Call Admission Control and Routing in Integrated Services Networks Using Neuro-Dynamic Programming*

Peter Marbach† Oliver Mihatsch‡ John N. Tsitsiklis§

February 11, 1999

Abstract

We consider the problem of call admission control and routing in an integrated services network that handles several classes of calls of different value and with different resource requirements. The problem of maximizing the average value of admitted calls per unit time (or of revenue maximization) is naturally formulated as a dynamic programming problem, but is too complex to allow for an exact solution. We use methods of neuro-dynamic programming (reinforcement learning), together with a decomposition approach, to construct dynamic (state-dependent) call admission control and routing policies. These policies are based on state-dependent link costs, and a simulation-based learning method is employed to tune the parameters that define these link costs. A broad set of experiments shows the robustness of our policy and compares its performance with a commonly used heuristic.

1 Introduction

We consider a communication network consisting of a set of nodes $\mathcal{N} = \{1, \ldots, N\}$ and a set of unidirectional links $\mathcal{L} = \{1, \ldots, L\}$, where each link l has a a total

*This research was supported by Siemens AG, Germany, Alcatel Bell, Belgium, and by the NSF under contract ECS-9873451. A preliminary version of this paper was presented at the 37th IEEE Conference on Decision and Control, Tampa, Florida, December 1998.

†Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; current affiliation: Center for Communication Systems Research, Cambridge University, UK; email: p.marbach@ccsr.cam.ac.uk

‡Siemens AG, Corporate Technology, Information and Communications 4, D-81730 Munich, Germany; email: oliver.mihatsch@mchp.siemens.de

§Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; email: jnt@mit.edu

capacity of $B(l)$ units of bandwidth. There is a set $\mathcal{M} = \{1, ..., M\}$ of different service classes, where each class m is characterized by its bandwidth requirement $b(m)$, its average call holding time $1/\nu(m)$, and the immediate reward (or value) $c(m)$ obtained whenever such a call is accepted. The bandwidth requirement $b(m)$ may reflect either the peak transmission rate requested by class m calls, or their "effective bandwidth" as defined and extensively studied in the context of ATM networks [WV96]. Furthermore, the reward $c(m)$ is not necessary a monetary one, but may reflect the importance of different classes and their desired quality of service (blocking probabilities). We assume that we are given the arrival rates $\lambda_{ij}(m)$ of class m calls with origin $i \in \mathcal{N}$ and destination $j \in \mathcal{N}$. We also assume that the arrivals are independent Poisson processes, and that the holding times of the calls are independent, exponentially distributed, and independent of the arrival processes.

When a new call of class m, with origin i and destination j arrives, it can be either rejected (with zero reward) or it can be admitted (with reward $c(m)$). In order to accept it, we need to choose a route out of a predefined list of possible routes from i to j. Furthermore, at the time that the call is accepted, each link along the chosen route must have at least $b(m)$ units of unoccupied bandwidth. The objective is to exercise call admission control and routing in such a way that the long term average reward is maximized. Ideally, this maximization should take place within the most general class of state-dependent policies, whereby the admission decision and the route choice are allowed to depend on the current state of the network.

The above defined call admission control and routing problem has been studied extensively; see e.g., [Kel91, Ros95] and the references therein. It is naturally formulated as an average reward dynamic programming problem, but is too complex to be solved exactly and suitable approximations have to be employed to compute control policies. One proposed approach in this context is the reduced load approximation (also called Erlang fixed-point method) [Kel91, CR93]. It relies on link independence and Poisson assumptions which allow to decompose the network into link processes where calls arrive according to independent Poisson processes. The corresponding arrival rates model the thinned (by blocking on other links) external traffic and are computed by iteratively solving a system of fixed-point equations. This approach has been used to analyze routing schemes such as probabilistic routing (also called proportional routing) [Kel88, MMR96, CR93] and dynamic alternative routing with trunk reservation [Key90, Laws95, GS97]. As its name suggests, state-independent probabilistic routing assigns routes to calls at random according to a given probability distribution. Using the concept of a state-independent link cost (link shadow price), gradient methods for tuning the routing probabilities can be devised [Kel88, MMR96, CR93]. Probabilistic routing can be shown to be asymptotically optimal, however in a "coarse sense": optimal routing schemes are sensitive to the model parameters, i. e. small modeling errors can severely degrade performance [Whi88]. More robust, but also more difficult to analyze and optimize, is the state-dependent dynamic alternative routing with trunk reservation. In the case of a single service class, a decomposition approach, that splits the reward associated with a call into link rewards, can be employed to compute state-dependent link costs (shadow prices) and to tune the trunk reservation parameters [Key90]. However, in the case of multiple service classes, a judicious choice of the trunk reservation parameters, that lead to near-optimal performance, can be difficult. An application of this approach is described in [Liu97, LBM98], for a relatively small problem, but can easily become intractable for larger networks. In [DM94] a variant of this approach was proposed which uses measurements in the network to determine the arrival rates associated with each link, thus avoiding the computational burden of solving fixed-point equations. Similar to [Key90], a decomposition approach of the call rewards can be employed to compute state-dependent link costs and to optimize the policy. This method can again become intractable unless further approximations such as link-state aggregations are employed. An application of this approach is given in [DM94].

The link independence and Poisson assumptions play an important role in the above described methods and allow to construct a simpler model of the network process and to compute implied link costs (shadow prices). These costs are then used to obtain an approximation of the true implied network costs (derived from the differential reward function of dynamic programming), and to optimize and implement a call admission control and routing policy. In this paper we develop a new approach which allows us to avoid the use of a reduced model, i. e. explicitly decomposing the network process into independent link processes. We start with a dynamic programming formulation (Section 2) and then use simulation-based approximate dynamic programming (also called reinforcement learning (RL) or neuro-dynamic programming (NDP)) [BT96, SB98] to construct an approximate differential reward function and to optimize the policy (Section 3). In the following, we will use the term NDP for simulation-based approximate dynamic programming. For these methods, performance guarantees exist only for special cases (see [BT96]), however recent case studies illustrate their ability to successfully address large-scale problems. In particularly, they have been applied to resources allocation problems in telecommunication systems such as the channel assignment problem in cellular telephone systems [SiB97], the link adaptation problem [NC95] and the single link admission control problem with self-similar traffic [CN97] or with statistical quality of service constraints [BTS99]. A successful application of NDP relies crucially on the choice of a suitable (parametric) architecture for the approximation of the differential reward function: it should be rich enough (i. e. involve enough parameters) to approximate closely the differential reward function, but also simple (i. e. involve not too many parameters) to limit the "training time" to obtain a good approximation. Typically, an approximation architecture is chosen by a combination of analysis, engineering insight, and trial and error. Motivated by the analysis carried out in connection with the reduced load approach and its variants, we rely on a function which depends quadratically on the number of active calls of each class on each link, and which leads to policies that rely on "trained" state-dependent link costs. Furthermore, we decompose the call reward into link rewards to allow a decentralized implementation of the optimization method and the resulting policies. We apply this approach to a large network, involving 62 links, and with 992 tunable parameters in our differential reward function approximator. To assess the method, we compare our call admission control and routing policies with the "Open-Shortest-Path-First" (OSPF) heuristic (Section 4.2–4.3). We show that the performance of our NDP policy is very robust with respect to changing arrival statistics. To investigate the accuracy of the quadratic approximator, we also provide a case study involving a single link (Section 4.1).

The main contributions of the paper are the following.

(a) We show that NDP can be applied to the call admission control problem in a manner that supports decentralized training and decentralized decision making. By using NDP, we are able to

(b) avoid the use of a reduced model, as it was introduced in previous approaches through the link independence and Poisson assumption, as well as to

(c) avoid the computational burden associated with the evaluation of the link reward functions, as it was encountered in [DM94, Key90].

2 Dynamic Programming Formulation

We will now formulate the problem of call admission control and routing as a continuous-time, average reward, finite-state dynamic programming problem [Ber95]. For any time t, let $n_t(r,m)$ be the number of class m calls that are currently active (have been admitted and have not yet terminated) and which have been routed along route r. The state x_t of the network at time t consists of a list of the numbers $n_t(r,m)$, for each r and m. The state space S (the set of all possible states) is defined implicitly by the requirements that each $n_t(r,m)$ be a nonnegative integer and that

$$\sum_{r \in R(l)} \sum_{m \in \mathcal{M}} n_t(r,m) b(m) \leq B(l), \quad \forall l \in \mathcal{L},$$

where $R(l)$ is the set of routes that use link l. Even though the process evolves in continuous time, we only need to consider the state of the network at the times when certain events take place. The events of interest are the arrivals of new call requests and the terminations of existing calls. Note that the nature of an event is completely specified by the class m, origin-destination pair (i, j), and if it corresponds to a call termination, the route r occupied by the call. We denote by Ω the (finite) set of all possible events.

If the state of the system is x and event ω occurs, a decision u has to be made. If ω corresponds to an arrival, the set of possible decisions $U(\omega, x)$ consists of the possible routes (subject to the capacity constraints and the current state of the network) and of the rejection decision. If ω corresponds to a departure, there are no decisions to be made, which amounts to letting $U(x, \omega)$ be a singleton. Given the present state of the network x, an event ω, and a decision $u \in U(x, \omega)$, the network moves to a new state which will be denoted by x'. The resulting reward will be denoted by $g(x, \omega, u)$: if ω corresponds to a class m arrival and u is a decision to admit along some route, then $g(x, \omega, u) = c(m)$; otherwise, $g(x, \omega, u) = 0$.

We define a *policy* to be a mapping μ whose domain is the set $S \times \Omega$ and which satisfies

$$\mu(x, \omega) \in U(x, \omega), \quad \forall\, x \in S, \omega \in \Omega.$$

We note that under any given policy μ, the state x_t evolves as a continuous-time finite-state Markov process. Let t_k be the time of the kth event, and let x_{t_k} be the state of the system just prior to that event. (This notation is equivalent to assuming that x_t is a left-continuous function of time.) We then define the average reward associated with a policy μ to be

$$v(\mu) = \lim_{N \to \infty} \frac{1}{t_N} \sum_{k=0}^{N-1} g(x_{t_k}, \omega_k, u_{t_k}) \tag{1}$$

where $u_{t_k} = \mu(x_{t_k}, \omega_{t_k})$. Note that the state corresponding to an empty system, to be denoted by $\hat{x}$, is recurrent. For this reason, the limit in Eq. (1) exists, is independent of the initial state, and is equal to a deterministic constant with probability 1.

A policy μ^* is said to be optimal if

$$v(\mu^*) \geq v(\mu)$$

for every other policy μ. We denote the average reward associated with an optimal policy μ^* as v^*.

An optimal policy can be obtained, in principle, by solving the Bellman optimality equation for average reward problems, which takes the form

$$\begin{aligned} v^* E_\tau\{\tau \mid x\} + h^*(x) &= E_\omega \left\{ \max_{u \in U(x,\omega)} [g(x, \omega, u) + h^*(x')] \right\}, \quad x \in S, && (2) \\ h^*(\hat{x}) &= 0. && (3) \end{aligned}$$

Here, τ stands for the time until the next event occurs and $E_\tau\{\tau \mid x\}$ is the expectation of τ given that the current state is x. Furthermore, $E_\omega\{\cdot\}$ stands for the expectation with respect to the next event ω, and x' stands for the state right after the event, which is a deterministic function of x, ω, and the chosen decision u. If $|S|$ is the cardinality of the state space, the Bellman equation is a system of $|S|+1$ nonlinear equations in the $|S|+1$ unknowns $h^*(x)$, $x \in S$, and v^*. Because the state $\hat{x}$ is recurrent under every policy, the Bellman equation has a unique solution and the function $h^*(\cdot)$, called the optimal differential reward, admits the following interpretation. If we operate the system under an optimal policy, then $h^*(x) - h^*(y)$ is equal to the expectation of the difference of the total rewards (over the infinite horizon) for a system initialized at x, compared with a system initialized at y.

Once the optimal differential reward function $h^*(\cdot)$ is available, an optimal admission control and routing policy μ^* is given by

$$\mu^*(x, \omega) = \arg \max_{u \in U(x,\omega)} \left[g(x, \omega, u) + h^*(x')\right]. \tag{4}$$

This amounts to the following. Whenever a new class m call requests a connection, consider admitting it along a permissible route and let x' be the resulting successor state. We compute the value of such a decision by adding the immediate reward $g(x, \omega, u) = c(m)$ to the merit $h^*(x')$ of x'. We pick a route that results in the highest value and route the call accordingly if that value is higher than the value $h^*(x)$ of the current state; otherwise, the call is rejected.

However, the dynamic programming approach is impractical because the state space S is typically so large that it is impossible to compute, or even store, the optimal differential reward $h^*(x)$ for each state $x \in S$. This leads us to consider methods that work with approximations to the function h^*.

3 Neuro-Dynamic Programming Solution

Neuro-dynamic programming (NDP) is a simulation-based approximate dynamic programming methodology for producing near-optimal solutions to large scale dynamic programming problems. The central idea is to approximate v^* and the function $h^*(\cdot)$ by a tunable scalar $\tilde{v}$ and an approximating function $\tilde{h}(\cdot, \theta)$, respectively, where θ is a tunable parameter vector. The structure of the function $\tilde{h}$ is chosen so that for any given x and θ, $\tilde{h}(x, \theta)$ is easy to compute. Once the general form of the function $\tilde{h}(\cdot, \cdot)$ is fixed, the next step is to set θ and $\tilde{v}$ so that the resulting function $\tilde{h}(\cdot, \theta)$ provides an approximate solution to Bellman's equation. Any particular choice of θ leads immediately to a policy μ^θ, given by

$$\mu^\theta(x, \omega) = \arg \max_{u \in U(x,\omega)} \left[g(x, \omega, u) + \tilde{h}(x', \theta)\right]. \tag{5}$$

This is similar to Eq. (4), which defines an optimal policy, except that the approximation $\tilde{h}(x', \theta)$ is used instead of $h(x')$.

There are two main ingredients in this methodology, to be discussed separately in the subsections that follow:

(a) Defining an "approximation architecture", that is, the general form of the function $\tilde{h}(\cdot, \cdot)$.

(b) Developing a method, usually simulation-based, for tuning θ and $\tilde{v}$.

3.1 Approximation Architecture

In defining suitable approximation architectures, one usually starts with a process of *feature extraction*. This involves a feature vector $f(x)$, which is meant to capture those "features" of the state x that are considered most relevant to the decision making process. Usually, the feature vector is handcrafted based on available insights on the nature of the problem, prior experience with similar problems, or experimentation with simple versions of the problem. Our choice of a feature vector will be described shortly.

Given the choice of the feature vector, a commonly used approximation architecture is of the form $\tilde{h}(f(x), \theta)$, where $\tilde{h}$ is a multilayer perceptron with input $f(x)$ and internal tunable weights θ (see for example [Hay94]). This architecture is powerful because it can approximate arbitrary functions of $f(x)$. The drawback, however, is that the dependence on θ is nonlinear, and tuning θ can be time consuming and unreliable.

An alternative is provided by a *linear* feature-based approximation architecture, in which we set

$$\tilde{h}(x, \theta) = \theta^T f(x).$$

Here, the superscript T stands for transpose, and the dimension of the parameter vector θ is set to be equal to the number of features – the dimension of the feature vector $f(x)$. Because of the linear dependence on θ, the problem of tuning θ resembles the linear regression problem, and is generally much more reliable.

Let $n_{l,m}$ be the number of class m calls that are active and which have been assigned to routes that go through link l. We view the variables $n_{l,m}$ and the products of the form $n_{l,m} n_{l,m'}$ as features, and we will work with a linear approximation architecture of the form

$$\tilde{h}(x, \theta) = \sum_{l \in \mathcal{L}} \left[\theta(l) + \sum_{m} \theta(l, m) n_{l,m} + \sum_{(m,m'):m' \leq m} \theta(l, m, m') n_{l,m} n_{l,m'} \right]. \tag{6}$$

A main reason for choosing a quadratic function of the variables $n_{l,m}$ is that it led to essentially optimal solutions to single link problems (see Section 4.1). Note that we have only included those products $n_{l,m}n_{l',m'}$ associated with a common link ($l = l'$). There are two reasons behind this choice: it opens up the possibility of a decomposable training algorithm (cf. Section 3.3). In addition, it results in policies with an appealing decentralized structure,which we now discuss.

Let $n_{l,m}$ be the variables associated with the current state x of the network and suppose that ω corresponds to an arrival of class m^*. Let us focus on a particular decision $u \in U(x, \omega)$ which assigns this call to route r, resulting in a new state x' and variables $n'_{l,m}$. Note that $n'_{l,m} = n_{l,m} + 1$ if $l \in r, m = m^*$, and $n'_{l,m} = n_{l,m}$, otherwise. With some straightforward algebra, the merit $Q(x, \omega, r, \theta)$ of this decision, in comparison to rejection, is given by

$$\begin{aligned} Q(x, \omega, r, \theta) &= g(x, \omega, u) + \tilde{h}(x', \theta) - \tilde{h}(x, \theta) \\ &= c(m^*) + \sum_{l \in r} \Bigg[\theta(l, m^*) + \theta(l, m^*, m^*)(2n_{l,m^*} + 1) \\ &\quad + \sum_{m < m^*} \theta(l, m^*, m) n_{l,m} + \sum_{m > m^*} \theta(l, m, m^*) n_{l,m} \Bigg]. \end{aligned}$$

The corresponding policy $\mu^\theta(\cdot, \cdot)$ [cf. Eq. (5)] amounts to choosing a route r^* for which $Q(x, \omega, r, \theta)$ is largest, using this route if $Q(x, \omega, r^*, \theta) > 0$, and rejecting the call if $Q(x, \omega, r^*, \theta) \leq 0$. This is equivalent to assigning a link cost (or shadow price)

$$\theta(l, m^*) + \theta(l, m^*, m^*)(2n_{l,m^*} + 1) + \sum_{m < m^*} \theta(l, m^*, m) n_{l,m} + \sum_{m > m^*} \theta(l, m, m^*) n_{l,m} \quad (7)$$

to each link, and using these link costs for admission control and shortest path routing. Note that these link costs (shadow prices) Eq. (7) are state-dependent and reflect the instantaneous congestion on each link which is in the spirit of [DM94, Key90]. However, the notion of a link cost results here from a specific choice of a approximation architecture, and not from an explicit decomposition of the networks process into independent link processes as in [DM94, Key90].

The family of policies μ^θ resulting from our approximation architecture can provide a fair amount of flexibility. It remains to assess:

(a) whether there are systematic methods for finding good policies within this family; this is the subject of the next subsection; and,

(b) whether they lead to significant performance improvement in comparison to more restricted families of policies; this is to be assessed experimentally in Section 4.

3.2 The Training Algorithm

There are several methods that can be used to tune the parameter θ, most of which rely on simulation runs (or on on-line observations of an actual system). We will use a variant of one of the most popular methods, namely, Sutton's TD(0) ("temporal differences") algorithm [Sut88]. The standard TD(0) algorithm has been designed for discrete-time problems with a discounted criterion (or for an undiscounted total reward criterion in systems that eventually terminate), where the goal is to maximize the so-called discounted reward-to-go of state x, given by

$$E\left[\sum_{k=0}^{\infty} e^{-\beta t_k} g(x_{t_k}, \omega_k, u_{t_k}) \mid x_0 = x\right],$$

simultaneously for every state of the system. Here, $\beta > 0$ is a discount factor. So, some modifications are necessary to apply TD(0) to our problem. The first one, going from discrete to continuous time is fairly straightforward. The second one, going from a discounted to an average reward criterion, is much more substantial, since average reward dynamic programming theory and algorithms are generally more complex. We will use the recently developed temporal difference method for average reward problems [TV97b], which preserves the same convergence properties and error guarantees of its discounted counterpart [TV97a]. It should be noted that this is the first time that this method is applied to an engineering problem.

In the simplest version of TD(0), the controlled Markov process x_t is simulated under a fixed policy μ. Let t_k be the time of the kth event ω_k, which finds the system at state x_{t_k}, and let $u_{t_k} = \mu(x_{t_k}, \omega_k)$ be the resulting decision. At such an event time, the vector θ and the scalar $\tilde{v}$ are updated according to

$$\begin{aligned} \theta_k &= \theta_{k-1} + \gamma_k d_k \nabla_\theta \tilde{h}(x_{t_{k-1}}, \theta_{k-1}), && (8) \\ \tilde{v}_k &= \tilde{v}_{k-1} + \eta_k \big(g(x_{t_{k-1}}, \omega_{k-1}, u_{t_{k-1}}) - (t_k - t_{k-1})\tilde{v}_{k-1}\big), && (9) \end{aligned}$$

where the "temporal difference" d_k is defined by

$$d_k = g(x_{t_{k-1}}, \omega_{k-1}, u_{t_{k-1}}) - (t_k - t_{k-1})\tilde{v}_{k-1} + \tilde{h}(x_{t_k}, \theta_{k-1}) - \tilde{h}(x_{t_{k-1}}, \theta_{k-1}),$$

and where γ_k and η_k are small step size parameters. The only difference from discrete-time average reward TD(0) is in the factor of $t_k - t_{k-1}$ that multiplies $\tilde{v}_{k-1}$ and which, in turn, is due to the factor $E_\tau\{\tau \mid x\}$ in Bellman's equation. Note that with our linear approximation architecture $\tilde{h}(x, \theta) = \theta^T f(x)$, we have $\nabla_\theta \tilde{h}(x, \theta) = f(x)$.

Under a fixed policy, and under the standard diminishing step size conditions, $\tilde{v}_k$ converges to the average reward $v(\mu)$, and θ_k converges to a limiting vector θ such that $\tilde{h}(\cdot, \theta)$ provides a "good" approximation of $h^\mu(\cdot)$. Here, $h^\mu(\cdot)$ is a function defined similar to $h^*(\cdot)$, but in a context in which there is a single possible decision at each state, the one prescribed by the policy μ. Furthermore, the approximation is "good" in the sense that the approximation error $\tilde{h}(\cdot,\theta) - h^\mu(\cdot)$, measured under a suitable norm, is of the same order of magnitude as the best possible approximation error under the given approximation architecture [TV97b].

One can start with a policy μ, run TD(0) until it converges, use the resulting limiting value of θ to define a new policy according to Eq. (5), and then repeat. This method has some (weak) theoretical guarantees [BT96], but it is common practice to keep changing the underlying policy with each update of the parameter vector θ_k. This *optimistic* TD(0) method is completely described by the update rule (8) together with

$$u_{t_k} = \arg \max_{u \in U(x_{t_k}, \omega_k)} \left[g(x_{t_k}, \omega_k, u) + \tilde{h}(x', \theta_k) \right], \tag{10}$$

where x' is the successor state that results from x_{t_k}, ω_k, and u.

Even though optimistic TD(0) has no convergence guarantees, its discounted variant has been found to perform well in a variety of contexts [SiB97, Tes88, ZD96].

3.3 A Decomposition Approach

The algorithm described in the preceding subsection can be very slow to converge, especially for networks with a substantial number of links. This led us to consider a decomposition approach that breaks the reward associated with a call into link rewards in the spirit of [DM94, Key90], and which led to much shorter training times. This improvement in terms of training time is essential for applying NDP to large networks (see Section 4.3).

For any link l, consider the local "state" $x^{(l)} = (n_{l,m} : \forall m)$. Of course, this is not a state in the true sense of the word because, in general, it does not evolve as a Markov process, but will be treated to some extent as if it were. We decompose the immediate reward $g(x_{t_k}, \omega_k, u_k)$ associated with the kth event, into a sum of rewards attributed to each link:

$$g(x_{t_k}, \omega_k, u_{t_k}) = \sum_{l \in \mathcal{L}} g^{(l)}(x_{t_k}, \omega_k, u_{t_k}).$$

In particular, whenever a new call (say, of class m) is routed over a route r that contains the link l, the immediate reward $g^{(l)}$ associated with link l is set to $c(m)/\#r$, where $\#r$ is the number of links along route r. For all other events, the immediate reward associated with link l is equal to 0.

Let us fix a policy μ, let $v^{(l)}(\mu)$ be the average reward attributed to link l, and note that

$$v(\mu) = \sum_{l \in \mathcal{L}} v^{(l)}(\mu).$$

For each link, we introduce a scalar $\tilde{v}^{(l)}$, which is meant to be an estimate of $v^{(l)}(\mu)$, as well an approximation architecture $\tilde{h}^{(l)}(x, \theta^{(l)})$ of the form

$$\tilde{h}^{(l)}(x^{(l)}, \theta^{(l)}) = \theta(l) + \sum_m \theta(l, m) n_{l,m} + \sum_{(m,m'):m' \leq m} \theta(l, m, m') n_{l,m} n_{l,m'},$$

where $\theta^{(l)}$ is the vector of parameters $\theta(l)$, $\theta(l, m)$, and $\theta(l, m, m')$, associated with link l. Note that

$$\tilde{h}(x, \theta) = \sum_{l \in \mathcal{L}} \tilde{h}^{(l)}(x^{(l)}, \theta^{(l)}),$$

and we are therefore dealing with the same approximation architecture as in Section 3.1. The key difference is that we will not update θ according to Eq. (8), but will use an update rule which is local to each link. The local TD(0) algorithm, for link l is given by

$$\begin{aligned} \theta_k^{(l)} &= \theta_{k-1}^{(l)} + \gamma_k^{(l)} d_k^{(l)} \nabla_{\theta^{(l)}} \tilde{h}^{(l)} \left(x_{t_{k-1}^{(l)}}^{(l)}, \theta_{k-1}^{(l)} \right), \\ \tilde{v}_k^{(l)} &= \tilde{v}_{k-1}^{(l)} + \eta_k^{(l)} \left(g^{(l)} \left(x_{t_{k-1}^{(l)}}^{(l)}, \omega_{k-1}^{(l)}, u_{t_{k-1}^{(l)}} \right) - (t_k^{(l)} - t_{k-1}^{(l)}) \tilde{v}_{k-1}^{(l)} \right), \end{aligned}$$

where

$$\begin{aligned} d_k^{(l)} &= g^{(l)} \left(x_{t_{k-1}^{(l)}}^{(l)}, \omega_{k-1}^{(l)}, u_{t_{k-1}^{(l)}} \right) - (t_k^{(l)} - t_{k-1}^{(l)}) \tilde{v}_{k-1}^{(l)} \\ &\quad + \tilde{h}^{(l)} \left(x_{t_k^{(l)}}^{(l)}, \theta_{k-1}^{(l)} \right) - \tilde{h}^{(l)} \left(x_{t_{k-1}^{(l)}}^{(l)}, \theta_{k-1}^{(l)} \right), \end{aligned} \tag{11}$$

$\gamma_k^{(l)}$ and $\eta_k^{(l)}$ are small step size parameters, and $t_k^{(l)}$ is the time of the kth event $\omega_k^{(l)}$ associated with link l. Here, we say that an event is associated with link l if it can potentially result in a change of $x^{(l)}$; this is the case if we have a departure of a call that was using link l, or if link l is part of a route in the predefined list of possible routes connecting the current origin-destination pair. This update rule is identical to the ordinary TD(0) update under the assumption that $x_t^{(l)}$ is a Markov process that receives rewards $g^{(l)}(x_{t_k^{(l)}}^{(l)}, \omega_k^{(l)}, u_{t_k^{(l)}})$ at the times $t_k^{(l)}$ of events associated with link l. Of course, $x_t^{(l)}$ is not Markov because its transitions are affected by the global state x_t. Although the update rules for different links are decoupled, they are to be carried out in the course of a single simulation of the entire system, which accurately reflects all dependencies involved. This is to be compared with [DM94, Key90], were the entire system was explicitly decomposed into independent link processes making $x_t^{(l)}$ truly Markov, however at the expense of ignoring certain dependencies and introducing an additional modeling error.

Table 1: Case study for 3 service classes and a link with a capacity of 12 units.

Service Class m	1	2	3
Bandwidth Demand $b(m)$	1.00	2.00	2.00
Average Holding Time $1/\nu(m)$	2.00	1.25	1.11
Arrival Rate $\lambda(m)$	3.00	2.00	2.50
Immediate Reward $c(m)$	4.00	15.00	12.00

	Performance	
	Average Reward	Lost Average Reward
Always Accept	40.09	31.86
Trunk Reservation	47.23	24.72
Dynamic Programming	47.23	24.72
TD(0): MLP	47.19	24.76
TD(0): Quadratic	47.19	24.76

4 Experimental Results

In this section, we report the results obtained in a broad set of experiments. We compare the policy obtained through NDP with the commonly used heuristic OSPF (Open Shortest Path First). For every pair of source and destination nodes, OSPF orders the list of predefined routes. When a new call arrives, it is routed along the first route in the corresponding list that does not violate the capacity constraint; if no such route exists, the call is rejected. For a single link problem, OSPF reduces to the naive policy that always accepts an incoming call, as long as the required bandwidth is available.

4.1 Single Link Problems

Our first set of experiments involved multiple classes but a single link. They were carried out in order to identify potential difficulties with this approach, and to validate the promise of the quadratic approximation architecture. Naturally, with a single link, no decomposition had to be introduced. Two case studies were carried out involving 3 and 10 service classes, respectively. For the latter case, three different scenarios were considered corresponding to a highly, medium and lightly loaded link, respectively. A more detailed account of these experiments and the results obtained can be found in [MT97].

The experiments were carried out using TD(0) for discounted problems. The

Table 2: Problem data of the case study for 10 service classes on a link with a capacity of 600 units.

Service Class m	1	2	3	4	5
Bandwidth Demand $b(m)$	2.00	2.00	4.00	4.00	6.00
Average Holding Time $1/\nu(m)$	1.00	1.00	1.25	1.25	1.67
Immediate Reward $c(m)$	2.00	1.40	5.00	2.50	10.00
Arrival Rate $\lambda(m)$ (high load)	15.00	15.00	12.00	12.00	10.00
Arrival Rate $\lambda(m)$ (medium load)	15.00	15.00	10.00	10.00	7.00
Arrival Rate $\lambda(m)$ (light load)	16.00	16.00	12.00	12.00	7.00

Service Class m	6	7	8	9	10
Bandwidth Demand $b(m)$	6.00	8.00	8.00	10.00	10.00
Average Holding Time $1/\nu(m)$	1.67	2.50	2.50	5.00	5.00
Immediate Reward $c(m)$	4.00	20.00	7.00	5.00	16.00
Arrival Rate $\lambda(m)$ (high load)	10.00	6.00	6.00	4.00	4.00
Arrival Rate $\lambda(m)$ (medium load)	7.00	3.00	3.00	1.80	1.80
Arrival Rate $\lambda(m)$ (light load)	7.00	2.40	2.40	1.10	1.10

performance of the resulting policies was evaluated on the basis of the average reward criterion. The discount factor was chosen to be very small, which makes the discounted problem essentially equivalent to an average reward problem. The evaluation of the average reward is based on a trajectory of $4{\cdot}10^6$ time steps.

Besides TD(0) with a quadratic approximation architecture, we also used TD(0) with a multilayer perceptron (MLP) [Hay94]. Furthermore, for the smaller problem, which only involved three classes, we also obtained an optimal policy through exact dynamic programming (DP), and used it as a basis of comparison. A comparison was also made with a naive policy that always accepts an incoming call, as long as the required bandwidth is available. By inspecting the nature of the best policy obtained using NDP, we observed that only some of the customer classes were ever deliberately rejected, and we were then able to use this knowledge to handcraft a trunk reservation (threshold) policy that attained comparable performance. However, in the absence of adequate tools for tuning trunk reservations parameters (as it is the case for large networks), the use of NDP can become very attractive. In addition, this illustrates that the quadratic approximator provides an adequate architecture for the differential reward function of a single link.

The parameters and results of the case studies are given in the Tables 1-5. One conclusion from these experiments is that NDP led to significantly better results

Table 3: Case study for 10 service classes and a highly loaded link with a capacity of 600 units.

	Performance	
	Average Reward	Lost Average Reward
Always Accept	412.77	293.15
Trunk Reservation	518.17	187.67
TD(0): MLP	505.46	200.34
TD(0): Quadratic	511.45	194.71

Table 4: Case study for 10 service classes and a medium loaded link with a capacity of 600 units.

	Performance	
	Average Reward	Lost Average Reward
Always Accept	389.53	34.31
Trunk Reservation	396.68	26.97
TD(0): MLP	393.68	29.87
TD(0): Quadratic	385.39	38.63

Table 5: Case study for 10 service classes and a lightly loaded link with a capacity of 600 units.

	Performance	
	Average Reward	Lost Average Reward
Always Accept	370.82	8.89
Trunk Reservation	372.40	7.70
TD(0): MLP	370.82	8.89
TD(0): Quadratic	370.82	8.89

Table 6: Service classes and arrival rates for the 4-node network.

Service Class m	1	2	3
Bandwidth Demand $b(m)$	1	3	5
Average Holding Time $1/\nu(m)$	10	10	2
Immediate Reward $c(m)$	1	2	50

	Arrival Rates		
Service Class	1	2	3
Origin-Destination Pairs (0-2)(2-0)(1-3)(3-1)	4.40	2.64	0.44
All Other Origin-Destination Pairs	2.00	1.20	0.20

than the heuristic "always accept" policy, except for the case of a lightly loaded link and 10 classes, where the performance of both approaches was the same. (This is understandable because for a lightly loaded system "interesting" events such as blocking are too rare to be able to fine-tune the policy.) In particular, for all cases, except for the one just mentioned, the rewards associated with calls that were blocked or deliberately rejected (these are the lost rewards), were reduced by 10–35%. For the case of three classes, essentially optimal performance was attained. It was also seen that the MLP architecture did not lead to performance improvements, and this was an important reason for not using it in larger problems.

4.2 A 4-Node Network

In this section, we present experimental results obtained for the case of an integrated services network consisting of 4 nodes and 12 unidirectional links. There are two different classes of links with a total capacity of 60 and 120 units of bandwidth, respectively (indicated by thick and thin arrows in Figure 1). We assume a set $\mathcal{M} = \{1, 2, 3\}$ of three different service classes. The corresponding parameters are given in Table 6. Note that the calls of type 3 are much more valuable than the one of type 1 and 2. Furthermore, for each pair of source and destination nodes, the list of possible routes consists of three entries: the direct path and the two alternative 2-hop-routes.

This case study is characterized by a high traffic load and by calls of one service class having a much higher immediate reward than calls of the other types. Clearly, for this case, a good call admission control and routing policy should give "priority" to calls of the service class with the highest reward. We chose this setting to determine the potential of our optimization algorithm, i. e. to find out if NDP indeed discovers a control policy which reserves bandwidth for calls of the most

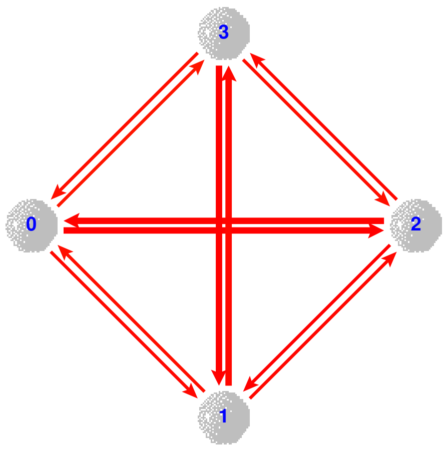

Figure 1: Telecommunication network consisting of 4 nodes and 12 unidirectional links.

valuable service type.

This experiment was carried out using TD(0) for discounted problems combined with the decomposition approach. However, the performance of the resulting policies was evaluated on the basis of the average reward criterion. Our value function approximator contains 120 tunable parameters. There are approximately $1.6 \cdot 10^{45}$ different link state (feature) configurations. Note that the cardinality $|S|$ of the underlying state space is even higher. We make the following observations.

(a) Employing the decomposition approach did not affect the the performance of our final NDP policy and reduced the training time by a factor of 2. (Note that the decomposed optimization updates the parameters corresponding to only five links instead of twelve at every time step.) This was an important reason for using it in larger problems (see Section 4.3).

(b) In order to assure convergence of the *discounted* TD(0) method we had to carefully handcraft some of the initial parameter values of our function approximator. In particular the magnitude of the parameter $\theta(l)$ associated with each link turned out to be critical. This procedure becomes rapidly impractical as the number of links increases. Larger problems can be solved much easier using *average reward* algorithms which are less sensitive in this respect (see Section 4.3).

(c) For this case study we could significantly improve the performance of the resulting policy by enforcing an explicit *exploration* of the state space during the training. At each state, with probability $p = 0.5$, we apply a random action, instead of the action recommended by the current value function, to

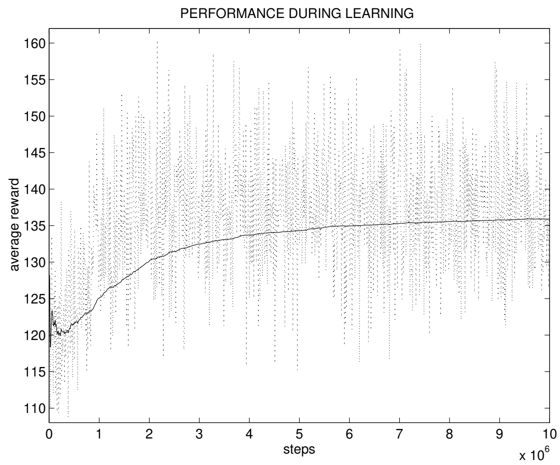

Figure 2: "Empirical" average reward per time unit during the whole training phase of 10^7 steps (solid) and during shorter time windows of 10^5 steps (dashed).

generate the next state in our training trajectory. However, the successor state $x^{(l)}_{t^{(l)}_k}$ that is used in update rule (11) is still chosen according to the greedy action given in (10). The importance of using a certain amount of exploration in connection with NDP methods is well-known (see for example [BT96]).

The results of the case study is given in Figure 2 (training phase), Figure 3 (performance) and Figure 4 (routing behavior). We give here a summary of the results.

Training phase: Figure 2 shows the performance improvement during the optimization phase. Here, the "empirical" average reward of the NDP policy (computed by averaging the rewards obtained during the whole training and during shorter time window of 10^5 steps) is depicted as a function of the training steps. Although this average reward increases during the training, it does not exceed 141, the average reward of the heuristic OSPF. This is due to the high amount of exploration in the training phase. We obtained the final control policy after 10^7 iteration steps.

Performance comparison: We used simulated trajectories of 10^7 time steps to evaluate our policies. The policy obtained through NDP gives an average reward of 212, which as about 50% higher than the one of 141 achieved by OSPF. Furthermore, the NDP policy reduces the number of rejected calls for all service classes. The most significant reduction is achieved for calls of service class 3, the service class, which has the highest immediate reward. Figure 3 also shows that the average reward of the NDP policy is close to the potential average reward of 242, which is the average reward we would obtain if all calls were accepted. This leaves us to believe that the NDP policy is close to optimal. Figure 4 compares the routing behavior of

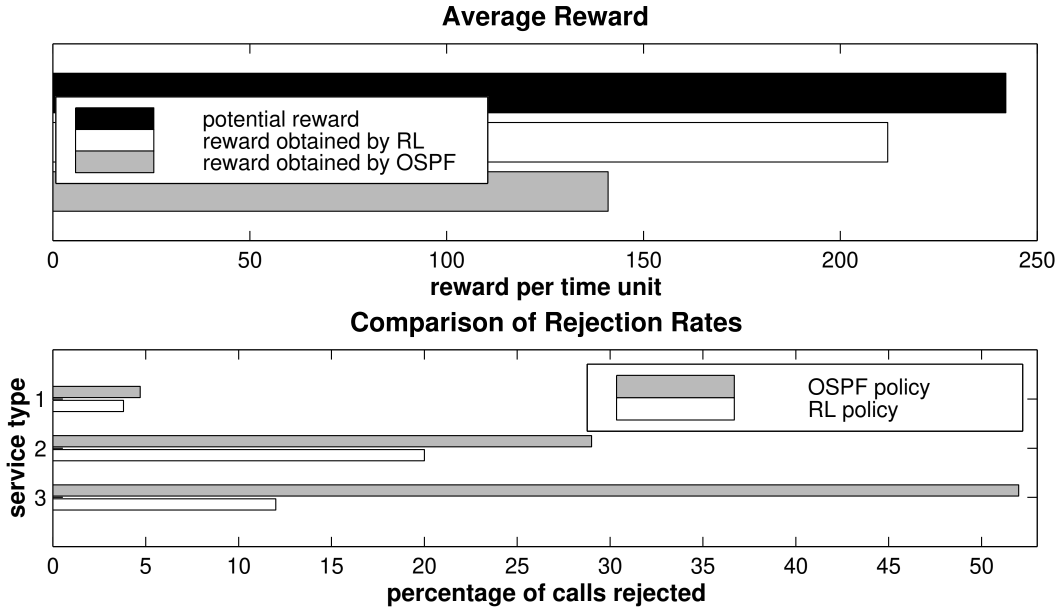

Figure 3: 4-node network: Comparison of the average rewards and rejection rates of the NDP and OSPF policies.

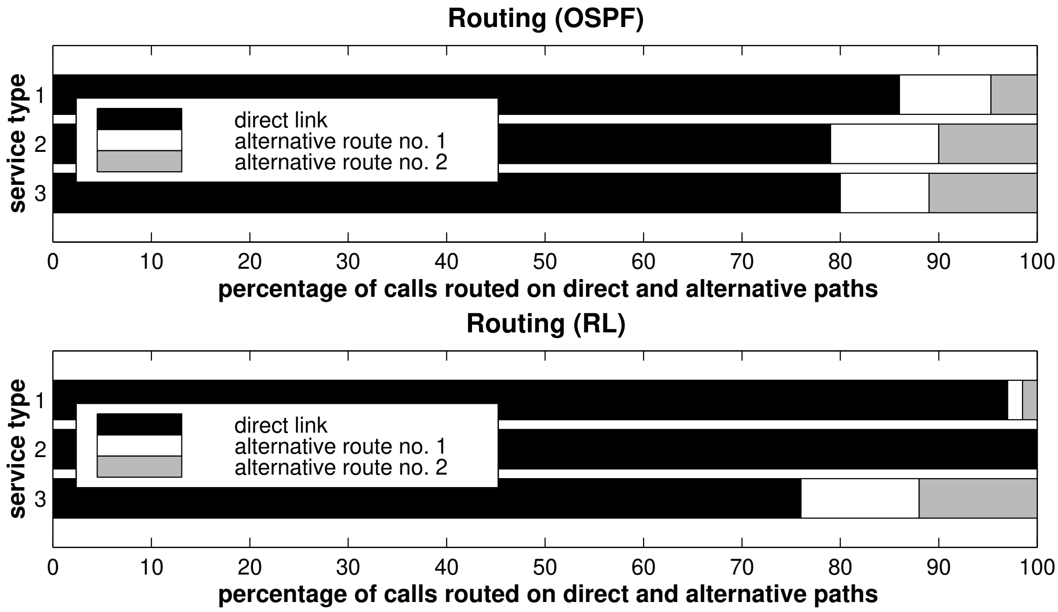

Figure 4: 4-node network: Comparison of the routing behavior of the NDP and OSPF policies.

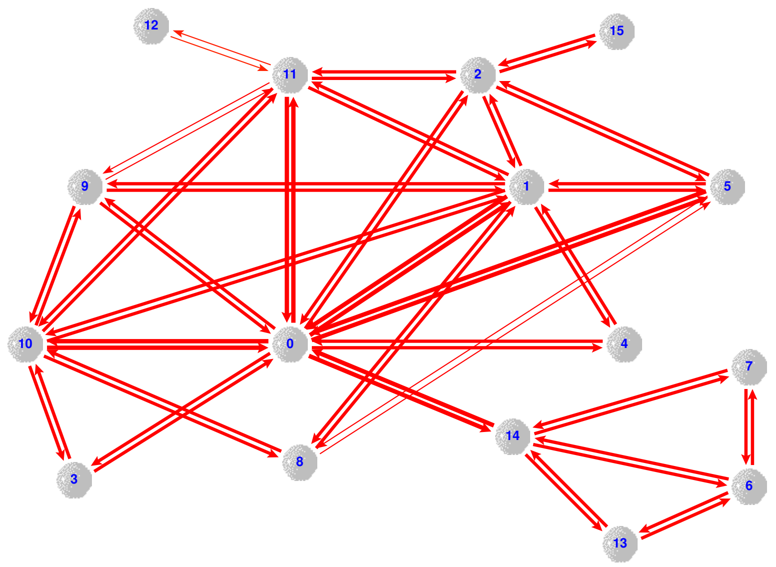

Figure 5: Telecommunication network consisting of 16 nodes and 62 unidirectional links.

the NDP control policy and OSPF. While OSPF routes about 15%-20% of all calls along one of the alternative 2-hop-routes, the NDP policy uses alternate routes for calls of type 3 (about 25%) and routes calls of the other two service classes almost exclusively over the direct route. This indicates, that the NDP policy uses a routing scheme, which avoids 2-hop-routes for calls of service class 1 and 2, and which allows us to use network resources more efficiently.

4.3 A 16-Node Network

In this section, we present experimental results obtained for a network consisting of 16 nodes and 62 unidirectional links (see Figure 5). The network topology is taken from [GS97]. The network consists of three different classes of links with a capacity of 60, 120 and 180 units of bandwidth, respectively. We assume four different service classes. Table 7 summarizes the corresponding bandwidth demands, average holding times and immediate rewards. The table of arrival rates is also taken from [GS97]. However, for our experiments we rescaled them by a factor of 2. The list of accessible routes consists of a maximum of six minimal hop routes for each pair of source and destination nodes. Routes with an equal number of hops are ordered by their absolute path length (in miles) which is also reported in [GS97].

For this experiment, there are approximately $1.4 \cdot 10^{256}$ different link state (feature) configurations and 992 tunable parameters. The results of the case study are summarized by Figure 6 (training), Figure 7 (performance), Figure 8 (routing),

Table 7: Service classes for the 16-node network.

Service Class m	1	2	3	4
Bandwidth Demand $b(m)$	1.00	2.00	3.00	4.00
Average Holding Time $1/\nu(m)$	1.25	1.25	1.25	1.25
Immediate Reward $c(m)$	0.25	1.00	6.00	15.00

and Figure 9 (robustness).

We make the following observations.

(a) Without using the decomposition approach, no substantial improvement over the initial policy is achieved within a reasonable amount of computation time (24 hours, say). This illustrates the importance of the decomposition approach in applying NDP to the call admission control and routing problem.

(b) *Discounted* reward algorithms failed due to their critical dependence on initial parameter values (see Section 4.2). This difficulty does not arise with *average* reward algorithms.

(c) Instabilities can occur during the training phase, even when exploration is employed (see the discussion below).

(d) Our NDP policies are very robust with respect to changes of the underlying arrival statistics.

Training phase: Figure 6 shows the "empirical" average reward of the NDP policy (computed by averaging the rewards obtained during the simulation run) as a function of the training steps. In contrast to the 4-node example the NDP policy does not converge towards a final policy better than OSPF, although the average reward significantly improved during the first $4{\cdot}10^5$ training steps. Afterwards, a sudden performance breakdown occurs, from which the system never recovers. This loss of stability did not disappear, even if we introduce explicit exploration during the training. For the subsequent performance comparison between NDP and OSPF we pick the best policy (given by the parameter values just before the loss of stability) generated in the course of the algorithm, not the last one.

Performance comparison: The policies are empirically evaluated based on simulated trajectories of 10^7 time steps. The OSPF policy almost exclusively routes all calls over the shortest path. This leads to an average reward of about 4254. The rate of rejected calls is positive for all service classes. The two most valuable service classes 3 and 4 receive the highest rejection rate. In contrast, the NDP policy comes up with a very different routing scheme that uses alternative paths

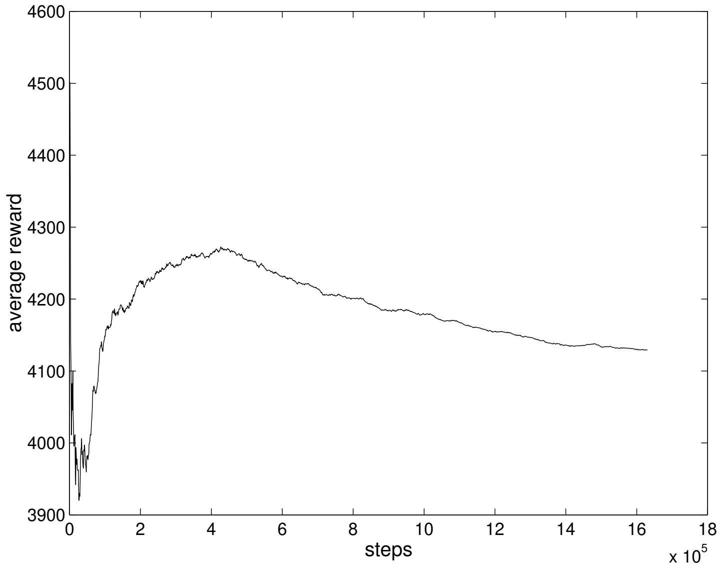

Figure 6: "Empirical" average reward obtained during the training as a function of training steps. The performance initially improves and then suddenly deteriorates.

for all types of services. Now, the rejection rates for calls of type 1, 3 and 4 vanish whereas that for service class 2 increases. The NDP policy rejects these calls in a strategic way, i. e. NDP is not forced to do so by the capacity constraint. Instead, it explicitly reserves bandwidth for the most valuable calls of type 3 and 4. The average reward of 4349 obtained through the NDP policy is about 2.2% higher than the one achieved by OSPF. A more significant performance measure illustrating the power of the NDP policy is the so-called lost average reward. It is defined as the difference between the potential average reward of 4438 and the actual one achieved by the current policy. NDP reduces the lost average reward by about 52% compared with OSPF.

Robustness: We applied our best policy obtained through training under the above mentioned arrival statistics to problems with randomly changed arrival rates in order to show the robustness of NDP policies. In particular, each arrival rate is multiplied by a factor $1+\rho$, where $\rho \in [-\alpha, \alpha]$ is independently drawn from a uniform distribution. An arrival rate is set to zero, if $1+\rho$ happens to be negative. We carried out a set of experiments by varying the magnitude $\alpha \in [0, 2]$ in steps of 0.1, which amounts to rather strong perturbations of the traffic statistics. Figure 9 shows the result of these experiments. The magnitude α of the relative perturbations of the arrival rates is depicted against the relative lost reward defined as

$$\frac{v(\mu_{\text{NDP}}) - v(\mu_{\text{OSPF}})}{v_{\text{potential}} - v(\mu_{\text{OSPF}})}.$$

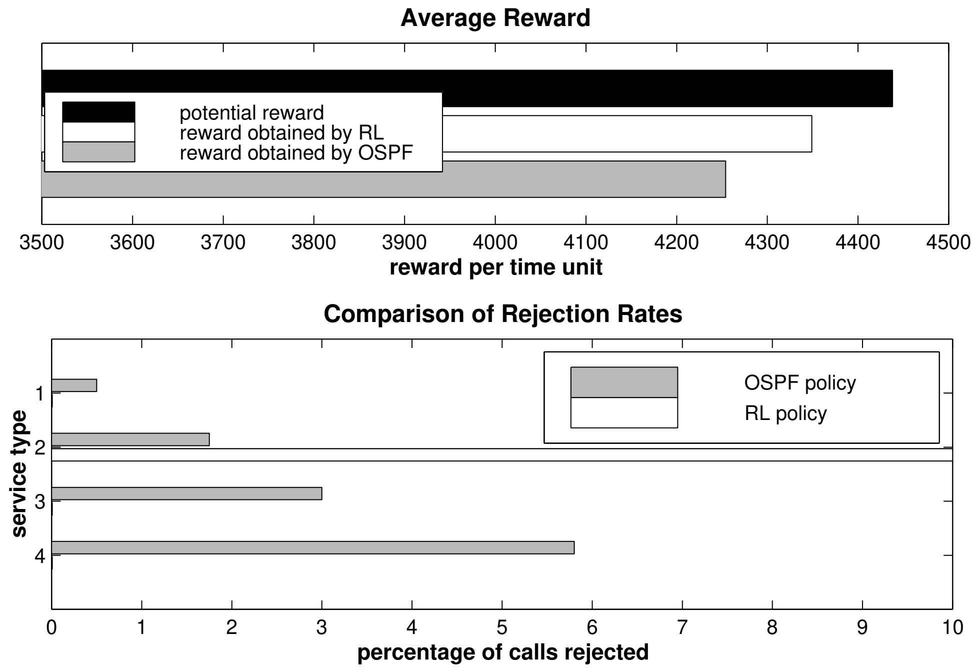

Figure 7: 16-node network: Comparison of the average rewards and rejection rates of the NDP and OSPF policies.

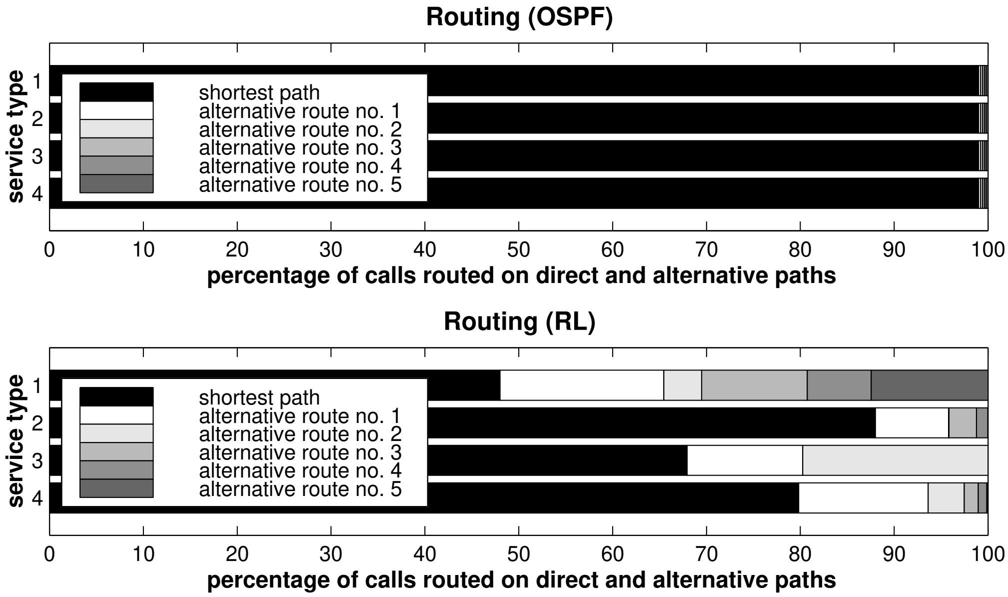

Figure 8: 16-node network: Comparison of the routing behavior of the NDP and OSPF policies.

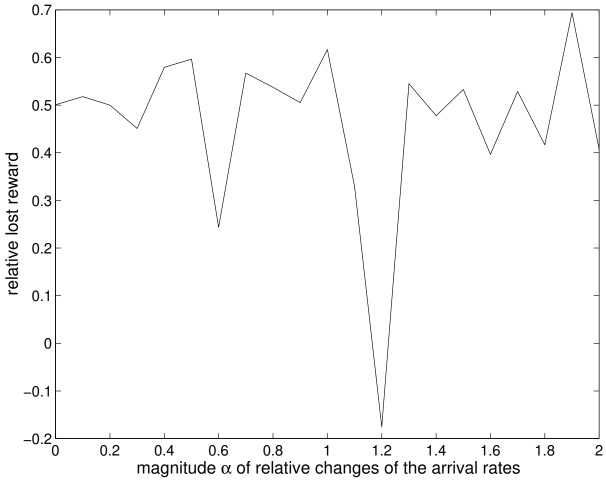

Figure 9: Relative lost reward of the NDP policy applied to networks with randomly changed arrival statistics.

Here, $v_{\text{potential}}$, μ_{NDP} and μ_{OSPF} denote the potential average reward, the NDP policy and the OSPF policy, respectively. The experiments show, that our NDP policy is indeed very robust against changes in the arrival rates. There is only one out of twenty experiments where the NDP policy happened to be worse then OSPF. (We did not average several experiments with equal perturbation parameter α.) For all other arrival statistics the NDP policy still outperforms OSPF with a relative lost reward between 25% and 70%.

5 Conclusion

The call admission control and routing problem for integrated service networks is naturally formulated as an average reward dynamic programming problem, but with a very large state space. Traditional dynamic programming methods are computationally infeasible for such large scale problems. We use neuro-dynamic programming, based on the average reward TD(0) method of [TV97b], combined with a decomposition approach that views the network as consisting of decoupled link processes. This decomposition has the advantage that it allows for decentralized decision making and decentralized training, which reduces significantly the training time. We have presented experimental results for several example problems, of different sizes. The case study involving a 16-node network shows that NDP can lead to sophisticated control policies involving strategic call rejections, and which are difficult to obtain through heuristics.

Concerning the practical applicability of this general methodology, there are two somewhat distinct issues. The first is whether dynamic policies based on state-dependent costs (depending linearly on the variables $n_{l,m}$) can lead to significant performance improvements. Our results suggest that this is indeed the case, although a comparison with alternative policies (such as dynamic alternative routing with trunk reservation) remains to be made. A somewhat related issue is whether efficient performance evaluation tools are possible (based on ideas similar to the reduced load approximation, that do not involve simulation) which apply to policies of the form considered in this paper.

The second issue refers to computational requirements. Simulation-based methods such as TD can be slow. For example, the computation times for our different experiments ranged from one to four hours of CPU time on Sun Sparc 20 workstation. On the other hand, once we can see promise in an application domain, a variety of ways of improving speed can be considered. Besides optimizing the code, these could include batch linear least squares methods for tuning θ (to replace small step size incremental training), or the use of a smaller set of tunable parameters after identifying those "features" that are most critical for improved performance.

References

[LBM98] M. Y. Liu, J. S. Baras, and A. Misra, "Performance evaluation in multi-rate, multi-hop communication networks with adaptive routing". *ARL Federated Laboratory 2nd Annual Conference*, University of Maryland, College Park, MD, February 1998.

[Ber95] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, 1995.

[BT96] D. P. Bertsekas, and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

[BTS99] T. X. Brown, H. Tong, and S. Singh, "Optimizing admission control while ensuring quality of service in multimedia networks via reinforcement learning". To appear in *Advances in Neural Information Processing Systems 11*, MIT Press, Cambridge, MA, 1999.

[CN97] J. Carlström, and E. Nordström, "Control of self-similar ATM call traffic through reinforcement learning". *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunication 3 (IWANNT*97)*, J. Alspector, R. Goodman, T. X. Brown (eds.), pp. 54–62, Lawrence Erlbaum, Melbourne, Australia, 1997.

[CR93] S. Chung, and K. W. Ross, "Reduced load approximations for multi-rate, multi-hop communication networks". *IEEE Trans. Commun.*, vol. 41, no. 8, August 1993.

[DM94] Z. Dziong, and L. G. Mason, "Call admission and routing in multi-service loss networks". *IEEE Trans. Commun.*, vol. 42, no. 2/3/4, February/March/April 1994.

[GS97] A. Greenberg, and R. Srikant, "Computational techniques for accurate performance evaluation in multirate, multihop communication Networks". *IEEE/ACM Transactions on Networking*, vol. 5, no. 2, April 1997.

[Hay94] S. Hayking, *Neural Networks: A Comprehensive Foundation*. McMillian, N. Y., 1994.

[Kel88] F. P. Kelly, "Routing in circuit switched networks: optimization, shadow prices and decentralization". *Adv. Appl. Prob.*, vol. 20, 1988.

[Kel91] F. P. Kelly, "Loss networks". *Annals of Applied Probability*, vol. 1, pp. 319–378, 1991.

[Key90] P. B. Key, "Optimal control and trunk reservation in loss networks". *Prob. Engrn. Inf. Sci.*,vol. 4, pp. 203–242, 1990.

[Laws95] C. N. Laws, "On trunk reservation in loss networks". In *Stochastic Networks*, F. P. Kelly and R. J. Williams (eds.), vol. 71 of IMA Volumes in Mathematics and its Applications, pp. 187–198. Springer-Verlag, New York, 1995.

[Liu97] M. Liu, *Performance evaluation in multi-rate multi-hop communication networks*. Master's thesis, U. Maryland, College Park, MD 1997. Available from http://www.isr.umd.edu/TechReports/ISR/1997.

[MT97] P. Marbach, and J. N. Tsitsiklis, *A Neuro-Dynamic Programming Approach to Call Admission Control in Integrated Service Networks: The Single Link Case*. Technical Report LIDS-P-2402, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, November 1997. Available from http://web.mit.edu/jnt/www/publ.html.

[MMR96] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, "ATM network design and optimization: a multirate loss network framework". *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, August 1996.

[NC95] E. Nordström, and J. Carlström, "A reinforcement learning scheme for adaptive link allocation in ATM networks". *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunication*

*2 (IWANNT*95)*, J. Alspector, R. Goodman, T. X. Brown (eds.), pp. 88–95, Lawrence Erlbaum, Stockholm, Sweden, 1995.

[Ros95] K. W. Ross, *Multiservice Loss Models for Broadband Communication Networks*. Springer-Verlag, Berlin, Heidelberg, New York, 1995.

[SiB97] S. Singh, and D. P. Bertsekas, "Reinforcement learning for dynamic channel allocation in cellular telephone systems". In M. C. Mozer and M. I. Jordan and T. Petsche (eds.), *Advances in Neural Information Processing Systems 9*, pp. 974–980. MIT Press, Cambridge, MA, 1997.

[Sut88] R. S. Sutton, "Learning to predict by the methods of temporal differences". *Machine Learning*, vol. 3, pp. 9–44, 1988.

[SB98] R. S. Sutton, and A. B. Barto, *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[Tes88] J. Tesauro, "Practical issues in temporal difference learning". *Machine Learning*, vol. 8, 1988.

[TV97a] J. N. Tsitsiklis, and B. Van Roy, "An Analysis of Temporal-Difference Learning with Function Approximation". *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, May 1997

[TV97b] J. N. Tsitsiklis, and B. Van Roy, *Average cost temporal-difference learning*. Lab. for Info. and Decision Systems Report LIDS-P-2390, Massachusetts Institute of Technology, Cambridge, MA; accepted for publication in *Automatica*, 1997.

[WV96] J. Walrand, and P. Varaiya, *High-Performance Communication Networks*. Morgan Kaufman, San Francisco, CA, 1996.

[Whi88] P. Whittle, "Approximation in large-scale circuit-switched networks". *Prob. Engrn. Inf. Sci.*,vol. 2, pp. 279–291, 1988.

[ZD96] W. Zhang, and T. G. Dietterich, "High performance job-shop scheduling with a time-delay TD(λ) network". In D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems 8*, pp. 1024–1030. MIT Press, Cambridge, MA, 1996.