Chapter 1

SCHEDULING ALGORITHMS FOR UNICAST, MULTICAST, AND BROADCAST

George N. Rouskas
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-7534
rouskas@csc.ncsu.edu

Abstract In this chapter we present a survey of algorithms for scheduling packet traffic in broadcast optical WDM networks. We first describe the context and motivations of the scheduling problem. We then review the current literature in the field with an emphasis on scheduling techniques for providing best-effort service as well as guaranteed service for both unicast and multi-destination traffic. We provide alternative formulations of the problem, and we compare the formulations and theoretical results, as well as algorithms and heuristics.

Keywords: Broadcast optical networks, Wavelength division multiplexing (WDM), Scheduling algorithms, Open shop scheduling, Quality of service (QoS), Multicast

1. INTRODUCTION

The broadcast WDM network architecture has been widely studied as an approach to building optically interconnected networks. Under one widely adopted scenario for the evolution of the optical network infrastructure (Kam et al., 1998), broadcast WDM subnetworks will be used to provide local access, and the subnetworks will be interconnected through wavelength routed MANs and WANs. One of the potentially difficult issues that arise in a broadcast WDM environment, is that of coordinating the various transmitters/receivers. Some form of coordination is necessary because (a) a transmitter and a receiver must both be tuned to the same channel for the duration of a packet's transmission, and (b) a

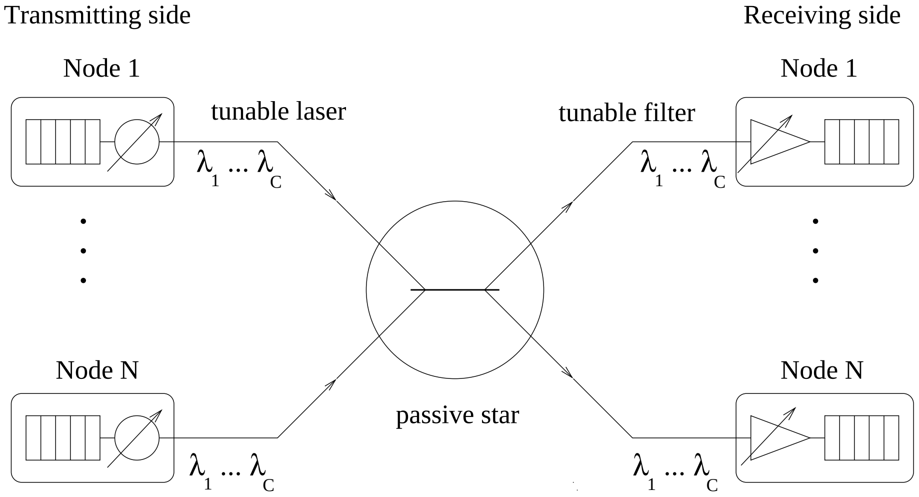
Transmitting side
Receiving side
Node 1
Node 1
tunable laser
tunable filter
$\lambda_1 \dots \lambda_C$
$\lambda_1 \dots \lambda_C$
Node N
Node N
passive star
$\lambda_1 \dots \lambda_C$
$\lambda_1 \dots \lambda_C$

Figure 1.1 A broadcast WDM network with N nodes and C channels

simultaneous transmission by one or more nodes on the same channel will result in a collision. The issue of coordination is further complicated by the fact that tunable transceivers may need a non-negligible amount of time to switch between wavelengths. Thus, at the heart of media access control (MAC) protocols for broadcast WDM subnetworks is a scheduling algorithm responsible for coordinating access to the available channels (wavelengths).

In this chapter we survey a number of algorithms for traffic scheduling in a packet-switched broadcast WDM with N nodes and C channels, $N \geq C$, as shown in Fig. 1.1. Unless otherwise specified, it is assumed that each node has exactly one tunable transmitter and one tunable receiver. We let Δ denote the transceiver tuning latency, i.e., the time it takes a transmitter or receiver to tune from one wavelength to another. Packets in the network are of fixed size. Time is slotted, with a slot time equal to the packet time plus some guard band, which depends on the MAC protocol and the corresponding scheduling algorithm.

The chapter is organized as follows. In Section 2. we discuss the role of scheduling within the context of broadcast WDM networks, and we examine its relationship to reservation protocols and load balancing. In Section 3. we review scheduling algorithms for both best-effort and guaranteed-service unicast traffic. In Section 4. we discuss approaches to scheduling multi-destination traffic. We conclude the chapter in Section 5..

2. LOAD BALANCING, RESERVATIONS, AND SCHEDULING

While in this chapter we are mainly concerned with scheduling algorithms, one should keep in mind that, in ensuring an acceptable level of network performance, scheduling is only one piece of the puzzle. In this section we briefly review two other components critical to the operation of broadcast WDM networks, namely, load balancing and reservation protocols, and we discuss their relationship to scheduling.

We distinguish two levels of network operation, differing mainly in the time scales at which they take place. At the *media access control* level, connectivity among the network nodes is provided by a reservation protocol, whose main function is to collect information regarding traffic demands, and a scheduling algorithm whose objective is to provide collision-free communication among the nodes while optimizing some performance measure of interest (e.g., schedule length). At the *network dimensioning* level, which takes place at significantly longer time scales, the objective is to allocate resources in a way that optimizes network performance. In this context, the shared resource of interest is bandwidth, and load balancing algorithms are needed to ensure good performance and fairness at this level of network operation.

2.1 LOAD BALANCING AND RECONFIGURATION

In optical WDM networks, each channel will have to be shared by multiple receivers, and the problem of assigning receive wavelengths arises. A wavelength assignment (hereafter referred to as WLA) implies an allocation of the bandwidth to the various network nodes. Intuition suggests that if the traffic load is not well balanced across the available channels, the result will be poor network performance. A recent study on the performance of the HiPeR-ℓ reservation protocol (Sivaraman and Rouskas, 1997) has confirmed this intuition. Let us define parameter ϵ_b such that no channel carries more than $\frac{(1+\epsilon_b)}{C}$ times the total traffic offered to the network. In other words, ϵ_b is a measure of the *degree of load balancing* of the network; under perfect load balancing, $\epsilon_b = 0$. It was shown in (Sivaraman and Rouskas, 1997) that the maximum sustained throughput γ (i.e., the number of packets successfully transmitted per packet time) is directly affected by ϵ_b through the following stability condition:

$$\gamma \quad < \quad \frac{C}{(1+\epsilon_b)(1+\epsilon_s)}. \tag{1.1}$$

It can be seen from (1.1) that the higher the degree of load balancing (i.e., the lower the value of ϵ_b is), the higher the overall arrival rate γ that the network can accommodate, and vice versa. Parameter ϵ_s is the guarantee on the schedule length and depends on the scheduling algorithm used. In other words, even an optimal algorithm (for which $\epsilon_s = 0$ in (1.1)) will achieve a very low throughput if the load is not well balanced. Although the stability condition (1.1) was derived specifically for HiPeR-ℓ, we believe that load balancing has a similar effect on the performance of any protocol for WDM broadcast networks.

Hence, the time-varying conditions expected in this type of environment call for mechanisms that periodically adjust the bandwidth allocation to ensure that each channel carries an almost equal share of the corresponding offered load. The problem of dynamic load balancing by retuning a subset of receivers in response to changes in the overall traffic pattern was studied in (Baldine and Rouskas, 1998; Baldine and Rouskas, 1999b). Assuming an existing WLA and some information regarding the new traffic demands, this work studied two approaches to obtaining a new WLA such that (a) the new traffic load is balanced across the channels, and (b) the number of receivers that need to be retuned is minimized. The latter objective is motivated by the fact that tunable receivers take a non-negligible amount of time to switch between wavelengths during which parts of the network are unavailable for normal operation. Since this variation in traffic is expected to take place over larger time scales (i.e., retuning will be a relatively infrequent event), employing slowly tunable devices can be a cost effective solution. An approximation algorithm for the load balancing problem was presented that provides for tradeoff selection, using a single parameter, between two conflicting goals, namely, the degree of load balancing and the number of receivers that need to be retuned.

The issues arising in the reconfiguration phase of broadcast networks were also studied in (Baldine and Rouskas, 1999a), where reconfiguration policies to determine *when* to reconfigure the network were developed, and an approach was presented to carry out the network transition by describing a class of strategies that determine *how* to retune the optical receivers. The problem was formulated as a Markovian Decision Process, yielding a systematic and flexible framework in which to view and contrast reconfiguration policies, and it was shown how an appropriate selection of reward and cost functions can be used to achieve the desired balance among various performance criteria of interest.

2.2 RESERVATION PROTOCOLS

Reservation protocols collect information about the short-term traffic demands between source-destination pairs, by having nodes periodically transmit reservation packets. The reservation packets are usually sent on a separate control channel dedicated to carrying signaling information. To prevent the control channel from becoming a bottleneck, it is also possible to use multiple control channels (Humblet et al., 1993) or employ in-band reservations, as in HiPeR-ℓ (Sivaraman and Rouskas, 1997), in which case all channels in the network may be used to carry both data and control packets. Reservation packets can contain information about the head-of-line packet only, as in PROTON (Levine and Akyildiz, 1995), or about the depth of all queues in a node, as in HiPeR-ℓ. The reservation information is used to build an identical (delayed) snapshot of the state of the queues in the network at each node (Foo and Robertazzi, 1995; Muir and Garcia-Luna-Aceves, 1996). This snapshot is the main input to the scheduling algorithms discussed in the remainder of this chapter.

Reservation protocols may employ pipelining techniques to mask the effects of the tuning latency (Tridandapani et al., 1994) or the propagation delay (Sivaraman and Rouskas, 1997), both of which may otherwise have severe impact on overall performance in ultra high-speed WDM environments. Some reservation protocols organize time in distinct reservation and data phases (Sivalingam and Dowd, 1995), while in others (Sivaraman and Rouskas, 1997) there is no separate reservation phase and reservation information is multiplexed with data packets. The latter approach has the advantage that data transmission is not interrupted for large periods of time (i.e., during a separate reservation phase). It also makes it possible to transmit data while at the same time computing the schedule for the next transmission phase, effectively masking the computation time of the schedule.

3. SCHEDULING OF UNICAST TRAFFIC

3.1 BEST-EFFORT TRAFFIC

The scheduling problem in a broadcast optical network with N nodes and $C,\ N \geq C$, data channels is typically formulated as a matrix clearing problem. Specifically, it is assumed that there exists a $N \times C$ *traffic demand matrix* $\mathbf{M}=[m_{ic}]$, where integer m_{ic} represents the number of packets to be transmitted from node $i,\ i = 1, \cdots, N$, on channel $\lambda_c,\ c = 1, \cdots, C$. The traffic demands m_{ic} may be derived as: $m_{ic} = \sum_{j \in R_c} a_{ij}$, where the number a_{ij} of packets to be transmitted from node i to node j can be obtained using a reservation protocol such as the ones discussed in the previous section, and the sets of receivers R_c listening on a certain wavelength λ_c, $c = 1, \cdots, C$, (i.e., the WLA) are determined by the load balancing algorithm.

Given a traffic demand matrix $\mathbf{M}$, the objective most commonly considered in the literature is to construct an *optimal finish time* (OFT) schedule, i.e., one which has the least finish time among all schedules for matrix $\mathbf{M}$. An OFT schedule is highly desirable since it both minimizes average packet delay and maximizes the aggregate network throughput (recall the effect of parameter ϵ_s in the stability condition (1.1)). However, if no restriction is imposed on the number of reservations submitted by the nodes (that is, the values that quantities m_{ic} may take), the length of even the OFT schedule can become very large under high loads. In this case, while the average packet delay and aggregate network throughput will be optimal, the length and variability of the schedules make it impossible to provide guarantees (e.g., on delay and/or delay jitter) to individual packet flows. Thus, this approach is appropriate for best-effort traffic, but not well-suited to support real-time services.

When the transceiver tuning latency Δ is assumed to be small relative to the packet transmission time, a padding equal to Δ time units can be included within each slot to allow the transceivers sufficient time to switch between wavelengths, with minimal effects on overall performance (Bogineni et al., 1993; Humblet et al., 1993). In this case, the matrix clearing problem is equivalent to the *open shop scheduling* problem where *preemption* is allowed (Gonzalez and Sahni, 1976). The open shop scheduling problem formulation prevents channel, transmitter, and receiver collisions by including constraints which guarantee that, within each time slot (a) two or more sources do not transmit on the same channel, (b) a given source transmits on at most one channel, and (c) a given receiver listens on at most one channel. In the context of broadcast WDM networks, a preemptive schedule is such that there may exist a node-channel pair (i, λ_c) for which the m_{ic} packets are not transmitted in contiguous slots. In other words, the node i of any such pair (i, λ_c) will have to tune to channel λ_c multiple times in order to transmit all its traffic demands m_{ic}. Preemptive OFT schedules for the open-shop scheduling problem can be constructed in polynomial time using techniques for maximum matching on bipartite graphs (Rouskas and Ammar, 1995; Gonzalez and Sahni, 1976).

For networks where the value of the tuning latency is comparable to, or greater than, the packet transmission time, including a padding of Δ time units within each slot would be highly inefficient in terms of both throughput and delay. A better approach is to keep the slot time equal to the packet time, and introduce a new set of constraints to account for the time it takes a transceiver to tune from one wavelength to another, during which it is taken off-line and is not available for transmitting or receiving packets. The objective, then, becomes that of minimizing the impact of the tuning requirements on the length of the schedule. However, adding the new constraints introduces significant difficulty to the scheduling problem, and it makes it impossible to obtain an OFT schedule in polynomial time.

One approach to alleviating the effects of the tuning latency is to insist on *non-preemptive* schedules, whereby each node i tunes its transceiver to each channel λ_c exactly once during the schedule, and remains at that channel until all m_{ic} packets have been transmitted. The non-preemptive open shop scheduling problem with $\Delta = 0$ admits a polynomial-time solution when the number of channels $C = 2$, but it is NP-complete for $C \geq 3$ (Gonzalez and Sahni, 1976). When the tuning latency Δ is non-zero, however, the problem becomes NP-complete even when $C = 2$ (Rouskas and Sivaraman, 1997). We now discuss several heuristics and approximation algorithms for constructing near-optimal schedules.

In (Rouskas and Sivaraman, 1997), the design of non-preemptive open-shop schedules for broadcast WDM networks with non-uniform traffic demands and arbitrary tuning latencies was undertaken. Two distinct regions of network operation were identified. The *tuning-limited* region is such that the schedule length is determined by the transceiver tuning requirements. When the network operates in the *bandwidth-limited* region, the length of the schedule is determined by the traffic demands. The point at which the network switches between the two regions was also identified in terms of system parameters such as the number of nodes and channels and the tuning latency. A special class of schedules was then introduced such that the order in which the various transceivers tune to each channel is the same for all channels. This class of schedules permits an intuitive formulation of the scheduling problem, and, under uniform traffic (i.e., when $m_{ic} = m \ \forall \ i, c$), an OFT schedule within this class can be readily constructed. Based on the new formulation, polynomial-time algorithms were developed to construct OFT schedules when the elements of the traffic demand matrix $\mathbf{M}$ satisfy certain optimality conditions. In essence, the optimality conditions impose an upper bound on the "degree of non-uniformity" of matrix $\mathbf{M}$ for the algorithm to construct an OFT schedule within this class. A set of heuristics was also developed which, in the general case (that is, when matrix $\mathbf{M}$ does not satisfy the optimality conditions), where shown to construct schedules of length very close to the lower bound. An important outcome of this work was the realization that algorithms which work well within the bandwidth-limited region may not work well within the tuning-limited region, and vice versa. Consequently, optimality conditions, optimal algorithms, and heuristics were developed for both bandwidth-limited and tuning-limited networks.

An important feature of the algorithms and heuristics in (Rouskas and Sivaraman, 1997) is that their running-time complexity is a function only of system parameters, namely, the number N of nodes and the number C of wavelengths, and is independent of the actual length of the schedule. This property makes it possible to allow the transmission of variable-length packets over the broadcast WDM network without any extra control overhead. This can be accomplished by letting the slot time be equal to one byte, and having the nodes send reservation requests for the number of *bytes* they wish to transmit, rather than the number of *fixed-size packets*. The algorithms will then schedule each node's transmission in a number of contiguous bytes (slots). Having the nodes make reservations in terms of number of bytes eliminates the problem of selecting the length for the fixed-size packets, and it also eliminates the overhead for segmenting and then reassembling the upper layer variable-length packets (e.g., IP datagrams). The problem of determining the "best" fixed length for packets is a difficult one since it strongly depends on the (mostly unknown) mix of applications that will be carried over the network, and may lead to non-optimal compromises (e.g., as in the size of ATM cells). On the other hand, it would be inefficient to use algorithms whose running time is a function of the schedule length (in slots) in a network where the slot size is equal to one byte.

An approximation algorithm for constructing non-preemptive open shop schedules was developed in (Choi et al., 1996). The algorithm is based on the well-known concept of list scheduling. Specifically, as soon as a transmitter completes its transmissions on a given wavelength, it tunes to the wavelength in which it can start transmitting at the *earliest time*. It was shown that the length of a schedule constructed by this algorithm is at most twice the length of the OFT schedule for a given matrix $\mathbf{M}$, for any value of the tuning latency Δ.

A different two-phase heuristic for the same problem was derived in (Borella and Mukherjee, 1996). In the first phase, nodes are assigned to transmit in contiguous slots on a given channel, in decreasing order of their demands for that channel. The assignment ensures that all channel, transmitter, and tuning constraints are satisfied. Since this approach may result in unused slots in which no node has been assigned to transmit, the second phase of the algorithm attempts to fill these slots. Specifically, for each unused slot, the nodes that are assigned to transmit in slots immediately before or after the unused slot are examined. If the tuning constraints allow, one of these nodes is assigned to transmit in the previously unused slot. The transmissions allocated during the second pass are in addition to the demands specified by matrix $\mathbf{M}$, and can greatly increase the efficiency of the schedule by decreasing the number of unused slots.

Special cases of the general open shop scheduling problem have also been addressed in (Pieris and Sasaki, 1994; Azizoglu et al., 1996; Choi et al., 1996). The *all-to-all* scheduling problem is a special case such that each transmitter has exactly one packet to send to each receiver. Under such uniform traffic, and assuming that the number N of nodes is a multiple of the number C of wavelengths, the optimal WLA is one in which exactly N/C receivers are tuned to each channel. Consequently, the traffic matrix $\mathbf{M}$ is such that $m_{ic} = N/C \ \forall \ i, c$. For this traffic matrix, lower and upper bounds on the schedule length were derived in (Pieris and Sasaki, 1994), and a scheduling algorithm was presented. This algorithm was shown in (Choi et al., 1996) to be optimal. It is interesting to note that the schedules constructed by this algorithm fall within the class of schedules considered in (Rouskas and Sivaraman, 1997). A different special case was studied in (Azizoglu et al., 1996). Specifically, the traffic demands were such that each transmitter has either one packet or no packet to send to each receiver (representing the existence or not, respectively, of a head-of-line packet at the various queues), and the value of the tuning latency Δ was restricted to be at most equal to the packet transmission time. A heuristic based on a variation of the list scheduling algorithm in (Choi et al., 1996) was analyzed through simulations and was shown to exhibit good average case behavior for this problem.

All the algorithms discussed so far have the same objective, namely, they attempt to construct OFT non-preemptive open shop schedules. In such schedules, the tuning and transmission periods are interleaved so as to minimize the overall finish time for a given traffic matrix. (A performance analysis of non-preemptive open-shop schedules under packet traffic has been carried out in (McKinnon et al., 1998b; McKinnon et al., 1999; McKinnon et al., 1998a).) Another approach to scheduling packet transmissions in WDM networks is to construct schedules satisfying the *tune-transmit separability constraint* (Pieris and Sasaki, 1994). Specifically, time is divided into alternating periods of transmission and tuning. Each transceiver operates on a fixed channel during a transmission period; no packets are transmitted during the tuning periods, which are reserved to retune transceivers to be ready for the next transmission period.

This version of the problem is closely related to the well-known scheduling problem in satellite-switched time division multiple access (SS/TDMA)

(Gopal and Wong, 1985; Inukai, 1979). More formally, the problem can be defined as follows. Given a traffic demand matrix $\mathbf{M}$, the objective is to decompose it into sub-matrices $\mathbf{M}_k,\ k = 1, \cdots, K$, such that (a) each row and each column of each sub-matrix $\mathbf{M}_k$ has at most one non-zero element, (b) $\sum_{k=1}^{K} \mathbf{M}_k = \mathbf{M}$, (c) the total time to sequentially transmit the individual sub-matrices is minimized, and (d) the number K of sub-matrices is minimized. In this formulation, each sub-matrix $\mathbf{M}_k$ corresponds to a transmission period within the schedule. Requirement (a) ensures that there are no receiver or transmitter collisions within each transmission period, while requirement (b) ensures that all the traffic demands of matrix $\mathbf{M}$ will be met by following the transmissions indicated by the sub-matrices. Objective (c) reflects the desire to minimize the time it takes to clear matrix $\mathbf{M}$, while objective (d) is necessary to keep the time spent tuning the transceivers between transmission as short as possible; together, the two objectives ensure that the total time spent transmitting and tuning is minimized. As defined, this problem is NP-hard (Gopal and Wong, 1985). Next, we discuss various heuristics in the context of broadcast WDM networks.

In (Ganz and Gao, 1994), the network was viewed as a bipartite graph of N sources and N destinations (in other words, the starting point for the decomposition is not the matrix $\mathbf{M}$ we have considered so far, but rather the $N \times N$ matrix of traffic demands between each source-destination node). A bipartite matching algorithm was used to decompose this graph into a number K of bipartite matchings, where each matching is constrained to have at most C arcs. No bounds on the performance of the algorithm were derived. In (Choi et al., 1996), a network with demand matrix $\mathbf{M}$ was modeled as a bipartite multi-graph with N sources and C destinations. The bipartite multi-graph is first edge-colored, and then decomposed into subgraphs consisting of edges of the same color. The transmissions corresponding to the edges of a subgraph can all take place simultaneously, similar to the transmissions in a sub-matrix $\mathbf{M}_k$ in the above formulation. A bound on the length of the schedule constructed by the algorithm was derived, and the average case behavior of this approach was studied through simulations. Also, a lower bound on the length of all-to-all schedules satisfying the tune-transmit separability constraints was derived in (Pieris and Sasaki, 1994).

A different approach to matrix decomposition was taken in (Sivalingam and Wang, 1996), where the focus was on the running-time efficiency and ease-of-implementation of the scheduling algorithm. Another important feature of the techniques developed in (Sivalingam and Wang, 1996) is that the schedule is built using partial information as it becomes available to the reservation protocol, without the need to wait until the entire traffic demand matrix $\mathbf{M}$ is complete. Specifically, as soon as a transmitter's reservation requests are received, all nodes in the network use the same greedy strategy to schedule the requests within an existing sub-matrix, if one that can accommodate the requests is found. Otherwise, a new sub-matrix is created for the transmitter's requests. A bound on the number K of sub-matrices generated by the algorithm was derived, and its performance in terms of average packet delay and network throughput was studied for both uniform and client-server traffic.

While most algorithms that have appeared in the literature attempt to minimize the finish time of a schedule, a different objective in scheduling packet transmissions was considered in (Kam et al., 1998). Achieving *max-min fairness* was the primary concern of this work, taking precedence over maximizing throughput or minimizing delay. The algorithm developed uses information on whether a source is back-logged or idle, and builds the schedule slot by slot using the following greedy approach. To determine the transmissions in a given slot, each source is considered in increasing order of the number of packets sent by the source so far. If the source is back-logged, it is assigned to transmit in the slot if no transmitter, receiver, or channel constraints are violated by doing so. Otherwise, the next source is considered until either all channels have been assigned transmissions or no back-logged sources remain. Simulation studies presented in (Kam et al., 1998) demonstrate that the algorithm has good fairness properties while also achieving high throughput.

3.2 GUARANTEED-SERVICE TRAFFIC

Packet-switched WDM networks will need to support a range of applications with varying quality of service (QoS) requirements, such as bandwidth and delay guarantees. The algorithms discussed in the previous section focus on minimizing the finish time of the schedule, and do not address the issue of supporting time-constrained communication. From the point of view of scheduling, the requirement to provide QoS guarantees necessitates algorithms which can transmit packets in some priority order, e.g., according to deadlines, virtual finish times, eligibility times, or other time-stamps associated with a packet (Liu and Layland, 1973).

Under the assumption of negligible tuning latency, the problem of scheduling real-time packet flows in WDM networks is related to the problem of scheduling periodic tasks in a real-time multiprocessor system (Dertouzos and Mok, 1989). In (Wang et al., 1997), the problem of scheduling *isochronous* message streams was considered, where each stream l is characterized by its deadline D_l, and the maximum number C_l of packets that can arrive in any time interval of length D_l (the "computation time" in multiprocessor scheduling terminology). A *feasible* schedule for a set of message streams is such that exactly C_l slots are allocated to stream l in any time window of size D_l. An algorithm based on the *rate-monotonic* principle (Liu and Layland, 1973) was applied to schedule a static set of isochronous message streams. The algorithm may not be successful in constructing feasible schedules when the deadlines D_l are not multiples of a basic value $D \geq 2$, however, no sufficient condition for schedulability was derived. The dynamic problem was also considered, and a set of algorithms was presented to schedule transmissions of new message streams, as well as to deallocate slots assigned to terminating streams.

The problem of optimally scheduling periodic tasks on multiprocessors was studied in (Jackson and Rouskas, 1998). The existence of a feasible schedule for this problem when the total task density $\rho = \sum_l (C_l/D_l) = C$, where C is the number of processors (channels in the corresponding WDM problem), has been an open problem since the work in (Dertouzos and Mok, 1989). It was shown in (Jackson and Rouskas, 1998) that the condition $\rho \leq C$ is both necessary and sufficient for the existence of a feasible schedule. A network flow formulation was also presented, based on which an algorithm to construct a feasible schedule was developed. In addition to broadcast networks, this algorithm can have applications to scheduling packet traffic on WDM point-to-point links between routers.

An algorithm which provides a minimum bandwidth guarantee to packet flows was presented in (Kam et al., 1998; Kam and Siu, 1998). This algorithm is in fact an extension of the max-min fair algorithm discussed in the previous section. The main difference is that traffic flows are considered for slot allocation in increasing order of the *excess* bandwidth they have used beyond their guaranteed bandwidth. By letting the guaranteed bandwidth of best-effort traffic be zero, this algorithm can be used to provide both bandwidth guarantees and max-min fairness. Thus, this approach represents a first step towards supporting integrated services in a broadcast WDM environment.

4. SCHEDULING OF MULTI-DESTINATION TRAFFIC

Many applications and telecommunication services, including teleconferencing, distributed data processing, and video distribution, require some form of multipoint communication. Traditionally, without network support for multicasting, a multi-destination message is replicated and transmitted individually to all its recipients. This method, however, consumes more bandwidth than necessary. Bandwidth consumption constitutes a problem since most multipoint applications require a large amount of bandwidth. An alternative solution is to broadcast a multi-destination message to all nodes in the network. The problem in this case is that nodes not addressed in the message will have to dedicate resources to receive and process the message. Thus, the ability to efficiently transmit messages addressed to multiple destinations has become increasingly important, and the issues associated with providing network support for multipoint communication have been widely studied within a number of different networking contexts (Cruz et al., 1996).

In WDM broadcast networks, information transmitted on any channel is broadcast to the entire set of nodes, but it is only received by those with a receiver listening on that channel. The broadcast feature, coupled with tunability at the receiving end, makes it possible to design scheduling algorithms (Rouskas and Ammar, 1997; Borella and Mukherjee, 1995) such that a *single* transmission of a multicast packet can reach all receivers in the packet's destination set simultaneously. The high degree of efficiency in using the network resources makes this approach especially appealing for transmitting multicast traffic. However, the design of appropriate receiver tuning algorithms is complicated by the fact that (a) tunable receivers take a non-negligible amount of time to switch between channels, and (b) different multicast groups may have several receivers in common. On the other hand, waiting until all receivers become available before scheduling a multicast packet may result in low wavelength throughput (i.e., low average number of packets transmitted per unit time), especially for medium to large size multicast groups. To improve the situation, it was proposed in (Jue and Mukherjee, 1997) to partition a multicast group into several sub-groups, and to transmit a packet once to each sub-group. This approach leads to higher wavelength throughput despite the fact that each packet is transmitted multiple times, indicating the existence of a tradeoff between wavelength throughput and the degree of efficiency in using the bandwidth.

An approach similar to the one in (Jue and Mukherjee, 1997) was presented in (Modiano, 1998). Specifically, a packet is also transmitted multiple times, until it is received by all members of its multicast group. Instead of partitioning the multicast group in advance, however, each receiver follows a set of rules to listen to a packet transmission in each slot. An analytical model for obtaining the average packet delay was developed for two schemes, one employing persistent transmissions and one that introduces a random back-off delay. Under the first scheme, a packet is continuously transmitted until it is received by all members of its multicast group. The random back-off scheme eliminates the head-of-line problem of persistent transmissions by retransmitting packets not received by all intended receivers after a random delay, and results in better performance. Also, it was shown that the algorithm used by the receiver to select one among multiple packets addressed to it can have a significant impact on performance. Specifically, an algorithm where the receiver selects the packet with the smallest number of intended receivers remaining outperforms one in which packets are selected based on the time of their initial transmission (i.e., a first-come first-served discipline).

The problem of scheduling multicast traffic was also considered in (Ortiz et al., 2000; Ortiz et al., 1997). Let a *multicast completion* denote the completion of the transmission of a multicast packet to all receivers in its multicast group. The *multicast throughput*, defined as the average number of multicast transmissions per slot, was introduced as the performance measure of interest, and it was shown that it depends on two measures that have previously been considered in isolation, namely, the degree of efficiency in using the channel bandwidth and wavelength throughput. Then, a new technique was presented for the transmission of multicast packets based on the concept of a *virtual* receiver, a set of physical receivers which behave identically in terms of tuning. It was demonstrated that the number of virtual receivers naturally captures the performance of the system in terms of multicast throughput. By partitioning the set of all physical receivers into virtual receivers, a multicast packet must be transmitted to each virtual receiver containing a physical receiver in the packet's multicast group, and the original network with multicast traffic is transformed into a new network with unicast traffic. This approach decouples the problem of determining how many times each multicast packet should be transmitted, from the problem of scheduling the actual packet transmissions. Thus, rather than developing new scheduling algorithms for multicast traffic, one may take advantage of the algorithms discussed in the previous section. Consequently, the focus of the work in (Ortiz et al., 1997) was on the problem of optimally selecting the virtual receivers to maximize multicast throughput, and it was proven that it is NP-complete. Finally, four heuristics of varying degree of complexity were presented for obtaining a set of virtual receivers that provide near-optimal performance in terms of multicast throughput.

In (Ortiz et al., 1998) the performance of various strategies for scheduling a combined load of unicast and multi-destination traffic was studied. The performance measure of interest was schedule length. Three different scheduling strategies were presented, namely: separate scheduling of unicast and multicast traffic, treating multicast traffic as a number of unicast messages, and treating unicast traffic as multicasts of size one. A lower bound on the schedule obtained by each strategy was first obtained. Subsequently, the strategies were compared against each other using extensive simulation experiments in order to establish the regions of operation, in terms of a number of relevant system parameters, for which each strategy performs best. The main conclusions were as follows. Multicast traffic can be treated as unicast traffic under very limited circumstances. On the other hand, treating unicast traffic as multicast traffic produces short schedules in most cases. Alternatively, scheduling and transmitting each traffic separately is also a good choice.

5. CONCLUDING REMARKS

We have reviewed algorithms for scheduling unicast and multi-destination traffic in broadcast WDM networks. A classification of the scheduling algorithms is presented in Tables 1.1 and 1.2.

Scheduling of best-effort traffic is a well-researched problem, and many efficient algorithms have been developed that give optimal or near-optimal results. More work is needed in the area of scheduling algorithms for providing QoS guarantees to real-time traffic, especially when the tuning latency must be taken into account. With the current interest on packet (especially IP) over WDM architectures, it would also be important to develop integrated approaches that fit within the Internet's differentiated services framework. The main challenge, however, is in the deployment of optical WDM packet network testbeds, such as the WDM LAN extension to the wideband all-optical network (Kaminow et al., 1996) at MIT and the Helios IP over WDM joint testbed between MCNC and North Carolina State University, which will provide opportunities for extensive experimentation with, and validation and extension of the proposed scheduling algorithms and heuristics.

Table 1.1 Classification of Scheduling Algorithms for Unicast Traffic

Objective	Δ	**Schedule Class**	**Algorithm**
QoS Guarantees	≈ 0	Periodic Task	(Jackson and Rouskas, 1998) (Wang et al., 1997)
Max-Min Fairness		Open Shop with preemption	(Kam and Siu, 1998) (Kam et al., 1998)
Minimize Length			(Rouskas and Ammar, 1995) (Choi et al., 1996, Sec. III.A)
	small	Open Shop without preemption	(Azizoglu et al., 1996)
	arbitrary		(Rouskas and Sivaraman, 1997) (Borella and Mukherjee, 1996) (Pieris and Sasaki, 1994)
		Tune Transmit Separability	(Choi et al., 1996, Sec. III.B) (Sivalingam and Wang, 1996) (Ganz and Gao, 1994)

Table 1.2 Classification of Scheduling Algorithms for Multi-Destination Traffic

Δ	**Approach**	**Algorithm**
≈ 0	Repeated transmissions	(Modiano, 1998)
	Single transmission of multicast packet	(Rouskas and Ammar, 1997)
arbitrary		(Borella and Mukherjee, 1995)
	Partition groups	(Jue and Mukherjee, 1997)
	Virtual Receivers (VR)	(Ortiz et al., 1997)
	VR for unicast & multicast	(Ortiz et al., 1998)

References

Azizoglu, M., Barry, R. A., and Mokhtar, A. (1996). Impact of tuning delay on the performance of bandwidth-limited optical broadcast networks with uniform traffic. *IEEE Journal on Selected Areas in Communications*, 14(5):935–944.

Baldine, I. and Rouskas, G. N. (1998). Dynamic load balancing in broadcast WDM networks with tuning latencies. In *Proceedings of INFOCOM '98*, pages 78–85. IEEE.

Baldine, I. and Rouskas, G. N. (1999a). Dynamic reconfiguration policies for WDM networks. In *Proceedings of INFOCOM '99*, pages 313–320. IEEE.

Baldine, I. and Rouskas, G. N. (1999b). Reconfiguration and dynamic load balancing in broadcast WDM networks. *Photonic Network Communications*, 1(1):49–64.

Bogineni, K., Sivalingam, K. M., and Dowd, P. W. (1993). Low-complexity multiple access protocols for wavelength-division multiplexed photonic networks. *IEEE Journal on Selected Areas in Communications*, 11(4):590–604.

Borella, M. and Mukherjee, B. (1995). A reservation-based multicasting protocol for WDM local lightwave networks. In *Proceedings of ICC '95*, pages 1277–1281.

Borella, M. S. and Mukherjee, B. (1996). Efficient scheduling of nonuniform packet traffic in a WDM/TDM local lightwave network with arbitrary transceiver tuning latencies. *IEEE Journal on Selected Areas in Communications*, 14(5):923–934.

Choi, H., Choi, H.-A., and Azizoglu, M. (1996). Efficient scheduling of transmissions in optical broadcast networks. *IEEE/ACM Transactions on Networking*, 4(6):913–920.

Cruz, R., Hill, G., Kellner, A., Ramaswami, R., Sasaki, G., and (Eds.), Y. Y. (1996). Special issue on optical networks. *IEEE Journal Selected Areas in Communications*, 14(5).

Dertouzos, M. L. and Mok, A. K.-L. (1989). Multiprocessor on-line scheduling of hard-real-time tasks. *IEEE Transactions on Software Engineering*, 15(12):1497–1506.

Foo, E. M. and Robertazzi, T. G. (1995). A distributed global queue transmission strategy for a WDM optical fiber network. In *Proceedings of INFOCOM '95*, pages 154–161.

Ganz, A. and Gao, Y. (1994). Time-wavelength assignment algorithms for high performance WDM star based networks. *IEEE Transactions on Communications*, 42(4):1827–1836.

Gonzalez, T. and Sahni, S. (1976). Open shop scheduling to minimize finish time. *Journal of the Association for Computing Machinery*, 23(4):665–679.

Gopal, I. and Wong, C. (1985). Minimizing the number of switchings in an SS/TDMA system. *IEEE Transactions on Communications*, 33(6):497–501.

Humblet, P. A., Ramaswami, R., and Sivarajan, K. N. (1993). An efficient communication protocol for high-speed packet-switched multichannel networks. *IEEE Journal on Selected Areas in Communications*, 11(4):568–578.

Inukai, T. (1979). An efficient SS/TDMA time slot assignment algorithm. *IEEE Transactions on Communications*, 27(10):1449–1455.

Jackson, L. E. and Rouskas, G. N. (1998). Optimal scheduling of periodic tasks on multiple identical processors. Technical Report TR-98-14, North Carolina State University, Raleigh, NC.

Jue, J. and Mukherjee, B. (1997). The advantages of partitioning multicast transmissions in a single-hop optical WDM network. In *Proceedings of ICC '97*, pages 427–431.

Kam, A. C. and Siu, K.-Y. (1998). A real-time distributed scheduling algorithm for supporting QoS over WDM networks. In *Proceedings of SPIE*, volume 3531, pages 181–193.

Kam, A. C., Siu, K.-Y., Barry, R. A., and Swanson, E. A. (1998). Toward best effort services over WDM networks with fair access and minimum bandwidth guarantee. *IEEE Journal Selected Areas in Communications*, 16(7):1024–1039.

Kaminow, I. P., Doerr, C. R., Dragone, C., Koch, T., Koren, U., Saleh, A. A. M., Kirby, A. J., Ozveren, C. M., Schoffield, B., Thomas, R. E., Barry, R. A., Castagnozzi, D. M., Chan, V. W. S., Hemenway, B. R., MArquis, D., Parikh, S. A., Stevens, M. L., Swanson, E. A., Finn, S. G., and Gallager, R. G. (1996). A wideband all-optical WDM network. *IEEE Journal Selected Areas in Communications*, 14(5):780–799.

Levine, D. A. and Akyildiz, I. F. (1995). PROTON: A media access control protocol for optical networks with star topology. *IEEE/ACM Transactions on Networking*, 3(2):158–168.

Liu, C. L. and Layland, J. W. (1973). Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the ACM*, 20(1):46–61.

McKinnon, M. W., Perros, H. G., and Rouskas, G. N. (1999). Performance analysis of broadcast WDM networks under IP traffic. *Performance Evaluation*, 36-37:333–358.

McKinnon, M. W., Rouskas, G. N., and Perros, H. G. (1998a). Performance analysis of a photonic single-hop ATM switch architecture with tunable transmitters and fixed frequency receivers. *Performance Evaluation*, 33(2):113–136.

McKinnon, M. W., Rouskas, G. N., and Perros, H. G. (1998b). Queueing-based analysis of broadcast optical networks. In *Proceedings of ACM SIGMETRICS/PERFORMANCE '98*, pages 121–130. ACM.

Modiano, E. (1998). Unscheduled multicasts in WDM broadcast-and-select networks. In *Proceedings of INFOCOM '98*.

Muir, A. and Garcia-Luna-Aceves, J. J. (1996). Distributed queue packet scheduling algorithms for WDM-based networks. In *Proceedings of INFOCOM '96*, pages 938–945.

Ortiz, Z., Rouskas, G. N., and Perros, H. G. (1997). Scheduling of multicast traffic in tunable-receiver WDM networks with non-negligible tuning latencies. In *Proceedings of SIGCOMM '97*, pages 301–310. ACM.

Ortiz, Z., Rouskas, G. N., and Perros, H. G. (1998). Scheduling of combined unicast and multicast traffic in broadcast WDM networks. In *Proceedings of PICS '98*, pages 137–150. Chapman & Hall.

Ortiz, Z., Rouskas, G. N., and Perros, H. G. (2000). Maximizing multicast throughput in WDM networks with tuning latencies using the virtual receiver concept. *European Transactions on Telecommunications*, 11(1):63–72.

Pieris, G. R. and Sasaki, G. H. (1994). Scheduling transmissions in WDM broadcast-and-select networks. *IEEE/ACM Transactions on Networking*, 2(2):105–110.

Rouskas, G. N. and Ammar, M. H. (1995). Analysis and optimization of transmission schedules for single-hop WDM networks. *IEEE/ACM Transactions on Networking*, 3(2):211–221.

Rouskas, G. N. and Ammar, M. H. (1997). Multi-destination communication over tunable-receiver single-hop WDM networks. *IEEE Journal on Selected Areas in Communications*, 15(3):501–511.

Rouskas, G. N. and Sivaraman, V. (1997). Packet scheduling in broadcast WDM networks with arbitrary transceiver tuning latencies. *IEEE/ACM Transactions on Networking*, 5(3):359–370.

Sivalingam, K. and Dowd, P. (1995). A multi-level WDM access protocol for an optical interconnected multi-processor system. *IEEE/OSA Journal of Lightwave Technology*, 13(11):2152–2167.

Sivalingam, K. and Wang, J. (1996). Media access protocols for WDM networks with on-line scheduling. *IEEE/OSA Journal of Lightwave Technology*, 14(6):1278–1286.

Sivaraman, V. and Rouskas, G. N. (1997). HiPeR-ℓ: A High Performance Reservation protocol with ℓook-ahead for broadcast WDM networks. In *Proceedings of INFOCOM '97*, pages 1272–1279. IEEE.

Tridandapani, S., Meditch, J. S., and Somani, A. K. (1994). The MaTPi protocol: Masking Tuning times through Pipelining in WDM optical networks. In *Proceedings of INFOCOM '94*, pages 1528–1535.

Wang, B., Hou, C.-J., and Han, C.-C. (1997). On dynamically establishing and terminating isochronous message streams in wdma-based local area lightwave networks. In *Proceedings of INFOCOM '97*, pages 1263–1271.