The stochastic single node service provision problem*

Shane Dye†
Email: s.dye@mang.canterbury.ac.nz

Leen Stougie‡
Email: leen@win.tue.nl

Asgeir Tomasgard§
Email: Asgeir.Tomasgard@indman.sintef.no

November 15, 2001

Abstract

The service provision problem described in this paper comes from an application of distributed processing in telecommunications networks. The objective is to maximize a service provider's profit from offering computational based services to customers. The service provider has limited capacity and must choose from a set of software applications those he would like to offer. This can be done dynamically taking into consideration that demand for the different services is uncertain. The problem is examined in the framework of stochastic integer programming.

Approximations and complexity are examined for the case when demand is described by a discrete probability distribution. For the deterministic counterpart a fully polynomial approximation scheme is known [2]. We show that introduction of stochasticity makes the problem strongly NP-hard, implying that the existence of such a scheme for the stochastic problem is highly unlikely. For the general case a heuristic with a worst-case performance ratio that increases in the number of scenarios is presented. Restricting the class of problem instances in a way that many reasonable practical problem instances will satisfy, allows for the derivation of a heuristic with a constant worst-case performance ratio. These worst-case results are the first results for stochastic programming problems that the authors are aware of in a direction that is classical in the field of combinatorial optimization.

*For this work, financial support has been received from Telenor and Leonardo Da Vinci. Dye and Tomasgard were employees at The Norwegian University of Science and Technology in Trondheim, Norway when the main part of the work was done.

†University of Canterbury, New Zealand

‡Eindhoven University of Technology and CWI Amsterdam, The Netherlands

§SINTEF Industrial Management, Norway

Keywords: distributed processing; telecommunications; service provision; stochastic (integer) programming; strong NP-hardness; approximation; worst-case analysis

1 Introduction

The service provision problem discussed in this paper comes from an application in telecommunications. It considers how to install different *processing based services* on a set of computer nodes in a network with distributed processing capabilities. The computers typically have limited resources such as memory, processing capacity and storage capacity. All the services are built from a set of subservices. The subservices are software applications, which run in a distributed manner in the network. The service provider must decide how to allocate computational resources to a set of subservices in order to meet customer demand for services. Because the resources are limited, it may be necessary to reject some customers. It is assumed that the service provider tries to maximize his profit.

From the prognosis that the problem of allocating node resources will be important in near future (as one can already see for the Internet) the authors were asked by the industrial financial contributor to examine the situation where transportation does not play a role. Further, because of the distributed processing capabilities of the network, it is possible to consider subservice demand independently of which service generated it.

Demand for services is dynamic and uncertain. At various times the demand for a single service peaks, affecting the demand for all subservices used by the service. Before the peak actually occurs, deviations from the normal demand patterns for subservices can be observed. These deviations can be used as a signal indicating that a peak is about to occur. The signals can be ambiguous but point to a limited number of possible services that might peak. For any possible signal a few scenarios often give sufficient description of the situation that is about to occur in terms of subservice demand.

The subservices typically take time and resouces for start-up and shutdown. The configuation of subservices can not react to changes in demand instantaneously. When the signal gives just enough time to re-configure the network before the peak occurs, a two-stage decision situation naturally emerges. In the first stage the decision is which subservices to install given only probabilistic information on demand for subservices. During the set-up time uncertainty resolves itself. The only possible recourse action in the second-stage concerns what demand should be met using the subservices installed in the first stage. The available capacity is restricted by the first stage decision. More information on the model can be found in Tomasgard et al [10].

This paper considers a variant of the problem with only one node on which to install subservices, and a single constraining resource. This is typically the situation a service provider faces when he rents capacity from a network provider. The service provider does not take into consideration whether the capacity he has rented is located on one or several computing nodes. He uses it as if it were one continuous block of capacity. The network provider on the other hand is free to replicate and move the various service providers' subservices on all the nodes he manages. For a further discussion of the roles in the network and a

discussion around distribution see [9, 10].

Here the underlying decision process is briefly described.

Demand is treated in terms of the limited resource used by the subservices. Let n be the number of subservices and s the resource capacity of the single node. q_j is the profit obtained from allocating one resource unit to meeting demand for subservice, j. In addition, each subservice uses a fixed amount of capacity just to be available, independent of the demand met. This *installation requirement* is denoted by r_j for subservice j. Subservice demand is uncertain and described by the probability space (Δ, σ, μ). Let $\underline{\delta} \in \Delta$ be a realization of the demand, where δ_j is the demand for subservice j for this random outcome.

The first stage decision variables z_j indicate whether subservice j is installed, in which case $z_j = 1$, or not, indicated by $z_j = 0$, $j = 1, \ldots, n$.

The objective of the first stage is to maximise expected profit, subject to a capacity constraint.

$$\begin{aligned} \max \quad & E_{\Delta}\left[Q(z, \delta)\right] \\ \text{s.t.} \quad & \sum_{j=1}^{n} r_j z_j \leq s \\ & z_j \in \{0, 1\} \qquad j = 1, \ldots, n. \end{aligned} \tag{1}$$

where $Q(z, \delta)$ is the second-stage cost for first stage decision z and demand δ. This is the optimal objective value of the second-stage linear program, where z and δ act as parameters. The second-stage variables x_j denote the resource used to meet demand for subservice j. The objective of the second stage is to maximise profit. There are two constraints. The *capacity constraint* ensures that node capacity is not exceeded. The *demand constraint* ensures that demand is only met for subservices that have been installed.

$$\begin{aligned} Q(z, \delta) = \max \quad & \sum_{j=1}^{n} q_j x_j \\ \text{s.t.} \quad & \sum_{j=1}^{n} x_j \leq s - \sum_{j=1}^{n} r_j z_j \\ & x_j \leq \delta_j z_j \qquad j = 1, \ldots, n, \\ & x_j \geq 0 \qquad j = 1, \ldots, n. \end{aligned} \tag{2}$$

When the node capacity, the installation requirements, and demands are integral, the x variables will automatically be integral.

When uncertain demand for subservices is described by a discrete distribution a *deterministic equivalent* [5] can be formulated, as discussed in [10].

The probability distribution of uncertain demand is described in stochastic programming terminology in terms of scenarios [5]. Denote by m the number of demand scenarios and by p_k the probability of scenario k occurring. A scenario can be viewed as a vector of demands with an assigned probability. Then, δ_{jk} is demand for the resource generated by subservice j in scenario k.

The second-stage variables become x_{jk}, denoting the resource allocated to subservice j in scenario k.

The deterministic equivalent of the stochastic single node service provision

problem (SSNP) will be a linear mixed integer programming model (MIP) [7].

$$\begin{array}{lll} \max & \sum_{k=1}^{m} p_k \sum_{j=1}^{n} q_j x_{jk} & \\ \text{s.t.} & \sum_{j=1}^{n} (r_j z_j \quad + x_{jk}) \leq s & k = 1, \ldots, m, \\ & \delta_{jk} z_j \quad - x_{jk} \geq 0 & j = 1, \ldots, n, \quad k = 1, \ldots, m, \\ & z_j \in \{0, 1\}, \; x_{jk} \quad \geq 0 & j = 1, \ldots, n, \quad k = 1, \ldots, m. \end{array} \tag{3}$$

In the remainder of this paper the expected demand for subservice j will be written in the following manner

$$E_k[\delta_{jk}] = \sum_{k=1}^{m} p_k \delta_{jk},$$

diverging slightly from customary notation for expectations in probability theory literature.

The mathematical program of interest is a stochastic integer program. As stated, the integrality is purely in the first stage. When the input data is integral, the second stage is naturally integer and the problem may be classified as having an integer second stage. From the stand point of stochastic integer programming, the formulation is interesting in and of itself. Our analysis highlights an interesting result. When the number of scenarios allowed is fixed, the problem may be solved in pseudo-polynomial time. However, for an arbitrary number of scenarios, the formulation is strongly NP-hard. For this problem, the better the description of uncertainty, the more difficult the problem becomes. This suggests that algorithms for general stochastic integer programming, or those that rely on the uncertainty structure, are unlikely to be "scalable".

To facilitate the exposition the assumption is made that no demand is higher than the node capacity minus the corresponding installation requirement. This can, if necessary, be ensured by preprocessing.

Assumption 1 *For any subservice j in any scenario k, the support of δ_{jk} is in the interval $[0, s - r_j]$.*

A consequence of this is that for any subservice the profit of meeting its expected demand is no greater than the optimal profit of the overall problem. Let π^{OPT} be the optimal value of (3). Then Assumption 1 ensures that

$$\pi^{\text{OPT}} \geq q_j E_k[\delta_{jk}], \qquad j = 1, \ldots, n. \tag{4}$$

Feasibility of the deterministic service provision problem with multiple nodes and the requirement that all demand must be met is strongly NP-complete, [3]. When demand is deterministic and profit is maximized, Dye et al [2] show that the single node problem is NP-hard and has a fully-polynomial time approximation scheme. In the same paper it is shown that the multiple node problem is strongly NP-hard and that there exists no fully polynomial approximation scheme even when the number of nodes is fixed. The analysis turned out to have many similarities with the well known knapsack problem [6]. The results do not follow straightforwardly from the deterministic counterparts of the problem.

We show in Section 4 that (SSNP) is strongly NP-hard, whereas as noted above the deterministic counterpart admits a fully polynomial approximation scheme. This is remarkable since the integer variables appear only in the first stage of the two-stage stochastic programming problem. When the number of scenarios is fixed the problem can be solved in pseudo-polynomial time by dynamic programming.

When the number of scenarios is considered as part of the input, there is little hope to find efficient algorithms that solve the problem to optimality or fully polynomial time approximation schemes. It is still possible to find good approximations. This is the motivation behind investigating the LP relaxation. The LP relaxation is discussed in Section 2 together with an approximation method directly based on the LP results. A worst case bound increasing in the number of scenarios is given. In Section 3, for a slightly restricted problem class (to which many reasonable practical problem instances belong) the bound on the ratio between the LP solution value and the optimal integer one is tightened and a constant bound approximation method based on the proof is presented. These are the first worst-case performance results known by the authors for approximation of stochastic integer programming problems.

2 The LP bound and a heuristic

The LP relaxation of (SSNP) replaces the requirement $z_j \in \{0,1\}$ in (3) by $0 \leq z_j \leq 1$ for $j = 1, \ldots, n$. This section describes an optimal basis for the LP relaxation of (SSNP) and uses it to give an upper bound on the ratio of the LP versus the optimal solution. A heuristic based on the bound is given subsequently in Subsection 2.2.

2.1 The LP bound

Relaxing the integrality constraints, consider the resulting LP. The following theorem bounds the number of fractional variables in an optimal LP solution. A variable z_j is fractional if $0 < z_j < 1$, and a variable x_{jk} is fractional if $0 < x_{jk} < \delta_{jk} z_j$. Note that if $z_j < 1$, then it is possible to have $0 < x_j < \delta_{jk}$ without x_{jk} being fractional, as long as it is equal to $\delta_{jk} z_{jk}$.

Theorem 1 *Any basic optimal solution to the LP relaxation of (SSNP) with m scenarios has at most m fractional z and x variables.*

Proof Let $(z^{\mathrm{LP}}, x^{\mathrm{LP}})$ be an optimal basic solution to the LP relaxation of (SSNP). Define the reduced problem to be the instance with problem data corresponding to the original, with the exception that subservices for which $z_j^{\mathrm{LP}} = 0$ are removed. The corresponding optimal solution of the reduced problem has the same number of fractional x and z variables. This means the only instances to consider have an LP relaxation with a basic optimal solution, $(z^{\mathrm{LP}}, x^{\mathrm{LP}})$, for which $z^{\mathrm{LP}} > 0$.

Introducing slacks t_k for the capacity constraints and u_{jk} for the demand constraints, results in the following reformulation of the LP relaxation.

$$\max \sum_{j=1}^{n} \sum_{k=1}^{m} p_k q_j x_{jk} \tag{5}$$

s.t.

$$\begin{aligned}
\sum_{j=1}^{n} r_j z_j + \sum_{j=1}^{n} x_{jk} + t_k &= s \quad k = 1, \ldots, m, && (6)\\
\delta_{jk} z_j - x_{jk} - u_{jk} &= 0 \quad j = 1, \ldots, n,\ k = 1, \ldots, m, && (7)\\
0 \le z_j \le 1,\ x_{jk}, u_{jk}, t_k &\ge 0 \quad j = 1, \ldots, n,\ k = 1, \ldots, m. && (8)
\end{aligned}$$

This LP has $m+nm$ functional constraints, so that at any basic solution at most $m+nm$ variables will lie strictly between their bounds. Let $(t^{\mathrm{LP}}, u^{\mathrm{LP}}, x^{\mathrm{LP}}, z^{\mathrm{LP}})$ be a basic optimal solution to the above for which $z^{\mathrm{LP}} > 0$. Now, count the number of variables lying strictly between their bounds.

Since $z_j^{\mathrm{LP}} > 0$, $j = 1, \ldots, n$, Constraints (7) imply that at least one of x_{jk}^{LP} or u_{jk}^{LP} will be positive for each pair (j,k), $j = 1, \ldots, n$, $k = 1, \ldots, m$. This accounts for at least nm variables strictly between their bounds. Define the following sets

$$\mathcal{F} = \{\, j \mid z_j^{\mathrm{LP}} < 1 \,\},$$

$$\mathcal{U} = \{\, (j,k) \mid x_{jk}^{\mathrm{LP}} > 0 \text{ and } u_{jk}^{\mathrm{LP}} > 0 \,\}$$

and

$$\mathcal{T} = \{\, k \mid t_k^{\mathrm{LP}} > 0 \,\}.$$

Notice that $\mathcal{U}$ is exactly the set of indices for which x_{jk}^{LP} are fractional because they are positive but not equal to $\delta_{jk} z_j^{\mathrm{LP}}$.

The number of fractional z^{LP} and x^{LP} is $|\mathcal{F}|+|\mathcal{U}|$ and the number of variables lying strictly between their bounds is $|\mathcal{F}|+|\mathcal{T}|+|\mathcal{U}|+nm$. From the above this is no greater than $m+nm$, implying $|\mathcal{F}|+|\mathcal{T}|+|\mathcal{U}| \le m$. □

Thus, if $(z^{\mathrm{LP}}, x^{\mathrm{LP}})$ is a basic optimal solution to the LP relaxation write its optimal value, π^{LP}, as

$$\pi^{\mathrm{LP}} = \sum_{j \in \mathcal{W}} \sum_{k=1}^{m} p_k q_j x_{jk}^{\mathrm{LP}} + \sum_{j \in \mathcal{F}} \sum_{k=1}^{m} p_k q_j x_{jk}^{\mathrm{LP}} \tag{9}$$

where $\mathcal{W} = \{j | z_j^{\mathrm{LP}} = 1\}$ and $\mathcal{F} = \{j | 0 < z_j^{\mathrm{LP}} < 1\}$

In particular, $|\mathcal{F}| \le \min\{m, n\}$. Under Assumption 1, the above theorem provides an immediate bound for the optimal value of the LP relaxation in terms of the optimal solution value π^{OPT} of (SSNP).

Corrollary 1 *If π^{OPT} is the optimal solution value of an instance of (SSNP) and π^{LP} is the optimal value of the LP relaxation, $\pi^{\mathrm{LP}} \le \min\{m+1, n\}\pi^{\mathrm{OPT}}$.*

Proof

$$\begin{aligned}
\pi^{\mathrm{LP}} &\le \sum_{j \in \mathcal{W}} \sum_{k=1}^{m} p_k q_j x_{jk}^{\mathrm{LP}} + \sum_{j \in \mathcal{F}} q_j \sum_{k=1}^{m} p_k \delta_{jk} && (10)\\
&\le \pi^{\mathrm{OPT}} + \sum_{j \in \mathcal{F}} \pi^{\mathrm{OPT}}\\
&\le (\min\{m, n\} + 1)\pi^{\mathrm{OPT}}
\end{aligned}$$

Notice that $\sum_{j\in\mathcal{W}}\sum_{k=1}^{m} p_k q_j x_{jk}^{\mathrm{LP}}$ is the value of an integer feasible solution and is therefore no greater than π^{OPT}. To obtain the second inequality is then a matter of applying (4). The last inequality is implied by $|\mathcal{F}| \leq \min\{m,n\}$, which is a direct consequence of Theorem 1. Finally, note that if $|\mathcal{F}| = n$, $\mathcal{W} = \emptyset$ so that (10) implies $\pi^{\mathrm{LP}} \leq n\pi^{\mathrm{OPT}}$. □

We have no example that shows tightness of this bound. The worst example we found so far has a ratio $\pi^{\mathrm{LP}}/\pi^{\mathrm{OPT}} = 4$.

2.2 The LP round-down heuristic

This section investigates a heuristic which amounts to rounding down the optimal solution of the LP relaxation of (SSNP). The worst-case performance ratio analysis is related to the analysis for the greedy heuristic of the knapsack problem [6, Subsection 2.4]. In the deterministic case the knapsack LP solution can be found in $O(n)$ time by a median finding algorithm using the price per unit criterion [2]. Here a similar approach is not known.

The previous section showed that any optimal LP solution of an m-scenario problem will have at most m subservices for which the z_j^{LP}-values are fractional. All remaining z are 0 or 1. This motivates the following LP round-down heuristic, which we call LPR: Install each subservice j for which $z_j^{\mathrm{LP}} = 1$ and no others, that is install all $j \in \mathcal{W}$. Afterwards the remaining capacity is allocated to serve demand of the installed subservices in a greedy manner, starting with the subservices with the highest q_j. Assume for simplicity that the subservices are sorted by non-increasing q_j. Then there will be a critical subservice j_k in each scenario k for which $x_{jk}^{\mathrm{LP}} = \delta_{jk} z_j^{\mathrm{LP}}, j < j_k$ and $x_{jk}^{\mathrm{LP}} = 0, j > j_k$. Let $(z^{\mathrm{LPR}}, x^{\mathrm{LPR}})$ be the heuristic solution and π^{LPR} the solution value.

Proposition 1 *A lower bound for the LP round-down heuristic (LPR) value is to allocate only the amount indicated by the LP solution to each installed subservice*

$$\pi^{\mathrm{LPR}} \geq \sum_{j\in\mathcal{W}}\sum_{k=1}^{m} p_k q_j x_{jk}^{\mathrm{LP}}.$$

Proof In the LPR heuristic all the space allocated to subservices $j \in \mathcal{F}$ in the LP is free, as these subservices are not installed. This free capacity can potentially be used to meet demand for subservices $j \in \mathcal{W}$. So $x_{jk}^{\mathrm{LPR}} \geq x_{jk}^{\mathrm{LP}}, \forall j \in \mathcal{W},\ \forall k$. □

LPR can for some instances of the problem be arbitrarily bad, because a better solution with an arbitrarily higher value may be to install one of the fractional subservices. The heuristic is now modified into a heuristic that we call $\lfloor \mathrm{LP} \rfloor$ to avoid this problem. If the value of installing the best of the services $j \in \mathcal{F}$ is higher than the value of installing all subservices $j \in \mathcal{W}$ then do that instead. Let $\pi^{\lfloor \mathrm{LP} \rfloor}$ be the optimal value of this heuristic. Then

$$\pi^{\lfloor \mathrm{LP} \rfloor} = \max\{\pi^{\mathrm{LPR}}, \max_{j\in\mathcal{F}} q_j E_k[\delta_{jk}]\}.$$

Theorem 2 *The modified LP round-down heuristic* $\lfloor \mathrm{LP} \rfloor$ *has a worst case performance ratio of*

$$\pi^{\mathrm{OPT}}/\pi^{\lfloor \mathrm{LP} \rfloor} \leq \min\{m+1, n\},$$

and this ratio is tight.

Proof

$$\begin{aligned}
\pi^{\mathrm{OPT}} &\leq \pi^{\mathrm{LP}} = \sum_{j\in\mathcal{W}}\sum_{k=1}^{m} p_k q_j x_{jk}^{\mathrm{LP}} + \sum_{j\in\mathcal{F}}\sum_{k=1}^{m} p_k q_j x_{jk}^{\mathrm{LP}}, \\
&\leq \pi^{\mathrm{LPR}} + \sum_{j\in\mathcal{F}} q_j E_k[\delta_{jk}], \\
&\leq \pi^{\lfloor \mathrm{LP}\rfloor} + |\mathcal{F}|\pi^{\lfloor \mathrm{LP}\rfloor} \leq (\min\{m,n\}+1)\pi^{\lfloor \mathrm{LP}\rfloor}.
\end{aligned}$$

Again, the case where $|\mathcal{F}| = n$ tightens the bound to $\min\{m+1, n\}\pi^{\lfloor \mathrm{LP}\rfloor}$.

A tight example is given here. The problem has $v+1$ subservices and v scenarios, $v \geq 2$. Let $q_j = v - \epsilon$, $r_j = \epsilon$, $j = 1, \ldots, v$, $q_{v+1} = v/\epsilon$ and $r_{v+1} = 0$ where $0 < \epsilon < 1$. The node size is $s = 1 + v\epsilon$ and all scenarios are equally likely. Demand is defined to be constant over all scenarios for subservice $v+1$, $\delta_{v+1k} = \epsilon/v$, $k = 1, \ldots, v$. For all other subservices $j = 1, \ldots, v$ demand is present only in scenario j, with $\delta_{jj} = 1$ and $\delta_{jk} = 0$ when $j \neq k$.

The optimal solution is to install all subservices. Demand for subservice $v+1$ is always met completely, while in scenarios $j = 1, \ldots, v$ the optimal solution has $x_{jj}^{\mathrm{OPT}} = 1 - \frac{\epsilon}{v}$. The profit from this is $\pi^{\mathrm{OPT}} = 1 + v - 2\epsilon + \frac{\epsilon^2}{v}$.

The optimal LP solution, $(z_{jk}^{\mathrm{LP}}, x_{jk}^{\mathrm{LP}})$ is $z_{v+1}^{\mathrm{LP}} = 1$, $x_{v+1k}^{\mathrm{LP}} = \epsilon/v$, $\forall k$, $x_{jj}^{\mathrm{LP}} = z_j^{\mathrm{LP}} = \frac{1+v\epsilon-\frac{\epsilon}{v}}{1+v\epsilon}$, $\forall j$, $x_{jk}^{\mathrm{LP}} = 0$, $k \neq j$. This solution has $m = v$ fractional z-values and no fractional x-values. The modified LP round-down solution value $\pi^{\lfloor \mathrm{LP}\rfloor}$ is the maximum of installing one of the fractional subservices or subservice $v+1$: $\pi^{\lfloor \mathrm{LP}\rfloor} = \max\{1, \frac{1}{v}(v-\epsilon)\} = 1$. As ϵ gets arbitrarily small,

$$\pi^{\mathrm{OPT}}/\pi^{\lfloor \mathrm{LP}\rfloor} = 1 + v - 2\epsilon + \frac{\epsilon^2}{v}$$

gets arbitrarily close to $v+1$ where v is the number of scenarios and $v+1$ is the number of subservices. □

Notice that the given bound on the performance ratio holds for any possible discrete distribution defined in terms of scenarios. It is increasing in the number of scenarios and if the number of scenarios is greater than the number of subservices the bound is even linear in the number of subservices, which, in general, is not a very favourable situation. For the considered application with a limited number of scenarios, it may still be useful. Yet, it would be better to have a constant performance ratio. Next, the bound on the LP ratio is tightened for a class of problem instances and a heuristic with a constant bound for this problem class is defined.

3 A constant bound

The results from Section 2 depend on the demand probability distribution in a fundamental way. It is directly dependent on the number of realizations the random variables may take. This section shows that for a class of service provision problems it is possible to find a worst-case ratio that is independent of the discrete demand distribution.

The class of problems examined are those for which it is feasible (but not necessarily optimal) to install all subservices concurrently. That is, the sum of the installation requirements is less than the node capacity. This assumption is reasonable in many cases for the problem setting. In order to facilitate the exposition, the node capacity is scaled to 1; $s = 1$. In this setting the class of problems has

$$\sum_{j=1}^{n} r_j \leq 1. \tag{11}$$

In Section 2 the bound was obtained by considering each fractional subservice individually. In this section the bound is improved by considering sets of these subservices together. The important aspect here is the trade-off between the number of sets and the capacity used by the installation requirements of the subservices in each set.

3.1 The LP ratio

Let $(z^{\mathrm{LP}}, x^{\mathrm{LP}})$ be a basic optimal LP relaxation solution. Let ℓ be the number of fractional z_j^{LP} and assume that ℓ_w of these subservices have $r_j \leq w$ for some $0 < w < 1$ to be chosen later. These subservices will be installed in groups while those with $r_j > w$ will be installed separately as before. Again, let $\mathcal{W}$ be the set of subservices with $z_j^{\mathrm{LP}} = 1$. Without loss of generality let $0 < z_j^{\mathrm{LP}} < 1$ and $r_j \leq w$ for $j = 1, \ldots, \ell_w$ and $0 < z_j^{\mathrm{LP}} < 1$ and $r_j > w$ for $j = \ell_w + 1, \ldots, \ell$. Write the optimal LP value as

$$\pi^{\mathrm{LP}} = \pi_0^{\mathrm{LP}} + \pi_1^{\mathrm{LP}} + \pi_2^{\mathrm{LP}} \tag{12}$$

where

$$\pi_0^{\mathrm{LP}} = \sum_{j \in \mathcal{W}} q_j E_k[x_{jk}^{\mathrm{LP}}],$$

$$\pi_1^{\mathrm{LP}} = \sum_{j=1}^{\ell_w} q_j E_k[x_{jk}^{\mathrm{LP}}],$$

and

$$\pi_2^{\mathrm{LP}} = \sum_{j=\ell_w+1}^{\ell} q_j E_k[x_{jk}^{\mathrm{LP}}].$$

Feasible solutions generated from the LP solution will be used to bound parts of (12). From Section 2

$$\pi^{\mathrm{OPT}} \geq \pi^{\mathrm{LPR}} \geq \pi_0^{\mathrm{LP}}. \tag{13}$$

Next π_1^{LP} is bound. First, define $\sum_{j=1}^{\ell} r_j z_j^{\mathrm{LP}} = A$ and note that $\sum_{j=1}^{\ell_w} x_{jk}^{\mathrm{LP}} \leq 1 - A$ for each $k = 1, \ldots, m$. Integer feasible solutions are generated for which the capacity used by the r_j's of the installed subservices is close to some constant β. First partition the set $\{1, \ldots, \ell_w\}$ into I subsets, $\{S_i\}_{i=1}^{I}$, where

$$\sum_{j \in S_i} r_j \leq \beta + w \qquad i = 1, \ldots, I$$

and

$$\sum_{j \in S_i} r_j \geq \beta \qquad i = 1, \ldots, I-1. \tag{14}$$

Notice that the last bound is not required for S_I. The LP relaxation had at most $1 - A$ units of capacity available for the x variables. Installing only the subservices in one of the sets S_i will leave at least $1 - \beta - w$ units of capacity available. The x-variable values from the LP relaxation solution corresponding to subservices in S_i may be scaled down, if necessary, to use a total of no more than $1 - \beta - w$ units of capacity in each scenario.

For each $i = 1, \ldots, I$ generate the integer feasible solution (z^{H_i}, x^{H_i}) for which $z_j^{H_i} = 1$ for $j \in S_i$, and $z_j^{H_i} = 0$ for all $j \notin S_i$. Set $x_{jk}^{H_i} = \gamma x_{jk}^{\mathrm{LP}}$ for $j \in S_i$, $k = 1, \ldots, n$ and $x_{jk}^{H_i} = 0$ for all other jk with

$$\gamma = \begin{cases} \dfrac{1 - \beta - w}{1 - A} & \text{if } \beta + w \geq A, \\ 1 & \text{otherwise.} \end{cases} \tag{15}$$

Now the objective value of the solution (z^{H_i}, x^{H_i}) is

$$\pi^{H_i} = \sum_{j \in S_i} q_j E_k[x_{jk}^{H_i}] = \gamma \sum_{j \in S_i} q_j E_k[x_{jk}^{\mathrm{LP}}]$$

and it follows that

$$\pi_1^{\mathrm{LP}} = \sum_{i=1}^{I} \sum_{j \in S_i} q_j E_k[x_{jk}^{\mathrm{LP}}] = \frac{1}{\gamma} \sum_{i=1}^{I} \pi^{H_i} \leq \frac{I}{\gamma} \pi^{\mathrm{OPT}} \tag{16}$$

Observe, that the size of I may be bound using (14) with the following construction.

$$1 \geq \sum_{j=1}^{n} r_j \geq \sum_{j=1}^{\ell_w} r_j = \sum_{i=1}^{I} \sum_{j \in S_i} r_j \geq (I-1)\beta. \tag{17}$$

This means that $I \leq 1 + 1/\beta$ leading to the following bound

$$\pi_1^{\mathrm{LP}} \leq \frac{\beta + 1}{\beta \gamma} \pi^{\mathrm{OPT}}. \tag{18}$$

where γ is given by (15).

For bounding π_2^{LP} consider installing each subservice $j = \ell_w + 1, \ldots, \ell$ individually. Note that from the definition of A, and since $r_j \geq w$ for $j = \ell_w + 1, \ldots, \ell$,

$$A = \sum_{j=1}^{\ell} r_j z_j^{\mathrm{LP}} \geq \sum_{j=\ell_w+1}^{\ell} r_j z_j^{\mathrm{LP}} \geq w \sum_{j=\ell_w+1}^{\ell} z_j^{\mathrm{LP}}$$

Thus,

$$\sum_{j=\ell_w+1}^{\ell} z_j^{\mathrm{LP}} \leq \frac{A}{w}.$$

The solution obtained by installing just subservice j from among subservices $\ell_w + 1, \ldots, \ell$ has an objective value of no more than $q_j E_k[\delta_{jk}]$.

From the demand constraint it follows that $E_k[x_{jk}^{\mathrm{LP}}] \leq E_k[\delta_{jk}]z_j^{\mathrm{LP}}$. By Assumption 1, this leads to the following bound.

$$\begin{aligned}
\pi_2^{\mathrm{LP}} &= \sum_{j=\ell_w+1}^{\ell} q_j E_k[x_{jk}^{\mathrm{LP}}] \\
&\leq \sum_{j=\ell_w+1}^{\ell} q_j E_k[\delta_{jk}] z_j^{\mathrm{LP}} \\
&\leq \pi^{\mathrm{OPT}} \sum_{j=\ell_w+1}^{\ell} z_j^{\mathrm{LP}} \leq \frac{A}{w}\pi^{\mathrm{OPT}} \qquad (19)
\end{aligned}$$

Combining (13), (18), and (19) gives

$$\pi^{\mathrm{LP}} \leq \left(1 + \frac{\beta+1}{\beta\gamma} + \frac{A}{w}\right)\pi^{\mathrm{OPT}} \qquad (20)$$

where γ is given by (15) and $w, \beta \in (0,1)$ may be chosen with $w + \beta < 1$. This leads to the following theorem.

Theorem 3 *Under the assumption that* $\sum_{j=1}^{n} r_j \leq 1$

$$\pi^{\mathrm{LP}} \leq (5 + 2\sqrt{3})\pi^{\mathrm{OPT}}.$$

Proof The choice of w and β is based on the value of A in (20), which depends on the LP solution. When $A < \frac{1}{2}$ take $w = 1 - \frac{1}{2}\sqrt{3}$ and $\beta = -\frac{1}{2} + \frac{1}{2}\sqrt{3}$ and when $A \geq \frac{1}{2}$ take $w = \beta = \frac{1}{2}A$. For both cases $w + \beta \geq A$ so that $\gamma = \frac{1-\beta-w}{1-A}$. For the former case the bound (20) leads to

$$\begin{aligned}
\pi^{\mathrm{LP}} &\leq \left(1 + \frac{2(1+\sqrt{3})(1-A)}{-1+\sqrt{3}} + \frac{A}{1-\frac{1}{2}\sqrt{3}}\right)\pi^{\mathrm{OPT}} \\
&= \left(1 + (1+\sqrt{3})^2(1-A) + 4(1+\frac{1}{2}\sqrt{3})A\right)\pi^{\mathrm{OPT}} = (5+2\sqrt{3})\pi^{\mathrm{OPT}}
\end{aligned}$$

while in the latter case (20) leads to the bound

$$\pi^{\mathrm{LP}} \leq \left(4 + \frac{2}{A}\right)\pi^{\mathrm{OPT}} \leq 8\pi^{\mathrm{OPT}} \leq (5 + 2\sqrt{3})\pi^{\mathrm{OPT}}.$$

□

Notice that we stated the theorem for the node capacity s being equal to 1. However, it is easy to see that the theorem holds for any value of s since scaling the problem so that $s = 1$ leaves the ratio unchanged.

We can show that in case $A \leq \frac{1}{2}$ there is no better choice of w and β in this analysis. In case $A > \frac{1}{2}$ a better choice of w and β seems possible though, so that in that case the analysis could lead to a slightly better constant bound.

3.2 A round and partition heuristic with constant worst-case ratio

Based on the previous LP bound a round and partition heuristic (RP) is developed with a worst case performance ratio bounded above by $5 + 2\sqrt{3}$.

Consider the class of heuristics that, given $S \subseteq \{1, \ldots, m\}$, produce the solution (z^S, x^S) with objective value π^S, by setting $z_j^S = 1$ if $j \in S$ or $z_j^S = 0$ if $j \notin S$ and choosing x^S to maximize the LP created by fixing z to z^S in (SSNP). Guided by the previous section, we will generate many such solutions by partitioning the set of services.

The two constants w and β of the previous subsection are chosen as in Theorem 3. That is, when $A < \frac{1}{2}$ choose $w = 1 - \frac{1}{2}\sqrt{3}$ and $\beta = -\frac{1}{2} + \frac{1}{2}\sqrt{3}$ and when $A \geq \frac{1}{2}$ choose $w = \beta = \frac{1}{2}A$. Regarding the remark following Theorem 3 in the previous subsection in case $A > \frac{1}{2}$ also here better choices of w and β seem possible.

Let $(z^{\mathrm{LP}}, x^{\mathrm{LP}})$ be a basic optimal LP relaxation solution with the optimal solution value given by (12). From this solution we generate a partition $\{W, Z, B, T_1, \ldots, T_K\}$ for some K of $\{1, \ldots, m\}$.

- $\mathcal{W} = \{j | z_j^{LP} = 1\}$;
- $\mathcal{Z} = \{j | z_j^{LP} = 0\}$;
- $\mathcal{B} = \{j | 0 < z_j^{LP} < 1, r_j > w\}$.

The remaining subservices with $z_j^{LP} > 0$ and $r_j \leq w$ are partitioned into the sets $\mathcal{T}_1, \ldots, \mathcal{T}_K$ in the following way. Consider these subservices in arbitrary order. Start by filling the set $\mathcal{T}_1$ with the first subservices until addition of the next subservice will raise the sum of the installation requirements above $w + \beta$. That subservice will be the first one to go into the set $\mathcal{T}_2$. Continue in the same way filling the set $\mathcal{T}_2$ and so on until the last set T_K is constituted by the last few items. Thus, the sets $\mathcal{T}_1, \ldots, \mathcal{T}_K$ have the properties

- $\sum_{j \in T_i} r_j \in [\beta, \beta + w]$ for $i = 1, \ldots, K-1$
- $\sum_{j \in T_K} r_j \leq \beta + w$

The partition generation takes $O(m)$ time once the LP solution is known.

The round and partition heuristic then chooses a solution (x^S, z^S) where S is: $\mathcal{W}$, one of the sets $\mathcal{T}_i$, or a single element of $\mathcal{B}$. That is, the round and partition heuristic solution, $(z^{\mathrm{RP}}, x^{\mathrm{RP}})$, is given by

$$(z^{\mathrm{RP}}, x^{\mathrm{RP}}) = \operatorname{argmax} \left\{ \pi^S \mid S \in \{W, T_1, \ldots, T_K\} \cup \bigcup_{j \in B} \left\{ \{j\} \right\} \right\}.$$

Let π^{RP} be the solution value of the round and partition heuristic.

Theorem 4 *The round and partition heuristic has a worst case performance ratio of*

$$\pi^{\mathrm{OPT}} / \pi^{\mathrm{RP}} \leq (5 + 2\sqrt{3}).$$

Proof This follows almost immediately from the proof of the bound for the LP-relaxation in Section 3.1 taking the $\{S_i\}_{i=1}^{I}$ as $\{T_i\}_{i=1}^{K}$. The w and β values used above are the same as in the proof.

Notice that $\pi^W = \pi^{\mathrm{LPR}}$ and for any $j \in B$ $\pi^{\{j\}} = q_j E_k[\delta_{jk}]$. Also, for each $i \in \{1, \ldots, K\}$ $\pi^{T_i} = \pi^{H_i}$. With this, from the definition of the heuristic,

$$\pi^{\mathrm{RP}} \geq \pi_0^{\mathrm{LP}}, \quad \pi^{\mathrm{RP}} \geq \pi^{H_i} \ \forall i = 1, \ldots, K \quad \text{and} \quad \pi^{\mathrm{RP}} \geq q_j E_k[\delta_{jk}], \forall j \in B.$$

From this π^{OPT} may be replaced by π^{RP} in (16) and (19). Following this through to the proof of the LP bound in Theorem 3 gives

$$\pi^{\mathrm{OPT}} \leq \pi^{\mathrm{LP}} \leq (5 + 2\sqrt{3})\pi^{\mathrm{RP}}.$$

□

It should be remarked that the derived bound might not be tight. The tightest bound discovered by the authors, from any instance, has a performance ratio of 2.

4 Computational complexity

This section gives evidence that the above results are interesting in the sense that one cannot hope to arrive at the optimal solution of (SSNP) in polynomial time. As indicated in the introduction the deterministic counterpart of the problem admits a fully polynomial approximation scheme for its solution. Here we show that this is unlikely to be achievable for (SSNP) by proving that it is strongly NP-hard.

Theorem 5 *The stochastic single node service provision problem is strongly NP-hard.*

Proof The natural recognition version of this problem obtained by introducing a number and asking if there is a feasible solution giving profit at least that number is obviously in NP, since the representation of the probabilistic input in scenarios allows the formulation of a deterministic equivalent mixed-integer programming problem. To see that the recognition version is strongly NP-Complete consider a reduction from the well-known strongly NP-Complete vertex cover problem (see [4]):

> Given a graph $G = (V, E)$ with $|V|$ vertices and $|E|$ edges and a constant K, does there exist a subset V' of the vertices, such that each edge in E is incident to at least one vertex in V', and such that $|V'| \leq K$?

For every vertex $j \in V$ introduce a subservice j with installation requirement $\alpha = \frac{1}{K|E|}$. For every edge introduce a scenario with demand 1 for the two subservices incident to it and demand 0 for all other subservices. Let $q_j = |E| \ \forall j \in V$, and let all scenarios have a probability $\frac{1}{|E|}$ of occurring. Then the expected profit from meeting one unit of demand in a single scenario is 1. Take $K\alpha + 1$ as capacity of the node in (SSNP). The question is whether there is a solution to this instance of (SSNP) with total expected profit at least $|E|$.

This transformation is obviously polynomial. In case there exists a vertex cover of size at most K then there is a service provision with total expected profit at least $|E|$. Install the subservices corresponding to the vertices in the vertex cover. Then for each scenario (edge) at least one of the subservices with demand 1 is installed. The total capacity used by the installation of the subservices is at most $K\alpha$ leaving at least capacity 1 to fill with the demands for each scenario.

The other direction is a bit more complicated. Suppose there does not exist a vertex cover of size K or less. Then installing all subservices corresponding to a vertex cover would use node capacity strictly greater than $K\alpha$ leaving strictly less than 1 for meeting demand in each of the $|E|$ scenarios, making a total expected profit of at least $|E|$ unattainable. Installing any set of subservices of size $L < K$ would leave $(K-L)\alpha + 1$ node capacity for meeting demand in each scenario. However, at least one edge will remain uncovered, implying that there is at least one scenario in which both subservices with a positive demand are not installed. With at most $|E| - 1$ scenarios the expected profit will be at most $(|E|-1)((K-L)\alpha+1) \leq (|E|-1)(K\alpha+1) = (|E|-1)(\frac{1}{|E|}+1) < |E|$. □

In case the number of scenarios is fixed a dynamic programming algorithm shows that the problem can be solved in pseudo-polynomial time. We argued in the introduction that this problem is not only of academic interest, but reflects a plausible real-world situation. For this it is assumed that all problem parameters are integers.

Theorem 6 *The stochastic single node service provision problem with a fixed number of scenarios can be solved in pseudo-polynomial time.*

Proof Consider the following DP that has the subservices as its stages. A state, $S \in \mathbb{Z}_+^m$, gives the capacity used in each scenario. Define $f_j(S)$ as the maximum profit that can be achieved from scenario capacities $S = (S_1, ..., S_m)$ using the subservices $1, ..., j$. Each S_k may take a value between 0 and s so there are at most $(s+1)^m$ states per stage. There are two types of transitions in every stage, either the subservice is not installed, or it is installed and some demand is met. There are fewer than $s+1$ possible choices concerning the demand to meet in each scenario, and overall there are then fewer than $(s+1)^m$ different feasible decisions in a state. The initial settings are

$$f_0(S) = \begin{cases} 0 & \text{if } 0 \leq S_i \leq s, \quad \forall i = 1, \ldots, m \\ -\infty & \text{otherwise.} \end{cases}$$

The recurrence is given by

$$f_j(S_1, \ldots, S_m) = \max_{0 \leq x_k \leq \delta_{jk}} \Big\{ f_{j-1}(S_1 - r_j - x_1, \ldots, S_m - r_j - x_m) + q_j \sum_{k=1}^{m} p_k x_k, \; f_{j-1}(S_1, \ldots, S_m) \Big\}.$$

From each state there are at most $(s+1)^m + 1$ possible transitions, at each stage there are at most $(s+1)^m$ states and there are n stages. The running time of the DP is therefore at most $O(ns^{2m})$, which implies the theorem. □

Thus, the conclusion is that the problem with a fixed number of scenarios is not strongly NP-hard. This suggests also the existence of a polynomial approximation scheme for the problem, a nice subject for future investigations. That this subclass of problems is still NP-hard is implied by the NP-hardness of the deterministic counterpart of the problem which has been proved in [2].

5 Conclusions

This paper considered a service provision problem on a distributed processing telecommunication network, under uncertain demand for the services. It was shown that the natural stochastic integer programming model is strongly NP-hard. It is worthwhile to stress this as its deterministic counterpart having the same number of binary decision variables is weakly NP-hard. Thus, the complexity of the problem increases by introducing stochasticity, even if it only means adding continuous decision variables for each scenario of the problem. This suggests that algorithms for general stocastic integer programming are unlikely to be "scalable".

Because of the strong NP-hardness, approximation algorithms were studied for this problem. A first algorithm based on the LP relaxation of the deterministic equivalent of the stochastic problem has worst-case performance ratio equal to the minimum of the number of services and the number of scenarios that describe the stochastic demand plus one. The second algorithm has a constant worst-case performance ratio for a more restricted class of problems. The assumption defining this subclass is, however, satisfied for many reasonable practical problem situations.

Moreover, the variable bound on the performance ratio of the first algorithm is not as bad as it may seem at first sight because (as indicated in the introduction) the number of scenarios may actually be small in our telecommunication application. In a situation with a small number of scenarios one might alternatively think of using the dynamic programming formulation of Section 4. However, it should be noted that if precision is required and the resource capacity and the resource requirements are large then the pseudo-polynomial nature of the method leads to excessive computation times.

References

[1] M. De Prycker. *Asynchronous Transfer Mode, Solution for Broadband ISDN.* Prentice Hall, New Jersey, 1995.

[2] S. Dye, L. Stougie, and A. Tomasgard. Approximation algorithms and relaxations for a service provision problem on a telecommunication network. Working paper #2-98, Department of industrial economics and technology management, Norwegian university of science and technology, N-7034 Trondheim, Norway, 1998., 1998.

[3] S. Dye, A. Tomasgard, and S.W. Wallace. Feasibility in transportation networks with supply eating arcs. *Networks*, 31:165–176, 1998.

[4] M.R Garey and D.S. Johnson. *Computers and Intractability, a guide to the theory of NP-completeness.* Freeman, New York, 1979.

[5] P. Kall and S.W. Wallace. *Stochastic Programming.* Wiley, Chichester, 1994.

[6] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations.* John Wiley & sons, 1990.

[7] G.L. Nemhauser and L.A. Wolsey. *Integer and Combinatorial Optimization.* Wiley-Interscience series in discrete mathematics and optimization. Wiley, New York, 1988.

[8] R. Onvural. *Asynchronous Transfer Mode Networks: Performance Issues.* Artech House, Boston, 1994.

[9] A. Tomasgard, S. Dye, S.W. Wallace, J.A. Audestad, L. Stougie, and M.H. van der Vlerk. Stochastic optimization models for distributed communication networks. Working paper #3-97, Department of industrial economics and technology management, Norwegian university of science and technology,N- 7034 Trondheim, Norway, 1997.

[10] A. Tomasgard, J.A. Audestad, S. Dye, L. Stougie, M.H. van der Vlerk, and S.W. Wallace. Modelling aspects of distributed processing in telecommunication networks. *Annals of Operations Research*, 82:161–184, 1998.