Dynamic Routing and Wavelength Assignment Using First Policy Iteration

Esa Hyytiä and Jorma Virtamo
Helsinki University of Technology
Laboratory of Telecommunications Technology
P.O.Box 3000, FIN-02150 HUT, Finland
{*esa.hyytia, jorma.virtamo*}*@hut.fi*

Abstract

With standard assumptions the routing and wavelength assignment problem (RWA) can be viewed as a Markov Decision Process (MDP). The problem, however, defies an exact solution because of the huge size of the state space. Only heuristic algorithms have been presented up till now. In this paper we propose an approach where, starting from a given heuristic algorithm, one obtains a better algorithm by the first policy iteration. In order to estimate the relative costs of states, we make a simulation on the fly studying, at each decision epoch, the consequences of all the alternatives actions. Being computationally intensive, this method can be used in real time only for systems with slow dynamics. Offline it can be used to assess how close the heuristic algorithms come to the optimal policy. Numerical examples are given about the policy improvement.

1. Introduction

The wavelength division multiplexing (WDM) is a promising technology for future all-optical networks. In WDM several optical signals using different wavelengths share same fibre. The capacity of such fibre links can be huge, even terabits per second. The routing in network nodes is based on wavelengths of incoming signals [1][2][3]. Generally, the routing and wavelength assignment (RWA) problem in WDM networks consists of choosing a route and a wavelength for each connection so that no two connection using the same wavelength share the same fibre [4][5].

When the traffic is not static, lightpath requests arrive randomly following some traffic pattern. Connection requests between a given source destination pair constitute a *traffic class*, which we index by k, $k \in \mathcal{K}$, where $\mathcal{K}$ is the set of all source destination pairs. The RWA algorithm configures the lightpaths in the network unless there is no enough resources available and the request is blocked (see e.g. [6][7][8]). Here we also assume that reconfiguration of the lightpaths is not possible.

Several heuristic algorithms have been proposed and studied (see e.g. [6][7][8]). In this paper we study this problem in the setting of Markov Decision Processes (MDP) and propose a new approach, where we try to improve any given heuristic algorithm by the first policy iteration [9][10]. The policy iteration, indeed, is known to lead to a new policy with better performance. In order to avoid dealing with huge size of the state space in calculating the relative state costs needed in the policy improvement step, we suggest to estimate these costs on the fly by simulations for the limited set of states that are relevant at any given decision epoch, i.e. when the route and wavelength assignment for an arriving call has to be made.

The rest of the paper is organized as follows. In section 2 we briefly review Markov Decision Processes and policy iteration in general, and the first policy iteration, in particular. In section 3, we consider the relative costs of states and how they are used in the policy iteration, and in the following section 4 we study how theses state costs can be estimated by simulations. Different heuristic RWA algorithms are presented in section 5. These are used as a starting point for policy iteration, and, in section 6 some numerical results obtained by simulations are presented. Finally, section 7 contains conclusions.

2. Policy Iteration

Routing and wavelength allocation constitute a typical decision making problem. When certain events occur, one has to decide on some action. In the RWA problem, in particular, upon arrival of a request for a new connection one has to decide whether or not to accept the request, and if accepted which resources to allocate for it, i.e. which of available routes and wavelengths are used for that connection. In general, one is interested in the optimal policy which maximizes or minimizes the expectation (infinite time horizon) of a given objective function. Here we assume that the ob-

jective is defined in terms of minimizing some cost function. The cost may represent e.g. the loss of revenue due to blocked calls, where different revenue may be associated to each type of call.

When the arrival process of type k calls is a Poisson process with intensity λ_k, the holding times of those calls are distributed exponentially with mean $1/\mu_k$ and the expected revenue per carried call is w_k, then the system constitutes a Markov Process and the problem of determining the optimal policy belongs to the class of Markov Decision Processes (MDP) described e.g. in [9] and [10]. Three main approaches for solving the optimal policy in the MDP setting are the policy iteration, value iteration and linear programming approach. In this paper, we concentrate on the iteration in the policy space, where, as the name says, one tries to find the optimal policy by starting from some policy and iteratively improving it. This policy iteration is known to converge rather quickly to the optimal policy.

At each decision epoch, i.e. arrival of a new request, there is a finite set of possible actions: either reject the call or accept it and assign a feasible combination of route and wavelength (RW) to it. A *policy* defines for each possible state of the system and for each class k of an arriving call which of the possible actions is taken. Many heuristic policies have been proposed in the literature such as the first-fit wavelength and most-used wavelength policies combined with shortest path routing or near shortest path routing. Some of them work reasonably well. Common to all heuristic policies is that they are simple. The choice of the action to be taken at each decision epoch can usually be described in simple terms and does not require much computation. We take one of the heuristic policies as a starting point and call it the *standard policy*. The policy resulting from the first policy iteration we refer to as the *iteration policy*.

By doing the first policy iteration we have two goals in mind, 1) finding a better RWA algorithm which, being computationally intensive, may or may not be calculable in real time, depending on the time scale of the dynamics of the system, 2) even in the case the algorithm is not calculable in real time, estimating how far the performance of a heuristic algorithm is from the optimal one.

Briefly, as explained in more detail below, our idea in the policy iteration is the following: at each decision epoch we make a decision analysis of all the alternative actions. For each of the possible actions, i.e. decision alternatives, we estimate the future costs by simulation. Thus, assuming that a given action is taken we let the system proceed from the state where it is after that action and use the standard policy to make all the subsequent decisions. The iteration policy is the policy which is obtained when at each decision epoch the action is chosen for which the estimated cost is the minimum. It can be shown that the iteration policy is always better or at least as good a policy as the standard policy, it often comes rather close to the optimal policy.

3. Relative costs of states

In the MDP theory, the first policy iteration consists of the following steps: With the standard policy one solves the Howard equations (see, e.g. [9][10]) to give the so called relative costs of the states, C_i, which for each possible state i of the system describe the difference in the expected cumulative cost from time 0 to infinity, given that the system starts from state i rather than from the equilibrium. Then, given that the current state of the system is j and a class-k call is offered, one calculates the cost $C_j + w_k$ for the action that the call is rejected, and the cost C_i, $i \in \mathcal{A}(j,k)$, for the case the call is accepted, where $\mathcal{A}(j,k)$ is the set of states reachable from state j by assigning class k call a feasible RW pair. By choosing always the action which minimizes the cost, one gets the iteration policy, i.e. the policy resulting from the first policy iteration.

Though the Howard equations are just a set of linear equations for relative costs C_i and the average cost rate c of the standard policy (see below), their solution cannot be obtained because of the prohibitive size of the state space for any realistic system. However, at any decision epoch the relative costs C_i are needed only for the current state j and a small set of states $\mathcal{A}(j,k)$ reachable from the current state. We propose to estimate these values on the fly by means of simulations. To this end, it is useful to consider the physical interpretation of the relative costs C_i.

Given that the system starts from state i at time 0 and standard policy is applied for all decisions, the cumulative costs are accrued at the expected rate $c_t(i)$ at time t,

$$c_t(i) = \sum \lambda_k w_k \mathrm{P}\{I_t \in \mathcal{B}_k | I_0 = i\}, \qquad (1)$$

i.e. the expected rate of lost revenue, where $\mathrm{P}\{I_t \in \mathcal{B}_k\}$ is the probability that at time t the state of the system I_t is a blocking state for class-k calls. When $I_t \in \mathcal{B}_k$ class-k calls arriving at time t are blocked by the standard policy because either no feasible RW pair exists or the policy otherwise deems the blocking to be advantageous in the long run. The expected cost rate $c_t(i)$ depends on the initial state i. However, no matter what the initial state is, as t tends to infinity, the expected cost rate tends to a constant c, which is specific to the standard policy, and corresponds to (1) with steady state blocking probabilities $\mathrm{P}\{I_t \in \mathcal{B}_k\}$.

A typical behavior of the function $c_t(i)$ is depicted in Figure 1 for two different initial values i_1 and i_2. The relative costs C_i is defined as the integral

$$C_i = \int_0^\infty (c_t(i) - c)\, dt,$$

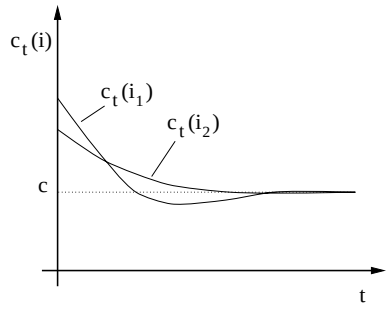

Figure 1. Expected costs with different initial choices as a function of time.

i.e. the area between the curve $c_t(i)$ and the line at level c. So we are interested in the transient behavior of $c_t(i)$; after the transient no contribution comes to integral. The length of the transient is of the order of $1/\mu$, where $1/\mu$ is the average holding time of a connection. After that the system essentially forgets the information about the initial state. So we can restrict ourselves to an appropriately chosen finite interval $(0, T)$. The actual choice of T is a tradeoff between different considerations as will be discussed later.

One easily sees that in the policy improvement step only the differences of the values C_i between different states are important. Therefore, we can neglect the c in the integral, as it is common to all states, and end up for thus redefined C_i,

$$C_i \approx C_i(T) = \int_0^T c_t(i)\, dt, \tag{2}$$

which is simply the expected cumulative cost in interval $(0, T)$ starting from initial state i.

4. Estimation of the state costs by simulation

In practice, it is not feasible to calculate the cost rate function $c_t(i)$ analytically even for the simplest policies. Therefore, we estimate the state costs C_i by simulations. In each simulation the system is initially set in state i and then the evolution of the system is followed for the period of length T, making all the RWA decisions according to the standard policy.

4.1. Statistics collection: blocking time vs. blocking events

In collecting the statistics one has two alternatives. Either one records the time intervals when the system is in a blocking state of class-k calls, for all $k \in \mathcal{K}$. If the cumulative time within interval $(0, T)$ when the system is in the blocking state of class-k calls is denoted by $\tau_k(i)$, then the integral is simply

$$\hat{C}_i = \sum \lambda_k w_k \tau_k(i). \tag{3}$$

Alternatively, one records the number $\nu_k(i)$ of blocked calls of type k in interval $(0, T)$. Then we have

$$\hat{C}_i = \sum w_k \nu_k(i). \tag{4}$$

In these equations we have written explicitly $\tau_k(i)$ and $\nu_k(i)$ in order to emphasize that the system starts from the state i. Both (3) and (4) give an unbiased estimate for C_i. In either case, the simulation has to be repeated a number of times in order to get an estimator with small enough confidence interval.

Denote the estimates of future costs obtained in the jth simulation run by $\hat{C}_i^{(j)}$, using (3) or (4) as the case may be. Then our final estimator for C_i is

$$\hat{C}_i = \frac{1}{N} \sum_{j=1}^{N} \hat{C}_i^{(j)}, \tag{5}$$

where N is the number of simulation runs. In fact, for the policy improvement the interesting quantity is the difference $E_{i_1,i_2} = C_{i_2} - C_{i_1}$, for which we have the obvious estimate

$$\hat{E}_{i_1,i_2} = \hat{C}_{i_2} - \hat{C}_{i_1}. \tag{6}$$

From the samples $\hat{C}_{i_1}^{(j)}$ and $\hat{C}_{i_2}^{(j)}$, $j = 1, \ldots, N$, we can also derive an estimate for the variance $\hat{\sigma}^2_{i_1,i_2}$ of the estimator $\hat{E}_{i_1,i_2}$

$$\hat{\sigma}^2_{i_1,i_2} = \frac{\hat{S}^2_{i_1,i_2} - (\hat{E}_{i_1,i_2})^2}{N-1},$$

where $\hat{S}^2_{i_1,i_2} = \frac{1}{N} \sum_j \left(\hat{C}_{i_2}^{(j)} - \hat{C}_{i_1}^{(j)} \right)^2$.

The choice between the alternative statistics collection methods is based on technical considerations. Though estimator (3) (blocking time) has a lower variance per one simulation run, it requires much more bookkeeping and the variance obtained with a given amount of computational effort may be lower for estimator (4) (blocking events).

4.2. Policy iteration with uncertain state costs

In order to deal with uncertainty of the estimators $\hat{C}_i$, we do not blindly accept the action with the smallest estimated cost, but give a special status for the decision which would be chosen by the standard policy. Let us give this policy the index 0. Based on the simulations we form estimates $\hat{E}_{0,i}$ for each possible action i. Then, as the decision we choose the action which minimizes the quantity

$$\hat{E}_{0,i} + k \cdot \hat{\sigma}_{0,i}, \tag{7}$$

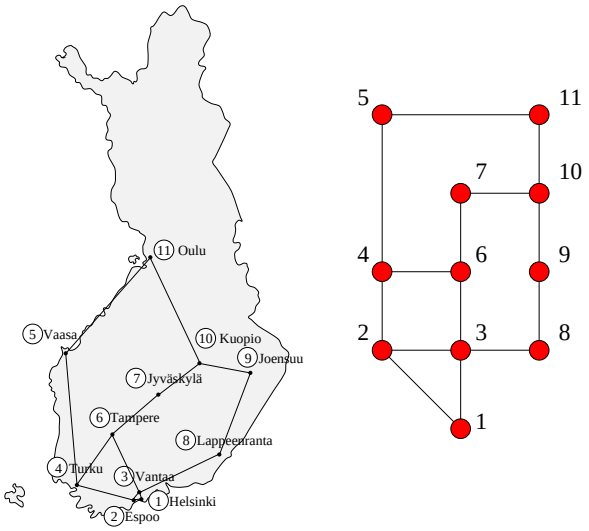

Figure 2. Hypothetical WDM-network in Finland.

where k is an adjustable parameter. Note that for $i = 0$ this quantity is equal to 0. Thus, in order for another action i to replace the action 0 of the standard policy, we must have $\hat{E}_{0,i} < -k \cdot \hat{\sigma}_{0,i}$, i.e. we require a certain minimum level of confidence for the hypothesis $C_i < C_0$. An appropriate value for k has to be determined experimentally.

The important parameters of the simulation now are the length of the simulation period T and the number of simulation runs N used for the estimation of each C_i. In practice, we are interested in the smallest possible values of T and N in order to minimize the simulation time. However, making T and N too small increases the simulation noise, i.e. the error in the estimates for C_i, occasionally leading to decisions that differ from that of the true iteration policy, consequently degrading the performance of the resulting algorithm.

5. Heuristic Algorithms

Several quick heuristic algorithms have been proposed in the literature. Here we briefly present some of them and study how iteration approach works with them. The first set of algorithms assumes that a fixed set of possible routes for each connection is given in advance. Some papers refer to this as alternate routing. In practice this set usually consists shortest or nearly shortest path of routes. Each algorithm accepts the first feasible RW pair found (first-fit).

- *basic* algorithm goes through all the routes in a fixed order and for each route tries all the wavelengths in a fixed order.
- *porder* algorithm is similar to *basic*-algorithm but it goes through all the wavelengths in a fixed order and for each wavelength tries all the routes in a fixed order.
- *pcolor* algorithm works like *porder* but wavelengths are gone through in order of the usage instead of a fixed order, so that the most used wavelength is tried first.
- *lpcolor* algorithm also tries to pack colors, but the primary target is to minimize the number of used links. So the algorithm first tries the most used wavelength with all the shortest routes, then the next often used wavelength and so on. If no wavelength works, the set of routes is expanded to include routes having one link more and wavelengths are tried again in the same order.
- *ll* or least loaded algorithm (see [6] is similar to *pcolor* but here the chosen RW pair is the one which leaves most capacity free after the assignment, i.e. the minimal number of free wavelengths over the links used is maximized.

The adaptive unconstrained routing (AUR) algorithms (see eg. [7]) search for a free route dynamically based on the current state of the network.

- *aurpack* is similar to *pcolor*, but without the limitations of a fixed set of routes.
- *aurexhaustive* finds a route with each wavelength (if possible) and chooses the shortest among them, i.e. it is identical to *lpcolor* except that the set of possible routes is not limited.

Also other heuristics are given in [7] like *random* (tries wavelengths in random order) or *spread* (tries least used wavelength first), but they were reported to work worse than the ones described above, and are not further discussed here.

6. Simulation results

Next we will present some numerical results from simulations. All tests were run for the small network shown in figure 2. The network was assumed to have 8 wavelengths available on each link. All the links contained one fibre. The offered load was uniform among all traffic classes (node pairs) and each rejected call represents an equal cost, $w_k = 1$ for all $k \in \mathcal{K}$. These assumptions simplify formulas (3) and (4). Note that the assumption $w_k = 1$ for all k means that the the objective is to minimize the long term blocking rate, i.e. the blocking probability.

It should also be noted that the results for the iteration policy were obtained by two levels of nested simulations. In order to assess the performance of the policy, an outer simulation is run, where connections arrive and leave the network and blocking times or events are recorded. Upon each arrival, a number of inner simulations are launched from the current state in order to make a comparison between different decision alternatives. Based on this comparison one alternative is chosen and used in the outer simulation, which

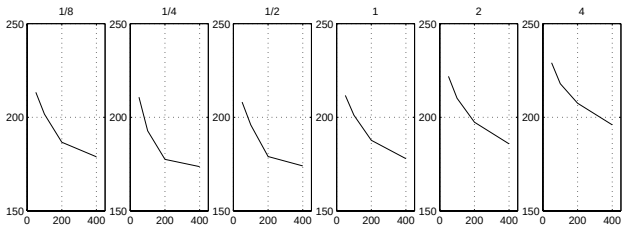

Figure 3. The effect of extending simulation period in the state cost estimation on the blocking probability of the iteration algorithm. The routing parameters are $\boldsymbol{\Delta l}= \mathbf{0}$ and *rmax*$= \mathbf{4}$, and *basic* is used as the standard policy. Load is $\boldsymbol{a} = \mathbf{0.4}$ for each traffic class.

then continues until upon the next arrival the decision analysis by the inner simulations is again started.

6.1. Routing algorithm parameters

All non-*aur* algorithms assume that a predefined set of possible routes per traffic class is given. This raises the question, which set of routes is optimal? Clearly too small a set of routes limits how well any algorithm can perform. But also it is not advantageous to use very long routes either. Here, the set of routes is specified with two parameters Δl and *rmax*. Parameter Δl defines how many links longer routes than the shortest one are included in set of routes. The second parameter *rmax* limits the total number of routes, i.e. only the *rmax* first routes are included in set. For example, with Δl=0 and *rmax*=10 only the shortest routes are included, and if there are more than 10 shortest routes for some node pair only the first 10 found are included.

6.2. Effect of simulation noise

After the transient period the cost rate $c_t(i)$ is very near to the long time average c of the standard policy. Simulating past it thus gives no new information, but actually only increases the noise resulting from the stochastic nature of the simulation. This can be seen from figure 3 where results clearly get worse as the simulation period grows (diagrams from left to right) while the number of simulation runs is kept the same.

The figure presents blocking probability obtained with the first policy iteration in the network with a moderate load of $a = 0.4$ for each traffic class. On the x-axis of each diagram is the number of simulation runs N, i.e. samples of future costs of a given initial decision. The results in the figure were obtained using estimator (3) in the estimation of the state costs. The conclusion is that the longer the simulation period is, the smaller is the 'signal to noise ratio' and the more simulation runs are needed to 'recover the signal'.

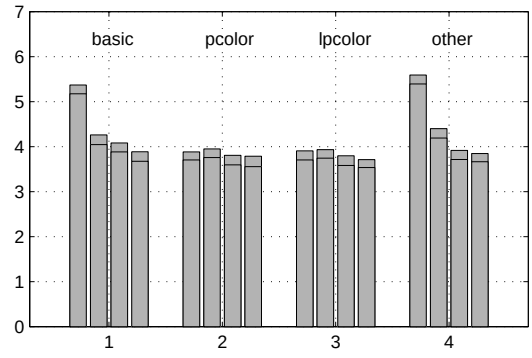

Figure 4. Results with quick heuristic algorithms and the first policy iteration. The routing parameters are: $\boldsymbol{\Delta l} = \mathbf{1}$ and *rmax* $= \mathbf{4}$. Each of the three groups of bars, *basic*, *pcolor*, *lpcolor*, contain results obtained with standard policy and iteration policy with $\boldsymbol{N} = \mathbf{50, 100}$ and $\mathbf{200}$. The last group of bars gives the results for *spread*, *porder*, *ll* and *aurpack*.

We cannot, however, make the simulation period T arbitrarily small, since if the whole transient period is not covered the signal becomes biased.

6.3. Iteration algorithm

The simulations were run for the same test network as was used before, i.e. the small network of figure 2. The network was assumed to have 8 wavelengths on each link and the offered load was uniform among all node pairs. Good running parameters for the inner simulations for this system were estimated from figure 3. Based on this we chose the simulation period $T = 0.25 \cdot 1/\mu$ and $N = 50 \ldots 200$ simulation runs for each alternative action.

Simulations were run with the quick heuristic algorithms as well as with the iteration algorithm with different parameters. The resulting blocking probabilities are shown in figure 4. The upper part of the bars (light gray) represent two times the standard deviation and the mean value is in the middle of upper part. The routing parameters here were Δl=1 and *rmax*=4 which clearly limit the set of routes. The first group of bars represents the blocking probability with *basic* algorithm and iteration policy with $N = 50, 100, 200$ using *basic* as the standard policy. The second group is the same but using *pcolor* instead of *basic*, and similarly in the third figure the *lpcolor* is used. The fourth group is obtained with quick heuristics *spread*, *porder*, *ll* and *aurpack*.

The improvement obtained by the first policy iteration starting with the *basic* algorithm was quite large, about 30%, while with *pcolor* the improvement is much less. Generally the results from iteration approach are always better than any of the heuristics which used same set of possible routes. The *aurpack* uses dynamic routing with routes of

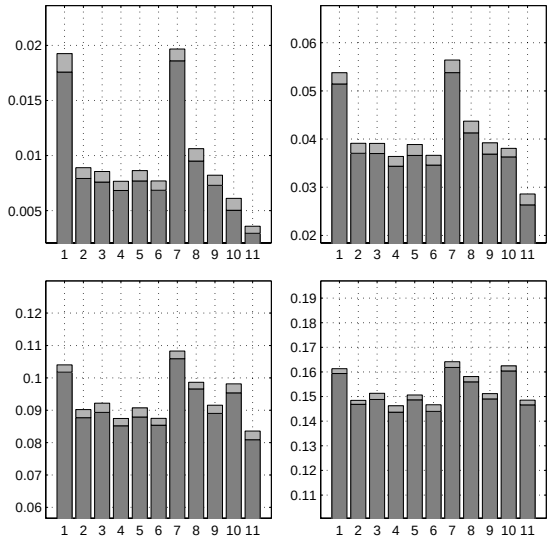

Figure 5. Blocking probability with loads ranging from $a = 0.3$ to $a = 0.6$. The set of routes were defined with Δl=1 and *rmax*=4. Algorithms are, from left to right, *basic*, *basic+iteration*, *pcolor*, *pcolor+iteration*, *lpcolor*, *lpcolor+iteration*, *spread*, *porder*, *ll*, *aurpack* and *aurexhaustive*.

any length in the search space and is here only for comparison.

In another set of simulations the iteration approach was applied with different standard policies to get some idea about how important the underlying algorithm is. In the four diagrams of figure 5 the results can be seen with four different offered loads, with the blocking probability varying from quite a low to a high value. The algorithms used were (in order) *basic*, *basic+iteration*, *pcolor*, *pcolor+iteration*, *lpcolor*, *lpcolor+iteration*, *spread*, *porder*, *ll*, *aurpack* and *aurexhaustive*. In these simulations the routing parameters were also the same Δl=1 and *rmax*=4. The number of simulations runs for each alternative action N was chosen to be 200. So *aurpack* and *aurexhaustive* have again much larger set of possible routes to choose from. It can be seen from the figure that in each case the iteration algorithm indeed gives slightly better results.

7. Conclusions

In this paper we have introduced the idea of applying the first iteration in the policy space to the RWA problem. With this method one can derive from any given heuristic policy an iteration policy, which theoretically always is a better policy, i.e. has lower average cost rate (e.g. blocking probability). The relative costs of states needed in the decision analysis are estimated on the fly by launching simulations from the current state of the system and trying different decision alternatives. The simulations introduce some noise in the cost estimates and careful control of the simulation parameters is required in order not to deteriorate the performance of the resulting iteration policy.

The performance improvement obtained by the policy iteration depends on the standard policy one starts with. The reduction of blocking probability in the numerical tests ranged from a few tens of percents to almost nothing. This suggests that algorithms like *lpcolor*, for which the improvement was small, is not far from optimal (with this particular network and setup). Also we can generally conclude that the order in which RW pairs are tried in a first-fit algorithm can be a very important factor for the performance. As the method of first policy iteration obviously is computationally intensive, it can be used in real time only if the dynamics of the system is slow. The inter-arrival times of the connection requests must be of the order of few seconds or more (depending on the number of possible decisions), but this may well be the case in WDM networks. However, the method can always be used off-line e.g. to evaluate heuristic algorithms to see how far they are from the optimum.

References

[1] B. Mukherjee, *Optical Communication Networks*. McGraw-Hill series on computer communications, McGraw-Hill, 1997.

[2] R. Ramaswami and K. Sivarajan, *Optical Networks, A Practical Perspective*. Morgan Kaufmann Series in Networking, Morgan Kaufmann Publishers, 1998.

[3] A. Willner, "Mining the optical bandwidth for a terabit per second," *IEEE Spectrum*, pp. 32–41, Apr. 1997.

[4] D. Banaree and B. Mukherjee, "A practical approach for routing and wavelength assignment in large wavelength-routed optical networks," *IEEE J. Select. Areas Commun.*, vol. 14, pp. 903–908, June 1996.

[5] S. Baroni, *Routing and wavelength allocation in WDM optical networks*. PhD thesis, University College London, May 1998.

[6] E. Karasan and E. Ayanoglu, "Effects of wavelength routing and selection algorithms on wavelength conversion gain in wdm optical networks," *IEEE/ACM Trans. Networking*, vol. 6, pp. 186–196, Apr. 1998.

[7] A. Mokhtar and M. Azizoğlu, "Adaptive wavelength routing in all-optical networks," *IEEE/ACM Trans. Networking*, vol. 6, pp. 197–206, Apr. 1998.

[8] R. Ramaswami and K. Sivarajan, "Routing and wavelength assignment in all-optical networks," *IEEE/ACM Trans. Networking*, vol. 3, pp. 489–500, Oct. 1995.

[9] H. C. Tijms, *Stochastic Models, An Algorithmic Approach*. John Wiley & Sons Ltd, 1994.

[10] Z. Dziong, *ATM Network resource management*. McGraw-Hill, 1997.