Managing Traffic Demand Uncertainty in Replica Server Placement with Robust Optimization

Kin-Hon Ho, Stylianos Georgoulas, Mina Amin, and George Pavlou

Centre for Communication Systems Research, University of Surrey, GU2 7XH, UK
{K.Ho, S.Georgoulas, M.Amin, G.Pavlou}@surrey.ac.uk

Abstract. The replica server placement problem determines the optimal location where replicated servers should be placed in content distribution networks, in order to optimize network performance. The estimated traffic demand is fundamental input to this problem and its accuracy is essential for the target performance to be achieved. However, deriving accurate traffic demands is far from trivial and uncertainty makes the target performance hard to predict. We argue that it is often inappropriate to optimize the performance for only a particular set of traffic demands that is assumed accurate. In this paper, we propose a scenario-based robust optimization approach to address the replica server placement problem under traffic demand uncertainty. The objective is to minimize the total distribution cost across a variety of traffic demand scenarios while minimizing the performance deviation from the optimal solution. Empirical results demonstrate that robust optimization for replica server placement can achieve good performance under all the traffic demand scenarios while non-robust approaches perform significantly worse. This approach allows content distribution providers to provision better and predictable quality of service for their customers by reducing the impact of inaccuracy in traffic demand estimation on the replica server placement optimization.

1 Introduction

Content Distribution Networks (CDNs) aim to efficiently deliver web content from servers to users through the Internet. In order to achieve this goal, CDN providers replicate their server infrastructure in multiple locations. The replication technique brings two major advantages to CDNs: first, it minimizes content download time since the replicated servers can be placed quite close to the requesting users; and, second, it allows the CDN providers to operate seamlessly if one of its servers is not available. For simplicity, we call each replicated server a *replica* in this paper.

To efficiently deliver web contents, a CDN provider needs to decide where to place replicas and how many replicas are required, so as to optimize network performance [1,2] and to support Quality of Service (QoS) guarantees [3,4]. This is known as the Replica Server Placement (RSP) problem. Achieving optimal and predictable RSP design is extremely important for the success of the CDN business as users will abandon a web site that fails to provide content in an acceptable response time[1]. In the

[1] According to Zona Research [19], about $4.35 billion may be lost in online business revenues in 1999 due to unacceptably slow response times.

F. Boavida et al. (Eds.): NETWORKING 2006, LNCS 3976, pp. 727 – 739, 2006.

literature, the RSP problem has been formulated as the *minimum P-median problem*, taking as input a set of estimated traffic demands. In theory, if the traffic demands are perfectly known, then an optimal and predictable performance for the RSP can be obtained. Unfortunately, deriving accurate traffic demands is far from trivial since Internet traffic patterns change over time as a result of unpredictable user behavior in traffic request. In addition, perfect (noiseless) flow measurements are rarely available on all links and egress/ingress points of the network [5]. Hence, traffic demands are usually derived with a degree of uncertainty whose consequence is to prevent conventional RSP optimization from producing optimal and predictable performance; this may subsequently lead to potential loss in business revenues. In this paper, we argue that it is insufficient to optimize CDNs performance for only a particular set of traffic demands given relevant inaccuracy. Instead, we need to fundamentally rethink the way in which we design CDNs for coping with uncertainty so as to avoid 'risky' solutions characterized by unsatisfactorily high traffic demand uncertainty in order to sustain, at least, stable business revenues. To the best of our knowledge, this important issue has not been investigated in the literature.

Rather than assuming accurate traffic demand estimation, which is not possible as explained, we propose an approach based on the principles of ***Scenario-based Robust Optimization*** (***SRO)*** [6,7]. When applied to the RSP, SRO structures uncertainty by using a set of potential traffic demand *scenarios*, each exhibits structural difference in traffic volume distribution. The objective of this robust RSP is thus to optimize performance objectives across a variety of such scenarios. We formulate the robust RSP problem as an integer programming problem and solve it by the MINLP solver. Simulation results demonstrate that the robust RSP approach achieves significantly better performance than non-robust optimization approaches under traffic demand uncertainty. In addition, the robust RSP approach guarantees the resulting performance to be within a specified envelope from the optimal solution.

The rest of this paper is organized as follows. In Section 2, we review the deterministic RSP problem formulation. We then present a robust version of RSP in Section 3. Section 4 presents three alternative strategies to tackle the robust RSP problem, which are used for performance comparison. In Sections 5 and 6, we present our evaluation methodology and simulation results. Finally, we conclude the paper in Section 7.

2 Deterministic Replica Server Placement Problem

Table 1 shows the notation used throughout this paper. The deterministic (i.e. non-robust) version of RSP [1] can be summarized as follows. A CDN network is modeled as a graph $G=(N,E)$ where N and E are network nodes and links. Given a set of user nodes $I \subseteq N$ and a set of potential server nodes $J \subseteq N$, select P out of J to be server nodes[2] and assign each user traffic demand to the closest server. A distribution cost $d_i c_{i,j}$ is incurred if traffic demand d_i is assigned to server node j, where $c_{i,j}$ is a general cost that may represent hop count, IGP cost, delay or any other performance metric.

[2] In line with [1,2], we assume a full replication for content distribution, i.e. each server has large storage capacities to hold the whole contents for serving any user request.

Table 1. Notation

NOTATION	DESCRIPTION
I	A set of user nodes, indexed by i
J	A set of potential server nodes, indexed by j
S	A set of traffic demand scenarios, indexed by s. This includes the base and the developed traffic demands
d_i	Traffic demand from user node i
$d_{s,i}$	Traffic demand of user node i under scenario s
$c_{i,j}$	Cost to transport one unit from server node j to user node i
X_j	A variable indicating whether node j is selected as server node
$Y_{i,j}$	A variable indicating whether traffic demand of user node i is assigned to server node j
z_s^*	The optimal total distribution cost under scenario s if input data is perfectly known for that scenario

Since the most concerned performance metric for RSP is content download time [8], we assume $c_{i,j}$, to be the delay between i and j over the shortest path in terms of hop count. The goal of RSP is thus to select P nodes as server nodes so as to minimize thetotal distribution cost[3] over all traffic demands. In [1], the RSP problem has been proven NP-hard by mapping it to the *uncapacitated minimum P-median problem*. The problem formulation of the deterministic RSP can be summarized as follows:

$$Minimize \sum_{i\in I}\sum_{j\in J} d_i c_{i,j} Y_{i,j} \tag{1}$$

subject to the following constraints:

$$\forall i \in I : \sum_{j\in J} Y_{i,j} = 1 \tag{2}$$

$$\forall i \in I, j \in J : Y_{i,j} \le X_j \tag{3}$$

$$\sum_{j\in J} X_j = P \tag{4}$$

$$\forall i \in I, j \in J : X_i, Y_{i,j} \in \{0,1\} \tag{5}$$

Objective function (1) minimizes the total distribution cost over all traffic demands. Constraint (2) ensures that each traffic demand is assigned to one server. Constraint (3) ensures that, whenever traffic demand d_i is assigned to node $j \in J$, then j must have been selected as server node. Constraint (4) states that P out of N servers are to be selected. Constraint (5) is the standard integrality constraint.

3 Robust Replica Server Placement Problem

In this section, we present a Scenario-based Robust Optimization (SRO) approach for RSP optimization to manage traffic demand uncertainty. SRO is a comprehensive

[3] Since the distribution cost takes into account the link delay, minimizing the total distribution cost over all traffic demands is effectively equivalent to minimizing the content download time for these traffic demands.

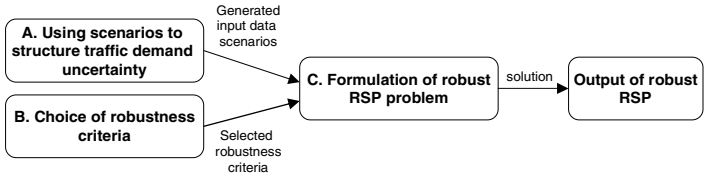


Fig. 1. Scenario-based robust optimization framework for the replica server placement problem

mathematical programming framework for robust decision making. The SRO framework applied to the RSP problem consists of three elements, as depicted on Figure 1.

A. Using Scenarios to Structure Traffic Demand Uncertainty

Decision makers may develop discrete scenarios that provide visions of alternative possible futures, and then use these them to structure their uncertain input data. Thus, scenarios are devised for representing the decision maker's perceptions about alternative environments in which their decisions might be applied, in the most appropriate manner based on internal knowledge and experience.

When applied to the RSP, SRO models uncertainty as a set of potential traffic demand scenarios. This set of traffic demand scenarios cover at least different critical views of possible traffic characteristics instant, e.g. morning, afternoon and evening. In fact, since the sources of errors or fluctuations in the traffic demands are well understood, their magnitude can be estimated within some known accuracy [5,21].

B. Choice of Robustness Criteria

Recall that the aim of the scenario-based robust optimization is to produce decisions that will have a reasonable objective value under any potential input data scenario. Different criteria can be used to select among robust decisions. We apply two criteria [6] to our robust RSP optimization, which make it more suitable for CDN providers to consider from a practical point of view.

The first criterion is ***minimax***, which aims at getting the best out of the worst possible conditions. This criterion is chosen based on a general observation that the decision makers are (in many cases) risk-averse, meaning that the RSP solution CDN providers want is neither the "optimal" for a particular traffic demand scenario nor the "worst" for any scenario but one that performs reasonably well across all the scenarios. Such risk-averse behavior may also be observed from capacity overprovisioning employed by top-tier Internet service providers as a means to provide good service to all IP traffic in their backbone networks [18]. Hence, CDN providers may want to optimize the worst-case network performance in order to prevent severe unpredicted performance degradation and the need for future expensive network capacity upgrading. The minimax criterion can thus be expressed by minimizing the worst-case total distribution cost across the set of traffic demand scenarios, i.e.

$$Minimize \ \underset{\forall s \in S}{Max} F(s) \tag{6}$$

where $F(s)$ is the resulting total distribution cost under traffic demand scenario s.

Although the minimax criterion can produce reasonably good performance across all the traffic demand scenarios, it may lead to RSP solutions that are overly conservative

Table 2. Example of total distribution cost under four different traffic demand scenarios

Scenario / Solution	s_1	s_2	s_3	s_4
x_1	89	90	87	**93**
x_2	79	81	75	**95**
optimal	74	76	70	82

or pessimistic, thereby accepting unnecessary high costs in non-worst-case scenarios. We illustrate this conservative effect by the example in Table 2.

Two solutions, x_1 and x_2, produce different total distribution costs for four traffic demand scenarios (s_1, s_2, s_3, s_4). The solution named "optimal" represents the optimal total distribution cost for each scenario if the traffic demands of that scenario are perfectly known. If only the minimax criterion of equation (6) is considered, x_1 is the best solution since it has lower worst-case total distribution cost than that of x_2 (93 vs. 95). However, x_1 has higher cost than x_2 under scenarios s_1, s_2 and s_3, and their costs deviate highly from the optimal ones. One may observe that if s_1, s_2 or s_3 occurs, x_1 will no longer be the best solution except only for the case where s_4 occurs. However, the occurrence probability (*prob*) of s_4 is likely going to be less than that of s_1, s_2 and s_3 altogether, i.e. $prob(s_4) < prob(s_1) + prob(s_2) + prob(s_3)$.

Ideally, CDN providers may want to obtain a RSP solution that not only has good worst-case total distribution cost but also has the total distribution cost as close as possible to the optimal in each scenario. We therefore employ as the second criterion the minimization of ***relative regret***. The relative regret of a solution in a given scenario is defined as the performance difference in percentage between the solution in that scenario and the optimal solution for that scenario. Thus, CDN providers may seek a RSP solution that keeps the worst-case total distribution cost as low as possible while minimizing the performance deviation of each scenario from optimality.

By jointly optimizing the minimax and relative regret criteria, a bi-criteria robust RSP problem is formulated. The solution that simultaneously optimizes both objectives is called *pareto-optimal*. However, as shown in the example of Table 2, the two objectives may conflict with each other and balancing relevant trade-offs is non-trivial, in particular how to determine their weighted importance. We thus resort to using the ϵ-constraint method [20], which is one of the most favored methods of generating pareto-optimal solutions. In this technique, one objective is selected for optimization, while the other one is constrained so as not to exceed a tolerance value (ϵ). We apply the ϵ-constraint method to the robust RSP by placing a constraint on the relative regret that may be attained by the solution while optimizing the worst-case total distribution cost across all the scenarios. More specifically, the constraint dictates that the relative regret in any scenario must be no greater than ϵ, where $\epsilon \geq 0$. In other words, the cost under each scenario must be within $100(1+\epsilon)\%$ of the optimal cost for that scenario z_s^*. By successively adjusting ϵ, one can obtain solutions with smaller relative regret but greater worst-case total distribution cost and vice versa. One objective of this paper is to demonstrate empirically that substantial improvements in robustness can be attained without large increases in the worst-case total distribution cost.

C. Problem Formulation

By taking into consideration the minimax and the relative regret criteria, we revise the deterministic RSP problem into a robust RSP problem. The optimization objective of the robust RSP problem is to

$$Minimize \quad \underset{s\in S}{Max} \sum_{i\in I} \sum_{j\in J} d_{s,i} c_{i,j} Y_{i,j} \tag{7}$$

subject to the following constraints:

$$\forall i \in I : \sum_{j\in J} Y_{i,j} = 1 \tag{8}$$

$$\forall i \in I, j \in J : Y_{i,j} \le X_j \tag{9}$$

$$\sum_{j\in J} X_j = P \tag{10}$$

$$\forall i \in I, j \in J : X_i, Y_{i,j} \in \{0,1\} \tag{11}$$

$$\forall s \in S : \sum_{i\in I} \sum_{j\in J} d_{s,i} c_{i,j} Y_{i,j} \le (1+\epsilon) z_s^* \tag{12}$$

Constraints (8)-(11) are identical to constraints (2)-(5). Constraint (12) enforces the ϵ-constraint condition. Compared to the deterministic RSP, which minimizes the total distribution cost for a particular traffic demand scenario, the robust RSP optimizes the worst-case total distribution cost across a variety of traffic demand scenarios, as expressed by the objective function (7). On the other hand, it is not surprising that the robust RSP problem is NP-hard since it is an extension of the deterministic RSP problem, which is itself NP-hard [1]. When the number of traffic demand scenarios $|S| = 1$ and $\epsilon = \infty$, the robust RSP problem reduces to the deterministic one.

4 Alterative Strategies for Managing Traffic Demand Uncertainty

Our implementation of the robust RSP is only one of several methods that can be used to help dimension a network under traffic demand uncertainty. Some common alternative approaches, such as mean-value model, worst-case model and stochastic optimization, can be considered for performance comparison. When applied to the RSP, these approaches differ in their structural traffic volume distribution.

A. Mean-Value Model

In the mean-value model, each element of the *mean traffic demand* is defined as:

$$\overline{d}_i = \sum_{s\in S} d_{s,i} / |S| \qquad \forall i \in I \tag{13}$$

where $|S|$ is number of traffic demand scenarios. The mean traffic demand is then taken as input to solve the deterministic RSP problem (Eq. 1-5).

B. Worst-Case Model

In the worst-case model, each element of the *worst-case traffic demand* is defined as:

$$\hat{d}_i = \underset{s\in S}{Max}\, d_{s,i} \quad \forall i \in I \tag{14}$$

In a similar fashion to the mean-value model, this worst-case traffic demand is then taken as input for solving the deterministic RSP. Note that the total traffic volume of the worst-case traffic demand serves as upper bound of the other traffic demands.

C. Stochastic Optimization (Expected Value Criterion Model)

Stochastic optimization is typically used for solving decision-making problems under risk situations. In the context of stochastic optimization, the expected value criterion model is commonly used. The model seeks the minimization of the expected total distribution cost over all traffic demand scenarios. The input data of the RSP assumes that each traffic demand scenario is probabilistic and the optimization objective is to

$$minimize \sum_{s \in S} \alpha_s \sum_{i \in I} \sum_{j \in J} d_{s,i} c_{i,j} Y_{i,j} \tag{15}$$

where α_s is the occurrence probability of traffic demand scenario s. Without loss of generality, we assume in this paper each traffic demand scenario has equal occurrence probability. This assumption is also known as *Laplace criterion* [9] for decision making under uncertainty. The Laplace criterion is based on the *principle of insufficient* reason. It asserts that, if one is completely unaware of which scenario will happen, then these scenarios may be treated as equally likely, since there is no reason to believe otherwise.

5 Evaluation Methodology

A. Network Topology

Our simulation is performed on 30-node AS-level topologies with node degree of 3, generated by the BRITE topology generator [10]. Each node and link in the topology represents an AS and a physical link between ASes respectively. Each link is associated with a propagation delay generated by BRITE. We assume that all ASes are user nodes and they are also considered as potential server nodes. The cost between two ASes (i.e. $c_{i,j}$) is the sum of link propagation delays along the shortest AS-hop path. Since the communication cost within an AS is often much better than between different ASes, we assume that at most one replica can be placed within each AS and we neglect the distribution cost generated by users attached to the same AS as the replica.

B. Web Content Traffic Demand

We generate synthetic traffic demands for our evaluation. We attach a traffic demand to each AS, which represents the total traffic demand requested at the AS. Previous work has shown that web traffic is not uniformly distributed. According to [11], the popularity of web content follows a Zipf-like distribution of $y \sim x^{-\alpha}$, which is a widely adopted model for real Web traces. The default value of popularity parameter α is set to be 0.75 with a reference to the analysis of real Web traces in [11].

We generate traffic demand scenarios using the methodology proposed in [6]: the traffic demand can vary within known ranges or can be estimated within known accuracy. This range is denoted by an error margin parameter $\omega \geq 1$. We consider ***base traffic demand*** which can be thought of as our best "guess" of the actual traffic

demand. The set of applicable traffic demand scenarios, which we call ***developed traffic demands***, includes each scenario s with error margin such that

$$d_{s,i} = \{\xi \in \mathbb{R} : d_i / \omega \leq \xi \leq \omega d_i\} \quad \forall i \in I \tag{16}$$

These developed traffic demands can be thought of as corresponding to traffic fluctuation or random errors in traffic estimation. The above method for generating traffic demand scenarios has also been used to evaluate many practical optimization problems [6] such as the robust Knapsack Problem. We remark that this traffic demand generation process is our best attempt to model web traffic fluctuation, as no synthetic model for the actual behavior of traffic in real networks can be found in the literature.

C. Comparison of RSP Approaches

We compare the performance of the following RSP approaches in our simulation:

- **Deterministic:** we run the deterministic RSP individually for each of the base and the developed traffic demands. We then use each of these RSP solutions to obtain the total distribution cost that would be achieved by the other traffic demand scenarios. In our simulation, we denote as "base" the deterministic optimization taking the base traffic demand as input. Likewise, the term "first" denotes the deterministic optimization taking the first of the developed traffic demands as input and so forth.
- **Statistical:** we run the mean-value and the worst-case models. These models roduce their traffic demands using the base and the developed traffic demands. We denote as "mean" and "worst" the two models respectively.
- **Robust:** we run the robust RSP approach by taking the base and the developed traffic demands as input data scenarios. We denote this approach as "robust".
- **Stochastic:** we run the stochastic optimization (i.e. the expected value criterion model) by taking the base and the developed traffic demands as input data scenarios. We denote the stochastic optimization as "stochastic".

D. Performance Metrics

The following two performance metrics [7] are used to evaluate different RSP approaches. For these metrics, lower values are better than high values.

- **Solution robustness:** an RSP solution is robust to the total distribution cost if it performs reasonably well for any realization of the traffic demand scenarios $s \in S$. For this metric, we capture the worst-case (i.e. the highest) total distribution cost under all the traffic demand scenarios for each RSP approach.
- **Relative robust deviation:** we capture the maximum relative regret under all the traffic demand scenarios for each RSP approach.

6 Simulation Results

All the RSP approaches presented in this paper have been implemented using the AMPL modeling language [12] and solved by the Mixed Integer Nonlinear Programming

(MINLP) solver [13][4]. The MINLP solver implements a branch and bound algorithm searching a tree whose nodes correspond to continuous nonlinear optimization problems. The continuous problems are solved using filterSQP, a Sequential Quadratic Programming solver, which is suitable for solving large nonlinear problems.

An important element in our simulation is the generation of various traffic demand scenarios. Following the methodology described in Section 5-B, we generate a base traffic demand and five developed traffic demands. Each simulation result takes approximately 10 minutes running time on average.

A. Evaluation of Solution Robustness

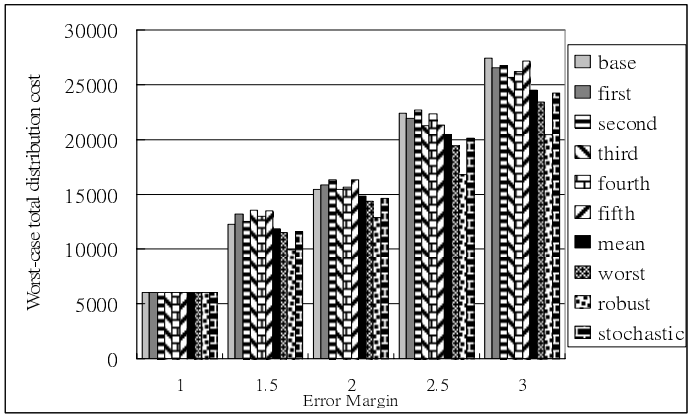
In this section, we evaluate the solution robustness of different RSP approaches. Regarding the value of ϵ, we initially set it to ∞ and then evaluate its impact on the worst-case total distribution cost and relative regret in the subsequent sections.

Figures 2(a) & (b) show the worst-case total distribution cost as a function of error margin for P=5 and P=10 respectively, where P is the number of servers to be selected. Similar result patterns for all the RSP approaches are exhibited in the two figures. An obvious difference between them is that the higher the P, the lower the worst-case total distribution cost because more servers can be located closer to the users. Therefore, we make a performance analysis that is applicable to both P results.

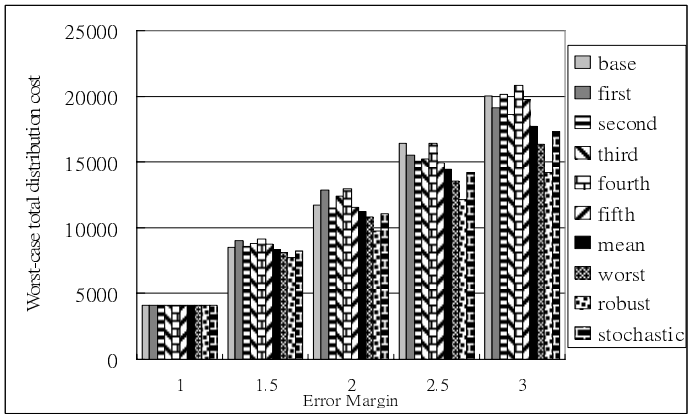
When ω=1.0, all the RSP approaches produce identical performance because they use identical traffic demand. At all other values of error margin, we observe a general phenomenon that the deterministic approach ("base", "first"…"fifth") is the worst performer. This result is expected: in fact the RSP solution optimized for a particular-traffic demand scenario may no longer maintain optimality for the other scenarios that have different structural traffic distribution. The performance gets worse when the error margin is large. In contrast, the statistical approach (i.e. "mean" and "worst") has slightly better performance than the deterministic approach. For the mean-value model, since the mean traffic demand is mixture of traffic characteristics from different traffic demand scenarios, it usually performs better than the deterministic approach that optimizes for only one traffic demand scenario. On the other hand, the worst-case model performs even slightly better than the mean-value model since the worst-case performance is optimized; this is close to the optimization objective of the robust RSP. This model, however, is overly conservative and does not produce truly optimal performance under traffic demand uncertainty. This is demonstrated by the superior performance of the robust approach.

Unlike the others, the worst-case total distribution cost of the robust approach increases smoothly as the error margin increases. This shows that the performance is not overly sensitive to errors in traffic demand estimation. Compared to the robust approach, the stochastic approach can only perform as well as the mean-value model. This phenomenon is expected because both approaches would behave optimally in the mean. However, they show poor performance at some particular realization of

[4] Ideally, heuristics are proposed to handle large-scale NP-hard optimization problems. However, since this paper aims at demonstrating the effectiveness of the robust RSP approach on coping with traffic demand uncertainty, we therefore solve the RSP problem using mathematical programming. Nevertheless, we are motivated to devise efficient heuristics to solve the problem as our future work.



(a) P=5



(b) P=10

Fig. 2. Worst-case total distribution cost vs. error margin

scenarios. On the whole, for coping with traffic demand uncertainty, the robust approach has significantly minimized the worst-case total distribution cost over the deterministic approach (about 15%-30% across different error margin values) and both the statistical and stochastic approaches (about 8%-20%).

B. Evaluation of Relative Robust Deviation

We would like to know for the results presented so far how much their performance deviates from the optimal one. We present the relative robust deviation results in Table 3. For brevity, we only show the results for P=10 and error margin equal to 2.0. For all other values of error margin, we have reached a similar conclusion.

The results in Table 3 can be interpreted as follows. Each row represents the solution of a given RSP approach, and each column represents a traffic demand scenario. The value of the table element α_{ij} (that is row i, column j) corresponds to the relative regret that would result for traffic demand scenario j if the solution of RSP approach i was implemented. Therefore, the values on the diagonal represent zero relative regret.

The maximum relative regret under all the traffic demand scenarios for each RSP solution in the row is shown in bold and underline face. The results show that the deterministic approach has the highest maximum relative regret. The robust approach is the best performer followed by the statistical and the stochastic approaches. This result is similar to that in Figure 2(b); in general, the higher the worst-case total distribution cost, the higher the maximum relative regret.

[5] **Table 3.** Relative regret (in %) for P=10 and ω = 2.0

Scenario / Solution	*base*	*first*	*second*	*third*	*fourth*	*fifth*	*mean*
base	0	17.23	18.23	23.04	**29.43**	17.54	15.21
first	22.24	0	**35.11**	25.32	26.51	19.24	14.31
second	21.45	19.92	0	27.34	24.12	**28.72**	17.9
third	22.45	15.33	19.37	0	**30.18**	14.35	12.45
fourth	19.26	27.21	28.23	19.62	0	**38.12**	20.59
fifth	**27.78**	23.57	25.12	19.43	27.27	0	18.56
mean	12.45	14.39	14.63	**20.21**	18.41	12.77	0
worst	10.22	16.32	9.56	14.56	**20.45**	14.13	7.72
robust	7.11	**11.67**	5.12	10.67	5.22	4.25	3.24
stochastic	14.15	12.21	14.13	16.24	**21.31**	13.74	9.43

C. Evaluation of ϵ

One of the objectives of the robust RSP is to minimize the maximum relative regret (by the choice of ϵ) with as little increase in the worst-case total distribution cost as possible. To illustrate this trade-off, we used the constraint method of multi-objective programming [14] to generate a trade-off table between the worst-case total distribution cost and the maximum relative regret. In particular, we first solved the problem with $\epsilon = \infty$ and recorded the two performance metrics; we then set ϵ equal to the maximum relative regret minus a small step down value (e.g. 0.2%) and re-solved the problem, continuing this process until no feasible solution could be found for a given value of ϵ. We performed this experiment using the traffic demand scenarios with error margin equal to 2.0 and P=10. The results are summarized in Table 4.

Table 4. Worst-case total distribution cost versus maximum relative regret

ϵ	*Total Distribution Cost*	*% Increase*	*Max Rel Reg*	*% Decrease*
∞	9753	0.0%	11.67%	0.0%
0.1147	9806	0.55%	11.27%	3.42%
0.1107	9863	1.13%	9.89%	15.25%
0.0969	10102	3.57%	8.54%	26.82%
0.0834	10187	4.46%	7.43%	36.33%

The column marked "ϵ" gives the value of ϵ used to solve the problem; "*Total Distribution Cost*" is the worst-case total distribution cost; "*% Increase*" is the percentage by which the worst-case total distribution cost is greater than that of the found

[5] Traffic demand produced from the worst case model is not included in the table column since it has higher traffic volume than the other traffic demand scenarios.

solution using $\epsilon = \infty$; "*Max Rel Reg*" is the maximum relative regret of the best found solution; and "*% Decrease*" is the percentage by which the maximum relative regret is smaller than that of the found solution using $\epsilon = \infty$. It is clear that large reductions in the maximum relative regret are possible with only small increases in the worst-case total distribution cost. For example, the last solution represents a 36.33% reduction in the maximum relative regret with only less than a 4.46% increase in the worst-case total distribution cost. These results justify the use of the ϵ-constraint approach since it costs very little to achieve robustness.

D. Performance Summary of the RSP Approaches

The simulation study in this section evaluated the performance of different RSP approaches. Simulation results have shown that the robust approach produces significantly better worst-case total distribution cost than non-robust approaches under traffic demand uncertainty. The robust approach also guarantees the performance of the solution to be within a specified envelope from the optimal solution, thereby improving robustness on RSP performance. We therefore conclude that the robust RSP approach can make RSP performance more robust and predictable.

7 Conclusions

In this paper we faced the problem of RSP, assuming that traffic demand uncertainty is handled by a set of traffic demand scenarios. By using the principles of scenario-based robust optimization, we proposed a novel integer programming formulation for robust RSP. We provided empirical results to assess the performance of several commonly used techniques for robust RSP. The results show that the robust RSP approach, whose optimization runs across the set of traffic demand scenarios, significantly improves the solution robustness while it also minimizes the performance deviation from the optimal solutions. We believe that our work provides insights to CDN providers on how to design robust CDNs by reducing the impact of inaccuracy in traffic demand estimation so as to provision better and predictable QoS for their users and avoid potential loss in business revenues.

We emphasize that our idea of SRO is not only limited to the RSP problem. In fact, the CDN-related design problems to which it can be applicable are *numerous*. Examples are web object replication [15,18], request routing [18], cache location [16] and topological design for service overlay networks [17]. A common characteristic of these CDN design problems is that their optimization objectives are influenced by the accuracy of estimated traffic demands. We believe that the robust approach can be adopted by CDN providers as a means to make their networks more robust.

Acknowledgement

This work was undertaken in the context of FP6 Information Society Technologies AGAVE (IST-027609) project, which is partially funded by the Commission of the European Union.

References

1. L. Qiu et al., "On the Placement of Web Server Replicas," Proc. *IEEE INFOCOM*, 2001.
2. S. Jamin et al., "Constrained Mirror Placement on the Internet," Proc. *IEEE INFOCOM*, 2001.
3. X. Tang and J. Xu, "QoS-Aware Replica Placement for Content Distribution," *IEEE Transactions on Parallel and Distribution Systems*, 16(10), 2005, pp. 921-932.
4. G. Rodolakis et al., "Replicaed Server Placement with QoS Constraints," Proc. *3rd International Workshop on QoS in Multiservice IP Networks (QoSIP)*, 2005.
5. A. Feldmann et al.., "Deriving Traffic Demands for Operational IP Networks: Methodology and Experience," *IEEE/ACM Transactions on Networking*, 9(3), 2001, pp. 265-280.
6. P. Kouvelis and G. Yu. *Robust Discrete Optimization and Its Applications*, Kluwer Academic Publishers, 1997.
7. J.M. Mulvey et al., "Robust optimization of large-scale systems," *Operations Research*, 43, 1995, pp. 264-281.
8. N. Hu et al., "Optimizing Network Performance in Replicated Hosting," Proc. *IEEE International Workshop on Web Caching and Content Distribution (WCW)*, 2005.
9. H.A. Taha. Operations Research, 7th edition, Prenticall Hall, 2003.
10. A. Medina et al., "BRITE: An Approach to Universal Topology Generation," Proc. *MASCOTS 2001*, 2001.
11. L. Breslau et al., "Web Caching and Zipf-like Distributions: Evidence and Implications," Proc. *IEEE INFOCOM*, 1999.
12. A Modeling Language for Mathematical Programming. Available at www.ampl.com.
13. The MINLP solver. University of Dundee, UK.
14. J.L. Cohon. *Multiobjective programming and planning*. Mathematics in Science and Engineering. Academic Press, New York, 1978.
15. J. Kangasharju et al., "Object replication strategies in content distribution networks," *Computer Communications*, 25(4), 2002, pp. 376-383.
16. P. Krishnan et al., "The cache location problem," *IEEE/ACM Transactions on Networking*, 8(5), 2000, pp.568-582.
17. S.L. Vieira and J. Liebeherr, "Topology design for service overlay networks with bandwidthi guarantees," Proc. *IEEE IWQOS*, 2004, pp. 211-220.
18. T. Bektas et al., "Designing cost-effective content distribution networks," to appear in *Computers & Operations Research*, 2006.
19. The economic impacts of unacceptable web-site download speeds. Zona Research, 1999.
20. V. Chankong and Y.V. Haimes. *Multiobjectve Decision Making–Theory and Methodology*, Elsevier, New York, 1983.
21. Y. Zhang et al., "An Information-Theoretic Approach to Traffic Matrix Estimation," Proc. *ACM SIGCOMM*, 2003.