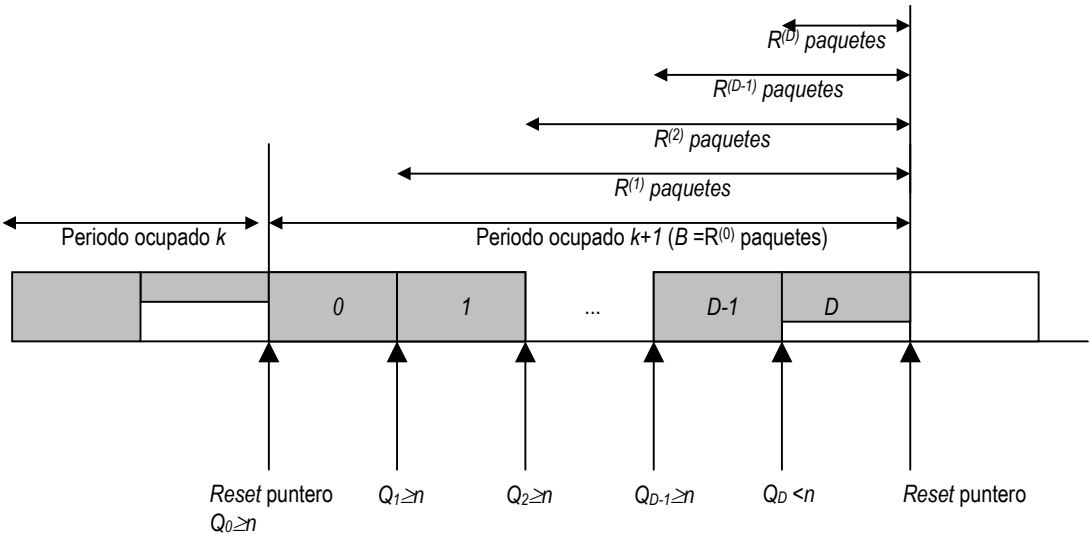


Figura 2-7. Periodo ocupado y periodo ocupado residual

Donde la variable aleatoria $R^{(i)}$, representa el periodo ocupado residual que comienza en la *i*-ésima ranura del periodo ocupado: el número de paquetes del periodo ocupado actual transmitidos desde la ranura *i* en adelante, *i=0,...,D* (en nuestro ejemplo). Nótese que $R^{(0)}=B=\{duración\ periodo\ ocupado\}$.

Para patrones de llegadas independientes e idénticamente distribuidas, la distribución de probabilidades de un periodo ocupado residual $R^{(i)}$, *i=0,1,...* depende *únicamente* del estado Q_i del sistema al comienzo de la *i*-ésima ranura. En especial, no depende del número *i* de ranuras temporales pasadas dentro de este periodo ocupado. Por ello, la distribución de probabilidad de un periodo ocupado con *P* paquetes en el estado inicial, y la de un periodo ocupado residual con *P* paquetes en el estado inicial, son la misma.

$$P\left[R^{(i)}=k\middle|Q_1=L\right]=P\left[R^{(0)}=k\middle|Q_0=L\right]=P\left[B=k\middle|Q_0=L\right],k=1,2,... \qquad \text{(Ec. 2.8)}$$

Aplicando el teorema de las probabilidades totales a la primera ranura temporal tenemos que:

$$\begin{aligned} B_k &= P[B=k]=\frac{1}{1-P[Q_0=0]}\sum_{L=1}^{nN}P\left[B=k\middle|Q_0=L\right]P[Q_0=L]= \\ &=\frac{1}{1-P[Q_0=0]}\sum_{L=1}^{nN}P\left[R^{(0)}=k\middle|Q_0=L\right]P[Q_0=L],k=1,2,... \end{aligned} \qquad \text{(Ec. 2.9)}$$

Donde Q_0 indica el número de paquetes en el sistema al comienzo del periodo ocupado, igual al número de paquetes llegados en la ranura temporal anterior A_{-1}. Los valores posibles de A_{-1} van desde *1* hasta *nN* (*0* llegadas implicaría que el periodo ocupado no comenzase en la ranura temporal *0*). Por ello, el teorema de las probabilidades totales requiere el factor $\frac{1}{1-P[Q_0=0]}$ en el sumatorio.

Las probabilidades condicionadas se calculan de la siguiente manera:

$$P\left[B=k\middle|Q_0=L\right]=\begin{cases}0\ si\ L>k\\0\ si\ L<n,k\neq L\\1\ si\ L<n,k=L\\P\left[R^{(1)}=k-n\right]si\ n\leq L\leq k\end{cases}\qquad k=1,2,...;L=0,...,nN \qquad \text{(Ec. 2.10)}$$

- La primera igualdad indica que la duración de un periodo ocupado no puede ser menor al número de llegadas en su primera ranura temporal.
- La segunda y tercera igualdades se aplican a periodos ocupados de duración *1* ranura temporal (menos de *n* paquetes). Esto proporciona los valores:

$$B_1=\frac{a_1}{1-a_0},B_2=\frac{a_2}{1-a_0},...,B_{n-1}=\frac{a_{n-1}}{1-a_0} \qquad \text{(Ec. 2.11)}$$

- La tercera igualdad se aplica para periodos ocupados en los que *n* paquetes son transmitidos en la primera ranura temporal. Por lo tanto, *k-n* paquetes son transmitidos a partir de la siguiente ranura temporal. La probabilidad de que esto suceda, depende del número de paquetes en el sistema después de la ranura temporal. Aplicando el teorema de las probabilidades totales

$$\begin{aligned}P\left[R^{(1)}=k-n\right]&=\sum_{i=0}^{nN}a_i\cdot P\left[R^{(1)}=k-n\middle|Q_1=min(max(0,L-n)+i,nM)\right]=\\&=\sum_{i=0}^{nN}a_i\cdot P\left[B=k-n\middle|Q_0=min(max(0,L-n)+i,nM)\right]\end{aligned} \qquad \text{(Ec. 2.12)}$$

Esta es la base del método, ya que permite calcular iterativamente las probabilidades condicionadas, en valores de *k* crecientes, a partir de los valores calculados para *k-n*. Para el caso en que *k=n*, se requiere calcular $P\left[B=0\middle|Q_0=min(max(0,L-n)+i,nM)\right]$, que vale *1* para el valor $Q_0=0$.

Las ecuaciones Ec. 2.13a y Ec. 2.13b resumen el método de cálculo iterativo para valores crecientes de *k* de las probabilidades condicionadas, y su aplicación en el cálculo de la distribución de probabilidad de periodo ocupado.

$$P\left[B=0\middle|Q_0=L\right]=\begin{cases}0\ si\ L>k\\0\ si\ L<n,k\neq L\\1\ si\ L<n,k=L\\\sum_{i=0}^{nN}a_i\cdot P\left[B=k-n\middle|Q_0=min(max(0,L-n)+i,nM)\right]si\ n\leq L\leq k\end{cases}$$

$$L=0,...,nN \qquad \text{(Ec. 2.13a)}$$

$$k=0,1,...$$

$$B_k=\frac{1}{1-a_0}\sum_{L=1}^{nN}P\left[B=k\middle|Q_0=L\right]a_L \qquad \text{(Ec. 2.13b)}$$

Nótese que el cálculo de la distribución de periodo ocupado se realiza sin necesidad de resolver previamente las probabilidades de estado de la cadena de Markov ($Q_0,...,Q_{nN}$). El método es válido para patrones de tráfico de entrada IID. Esto incluye el caso de conmutadores con carga Bernouilli uniforme, y carga Bernouilli no uniforme (*hot-spot*).

2.3.2.3 Resultados de la distribución de la selección de longitud de onda

En esta sección se realiza un estudio de la desigualdad en la selección de longitud de onda de salida para conmutadores SCWP de *N* fibras de entrada y salida, *n* longitudes de onda por fibra, y un número de retardos suficiente para provocar unas pérdidas no significativas. Se toma esta decisión para mostrar con mayor claridad los efectos en la probabilidad de selección, de la variación de la carga de entrada, fibras y longitudes de onda de los conmutadores.

El tráfico de entrada se asume de tipo Bernouilli uniforme de carga por puerto de entrada ρ.

$$a_k = P[A = k] = \binom{nN}{k}\left(\frac{\rho}{N}\right)^k\left(1-\frac{\rho}{N}\right)^{nN-k}, k = 0,...,nN \qquad \text{(Ec. 2.14)}$$

Que para $N = \infty$ se convierte en

$$a_k = P[A = k] = \frac{\left(\frac{\rho}{n}\right)^k e^{-\frac{\rho}{n}}}{k!}, k = 0,1,... \qquad \text{(Ec. 2.15)}$$

La figura 2-8 muestra un ejemplo de la variación de la probabilidad de selección de longitud de onda, para distintos valores de la carga de entrada, en un conmutador *64x64* de *4* fibras de entrada y salida (*N*), y *16* longitudes de onda (*n*). El efecto observado es una selección muy desbalanceada para cargas bajas, que corresponden con tamaños de periodo ocupado menores, donde el *reset* del puntero se produce más frecuentemente. En cargas altas, la selección de longitud de onda de salida se va igualando, hasta valores cercanos a la distribución uniforme en cargas superiores a *0.8*.

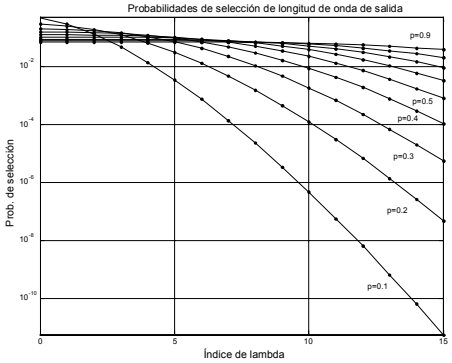
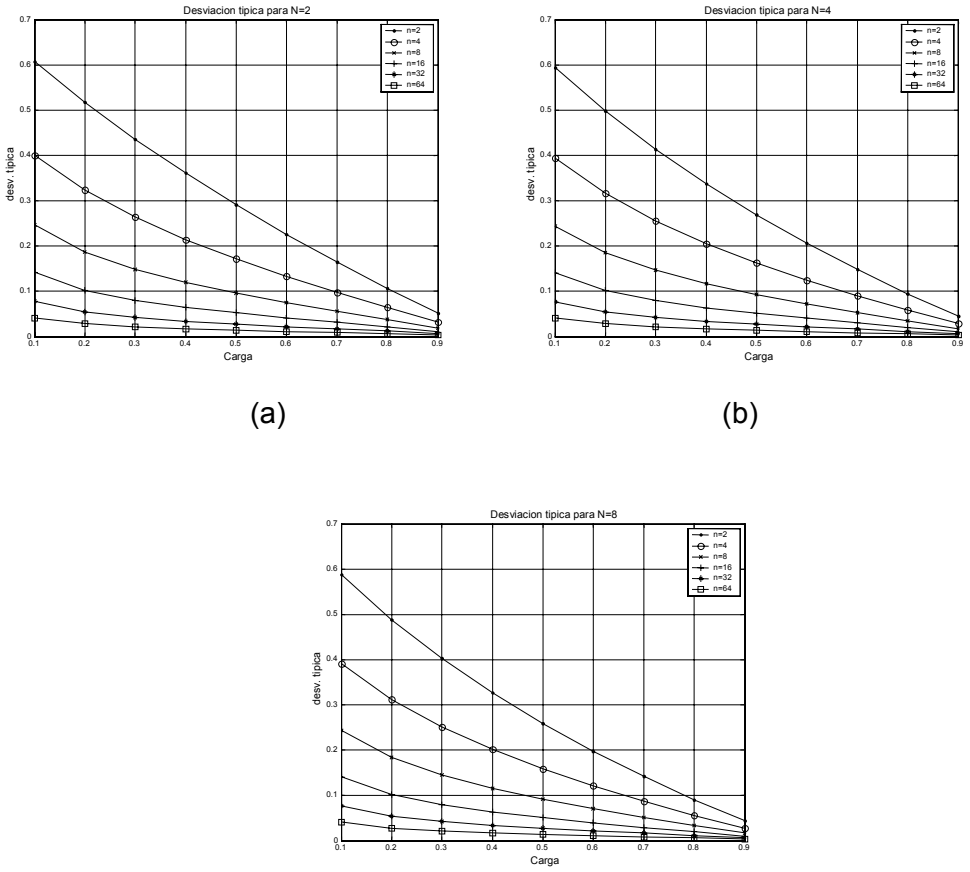


Figura 2-8. Distribución de la probabilidad de selección de longitud de onda de salida para un conmutador *64x64, N=4, n=16, ρ={0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, M=∞*

Con el objetivo de facilitar el estudio de la evolución de la desigualdad con los parámetros *N*, *n* y ρ, se define una métrica de la desigualdad como la desviación típica (con el estimador normalizado a *n-1* muestras) de las probabilidades de selección para todas las longitudes de onda Ec. 2.16:

$$\sigma(N,n,\rho)=\sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}\left(P[\lambda_i]-\frac{1}{n}\right)^2} \qquad \text{(Ec. 2.16)}$$

Un valor de σ próximo a *0* indica por tanto una distribución semejante a la uniforme.Las figura 2-9 muestra los valores de σ, para distintos valores de {*N,n,*ρ}.



(a) (b)

(c)

Figura 2-9. Desviación típica σ(N,n,ρ), n={2,4,8,16,32,64}, ρ={0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, (a) *N*=2, (b) *N=4*, (c) *N=8*

De los valores mostrados puede deducirse lo siguiente:

- A igualdad del resto de parámetros, los valores son ligeramente mejores a medida que *N* crece.

  $\sigma(N-1,n,\rho)>\sigma(N,n,\rho),\forall n,\rho$

- Se observa un acercamiento asintótico a medida que *N* crece. La diferencia es muy pequeña para valores de *N>16*.

$$\lim_{N\to\infty} \sigma(N-1,n,\rho) - \sigma(N,n,\rho) = 0, \forall n,\rho$$

- Se observa una variación relevante de σ con el parámetro *n*, con valores más altos a medida que *n* decrece. Para un tamaño de conmutador *nN* fijo, valores más altos de *n* producen mayor uniformidad en la asignación.

$$\sigma(N,n-1,\rho) > \sigma(N,n,\rho), \forall N,\rho$$

- Se confirma para todos los valores de *N*, *n*, el comportamiento más desigual en la selección de longitudes de onda de salida para cargas bajas.

$$\sigma(N,n-1,\rho_1) > \sigma(N,n,\rho_2), \forall N,n,\rho_1 < \rho_2$$

2.3.2.4 Algoritmo de planificación SCWP uniforme

El algoritmo presentado en [Pav03-2], y empleado en [Pav03-3] y [Pav03-5], presenta los siguientes inconvenientes:

1) Provoca una injusticia con las fuentes de tráfico (*unfairness*), al visitar siempre las fibras de entrada comenzando en la *0* hasta la *N-1*. Como consecuencia, los retardos medios y probabilidades de pérdida son mejores que la media, para los paquetes pertenecientes a OPPs entrantes por fibras de índice bajo, y peores que la media, para los paquetes pertenecientes de OPPs entrantes por fibras de índice alto.
2) Presenta una no uniformidad en la selección de longitud de onda de salida, como ha sido mostrado en la sección anterior.

Las consecuencias al primer problema son graves en cualquier arquitectura de conmutación. Afortunadamente, su solución es trivial: en cada ranura temporal, debe rotarse la fibra de entrada que es visitada primero (variable `f0` en figura 2-10).

El segundo problema es totalmente transparente para las fuentes de tráfico: las longitudes de onda de transmisión de $n' \leq n$ paquetes por una fibra no son relevantes para arquitecturas con capacidad de emular colas a la salida como las indicadas, siempre que exista un mecanismo que permita mantener el orden entre paquetes. Resumiendo, esta desigualdad no provoca una injusticia de tráfico (*unfairness*) entre distintos OPP del estilo del punto (1).

Sin embargo, la selección de longitud de onda de salida sí es relevante (al menos) en tres casos:

- Desde un punto de vista de implementación, una selección no uniforme puede provocar una utilización no uniforme de los dispositivos que forman parte del conmutador. En ciertas arquitecturas basadas en dispositivos SOA, como el conmutador KEOPS o el conmutador OB-WR, esta no uniformidad tiene como consecuencia una disipación no uniforme de la potencia en

distintas partes del conmutador. Esto es debido a que los SOA están alimentados con corrientes de inyección altas para provocar el estado ON, y corrientes de inyección bajas para provocar el estado OFF. Mientras la absorción de la señal luminosa durante los estados OFF no requiere una gran disipación de calor, las altas corrientes de inyección en el estado ON sí. Los conmutadores KEOPS y OB-WR tienen simetría por puerto de salida. Los SOAs en las puertas ópticas de conmutación espacial y en los dispositivos FWC de salida, asociados a los puertos de longitudes de onda más seleccionados, tendrán unos ratios de estado ON superiores, y por tanto una necesidades de disipación de potencia superiores. Esta no uniformidad en la utilización y en la disipación debe ser generalmente evitada por motivos de diseño del sistema de refrigeración, y de fiabilidad de componentes.

- En el capítulo *3* se estudiará la arquitectura de conmutación *Input-Buffered Wavelength-Routed Switch* (IB-WR), para la que la distribución de longitudes de onda en las fibras de entrada sí es relevante a efectos de prestaciones. Por ello, en el caso de conectar un nodo de colas a la salida con un conmutador IB-WR, las prestaciones de este último se verán afectadas por el patrón de longitudes de onda transmitidas. En el capítulo *3*, se argumentará la ventaja de nuevo de una distribución que reparta de manera uniforme la ocupación de las longitudes de onda.
- En el capítulo *4* se propondrán un conjunto de arquitecturas *knock-out* para conmutadores OPS de gran escala. La arquitectura propuesta que ofrece prestaciones más prometedoras, emplea un módulo para cada longitud de onda de salida, en modo de operación SCWP. Las prestaciones de este conmutador en términos de pérdidas *knock-out* son mejores para el caso de una selección uniforme de longitud de onda de salida.

A continuación se propone una variación del algoritmo, que elimina los inconvenientes (1) y (2) anteriores, que llamamos algoritmo de planificación SCWP uniforme. Para ello, los punteros de selección de longitud de onda siguen una evolución *round-robin* pura, eliminándose su posible puesta a cero al final de una ranura temporal. Inicialmente, esto provocaría dos problemas:

- Pérdida de prestaciones, ya que en el comienzo de un periodo ocupado, el primer retardo podría no llenarse completamente.
- Incumplimiento del criterio utilizado en [Niz98], como referencia para mantener el orden entre paquetes transmitidos simultáneamente.

El primer problema se elimina incluyendo una variable que cuente el número de paquetes en el último retardo (`lastDelayOccup`). No se comenzará el llenado del retardo *R* hasta que el retardo *R-1* esté ocupado en las *n* colas asociadas.

El mantenimiento del orden entre paquetes, exige que el orden de lectura de los puertos de entrada del nodo siguiente, siga el orden de escritura *round-robin* puro del nodo anterior. Esto requiere un puntero *round-robin* de lectura para cada fibra de entrada ($\lambda_{in\ [fin]}$), que conserve la última longitud de onda utilizada en la ranura temporal anterior. Obsérvese que esto implica utilizar otro criterio distinto al expresado en [Niz98], para el ordenamiento entre paquetes simultáneos.

```
/* N = nº de fibras entrada/salida */
/* n = nº de long. de onda por fibra */
/* M = nº de retardos */
/* delay [f] = próximo retardo a asignar a paquetes destinados a fibra f */


for fiberCounter = 0 to N-1 do
  fin = (f0 + fiberCounter) mod N
  for wavCounter = 0 to n-1 do
    if (packet p in input (fin , λin [fin]) then
      fout = output fiber p /* ( dependiente de opp (p) ) */
      if (delay [fout] < M) then
        associate delay [fout] to packet p
        associate λout [fout] to packet p
        λout [fout] = (λout [fout] + 1) mod n
        lastDelayOccup [fout] ++
        if (lastDelayOccup [fout] == n) then
          lastDelayOccup [fout] = 0
          delay [fout] ++
        endif
      endif
      λin [fin] = (λin [fin] + 1) mod n //
    else
      /* se comprueban los índices de longitud de onda, hasta encontrar una
      longitud de onda libre, o se han recorido todas */
      break;
    end
  endfor
endfor

/* Proceso a realizar al final de esta ranura temporal */

f0 = (f0 + 1) mod N  // para garantizar la justicia entre fibras de entrada

for fout = 0 to N-1 do
  if (delay [fout] == 0)
    lastDelayOccup [fout] = 0
  else
    delay [fout] --
  endif
endfor
```

Figura 2-10. Algoritmo de planificación SCWP uniforme (pseudocódigo)

El resultado de añadir estas modificaciones se muestra en la figura 2-10. El *algoritmo SCWP uniforme* propuesto es justo en términos de tráfico, y uniforme en términos de selección de longitud de onda $\sigma(N,n,\rho)=0, \forall N,n, \rho$, para cualquier patrón de tráfico de entrada, con poca complejidad añadida. Asimismo, las prestaciones siguen siendo óptimas, al seguir existiendo la misma equivalencia con la cola multiservidor. Finalmente, el orden entre paquetes del mismo OPP también se mantiene, sin necesidad de un campo contador dentro de la cabecera del paquete.

2.4 Evaluación de arquitecturas

En este apartado se detalla el resultado del proceso de evaluación SHWP y SCWP para arquitecturas OPS con capacidad de emular colas a la salida, que corresponde con el trabajo compendiado en [Pav03-5].

2.4.1 Análisis de prestaciones

En esta sección se presenta la evaluación de conmutadores OPS con capacidad de emular colas a la salida bajo los modos de operación SHWP y SCWP. La comparación se basa en análisis por teoría de colas, validado a través de simulaciones.

El conmutador bajo estudio se asume como simétrico, con *N* fibras de entrada y de salida, *n* longitudes de onda $\lambda_0,..,\lambda_{n-1}$ por fibra, y *M* retardos. El análisis se realiza asumiendo tráfico de entrada Bernouilli uniformemente distribuido, de parámetro $\rho \leq 1$.

El proceso de selección de retardo para la conmutación SHWP ofrece el tradicional comportamiento FIFO con colas a la salida de tamaño *M*. La evaluación de este modelo se basa en el estudio de una cola de salida fijada, alimentada por la agregación de *Nn* fuentes Bernouilli de carga ρ/Nn. Bajo estas consideraciones, la solución se obtiene mediante el análisis tradicional de conmutadores electrónicos con colas a la salida, que no será reproducido en este documento (ver [Hlu88] para más detalles).

El análisis de prestaciones para el conmutador en modo SCWP, aplicando el algoritmo de planificación SCWP visto en la sección anterior (en su versión uniforme o no-uniforme), requiere la evaluación de prestaciones de la cola multiservidor equivalente (figura 2-5-(b)).

El tráfico de entrada a la cola multiservidor es el creado por la agregación de *nN* fuentes de tráfico, de carga por fibra de salida ρ/N, de la misma manera que en el estudio de periodo ocupado

$$a_k = P[A = k] = \binom{nN}{k}\left(\frac{\rho}{N}\right)^k\left(1-\frac{\rho}{N}\right)^{nN-k}, k = 0,...,nN \qquad \text{(Ec. 2.17)}$$

que para $N = \infty$ se convierte en

$$a_k = P[A = k] = \frac{\left(\frac{\rho}{n}\right)^k e^{-\frac{\rho}{n}}}{k!}, k = 0,1,... \qquad \text{(Ec. 2.18)}$$

Denotando Q_m como el número de paquetes en la cola al final de la ranura temporal *m*, y A_m como el número de llegadas durante la ranura temporal *m*, se obtiene

$$Q_{m+1} = min\{max\{0, Q_m - n\} + A_m, nM\} \qquad \text{(Ec. 2.19)}$$

Q_m se modela como una cadena de Markov finita con probabilidades de transición $P_{i,j} = P[Q_{m+1} = j | Q_m = i]$ dadas por Ec. 2.20, cuando $M \geq N$.

$$P_{i,j} = \begin{cases} a_j \quad si\ i \le n, j \le nN \\ a_{j-i+n} \quad si\ n+1 \le i \le nM, i-n \le j \le min(nM-1, nN+i-n) \\ \sum_{s=nM+n-i}^{nN} a_s \quad si\ n(M-N+1) \le i \le nM, j = nM \end{cases} \qquad \text{(Ec. 2.20)}$$

Para $N \geq M$ tenemos

$$P_{i,j} = \begin{cases} a_j \quad si\ i \le n, j < nM \\ \sum_{s=nM}^{nN} a_s \quad si\ i \le n, j = nM \\ a_{j-i+n} \quad si\ n+1 \le i \le nM, i-n \le j < nM \\ \sum_{s=nM+n-i}^{nN} a_s \quad si\ n+1 \le i \le nM, j = nM \end{cases} \qquad \text{(Ec. 2.21)}$$

Las probabilidades de estado en estado estacionario q_i, $i=0..nM$ pueden ser, por tanto, obtenidas directamente de las ecuaciones de balance de la cadena de Markov [Kle75]. El *throughput* del sistema ρ_{out} se calcula observando el número de paquetes transmitidos durante una ranura temporal,

$$\rho_{out} = \sum_{s=1}^{nM} q_s \cdot min(s,n) < n \qquad \text{(Ec. 2.22)}$$

La probabilidad de pérdida de paquete se calcula en función de los paquetes ofrecidos a la cola y los servidos por la misma (Ec. 2.22).

$$P[packet\ loss] = 1 - \frac{\rho_{out}}{n\rho} \qquad \text{(Ec. 2.23)}$$

El tiempo medio de espera en cola $\overline{W}$ para un paquete puede ser calculado aplicando la ley de Little.

$$\overline{W} = \frac{\overline{Q}}{\rho_{out}} = \frac{\sum_{s=1}^{nM} s \cdot q_s}{\sum_{s=1}^{nM} min(s,n) \cdot q_s} \qquad \text{(Ec. 2.24)}$$

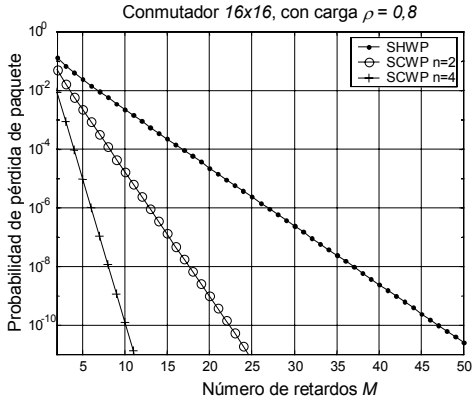
La obtención de las probabilidades de estado nos permite también el cálculo de la distribución del retardo D sufrido por un paquete, aplicando el resultado presentado en [Vin96]:

$$P[D=d]=\frac{1}{\rho_{out}}\sum_{p=-n+1}^{n-1}(c-|p|)q_{nd+p},\ d=1,2,... \qquad \text{(Ec. 2.25)}$$

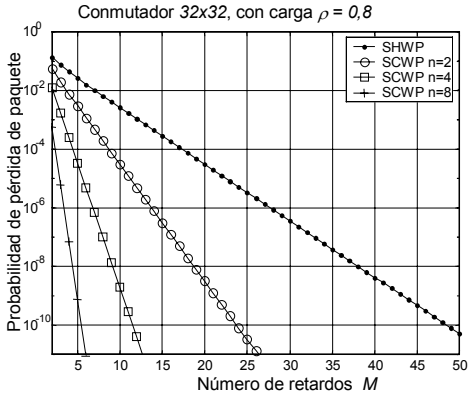
2.4.2 Dimensionamiento del conmutador

En esta sección, ambos procedimientos de análisis SHWP y SCWP serán empleados para el dimensionamiento de conmutadores OPS con capacidad de emular colas a la salida. La evaluación se centrará en el caso de conmutador simétrico, con *N* fibras de entrada y salida, *n* longitudes de onda por fibra y *M* líneas de retardo.

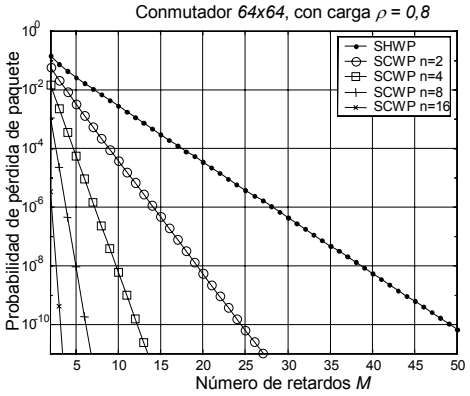
En un primer paso, la influencia del modo de operación se ha estudiado en conmutadores de tamaño *16x16*, *32x32* y *64x64*, donde el número de puertos viene determinado por el parámetro *nN*. La figura 2-11-(a,b,c), muestra la probabilidad de pérdida de paquete para el conmutador SHWP y SCWP en función del número de retardos *M,* para diferentes valores de *n,* y carga ofrecida ρ = *0.8*. La conmutación SHWP no obtiene ningún aprovechamiento del número de longitudes de onda *n*, por lo que las prestaciones de un conmutador de este tipo son independientes de este parámetro, para un tamaño de conmutador *nN* fijo. Por otro lado, se observa un fuerte impacto del parámetro *n* en las gráficas para los conmutadores SCWP, que claramente suponen una mejora frente a la versión SHWP. Como ejemplo, bajo el modo de operación SHWP, en un conmutador *16x16* con *4* fibras y *4* longitudes de onda por fibra, son necesarios *42* retardos para obtener una probabilidad de pérdida de paquete $<10^{-9}$, mientras sólo se requieren *11* en el modo de operación SCWP. Asimismo, la figura 2-11 muestra un leve incremento de la probabilidad de pérdida a medida que el tamaño del conmutador aumenta (en términos de número de fibras de entrada/salida). Para ambos modos de operación, la curva para el valor $N=\infty$ puede considerarse una aproximación pesimista precisa cuando *nN>32*.



(a)



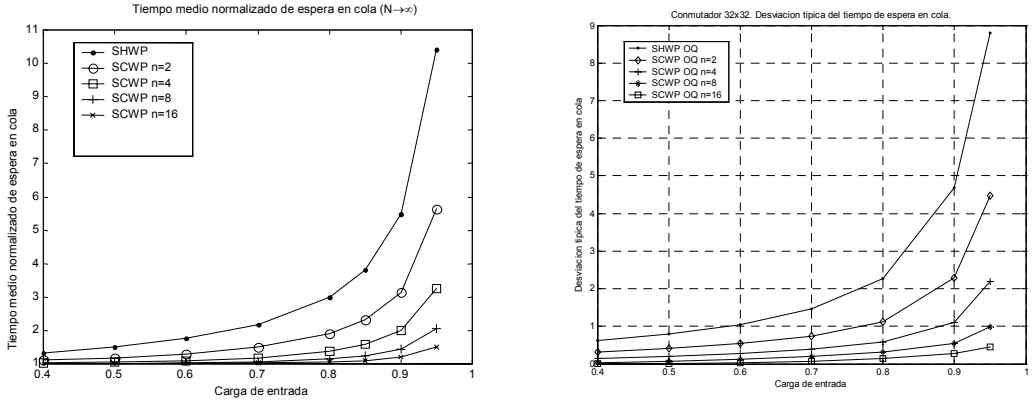
(b)



(c)

Figura 2-11. Probabilidad de pérdida de paquete respecto a número de retardos (*M*), para un conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida, carga ρ*=0.8*, *n={2,4,8,16,32}*, y tamaños de conmutador *nN* (a) *16x16*, (b) *32x32*, (c) *64x64*

En la figura 2-12-(a) se muestra la comparación del impacto de la carga de entrada en el retardo medio, de los modos de operación SHWP y SCWP, normalizado en número de ranuras temporales, asumiendo un tamaño infinito del conmutador, y una probabilidad de pérdida de paquete despreciable. De nuevo, se observa una fuerte mejora en el conmutador SCWP respecto a la versión SHWP. Esta mejora se hace más evidente para mayores valores de *n*, incluso en carga altas. La figura 2-12-(b) muestra la variación de la desviación típica de la variable aleatoria retardo (Ec. 2.25) en función de la carga en las mismas condiciones. Los resultados obtenidos muestran cómo la desviación típica se incrementa con la carga en todos los casos, aunque de manera mucho menor para valores altos de *n* en modo SCWP.



(a) (b)

Figura 2-12. Conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida (asumiendo *N*→∞), en función de la carga de entrada ρ, (a) tiempo medio normalizado de espera en cola, (b) desviación típica del tiempo medio normalizado de espera en cola.

En la tabla 2-1 se calcula el número de retardos *M* necesarios para obtener una probabilidad de pérdida de paquete < 10^{-9} con carga de entrada *0.8*, en conmutadores de distintos tamaños. La última columna muestra el impacto del parámetro *n*, en conmutadores de gran tamaño (*N*→∞), que sirven como una aproximación pesimista precisa. Como ejemplo, los valores obtenidos muestran un *50%* de ahorro en número

de retardos en el conmutador SCWP, para un número de longitudes de onda por fibra igual a *2*. Valores mayores de *n*, conllevan requisitos de buffer decrecientes, que confirman los beneficios de este modo de operación. Por ejemplo, es necesario un tamaño de buffer de sólo *5* posiciones en un conmutador *32x32* con *8* longitudes de onda por fibra, o de sólo *2* para un conmutador con *32* longitudes de onda por fibra.

		16x16	32x32	64x64	N→∞
	n				
SHWP	**---**	42	44	44	45
SCWP	**2**	19	22	22	23
SCWP	**4**	10	11	11	12
SCWP	**8**	4	5	6	7
SCWP	**16**	***	3	3	4
SCWP	**32**	***	***	2	3
SCWP	**64**	***	***	***	2

Tabla 2-1. Número de retardos necesarios para conmutadores OPS con colas a la salida SHWP y SCWP, para una probabilidad de pérdida de paquete <10^{-9}, *n={2,4,8,16,32,64}*, tamaños de conmutador *16x16, 32x32, 64x64*, y N→∞

2.4.3 Comparativa de costes entre arquitecturas

En esta sección, los requisitos de almacenamiento (*M*), descritos en el apartado 2.2, son empleados como base de una comparativa de coste *hardware* de los conmutadores seleccionados. Nuestro objetivo es comparar las distintas arquitecturas bajo ambos modos de operación, algo viable tras el proceso de adaptación seguido, y que contribuye a una comparativa global SCWP vs. SHWP.

En un primer paso, la tabla 2-2 resume el número de los distintos componentes para cada una de las arquitecturas, en función del número de fibras de entrada y de salida (*N*), número de longitudes de onda por fibra (*n*), y número de líneas de retardo (*M*). Los cálculos se realizan separadamente para cada uno de los componentes. Como se ha argumentado en el capítulo *1*, una función de coste total, combinando los costes de todos los componentes no tiene una utilidad clara en un escenario de tecnología tan cambiante. Posteriormente, nos fijaremos en el número de puertas ópticas, número y rango de sintonización de los dispositivos TWC, y en los kilómetros de fibra óptica necesarios, como factores limitantes fundamentales.

Por otro lado, existen fuertes limitaciones en las tres arquitecturas descritas en cuanto al tamaño máximo de los conmutadores, en número de puertos. En esta sección pondremos un límite al estudio de conmutadores de hasta *32x32* puertos. Como mostrarán los números, es incluso un límite lejano para el desarrollo comercial, observando el actual estado de integración de los procesos de fabricación, sobre todo de puertas ópticas. En el capítulo *4* se mostrarán distintas soluciones propuestas para el diseño de conmutadores de más alto número de puertos.

	FWC	Puertas ópticas	TWC (rango sintoniz. máx)	Delay loops	Tamaño máx. AWG
KEOPS switch	2nN	$MnN+n^2N^2$	0	1...M	0
OB-WR switch	nN	n^2N^2	nN (max(nN,M))	1...M	max (nN,M)
Space switch	0	nN^2M	nN (n)	N·(1...M)	0

Tabla 2-2. Cómputo de componentes *hardware*

Los valores en la Tabla 2-1 y en la Tabla 2-2 han sido empleados para rellenar la Tabla 2-3, que recoge el número de componentes en la construcción de las distintas arquitecturas para un tamaño de conmutador *32x32* (*nNxnN*), los modos SHWP y SCWP, y diversos valores del parámetro *n*. Para el conmutador OB-WR, el cálculo del número de puertas ópticas se realiza asumiendo una configuración de *crossbar* en la etapa de conmutación espacial. Se extraen las siguientes conclusiones:

- Para el conmutador KEOPS, el modo de operación impacta fuertemente en el número de puertas ópticas. Asumiendo un tamaño de conmutador constante, este valor muestra un incremento *lineal* con el factor *M*. Por esta razón, el fuerte decrecimiento (no lineal) observado en los requisitos de almacenamiento (*M*) en el modo de operación SCWP, es linealmente traducido a una simplificación en el *hardware*. Como ejemplo, la reducción en el número de puertas ópticas en el modo de operación SCWP frente a SHWP, alcanza el *50%* para *n=8*. La reducción en el número de kilómetros de fibra que supone el menor número de retardos es muy destacable también (por ejemplo, de *198* km a *1.2* km en *n=16*).
- En la arquitectura OB-WR, se obtienen mejoras menores en el tamaño de los dispositivos AWG y en los rangos de sintonización de los convertidores TWC bajo el modo de operación SCWP. Se destaca que el número de puertas ópticas (previsible factor crítico en el coste de esta arquitectura) no varía con la disminución del número de retardos. La variación de los kilómetros de fibra sigue los mismos parámetros que para el conmutador KEOPS.
- En el conmutador *space switch*, asumiendo un tamaño de conmutador *nN* fijo, el número de puertas ópticas y el número de retardos necesarios crece linealmente con *M/n*. En el modo de operación SHWP, el parámetro *M* es alto, y constante con *n*. El efecto combinado de reducción de *M* con el incremento de *n* en modo SCWP, tiene doble efecto en el factor *M/n*, apuntando a una mayor simplificación en el número de puertas ópticas (en el orden de *15* a *1* para *n=16*). La aplicación del modo de operación SHWP es prácticamente descartable, por el número de puertas ópticas necesarias, y la multiplicación en los kilómetros de fibra necesarios (por ejemplo, más de *3000* km para *n=2, N=16*).

Analizando los costes *hardware* de cada conmutador para el modo de operación SHWP, las arquitecturas KEOPS y OB-WR parecen los dos candidatos más prometedores. El conmutador OB-WR requiere la mitad de convertidores FWC, y un número menor de puertas ópticas que el conmutador KEOPS. Por otro lado, la arquitectura KEOPS no requiere de dispositivos sintonizables, ni encaminadores AWG. Además, permite de forma natural la transmisión de tráfico *multicast* (debido a su operación difusión-selección), y una posible aplicación de técnicas de priorización de tráfico.

La comparativa entre conmutadores en modo SCWP, muestra que la arquitectura *space switch* debe ser la alternativa preferida para valores altos de *n* (previsibles en una red troncal DWDM, *Dense Wavelength Division Multiplexing*).

- Es la arquitectura que mayor reducción en el número de puertas ópticas obtiene de la disminución del parámetro *M*.
- No requiere de convertidores FWC (se realiza una única conversión de longitud de onda, a la longitud de onda de transmisión final).
- El rango de sintonización requerido en los dispositivos TWC es siempre el menor posible, igual a *n*.

La posible desventaja de esta arquitectura estriba en la multiplicación por *N* de la longitud total de fibra necesaria, algo asumible en el caso de valores altos de *n* y por tanto bajo número de retardos. Hipotéticos requisitos en las redes troncales OPS, de funcionamiento *multicast* o -mucho más improbable en la opinión del autor de esta tesis doctoral- de técnicas de calidad de servicio, podrían favorecer a la arquitectura KEOPS frente la arquitectura *space switch*.

		KEOPS		OB-WR		SPACE SWITCH	
		SHWP	**SCWP**	**SHWP**	**SCWP**	**SHWP**	**SCWP**
n=2 ***N=16***	**FWCs**	64	64	32	32	0	0
	Puertas ópticas	2432	1728	1024	1024	22528	11264
	TWCs (rango máx.)	0 (0)	0 (0)	32 (44)	32 (32)	32 (2)	32 (2)
	Nº retardos	1..44	1..22	1..44	1..22	16·(1..44)	16·(1..22)
	km. fibra totales	198	50.6	198	50.6	3168	809.6
	Tamaño máx. AWG	0	0	44	32	0	0
n=4 ***N=8***	**FWCs**	64	64	32	32	0	0
	Puertas ópticas	2432	1376	1024	1024	11264	2816
	TWCs (rango máx.)	0 (0)	0 (0)	32 (44)	32 (32)	32 (4)	32 (4)
	Nº retardos	1..44	1..11	1..44	1..11	8·(1..44)	8·(1..11)
	km. fibra totales	198	13.2	198	13.2	1584	105.6
	Tamaño máx. AWG	0	0	44	32	0	0
n=8 ***N=4***	**FWCs**	64	64	32	32	0	0
	Puertas ópticas	2432	1216	1024	1024	5632	768
	TWCs (rango máx.)	0 (0)	0 (0)	32 (44)	32 (32)	32 (8)	32 (8)
	Nº retardos	1..44	1..6	1..44	1..6	4·(1..44)	4·(1..6)
	km. fibra totales	198	4.2	198	4.2	792	16.8
	Tamaño máx. AWG	0	0	44	32	0	0
n=16 ***N=2***	**FWCs**	64	64	32	32	0	0
	Puertas ópticas	2432	1120	1024	1024	2816	192
	TWCs (rango máx.)	0 (0)	0 (0)	32 (44)	32 (32)	32 (16)	32 (16)
	Nº retardos	1..44	1..3	1..44	1..3	2·(1..44)	2·(1..3)
	km. fibra totales	198	1.2	198	1.2	396	2.4
	Tamaño máx. AWG	0	0	44	32	0	0

Tabla 2-3. Evaluación de coste *hardware,* para las arquitecturas seleccionadas, para un tamaño de conmutador *32x32*, *n={2,4,8,16}*, aplicando los tamaños de buffer detallados en la tabla I. Para el cálculo de la longitud de fibra se considera un tamaño de paquete de *1* μs, y una velocidad de propagación de *200* m/μs

2.5 Conclusiones

En este capítulo se han evaluado las prestaciones de los conmutadores OPS de colas a la salida, para los modos de operación SHWP y SCWP. Los algoritmos de planificación empleados en ambos casos son óptimos en cuanto a las prestaciones del conmutador. Para el modo de operación SCWP se han propuesto dos variantes, la segunda de las cuales produce una selección uniforme de la longitud de onda de transmisión, manteniendo sus prestaciones óptimas, por lo que es preferida. Para la evaluación de la no uniformidad de asignación del algoritmo de planificación SCWP propuesto en [Pav03-2], se ha obtenido analíticamente la distribución de probabilidad de periodo ocupado de una cola multiservidor finita, para tráfico independiente e idénticamente distribuido (IID), igual a la distribución de periodo ocupado de un conmutador SCWP de colas a la salida. El método presentado tiene la ventaja de ser computacionalmente sencillo, y no requerir la resolución de la ecuación de estados de Markov.

Las prestaciones de las arquitecturas de colas a la salida han sido analizadas mediante un modelo de colas, y resueltas para procesos de llegadas IID. Esto ha proporcionado el mecanismo para el dimensionado del número de retardos necesarios en el caso de tráfico Bernouilli uniforme. Los valores obtenidos han servido, a su vez, de base para un estudio comparativo entre tres arquitecturas de conmutación con colas a la salida adaptadas: conmutador KEOPS, conmutador OB-WR y conmutador *space switch*. La conclusión fundamental de este estudio ha sido el fuerte ahorro en número de componentes que se obtiene con el modo de operación SCWP, mejorando para valores altos del número de longitudes de onda de transmisión (escenario DWDM, *Dense Wavelength Division Multiplexing*). Ante una situación de este tipo, atendiendo a los resultados obtenidos para estas tres arquitecturas de conmutación, el modo de operación SCWP aparece como claramente ventajoso.

Capítulo 3. Arquitectura *Input-Buffered Wavelength-Routed Switch*

3.1 Introducción

En este capítulo nos centraremos en el estudio de la arquitectura de conmutación OPS *Input-Buffered Wavelength-Routed Switch,* propuesta en [Zho98]. El elemento distintivo de esta arquitectura es la ausencia de puertas ópticas en su diseño, y un coste *hardware* comparativamente menor a otras alternativas.

La arquitectura será adaptada al entorno WDM, y se estudiará su comportamiento en los modos de operación SHWP y SCWP. El problema de asignación de retardo será modelado formalmente como un problema de programación dinámica entera. Siguiendo lo publicado en [Pav03-1], el problema en cada ranura temporal será caracterizado como un problema de maximización en grafos bipartitos. Esto requiere el diseño de planificadores que exploren de manera eficiente en el conjunto de soluciones posible. En este sentido, se propondrá el algoritmo de planificación PDBM (*Parallel Desynchronized Block Matching*) para el modo de operación SCWP, que admite una implementación viable a las velocidades requeridas. Sus prestaciones serán comparadas con las de un planificador secuencial SCWP (evolución trivial del planificador propuesto en [Zho98], y no implementable a las velocidades demandadas), y con las prestaciones óptimas alcanzables, determinadas por las arquitecturas de colas a la salida.

3.2 Descripción de la arquitectura

3.2.1 Trabajo previo

La arquitectura original *Input-Buffered Wavelength-Routed Switch* (IB-WR), tal y como fue propuesta en [Zho98], se muestra en la figura 3-1-(a) para un conmutador con *N* puertos de entrada y *N* puertos de salida no WDM. Se compone de la interconexión de una sección de almacenamiento (*buffering section*) semejante a la empleada en la arquitectura OB-WR, y una etapa de conmutación (*switching section*), ambas basadas en conversores de longitud de onda sintonizables TWC y encaminadores AWG (nótese la ausencia de puertas ópticas en esta arquitectura).

Los paquetes entrantes por cada puerto son convertidos a un longitud de onda que determina el retardo *0...M-1* que sufrirán en la sección de almacenamiento (*buffering section*). Debido al modo cíclico de operación de los dispositivos AWG, el puerto de salida del módulo de almacenamiento es siempre igual al puerto de entrada, independientemente de la longitud de onda de conversión. Esta longitud de onda se emplea para determinar el retardo atravesado siguiendo la regla λ*=(puerto E/S + retardo) mod K*, donde *K=max(N,M)* es igual al tamaño necesario de los dispositivos AWG, e igual al rango de sintonización necesario de los conversores TWC.

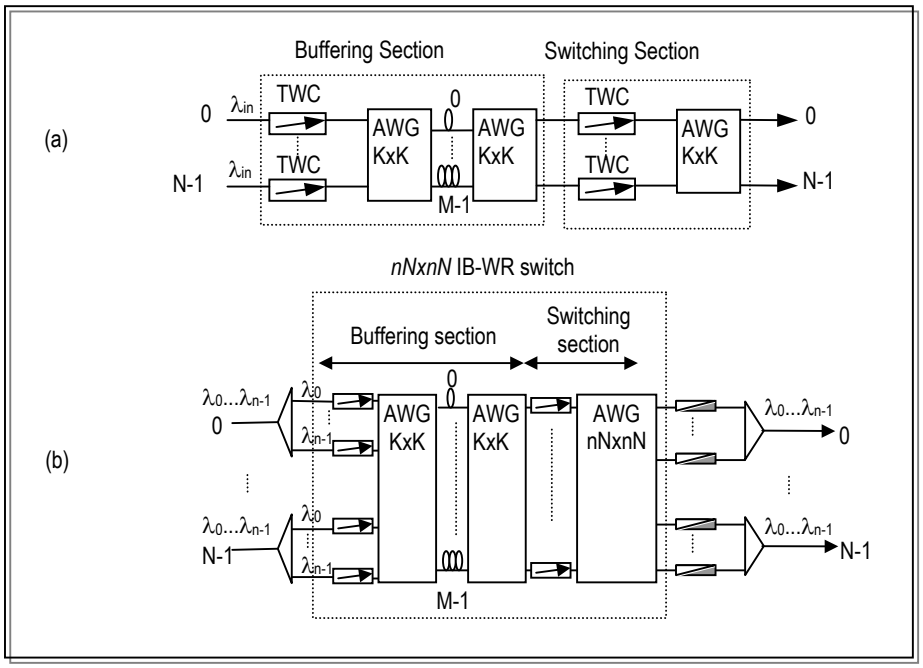


Figura 3-1. Arquitectura IB-WR, (a) diseño original, (b) adaptación WDM

La sección de conmutación es de tamaño *NxN,* y está formada por *N* dispositivos TWC y un encaminador AWG. Los paquetes sufren una conversión de longitud de onda de salida que determina el puerto de salida por el que serán transmitidos.

En una ranura temporal, a lo sumo un paquete puede llegar a cada uno de los *N* puertos de entrada de la etapa de conmutación, ya que los dispositivos TWC pueden operar con un único paquete por ranura temporal. Esto ha de ser tenido en cuenta por el planificador del conmutador en la asignación de retardos a los paquetes entrantes. El proceso de planificación propuesto en [Zho98] se basa en la expresión del estado de conmutador mediante dos conjuntos de vectores de dimensión *M*.

- La familia de vectores X_i, *i=0...N-1*, con un vector para cada puerto de entrada, se asocia a la ocupación de los *N* dispositivos TWC de la etapa de conmutación. La coordenada *b* de un vector X_i (=$X_i[b]$, *b=0,..,M-1*) tendrá un valor *1* si un paquete entrante por el puerto *i* ocupa el TWC *i* de la sección de conmutación en la ranura temporal *b*. $X_i[b]$ tendrá el valor *0* en caso contrario.
- La familia de vectores Y_j, *j=0...N-1*, con un vector para mostrar la ocupación de cada puerto de salida. La coordenada *b* de un vector Y_j (=$Y_j[b]$, *b=0,..,M-1*) tendrá un valor *1* si un paquete es transmitido por el puerto de salida *j*, en la ranura temporal *b*. $Y_j[b]$ tendrá el valor *0* en caso contrario.

Para la planificación de un paquete arbitrario, proveniente del puerto de entrada *i,* destinado al puerto de salida *j*, debe inspeccionarse el contenido de los vectores X_i e Y_j. Un retardo *b* es asignable al paquete entrante, en el caso de que el TWC *i*-ésimo de la etapa de conmutación, y el puerto de salida *j*-ésimo, se encuentren libres dentro de *b* ranuras temporales, o lo que es lo mismo $X_i[b]=Y_j[b]=0$. Un paquete será descartado en el caso de que esta doble condición no se cumpla para ninguno de los *M* retardos. En la figura 3-2, un paquete entrante por la entrada *2* demandando el

puerto de salida *5* puede ser únicamente transmitido a través de las líneas de retardo *2* ó *4*.

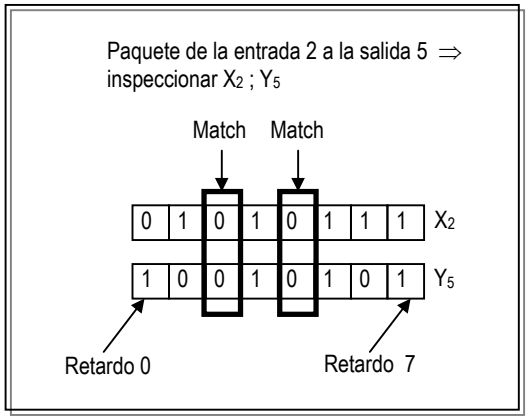


Figura 3-2. Ejemplo de planificación en la arquitectura original IB-WR

En el trabajo original [Zho98], se propuso que el problema de planificación a resolver era semejante al problema de asignación de ranura temporal de un dispositivo TSI (*time-slot-interchange*) [Hui91], o a la elección de etapa intermedia en una red de interconexión telefónica basada en una red de Clos de *3* etapas. Como se mostrará en este capítulo, esta semejanza no es oportuna, ya que no maneja la correlación entre los vectores de estado del sistema en ranuras temporales sucesivas.

Para la planificación del conmutador, se propuso un algoritmo simple, que visita secuencialmente los puertos de entrada, y asigna el retardo más pequeño disponible para cada paquete entrante. Este algoritmo se reproduce en la figura 3-3.

```
/* N = n° de puertos entrada/salida */
/* M = n° de retardos */

for input i = 0 to N-1 do
  if (packet is present on input i) then
    j = output port required by input packet i
    for timeslot t = 0 to M-1 do
      if (t is idle in Xi and Yj) then
        assign delay t to packet
        allocate time slot t in Xi and Yj: Xi[t]=Yj[t]=1
        break
      endif
    endfor
  endif
endfor


/* Actualizar valores de Xi, Yj tras cada ranura temporal */

for i=0 to N-1 do
  shift Xi vector to the left. Xi[M-1] = 0
  shift Yi vector to the left. Yi[M-1] = 0
endfor
```

Figura 3-3. Algoritmo secuencial de planificación para la arquitectura IB-WR [Zho98]

3.2.2 Adaptación WDM

El conmutador IB-WR original no fue diseñado para escenarios con fibras de entrada y salida WDM. Es de nuevo necesario realizar el paso *1* del proceso de adaptación, previo a la aplicación de los modos de operación SHWP y SCWP. La figura 3-1-(b) muestra los cambios propuestos en la arquitectura:

- Se ha añadido una etapa de demultiplexación WDM en las fibras de entrada, que separan los paquetes entrantes por su longitud de onda, en *n* fibras distintas.
- Las *nN* fibras resultantes son conectadas a un conmutador IB-WR de tamaño *nNxnN*.
- En el conmutador original, la longitud de onda de salida de un paquete viene determinada por el puerto de salida del mismo. Para fijar la longitud de onda de transmisión $\lambda_0,...,\lambda_{n-1}$, cada uno de los *nN* puertos de salida del conmutador están conectados a un convertidor a longitud de onda fija, y posteriormente a un multiplexor para cada fibra de salida. De esta forma, la decisión sobre la fibra y longitud de onda de transmisión de un paquete se traduce en una decisión sobre su puerto de salida del conmutador IB-WR subyacente, de manera similar a lo descrito para el conmutador OB-WR:

  - En el modo de operación SHWP, el OPP al que pertenece un paquete establece su fibra y longitud de onda de transmisión, y por tanto, el puerto de salida del conmutador IB-WR subyacente.

  - Para el modo SCWP, la libertad en la selección de longitud de onda de salida se traduce en una libertad para conmutar el paquete hacia uno de los *n* puertos del conmutador IB-WR, asociados a la fibra de salida de destino.

3.2.3 Planificación del conmutador

En esta sección se analiza formalmente la planificación de los conmutadores IB-WR adaptados bajo los modos de operación SHWP y SCWP. En ambos casos, esta planificación se caracterizará como un problema de programación dinámica (y por tanto de dimensión infinita). Posteriormente este problema se simplificará a un problema de programación (optimización) finita, que se mostrará equivalente a un problema emparejamiento en grafos bipartitos, siguiendo la propuesta publicada en [Pav03-1]. Esta simplificación marca un camino a seguir para la obtención de algoritmos eficientes de planificación, que desembocará en la propuesta del algoritmo PDBM (*Parallel Desynchronized Block Matching*) para el modo de operación SCWP.

3.2.3.1 Planificación SHWP

3.2.3.1.1 Expresión como un problema de optimización dinámica

En la versión SHWP de un conmutador IB-WR adaptado como el que muestra la figura 3-1-(b), el puerto de salida de un paquete está determinado por el circuito virtual OPP al que pertenece. En estas condiciones, la planificación del conmutador IB-WR original es equivalente a la planificación IB-WR SHWP del conmutador adaptado.

En esta tesis doctoral se define un planificador SHWP para la arquitectura IB-WR, como un dispositivo que resuelve el problema de optimización dinámica expresado en (Ec. 3.1).

$$\text{Obtener } A_{ibk}^{(T)} \text{ que maximice } \left\{ \sum_{T,i,b,j} A_{ibk}^{(T)} \right\} \text{ sujeto a las restricciones}$$

$$(1) \sum_{b=0}^{M-1} A_{ibj}^{(T)} \le R_{ij}^{(T)}, \forall i = 0,...,nN-1, \forall j = 0,...,nN-1, \forall T = 0,...$$

$$(2) \sum_{i=0}^{nN-1} A_{ibj}^{(T)} + E_{ibk}^{(T)} \le 1, \forall b = 0,...,M-1, \forall j = 0,...,nN-1, \forall T = 0,...$$

$$(3) \sum_{j=0}^{nN-1} A_{ibj}^{(T)} + E_{ibk}^{(T)} \le 1, \forall i = 0,...,nN-1, \forall b = 0,...,M-1, \forall T = 0,...$$

$$(4) E_{ibj}^{(0)} = 0, \forall i = 0,...,nN-1, \forall b = 0,...,M-1, \forall j = 0,...,nN-1$$

$$(5) E_{ibj}^{(T+1)} = E_{i(b+1)j}^{(T)} + X_{i(b+1)j}^{(T)}, \forall i = 0,...,nN-1, \forall b = 0,...,M-2, \forall j = 0,...,nN-1, \forall T = 0,...$$

$$E_{i(M-1)j}^{(T+1)} = 0, \forall i = 0,...,nN-1, \forall j = 0,...,nN-1, \forall T = 0,...$$

$$A_{ibj}^{(T)} = \begin{cases} 1 \text{ si a un paquete entrante en tiempo } T, \text{ por el puerto } i, \text{ destinado al puerto } j, \\ \text{se le asigna el retardo } b;\ i,j = 0,..,nN-1, b = 0,...,M-1, T = 0,... \\ 0 \text{ en caso contrario} \end{cases}$$

$$R_{ij}^{(T)} = \begin{cases} 1 \text{ si un paquete es recibido en tiempo } T, \text{ por el puerto de entrada } i, \text{ destinado} \\ \text{al puerto } j; i,j = 0,...,nN-1 \\ 0 \text{ en caso contrario} \end{cases}$$

$$E_{ibj}^{(T)} = \begin{cases} 1 \text{ si al comienzo de la ranura temporal } T, \text{ el retardo } b \text{ para el puerto de entrada } i, \\ \text{está ocupado por un paquete destinado al puerto } j; i,j = 0,...,nN-1 \\ 0 \text{ en caso contrario} \end{cases} \quad \text{(Ec. 3.1)}$$

- La función objetivo pretende maximizar el número de paquetes entrantes a los que se les asigna un retardo, en todas las ranuras temporales. Los paquetes entrantes a los que no se les asigna retardo, son paquetes descartados. Por ello, esto equivale a maximizar el *throughput* medio del conmutador a lo largo de las *T=0,...* ranuras temporales.
- La familia (1) de $(nN)^2$ restricciones en cada ranura temporal, indica que es posible asignar un retardo a un paquete en un puerto de entrada *i*, destinado a un puerto de salida *j*, si efectivamente existe la llegada de un paquete con esas condiciones. Esto viene indicado por el proceso estadístico de llegadas, con la condición $R_{ij}^{(T)} = 1$.
- La familia (2) de *nNM* restricciones en cada ranura temporal, describe la contención por cada puerto de salida, asociada al concepto de conmutación de paquetes: en cada ranura temporal, cada puerto de salida puede transmitir un máximo de un paquete.
- La familia (3) de *nNM* restricciones en cada ranura temporal, describe la contención por cada dispositivo TWC de la etapa de conmutación, peculiar a esta arquitectura: puede llegar un máximo de un paquete en cada ranura temporal a cada uno de estos dispositivos.
- Las familias de restricciones (4) y (5) describen la relación existente entre ranuras temporales sucesivas, a través de la variable auxiliar $E_{ibj}^{(T)}$ que simboliza el estado del sistema al comienzo de la ranura temporal *T*. La familia (4) indica que el conmutador se encuentra vacío en la ranura temporal *T=0*. La familia (5) expresa la relación entre las decisiones de planificación en el instante *T* ($A_{ibj}^{(T)}$), el estado del sistema en ese instante ($E_{ibj}^{(T)}$), y el estado

del sistema en un instante posterior ($E_{ibj}^{(T+1)}$), donde todos los paquetes almacenados han avanzado una ranura temporal.

La formulación completa del problema muestra las diferencias existentes con el problema de asignación en conmutadores TSI (según se indica en [Zho98]), donde no se tiene en cuenta la evolución peculiar en ranuras temporales sucesivas, característica de las líneas de retardo.

3.2.3.1.2 Expresión como un problema de optimización entera finita

La programación dinámica [Inf91] es una herramienta que nos permite describir de manera elegante ciertos problemas como los de planificación de conmutadores de paquetes, definir el concepto de planificador como un algoritmo que proporciona una solución a este problema, e ilustrar claramente mediante restricciones las limitaciones que el *hardware* de la arquitectura implica. Por ejemplo, eliminando la familia de restricciones (3), obtenemos directamente la descripción del problema de planificación de los conmutadores con colas a la salida. Sin embargo, no existen algoritmos de tipo general para la resolución de esta clase de problemas, sino que cada problema debe tener un tratamiento especial [Inf91]. La complicación se incrementa teniendo en cuenta que nos hallamos ante un problema de programación dinámica no determinista (llamada *estocástica*), debido a la componente probabilística del proceso de llegadas $R_{ij}^{(T)}$.

Nuestro objetivo a continuación, será llegar a una expresión del problema más simple, que incluya las características inherentes a esta arquitectura, pero que nos guíe en el camino de diseñar algoritmos de planificación que ofrezcan buenas prestaciones. Estas prestaciones del planificador serán lógicamente sub-óptimas. No se aborda la búsqueda de un algoritmo óptimo, que por otro lado sería presumiblemente distinto para distintos patrones de llegadas $R_{ij}^{(T)}$.

La simplificación elegida tiene las siguientes características:

1. Enfocaremos el problema en cada ranura temporal *T* como una maximización del *throughput instantáneo* del sistema, lo que equivale a una minimización en cada ranura temporal del número de paquetes descartados de entre los paquetes entrantes. El problema se transforma en un problema de optimización entera de número finito de variables. Nótese que la optimización del caudal instantáneo en cada ranura temporal no implica la optimización del caudal medio.
2. En cada ranura temporal *T*, de entre las posibles planificaciones que proporcionen el mismo máximo de paquetes asignados (mínimo de paquetes descartados), seleccionaremos aquella que minimice el *retardo medio* de los paquetes asignados. El objetivo de este criterio es optar por la planificación que intente minimizar la ocupación de recursos en ranuras temporales posteriores. Como se desprende de la familia de restricciones (5) en (Ec. 3.1), la asignación de un retardo *b* para un paquete entrante por el puerto *i*, y destinado al puerto de salida *j*, implica la ocupación de recursos en *b+1* ranuras temporales sucesivas:

- $X_i[b]=Y_j[b]=1$, en esta ranura temporal *T*.
- $X_i[b-1]=Y_j[b-1]=1$, en la siguiente ranura temporal *T+1*.
- ...

- $X_i[0]=Y_j[0]=1$, en la ranura temporal $T+b$.

La ecuación 3.2 detalla el problema de planificación simplificado en que se centra nuestro estudio.

$$
\begin{aligned}
&\textit{En cada ranura temporal } T\text{, dado un estado del sistema}\\
&E_{ibj}=\begin{cases}1 \textit{ el retardo } b \textit{ para el puerto de entrada } i,\\ \textit{está ocupado por un paquete destinado al puerto } j; i,j=0...nN-1\\ 0 \textit{ en caso contrario}\end{cases}\\
&\textit{que determinan los vectores de estado}\\
&X_i[b]=\sum_{j=0}^{nN-1}E_{ibj}\in\{0,1\},\ \textit{ocupación de puerto de entrada i de}\\
&\qquad\qquad \textit{la etapa de conmutación}, i=0...nN-1\\
&Y_j[b]=\sum_{i=0}^{nN-1}E_{ibj}\in\{0,1\},\ \textit{ocupación de puerto de salida } j=0...nN-1\\
&\textit{obtener } X_{ibj} \textit{ que}\begin{cases}\textit{entre las soluciones que maximicen el throughput}\left\{\sum_{i,b,j}X_{ibj}\right\}\\ \textit{sea la que minimice el retardo medio asignado}\left\{\sum_{i,b,j}b\cdot X_{ibj}\right\}\end{cases}\\
&\textit{sujeto a las restricciones}\\
&(1)\sum_{b=0}^{M-1}A_{ibj}\le R_{ij},\forall i=0,...,nN-1,\forall j=0,...,nN-1\\
&(2)\sum_{i=0}^{nN-1}A_{ibj}+E_{ibj}\le 1,\forall b=0,...,M-1,\forall j=0,...,nN-1\\
&(3)\sum_{j=0}^{nN-1}A_{ibj}+E_{ibj}\le 1,\forall i=0,...,nN-1,\forall b=0,...,M-1\\
&A_{ibj}^{(T)}=\begin{cases}1 \textit{ si a un paquete entrante por el puerto } i, \textit{ destinado al puerto } j,\\ \textit{se le asigna el retardo } b;\ i,j=0,..,nN-1, b=0,...,M-1\\ 0 \textit{ en caso contrario}\end{cases}\\
&R_{ij}^{(T)}=\begin{cases}1 \textit{ si un paquete es recibido por el puerto de entrada } i, \textit{ destinado}\\ \textit{al puerto } j; i,j=0,...,nN-1\\ 0 \textit{ en caso contrario}\end{cases}
\end{aligned}
\qquad \text{(Ec. 3.2)}
$$

3.2.3.1.3 Expresión como un problema de emparejamiento máximo en grafos bipartitos

En esta sección mostraremos cómo el problema de optimización descrito en la Ec. 3.2 puede hacerse equivalente a un conjunto de problemas de Emparejamiento en Grafos Bipartitos. Resulta conveniente por tanto formalizar las siguientes definiciones:

- **Grafo**. Un grafo *G=(N,A)* consiste en un conjunto *N* de nodos y un conjunto *A* de arcos, donde un arco viene indicado por el par de nodos que relaciona *(i,j), i,j∈N*. En nuestro caso, nos referiremos a grafos no dirigidos, donde los arcos *(i,j)* y *(j,i)* son equivalentes.

- **Grafo bipartito.** Un grafo $G=(N,A)$ se dice bipartito cuando existe una partición de los nodos en dos conjuntos $N=N_1 \cup N_2$, $N_1 \cap N_2=\varnothing$, tal que no existen arcos entre elementos del mismo conjunto.
- **Emparejamiento en un grafo bipartito**. Un emparejamiento en un grafo bipartito $G=(N_1 \cup N_2,A)$ es un grafo bipartito subconjunto $G'=(N_1 \cup N_2,A')$, $A' \subset A$, tal que

  (1) de cada nodo de N_1 salga a lo sumo un arco, y

  (2) de cada nodo de N_2 llegue a lo sumo un arco.

La caracterización propuesta se basa en las siguientes consideraciones:

- Dentro de una ranura temporal T, la asignación de retardos a paquetes destinados al mismo puerto de salida j, se encuentran mutuamente afectados por la inspección del vector Y_j.
- En consecuencia, una asignación óptima de retardos (en el sentido de la Ec. 3.2) a estos paquetes de destino común, debe realizarse de manera global.
- Dentro de una ranura temporal T, la asignación de retardos a paquetes destinados a puertos de salida diferentes, son independientes.

El último punto nos permite dividir y paralelizar el problema en nN problemas independientes, uno para cada puerto de salida del conmutador. Cada uno de estos problemas individuales puede ser expresado como un grafo bipartito. Para un puerto de salida j, su grafo bipartito es tal que:

(1) Contiene nN nodos del lado izquierdo del grafo (N_1), uno para cada puerto de entrada del conmutador.

(2) Contiene M nodos del lado derecho del grafo (N_2), uno para cada retardo del conmutador.

(3) Existe un arco entre el puerto de entrada i, y el retardo b, en el caso de que

- en el puerto de entrada i exista un paquete destinado al puerto de salida j ($R_{ij}=1$)
- y que el retardo b sea un retardo elegible para ser asignado a ese paquete, $X_i[b]=Y_j[b]=0$.

Bajo estas condiciones, un emparejamiento es una solución válida al problema de planificación, ya que a cada paquete de entrada se le asignará a lo sumo un retardo, y a cada retardo se le asignará a lo sumo un paquete de entrada.

Como consecuencia, la función objetivo indicada en la Ec. 3.2, se traduce en encontrar un emparejamiento en cada uno de los nN grafos bipartitos asociados, tal que cada uno de ellos:

- Sea una solución de tamaño máximo posible al problema de emparejamiento (maximice el *throughput*). Este tipo de problemas de emparejamiento máximo en grafos bipartitos se conocen como MSM (*Maximum Size Matching*).

- En el caso de existir varias soluciones en un grafo de tamaño igual al MSM, debe escogerse aquella con la media de los retardos asignados menor.

Para ilustrar mejor el problema, en la figura 3-4 se muestra un ejemplo de aplicación en un conmutador IB-WR SHWP con *4* puertos de entrada y *4* puertos de salida, *M=3* retardos d_0,d_1,d_2. En una determinada ranura temporal *T*, el estado del sistema es el mostrado por los vectores X_i, *i=0,...,3,* Y_j, *j=0,...,3*. Las llegadas al sistema en esa ranura temporal son las indicadas por R_{ij}. Los puertos de entrada *0, 1 y 3,* reciben paquetes destinados al puerto de salida *0*. El puerto *2* recibe un paquete destinado a la salida *2*. El estado del sistema y el proceso de llegadas producen los *4* grafos bipartitos mostrados en la figura 3-4-(b).

La aplicación de un algoritmo de planificación secuencial como el propuesto en [Zho98], comenzando por el puerto de entrada *0*, produce en este caso un emparejamiento de menor tamaño al MSM en el puerto de salida *0*, donde el paquete entrante por el puerto *3* es descartado, como se muestra en la figuras 3-4-(c).

La aplicación de un algoritmo MSM produciría un emparejamiento como el de la figura 3-4-(d), donde el paquete entrante por el puerto *3* no es descartado.

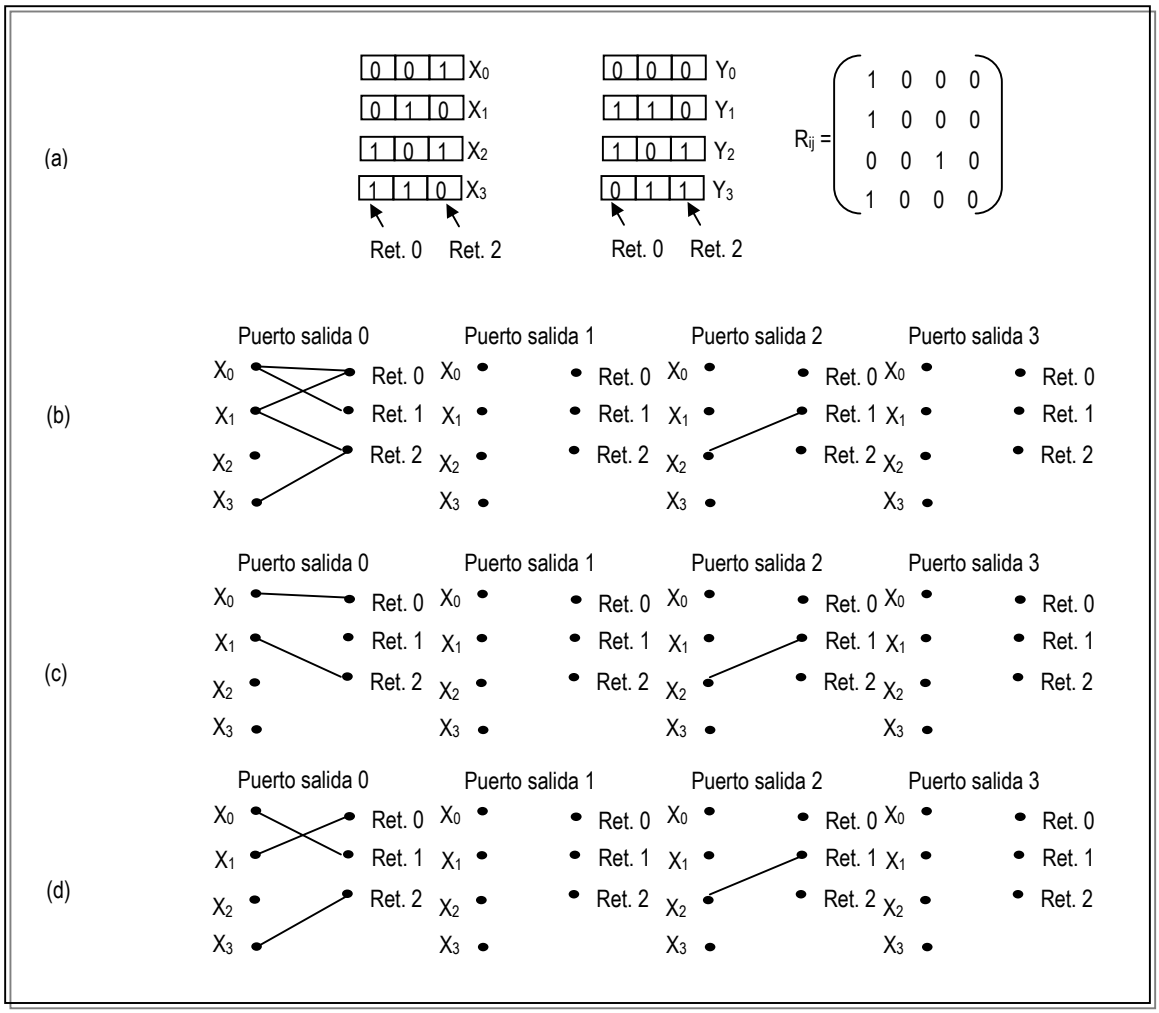


Figura 3-4. Ejemplo de emparejamiento planificación SHWP, (a) estado del sistema y proceso de llegadas, (b) planteamiento del problema como un conjunto de grafos bipartitos, (c) planificación por algoritmo secuencial [Zho98], (d) planificación por algoritmo MSM

Nota: El problema de planificación SHWP se puede plantear de manera equivalente como un único grafo bipartito, con *nN* nodos del lado izquierdo (uno para cada puerto de entrada), y *nNM* nodos del lado derecho (*M* por cada puerto de salida). Se ha aprovechado en esta sección la condición de poder ser descompuestos en *nN* grafos aislados (uno por cada puerto de salida), cuyo problema de emparejamiento puede resolverse en paralelo.

3.2.3.2 Planificación SCWP

3.2.3.2.1 Expresión como un problema de optimización dinámica

En la versión SCWP de un conmutador IB-WR adaptado como el que muestra la figura 3-1-(b), la fibra de salida de un paquete está determinada por el OPP al que pertenece. Sin embargo, el planificador es libre de seleccionar la longitud de onda de transmisión. Esto se traduce en que el planificador tiene libertad para seleccionar uno de los *n* puertos de salida asociados a la fibra de salida demandada. Las implicaciones de esta libertad de selección en el planificador del conmutador no han sido contempladas en la propuesta original [Zho98].

En esta tesis doctoral, definiremos un planificador SCWP para la arquitectura IB-WR, como un dispositivo que resuelve el problema de optimización dinámica expresado en la Ec. 3.3.

$$
\begin{aligned}
&\textit{Obtener } A_{ibk}^{(T)} \textit{ que maximice } \left\{ \sum_{T,i,b,j} A_{ibk}^{(T)} \right\} \textit{ sujeto a las restricciones} \\
&(1) \sum_{b=0}^{M-1} A_{ibj}^{(T)} \le R_{ij}^{(T)}, \forall i = 0,...,nN-1, \forall j = 0,...,N-1, \forall T = 0,... \\
&(2) \sum_{i=0}^{nN-1} A_{ibj}^{(T)} + E_{ibk}^{(T)} \le n, \forall b = 0,...,M-1, \forall j = 0,...,N-1, \forall T = 0,... \\
&(3) \sum_{j=0}^{N-1} A_{ibj}^{(T)} + E_{ibk}^{(T)} \le 1, \forall i = 0,...,nN-1, \forall b = 0,...,M-1, \forall T = 0,... \\
&(4) E_{ibj}^{(0)} = 0, \forall i = 0,...,nN-1, \forall b = 0,...,M-1, \forall j = 0,...,N-1 \\
&(5) E_{ibj}^{(T+1)} = E_{i(b+1)j}^{(T)} + X_{i(b+1)j}^{(T)}, \forall i = 0,...,nN-1, \forall b = 0,...,M-2, \forall j = 0,...,N-1, \forall T = 0,... \\
&\quad E_{i(M-1)j}^{(T+1)} = 0, \forall i = 0,...,nN-1, \forall j = 0,...,N-1, \forall T = 0,... \\
&A_{ibj}^{(T)} = \begin{cases} 1 \textit{ si a un paquete entrante en tiempo } T, \textit{ por el puerto } i, \textit{ destinado a la fibra } j, \\ \textit{se le asigna el retardo } b;\ i,j = 0,..,nN-1, b = 0,...,M-1, T = 0,... \\ 0 \textit{ en caso contrario} \end{cases} \\
&R_{ij}^{(T)} = \begin{cases} 1 \textit{ si un paquete es recibido en tiempo } T, \textit{ por el puerto de entrada } i, \textit{ destinado} \\ \textit{a la fibra } j; i,j = 0,...,nN-1 \\ 0 \textit{ en caso contrario} \end{cases} \\
&E_{ibj}^{(T)} = \begin{cases} 1 \textit{ si al comienzo de la ranura temporal } T, \textit{ el retardo } b \textit{ para el puerto de entrada } i, \\ \textit{está ocupado por un paquete destinado a la fibra } j; i,j = 0,...,nN-1 \\ 0 \textit{ en caso contrario} \end{cases}
\end{aligned}
\qquad \text{(Ec. 3.3)}
$$

- De manera similar a la planificación SHWP, la función objetivo pretende maximizar el número de paquetes entrantes a los que se les asigna un retardo, en todas las ranuras temporales. Los paquetes entrantes a los que no se les asigna retardo, son paquetes descartados. Por ello, esto equivale a

maximizar el *throughput* medio del conmutador a lo largo de las *T=0,...* ranuras temporales.

- De manera similar a la planificación SHWP, la familia (1) de nN^2 restricciones en cada ranura temporal, indica que es posible asignar un retardo a un paquete en un puerto de entrada *i*, destinado a una fibra *j*, si efectivamente existe la llegada de un paquete con esas condiciones. Esto viene indicado por el proceso estadístico de llegadas, con la condición $R_{ij}^{(T)} = 1$.
- La familia (2) de *NM* restricciones en cada ranura temporal, describe la contención por cada fibra de salida, peculiar a la conmutación óptica de paquetes en modo SCWP: en cada ranura temporal, cada fibra de salida puede transmitir hasta *n* paquetes. Esta familia de restricciones constituye la diferencia fundamental respecto a la planificación SHWP.
- De manera similar a la planificación SHWP, la familia (3) de *nNM* restricciones en cada ranura temporal, describe la contención por cada dispositivo TWC de la etapa de conmutación, peculiar a esta arquitectura: a cada uno de estos dispositivos puede llegar un máximo de un paquete en cada ranura temporal. Eliminando esta familia de restricciones, se obtiene directamente la descripción del problema de planificación de los conmutadores SCWP de colas a la salida.
- De forma similar a la planificación SHWP, las familias de restricciones (4) y (5) describen la correlación existente entre ranuras temporales sucesivas, a través de la variable auxiliar $E_{ibj}^{(T)}$ que simboliza el estado del sistema al comienzo de la ranura temporal *T*. La familia (4) indica que el conmutador se encuentra vacío en la ranura temporal *T=0*. La familia (5) expresa la relación entre las decisiones de planificación en el instante *T* ($A_{ibj}^{(T)}$), el estado del sistema en ese instante ($E_{ibj}^{(T)}$), y el estado del sistema en un instante posterior ($E_{ibj}^{(T+1)}$), donde todos los paquetes almacenados han avanzado una ranura temporal.

En resumen, las diferencias respecto a la planificación SHWP provienen de la sustitución del concepto de puerto de salida *j=0,...,nN-1*, por el de fibra de salida *j=0,...,N-1,* que afecta a las variables $R_{ij}^{(T)}$, $E_{ibj}^{(T)}$ y $X_{ibj}^{(T)}$, y de la familia de restricciones (2).

3.2.3.2.2 Selección de longitud de onda de transmisión

Nótese que la formulación del problema de planificación SCWP atañe únicamente a la asignación de retardos. No se refleja la selección de la longitud de onda de transmisión de los paquetes salientes. Esto es así ya que, en esta arquitectura, la selección de longitud de onda de transmisión es un aspecto independiente, que atañe a la etapa de conmutación. Los criterios para la distribución de las *n* longitudes de onda de transmisión para cada fibra de salida que realiza la etapa de conmutación no son valorados en esta tesis doctoral. Se hacen simplemente las siguientes consideraciones:

- Esta elección puede implementar un esquema que permita mantener el orden entre paquetes transmitidos *simultáneamente*. No se estudiarán los problemas de desorden en esta arquitectura de conmutación.

- Esta elección influye en las llegadas de paquetes en los nodos de conmutación vecinos hacia los que se dirige nuestro tráfico saliente:
  - En el caso de tratarse de un nodo con capacidad de emular colas a la salida, la distribución de longitudes de onda en cada fibra de entrada es irrelevante a efectos de prestaciones, salvo para efectos de desorden, que requieren un criterio común entre nodos como el expresado en el capítulo *2*.
  - En el caso de tratarse de un nodo IB-WR, este proceso de llegadas sí es relevante para las prestaciones del nodo. Esto se ve reflejado en que la longitud de onda de entrada, determina el vector X_i a emplear como restricción. Se plantea una cuestión interesante, que es decidir, en este caso, cuál es la asignación de longitudes de onda de transmisión de un nodo, que beneficiaría más a las prestaciones del nodo posterior IB-WR. Parece claro que un principio a seguir, es intentar distribuir el tráfico entre todas las longitudes de onda de manera equilibrada, para "ocupar" de manera más uniforme los vectores X_i de los puertos de entrada. En este sentido, resulta más favorable un reparto lo más equilibrado de las longitudes de onda de transmisión, al estilo del algoritmo de planificación SCWP uniforme descrito en el capítulo *2*.

3.2.3.2.3 Expresión como un problema de optimización entera finita

A continuación, nuestro objetivo será, al igual que en la planificación SHWP, llegar a una expresión del problema más simple, abandonando la programación dinámica. Esto se hará aplicando exactamente los mismos criterios de simplificación que los indicados para la planificación SHWP:

1. Enfocaremos el problema en cada ranura temporal *T* como una maximización del *throughput instantáneo* del sistema, obteniendo un problema de optimización entera de número finito de variables. De nuevo, nótese que la optimización del caudal instantáneo en cada ranura temporal no implica la optimización del caudal medio.
2. En cada ranura temporal *T*, de entre las posibles planificaciones que proporcionen el mismo máximo de paquetes asignados (mínimo de paquetes descartados), seleccionaremos aquélla que minimice el *retardo medio* de los paquetes. De nuevo, este objetivo pretende optar por la planificación que minimice la ocupación de recursos en ranuras temporales posteriores.

El conjunto de expresiones (Ec. 3.4) enuncian el problema de optimización finita obtenido:

En cada ranura temporal T, dado un estado del sistema

$$E_{ibj} = \begin{cases} 1 \text{ el retardo } b \text{ para el puerto de entrada } i, \text{ está ocupado} \\ \text{por un paquete destinado a la fibra } j; i = 0...nN-1, j = 0...N-1 \\ 0 \text{ en caso contrario} \end{cases}$$

que determinan los vectores de estado

$$X_i[b] = \sum_{j=0}^{N-1} E_{ibj} \in \{0,1\}, \text{ ocupación de puerto de entrada } i \text{ de}$$

$$\text{la etapa de conmutación}, i = 0...nN-1$$

$$Y_j[b] = \sum_{i=0}^{nN-1} E_{ibj}, \in \{0,...,n\} \text{ ocupación de fibra de salida } j = 0...N-1$$

$$\text{obtener } X_{ibj} \text{ que} \begin{cases} \text{entre las soluciones que maximicen el throughput} \left\{ \sum_{i,b,j} X_{ibj} \right\} \\ \text{sea la que minimice el retardo medio asignado} \left\{ \sum_{i,b,j} b \cdot X_{ibj} \right\} \end{cases}$$

sujeto a las restricciones

$$(1) \sum_{b=0}^{M-1} A_{ibj} \le R_{ij}, \forall i = 0,...,nN-1, \forall j = 0,...,N-1$$

$$(2) \sum_{i=0}^{nN-1} A_{ibj} + E_{ibj} \le n, \forall b = 0,...,M-1, \forall j = 0,...,N-1$$

$$(3) \sum_{j=0}^{N-1} A_{ibj} + E_{ibj} \le 1, \forall i = 0,...,nN-1, \forall b = 0,...,M-1 \qquad \text{(Ec. 3.4)}$$

$$A_{ibj}^{(T)} = \begin{cases} 1 \text{ si a un paquete entrante por el puerto } i, \text{ destinado a la fibra } j, \\ \text{se le asigna el retardo } b; i = 0,..,nN-1, j = 0,...,N-1, b = 0,...,M-1 \\ 0 \text{ en caso contrario} \end{cases}$$

$$R_{ij}^{(T)} = \begin{cases} 1 \text{ si un paquete es recibido por el puerto de entrada } i, \text{ destinado} \\ \text{a la fibra } j; i = 0,...,nN-1, j = 0,...,N-1 \\ 0 \text{ en caso contrario} \end{cases}$$

3.2.3.2.4 Expresión como un problema de emparejamiento máximo en grafos bipartitos

En esta sección mostraremos cómo el problema de optimización descrito en la Ec. 3.4 puede hacerse equivalente a un conjunto de problemas de Emparejamiento en Grafos Bipartitos. Teniendo en cuenta la definición de los vectores X_i, *i=0...nN-1*, Y_j, *j=0...N-1*, en la Ec. 3.4 (diferente en el caso Y_j a lo definido en la planificación SHWP):

- Dentro de una ranura temporal *T*, la asignación de retardos a paquetes destinados hacia la misma fibra de salida *j*, se encuentran mutuamente afectados por la inspección del vector Y_j.
- En consecuencia, debe realizarse de manera global una asignación óptima de retardos (en el sentido de la Ec. 3.4) a estos paquetes.
- Dentro de una ranura temporal *T*, la asignación de retardos a paquetes destinados a fibras de salida diferentes, son independientes.

El último punto nos permite dividir y paralelizar el problema en *N* problemas independientes, uno para cada fibra de salida del conmutador. Cada uno de estos problemas individuales puede ser expresado como un grafo bipartito. Para una fibra de salida *j*, su grafo bipartito es tal que:

(1) Contiene *nN* nodos del lado izquierdo del grafo (N_1), uno para cada puerto de entrada del conmutador.

(2) Para cada retardo *b, b=0...M-1*, contiene $n\text{-}Y_j[b]$ nodos del lado derecho del grafo, es decir, tantos como longitudes de onda libres en esa fibra de salida y ese retardo. Esto hace un total de $nM - \sum_{b=0}^{M-1} Y_j[b]$ nodos del lado derecho del grafo (N_2). Nótese que no existe una relación fija entre los nodos de N_2 y las longitudes de onda de transmisión finales, sino un método para asegurar que un máximo de *n* paquetes son destinados a la misma fibra de salida en cada ranura temporal.

(3) Existe un arco entre el puerto de entrada *i*, y *cada uno* de los $(n\text{-}Y_j[b])$ nodos asociados al retardo *b*, en el caso de que:

- en el puerto de entrada *i* exista un paquete destinado la fibra de salida *j* ($R_{ij}=1$)
- y que el retardo *b* sea un retardo elegible para ser asignado a ese paquete, $X_i[b]=0; Y_j[b]<n$.

Bajo estas condiciones, un emparejamiento de este grafo bipartito es una solución válida al problema de planificación en la que (1) a cada paquete de entrada se le asigna a lo sumo un retardo con una longitud de onda libre, y (2) a cada longitud de onda libre en un retardo, se le asigna a lo sumo un paquete. Como consecuencia, la función objetivo indicada en la Ec. 3.4, se traduce en encontrar un emparejamiento en cada uno de los *N* grafos bipartitos independientes, tal que cada uno de ellos:

- Sea una solución de tamaño máximo (MSM) al problema de emparejamiento (maximice el *throughput*).
- En el caso de existir varias soluciones en un grafo de tamaño igual al MSM, debe escogerse aquella con la media de los retardos asignados menor.

En la figura 3-5 se muestra un ejemplo de aplicación en un conmutador IB-WR SCWP con *8* puertos de entrada y *8* puertos de salida (*N=4* fibras de salida y *n=2* longitudes de onda por fibra), *M=3* retardos d_0,d_1,d_2. En una determinada ranura temporal *T*, el estado del sistema es el mostrado por los vectores X_i, *i=0,...,7,* Y_j, *j=0,...3*. Las llegadas al sistema en esa ranura temporal son las indicadas por R_{ij}. Los puertos de entrada *0, 2 y 5,* reciben paquetes destinados a la fibra de salida *0*. El puerto *4* recibe un paquete destinado a la fibra de salida *1*. Las fibras de salida *2* y *3* no tienen puertos destinados a ellas. El estado del sistema y el proceso de llegadas producen los *2* grafos bipartitos mostrados en la figura 3-5-(b) (no se muestran los grafos vacíos para la fibras de salida *2* y *3*).

En [Pav03-1] se propuso un algoritmo de selección secuencial para el modo de operación SCWP, modificación trivial del empleado para el modo SHWP. Este algoritmo, inspecciona secuencialmente los puertos de entrada, asignando en cada caso el menor retardo disponible, como se observa en la figura 3-6. Su interés radica

en mostrar cómo la aplicación de un algoritmo de planificación secuencial de este tipo, produce en este caso un emparejamiento de menor tamaño al MSM en la fibra de salida *0*, donde el paquete entrante por el puerto *5* es descartado, como se muestra en la figuras 3-5-(c). En esta misma figura se revela que la aplicación de un algoritmo MSM produciría un emparejamiento donde el paquete entrante por el puerto *5* no es descartado.

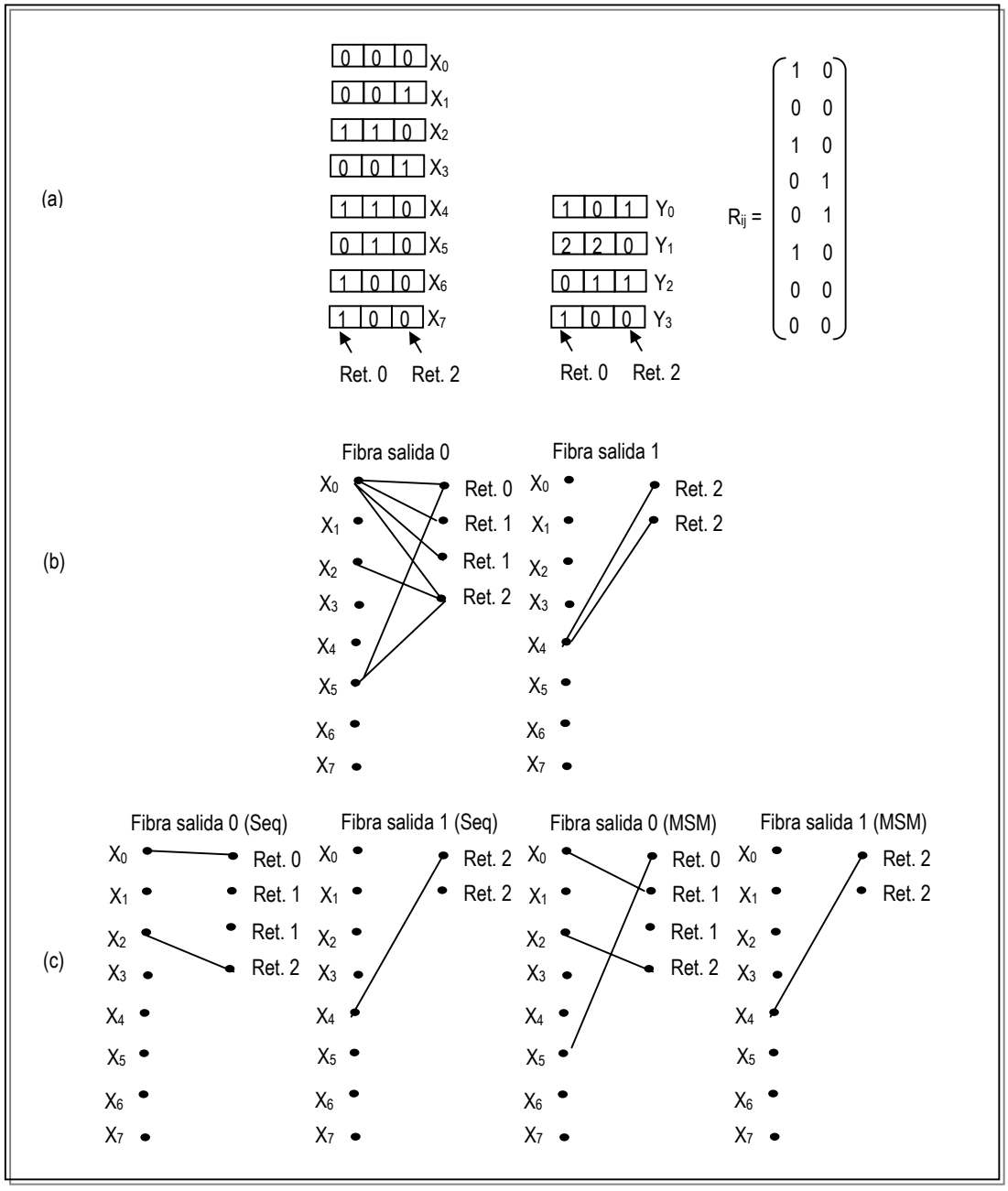


Figura 3-5. Ejemplo de emparejamiento planificación SCWP, (a) estado del sistema y proceso de llegadas, (b) planteamiento del problema como un conjunto de grafos bipartitos, (c) planificación por algoritmo secuencial y por algoritmo MSM

```
/* N = nº de fibras de entrada/salida */
/* n = nº de longitudes de onda por fibra */
/* M = nº de retardos */
/* Xfin,win = vector de estados del puerto de entrada [fin,win]
/* Yfout = vector de estados de la fibra de salida fout

for fiber input fin = 0 to N-1 do
  for wavelength input win = 0 to n-1 do
    if (packet is present on input [fin,win]) then
      fout = output fiber required by input packet
      for timeslot t = 0 to M-1 do
        if Yfout[t] < n and Xfin,win[t] == 0 then
          Xfin,win[t] = 1
          Yfout[t] ++
          assign time slot t to packet
          break
        endif
      endfor
    endif
  endfor
endfor

/* Actualizar valores de Xi, Yj tras cada ranura temporal */

for i=0 to N-1 do
  shift Xi vector to the left. Xi[M-1] = 0
  shift Yi vector to the left. Yi[M-1] = 0
endfor
```

Figura 3-6. Algoritmo secuencial de planificación SCWP

Nota: El problema de planificación SCWP puede ser planteado de manera equivalente como un único grafo bipartito, con *nN* nodos del lado izquierdo (uno para cada puerto de entrada), y hasta *nNM* nodos del lado derecho (*nM* por cada fibra de salida). En esta sección se ha aprovechado la condición de poder ser descompuestos en *N* grafos aislados (uno por cada fibra de salida), cuyo problema de emparejamiento puede resolverse en paralelo.

3.2.3.3 Expresión como un problema de emparejamiento máximo en grafos bipartitos ponderados

En las secciones anteriores se mostró que los problemas de planificación entera para la planificación SHWP y SCWP pueden ser expresados como un problema de emparejamiento máximo en grafos bipartitos. La inclusión de la condición de minimización de retardo medio, obligaba a la búsqueda entre todas las soluciones de tamaño máximo, de aquella con menor retardo medio.

Esta condición puede expresarse de manera más simple empleando el concepto de grafo bipartito ponderado (*bipartite weighted graph*). Un grafo bipartito ponderado es un grafo bipartito en el que cada arco *(i,j)* tiene asociado un número real w_{ij}, llamado peso del arco. Un algoritmo MWM (*Maximum Weight Matching*) aplicado a un grafo ponderado de este tipo, es aquél que proporciona el emparejamiento que maximice la suma de los pesos de los arcos involucrados. Lógicamente, en el caso de que todos los pesos w_{ij} de un grafo bipartito sean iguales, el problema es equivalente a encontrar el emparejamiento de mayor tamaño (MSM) del grafo.

El peso de los arcos nos permitirá incluir información de retardo directamente dentro del grafo. El objetivo de esta sección es definir los pesos asociados a los arcos de los grafos indicados en las secciones SHWP y SCWP anteriores, de forma tal que la aplicación de un algoritmo MWM produzca la solución buscada en la que: (1) se

maximice el *throughput*, (2) en caso de que exista más de un emparejamiento de tamaño máximo, el de mayor peso sea aquél con el retardo medio más bajo.

Nótese que el único objetivo es encontrar una manera *equivalente* de expresar el mismo problema de optimización, pero que ahora sea resoluble mediante algoritmos MWM.

En [Pav03-1] se dedujeron los valores sobre los pesos w_{ij}, que provienen de los siguientes condicionantes:

(1) Los pesos asociados a arcos del mismo retardo deben ser iguales, independientemente del puerto de entrada.

$$w_j = w_{i_1 j} = w_{i_2 j} \forall i_1, i_2 \in \{0,...,nN-1\} \text{ (modo SHWP y SCWP)}$$

$$w_{j_1} = w_{j_2}, \forall j_1, j_2 \text{ nodos asociados al mismo retardo (condición extra en modo SCWP)}$$

(2) Los pesos deben favorecer (peso mayor) los retardos más cortos.

$$w_{j_1} > w_{j_2} \Leftrightarrow j_1 \text{ asociado a retardo menor que } j_2 \text{ (en lo sucesivo } j_1 < j_2)$$

(3) Debe cumplirse siempre, que cualquier emparejamiento de tamaño *s*, tenga un peso mayor que cualquier emparejamiento de tamaño *s-1*.

$$s \cdot w_{M-1} > (s-1) \cdot w_0, \forall s = 2,...,nN$$

(4) La maximización del peso de un emparejamiento debe ser equivalente a la minimización del retardo medio del mismo.

$$w_j = w_0 - k \cdot j; k = \frac{w_0 - w_{M-1}}{M-1}; j = 0,...,M-1$$

La condición (3) describe el caso peor, que se alcanza en modo SCWP, y que sirve como cota para el modo SHWP. El caso consiste en asegurar que un emparejamiento de tamaño *s* con los arcos de menor peso (mayor retardo) sea preferible siempre a un emparejamiento de tamaño menor *(s-1)*, aunque sea con los arcos de mayor peso (menor retardo). El condicionante más restrictivo se alcanza para *s=nN*, que lleva a la condición:

$$w_{M-1} > \frac{(nN-1)w_0}{nN} \qquad \text{(Ec. 3.5)}$$

La condición (4) se requiere para asegurar que el emparejamiento de mayor peso minimice el retardo medio, como se deduce de la (Ec. 3.6), donde *S* simboliza el conjunto de *s* arcos del emparejamiento.

$$\sum_{j \in S} w_j = \sum_{j \in S} w_0 - k \cdot j = s \cdot w_0 - \sum_{j \in S} j = s \cdot w_0 - k \cdot s \sum_{j \in S} \frac{j}{s} = s \cdot w_0 - k \cdot s \cdot \overline{W} \qquad \text{(Ec. 3.6)}$$

Las condiciones (1)-(4) dejan un grado de libertad para el cómputo de los pesos de los distintos arcos. Un posible mecanismo para la obtención de los mismos es el descrito en [Pav03-1]. Mostraremos este mecanismo calculando los pesos del conmutador SCWP ejemplo de la figura 3-5, de *4* fibras de entrada y salida, *2* longitudes de onda por fibra, y *3* retardos. Inicialmente, se fija el valor de w_0 a *1*. La aplicación de las condiciones (2) y (3) proporciona los valores $w_{M-1}=w_2>7/8=0.875$. Se escoge el valor $w_2=0.9$. La aplicación de la condición (4) determina que $k=0.1/2=0.05$, por lo que $w_1=w_0+k=0.85$.

La figura 3-7 presenta un segundo ejemplo de un conmutador *8x8* SCWP con *4* fibras de entrada y salida, *2* longitudes de onda por fibra, y *5* retardos $d_0,...,d_4$. En las figuras 3-7-(a) y (b) se describe el planteamiento del problema para la fibra de salida *2*, mostrando únicamente la información relativa a los puertos de entrada *0, 3* y *4*, que contienen paquetes destinados a la fibra *2*. Las figuras 3-7-(c) y 3-7-(d) muestran dos soluciones de tamaño máximo *3*. Los pesos asociados a cada arco, siguiendo el mismo método que en el ejemplo anterior son: $w_0=1, w_4=0.9>7/8=0.875$, $k=0.025$, $w_1=0.975$, $w_2=0.95$, $w_3=0.925$. La opción mostrada en la figura 3-7-(c) acumula un peso de (*1+0.925+0.9)=2.825*, y tiene asociado un retardo de *(0+1+4)/3=0.166*. La opción mostrada en la figura 3-7-(d) acumula un peso de *(0.95+0.975+1)=2.925*, y tiene asociado un retardo menor de *(0+1+2)/3=1*. Esta sería la opción elegida por un algoritmo MWM, igual que por un algoritmo MSM que convergiese a la solución de menor retardo medio, cumpliendo la función objetivo deseada para el problema de optimización.

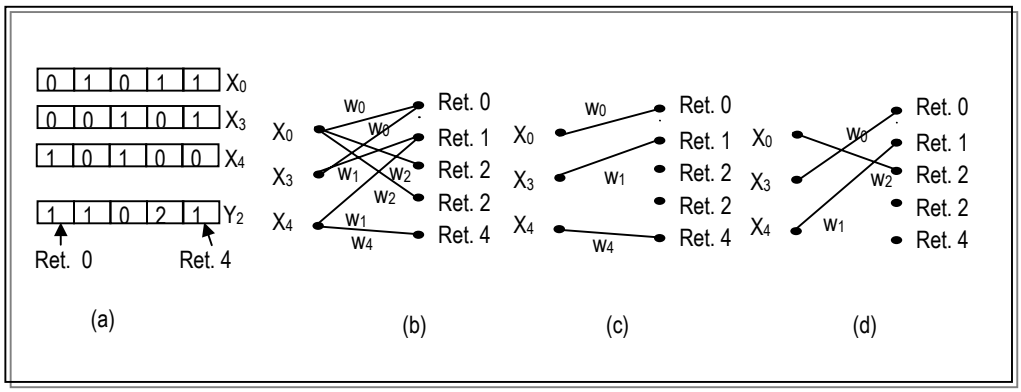


Figura 3-7. Ejemplo de aplicación de algoritmo MWM a la planificación SCWP, (a) vectores de estado relevantes, (b) planteamiento del grafo para la fibra de salida *2,* (c) solución 1, (d) solución 2

Es importante destacar los siguientes puntos:

- Se remarca que este planteamiento del problema como un grafo bipartito ponderado con los pesos calculados, es equivalente al problema de optimización estudiado en este capítulo.
- El diseño de los pesos puede tener en cuenta otros aspectos como la priorización de tráfico, asignando mayor peso a los paquetes entrantes con una determinada marca de prioridad.
- En general, los algoritmos MWM trabajan por motivos de eficiencia con pesos enteros. En el caso de aplicar directamente un algoritmo de este tipo para la implementación de un planificador, es posible transformar los pesos calculados a números enteros, teniendo en cuenta que la suma de un

número constante a todos los pesos, o la multiplicación de todos los pesos por un número positivo constante, proporciona un grafo ponderado con la misma solución MWM.

3.2.4 Diferencias con la planificación en conmutadores VOQ

Las arquitecturas de conmutación basadas en Colas Virtuales a la Salida (VOQ, *Virtual Output Queueing*) son empleadas actualmente en los conmutadores electrónicos de paquetes de altas prestaciones (como por ejemplo la serie Cisco 12000), y constituyen, hoy en día, un campo activo de investigación. Una arquitectura VOQ para un conmutador de *N* puertos de entrada y *N* puertos de salida, se basa en la existencia de memorias en los puertos de entrada del conmutador, organizadas en *NxN* colas virtuales, una para cada par origen-destino. Un paquete proveniente del puerto *i*, destinado al puerto *j* se introduce en la cola *VOQ(i,j)*. Estas memorias se conectan a una matriz de conmutación sin memoria *NxN*, que encamina los paquetes desde los puertos de entrada hacia sus respectivos puertos de salida. El número de paquetes que pueden ser leídos de las memorias y transmitidos en cada ranura temporal es igual al factor de aceleración (*s*, *speed-up).* El control del conmutador es el encargado de decidir desde qué cola VOQ de cada puerto de entrada se extraerán paquetes en cada ranura temporal. El problema de planificación (reducido a una ranura temporal) que este control debe solucionar es el siguiente:

$$
\begin{aligned}
&\textit{En cada ranura temporal } T\text{, } \textit{maximizar } \sum_{i=0}^{N-1}\sum_{k=0}^{N-1} w_{ij} A_{ij} \\
&\textit{sujeto a las restricciones} \\
&(1)\, A_{ij} \le R_{ij}, A_{ij}, R_{ij} \in \{0,1\}; i, j = 0,...,N-1 \\
&(2)\sum_{i=0}^{N-1} A_{ij} \le s,\ s = \textit{speed} - \textit{up del sistema} \\
&(3)\sum_{j=0}^{N-1} A_{ij} \le s,\ s = \textit{speed} - \textit{up del sistema} \\
&A_{ibj}^{(T)} = \begin{cases} 1 \textit{ si la cola } VOQ(i,j) \textit{ es elegida, y el paquete} \\ \textit{ en la cabeza de la cola es transmitido} \\ 0 \textit{ en caso contrario} \end{cases} \\
&R_{ij}^{(T)} = \begin{cases} 1 \textit{ si la cola VOQ no está vacía} \\ 0 \textit{ en caso contrario} \end{cases}
\end{aligned}
\qquad \text{(Ec. 3.7)}
$$

- Los pesos w_{ij} se incluyen para potenciar el servicio de VOQs más cargadas, frente a VOQs con pocos paquetes.
- El *speed-up* del sistema indica el número de paquetes que pueden ser leídos de un puerto de entrada en una ranura temporal, y el número de paquetes que pueden llegar a un puerto de salida en una ranura temporal. Un valor mayor que *1* para este factor, implica por tanto, (1) la necesidad de memorias en los puertos de salida (si la velocidad de operación de los puertos de entrada y salida es la misma), (2) la aceleración de la matriz de conmutación respecto a los puertos de entrada y salida, y (3) la aceleración de todas las memorias, capaces de funcionar a *s* veces la velocidad de línea.

El problema de planificación indicado se puede expresar como un único problema de maximización del emparejamiento en grafos bipartitos ponderados (por el peso w_{ij}), definido de la siguiente manera:

- Un nodo del lado izquierdo para cada puerto de entrada del conmutador.
- Un nodo del lado derecho para cada puerto de salida del conmutador.
- Un arco entre el nodo *i* y el nodo *j* si existen paquetes originados en el puerto de entrada *i* destinados al puerto *j* (es decir, si la cola *VOQ(i,j)* no está vacía).

Este hecho ha impulsado el estudio de algoritmos que sean capaces de resolver este tipo de problemas de manera sub-óptima, pero implementable a las velocidades requeridas. Por ello, es muy interesante destacar las diferencias y parecidos existentes entre los problemas de planificación de los conmutadores VOQ y los problemas descritos para los conmutadores IB-WR.

- **Nodos del grafo.** En ambos casos, los nodos de la izquierda representan los puertos de entrada. Los nodos de la derecha representan retardos libres en el conmutador IB-WR, y puertos de salida en el conmutador VOQ.
- **Número de grafos.** En cada ranura temporal, el problema VOQ equivale a la resolución del emparejamiento de un grafo bipartito. En el caso IB-WR, a pesar de poder ser expresado todo el problema en un único grafo, se trata de *nN* (SHWP) o *N* (SCWP) problemas independientes (de menor tamaño), que pueden ser resueltos en paralelo.
- **Evolución temporal.** En un conmutador VOQ, los paquetes no seleccionados continúan almacenados en cola (no se pierden), y se añaden a las llegadas en siguientes ranuras temporales. En un conmutador IB-WR, los paquetes no seleccionados son descartados. Los paquetes a los que se les asigna un retardo *b* afectan al estado del conmutador en las siguientes ranuras temporales de una manera muy determinada.
- **Función objetivo.** Por la distinta naturaleza del problema a resolver, las diferencias deben aparecer reflejadas en la función objetivo. Estas pueden concentrarse en el significado de los pesos w_{ij}:

- En un conmutador VOQ, todos los puertos de salida son iguales para el conmutador. Los pesos w_{ij} se seleccionan para favorecer el vaciado de las colas VOQ más cargadas.
- En los conmutadores IB-WR, no todos los nodos de la derecha tienen el mismo significado. Como se ha indicado en la sección anterior, deben preferirse los nodos asociados a retardos más cortos, lo cual se refleja en su mayor peso en el grafo. Asimismo, los emparejamientos de tamaño sub-óptimo tienen mayor coste en prestaciones (los paquetes son descartados), por lo que la ponderación debe favorecer de manera especial los emparejamientos de mayor tamaño.

En definitiva, la expresión del problema de planificación en los conmutadores IB-WR como de maximización del emparejamiento, nos permite observar que, dos escenarios de planificación de características y criterios distintos, concentran sus diferencias fundamentalmente en el planteamiento del problema (construcción del

grafo bipartito), pero pueden compartir estrategias de obtención de algoritmos eficientes, implementables a las velocidades requeridas (en el orden de la decenas de nanosegundos en conmutadores VOQ, y de la duración de paquete en arquitecturas OPS). El algoritmo PDBM (*Parallel Desynchronized Block Matching*) propuesto para la planificación IB-WR SCWP, es una prueba de esta afirmación.

3.2.5 Equivalencia con la planificación en arquitecturas WASPNET

El proyecto WASPNET (*WAvelength Switched Packet NETwork*) [Hun99], del que surgieron las definiciones de los modos de operación SHWP/SCWP, fue asimismo origen de propuestas de diversas arquitecturas de conmutación. Para todas ellas, se estableció un formato de paquete formado por una cabecera de *4* bytes, incluyendo un identificador de OPP de *24* bits, un campo de tipo de datos de *2* bits, un campo de prioridad de paquete de *2* bits, y un campo de datos de *256* bytes a *10* Gbps.

En la arquitectura WASPNET sin retroalimentación (*feed-forward*) mostrada en la figura 3-8-(a), los paquetes se convierten a una longitud de onda que les dirige hacia uno de los *N* módulos de almacenamiento. Cada uno de ellos está compuesto por un convertidor TWC, que dirige el paquete hacia uno de los *0,...,M-1* retardos. Los paquetes salientes de los módulos de almacenamiento son conmutados hacia el puerto de salida mediante una nueva conversión de longitud de onda y un encaminamiento AWG.

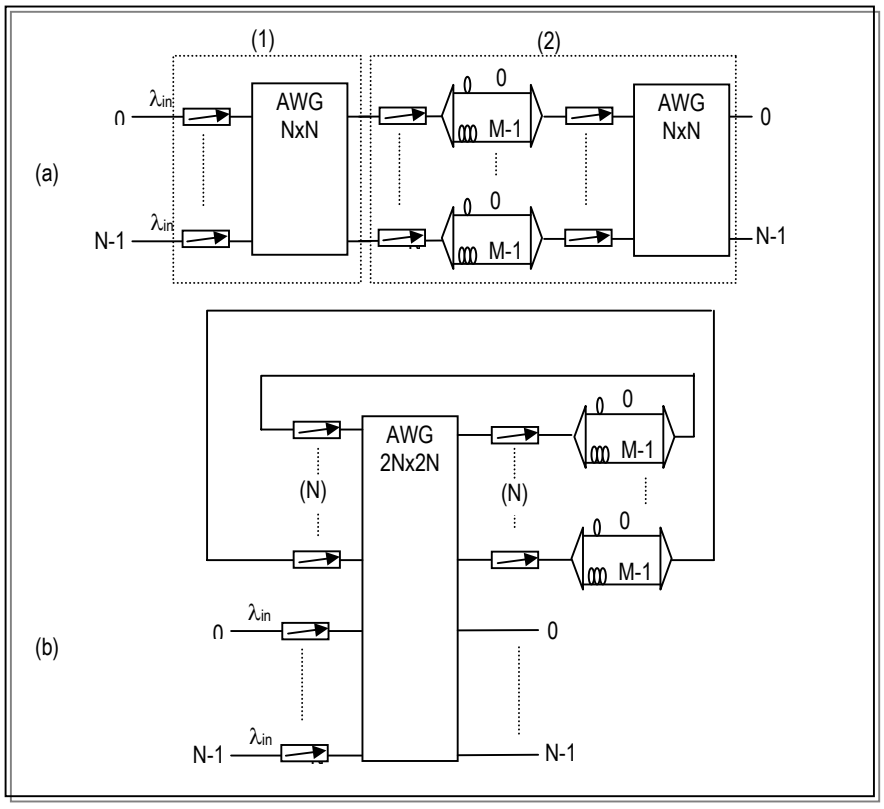


Figura 3-8. (a) Arquitectura WASPNET *feed-forward*, (b) Arquitectura WASPNET *feed-back*

Fijándonos en las restricciones a la planificación de la arquitectura *feed-forward* se observa lo siguiente:

- En la sección etiquetada como (1), la elección de la primera longitud de onda de conversión está restringida por el hecho de que dos paquetes no pueden ser encaminados en la misma ranura temporal al mismo módulo de almacenamiento.
- En la sección etiquetada como (2), la elección de retardo en cada módulo de almacenamiento *i* está condicionada por dos aspectos:

- Un máximo de un paquete puede alcanzar la salida de cada módulo de almacenamiento en una ranura temporal, ya que un máximo de un paquete puede llegar al convertidor ahí situado.
- El retardo está también restringido por la contención por el puerto de salida final *j* al que deberá dirigirse el paquete: por cada puerto de salida puede transmitirse un máximo de un paquete en una ranura temporal.

La descripción anterior muestra lo siguiente:

- La sección (2) del conmutador WASPNET *feed-forward* es exactamente equivalente a un conmutador IB-WR, en su versión original [Zho98]. De hecho, la sustitución de esta sección por un conmutador IB-WR compite con ventaja en simplicidad *hardware,* en cuanto a kilómetros de fibra.
- La sección de conmutación sin memoria (1) consituye un caso de pre-etapa para el equilibrado de carga, a la entrada del conmutador. Su objetivo debe por tanto, ser la redistribución de los paquetes entrantes de tal manera que su llegada al conmutador IB-WR-equivalente, maximice sus prestaciones. Se trata de una línea de investigación cercana a la selección de la longitudes de onda de transmisión que maximice las prestaciones de un nodo IB-WR posterior, expresada en la sección 3.2.3.2.2 *Selección de longitud de onda de transmisión*.

Dentro de las publicaciones del proyecto WASPNET, se ha descrito un algoritmo de planificación secuencial, sub-óptimo y presumiblemente no implementable a las velocidades deseadas. No se ha hecho ninguna referencia al paralelismo existente con el conmutador IB-WR, ni se ha descrito la primera sección del conmutador como una sección de equilibrado de carga.

El equilibrado de carga como etapa previa a conmutadores electrónicos con colas virtuales a la salida (*Virtual Output Queueing, VOQ*) es un campo de floreciente estudio actual, debido al cuello de botella en el control del conmutador que existe en el escenario electrónico. A pesar de que el problema de planificación asociado a un conmutador VOQ es distinto al de un planificador IB-WR, y que es dudoso que ese cuello de botella exista en el escenario OPS (hoy en día al menos), resulta interesante el estudio de qué métodos de equilibrado de carga mejoran las prestaciones de los conmutadores IB-WR SHWP/SCWP. Este campo está sin duda relacionado (sin ser equivalente) con la decisión de asignación de longitud de onda de transmisión en la etapa de conmutación en la arquitectura IB-WR SCWP, o en la elección de longitudes de onda de transmisión en los nodos frontera en modo SCWP.

En esta tesis doctoral no se profundiza en la problemática de este equilibrado de carga. Asimismo, no se considera la arquitectura WASPNET como candidata. Esto es debido a que las prestaciones de una arquitectura WASPNET con *M* retardos están acotadas:

- Superiormente, por los conmutadores con colas a la salida con *M* retardos, cuyas prestaciones son óptimas.
- Inferiormente, por el conmutador IB-WR sin etapa de equilibrado.

Como se mostrará en este capítulo, las prestaciones de los conmutadores IB-WR son aceptablemente cercanas al máximo alcanzable con conmutadores que emulan colas a la salida. Por ello, la posible mejora que proporcionaría una etapa de equilibrado, no ha sido considerada. Sin embargo, se destaca que la evolución tecnológica puede situarnos en el futuro en un escenario en el que debamos replantear esta afirmación.

Un caso a parte requiere la arquitectura WASPNET con retroalimentación (*feed-back*), mostrada en la figura 3-8-(b). Esta arquitectura es equivalente a su versión *feed-forward* salvo en el caso de que un paquete pueda ser encaminado más de una vez hacia las líneas de retardo. Dentro del proyecto WASPNET se destacó esta arquitectura, argumentando la posible aplicación de técnicas de calidad de servicio empleando la marca de prioridad de paquete en la cabecera del mismo (aunque no se ha descrito ninguna técnica ni algoritmo concreto para su aplicación). Esta arquitectura no ha sido objeto de estudio en esta tesis doctoral.

3.3 Algoritmo de planificación SCWP PDBM

En esta sección se describirá el algoritmo de planificación PDBM (*Parallel Desynchronized Block Matching*) propuesto en esta tesis doctoral, para arquitecturas IB-WR (figura 3-1-(b)) en modo SCWP. Sin embargo, todo lo indicado puede ser aplicado igualmente bajo el modo de operación SHWP, considerando un conmutador con *nN* fibras de salida, y *1* longitud de onda por fibra. El interés de este algoritmo reside en que su implementación es viable para las velocidades de planificación demandadas.

3.3.1 Antecedentes

La definición del algoritmo PDBM, se inicia con un estudio de distintas propuestas de algoritmos de planificación para conmutadores electrónicos de paquetes VOQ. Estos algoritmos consisten en heurísticos que tratan de aproximar una solución MSM o MWM al grafo bipartito planteado en cada ranura temporal. Los algoritmos que aseguran una solución óptima MSM o MWM no son aplicables por: (1) tener un tiempo de convergencia del orden $O(N^{2.5})$ iteraciones para MSM y $O(N^{2.5}log_2N)$ iteraciones para MWM, en sus versiones más eficientes conocidas, (2) ser de implementación *hardware* muy costosa. Comenzaremos la descripción con el algoritmo de planificación *iSLIP* [McK99], considerado un estándar *de facto* en este campo. El algoritmo *iSLIP* es implementable mediante circuitos electrónicos de complejidad acotada, estando presente en arquitecturas comerciales (como por ejemplo la serie Cisco 12000).

Cada una de las iteraciones del algoritmo *iSLIP* establece un conjunto de tres pasos consecutivos, realizados cada uno de ellos en paralelo por los puertos de entrada y de salida. Cada una de estas fases es implementable mediante circuitos electrónicos combinacionales. Para un algoritmo *iSLIP* de una iteración, en un conmutador con *N* puertos de entrada y *N* puertos de salida, los pasos definidos son:

- **Paso 1. *Request*.** Cada puerto de entrada envía una señal de petición (*request*) a cada puerto de salida para los que tiene un paquete en espera.

- **Paso 2. *Grant.*** Cada puerto de salida mantiene un puntero *round-robin,* que apunta a los puertos de entrada. Comenzando la búsqueda por el puerto de entrada al que apunta el puntero, se otorga una aprobación (*grant*) a la primera petición encontrada. En el caso de que se envíe una aprobación a un puerto de entrada, y ésta sea aceptada, el puntero se moverá al siguiente puerto de entrada (módulo *N*). En caso contrario, el puntero mantiene su posición.
- **Paso 3. *Accept***. Cada puerto de entrada mantiene un puntero *round-robin,* que apunta a los puertos de salida. Cada puerto de entrada acepta (*accept*), entre las aprobaciones que reciba de los puertos de salida, la primera que encuentra, comenzando la búsqueda por la posición indicada por su puntero. Tras una aceptación, el puntero se fijará al puerto de salida siguiente (módulo *N*) al aceptado.

El funcionamiento del algoritmo requiere, por tanto, la utilización de un puntero por cada puerto de entrada, y un puntero por cada puerto de salida. Nos interesa estudiar las prestaciones del algoritmo, que vienen determinadas principalmente por lo sucedido en el paso *2*. En este punto, cada puerto de salida recibe peticiones de los puertos de entrada que disponen de paquetes almacenados destinados a él. Supongamos un conmutador cargado, en el que todos los puertos de entrada tienen paquetes destinados a todos los puertos de salida (todas las colas VOQ con algún paquete almacenado). Por tanto, todos los puertos de salida reciben *N* peticiones, una de cada puerto de entrada (total N^2 peticiones), y se genera una aprobación por cada puerto de salida (total *N* aprobaciones). En este momento, si varios puertos de salida eligen otorgar su aprobación (*grant*) al mismo puerto de entrada, sucede que:

- El puerto de entrada en cuestión aceptará (*accept*) una de las aprobaciones recibidas, en función de la posición de su puntero *round-robin*, rechazando el resto.
- Al menos un puerto de entrada no recibirá ninguna aprobación, dado que en total *N* han sido generadas, y *2* han coincidido en un solo puerto. Los puertos que no reciben ninguna aprobación no transmiten ningún paquete, a pesar de tener paquetes esperando a ser transmitidos hacia todos los puertos de salida.
- De los dos puertos de salida coincidentes, aquél que no reciba la aceptación (*accept*) no transmitirá ningún paquete, a pesar de que cualquier puerto de entrada tiene paquetes destinados para él.

Este fenómeno se conoce como *sincronización*. Se trata de un efecto a evitar, por el empobrecimiento de las prestaciones del conmutador que implica. La ventaja del algoritmo *iSLIP* reside en que los punteros *grant* de los distintos puertos de salida, tienen tendencia a desincronizarse. Una sincronización de dos punteros en una ranura temporal, se rompe si uno de los dos puertos de salida recibe un *accept* (ya que avanzará su puntero). Para el algoritmo *iSLIP,* la desincronización puede llegar a ser completa a cargas altas del conmutador.

En función del número de veces que se aplica el algoritmo en una ranura temporal, obtenemos el algoritmo *1-SLIP*, *2-SLIP*, etc. Cada iteración consiste en la aplicación del algoritmo a un grafo bipartito al que se le ha eliminado los emparejamientos ya realizados, con sus puertos de entrada y salida involucrados. La única diferencia entre las iteraciones distintas a la primera, es que los punteros *round-*

robin de los puertos de salida no serán incrementados. Se dice que el algoritmo ha convergido, cuando una iteración más del mismo no incrementa el emparejamiento. El algoritmo *iSLIP* convergerá en un máximo de *N* iteraciones, a una solución de tamaño maximal (máximo local), menor o igual al MSM. Los resultados muestran buenas prestaciones del algoritmo para un número bajo de iteraciones [McK99].

En [Jia01] se propone una familia de modificaciones al algoritmo basada en la idea de provocar artificialmente una desincronización completa de los punteros *grant*. Para ello, se obliga a los punteros *round-robin* a apuntar a puertos de entrada distintos, y moverse síncronamente cada ranura temporal, independientemente de las peticiones recibidas, o aprobaciones otorgadas, manteniendo por tanto, la falta de sincronismo (desincronismo) entre todos ellos. Entre las propuestas destacamos la llamada RDSRR (*Rotating Double Static Round Robin*), cuyas prestaciones mejoran las de *iSLIP*, como se muestra en [Jia01]:

- **Paso 0. Inicialización de punteros *grant* y *accept* al arranque del conmutador.** Los punteros *grant* y *accept* se inicializan con el mismo patrón, tal que no hay duplicidad entre punteros.

- **Paso 1. *Request*.** Cada puerto de entrada envía una señal de petición (*request*) a cada puerto de salida para los que tiene un paquete en espera.
- **Paso 2. *Grant*.** Cada puerto de salida mantiene un puntero *round-robin,* que apunta a los puertos de entrada. Comenzando la búsqueda por el puerto de entrada al que apunta el puntero, se otorga una aprobación (*grant*) a la primera petición encontrada, buscando en puertos de orden mayor módulo *N* (lo que llamaremos “en sentido horario”). En la siguiente ranura temporal, esta búsqueda partirá del puerto al que apunta el puntero, continuando en puertos de orden menor módulo *N* (lo que llamamos “en sentido antihorario”). En cada ranura temporal, todos los punteros *grant* se incrementan en una unidad.
- **Paso 3. *Accept***. Cada puerto de entrada mantiene un puntero *round-robin,* que apunta a los puertos de salida. Cada puerto de entrada acepta (*accept*) entre las aprobaciones que reciba de los puertos de salida, la primera que encuentra, comenzando la búsqueda por la posición indicada por su puntero, siempre buscando en el sentido de las agujas del reloj. En cada ranura temporal, todos los punteros *accept* se incrementan módulo *N* en una unidad.

Las características que destacamos de este algoritmo, que serán introducidas en el algoritmo PDBM son:

- Desincronización total de los punteros en su inicialización, que se mantiene al moverse todos ellos de manera síncrona.
- La utilización alternativa de una búsqueda en “sentido horario” y en “sentido antihorario” se realiza para tratar de manera justa todas las fuentes de tráfico, ante cualquier patrón. Esto se explica con un ejemplo de conmutador *4x4* en el que los puertos de entrada *1* y *4* son los únicos con tráfico hacia el puerto de salida *3*, y el conmutador realiza siempre la búsqueda en sentido horario. Si el puntero *grant* de este puerto *3* está en las posiciones *2*, *3* ó *4*, es el puerto *4* el atendido. Únicamente cuando el puntero *grant* del puerto tiene el valor *1*, es este puerto *1* el atendido. Alternando las búsquedas en

sentido horario y antihorario en cada ranura temporal, se elimina este tipo de desigualdad en la asignación de caudal a los puertos de entrada.

3.3.2 Descripción del algoritmo

La estrategia seguida para la definición del algoritmo PDBM ha sido la de un algoritmo iterativo con tres fases (*request, grant, accept*), aprovechando la idea de desincronización forzada de los punteros al estilo de RDSRR. Esta estrategia se ha adaptado en todos los pasos a las características del problema de planificación IB-WR objetivo.

El algoritmo será descrito para un conmutador IB-WR con *N* fibras de entrada y salida, *n* longitudes de onda por fibra, y *M* retardos. Las fases *request, grant, accept* involucran *nN* nodos de entrada (uno para cada puerto de entrada del conmutador), y *NM* nodos de salida (*M* para cada fibra de salida, cada uno de ellos asociado a un retardo distinto). Cada nodo de salida maneja un puntero *grant round-robin* independiente, que apunta a uno de los *nN* puertos de entrada. Nótese que el número de nodos de entrada y salida son distintos, así como la no existencia de punteros *accept* en los puertos de entrada.

- **Paso 0. Inicialización de punteros *grant* al arranque del conmutador.** Cada uno de los *M* punteros *grant* asociados a la misma fibra de salida *j, p(f,b), b=0...M-1* deben apuntar a un puerto de entrada *0,...,nN-1*, de tal manera que se maximice la distancia módulo *nN* entre ellos. No se exige ninguna relación especial entre los punteros *grant* de fibras de salida distintas, para el mismo retardo. En las pruebas realizadas $p(f_1,b)=p(f_2,b), \forall f_1,f_2=0,...,N-1, \forall b=0,...,M-1$.

En cada ranura temporal se producen un número acotado de iteraciones del algoritmo, cada una de ellas como la que sigue:

- **Paso 1. *Request*.** (Cada uno de los *nN* puertos de entrada del conmutador en paralelo). Para el puerto de entrada *i,* en el caso de recibir un paquete destinado a la fibra de salida *j*, se observa el vector de estado X_i y se envía una señal *request* a todos los $k \leq M$ nodos de la fibra de salida *j,* asociados a retardos elegibles ($X_i[k]=0$).
- **Paso 2. *Grant.*** (Cada uno de los *NM* nodos de salida en paralelo). Cada nodo de salida $(f,b),\ 0 \leq f \leq N-1,\ 0 \leq b \leq M-1$, maneja la información (actualizada cada iteración) de la coordenada $Y_j[b] \leq n$ del número de paquetes que ya han sido asignados para ser transmitidos por la fibra de salida *f* dentro de *b* ranuras temporales . Por ello, conoce el número $d=n-Y_j[b]$ de longitudes de onda de transmisión aún disponibles. La fase *grant* para cada nodo, consiste en otorgar un bloque de hasta *d* aprobaciones, una para cada *request* recibida, comenzando la búsqueda a partir del puerto de entrada marcado por el puntero del nodo. La búsqueda en todos los nodos se produce en sentido horario, si en la iteración anterior se produjo en sentido antihorario, y viceversa. Cada dos iteraciones (una en cada sentido), los punteros de todos los nodos son incrementados síncronamente en *1* posición (módulo *nN*).
- **Paso 3. *Accept*.** (Cada uno de los *nN* puertos de entrada del conmutador en paralelo). De todas las aprobaciones recibidas (un máximo de *M*, una para

cada retardo de la fibra de salida del paquete), acepta la de menor retardo, que asigna al paquete de entrada en el puerto.

3.3.3 Justificación y propiedades del algoritmo

3.3.3.1 Inicalización de los punteros *grant*

Como se deduce de los pasos indicados, al igual que en el algoritmo RDSRR, la posición de los punteros se modifica de manera completamente independiente de las peticiones recibidas, aprobadas o aceptadas, o cualquier otra situación que atañe al tráfico. La posición $p=0,..,nN-1$ de un puntero al comienzo de una iteración T depende únicamente de su posición inicial p_0, $p=(p_0+ floor\ (T/2))\ mod\ nN$. Por ello, la posición inicial de cada uno de los MN punteros $p_0(f,b),\ f=0,...,N-1,\ b=0,...,M-1$, debe ser escogida cuidadosamente. El criterio empleado en esta tesis doctoral es el de la *minimización del solapamiento*.

Para una fibra de salida j concreta, el número máximo de aprobaciones que cada retardo b puede otorgar es igual al número de longitudes de onda libres $d(j,b),\ b=0,...,M-1$. Si un retardo (j,b) recibe un número de peticiones menor o igual a $d(j,b)$, todas ellas recibirán aprobación, sin que importe la posición del puntero *grant* $p(j,b)$. En caso de recibir más peticiones que posibles aprobaciones, interesa que aquellas peticiones que no reciban aprobación de un retardo b, por estar "lejos" del puntero $p(j,b)$, lo reciban de algún otro retardo b' de la misma fibra de salida. Esto se ve favorecido si las posiciones de los punteros $p(j,b),\ b=0,...,M-1$ están lo más separadas posible módulo nN, cubriendo los nN puertos de entrada.

Como ejemplo, supongamos el caso de que k paquetes son recibidos con destino una misma fibra de salida j. Supongamos que todos los vectores X_i de los puertos de entrada en cuestión están vacíos, por lo que todos los puertos de entrada generan M peticiones, una a cada nodo de salida $(j,b),\ b=0,..,M-1$. El número de longitudes de onda libres en cada uno de estos puertos es $d(j,b),\ b=0,...,M-1$. Si en este punto, los M punteros de los nodos fueran iguales, las aprobaciones de retardo 0 se otorgarían a las $d(j,0)$ primeras peticiones, las de retardo 1 a los $d(j,1)$ primeras peticiones, ..., estando todas ellas solapadas. El número de puertos que recibiría alguna aprobación de algún puerto sería igual a $max\{d(j,b)\},\ b=0,...,M-1$. Un número de paquetes entrantes igual a $k-max\{p_b\}$ serían descartados. Esto se evitaría disminuyendo el solapamiento entre los bloques otorgados como se ha indicado previamente.

El proceso de inicialización empleado en esta tesis doctoral se describe en la figura 3-9. Obsérvese que, justificado por la independencia de decisión entre los paquetes destinados a fibras de salida distintas (en una misma ranura temporal), se ha empleado la misma distribución de punteros en todas las fibras de salida $p(f_1,b)=p(f_2,b),\ \forall f_1,f_2=0,...,N-1,\ \forall b=0,...,M-1$. Por supuesto, otras posibles inicializaciones podrían mejorar las prestaciones del algoritmo, siendo éste un posible campo de estudio.

```
/*************************************************************************
 Inicialización de las posiciones de los punteros.
- Los punteros de fibras de salida distintas tienen igual distribución.
- Los punteros de distintos retardos dentro de la misma fibra de salida, buscan
maximizar la distancia entre ellos, para minimizar el solapamiento.
    - Si M>=nN (improbable) cada puntero al puerto de entrada anterior mas uno.
    - Si M<nN maximizar la distancia entre punteros
- La variable rrGrantPointer [j][b] contiene el valor del puntero de la fibra de
salida j, y el retardo b, j=0,...,N-1 ; b=0,...,M-1
*************************************************************************/

if (M >= nN) // más punteros que puertos de entrada (escenario improbable)
{
  rrGrantPointer [0][0] = 0;

  // posicion inicial punteros van creciendo de 1 en 1
  for (delay  = 1; delay < M ; delay ++)
  {
    rrGrantPointer [0][delay] = (rrGrantPointer [0][delay - 1] + 1) % nN;
  }
}
else // más puertos de entrada que punteros (minimizar solapamiento)
{
 rrGrantPointer [0][0] = 0;

  int cociente = (int) (nN / M);
  int resto = (nN % M);

  // M-resto-1 punteros separados "cociente" puertos de entrada del anterior puntero
  for (delay = 1; delay < M-resto ; delay ++)
    rrGrantPointer [0][delay] = (rrGrantPointer [0][delay - 1] + cociente) %  nN;

  // "resto" punteros separados "cociente+1"  puertos de entrada del anterior puntero
  for (delay = M - resto; delay < M ; delay ++)
    rrGrantPointer [0][delay] = (rrGrantPointer [0][delay - 1] + cociente + 1) % nN;
}


/* Todas las fibras de salida con la misma distribucion de punteros */
for (fout = 1 ; fout < N ; fout ++)
  for (delay = 0 ; delay < M ; delay ++)
    rrGrantPointer [fout][delay] = rrGrantPointer [fout - 1][delay];
```

Figura 3-9. Inicialización de punteros *grant* empleada en el algoritmo PDBM (extracto código C++)

3.3.3.2 Convergencia del algoritmo

El algoritmo PDBM tal y como ha sido descrito se basa en la repetición secuencial de *3* pasos *request, grant, accept*, que conforman una iteración del algoritmo. Cuando varias iteraciones se producen en una misma ranura temporal, los emparejamientos puerto de entrada – retardo ya aceptados en una iteración, se eliminan del grafo para el planteamiento de la siguiente iteración. Por tanto las iteraciones posteriores pretenden agregar pares al emparejamiento acumulado, mejorando las prestaciones. El algoritmo converge en una ranura temporal, cuando tras una iteración no se ha producido ninguna aceptación.

Con el objetivo de estudiar las propiedades de la convergencia del algoritmo, se realizan las siguientes definiciones:

$r^{(i)}(f,b)$ Conjunto de puertos de entrada que realizan una petición hacia el nodo de salida *(f,b)*, *f=0,...,N-1, b=0,...,M-1*, en la iteración *i=0,1,...* Al número de elementos de este conjunto se le denota como $\#r^{(i)}(f,b)$.

$g^{(i)}(f,b)$ Conjunto de puertos de entrada que reciben una aprobación de una petición realizada al nodo de salida *(f,b)*, *f=0,...,N-1, b=0,...,M-1*, en la iteración *i=0,1,...* Al número de elementos de este conjunto se le denota como $\#g^{(i)}(f,b)$.

$a^{(i)}(f,b)$ Conjunto de puertos de entrada que aceptan una aprobación recibida del nodo de salida *(f,b)*, *f=0,...,N-1, b=0,...,M-1*, en la iteración *i=0,1,...* Al número de elementos de este conjunto se le denota como $\#a^{(i)}(f,b)$.

$d^{(i)}(f,b)$ Número de longitudes de onda de transmisión disponibles en la fibra de salida *f*, retardo *b*, *f=0,...,N-1, b=0,...,M-1*, al comienzo de la iteración *i=0,1,...*

El modo de operación del algoritmo en iteraciones sucesivas se describe empleando la nomenclatura anterior de la siguiente manera:

- El número de aprobaciones generadas por un nodo de salida está limitado por el número de peticiones y por el número de longitudes de onda libres.

$$\#g^{(i)}(f,b) = min\left\{d^{(i)}(f,b), \#r^{(i)}(f,b)\right\}, \forall f = 0,...,N-1; b = 0,...,M-1; i = 0,1,... \quad \text{(Ec. 3.8)}$$

- Los puertos de entrada que aceptan una aprobación en una iteración, no generan peticiones en las iteraciones posteriores.

$$r^{(i+1)}(f,b) = r^{(i)}(f,b) \mid \bigcup_{j=0}^{M-1} a^{(i)}(f,b), \forall f = 0,...,N-1; b = 0,...,M-1; i = 0,1,...$$
$$\#r^{(i+1)}(f,b) \leq \#r^{(i)}(f,b) - \#a^{(i)}(f,b), \forall f = 0,...,N-1; b = 0,...,M-1; i = 0,1,... \quad \text{(Ec. 3.9)}$$

- El número de longitudes de onda libres disminuye con las aceptaciones de las aprobaciones otorgadas.

$$\#d^{(i+1)}(f,b) = \#d^{(i)}(f,b) - \#a^{(i)}(f,b), \forall f = 0,...,N-1; b = 0,...,M-1; i = 0,1,... \quad \text{(Ec. 3.10)}$$

De todo ello, se extraen las siguientes propiedades:

- **Prop. 1.** Si en una iteración todas las aprobaciones generadas por un nodo de salida son aceptadas, ese nodo no volverá a generar aprobaciones en iteraciones posteriores.

$$\left[Si\ \#a^{(i)}(f,b) = \#g^{(i)}(f,b) \Rightarrow \#g^{(j)}(f,b) = 0, \forall j > i\right],$$
$$\forall f = 0,...,N-1; b = 0,...,M-1; i = 0,1,... \quad \text{(Ec. 3.11)}$$

Demostración. El número de aceptaciones generadas en una iteración es igual al mínimo entre las peticiones recibidas y las longitudes de onda libres. Si todas las aprobaciones son aceptadas, (1) o bien se agotan las longitudes de onda libres, (2) o bien se agotan los puertos de entrada que generan peticiones hacia ese nodo de salida. Más formalmente:

$$\begin{aligned}
&\#g^{(i+1)}(f,b) = min\left\{d^{(i+1)}(f,b), \#r^{(i+1)}(f,b)\right\} = \\
&min\left\{d^{(i)}(f,b) - \#a^{(i)}(f,b), \#r^{(i+1)}(f,b)\right\} \leq \\
&min\left\{d^{(i)}(f,b) - \#a^{(i)}(f,b), \#r^{(i)}(f,b) - \#a^{(i)}(f,b)\right\} \overset{(hipótesis)}{=} \\
&min\left\{d^{(i)}(f,b) - \#g^{(i)}(f,b), \#r^{(i)}(f,b) - \#g^{(i)}(f,b)\right\} = \\
&min\left\{d^{(i)}(f,b) - min\left\{d^{(i)}(f,b), \#r^{(i)}(f,b)\right\}, \#r^{(i)}(f,b) - min\left\{d^{(i)}(f,b), \#r^{(i)}(f,b)\right\}\right\} = 0
\end{aligned} \qquad \text{(Ec. 3.12)}$$

- **Prop. 2.** Si en una iteración un nodo de salida no genera ninguna aprobación, no lo hará en iteraciones posteriores.

Demostración. Se deduce de la propiedad anterior, ya que en ese caso el número de aceptaciones es igual al de aprobaciones (0).

- **Prop. 3.** Los nodos de salida de la forma *(f,0), f=0,...,N-1*, no generan ninguna aprobación tras la primera iteración (iteración *0*).

$$\#g^{(i)}(f,0) = 0, \forall f = 0,...,N-1; i > 0 \qquad \text{(Ec. 3.13)}$$

Demostración. Todas las aprobaciones generadas para el retardo menor son siempre aceptadas. Por ello, el número de aprobaciones y de aceptaciones es igual en la primera iteración. Aplicando la propiedad *1*, se sabe que es igual a *0* en iteraciones posteriores.

- **Prop. 4.** Los nodos de salida de la forma *(f,b), f=0,...,N-1, b=0,...,M-1* no generan ninguna aprobación tras la iteración *b*.

$$\#g^{(i)}(f,b) = 0, \forall f = 0,...,N-1; i > b \qquad \text{(Ec. 3.14)}$$

Demostración. La propiedad *2* demuestra el caso *b=0*. Los casos posteriores se demuestran de manera similar.

- En la iteración *i=1*, todas las aprobaciones generadas por los nodos *(f,1)* serán aceptadas, ya que serán las de retardo menor (no existen aprobaciones de *b=0* por la propiedad *2*). En consecuencia,

  $$\#g^{(i)}(f,1) = 0, \forall f = 0,...,N-1; i > 1$$

- En la iteración *i=2*, todas las aprobaciones generadas por los nodos *(f,2)* serán aceptadas, ya que serán las de retardo menor (no existen aprobaciones de *b=0,1*). En consecuencia,

  $$\#g^{(i)}(f,2) = 0, \forall f = 0,...,N-1; i > 2$$

- ...

- En la iteración *i=M-1*, todas las aprobaciones generadas por los nodos *(f,M-1)* serán aceptadas, ya que serán las de retardo menor (no existen aprobaciones de *b=0,1,..,M-2*). En consecuencia,

$$\#g^{(i)}(f, M-1) = 0, \forall f = 0,..., N-1; i > M-1$$

- **Prop. 5 (Convergencia del algoritmo PDBM).** El algoritmo PDBM converge, a lo sumo, en *M* iteraciones.

Demostración. Directamente de la propiedad *3*, se observa que en la iteración *i*, los nodos que pueden generar alguna aprobación son los de la forma *(f,b), b≥i*. Para un tamaño de almacenamiento de conmutador en número de retardos *M*, b_{MAX}=*M-1*, por lo que ningún nodo genera aprobaciones tras la *M*-ésima iteración.

Es interesante destacar que esta cota al número de iteraciones para la convergencia es *independiente del tamaño del conmutador nN*, número de fibras de entrada y salida *N*, longitudes de onda por fibra *n*, o cualquier parámetro de tráfico.

3.3.3.2.1 Implementación

La figura 3-10 muestra el diagrama de bloques de una posible implementación electrónica del planificador PDBM. Por claridad, las señales de reloj no han sido incluidas. La implementación se basa en los siguientes módulos:

- **Controladores *Request/Accept*.** Un módulo para cada puerto de entrada *0,...,nN-1* del conmutador. En cada iteración, estos módulos trabajan en dos posibles modos, marcado por la señal de entrada *R/A*:

  - **Modo *Request*.** Para la primera iteración se observan las entradas *P* (presencia o no de un paquete en el puerto de entrada), y *D* (fibra de destino del paquete de entrada). De las *M* señales I/O asociadas a la fibra de salida indicada por *D*, se activan aquellas en las que exista un valor *0* en el registro interno *X*. En consecuencia, estarán activas simultáneamente un máximo de *M* señales I/O, lo que puede ser aprovechado para la simplificación del circuito combinacional. Iteraciones sucesivas dentro de la misma ranura temporal repiten las señales *request* emitidas si el paquete no ha sido aceptado.

  - **Modo *Accept*.** Para la primera iteración, recibe de los puertos I/O las aprobaciones recibidas. Todas las aprobaciones llegarán de una única fibra de salida (de nuevo, estarán simultáneamente activas un máximo de *M* señales), seleccionándose aquélla de menor retardo asociado. En caso de ser aceptada alguna, debe actualizarse apropiadamente el registro interno *X* para futuras ranuras temporales. Asimismo, la señal correspondiente al retardo aceptado debe permanecer activa para ser recibida por el controlador *Grant* adecuado. Las señales de salida *Lost* (paquete perdido), *Delay* (retardo asignado) deben activarse apropiadamente.

  - **Tras cada ranura temporal.** Los registros *X* actúan como registros de desplazamiento, siguiendo la propagación de los paquetes dentro de las líneas de retardo. Por ello, el bit asociado al retardo *i* es sobreescrito por el bit asociado al retardo *i+1*. Al bit *M-1* se le asigna al valor *0*.

- **Controladores *Grant*.** Un módulo para cada nodo *(f,b), f=0,...,N-1, b=0,...,M-1* del conmutador. La sección de evaluación de prestaciones mostrará que el número de módulos *Grant* puede ser muy inferior al de

módulos *Request/Accept* (*M*<<*n*). Cada uno de estos módulos maneja un registro *Y* con el número de longitudes de onda disponibles, un registro puntero de log_2nN bits (no mostrado en la figura), y un registro de sentido de búsqueda de *1* bit (no mostrado en la figura). En cada iteración, en función de los valores del registro puntero y del sentido de búsqueda, se otorgan hasta *Y* aprobaciones de entre las peticiones recibidas. Esto se realiza manteniendo activas dichas señales para que sean recibidas por los módulos *R/A*. Tras cada iteración, este módulo necesita obtener la información de cuáles de sus aprobaciones han sido aceptadas, para la actualización del registro *Y*.

- **Tras cada ranura temporal.** Los registros *Y* deben ser actualizados apropiadamente. Para ello, un nodo graba en su registro *Y*, el contenido del registro del nodo correspondiente al retardo inmediatamente superior, $0 \rightarrow Y(f,M-1) \rightarrow Y(f,M-2) \rightarrow ... \rightarrow Y(f,1) \rightarrow Y(f,0), \quad \forall f=0,...,N-1$, empleando las señales *Prev* y *Next*.

- **Red de interconexión.** Cada iteración del algoritmo involucra secuencialmente (1) señales *request* desde los módulos R/A hacia los módulos (2) aprobaciones desde los módulos *grant* hacia los módulo R/A, y (3) señales de aceptación realizadas, en el mismo sentido que (1).

Las velocidades de planificación necesarias, requieren la implementación de cada uno de los módulos mediante un circuito combinacional. El diseño de cada uno de estos circuitos no ha sido abordado en esta tesis doctoral. La complejidad se estima intuitivamente similar a la de los módulos constituyentes de planificadores VOQ comerciales, como *iSLIP*.

3.4 Evaluación de prestaciones

En esta sección se presentará la evaluación de prestaciones de la arquitectura IB-WR, llevada a cabo mediante simulaciones sobre la herramienta de libre distribución OMNET [Var99]. La primera serie de simulaciones realizadas se centran en los algoritmos de planificación secuenciales SHWP (figura 3-3) y SCWP (figura 3-6), ampliando los resultados presentados en [Pav03-1]. Las prestaciones del conmutador se comparan con las de las arquitecturas de colas a la salida (óptimo alcanzable), observándose una gran aproximación de las prestaciones del conmutador IB-WR respecto a esta cota superior. La no viabilidad de estos algoritmos de planificación secuenciales para las velocidades de planificación requeridas, motivaron el estudio de esquemas implementables. Este esfuerzo ha fructificado en el algoritmo PDBM, para el modo de operación SCWP. En este algoritmo se centra la segunda serie de simulaciones realizadas. Los resultados ofrecen unas prestaciones casi exactamente iguales a las del algoritmo secuencial, muy cercanas a la cota impuesta por las prestaciones de las arquitecturas de colas a la salida.

Los criterios aplicados para la duración de las simulaciones han sido:

- Para la evaluación de la probabilidad de pérdida: tiempo de simulación dos órdenes de magnitud mayor a la probabilidad de pérdida a evaluar (10^{-6} en este capítulo), o la pérdida de al menos *100* paquetes en el puerto de salida *0* (SHWP) o fibra de salida *0* (SCWP). Fin anticipado de la simulación en el caso de haberse alcanzado el valor de *1000* paquetes perdidos por el puerto/fibra de salida *0* (SHWP/SCWP). Tiempo transitorio 10^4 ranuras temporales.

- Para la evaluación del retardo/convergencia del algoritmo: tiempo de simulación de 10^8 ranuras temporales. Tiempo transitorio 10^4 ranuras temporales.

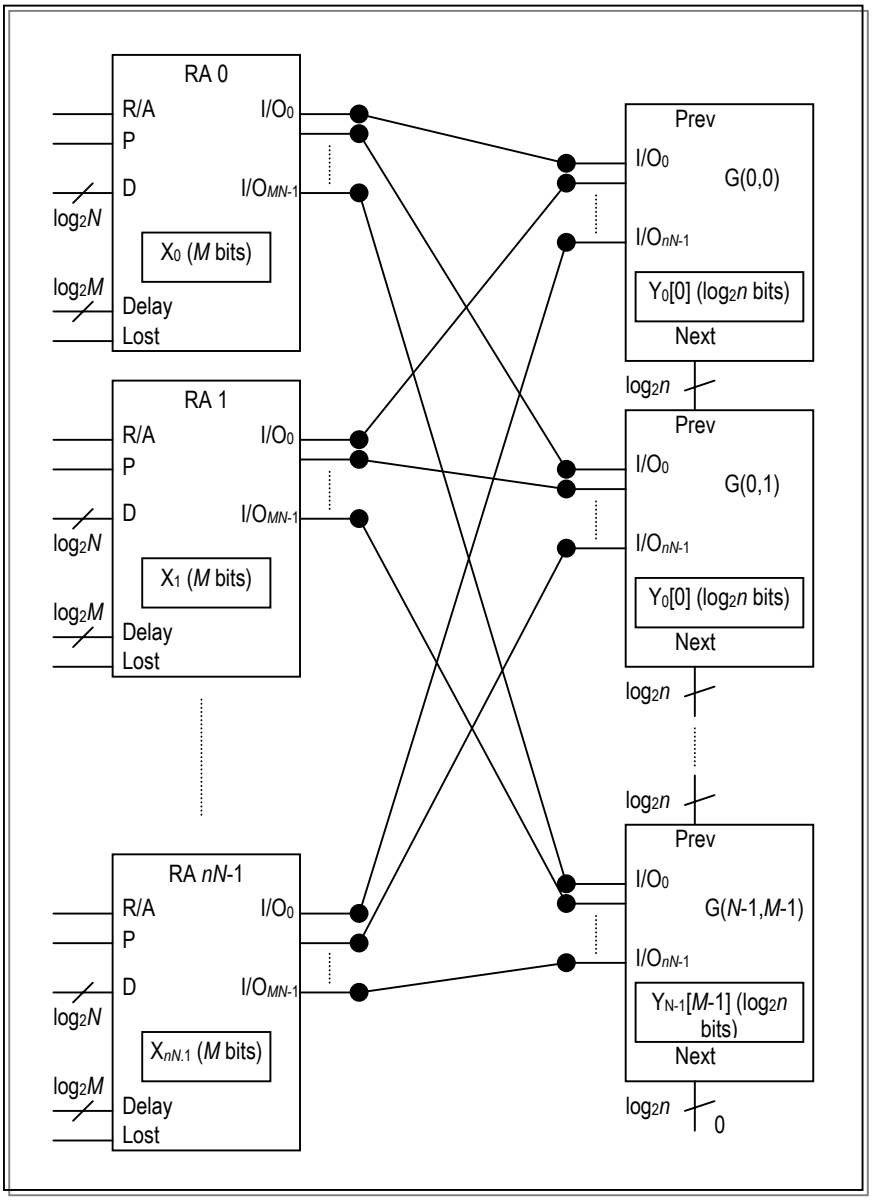


Figura 3-10. Diagrama de implementación del planificador PDBM

3.4.1 Algoritmo secuencial SHWP y SCWP

La figura 3-11-(a,b,c) muestra las probabilidades de pérdida de paquete para un conmutador IB-WR como el mostrado en la figura 3-1-(b), con *N* fibras de entrada y salida, *n* longitudes de onda por fibra, *M* retardos, tráfico de entrada Bernouilli uniforme de parámetro ρ*=0.8*. Las gráficas ilustran la variación de la probabilidad de pérdida de paquete en función del tamaño del buffer, para distintas longitudes de onda por fibra *n*. Se comparan las prestaciones de los algoritmos secuenciales en la arquitectura IB-WR, con respecto a las prestaciones para un conmutador con colas a la salida (capítulo *2*), en modo SHWP y en modo SCWP. Los resultados obtenidos muestran cómo las prestaciones se acercan al óptimo alcanzable, también en el modo de

operación SCWP. Como ejemplo, en todos los casos se ha obtenido una probabilidad de pérdida, a lo sumo, un orden de magnitud mayor respecto a los conmutadores con colas a la salida.

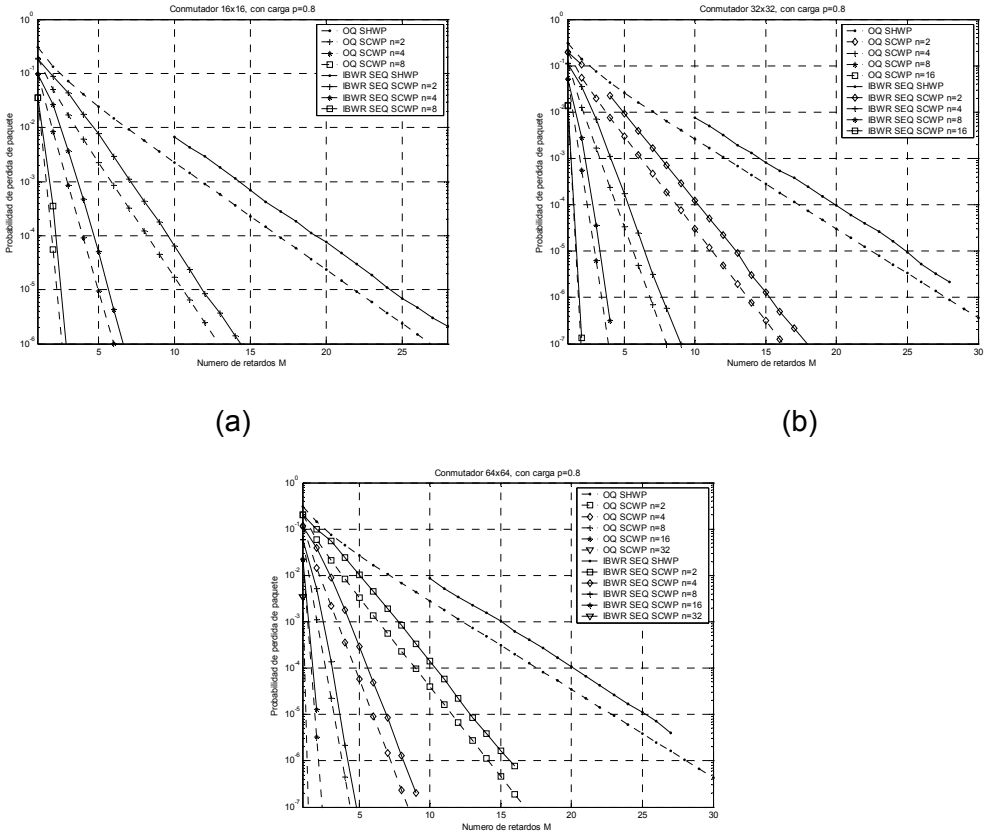


Figura 3-11. Comparación de la probabilidad de pérdida de paquete respecto a número de retardos (*M*), para un conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida vs. IB-WR planificación secuencial, carga ρ*=0.8*, *n={2,4,8,16,32}*, y tamaños de conmutador *nN* (a) *16x16*, (b) *32x32*, (c) *64x64*

En la figura 3-12 se muestra la comparación del impacto de la carga de entrada en el retardo medio de paquete, normalizado en número de ranuras temporales, y un tamaño de buffer que implique una probabilidad de pérdida de paquete despreciable. Los resultados para cargas bajas son prácticamente idénticos entre los conmutadores con colas a la salida y la arquitectura IB-WR. De nuevo, se observa una fuerte mejora en el conmutador SCWP respecto a la versión SHWP. Para cargas altas, y menor número de longitudes de onda por fibra (*n*), el retardo medio de la arquitectura IB-WR se separa en mayor medida del óptimo. Obsérvese sin embargo que, para el valor *n=8,* incluso en las cargas más altas, el retardo medio no supera el valor de *3* ranuras temporales.

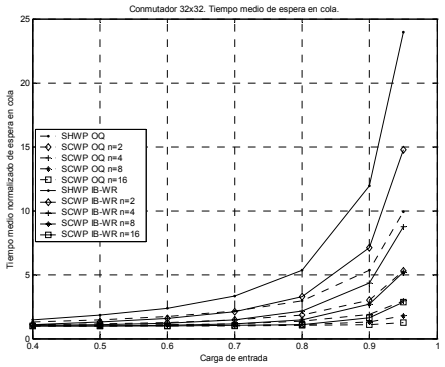


Figura 3-12. Comparación del retardo medio normalizado de paquete respecto a la carga de entrada (ρ),para un conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida vs. IB-WR planificación secuencial, *n={2,4,8,16}*, *nN=32*

La figura 3-13 muestra la variación de la desviación típica del retardo, en las mismas condiciones que la gráfica anterior. En este caso, se aprecia una gran diferencia en la evolución de esta parámetro. En las arquitecturas con colas a la salida la desviación típica es siempre creciente con la carga, y en mayor medida para valores bajos de *n*. Sin embargo, los valores obtenidos con el planificador secuencial IB-WR muestran una desviación típica casi constante, con valores, en general, menores (especialmente para cargas altas) que las arquitecturas de colas a la salida. Por tanto, en estas cargas altas, el planificador secuencial IB-WR asigna retardos en media mayores, pero con muy poca variabilidad entre distintos paquetes: es decir, un comportamiento en media peor, aunque más predecible. La razón de esta menor variabilidad debe buscarse en el hecho de que en al arquitectura IB-WR, los efectos de las cargas altas afectan a los paquetes en los puertos de entrada, a través de los vectores X_i, *i=0,...,nN-1,* independientemente del puerto o fibra de salida demandado. Sin embargo, en los conmutadores de colas a la salida, las fluctuaciones en los destinos de los paquetes entrantes, pueden producir situaciones temporales de saturación de unos puertos/fibras de salida (que asignan retardos altos), y mayor descarga de otros puertos/fibras de salida (que asignan retardos bajos). Estas situaciones se traducen en una mayor variabilidad media en los retardos asignados.

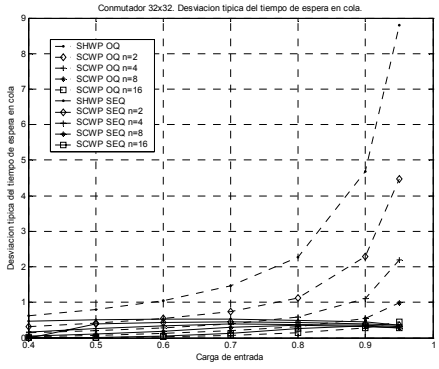


Figura 3-13. Comparación de la desviación típica del retardo normalizado de paquete respecto a la carga de entrada (ρ),para un conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida vs. IB-WR planificación secuencial, *n={2,4,8,16}*, *nN=32*

3.4.2 Algoritmo PDBM

Para la evaluación de prestaciones del planificador PDBM (*Parallel Desynchronized Block Matching*), se seguirá una misma secuencia de pruebas que las realizadas para el planificador secuencial. La figura 3-14-(a,b,c) muestra las probabilidades de pérdida de paquete para un conmutador IB-WR, con *N* fibras de entrada y salida, *n* longitudes de onda por fibra, *M* retardos, tráfico de entrada Bernouilli uniforme de parámetro ρ*=0.8*. De nuevo se muestran, a modo de comparación, las prestaciones para un conmutador con colas a la salida (capítulo *2*), en modo SCWP. Los resultados obtenidos son prácticamente idénticos a los del planificador secuencial, y por tanto, aceptablemente semejantes al óptimo marcado por los conmutadores de colas a la salida. En concreto, se ha obtenido también en todos los casos una probabilidad de pérdida a lo sumo un orden de magnitud mayor respecto al óptimo alcanzable.

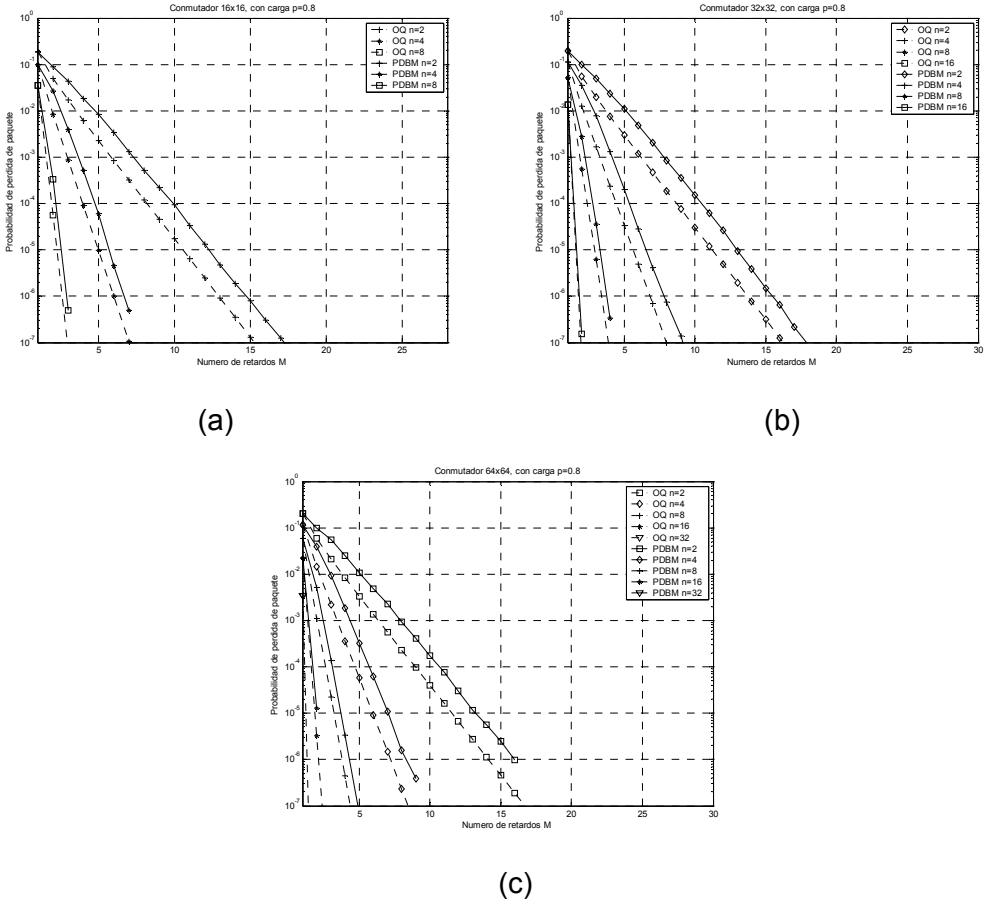


Figura 3-14. Comparación de la probabilidad de pérdida de paquete respecto a número de retardos (*M*), para un conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida vs. IB-WR planificación PDBM, carga ρ*=0.8*, *n={2,4,8,16,32}*, y tamaños de conmutador *nN* (a) *16x16*, (b) *32x32*, (c) *64x64*

En la figura 3-15 se comparan los valores de retardo medio normalizado en función de la carga de entrada, para un número de retardos que hacen despreciable la probabilidad de pérdida de paquete. Para cargas bajas, los resultados son prácticamente idénticos entre los conmutadores con colas a la salida, el planificador secuencial, y el planificador PDBM. Para cargas altas, los retardos medios se alejan

del óptimo en ambos casos. Comparando ambos planificadores IB-WR en cargas altas, se aprecia una leve diferencia a favor del planificador secuencial en valores bajos de *n,* que se anula para *n≥8*. Los reducidos valores de retardo obtenidos en esta situación incluso en cargas altas, muestran el aprovechamiento que consigue este algoritmo de la ganancia de multiplexación posible en el modo SCWP.

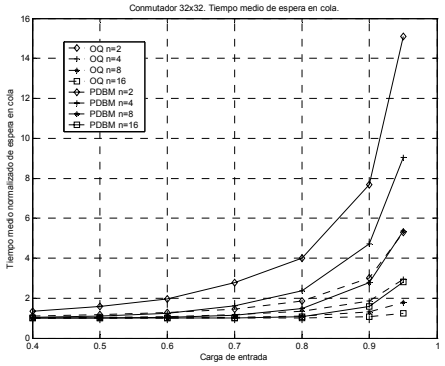


Figura 3-15. Comparación del retardo medio normalizado de paquete respecto a la carga de entrada (ρ),para un conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida vs. IB-WR planificación PDBM, *n={2,4,8,16}*, *nN=32*

La figura 3-16 muestra la variación de la desviación típica del retardo, en las mismas condiciones que la gráfica anterior. La evolución es similar a la encontrada para el planificador secuencial: la variabilidad del retardo es mínima para valores altos de la carga de entrada, aún menor con valores altos de *n*. Para cargas medias y bajas, y valores de *n=2* y *n=4*, los valores de la desviación siguen siendo bajos (menores a una ranura temporal), aunque peores que los obtenidos para el planificador secuencial.

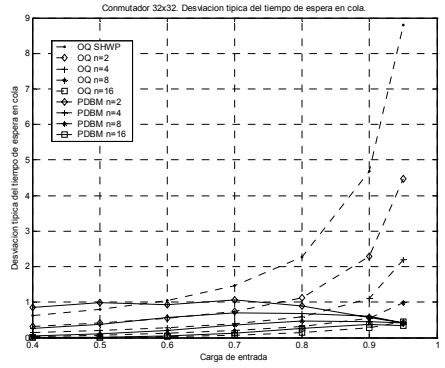


Figura 3-16. Comparación de la desviación típica del retardo normalizado de paquete respecto a la carga de entrada (ρ),para un conmutador OPS *nNxnN* SHWP y SCWP con emulación de colas a la salida vs. IB-WR planificación PDBM, *n={2,4,8,16}*, *nN=32*

3.4.2.1 Convergencia del planificador

En la sección anterior se demostraron varias propiedades relativas a la convergencia del algoritmo PDBM. Entre ellas, la convergencia del mismo en un máximo de *M* iteraciones (número de retardos del conmutador). Los estudios realizados muestran, sin embargo, que la convergencia se produce realmente en un número mucho menor de iteraciones.

Para ilustrar este hecho se ha confeccionado la tabla 3-1, que muestra la probabilidad de que un paquete aceptado, *no* lo sea en la primera iteración, para un tamaño de conmutador *nN=32, n= {2,4,8,16}*, distintos valores de la carga de entrada ρ, y tamaño de buffer *M* que hace despreciable la probabilidad de pérdida de paquete. Se observa que en gran parte de las situaciones, el algoritmo converge en una iteración, especialmente para valores altos del parámetro *n*. Los peores resultados se dan para *n=2* y *n=4*. En estos casos, la probabilidad de ser aceptado en la primera iteración es menor para las cargas más bajas y las cargas más altas. Para entender el origen de este (leve) efecto, profundizamos en la situación evaluada, con un ejemplo de dos paquetes involucrados y dos aprobaciones:

- Un paquete p_1 recibe en la primera iteración dos aprobaciones para su fibra de salida, una del retardo M_1 y otra del retardo M_2.
- El paquete p_2, destinado a la misma fibra de salida no recibe ninguna aprobación, por (1) encontrarse situado tras el paquete p_1 según el orden marcado por los punteros *round-robin* de M_1 y M_2, y (2) porque estos retardos tienen únicamente una longitud de onda libre.
- En la segunda iteración, p_2 recibe la aprobación no aceptada por p_1 en la primera iteración.

Esta situación es fruto de la combinación de efectos que se acentúan en cargas altas y efectos que se acentúan en cargas bajas, lo que permite explicar la variación del comportamiento con la carga:

- No existe más que una longitud de onda libre en M_1 y en M_2 (efecto acentuado en cargas altas, y valores bajos de *n*).
- Existen dos (o más) paquetes destinados a la misma fibra de salida p_1 y p_2 (efecto acentuado en cargas altas).
- Entre los puertos de entrada apuntados por los punteros *round-robin* de M_1 y M_2, no existen más paquetes destinados a la misma fibra de salida (efecto acentuado a cargas bajas).

It >1	**ρ=0.4**	**ρ=0.5**	**ρ=0.6**	**ρ=0.7**	**ρ=0.8**	**ρ=0.9**	**ρ=0.95**
n=2	$1.62\cdot10^{-4}$	$1.547\cdot10^{-4}$	$7.6\cdot10^{-5}$	$4.244\cdot10^{-5}$	$4.34\cdot10^{-7}$	$1.99\cdot10^{-6}$	$7.27\cdot10^{-4}$
n=4	$3.30\cdot10^{-6}$	$1.12\cdot10^{-5}$	$1.21\cdot10^{-6}$	$2.50\cdot10^{-7}$	$1.95\cdot10^{-7}$	$3.05\cdot10^{-7}$	$1.97\cdot10^{-7}$
n=8	0	$1.25\cdot10^{-8}$	0	0	0	$5.56\cdot10^{-8}$	$3.95\cdot10^{-8}$
n=16	0	0	0	0	0	0	$1.31\cdot10^{-8}$

Tabla 3-1. Probabilidad de paquete aceptado en una iteración, condicionado a que el paquete no se pierde, para un conmutador *nN=32, n={2,4,8,16}, ρ={0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95}*.

Los resultados obtenidos en posteriores iteraciones indican que en todos los casos, salvo en el punto de simulación *{n=2, ρ=0.95}*, el algoritmo converge en *2* iteraciones (cuando no lo ha hecho en *1*). En el caso anteriormente indicado, se ha observado que el algoritmo converge en *4* iteraciones, con una probabilidad de paquete aceptado en la *3º* iteración y *4º* iteración despreciables.

Estos resultados indican para los casos observados, que el número de iteraciones necesarias del algoritmo PDBM manteniendo las prestaciones del mismo, es de *1* para valores de $n \geq 8$, y de *2* en el resto de los casos observados. Lógicamente, estos valores resultan de interés por la simplificación de la implementación que conllevan.

3.5 Conclusiones

Este capítulo se ha centrado en el estudio de la arquitectura de conmutación *Input-Buffered Wavelength-Routed swtich* (IB-WR), de interés por su menor complejidad *hardware* respecto a otras arquitecturas (como se deduce de la tabla 3-2). En un primer paso, se ha realizado la adaptación de la arquitectura para la aplicación de los modos de operación SHWP y SCWP, no considerados en la propuesta original [Zho98]. La aplicación de ambos modos de operación ha puesto en evidencia un problema peculiar de planificación para la asignación del retardo, no abordado en la literatura, hasta la propuesta realizada en [Pav03-1]. Debemos destacar que lo descrito en este capítulo respecto a la planificación del conmutador, en ambos modos de operación, es una contribución original fruto del trabajo de esta tesis doctoral.

	FWC	**Puertas ópticas**	**TWC (rango sintoniz. máx)**	**Delay loops**	**Tamaño máx. AWG**
KEOPS switch	2nN	$MnN+n^2N^2$	0	1...M	0
OB-WR switch	nN	n^2N^2	nN (max(nN,M))	1...M	max (nN,M)
Space switch	0	nN^2M	nN (n)	N·(1...M)	0
IB-WR	nN	0	2nN max(nN,M')	(1...M')	max (nN,M')

Tabla 3-2. Cómputo de componentes *hardware*

En un primer paso, se han formalizado los problemas de planificación en ambos modos de operación, como problemas de programación entera dinámica. La imposibilidad de resolución general de este tipo de problemas, ha llevado a proponer la simplificación de los mismos, entablando la optimización independiente en ranuras temporales sucesivas. El resultado ha sido un problema de optimización entera de dimensión finita, cuya función objetivo se ha diseñado para buscar la asignación de retardo más favorable para ranuras temporales sucesivas. Como se ha mostrado en [Pav03-1], este problema de planificación a resolver en cada ranura temporal puede expresarse como un problema de maximización en grafos bipartitos.

En un segundo paso, nuestro interés se enfoca en encontrar algoritmos que sean capaces de resolver el problema de asignación de retardo en conmutadores IB-WR, a las velocidades requeridas. En este punto, la equivalencia encontrada del problema de planificación con el de maximización del emparejamiento en grafos bipartitos, constituye sin duda un beneficio. La razón estriba en que este tipo de problemas aparecen también en la planificación de conmutadores electrónicos VOQ (*Virtual Output Queueing*). El estudio de algoritmos que resuelvan velozmente y de manera eficiente este tipo de problemas, es un campo de estudio intenso en estos conmutadores electrónicos. Esto nos ha hecho encaminar nuestro estudio hacia las similitudes y diferencias entre el problema de planificación IB-WR, y el problema de planificación de los conmutadores electrónicos VOQ. El resultado ha sido la propuesta del algoritmo PDBM (*Parallel Desynchronized Block Matching*) para la planificación del conmutador IB-WR en modo SCWP, aprovechando los trabajos en la línea de

algoritmos VOQ de asignación paralela, como *iSLIP* o *RDSRR*. Las prestaciones obtenidas para el algoritmo PDBM se encuentran muy cerca del óptimo marcado por los conmutadores con colas a la salida. Asimismo, se ha conseguido demostrar la convergencia del mismo en un número de iteraciones independiente del tamaño de conmutador, y una convergencia práctica en *1* ó a lo sumo *2* iteraciones para los casos probados, lo que constituye una ventaja por la simplificación en el *hardware* del circuito planificador que conlleva.

Capítulo 4. Arquitecturas OPS de gran escala

4.1 Introducción

Este capítulo se centra en el estudio de las arquitecturas de Conmutación Óptica de Paquetes de gran escala, es decir, arquitecturas de alto número de puertos de entrada y salida. La necesidad de estas arquitecturas (no prevista en el corto plazo) se encuentra en la aplicación de OPS a redes troncales DWDM (*Dense Wavelength Division Multiplexing*), con un número de longitudes de onda por fibra elevado.

El crecimiento en número de puertos, usualmente implica un crecimiento no sostenible de la complejidad y requisitos de los dispositivos ópticos de las arquitecturas OPS convencionales. Se hace necesario, por tanto, la aplicación de técnicas que permitan diseños viables, y a ser posible escalables. En este capítulo, se comenzará realizando un repaso de las propuestas realizadas hasta la fecha, recopilando fundamentalmente el trabajo publicado en [Pav02]. Posteriormente, se propondrán dos estrategias de crecimiento basadas en el aprovechamiento del efecto *knock-out* [Yeh87], presentadas, en [Pav03-4]. Ambas estrategias interconectan una etapa de conmutación sin memoria, con módulos de menor tamaño basados en arquitecturas OPS. La propuesta OFD (*Output Fiber Distributed*) asocia uno de estos módulos para cada fibra de salida. La propuesta OWD (*Output Wavelength Distributed*), asocia un módulo por cada longitud de onda de transmisión. Se abordará la evaluación de pérdidas *knock-out* para ambas propuestas, bajo el modo de operación SCWP. La evaluación de estas pérdidas para la arquitectura OWD depende del algoritmo de planificación SCWP empleado. Se presenta el análisis matemático que permite la obtención de las pérdidas *knock-out*, para tráfico independiente e idénticamente distribuido, bajo el algoritmo de planificación SCWP uniforme descrito en el capítulo *2*. Se demuestra asimismo, la cota superior que permite la eliminación de este tipo de pérdidas, mostrando su excelente precisión para el dimensionamiento de conmutadores en cargas altas, y el escenario DWDM previsto.

El capítulo finaliza con una comparativa de costes de las distintas propuestas, aprovechando los métodos de evaluación descritos en esta tesis doctoral.

4.2 Estado de la técnica

4.2.1 Output-Buffered Wavelength-Routed Switch

El conmutador OB-WR en su arquitectura original fue descrito en el capítulo *2* (figura 2-2-(a)). La limitación al crecimiento de esta arquitectura se encuentra en que el rango de longitudes de onda requerido por los dispositivos TWC, debe ser igual a *max(N,M)*, donde *N* es el número de puertos de salida del conmutador y *M* el número de retardos. Para conmutadores de gran escala, el parámetro dominante es *N*, que determina el rango de los TWCs, lo que hace inviable esta estrategia según el estado de la tecnología expresado en el capítulo *1*.

En [Zho98] se propuso una arquitectura OB-WR de gran escala, que permite eliminar esta dependencia, disminuyendo asimismo el tamaño de los dispositivos AWG, a costa de aumentar el número de kilómetros de líneas de retardo necesarios. Para ello, el módulo de almacenamiento se convierte en un conjunto de *R* módulos de menor tamaño, como muestra la figura 4-1-(a). La viabilidad de esta estrategia depende ahora de la construcción de etapas de conmutación espacial basadas en puertas ópticas de tamaño *NxN*. En [Zho98] se propone el diseño de estas etapas mediante una red de Benes de puertas ópticas que requiere $2N(2\log_2 N-1)$ dispositivos frente a los N^2 de la configuración en *crossbar*.

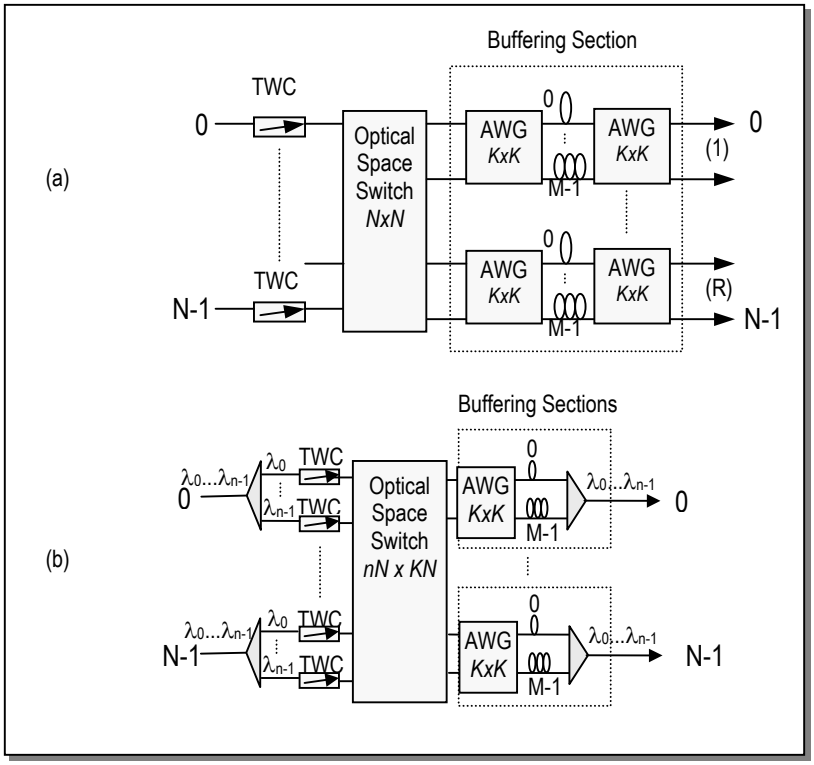


Figura 4-1. (a) Arquitectura OB-WR de gran escala [Zho98], (b) adaptación WDM propuesta

El conmutador OB-WR de gran escala propuesto en [Zho98] debe adaptarse al escenario WDM para la aplicación de los modos de operación SHWP/SCWP. La adaptación propuesta de la arquitectura OB-WR de gran escala se muestra en la figura 4-1-(b).

- Se ha añadido una etapa de demultiplexación WDM en las fibras de entrada, que separan los paquetes entrantes por su longitud de onda, en *n* fibras distintas.
- Se ha asociado una sección de almacenamiento a cada fibra de salida. Esta decisión permite la simplificación de la arquitectura, sustituyendo el AWG final por un multiplexor, y eliminando la necesidad de los dispositivos FWC de salida. En el diseño OB-WR adaptado, propuesto en el capítulo *2*, el puerto de entrada al módulo de almacenamiento debía ser igual al puerto de salida del conmutador, mientras que el retardo de paquete determinaba la longitud de onda de conversión. Según la adaptación propuesta para la arquitectura de gran escala, el puerto de entrada a un módulo de

almacenamiento no tiene ningún condicionante, ya que todos los paquetes entrantes al mismo módulo de almacenamiento serán finalmente transmitidos por la misma fibra de salida. Por ello:

- La longitud de onda de conversión será igual a la longitud de onda final de transmisión del paquete, eliminando la necesidad de una etapa final de convertidores FWC. Esta longitud de onda vendrá determinada por el OPP al que pertenece el paquete (SHWP), o por el algoritmo de planificación (SCWP).

- El grado de libertad disponible en cuanto al puerto de entrada al módulo de almacenamiento se emplea para seleccionar el retardo de paquete, siguiendo la regla impuesta por el funcionamiento de los dispositivos AWG: *input_port_to_Buffering_Section = (wavelength – delay) mod K.*

En definitiva, estas simplificaciones en la adaptación WDM atañen únicamente al *hardware*. La planificación SHWP y SCWP de los conmutadores sigue siendo la misma, obteniéndose las mismas prestaciones óptimas del conmutador.

Es importante destacar que el tamaño de los dispositivos AWG de la etapa de almacenamiento debe ser igual a *K=max(n,M)*. Esto implica el diseño de etapas de conmutación espacial de tamaño *nNxnN* para *n>M*, y de tamaño *nNxMN* para *M>n*. Ambas secciones de almacenamiento se muestran en la figura 4-2-(a) y 4-2-(b). El caso *n>M* es el más habitual en el escenario de aplicación de gran escala, especialmente en el modo de operación SCWP, y se trata asimismo del caso más favorable en esta arquitectura, al conllevar etapas de conmutación espacial *nNxnN* cuadradas, más fáciles de construir y controlar.

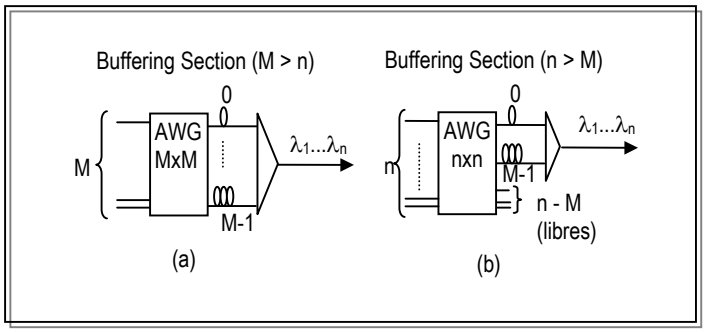


Figura 4-2. Sección de almacenamiento del conmutador WDM OB-WR de gran escala, (a) *M≥n*, (b) *n≥M*

4.2.2 KEOPS

El crecimiento en número de puertos de la arquitectura de conmutación KEOPS tiene dos limitaciones fundamentales: el número de puertas ópticas necesarias, igual a $NM+N^2$ (donde *N* es el número de puertos de entrada y salida, y *M* el número de retardos), y las pérdidas de potencia debidas al funcionamiento *broadcast*, proporcionales a NM^2. El límite *hardware* barajado por los investigadores del proyecto es de arquitecturas de hasta *32x32* puertos y pocas decenas de retardos [Raf00-1][Raf00-2].

El diseño de conmutadores de gran escala fue estudiado dentro del proyecto KEOPS, mediante la interconexión en varias etapas de arquitecturas KEOPS de menor tamaño. Los cálculos realizados mostraban que la interconexión de conmutadores KEOPS era viable sin pérdida de sensibilidad de los detectores, si se

realizaba una regeneración de la señal adecuada a la entrada de cada módulo conmutador. El esfuerzo investigador se enfocó en la evaluación mediante simulación y análisis de teoría de colas, de arquitecturas de interconexión de dispositivos KEOPS según una topología de red de Clos de 3 etapas [Raf00-1][Raf00-2][Jac96] (figura 4-3). Este proceso de evaluación se benefició de los resultados previos en la aplicación de las redes de Clos para arquitecturas de conmutación electrónica ATM.

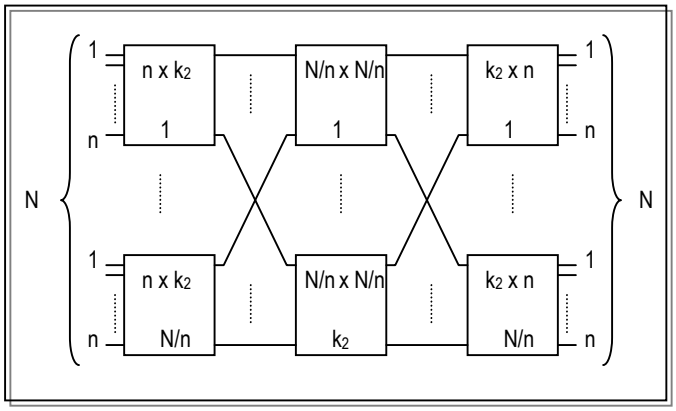


Figura 4-3. Red de Clos de tres etapas

Los parámetros a dimensionar en una red de este tipo son los requisitos de almacenamiento de los módulos en cada una de las tres etapas, y el número de módulos (*switching elements, SE*) de segunda etapa. Las arquitecturas de Clos se denominan "sin expansión" cuando el número de SE en la segunda etapa es igual al número de entradas de un SE de primera etapa ($k_2 = n$). Esto implica la utilización de arquitecturas de conmutación KEOPS cuadradas en todas la etapas. Si, en caso contrario, $k_2 \geq n$, la arquitectura de Clos resultante se conoce como "con expansión".

Las arquitecturas sin expansión requieren mayor número de líneas de retardo en las tres etapas para alcanzar una probabilidad de pérdida de paquete aceptable. Existen tres grandes alternativas estudiadas dentro del proyecto KEOPS para abordar esta limitación:

- La utilización de arquitecturas con expansión, dimensionando el número de SE en la etapa intermedia conjuntamente a los tamaños de buffer en las tres etapas, para obtener el deseado compromiso coste/prestaciones. Esta alternativa obliga a realizar concentración de tráfico en la última etapa, donde los requisitos de almacenamiento son mayores. Un método de dimensionamiento para este esquema de interconexión fue propuesto en [Raf00-1]. El interés de esta alternativa se encuentra en que el comportamiento de la red de tres etapas se asemeja al de un conmutador con colas a la salida.
- Distribuir la carga de tráfico de entrada, entre *R* planos idénticos. Cada plano se diseña como una red de Clos sin expansión. La menor carga por plano ($<1/R$) permite implementar bufferes de menor número de retardos en cada SE. Por otro lado, el tráfico proveniente de cada uno de los *R* planos se concentra en un multiplexor con memoria, que debe ser dimensionado. Un método de dimensionamiento para este esquema de interconexión fue propuesto en [Jac96].
- Otra opción estudiada ha sido la introducción de técnicas de control de flujo entre etapas, de manera similar a lo considerado en conmutadores electrónicos multietapa [Tur88]. En el escenario electrónico esto se consigue mediante el

almacenamiento de paquetes en una etapa, si una señal de control de la etapa posterior informa sobre saturación de sus bufferes (*back-pressure*). Esta técnica puede ser aplicada con los conmutadores KEOPS *broadcast-and-select*, ya que un mismo paquete se encuentra disponible *M* ranuras temporales consecutivas en todos los puertos de salida, lo que no es cierto para la mayoría de arquitecturas de conmutación OPS basadas en líneas de retardo. Sin embargo, el tiempo en el que un paquete puede permanecer almacenado a la espera está limitado a *M* ranuras temporales, limitación que no existe en los conmutadores electrónicos. Por ello, en [Raf00-1] fueron estudiadas distintas técnicas de control de flujo entre etapas, adaptadas a esta peculiaridad.

Las tres alternativas resumidas en los puntos anteriores pueden ser combinadas para lograr las prestaciones objetivo.

4.2.2.1 Adaptación WDM

Las técnicas para el escalado del conmutador KEOPS resumidas en los puntos anteriores, contemplan la utilización del conmutador original, como el mostrado en la figura 2-1-(a). Sin embargo, la adaptación al entorno WDM necesaria para la aplicación de los modos de operación SHWP/SCWP implica la necesidad de cambios. La figura 4-4 muestra los cambios propuestos para un conmutador en tres etapas monoplano con expansión:

- Se asocia un módulo *switch element* para cada fibra de entrada. Esto nos permite eliminar los dispositivos FWC de la primera etapa (cada paquete ya es recibido en una longitud de onda distinta).
- En la segunda etapa se emplea un conmutador KEOPS sin adaptación WDM para cada *switch element*.
- Se asocia un módulo *switch element* en la tercera etapa para cada fibra de salida. En cada uno de ellos, se añade una etapa de multiplexación con *n* dispositivos FWC de manera similar a la adaptación WDM propuesta en el capítulo *2*.

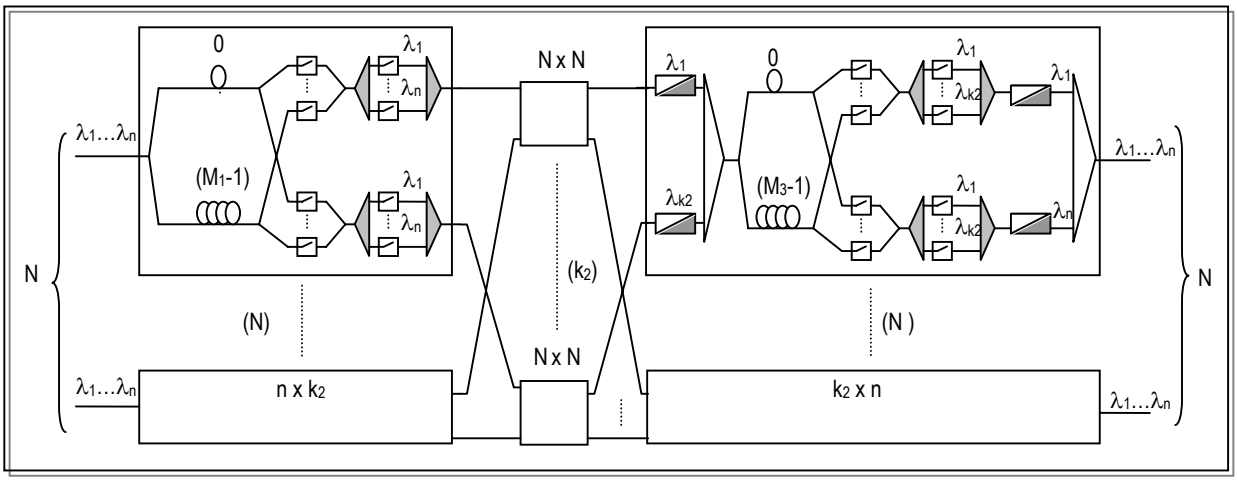


Figura 4-4. Adaptación WDM de una red de dispositivos KEOPS interconectados según una topología de red de Clos

La aplicación de los modos de operación SHWP/SCWP a esta arquitectura de gran escala afecta únicamente a la planificación de la última etapa. En modo SHWP, los paquetes demandan en la tercera etapa un puerto de salida concreto, determinado por la longitud de onda asociada al OPP al que pertenecen. En modo SCWP, los paquetes pueden ser transmitidos por cualquier puerto de salida. Los algoritmos de planificación SHWP y SCWP empleados por los conmutadores de tercera etapa son los mismos a los descritos en el capítulo *2*.

A continuación, se describe un método de dimensionamiento de los parámetros k_2, M_1, M_2, M_3 para una arquitectura como la mostrada en los modos de operación SHWP/SCWP, para tráfico Bernouilli uniforme, basado en el método propuesto en [Raf00-1]:

1. Para la primera etapa, el tamaño de buffer M_1 y el factor de expansión k_2 debe dimensionarse de manera conjunta de tal manera que la probabilidad de pérdida de paquete sea despreciable (por ejemplo, un orden de magnitud menor a nuestro objetivo de diseño). Los cálculos en esta etapa se realizan para ambos modos de operación aplicando análisis de prestaciones de un conmutador con colas a la salida ante tráfico uniforme [Hlu88].
2. La segunda y tercera etapa reciben tráfico a ráfagas, debido a la correlación añadida por las etapas anteriores. Sin embargo, en [Raf00-1] se muestra que, para arquitecturas con expansión, esta correlación puede ser despreciada, asumiendo tráfico Bernouilli uniforme a la entrada de la segunda y tercera etapas:
   - Los elementos de segunda etapa son de tamaño *NxN*, y reciben un tráfico por puerto de entrada $\rho_2 = \rho_1 \cdot n/k_2 < \rho_1$. El número de líneas de retardo debe calcularse siguiendo el mismo análisis descrito en [Hlu88], aunque en [Raf00-1] se recomienda dimensionar $M_2=M_1$, lo que proporciona una probabilidad de pérdida incluso menor a la obtenida en la primera etapa.
   - El dimensionado del número de líneas de retardo de los elementos de tercera etapa depende del modo de operación del conmutador. El algoritmo de planificación SHWP y SCWP de cada uno de estos dispositivos será igual al descrito en el capítulo *2*. El cálculo del número de retardos necesario debe realizarse considerando una carga por puerto de entrada $\rho_3=\rho_2=\rho_1 n/k_2<\rho_1$. Para el modo SHWP, se obtiene una profundidad de buffer mayor, debido a la concentración de tráfico existente. Para el modo SCWP, este efecto es invertido por la multiplexación estadística obtenida (mayor para valores altos de *n*). Como ejemplo, un valor de M_3 de 2 retardos proporciona una probabilidad de pérdida $<10^{-14}$ para una arquitectura de parámetros *n=16,* k_2*=20, nN=1024* con carga *0.8*.

4.2.3 Frontiernet

4.2.3.1 Arquitectura Frontiernet

Dentro de la investigación en Conmutación Óptica de Paquetes llevada a cabo en Japón, destacan las propuestas procedentes de los laboratorios *NTT Optical Network Systems*, centradas en la arquitectura de conmutación FRONTIERNET. De esta arquitectura fue construido un prototipo de *16x16* puertos de entrada y salida, y duración de la ranura temporal *51.2* ns [Sas93][Sas97][Yam98]. El conmutador original FRONTIERNET puede verse en la figura 4-5.

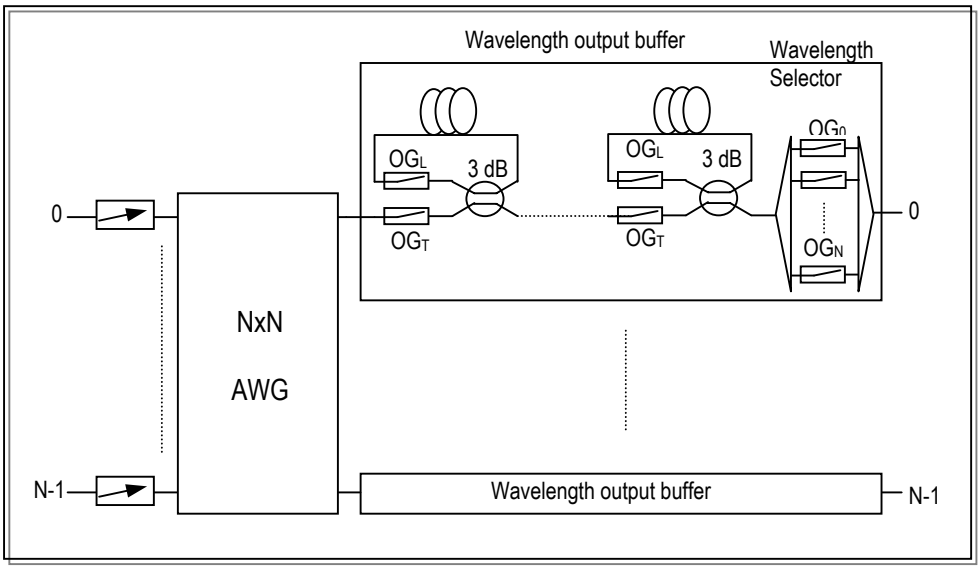


Figura 4-5. Arquitectura original FRONTIERNET

Un conmutador FRONTIERNET *NxN* interconecta *N* puertos de entrada y *N* puertos de salida. Para ello, son necesarios *N* dispositivos TWC con un rango de *N* longitudes de onda sintonizables. El puerto de salida de cada paquete determina la longitud de onda de conversión. El dispositivo AWG encamina los paquetes hacia un dispositivo de almacenamiento WDM, que puede recibir hasta *N* paquetes simultáneos (a distintas longitudes de onda). El diseño de los dispositivos WOB (*Wavelength Output Buffer*) propuestos, no se basa en líneas de retardo, sino en líneas recirculantes controladas por puertas ópticas. Las pruebas realizadas dentro del proyecto FRONTIERNET aseguran que el ruido ASE (*Amplified Spontaneous Emission*) acumulado por la señal en cada recirculación es aceptable, proporcionando un BER (*Bit Error Rate*) menor a 10^{-9} tras *10* recirculaciones.

La arquitectura FRONTIERNET permite emular un comportamiento de colas a la salida. Sin embargo, debido a la complejidad *hardware* asociada, comparativamente superior a otras alternativas, no ha sido estudiada en el capítulo *2*. Sí resulta de interés la descripción del mecanismo propuesto dentro del proyecto FRONTIERNET para el escalado en número de puertos del conmutador.

4.2.3.2 Frontiernet Multihop

La principal limitación para el escalado de la arquitectura FRONTIERNET se encuentra en el rango de sintonización necesario de los dispositivos TWC, igual al número de puertos del conmutador. Las propuestas para evitar este problema incluyen la interconexión en topologías multietapa, ya investigadas en el proyecto KEOPS. Sin embargo, en esta sección nos centraremos en otra de las alternativas descritas, por resultar un ejemplo interesante de cómo las estrategias de escalado presentes en conmutación electrónica, se encuentran también aplicadas en arquitecturas de conmutación OPS. En este caso, se trata de las redes de interconexión multi-salto mediante una topología *perfect-shuffle* [Sto71], en las que un paquete es encaminado entre distintos dispositivos de la red, dando un número de saltos variable, hasta encontrar su salida.

La arquitectura *Multihop* FRONTIERNET fue propuesta en [Sas95]. Como muestra la figura 4-6, el conmutador está compuesto por un conjunto de módulos de entrada (con almacenamiento), un dispositivo AWG, módulos de salida (con almacenamiento) y enlaces de recirculación. Un módulo de entrada está compuesto

por un buffer de entrada, un *array* de dispositivos TWC, y un combinador. Un módulo de salida está compuesto por un filtro de salida y un buffer de salida WDM. El aspecto diferenciador de esta arquitectura es que cada módulo de entrada puede acceder únicamente a un pequeño número de canales de salida. Por ello, para alcanzar su puerto de salida, los paquetes deben ser encaminados a través de distintos módulos y lazos recirculantes necesitando múltiples saltos en distintas longitudes de onda. El algoritmo de enrutado establece las conversiones de longitudes de onda en cada salto con el objetivo de minimizar los retardos origen-destino. El esquema de interconexión propuesto para *Multihop* FRONTIERNET es el *perfect-shuffle* [Sto71]. El número de puertos de entrada y de salida es $N=ar^a$, que se divide en *a* grupos de r^a puertos cada uno. El buffer de entrada debe resolver los conflictos por la misma longitud de onda entre los paquetes entrantes y los paquetes provenientes de la línea de recirculación. Se estudiaron dos esquemas de resolución de contención: *store-and-forward* y *hot-potato*. Para el primero, únicamente uno de los paquetes es convertido a la longitud de onda adecuada para minimizar los saltos hasta el puerto de salida. El resto de los paquetes son almacenados en el módulo de entrada y enviados en posteriores ranuras temporales. El esquema *hot-potato* establece que todos los paquetes en contienda son transmitidos (no existe almacenamiento en los bufferes de entrada), uno de ellos al módulo de salida óptimo, y el resto a otros módulos distintos.

Esta estrategia proporciona varias simplificaciones en los requisitos *hardware.* En primer lugar, el rango de sintonización de los dispositivos TWC se reduce a *r* canales. También, el buffer WDM de salida puede recibir un máximo de *r* paquetes, y no tiene que trabajar sobre anchos de banda elevados, lo que simplifica los requisitos de los amplificadores ópticos. Por último, el número de longitudes de onda empleadas en total es $2(r-1)r^a+1$, que depende del parámetro *a* para un tamaño de conmutador dado. Por ejemplo, el número de longitudes de onda se reduce a una tercera parte para *a=4*. En contra, el número de saltos necesarios de un paquete para alcanzar su puerto de salida también se incrementa con *a.* Para el esquema *store-and-forward*, el número máximo de saltos está acotado a *2a-1*. Para el esquema *hot-potato*, se ha propuesto la posibilidad de permitir un máximo de *1* encaminamiento por defecto, para garantizar un máximo de *3a-1* saltos. Esta decisión implica la presencia de almacenamiento dentro de los módulos de entrada.

Como conclusión para esta arquitectura, indicar que su interés fuera de la "novedosa" (en OPS) aplicación de multi-salto, es dudoso, debido al mucho mayor coste *hardware* incurrido, con respecto a otras alternativas. Por otro lado, no se han presentado resultados de evaluación de prestaciones de esta arquitectura ante ningún tipo de tráfico.

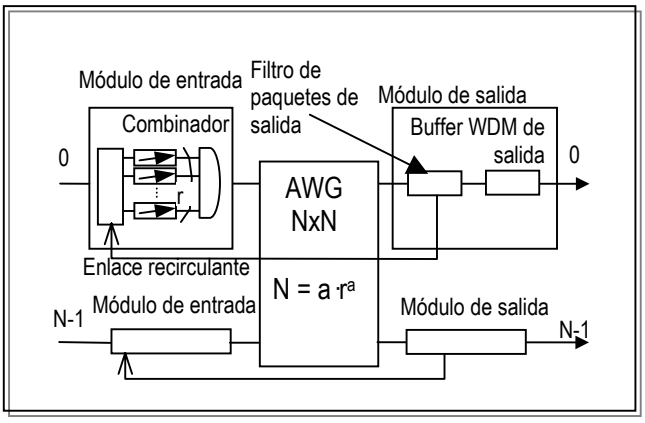


Figura 4-6. Arquitectura *Multihop* FRONTIERNET

4.2.4 WASPNET multiplano

La arquitectura de conmutación WASPNET fue presentada en el capítulo anterior (figuras 3-8-(a), 3-8-(b)). El crecimiento en número de puertos de esta arquitectura fue investigado dentro del proyecto WASPNET para el modo de operación SCWP únicamente. La arquitectura multiplano propuesta en [Hun99] se muestra en la figura 4-7, para un conmutador con *N* fibras de entrada y salida, y *n* longitudes de onda por fibra.

- Los paquetes entrantes son distribuidos pasivamente por un demultiplexor entre cada uno de los *n* planos, en función de su longitud de onda de entrada.
- Cada uno de los planos se compone de un conmutador *feedback NxN*, que requiere un encaminador AWG *2Nx2N* y *4N* dispositivos sintonizables TWC. La utilización de esta arquitectura con encaminamiento por defecto a través de las líneas recirculantes permite implementar políticas de calidad de servicio a un coste razonable (paquetes de menor prioridad pueden ser encaminados por las líneas recirculantes frente a paquetes de mayor prioridad).
- A la salida de cada plano, se dispone de una etapa de conmutación espacial, de manera que cada plano es capaz de transmitir hasta *N* paquetes a través de la misma fibra de salida, controlando la longitud de onda de transmisión de los mismos.

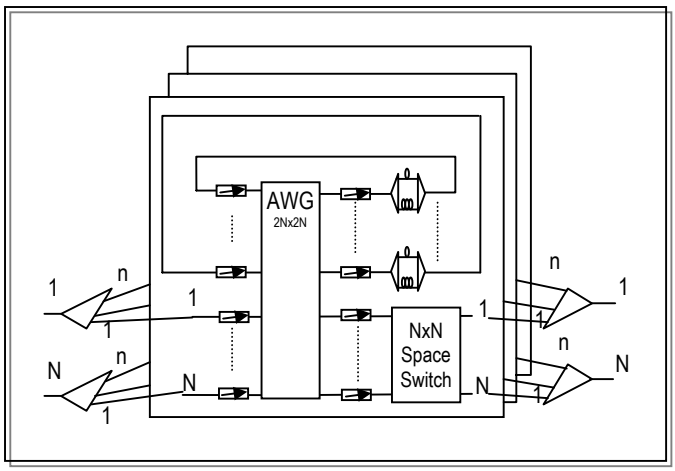


Figura 4-7. Arquitectura WASPNET multiplano

La planificación SCWP del conmutador requiere una coordinación entre todos los planos. Las restricciones que acotan las decisiones del planificador son:

1. **Contención de fibra de salida.** Para cada fibra de salida, puede ser transmitido un máximo de *n* paquetes en cada ranura temporal. Esta familia de *N* restricciones es inherente a la conmutación de paquetes en modo SCWP.

2. **Contención de salida de plano de conmutación.** Para cada plano, puede ser transmitido un máximo de *N* paquetes en cada ranura temporal. Esta familia de *n* restricciones es inherente a la distribución multiplano, donde cada plano es un conmutador *NxN*.

3. **Contención interna a cada plano de conmutación.**

- Para cada plano, en la entrada de cada línea recirculante puede haber un máximo de un paquete cada ranura temporal.

- Para cada plano, en la salida de cada línea recirculante puede haber un máximo de un paquete en cada ranura temporal.

Ambas familias de restricciones son impuestas por el hecho de emplear una arquitectura WASPNET *feedback* en cada plano. Como ejemplo, no existirían si cada plano estuviese formado por un conmutador OPS con capacidad de emular colas a la salida.

Las prestaciones de la arquitectura han sido evaluadas en [Chi01-1] asumiendo un planificador sub-óptimo, que permite mantener el orden entre paquetes.

4.2.5 Input-Buffered Wavelength-Routed Switch

La limitación al crecimiento de la arquitectura IB-WR estudiada en el capítulo *3* se debe al requisito de tamaño de encaminadores AWG y rangos de sintonización de dispositivos TWC iguales al número de puertos. En [Zho98] se propone una arquitectura IB-WR de gran escala que elimina los problemas anteriores, al coste de incluir conmutación espacial mediante puertas ópticas. La arquitectura propuesta (figura 4-8-(a)) se basa en:

- Utilización de *R* módulos de almacenamiento, donde el tamaño de los dispositivos AWG y el rango de sintonización de la primera etapa de dispositivos TWC son iguales a $K=max\ (N/R,M)$.

- La combinación de conmutación espacial mediante puertas ópticas y mediante encaminamiento dentro de un AWG en la sección de conmutación (*switching section*). Como consecuencia, el rango de sintonización de la segunda etapa de TWCs y el tamaño de los AWG empleados se recuce a *N/R*, a costa de la utilización de $N \cdot R$ puertas ópticas y *N* combinadores pasivos de *R* entradas.

Los paquetes entrantes son convertidos a una longitud de onda que determinará su retardo en la sección de almacenamiento (*buffering section*). La selección del retardo se realiza siguiendo exactamente las mismas restricciones que el conmutador IB-WR estudiado en el capítulo *3*: contención en el dispositivo TWC de la sección de conmutación (*switching section*) y contención de salida del conmutador. La utilización o no de la arquitectura en su versión de gran escala, es transparente a efectos de planificación.

El conmutador IB-WR de gran escala propuesto en [Zho98] debe adaptarse al escenario WDM para la aplicación de los modos de operación SHWP/SCWP. Las modificaciones aplicadas a un conmutador con *N* fibras de entrada y salida, y *n* longitudes de onda por fibra, son las siguientes:

- Se ha añadido una etapa de demultiplexación WDM en las fibras de entrada, que separan los paquetes entrantes por su longitud de onda, en *n* fibras distintas.
- Se ha asociado una sección de almacenamiento a cada fibra de entrada. El tamaño de los dispositivos AWG y el rango de sintonización de los TWC de esta sección son *K=max(n,M)*.
- Se ha asociado un módulo de la etapa de conmutación a cada fibra de salida. Esta decisión permite la simplificación de la arquitectura, sustituyendo el AWG final por un multiplexor. Respecto a la arquitectura IB-WR adaptada, se elimina la necesidad de un conversor FWC de salida. Por ello, los TWC de la etapa de conmutación deben convertir los paquetes a su longitud de onda de salida final.

Estas simplificaciones en la adaptación WDM atañen únicamente al *hardware*. El problema de optimización que deben resolver los conmutadores para ambos modos de operación es el mismo al expresado en el capítulo *2*. Por ello, los algoritmos de planificación estudiados son también de aplicación para la versión de gran escala de este conmutador. Esto resalta el interés en el algoritmo PDBM (*Parallel Desynchronized Block Matching*), propuesto en esta tesis doctoral, para el que se ha demostrado que el número de iteraciones para la convergencia es independiente del tamaño del conmutador.

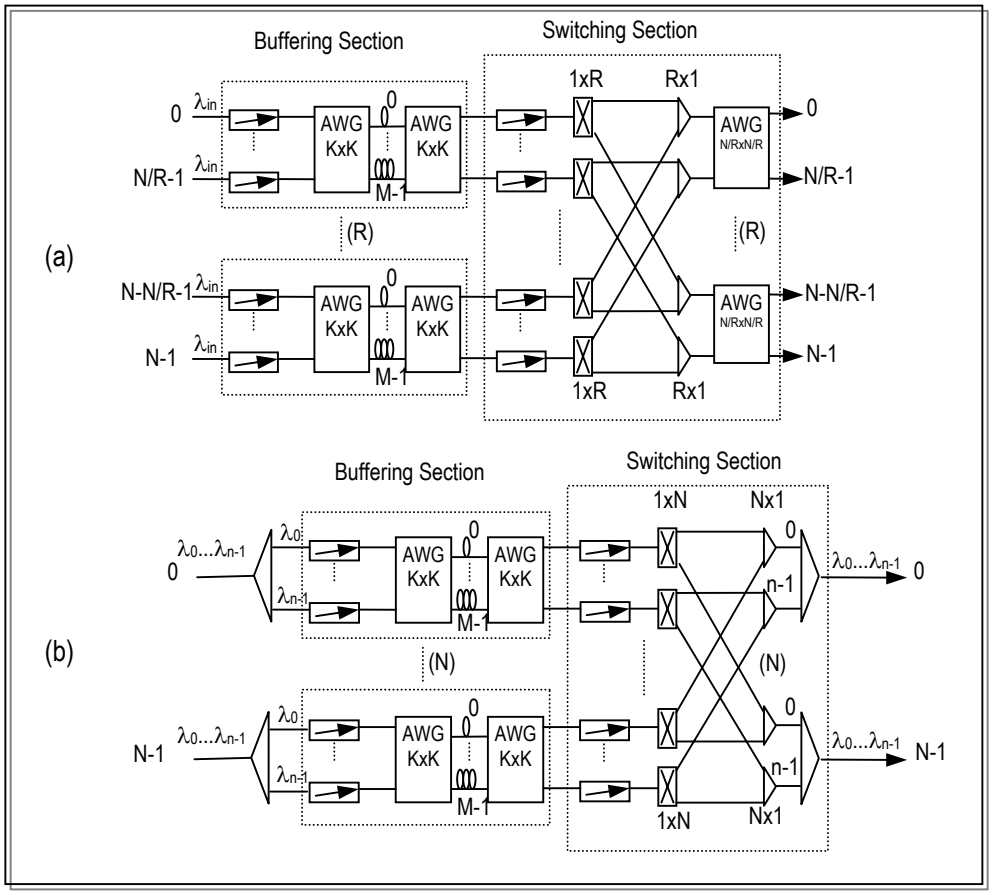


Figura 4-8. Arquitectura IB-WR de gran escala, (a) arquitectura original [Zho98], (b) adaptación WDM

4.2.6 Conclusiones a la revisión del estado de la técnica

La revisión del estado de la técnica descrito en esta sección incluye lo publicado en [Pav02]. Las conclusiones que se pueden extraer son:

- Se observa en muchos casos la repetición de tendencias similares a las aplicadas en conmutación electrónica de gran escala, para abordar el crecimiento cuadrático de la complejidad de los conmutadores: interconexión de dispositivos de menor tamaño organizados en redes multi-etapa, redes multi-salto, o arquitecturas multiplano.
- Ninguna de las propuestas encontradas, salvo el conmutador multiplano WASPNET, han sido diseñadas para trabajar con puertos de entrada y salida WDM, y por tanto, requieren algún tipo de adaptación *hardware*. En algunos casos, esta adaptación permite introducir simplificaciones. Por otro lado, la reducción de la necesidad de almacenamiento que implica el modo SCWP no ha sido tenido en cuenta en el diseño de muchas arquitecturas.

En la siguiente sección se describirán y evaluarán varias estrategias de crecimiento de arquitecturas OPS, basadas en el aprovechamiento del efecto *knock-out*, para el modo de operación SCWP. El diseño de las mismas busca explotar las ventajas que este modo de operación proporciona. Estas arquitecturas, y el análisis matemático que permite la evaluación de sus prestaciones, han sido propuestas como parte de esta tesis doctoral. Esta evaluación de prestaciones permitirá incluir dichas arquitecturas en una posterior comparativa de costes.

4.3 Arquitecturas de conmutación OPS *Knock-out*

4.3.1 Descripción de las arquitecturas

En esta sección se presentan dos estrategias de crecimiento para arquitecturas de conmutación OPS de gran escala, propuestas dentro de esta tesis doctoral. Ambas estrategias se basan en el aprovechamiento del efecto *knock-out* [Yeh87], en la interconexión de una etapa de conmutación sin memoria OPS (etapa de distribución, *distribution stage*), y un conjunto de módulos conmutadores OPS de menor tamaño (etapa de almacenamiento, *buffering stage*). Las dos versiones bajo estudio se muestran en las figuras 4-9-(a) y 4-9-(b), para un conmutador con N fibras de entrada y salida y n longitudes de onda por fibra. En la primera de ellas (que llamaremos OFD, *Output-Fiber-Distributed*), se asocia un módulo de almacenamiento para cada fibra de salida del conmutador. En la segunda versión (que llamaremos OWD, *Output-Wavelength-Distributed*), se emplea un módulo de almacenamiento para cada una de las longitudes de onda de salida del conmutador.

4.3.1.1 Etapa de distribución

La etapa de distribución consiste en una conmutador OPS sin memoria, con nN puertos de entrada y LN puertos de salida en la arquitectura OFD, $L'n$ puertos de salida en la arquitectura OWD. En el diseño mostrado en la figura 4-9, esta etapa se implementa mediante la interconexión de nN conversores TWC y un encaminador AWG. La conversión de longitud de onda de los paquetes entrantes se emplea para seleccionar el puerto de salida del AWG, entre los L (L') puertos que conectan la etapa de distribución con la fibra (longitud de onda) de salida demandada. Por lo tanto, las pérdidas *knock-out* surgen cuando en una ranura temporal, más de L paquetes

demandan la misma fibra (OFD), o más de *L'* paquetes seleccionan la misma longitud de onda de salida (OWD).

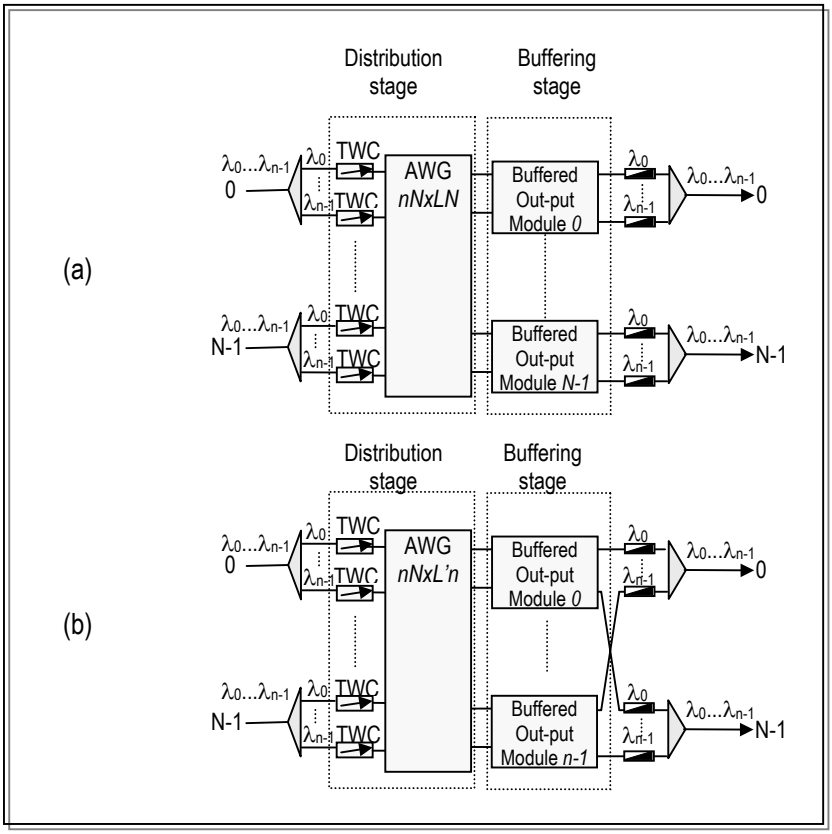


Figura 4-9. Arquitecturas *Knock-out* de gran escala [Pav03-4], (a) OFD, (b) OWD

Los límites de implementación para las etapas de distribución así diseñadas se encuentran en el tamaño del dispositivo AWG y el rango de sintonización de los dispositivos TWC, ambos iguales a *LN* (*L'n*). Este límite puede abordarse mediante la distribución en dos etapas de esta sección, tal y como muestran las figuras 4-10-(a) y 4-10-(b). La estructura final se convierte en una red de Clos de *3* etapas con expansión, con memoria únicamente en la etapa final.

En la versión de *3* etapas, existen dos fuentes de pérdida de paquetes en las etapas de distribución:

(1) Debidas al efecto *knock-out*, cuando más de *L* (*L'*) paquetes demandan el mismo módulo de salida.

(2) Debidas el encaminamiento a través de etapa intermedia, cuando dos paquetes destinados al mismo módulo de tercera etapa, son encaminados a través del mismo módulo de segunda etapa. Para arquitecturas simétricas, y para arquitecturas con expansión (como las que nos ocupa), la propiedad de no bloqueo de la red de Clos asegura que es posible eliminar este tipo de pérdidas empleando un algoritmo de encaminamiento óptimo. La aplicación de distintos algoritmos de encaminamiento subóptimos genera una cierta probabilidad de pérdida por encaminamiento (*routing loss*) a añadir a la probabilidad de pérdida *knock-out*, aunque de menor magnitud que ésta.

En esta tesis doctoral, el dimensionamiento del número de módulos de segunda etapa para este tipo de redes se hará atendiendo únicamente al efecto *knock-out*, suponiendo un encaminamiento óptimo, o despreciando las posibles pérdidas en el caso de aplicar encaminamiento subóptimo.

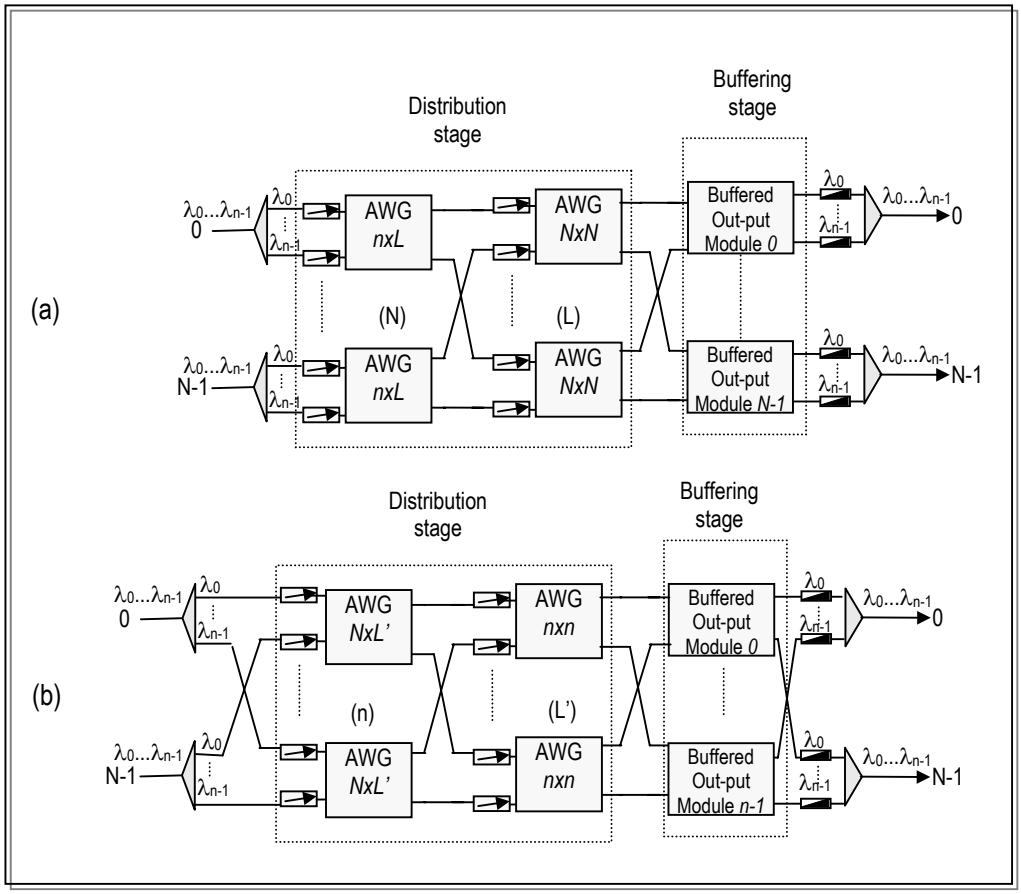


Figura 4-10. Arquitecturas *Knock-out* de gran escala [Pav03-4], con etapa de distribución multi-etapa (a) OFD, (b) OWD

4.3.1.2 Etapa de almacenamiento

Cualquiera de las arquitecturas de conmutación OPS descritas en esta tesis doctoral puede ser empleada en la implementación de los módulos de almacenamiento de salida. El objetivo de cada una de estas etapas es doble: (1) almacenar apropiadamente los paquetes entrantes, (2) resolviendo la contención en el proceso de conmutación hacia los puertos de salida demandados. La saturación de bufferes es la fuente de pérdidas de paquetes en esta etapa.

En la comparativa de la sección 4.4, se emplea la arquitectura KEOPS WDM (figura 2-1-(b)) como módulo de almacenamiento de salida. Es interesante destacar que, para la arquitectura distribuida por longitud de onda de salida (OWD), la utilización de conmutadores KEOPS permite mantener la funcionalidad de *multicast*: un paquete entrante a un conmutador KEOPS puede alcanzar desde ese módulo todas las fibras de salida, siempre empleando la misma longitud de onda. La exploración de las prestaciones que esta opción ofrece para tráfico *multicast* no será objeto de estudio en esta tesis doctoral.

4.3.1.3 Planificación SCWP

El algoritmo de planificación SCWP que aplicaremos a nuestra arquitectura es el algoritmo uniforme propuesto, mostrado en la figura 2-10. Como ya fue descrito en el capítulo *2*, este algoritmo asigna cíclicamente longitudes de onda a los paquetes destinados a la misma fibra de salida, mediante un puntero *round-robin* (que no es reiniciado en ningún momento). En el caso de emplear módulos de salida basados en conmutadores con capacidad de emular colas a la salida (como por ejemplo el conmutador KEOPS), este algoritmo proporciona las prestaciones óptimas en términos de probabilidad de pérdida y retardo. Por ello, en el caso de ser capaces de eliminar (o hacer despreciables) las pérdidas *knock-out*, las arquitecturas descritas en este capítulo proporcionan a su vez las prestaciones óptimas alcanzables.

4.3.2 Evaluación de las pérdidas *knock-out*

Esta sección analiza las pérdidas *knock-out* de las arquitecturas OPS *knock-out* OFD y OWD descritas en la sección 4.1. La evaluación se realiza asumiendo tráfico de entrada Bernouilli de parámetro ρ.

Se define la variable aleatoria *A*, con valores entre *0* y A_{MAX}, como el número de llegadas de paquetes destinados al mismo módulo de salida en una ranura temporal. La probabilidad de pérdida causada por el efecto *knock-out* (P_{loss}) en ambas arquitecturas viene dada por la relación entre los paquetes perdidos y recibidos por un módulo de salida en una ranura temporal (Ec. 4.1).

$$P_{loss} = \frac{E[paquetes\ perdidos]}{E[paquetes\ recibidos]} = \frac{\sum_{i=L+1}^{A_{MAX}} P(A=i)(i-L)}{\sum_{i=1}^{A_{MAX}} P(A=i) i} \qquad \text{(Ec. 4.1)}$$

La obtención de la función de densidad de probabilidad (discreta) del número de llegadas a un módulo de salida *A*, $P(A=i),\ i=0,...,A_{MAX}$, es un paso necesario para el cálculo de las pérdidas *knock-out*. Para la versión del conmutador distribuido por fibra de salida (OFD), todos los paquetes destinados a la misma fibra, son asimismo destinados al mismo módulo de salida. Esto no se ve afectado por la selección de longitud de onda que aplica el algoritmo de planificación SCWP. Por ello, para el patrón de llegadas Bernouilli uniforme, la distribución de probabilidad buscada *A* es directamente la proporcionada por la fórmula binomial:

$$P[A=k] = \binom{nN}{k}\left(\frac{\rho}{N}\right)^k\left(1-\frac{\rho}{N}\right)^{nN-k} ;k = 0,1,...,nN = A_{MAX} \qquad \text{(Ec. 4.2)}$$

Por otro lado, el cálculo de la función de densidad de probabilidad de *A* para la arquitectura distribuida por longitud de onda de salida (OWD), sí se ve afectada por la manera en que el algoritmo de planificación SCWP selecciona la longitud de onda de transmisión de los paquetes entrantes. La obtención de la función de distribución buscada requiere incorporar esta información.

Para un módulo de salida fijado (por ejemplo el asociado a la longitud de onda λ_0), definimos la variable aleatoria A_i (*i=0,...,N-1*) como el número de paquetes que están destinados a la fibra de salida *i*, *y a los que se les asigna la longitud de onda* λ_0. Por su parte, las variables aleatorias A_i son función de otras dos variables aleatorias:

- a_i (*i=0,...,N-1*). Indica el número total de paquetes destinados a la fibra de salida *i* en esta ranura temporal. Cada una de las variables aleatorias a_i puede tomar los valores desde *0* a *nN*.
- p_i (*i=0,...,N-1*). Indica la posición al comienzo de esta ranura temporal del puntero *round-robin* asociado a la fibra de salida *i*. Cada variable aleatoria p_i puede tomar los valores desde *0* a *n-1*. En nuestro caso, supondremos que un valor de *0* significa que el puntero señala a nuestro módulo de salida λ_0, por lo que el primer paquete destinado a la fibra de salida *i*, será encaminado hacia este módulo. Un valor de *n-1* indica que los *n-1* primeros paquetes recibidos con destino la fibra de salida *i*, serán destinados a otros módulos de salida.

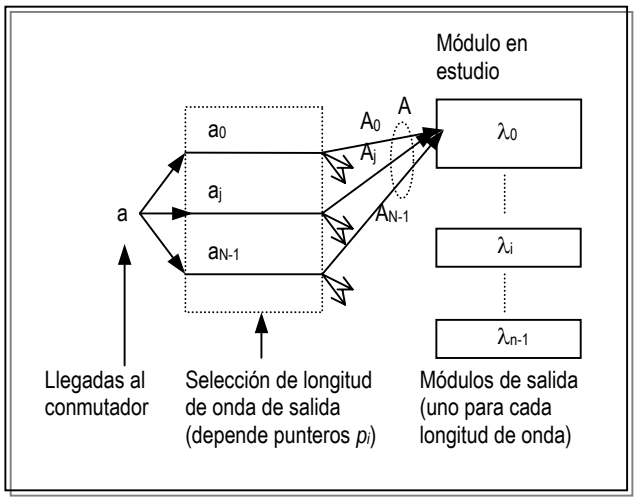


Figura 4-11. Variables aleatorias involucradas en el análisis de pérdidas *knock-out* para la arquitectura OWD

La relación entre las variables aleatorias se ilustra en la figura 4-11. Bajo estas consideraciones, las variables aleatorias A_i vienen especificadas usando la función techo:

$$A_i = \left\lceil \frac{a_i - p_i}{n} \right\rceil, i = 0,...,N-1 \qquad \text{(Ec. 4.3)}$$

Por tanto, la variable aleatoria de interés *A* se expresa como la suma de las variables aleatorias A_i, *i=0,...,N-1*.

$$\begin{aligned} &A = A_0 + ... + A_{N-1} = A(\vec{a}, \vec{p}) \\ &\vec{a} = (a_0,...,a_{N-1}),\ vector\ de\ llegadas\ a\ cada\ fibra\ de\ salida \\ &\vec{p} = (p_0,...,p_{N-1}),\ vector\ de\ posiciones\ de\ punteros\ de\ cada\ fibra\ de\ salida \end{aligned} \qquad \text{(Ec. 4.4)}$$

En cada ranura temporal, las variables aleatorias p_i son mutuamente independientes, al depender del número de paquetes recibidos en todas las ranuras temporales anteriores, destinadas a fibras de salida distintas. Asimismo, son

independientes de las variables aleatorias a_i, correspondientes a las llegadas en esta ranura temporal. Sin embargo, las variables aleatorias a_i, $i=0,...,N-1$ no pueden considerarse mutuamente independientes. Por ejemplo, la suma de las llegadas a todas las fibras de salida $a_0+...+a_{N-1}$ está acotada por el número de puertos de entrada nN.

La ecuación exacta que proporciona la función de densidad de probabilidad de A viene dada por la Ec. 4.5.

$$P[A=k]=\sum_{D}P[\vec{a}=\vec{a}_d,\vec{p}=\vec{p}_d]=\sum_{D}P[\vec{a}=\vec{a}_d]P[p_1=p_{1_d}]\cdot...\cdot P[p_{N-1}=p_{N-1_d}]$$
$$D=\left\{\vec{a}_d,\vec{p}_d \ tal\ que\ A=\sum_{i=0}^{N-1}\left\lceil\frac{a_i-p_i}{n}\right\rceil=k\right\} \qquad \text{(Ec. 4.5)}$$

Como se observa, D simboliza el conjunto de posibles llegadas $\vec{a}_d$ y posibles posiciones de punteros $\vec{p}_d$ que implican que el número de llegadas al módulo de salida bajo estudio sea igual a k.

La distribución de probabilidad conjunta del proceso de llegadas $\vec{a}_d$ viene dado por la fórmula de la distribución multinomial:

$$P[\vec{a}=\vec{a}_d=(a_{d_0},...,a_{d_{N-1}})]=\frac{(nN)!}{a_{d_0}!\cdot...\cdot a_{d_{N-1}}!\cdot E!}(1-\rho)^E\left(\frac{\rho}{N}\right)^{nN-E}$$
$$E=nN-\sum_{i=0}^{N-1}a_{d_i}=número\ de\ puertos\ vacíos \qquad \text{(Ec. 4.6)}$$

Para el algoritmo de planificación SCWP uniforme empleado, las probabilidades de encontrar un puntero en una posición concreta se distribuyen uniformemente entre todas las longitudes de onda de salida.

$$P[p_i=k]=\frac{1}{n},\forall k=0,...,n-1;\forall i=0,...,N-1 \qquad \text{(Ec. 4.7)}$$

4.3.2.1 Método simplificado de cálculo

La resolución por "fuerza bruta" de la ecuación (Ec. 4.5), recorriendo todos los posibles valores $\vec{a}_d$ y $\vec{p}_d$ y aplicando las probabilidades de cada combinación mediante las distribuciones de las Ec. 4.6 – 4.7, es inviable incluso para pequeños tamaños de conmutador. Como medida de la complejidad, el número de combinaciones a explorar viene dado por:

$$C(nN,N+1)\cdot n^N \ tal\ que$$
$$C(X,1)=1;C(X,F)=\sum_{i=0}^{X}C(X-i,F-1),F>1 \qquad \text{(Ec. 4.8)}$$

Donde el factor *C(nN,N+1)* computa el espacio a explorar por los posibles valores del vector $\vec{a}_d$, como el número de formas diferentes de agrupar *nN* puertos de entrada entre *N+1* destinos (*N* fibras de salida, más una fibra auxiliar para los puertos vacíos). El factor n^N computa el espacio a explorar por el vector $\vec{p}_d$, con *N* variables que pueden tomar *n* valores cada una.

El método de simplificación seguido en esta tesis doctoral, se basa en la definición de las variables aleatorias condicionadas:

$$\begin{aligned} A' &= \left(A \middle| \vec{a} = \vec{a}_d\right) \\ A_i' &= \left(A_i' \middle| \vec{a} = \vec{a}_d\right), i = 0,...,N-1 \end{aligned} \qquad \text{(Ec. 4.9)}$$

Donde se mantiene la relación

$$A' = A_0' + ... + A_{N-1}' \qquad \text{(Ec. 4.10)}$$

Aplicando el teorema de las probabilidades totales, expresamos la variable aleatoria objetivo *A*, como función de las variables aleatorias condicionadas:

$$A^{(k)} = P[A' = k] = \sum_{\vec{a}_d} P\left[A = k \middle| \vec{a} = \vec{a}_d\right] P[\vec{a} = \vec{a}_d] = \sum_{\vec{a}_d} P[A' = k] P[\vec{a} = \vec{a}_d] \qquad \text{(Ec. 4.11)}$$

El punto fundamental del método se basa en que ahora, la independencia entre las variables *condicionadas A_i', i=0,...,N-1*, sí nos permite la aplicación del teorema de la convolución para el cálculo de la variable aleatoria *A',* como suma de las variables *A_i', i=0,...,N-1* (como muestra la Ec. 4.10),

$$A' = A_0' + ... + A_{N-1}' \Rightarrow A'^{(k)} = A_0'^{(k)} * ... * A_{N-1}'^{(k)} \qquad \text{(Ec. 4.12)}$$

Aplicando las ecuaciones Ec. 4.3 y 4.7 obtenemos la función de distribución de probabilidad de las variables *A_i'* de la siguiente manera:

$$A_i'^{(k)} = P[A_i' = k] = \begin{cases} \left\lfloor \frac{a_i}{n} \right\rfloor \text{ con probabilidad } 1 - frac\left(\frac{a_i}{n}\right) \\ 1 + \left\lfloor \frac{a_i}{n} \right\rfloor \text{ con probabilidad } frac\left(\frac{a_i}{n}\right) \end{cases} \qquad \text{(Ec. 4.13)}$$

$$0 \le frac(x) < 1, frac(x) = \text{parte fraccional del número real } x$$

El segundo punto simplificador del método es que, analizando la expresión 4.11 para distintos valores de $\vec{a}_d = \left(a_{0_d},...,a_{N-1_d}\right)$, se observa que para aquellos vectores $\vec{a}_d$ con valores permutados de sus coordenadas a_{i_d}, se obtienen las mismas

probabilidades *P[A'=k]* y las mismas probabilidades $P[\vec{a} = \vec{a}_d]$. Por tanto, la ecuación 4.11 puede reescribirse de la siguiente manera:

$$A^{(k)} = P[A = k] = \sum_{\vec{a}_d} perm(a_{0_d},...,a_{N-1_d}) P[\vec{a} = \vec{a}_d] P[A' = k] \qquad \text{(Ec. 4.14)}$$

Donde la función *perm* computa el número de permutaciones diferentes de los valores a_{i_d} *, i=0,...,N-1*, dado por:

$$perm(a_{0_d},...,a_{N-1_d}) = \frac{N!}{r_0! \cdot ... \cdot r_{v-1}!}$$
$$v = \text{número de coordenadas } a_{i_d} \text{ con valores diferentes} \qquad \text{(Ec. 4.15)}$$
$$r_i, i = 0,...,v-1 = \text{número de veces que una coordenada se repite}$$

La programación del método, requiere:

- Un proceso que enumere todos los sumandos de la expresión 4.14, correspondientes al subconjunto de los posibles valores de $\vec{a}_d$, donde las coordenadas no sean una permutación de un sumando anterior. La implementación de este proceso de enumeración es computacionalmente simple, mediante actualizaciones de un vector de coordenadas ordenado. Cada uno de estos sumandos de la enumeración cubre $n^N \cdot perm(a_{0_d},...,a_{N-1_d})$ sumandos de la expresión 4.5.
- Para cada sumando, debe calcularse el valor de
  - $P[\vec{a} = \vec{a}_d]$, a partir de la ecuación 4.6
  - *P[A'=k]*, a partir de las ecuaciones 4.13 y 4.12. El proceso de cálculo de las *N* convoluciones puede acelerarse mediante el almacenamiento de resultados parciales de iteraciones anteriores.
- Una vez obtenida la distribución de probabilidades de la variable aleatoria *A*, aplicando la ecuación 4.1 (empleando *L'* en lugar de *L*), obtenemos las probabilidades de pérdida *knock-out* para distintos valores de *L'*.
- Es importante destacar que este método es directamente aplicable a procesos de llegadas independientes e idénticamente distribuidos (como Bernouilli uniforme o Bernouilli *hot-spot*), y adaptable a todos los procesos en los que se pueda extraer o acotar los procesos de llegadas en una ranura temporal.

El método, tal y como fue implementado (herramienta MATLAB) para la obtención de resultados, produjo los resultados presentados en esta sección en un tiempo variable desde unos segundos o minutos (para arquitecturas de hasta *32* puertos de entrada y salida), hasta horas o días (para arquitecturas mayores), en un sistema compartido AlphaServer HPC160 de *32* Gigaflops de rendimiento teórico.

4.3.2.2 Cota superior A_{MAX}

En la arquitectura OFD, el parámetro A_{MAX} (cota superior de llegadas a un módulo de salida), es igual al número de puertos de entrada del conmutador. Para la arquitectura OWD, resulta de interés calcular el número máximo de llegadas A_{MAX} que produce la aplicación del algoritmo de planificación SCWP. La razón es doble:

- El método de cálculo descrito en la sección anterior proporciona valores exactos, pero es computacionalmente muy costoso para arquitecturas de *128* puertos o más.
- Dimensionando el parámetro *L'* como A_{MAX}, es posible eliminar las pérdidas *knock-out* para cualquier patrón de tráfico de entrada.

Para la arquitectura OWD, el cálculo de A_{MAX} debe basarse en el estudio del caso peor al comienzo de una ranura temporal:

1) Todos los punteros p_i, *i=0,...,N-1*, apuntan al módulo de salida en estudio.
2) Llegan *nN* paquetes, de los cuales

- *N* van destinados a cada una de las *N* fibras de salida, y por tanto son todos encaminados al módulo de salida en estudio.
- Los restantes (*nN-N*) paquetes son destinados todos a una misma fibra de salida, y por tanto $\left\lceil \frac{nN-N-n+1}{n} \right\rceil$ llegan al módulo de salida en estudio.

Por tanto, el valor de A_{MAX} que se obtiene es:

$$A_{MAX} = min\left(nN, N + \left\lceil \frac{nN-N-n+1}{n} \right\rceil \right) \qquad \text{(Ec. 4.16)}$$

4.3.2.3 Precisión de la cota superior

Una vez obtenida la expresión de cota superior, es de interés valorar su posible aplicación para el dimensionamiento de arquitecturas, que suele realizarse asumiendo cargas altas en el conmutador. Para ello observamos los valores de la siguiente tabla:

L' / A_{MAX}	**N=2**	**N=4**	**N=8**	**N=16**	**N=32**
n=2	3/3	6/6	12/12	21/24	38/48
n=4	3/3	7/7	12/14	21/28	***
n=8	3/3	7/7	12/15	***	***
n=16	3/3	6/7	***	***	***
n=32	3/3	***	***	***	***

Tabla 4-1. Valores de *L'* y A_{MAX} para distintos tamaños de conmutador. Los valores *L'* han sido calculados para obtener una probabilidad de pérdida *knock-out* < 10^{-8}, para tráfico Bernouilli uniforme de carga *0.8*.

En ella se muestran los valores de *L'* y A_{MAX} para distintos tamaños de conmutador. Los valores *L'* han sido calculados para proporcionar una probabilidad de pérdida *knock-out* menor que 10^{-8}, ante tráfico Bernouilli de carga *0.8*. Los valores obtenidos indican lo siguiente:

- En todos los casos, la cota superior A_{MAX} que asegura la eliminación de las pérdidas knock-out es relativamente precisa.
- Esta precisión es muy significativa para valores pequeños de *N* (fibras de entrada y salida).
- La precisión parece ligeramente peor para valores mayores del parámetro *n*.

Por tanto, se deduce que la cota superior obtenida es una buena aproximación para el dimensionamiento de estos conmutadores en cargas altas de trabajo, en un escenario de red troncal con un bajo número de fibras de entrada y salida (*N*), y alto número de longitudes de onda (*n*) en el nodo. En el caso de tener como objetivo dimensionar un sistema para operar, por ejemplo, con cargas controladas menores a *0.5*, es el método exacto el que, lógicamente, debe ser empleado.

4.4 Comparativa de arquitecturas

En esta sección, se aborda una comparativa entre arquitecturas de conmutación OPS de gran escala. El escenario en el que se enfoca esta comparativa es el de una red troncal DWDM, por tanto con un bajo grado de interconexión de los nodos (*N*), y con un elevado número de longitudes de onda por fibra (*n*). El único modo de operación considerado es SCWP, por sus ineludibles mejores prestaciones. El dimensionamiento de los conmutadores se calcula para asegurar una probabilidad de pérdida de paquetes menor a 10^{-8} bajo carga Bernouilli uniforme de parámetro ρ=*0.8*.

Entre todas las arquitecturas descritas, serán excluidas del proceso de comparación aquellas que resulten claramente inferiores en cuanto a relación coste *hardware*/prestaciones:

- La arquitectura de conmutación Frontiernet multisalto. Los costes en cuanto a número de kilómetros de fibra (al tener módulos de almacenamiento para cada puerto de salida), y dispositivos TWC se encuentran claramente lejos de otras arquitecturas descritas.
- La arquitectura de conmutación WASPNET multiplano es inferior a la arquitectura IB-WR de gran escala. Esto es debido a que en todos los escenarios requiere el doble de dispositivos TWC, con mayor rango de sintonización, y una mayor cantidad de kilómetros de fibra.
- Las estrategias de crecimiento mediante redes de Clos propuestas por el proyecto KEOPS, con arquitecturas KEOPS con memoria en todas las etapas, son descartadas frente a las alternativas OFD y OWD con memoria únicamente en la última etapa.

Por otro lado, respecto a las arquitecturas incluidas, se destaca lo siguiente:

- Arquitecturas *knock-out* OWD: por las razones expuestas en la sección anterior, se empleará la cota superior A_{MAX} para el dimensionamiento de *L',* lo que garantiza la eliminación de las pérdidas *knock-out*. Se desprecia el posible efecto de las pérdidas por encaminamiento en la versión de *3* etapas. Para el cómputo de número de componentes, se considera la utilización de arquitecturas KEOPS WDM (figura 2-1-(b)) como módulos de almacenamiento de salida, de *L'* entradas y *N* salidas.
- Arquitecturas *knock-out* OFD: para el dimensionamiento del parámetro *L*, de número de entradas a cada módulo de salida, se aplica la condición de hacer las pérdidas *knock-out* menores a 10^{-9}, un orden de magnitud menores que nuestro objetivo de dimensionamiento (10^{-8}). Para el cómputo de número de componentes, se considera la utilización de arquitecturas KEOPS WDM (figura 2-1-(b)) como módulos de almacenamiento de salida, de *L* entradas y *n* salidas.
- Se incluye en la comparativa la arquitectura *space switch*, por su reducción del ratio coste/prestaciones para arquitecturas con un alto número de longitudes de onda por fibra (*n*).
- Se incluye la versión de gran escala (g.e.) de la arquitectura OB-WR. El número de puertas ópticas de la etapa de conmutación espacial *nNxnN* computado en la comparativa, será el obtenido empleando la topología de red de Benes, propuesta como solución factible en [Zho98]. Para los tamaños de conmutador barajados, esto permite reducir apreciablemente el número de componentes: $2nN\,(2\log_2 nN-1) < n^2N^2$.
- Se incluye la versión de gran escala (g.e.) de la arquitectura IB-WR, aplicando el algoritmo de planificación SCWP PDBM. El dimensionamiento del número de retardos *M'* se ha realizado mediante simulaciones, siguiendo los criterios de tiempo de simulación descritos en el capítulo *3*.

A continuación se muestra una tabla con el número de componentes requeridos para cada una de las arquitecturas en estudio, en función de los parámetros *n, N, M, M'* (tamaño de buffer para la arquitectura IB-WR).

	FWC	Puertas ópticas	TWC [rango sintoniz. máx.]	Delay loops	Tamaño máx. AWG
OWD	n(N+L')	nN(M+L')	nN [nL']	n·(1...M)	nL'
OFD	N(n+L)	nN(M+L)	nN [LN]	N·(1...M)	LN
OWD 3 etapas	n(N+L')	nN(M+L')	nN+nL' [max (n,L')]	n·(1...M)	max (n,L')
OFD 3 etapas	N(n+L)	nN(M+L)	nN+LN [L]	N·(1...M)	L
IB-WR g.e.	0	N^2n	2nN [max (M',n)]	N·(1...M')	max (M',n)
OB-WR g.e.	0	$2nN(\log_2 nN-1)$	nN [n]	N·(1...M)	n
Space switch	0	nN^2M	nN [n]	N·(1...M)	0

Tabla 4-2. Cómputo de componentes de arquitecturas de conmutación de gran escala

Los valores de los parámetros *M, M', L, L'* para los valores de *N* y *n* objeto de nuestra comparativa se muestran en la tabla 4-3.

		M	M'	L	L'=A_{MAX}
N=4	**n=16**	3	4	33	7
	n=32	2	3	53	7
	n=64	2	3	88	7
	n=128	2	3	153	7
N=8	**n=16**	3	4	35	15
	n=32	2	3	56	15
	n=64	2	3	92	15
	n=128	2	3	157	15

Tabla 4-3. Dimensionamiento de las arquitecturas de gran escala en estudio

A partir de las expresiones y valores mostrados en las tablas anteriores, se confecciona la tabla 4-4 y 4-5, para los valores *n={16,32,64,128}, N={4,8}*.

		n=16	n=32	n=64	n=128
OWD	**P. Ópt.**	640	1152	2304	4608
	Nº TWC	64	128	256	512
	Rango máx.	112	224	448	896
	Nº FWC	176	352	704	1408
	km. fibra	19.2	19.2	38.4	76.8
OFD	**P. Ópt.**	2304	7040	23040	79360
	Nº TWC	64	128	256	512
	Rango máx.	132	212	352	612
	Nº FWC	196	340	608	1124
	km. fibra	4.8	2.4	2.4	2.4
OWD 3 etapas	**P. Ópt.**	640	1152	2304	4608
	Nº TWC	176	352	704	1408
	Rango máx.	16	32	64	128
	Nº FWC	176	352	704	1408
	km. fibra	19,2	19.2	38.4	76.8
OFD 3 etapas	**P. Ópt.**	2304	7040	23040	79360
	Nº TWC	196	340	608	1124
	Rango máx.	33	53	88	153
	Nº FWC	196	340	608	1124
	km. fibra	4.8	2.4	2.4	2.4
IB-WR g.e.	**P. Ópt.**	256	512	1024	2048
	Nº TWC	128	256	512	1024
	Rango máx.	16	32	64	128
	Nº FWC	0	0	0	0
	km. fibra	8	4.8	4.8	4.8
OB-WR g.e	**P. Ópt.**	4096	3328	7680	17408
	Nº TWC	64	128	256	512
	Rango máx.	16	32	64	128
	Nº FWC	0	0	0	0
	km. fibra	4.8	2.4	2.4	2.4
Space switch	**P. Ópt.**	768	1024	2048	4096
	Nº TWC	64	128	256	512
	Rango máx.	16	32	64	128
	Nº FWC	0	0	0	0
	km. fibra	4.8	2.4	2.4	2.4

Tabla 4-4. Cómputo de componentes para arquitecturas de gran escala, *N=4*, *n={16,32,64,128}*, tráfico Bernouilli uniforme ρ=*0.8*, prob. de pérdida <10^{-8}

		n=16	n=32	n=64	n=128
OWD	**P. Ópt.**	2304	4352	8704	17408
	Nº TWC	128	256	512	1024
	Rango máx.	240	480	960	1920
	Nº FWC	368	736	1472	2944
	km. fibra	19.2	19.2	38.4	76.8
OFD	**P. Ópt.**	4864	14848	48128	162816
	Nº TWC	128	256	512	1024
	Rango máx.	280	448	736	1256
	Nº FWC	408	704	1248	2280
	km. fibra	9.6	4.8	4.8	4.8
OWD 3 etapas	**P. Ópt.**	2304	4352	8704	17408
	Nº TWC	368	736	1472	2944
	Rango máx.	16	32	64	128
	Nº FWC	368	736	1472	2944
	km. fibra	19.2	19.2	38.4	76.8
OFD 3 etapas	**P. Ópt.**	4864	14848	48128	162816
	Nº TWC	408	704	1248	2280
	Rango máx.	35	56	92	157
	Nº FWC	408	704	1248	2280
	km. fibra	9.6	4.8	4.8	4.8
IB-WR g.e.	**P. Ópt.**	1024	2048	4096	8192
	Nº TWC	256	512	1024	2048
	Rango máx.	16	32	64	128
	Nº FWC	0	0	0	0
	km. fibra	16	9.6	9.6	9.6
OB-WR g.e	**P. Ópt.**	3328	7680	17408	38912
	Nº TWC	128	256	512	1024
	Rango máx.	16	32	64	128
	Nº FWC	0	0	0	0
	km. fibra	9.6	4.8	4.8	4.8
Space switch	**P. Ópt.**	3072	4096	8192	16384
	Nº TWC	128	256	512	1024
	Rango máx.	16	32	64	128
	Nº FWC	0	0	0	0
	km. fibra	9.6	4.8	4.8	4.8

Tabla 4-5. Cómputo de componentes para arquitecturas de gran escala, *N=8*, *n={16,32,64,128}*, tráfico Bernouilli uniforme ρ=*0.8*, prob. de pérdida <10^{-8}

La cantidad de valores a considerar ilustran la dificultad de extraer conclusiones definitivas. Más aún, encontrándonos en un escenario de medio plazo, para el que no se conoce el coste relativo y la viabilidad de los componentes implicados. De los valores obtenidos se destaca lo siguiente:

- La arquitectura OWD en sus dos versiones se muestra superior a su homóloga OFD en todos los parámetros, salvo en el de kilómetros de fibra. Más aún teniendo en cuenta que es capaz de eliminar completamente las pérdidas de *knock-out*, frente a una reducción estadística proporcionada por la arquitectura OFD.
- La arquitectura *space switch* se muestra superior a la arquitectura OB-WR en todas las situaciones, a pesar de la reducción de puertas ópticas obtenida de la aplicación de una configuración en red de Benes en el conmutador OB-WR.

- La arquitectura IB-WR g.e. se muestra superior en mayor o menor medida a la arquitectura OWD en *3* etapas. La necesidad de *multicast* o priorización de tráfico puede favorecer la arquitectura OWD con módulos KEOPS en la última etapa. Asimismo, queda abierta la comparativa empleando otras arquitecturas para estos módulos de almacenamiento.
- La arquitectura IB-WR g.e. se muestra, en general, superior a la arquitectura OWD con etapa de distribución monolítica.
- La arquitectura IB-WR g.e. requiere la mitad de puertas ópticas, pero el doble de dispositivos TWC que la arquitectura *space switch*. La ventaja de una u otra alternativa está claramente marcada por el coste relativo de ambos componentes.

4.5 Conclusiones

En este capítulo se ha realizado una aproximación a las arquitecturas de Conmutación Óptica de Paquetes de gran escala (alto número de puertos), barajando un escenario de red troncal DWDM (*Dense Wavelength Division Multiplexing*), con nodos interconectados con un número reducido de fibras de entrada y salida, y alto número de longitudes de onda por fibra.

Se ha realizado un repaso del estado del arte en este campo, y se han descrito las arquitecturas propuestas más relevantes. Posteriormente se han presentado dos estrategias de crecimiento de estas arquitecturas basadas en el aprovechamiento del efecto *knock-out*. Se ha presentado el análisis matemático para la evaluación de las pérdidas *knock-out* de la arquitectura OWD (*Output-Wavelength-Distributed*), junto con una cota superior que permite eliminar completamente este tipo de pérdidas. Se ha mostrado la gran precisión de esta cota para cargas altas, en los escenarios de gran escala planteados.

Finalmente, se ha realizado una comparativa de las arquitecturas descritas, descartando las claramente inferiores. Las conclusiones de esta comparativa deben ser forzosamente muy prudentes. Las arquitecturas *knock-out* OWD, IB-WR y *space switch* parecen las más prometedoras hasta la fecha, para este escenario. Debe sin embargo destacarse las muchas alternativas por explorar, como la de arquitecturas multiplano o arquitecturas *knock-out* basadas en conmutadores distintos a WASPNET y KEOPS.

Capítulo 5. Conclusiones y líneas futuras

5.1 Conclusiones

El objetivo de esta tesis doctoral ha sido la evaluación comparativa de prestaciones de arquitecturas de Conmutación Óptica de Paquetes (*Optical Packet Switching*, OPS), sobre redes troncales WDM (*Wavelength Division Multiplexing*). El trabajo ha comenzado con una revisión del estado de la técnica, cuyas conclusiones han sido el punto de partida para el trabajo desarrollado:

- La Conmutación Óptica de Paquetes se plantea como una solución definitiva para las redes troncales WDM, por sus ventajas inherentes en el control y reparto del ancho de banda de transmisión de los enlaces.
- Es viable tecnológicamente (se han construido prototipos con éxito desde la década de los 90), aunque está lejos del estado comercial, por el elevado coste *hardware* actual de los nodos de conmutación.
- Numerosos aspectos relativos a la aplicación de la Conmutación Óptica de Paquetes sobre redes WDM, no han sido definidos. Los comités estandarizadores han prestado hasta la fecha una atención prácticamente nula a esta alternativa, debido a la lejanía estimada en su aplicación comercial.

Este último punto ha constituido el primer obstáculo en el objetivo de realizar un estudio sistemático de las arquitecturas de conmutación OPS. El punto esencial ha sido la falta de consenso respecto a lo que se ha denominado *modo de operación de la red*: el mecanismo de asociación entre los distintos circuitos de tráfico que atraviesan la red troncal, y las longitudes de onda de transmisión en las fibras atravesadas. Como se ha demostrado posteriormente, el modo de operación de la red afecta a la planificación de los conmutadores y a sus prestaciones de manera determinante. Por ello, la primera decisión ha sido la adopción de los modos de operación *Shared Wavelength Path* (SHWP) y *Scattered Wavelength Path* (SCWP) estudiados dentro del proyecto WASPNET, como criterio de clasificación entre arquitecturas de conmutación. Esta decisión posibilita una comparativa válida entre diseños, y ayuda a sistematizar el estudio, de una manera no realizada hasta la fecha. La novedad de la aproximación tiene como contrapartida el hecho de que supone, en algunos casos, la necesidad de modificar y completar las propuestas originales de arquitecturas descritas en la literatura, para que los modos de operación puedan ser aplicados. Para ello se ha seguido una secuecia de tres pasos sencillos, pero necesarios: (1) adaptación *hardware* de las arquitecturas al escenario WDM, (2) definición de la planificación SHWP, (3) definición de la planificación SCWP. Como se ha observado en esta tesis doctoral, estos pasos han supuesto, en no pocas ocasiones, nuevos enfoques no planteados en las arquitecturas originales.

El capítulo *2* aborda le evaluación de prestaciones de arquitecturas OPS con capacidad de emular colas a la salida. Lógicamente, estas prestaciones son dependientes del algoritmo de planificación empleado. Para el modo de operación SCWP, se han propuesto dos algoritmos de planificación equivalentes a efectos de prestaciones, que proporcionan los valores óptimos de retardo y probabilidad de pérdida de paquetes alcanzables en este tipo de arquitecturas. Ambos algoritmos son capaces de mantener el orden entre paquetes, aplicando dos criterios distintos. La aplicación del criterio de orden propuesto en [Niz98], lleva a una selección no uniforme de la longitud de onda de transmisión. Las consecuencias negativas de esta no uniformidad que se han destacado son *3*: (1) la no uniformidad de disipación de calor cuando es aplicado en arquitecturas de conmutación basadas en puertas ópticas, (2) la merma en prestaciones cuando se aplica en arquitecturas de conmutación *knock-out*, (3) el proceso de tráfico de salida creado provoca la merma en prestaciones de los nodos vecinos, en el caso de estar basados en la arquitectura de conmutación IB-WR. Por estas razones, se prefiere el algoritmo de planificación SCWP uniforme propuesto. La variación del criterio para el mantenimiento de la secuencia de paquetes, implica la necesidad de un puntero *round-robin* en cada fibra de entrada del conmutador, sin añadir significativamente complejidad al algoritmo.

Las prestaciones de las arquitecturas de conmutación con capacidad de emular colas a la salida han sido evaluadas para ambos modos de operación. Para el modo SHWP, se ha empleado el tratamiento clásico de conmutadores electrónicos de colas a la salida [Hlu88]. Para el modo SCWP, se ha mostrado la equivalencia del sistema a evaluar, con una cola multiservidor finita. El análisis realizado ha permitido obtener las probabilidades de pérdida, la distribución del retardo, y la distribución del periodo ocupado de salida para este tipo de conmutadores, ante tráfico independiente e idénticamente distribuido (IID). Este análisis ha permitido la realización de una comparativa de costes entre las arquitecturas con capacidad de emulación de colas a la salida más relevantes: KEOPS *switch*, *Ouput-Buffered Wavelength-Routed swtich* (OB-WR) y *Space switch*. El *hardware* de las dos primeras arquitecturas ha sido previamente adaptado al escenario WDM, para la aplicación de los modos de operación SHWP/SCWP. Los resultados muestran la reducción de costes que implican los menores requisitos de almacenamiento del modo de operación SCWP. Para valores altos del número de longitudes de onda por fibra (escenario DWDM), este número de retardos decrece hasta *2* ó *3* ranuras temporales, incluso para cargas altas. Esto reduce determinantemente la longitud total de fibra a emplear en las líneas de retardo, punto limitante crítico en la aplicabilidad de la Conmutación Óptica de Paquetes: se han puesto ejemplos significativos que requieren centenares de km. de fibra de retardo en modo SHWP, y únicamente centenares de metros en modo SCWP. En consecuencia, y en todos los casos, el modo de operación SCWP se muestra claramente ventajoso.

El capítulo *3* se centra en la arquitectura de conmutación *Input-Buffered Wavelength-Routed switch* (IB-WR), de especial interés por su menor coste *hardware* respecto a otras arquitecturas. En un primer paso, el diseño de la arquitectura original es adaptado al escenario WDM, para permitir la aplicación de los modos de operación SHWP/SCWP, no contemplados en la propuesta original [Zho98]. La aplicación de ambos modos de operación, plantea un problema singular de planificación para la asignación del retardo a paquetes entrantes, no abordado en la literatura, hasta la propuesta realizada en [Pav03-1].

En esta tesis doctoral, el problema de planificación se formula para ambos modos de operación como un problema de programación dinámica estocástica entera.

Debido a la imposibilidad de encontrar una solución general a este tipo de problemas, se propone el diseño de planificadores SHWP y SCWP que resuelvan un problema simplificado de optimización no dinámica, que encierra los objetivos del problema de asignación peculiar en esta arquitectura. Siguiendo lo publicado en [Pav03-1], se muestra la posibilidad de modelar de manera equivalente esta optimización simplificada, como un conjunto independiente de problemas de emparejamiento en grafos bipartitos. Esta equivalencia, nos ha llevado al estudio de algoritmos de planificación VOQ (*Virtual Output Queueing*), de gran actualidad en la conmutación electrónica de paquetes de altas prestaciones. Aprovechando y adaptando las técnicas de una familia de algoritmos VOQ de exploración paralela, surge la propuesta del algoritmo de planificación SCWP PDBM (*Parallel Desynchronized Block Matching*). Las prestaciones de este algoritmo, junto con las prestaciones de un algoritmo de planificación SCWP de asignación secuencial (no implementable a las velocidades requeridas), son comparadas con el óptimo alcanzable, marcado por las arquitecturas de colas a la salida. El análisis se ha basado en simulaciones empleando la herramienta de libre distribución OMNET. Los resultados obtenidos muestran unas prestaciones muy similares para el algoritmo de asignación secuencial y el algoritmo PDBM, cercanas a las prestaciones de los conmutadores SCWP de colas a la salida, especialmente para un número alto de longitudes de onda por fibra. Asimismo, se ha conseguido demostrar la convergencia del algoritmo PDBM en un máximo de *M* iteraciones, igual al número de retardos, independiente del tamaño del conmutador. Las simulaciones muestran incluso mejores resultados, con una convergencia real en una o a lo sumo dos iteraciones para los casos observados. Estos valores, unidos a los costes comparativamente menores de esta arquitectura, también en su versión de gran escala, intensifican el interés en el estudio de éste y otros algoritmos de planificación. También se confirma la ventaja del modo de operación SCWP en esta arquitectura, para la que no había sido considerado hasta la fecha.

Finalmente, como contribución en este capítulo, se ha mostrado la equivalencia entre la arquitectura WASPNET, y un conmutador IB-WR con una etapa previa de equilibrado de carga. Las ventajas que se obtienen de este equilibrado de carga no son abordadas en este documento de tesis. Sin embargo, resulta curiosa e interesante la simplificación conceptual que se obtiene, al mostrar la equivalencia en la planificación de dos arquitecturas de *hardware* tan aparentemente distinto.

El capítulo *4* versa sobre el estudio de las arquitecturas de Conmutación Óptica de Paquetes de gran escala. El escenario previsto de aplicación es el de red troncal DWDM, con un alto número de longitudes de onda por fibra. Nuestra investigación se concentra en el modo de operación SCWP, de ineludibles ventajas en este escenario por la ganancia de multiplexación que puede proporcionar. El trabajo realizado comienza con un repaso del estado de la técnica, ampliando lo publicado en [Pav02]. En este repaso, resulta curioso observar cómo algunas de las soluciones propuestas para el crecimiento de arquitecturas OPS, aplican técnicas ya empleadas en el crecimiento de conmutadores electrónicos. A continuación, se describen dos tipos de estrategias de crecimiento de arquitecturas, propuestas en [Pav03-4], basadas en el aprovechamiento del efecto *knock-out* (también estudiado en conmutadores electrónicos ATM). Ambas estrategias interconectan una etapa de conmutación sin memoria, con módulos de menor tamaño basados en arquitecturas OPS. La propuesta OFD (*Output Fiber Distributed*) asocia uno de estos módulos para cada fibra de salida. La propuesta OWD (*Output Wavelength Distributed*), asocia un módulo por cada longitud de onda de transmisión. En el caso de emplear una arquitectura con capacidad de emular colas a la salida como módulo de almacenamiento, las prestaciones de ambos conmutadores en modo SCWP son óptimas, aplicando el mismo algoritmo de planificación propuesto en el capítulo *2*, siempre y cuando se

puedan hacer despreciables las pérdidas *knock-out* de la arquitectura. Esto ha motivado el interés en la evaluación de este tipo de pérdidas para ambas estrategias de crecimiento. La evaluación de las pérdidas para la arquitectura OFD puede realizarse fácilmente, independientemente del algoritmo de planificación finalmente considerado. La evaluación para la arquitectura OWD, requiere un análisis más profundo, específico para el algoritmo de planificación SCWP uniforme. Este análisis de la probabilidad de pérdida *knock-out* se ha presentado para tráfico independiente e idénticamente distribuido. Asimismo, se demuestra la cota superior que permite la eliminación de este tipo de pérdidas, señalando su excelente precisión para el dimensionamiento de conmutadores en cargas altas, y el escenario DWDM previsto.

El capítulo *4* finaliza con una comparativa de costes *hardware* de las arquitecturas de gran escala más relevantes, asumiendo el modo de operación SCWP. Las conclusiones de esta comparativa, que deben ser forzosamente muy prudentes, destacan las arquitecturas *knock-out* OWD, IB-WR y *space switch* como las más prometedoras para este escenario.

5.2 Líneas futuras

El futuro despliegue comercial de una red troncal de Conmutación Óptica de Paquetes (*Optical Packet Switching*, OPS) depende de numerosos factores. En esta sección aventuraremos algunos de ellos, junto con las líneas de investigación asociadas, basándonos en la experiencia reunida a lo largo de esta tesis doctoral.

La Conmutación Óptica de Paquetes, ofrece a las operadoras de las redes troncales un aprovechamiento del ancho de banda de transmisión de los enlaces mucho mayor que otras alternativas. Asimismo, el control del reparto del ancho de banda entre fuentes de tráfico con muy distintas demandas, puede realizarse de manera unificada con esquemas simples de control de acceso, frente a por ejemplo las rígidas jerarquías de multiplexación SDH, o la conmutación *Wavelength-Routing*. Más aún, en un escenario previsto en el corto plazo de integración de servicios sobre el protocolo IP, los patrones característicos al tráfico de datagramas serán los dominantes. Esto creará una ventaja competitiva para aquellas opciones que soporten eficientemente este tipo de tráfico. En la red troncal, el esquema que indudablemente mejor se adapta a este escenario es la Conmutación Óptica de Paquetes. En este sentido, una propuesta de transmisión de datagramas IP sobre redes OPS, no existente hasta la fecha, que reúna las características de eficiencia y simplicidad, es un punto de máximo interés.

Las propiedades descritas en el párrafo anterior, son ventajas inherentes a la conmutación OPS. Sin embargo, la conmutación de paquetes tiene asociadas unas desventajas también inherentes a su operación: el mayor coste de los nodos de conmutación de la red, que deben realizar la función de conmutación paquete a paquete. En un nodo de conmutación OPS, esto afecta a la velocidad de operación de los dispositivos ópticos, y a la planificación y control electrónico del conmutador. Las implicaciones en el coste de los dispositivos ópticos, son críticas *en el actual estado de la técnica de estos dispositivos*. Las implicaciones en el coste del control y planificación electrónica del nodo, dependen directamente de la duración temporal de los paquetes, y del tamaño de los conmutadores. No existe una duración de paquete universalmente aceptada hasta la fecha. La aplicación del valor del orden de *1* μ*s* propuesto en el proyecto DAVID [Dit03], supondría que el número de decisiones de planificación por segundo se situase aproximadamente dos órdenes de magnitud por debajo de las posibilidades de los conmutadores electrónicos actuales. Si esta situación de costes relativos y velocidades se mantiene, la planificación electrónica no sería un cuello de botella, ni el coste del equipamiento electrónico un aspecto crítico en los conmutadores OPS. En conclusión, se estima que es la bajada de costes, y la

mejora de la capacidad de integración de dispositivos en sistemas ópticos fiables (con tiempos entre fallos asimilables a los existentes en el equipamiento electrónico), la condición crítica para hacer atractiva la Conmutación Óptica de Paquetes a las empresas operadoras.

A medida que esta posibilidad se acerque, existirá una creciente necesidad de definir finalmente numerosos procesos de la red OPS. A continuación se enumeran algunos de estos aspectos, junto con temas todavía inexplorados que constituyen, a juicio del autor de esta tesis doctoral, líneas futuras de interés investigador.

- **Modo de operación de la red.** El modo de operación es un punto influyente en todos los procesos de la red. En esta tesis doctoral se han mostrado las ineludibles ventajas del modo de operación SCWP, en cuanto a la simplificación de los requisitos de almacenamiento de las arquitecturas de conmutación, y la simplificación de costes *hardware* que implica. Asimismo, la ganancia de multiplexación obtenida por este modo de operación, permite para valores altos del número de longitudes de onda por fibra (*n*), la operación de la red a cargas altas sin por ello aumentar determinantemente el retardo sufrido por los paquetes ópticos. Los resultados muestran el modo SCWP como el previsible candidato elegido.
- **Arquitecturas de conmutación.** La arquitectura de conmutación es un componente crítico en coste. El estudio de distintas alternativas ha sido objeto de investigación en esta tesis doctoral. Como se ha argumentado, el interés en una u otra de las arquitecturas descritas dependerá de la evolución del coste relativo de los componentes fotónicos, el tamaño requerido de los conmutadores (escenario WDM/DWDM), y la necesidad o no, de aplicar técnicas de calidad de servicio, y/o *multicast*. Es de interés el estudio de las prestaciones de éstas y otras arquitecturas, para los patrones de tráfico reales esperados en una red de estas características, punto aún por determinar. Mención especial merece la arquitectura IB-WR, por su menor coste *hardware* asociado, también en arquitecturas de gran escala. Esto intensifica el interés en algoritmos de planificación implementables, que ofrezcan las mejores prestaciones. En este sentido, una línea inexplorada es la posible aplicación de técnicas que recorran el espacio de soluciones al problema de emparejamiento, donde cada iteración consista en permutaciones sencillas, al estilo de algoritmos de planificación VOQ como WWFA [Tam93], o 2DRR [LaM94].

  Por otra parte, las arquitecturas de conmutación OPS de gran escala recibirán en el futuro una creciente atención. Entre las posibles líneas de trabajo se encuentra la evaluación de prestaciones de arquitecturas multiplano, como posibles competidoras de las arquitecturas destacadas en el capítulo *4* de esta tesis doctoral.
- **Tráfico.** La forma en la que el tráfico que debe atravesar la red troncal es ensamblado en paquetes ópticos, marcará el patrón de tráfico que será conmutado por los nodos OPS. Además de para el control de la tasa contratada, los nodos frontera son lugares perfectos para la inclusión de técnicas de conformado de tráfico. ¿Qué criterios deben aplicarse para este conformado? ¿Es necesario? ¿Qué tráficos resultan más favorables para los nodos de conmutación? Respecto a la elección de longitud de onda de transmisión en modo SCWP, ¿qué criterios son de aplicación, teniendo en cuenta su efecto en las prestaciones de arquitecturas como IB-WR?

- **Desorden extremo a extremo.** No existe una respuesta a la siguiente pregunta: ¿es la condición de mantener el orden extremo a extremo una mayor fuente de ventajas que de problemas? ¿Qué criterios pueden ser aplicados para mantener el orden entre paquetes simultáneos en modo SCWP? Hemos visto criterios sencillos y válidos que permiten mantener el orden en arquitecturas de colas a la salida empleando la información de longitud de onda de transmisión. Estos mismos criterios sin embargo, no parecen de aplicación sencilla para las arquitecturas IB-WR. ¿Debe atribuirse a los nodos de interconexión de la red troncal la condición de mantener el orden, deben ser los nodos frontera los encargados de un reordenamiento basado en un contador en la cabecera del paquete, o debe eliminarse completamente de la red troncal la obligación de mantener la secuencia entre paquetes ópticos?
- **Control de la red.** A juicio de este autor, los mecanismos de señalización de la red OPS no han recibido suficiente atención hasta la fecha. La familia de especificaciones GMPLS (*Generalized Multi-Protocol Label Switching*) se baraja como el mecanismo de control para el establecimiento de las conexiones y sus rutas. Esto permitiría el aprovechamiento de las investigaciones realizadas en distintos protocolos de señalización y control, para poder maximizar las ventajas de asignación de ancho de banda de OPS.
- **Supervivencia de red.** La existencia de algoritmos efectivos de recuperación ante fallos de la red, es condición *sine qua non* para la llegada de la Conmutación Óptica de Paquetes a una red troncal comercial. Este tema ha sido objeto de argumentación en [Niz98], estimando una posible mayor complejidad al proceso de recuperación en redes OPS para el modo SCWP. Sin embargo, en opinión del autor de esta tesis doctoral, el modo de operación SCWP tiene también ciertas ventajas en la recuperación de redes OPS frente al modo SHWP. La caída de un canal (longitud de onda) de una fibra en modo SHWP obliga al reprovisionamiento inmediato de todos los circuitos virtuales establecidos a través de ella. Esta misma caída, para el modo SCWP, es equivalente a la pérdida de *1/n* parte del ancho de banda del enlace, que no impide el tráfico de paquetes de todos los circuitos virtuales que atraviesan la fibra. Por ello, en caso de ser necesario, el reprovisionamiento de estos circuitos no estaría sujeto a límites temporales tan estrictos. La supervivencia de redes operando en modo SCWP es, en opinión del autor, un punto de gran interés investigador.
- **Calidad de servicio.** Las redes troncales se caracterizan generalmente por unas distancias geográficas entre nodos, que implican tiempos de propagación en el orden de milisegundos. En las redes OPS, un tamaño de paquete óptico del orden de *1* μs, y un retardo para el modo de operación SCWP menor en muchos a casos a *5* ranuras temporales, hacen despreciable el retardo de conmutación frente al tiempo de propagación. Por ello, la implementación de mecanismos de calidad de servicio en los nodos de interconexión, que puedan ofrecer retardos de conmutación menores a tráficos prioritarios, no parece de gran interés. Estos mecanismos sí pueden ser aplicados en los nodos frontera, donde su impacto en las prestaciones será mayor, tanto en la conformación de paquetes ópticos, como en la extracción de los datos en el nodo de salida de la red troncal. Hasta el conocimiento del autor, no existe ningún trabajo en este sentido.
- **Multicast.** El conmutador óptico KEOPS admite de forma natural la transmisión de tráfico *multicast*, sin cambios en el *hardware*. Resulta de interés plantear algoritmos de planificación que permitan explotar estas

funcionalidades. Los estudios en el campo de *multicast* para arquitecturas OPS son marginales hasta la fecha [Fis02], e inexistentes (hasta el conocimiento del autor) en el caso del modo de operación SCWP.

- **Convergencia OPS-OBS.** La Conmutación Óptica de Ráfagas (*Optical Burst Switching*) se encuentra en un estado de investigación más avanzado que la Conmutación Óptica de Paquetes. Sin embargo, esta alternativa es vista en distintos foros como una solución intermedia, que puede no llegar a ser implantada. En este sentido, se ha argumentado [COST266] a favor de una posible convergencia entre las especificaciones OBS y OPS. La realidad de esta afirmación dependerá en gran medida de las propuestas que se realicen para la transmisión de tráfico IP en la red de transporte OPS/OBS.

Referencias bibliográficas

[And93] Anderson T., Owicki S., Saxe J., Thacker C., “High speed switch scheduling for local area networks”, *ACM Transactions on Computer Systems*, vol. 11, no. 4, November 1993, pp. 319-352.

[Awd99] Awduche W. *et al.*, "Multiprotocol Lambda Switching: Combining MPLS Traffic Engineering Control with Optical Cross-Connects", Internet Draft, draft-awduche-mpls-te-optical-01.txt, Nov. 1999.

[Bar93] Barry R. A., Humblet P., “Latin routers, design and implementation”, *Journal of Lightwave Technology*, vol. 11, no. 5/6, May/June 1993, pp. 891-899.

[Bat03] Battestilli T., Perros H., “An Introduction to Optical Burst Switching”, *IEEE Communications Magazine,* vol. 41, no. 8, August 2003, pp. S10-S15.

[Bon76] J. Bondy, U. Murty, *Graph theory with applications,* North-Holland, New York, 1976.

[Bon01] Bonenfant P., Rodriguez-Moral A., “Framing Techniques for IP over Fiber”, *IEEE Networks*, vol. 15, no. 4, July/August 2001, pp. 12-18.

[Bos97] Bostica B., “Synchronization issues in optical packet switched networks”, *Photonic Networks*

[Bre03] Bregni S., Pattavina A., Vegetti G., “Architectures and Performance of AWG-Based Optical Switching Nodes for IP Networks”, *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 7, September 2003, pp. 1113-1025.

[Bru02] Bruce E., “Tunable Lasers”, *IEEE Spectrum,* February 2002, pp. 35-39.

[Cal97] Callegati F., “Which packet length for a transparent optical network?”, in *Broadband Networking Technologies* SPIE Proceedings Vol. 3233, 1997, pp. 260-271.

[Cal00-1] Callegati F., “Optical Buffers for variable length packets”, *IEEE Communication Letters*, vol. 4, no.9, September 2000, pp. 292-294.

[Cal00-2] Callegati F., Corazza G., Raffaelli C., “Design of a WDM Optical Packet Switch for IP Traffic”, Proc. IEEE GLOBECOM 2000, San Francisco, November 2000, pp. 1283-1287.

[Cal02] Callegati F., Corazza G., Raffaelli C., “Exploitation of DWDM for Optical Packet Switching With Quality of Service Guarantees”, *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 1, January 2002, pp. 190-201.

[Chi99] Chia M. C., Hunter D. K., Andonovic I., Ball P., Wright I., "Feedback Arrayed-Waveguide Gratings-Based Optical Packet Switch With Improved Homowavelength Crosstalk Performance", Fifth Asia-Pacific Conference on Communications/Fourth Optoelectronics and Communications Conference, paper 151October 1999, Beijing, China.

[Chi01-1] Chia M., Hunter D., Andonovic I., Ball P., Wright I., Ferguson S., Guild K., O’Mahony M., "Packet loss and delay performance of feedback and feed-forward arrayed-waveguide gratings-based optical packet switches with WDM inputs-outputs", *IEEE Journal of. Lightwave Technology*, vol. 19, no. 9, Sept. 2001, pp. 1241-1254.

[Chi01-2] Chiaroni D., Le Sauze N., Zami T., Emery J.Y., “Semi-conductor optical amplifiers: A key technology to control the packet power variation”, in Proc. 27th European Conf. Optical Communication (ECOC’01), Amsterdam, The Netherlands, Oct. 2001, pp. 314-315.

[Chl96] Chlamtac I., Fumagalli A., Kazovsky L. G., Melman P., Nelson W. H., Poggiolini P., Cerisola M., Masum A., Fong T. K., Hofmeister R. T., Lu C., Mekkittikul A., Sabido D., Suh C., Wong E., “CORD: Contention Resolution by Delay Lines”, *IEEE Journal on Selected Areas in Communications,* vol. 14, no. 5, June 1996, pp. 1014-1029.

[Cho95] Choi Y., Tode H., Okada H., Ikeda H., “A Large Capacity Photonic ATM Switch for Wavelength Division Multiplexing Networks”, Proc. International Conference on Computer Communications and Networks (ICCCN’95), Las Vegas (USA), pp. 414-419.

[COST266] Report on the COST266/OPTIMIST [WWW]. Workshop.http://www.ist-optimist.org/pdf/workshops/ ONDM2003/WS_ONDM2003_minutes.pdf.

[Dan98] Danielsen S. L., Joergensen C., Mikkelsen B., Stubkjaer K., "Analysis of a WDM packet switch with improved performance under bursty traffic conditions due to tunable wavelength converters", *IEEE J. Lightwave Technol.*, vol. 16, no. 5, May 1998, pp. 729-735.

[Dit03] Dittman L., Develder C., Chiaroni F., Neri F., Callegati F., Koerber W., Stavdas A., Renaud M., Rafel A., Sole-Pareta J., Cerroni W., Leligou N., Dembeck L., Mortensen B., Pickavet M., Le Sauze N., Mahony M., Berde B., Eilenberger G., “The European IST Project DAVID: A Viable Approach Toward Optical Packet Switching”, *IEEE Journal on Selected Areas in Communications,* vol. 21, no. 7, Sep. 2003, pp. 1026-1040.

[Dur96] T. Durhuus, Mikkelsen B., Jorgensen C., Danielsen S.L., Stubkjaer K.E., "All-optical wavelength conversion by semiconductor optical amplifiers", *IEEE Journal of Lightwave Technology*, vol. 14, June 1996, pp. 942-954.

[Elb02] El-Bawab T. S., Shin J., "Optical Packet Switching in Core Networks: Between Vision and Reality", *IEEE Communications Magazine*, September 2002, pp. 60-65.

[Fis02] T. Fischer, “On Optical Packet-Switched Support for IP Multicast”. En Proc. of 3rd VDE-ITG Symposium on Photonic Networks 2002, pp. 19-24.

[Gui97] Guillemot C., Le Corre A., Kervarec J., Henry M., Simon J.C., Luron A., Vuchener C., Lamouler P., Gravey P., “Optical packet switch demonstrator assessment: Packet delineation and fast wavelength routing”, in Proc. IOOC/ECOC’97, vol. 3, Edinbourgh, U.K., 1997, pp. 343-346.

[Gui98] Guillemot C., Renaud M., Gambini P., Janz C., Andonovic I., Bauknecht R., Bostica B., Burzio M., Callegati F., Casoni M., Chiaroni D., Clérot F., Danielsen S., Dorgeuille F., Dupas A., Franzen A., Hansen P., Hunter D., Kloch A., Krähenbühl R., Lavigne B., Le Corre A., Raffaelli C., Schilling M., Simon J., Zucchelli L., "Transparent optical packet switching: the European ACTS KEOPS project approach", *IEEE Journal of Lightwave Technology*, vol. 16, no. 12, Dec. 1998, pp. 2117-2134.

[Gui99] Guild K. M., O’Mahony M. J., “A Novel Routing and Buffering Architecture in an All-Optical Switching Node”, FT3/CLEO/PACRIM ’99, pp. 1278-1280.

[Haa93] Haas Z., “The “Staggering Switch”: An Electronically Controlled Optical Packet Switch”, *IEEE Journal of Lightwave Technology,* vol. 11, no. 5/6, May/June 1993.

[Han98] Hansen P. B., Danielsen S. L., Stubkjaer E., “Optical Packet Switching without Packet Alignment”, Proc. ECOC 1998, Madrid, Spain, September 1998, paper WdD13.

[Hlu88] Hluchyj M., Karol M., “Queueing in high-performance packet switching”, *IEEE Journal on Selected Areas in Communications.,* vol. 6, no. 9, Dec. 1988, pp. 1587-1597.

[Hui90] Hui J., *Switching and traffic theory for integrated broadband networks*, Kluwer Academic Publishers, 1991.

[Hun98-1] Hunter D. K., Cornwell W. D., Gilfedder T. H., Franzen A., Andonovic I., “SLOB: A Switch with Large Optical Buffers for Packet Switching”, *IEEE Journal of Lightwave Technology*, vol. 16, no. 10, October 1998, pp. 1725-1736.

[Hun98-2] Hunter D. K., Chia C., Andonovic I., "Buffering in Optical Packet Switches", *Journal of Lightwave Technology*, vol. 16, no. 12, December 1998, pp. 2081-2094.

[Hun99] Hunter D., Nizam M., Chia M., Andonovic I., Guild K., Tzanakaki A., O’Mahony J., Bainbridge J., Stephens M., Penty R., White I., "WASPNET: A Wavelength Switched Packet Network", *IEEE Communications Magazine,* vol. 37, no. 3, March 1999, pp. 120-129.

[Hun00] Hunter D. K., Andonovic I., "Approaches to Optical Internet Packet Switching", *IEEE Communications Magazine*, September 2000, pp. 116-122.

[IETF03-1] Choi J. K., Kang M. H., Choi J. Y., Lee G. M., Cha Y. W., “General Switch Management Protocol (GSMP) v3 for Optical Support”, GSMP IETF Working Group Internet Draft, *draft-ietf-gsmp-optical-spec-o2.txt*, Junio 2003.

[IETF03-2] Choi J. K., Kang M. H., Choi J. Y., Noh T., Hahm K. R., “Requirements of Optical Burst-Level Control in Optical Networks”, CCAMP IETF Working Group Internet Draft, *draft-choi-burst-control-00.txt*, Junio 2003.

[Inf91] Infante Macías, R., *Métodos de programación matemática (Tomo II), 2º Edición* Universidad Nacional de Educación a Distancia, 1991.

[Jac96] Jacob J. B., Casoni M., Corazza G., Raffaelli C., Masetti F., Parmentier P., "System Design and Evaluation of a Large Modular Photonic ATM Switch", *European Transactions on Telecommunications*, vol. 7, no. 6, Nov.-Dec. 1996, pp. 565-573.

[Jia01] Jiang Y., Hamdi M., “A fully-desynchronized round-robin matching scheduler for VOQ packet switched architecture”, Proc. of 2001 Workshop on High Performance Switching and Routing, May 2001, Dallas (USA), pp. 407-411.

[Jou01] Jourdan A., Chiaroni D., Dotaro E., Eilenberger G. J., Masetti F., Renaud M., "The Perspective of Optical Packet Switching in IP-Dominant Backbone and Metropolitan Networks", *IEEE Communications Magazine*, March 2001, pp. 136-141.

[Kal92] Kalman R.F., Kazovsky L.G., Goodman J.W., "Space division switches based on semiconductor optical amplifiers", *IEEE Photon. Technol. Lett.*, vol. 4, Sept. 1992, pp. 1048-1051.

[Kar87] Karol M., Hluchyj M., “Input versus output queueing on a space division switch”, *IEEE Transactions on Communications*, vol. 35, no. 12, December 1987, pp. 1347-1356.

[Ker97] Kervarec J., Guillemot C., Henry M., Kandouci M., Simon J.C., Gravey P., “Optical packet wavelength-routing demostrator: Architecture and synchronization issues”, Proc. of Photon. Switching, Stockholm, Sweden, 1997, pp. 184-187.

[Kle75] Kleinrock L., *Queueing systems. Volume I: Theory*, John Wiley & Sons, 1975.

[LaM94] LaMaire R.O., Serpanos D.N., “Two-Dimensional Round-Robin Schedulers for Packet Switches with Multiple Input Queues”, *IEEE/ACM Transactions on Networking,* vol. 2, no. 5, October 1994, pp. 471 –481.

[McK99] McKeown N., “The iSLIP Scheduling Algorithm for Input-Queued Switches”, *IEEE/ACM Transactions on Networking*, vol. 7, no. 2, April 1999, pp. 188-201.

[Muk97] Mukherjee B., *Optical Communications Networks*, McGraw-Hill, 1997.

[Mur02] Murthy C., Gurusamy M., *WDM optical networks. Concepts, design and algorithms,* Prentice Hall PTR, 2002.

[Neu01] Neukermans A., Ramaswami R., "MEMS Technology for Optical Networking Applications", *IEEE Communications Magazine* , vol. 39, no. 1, Jan. 2001, pp. 62-69.

[Niz98] Nizam M.H.M., Hunter D.K., Andonovic I., "Designing an optimum WDM transport network: control architectures, node requirements and performance", *Proc. Soc. Photo-Optical Instrumentation Engineers (SPIE),* vol. 3531, Oct. 1998, pp. 244-255.

[Nog98] Noguchi K., "Optical Free-Space Multichannel Switches Composed of Liquid-Crystal Light Modulator Arrays and Birefringent Crystals", *IEEE Journal of Lightwave Technology*, vol. 16, no. 8, Aug. 1998, pp. 1473-1481.

[NTT00] "Very-Large-Scale 256-Channel WDM Filter for Future Photonic Networks", NTT Research and Development Review of Activities, 2000. http://www.ntt.co.jp/RD/OFIS/active/2000pdfe/ct31_e.pdf

[OMa01] O'Mahony M. J., Simeonidou D., Hunter D. K., Tzanakaki A., "The Application of Optical Packet Switching in Future Communication Networks", *IEEE Communications Magazine*, March 2001, pp. 128-135.

[Pav02] Pavon-Marino P., Garcia-Haro J., Malgosa-Sanahuja J., "Scaling strategies survey for envisaged backbone optical packet switches", Proc. of IASTED Comm. Systems and Networks (CSN 2002), Malaga, España, Sep. 2002, pp. 178-183.

[Pav03-1] Pablo Pavon-Marino, Joan Garcia-Haro, Josemaria Malgosa-Sanahuja, Fernando Cerdan, "Maximal Matching Characterization of Optical Packet Input-Buffered Wavelength Routed Switches", Proc. of 2003 IEEE Workshop on High Performance Switching and Routing (HPSR 2003), Torino (Italy), June 2003, pp. 55-60.

[Pav03-2] Pavon-Marino P., Garcia-Haro J., Malgosa-Sanahuja J., Cerdan F., "Optical Packet Switching Fabrics Comparison Under SCWP/SHWP Operational Modes", Proc. of 8th IEEE Symposium on Computers and Communications (ISCC'2003), vol. 1, Antalya, Turkey, July 2003, pp. 547-553.

[Pav03-3] Pablo Pavón Mariño, Joán García Haro, Josemaría Malgosa Sanahuja, Fernando Cerdán, "Algoritmo Óptimo de Selección de Longitud de Onda en Arquitecturas de Conmutación Óptica de Paquetes SCWP", Proc. de las las IV Jornadas de Ingeniería Telemática (Jitel 2003), Gran Canaria (España), Septiembre 2003, pp. 153-160.

[Pav03-4] Pavon-Mariño P., Garcia-Haro J., Malgosa-Sanahuja J., Cerdan F., "A Performance Study of a Knock-Out Large-Scale Optical Packet Switching Architeture Under Scattered Wavelength Path Operational Mode", Proc. of 2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM 2003), Victoria (Canada), August 2003, pp. 478-481.

[Pav03-5] P. Pavon-Marino, J. Garcia-Haro, J. Malgosa-Sanahuja, F. Cerdan, "Scattered Versus Shared Wavelength Path Operation, Application to Output Buffered Optical Packet Switches. A comparative study", *SPIE/Kluwer Optical Networks Magazine,* vol. 4, no. 6, November/December 2003, pp. 134-145.

[Pav03-6] P. Pavon-Marino, J. Garcia-Haro, J. Malgosa-Sanahuja, F. Cerdan, "A computational efficient method for the Study of the per-wavelength arrival process in SCWP Optical Packet Switching Architectures", enviado a *IEEE Communication Letters,* Septiembre 2003.

[Qia99] Qiao C., Yoo M., "Optical Burst Switching (OBS): A New Paradigm For an Optical Internet", *J. High Speed Networks, Special Issues on Optical Networking*, vol. 8, no. 1, Jan. 1999, pp. 69-84.

[Qia00] Qiao C., Yoo M., "Choices, Features and Issues in Optical Burst Switching (OBS)", *Optical Networks Magazine*, vol. 1, no. 2, April 2000, pp. 36-44.

[Raf00-1] Raffaelli C., "Design of a multistage optical packet switch", *European Transactions on Telecommunications,* vol. 11, no. 5, Septiembre 2000, pp. 443-451.

[Raf00-2] Raffaelli C., "Design of a Core Switch for an Optical Transparent Packet Network", *Photonic Network Communications,* vol. 2, no. 2, December 2000, pp. 123-133.

[Ram98] Ramaswami R., Sivarajan K., *Optical Networks: A Practical Perspective*, Morgan Kaufmann, 1998.

[Ram01] Ramamurthy B., *Design of Optical WDM Netowrks. LAN, MAN and WAN Architectures*, Kluwer Academic Publishers, 2001.

[Ren98] Renaud M., Keller D., Sahri N., Silvestre S., Prieto D., Dorgeuille F., Pommereau F., Emery J.Y., Grard E., Mayer H.P., "SOA-based optical network components", in Proc. 51st Electronic Components and Technology Conf. ECTC'01, Lake Buena Vista, FL, May 1, 2001, pp. 433-438.

[Sas93] Sasayama K., Habara K., Zhong W., Yukimatsu K., "Photonic ATM switch using frequency-routing-type time-division interconnection network", *Electronic Letters*, vol. 29, no. 20, September 2003, pp. 1778-1780.

[Sas95] Sasayama K., "Multihop Frontiernet using generalised perfect shuffle interconnection topology", *Electronic Letters,* vol. 31, no. 13, June 1995, pp. 1087-1088.

[Sas97] Sasayama K., Yamada Y., Habara K., Yukimatsu K., "FRONTIERNET: frequency-routing-type time-division interconnection network", *IEEE Journal of Lightwave Technology*, vol. 15, no. 3, March 1997, pp. 417-419.

[Sex97] Sexton M., *Broadband networking: ATM, SDH and SONET*, Artech House, 1997.

[Sha00] N. Shari et al., "A highly integrated 32-SOA gates optoelectronic module suitable for IP multi-terabit optical packet routers", in Proceedings OFC 2001, Anaheim, CA, USA, Feb. 2001.

[Siv00] Sivalingam K., Subramaniam S., *Optical WDM Networks. Principles and Practice*, Kluwer Academic Publishers, 2000.

[Ste99] Stern T., Bala K., Multiwavelength Optical Networks: A Layered Approach, Addison-Wesley, 1999.

[Sto71] Stone H. S., "Parallel processing with the perfect shuffle", *IEEE Trans. Comput.,* vol. C-20, February 1971, pp. 153-161.

[Stu00] Stubkjaer K. E., "Semiconductor Optical Amplifier-based All-Optical Gates for High-Speed Optical Processing", *IEEE Journal Sel. Quant. Elec.*, vol. 6, 2000, pp. 1428-35.

[Tan00] Tancevski L., Yegnanarayanan S., Castañon G., Tamil L., Masetti F., McDermott T., "Optical Routing of Asynchronous Variable Length Packets", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, October 2000, pp. 2084-2093.

[Tak90] Takahashi H., Suzuki S., Kato K., Nishi I., "Arrayed-waveguide grating for wavelength division multi/demultiplexer with nanometre resolution", *Electronic Letters*, vol. 26, 1990, pp. 87-88.

[Tam93] Tamir Y., Chi H.-C. "Symmetric crossbar arbiters for VLSI communication switches", *IEEE Transactions on Parallel and Distributed Systems,* vol. 4, no. 1, 1993, pp. 13–27.

[Tur88] Turner J. S., "Design of a Broadcast Packet Switching Network", *IEEE Transactions on Communications*, vol. 36, no. 6, June 1988, pp. 734-743.

[Tur97] Turner J. S., "Terabit Burst Switching", Tech. report WUCS-97-49, Dept. of Comp. Sci., Washington University, St. Louis, MO, Dec. 1997.

[Var99] Varga A., “Using the OMNeT++ Discrete Event Simulation System in Education”, *IEEE Transactions on Education*, vol. 42, no. 4, November 1999, pp. 372.

[Ven01] Venkatesh S., Fouquet J.E., Son J.W., Hoffman P.F., Guo H., Price K., Hengstler S., "Recent advances in bubble-actuated cross-connect switches", Proc. of CLEO/Pacific Rim 2001, vol. 1, pp. I_414-I_415.

[Vin96] Vinck B., Bruneel H., “Delay Analysis of Multiserver ATM buffers”, *IEE Electronic Letters*, vol. 32, no. 15, July 1996, pp. 1352-1353.

[Whi02] White I., Penty R., Webster M., Chai Y., Wonfor A., Shahkooh S., "Wavelength Switching Components for Future Photonic Networks", *IEEE Communications Magazine*, September 2002, pp. 74-81.

[Xu01] Xu L., Perros H. G., Rouskas G.m, "Techniques for Optical Packet Switching and Optical Burst Switching", *IEEE Communications Magazine*, January 2001, pp. 136-142.

[Yam98] Yamada Y., Sasayama K., Habara K., Misawa A., Tsukada M., Matsunaga T., Yukimatsu K., “Optical Output Buffered ATM Switch Prototype Based no FRONTIERNET Architecture”, *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 7, September 1998, pp. 1298-1308.

[Yao00] Yao S., Mukherjee B., Dixit S., “Advances in photonic packet switching: an overview”. *IEEE Communications Magazine*, Feb. 2000, pp. 84-94.

[Yao02] Yao S., Xue F., Mukherjee B., Yoo B., Dixit S., “Electrical ingress buffering and traffic aggregation for optical packet switching and their effect on TCP-level performance in optical mesh networks”, *IEEE Communications Magazine*, vol. 40, no. 9, Sep. 2002, pp. 66-72.

[Yeh87] Y. S. Yeh, M. G. Hluchyj, A. S. Acampora, "The knock-out switch: A simple, modular architecture for High Performance Packet Switching", *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, no. 8, October 1987, pp. 223-231.

[Zho98] Zhong W., Tucker R., "Wavelength routing-based photonic packet buffers and their applications in photonic packet switching systems", *IEEE Journal of Lightwave Technology*, vol. 16, no. 10, Oct. 1998, pp. 1737-1745.

[Zuc96] Zuccelli L., Burzio M., Gambini P., “New solutions for optical packet delineation and synchronization in optical packet switched networks”, in Proc. ECOC’96, vol. 3, Oslo, Norway, 1996, pp. 301-304.