Algorithmic Aspects of Planning under Uncertainty for Service Delivery Organizations

Sreyash Kenkre[1], Ranganath Kondapally[2,⋆], and Vinayaka Pandit[1]

[1] IBM India Research Laboratory
{srekenkr,pvinayak}@in.ibm.com
[2] Computer Science Dept., Dartmouth College, Hanover, USA
rangak@cs.dartmouth.edu

Abstract. Remote delivery of services using geographically distributed service delivery locations has emerged as a popular and viable business model. Examples of services delivered in this manner are software services, business process outsourcing services, customer support centers, etc. The very nature of services and the fragile nature of the business environments in global delivery locations accentuates the role of uncertainty in planning for business continuity. We model the problem of critical service contingency planning based on *recourse actions*. We present an $O(\log n)$-approximation algorithm, generalizations to other planning problems under uncertainty, and present preliminary empirical results.

1 Introduction

Business continuity is an important aspect of service delivery. This entails service provider's commitment of continuity of business operations to the service seeker. The service provider could be IT enabled service provider, governance body delivering services to its citizens, public utility serving the citizens, etc. In this paper, we present examples from IT-enabled service delivery. But, the concepts are applicable much more broadly.

Recently, countries like India, China, Brazil, etc. have emerged as popular destinations to deliver software services, back-office services, remote infrastructure management, etc. due to the investor friendly policies and access to talent. Typically, the delivery centers and the consumers of the delivery are geographically separated. Such service delivery is enabled by setting up of large-scale, geographically distributed IT infrastructures consisting of heterogenous resources. Although this model is attractive, it also faces the challenges of heightened uncertainties in the operating environments of these geographies. We review the relevant issues taking example of a hypothetical organization X.

Operational Setting: Let us say X is an organization that delivers software services (ex: support, maintenance, testing, feature developments, etc.) to a large number of customers worldwide. X delivers its services from multiple countries, with multiple campuses in each country which allows it to tap into appropriate

⋆ The work was done when the author was visiting IBM Research - India

C. Liu et al. (Eds.): ICSOC 2012, LNCS 7636, pp. 615–622, 2012.

workforce with required skills. Furthermore, it deploys a complex infrastructure of servers, communication networks, buildings, utilities, transportation logistics, etc. Each customer is treated as a customer account (or project) and is characterized by the combination of resources it requires. For instance, a customer account could be characterized by the physical security feature of the workplace it requires (ex: seats with secure badge access), the cyber security features of the WANs and LANs it uses, power requirements, and access features to the client environment. Therefore, enabling the service delivery for a customer account essentially involves making available the right combination of the resources. One of the reasons for the feasibility and profitability of X is the fact that there is lot of similarity in the services it delivers to different accounts. Hence, it can achieve economies of scale for resources that are commonly used across accounts (one example of such an infrastructure element is "Wide Area Network"). See [4] to understand why service providers prefer this model.

Motivation for Contingency Planning: Customers sourcing such global service delivery would naturally be worried about the uncertain business environment of the emerging geographies. Therefore, they put in place stringent SLAs on the continuity of service delivery. Typically, the customer identifies a subset of the procured services as "critical" and demands that the service provider provide round the clock continuity for at least the critical services. Examples of critical services in case of X could be "fix all high priority bugs", "fix bugs reported on security features of the product", and so on. See [7] for a client perspective of global sourcing of services. Even for the service provider, a strong commitment to business continuity not only helps meet the SLAs, but also build brand reputation for future business. Therefore, service delivery organizations are increasingly making contingency planning an integral part of overall operations.

Disruptions and Rerouting: Majority of the disruptions that arise in emerging geographies are local to a city, suburban area, etc. Examples of such disruptions are strikes, societal unrest, urban flooding, natural disasters, below par supply of utilities, etc. When a disruption happens at a location, the part of the organization's infrastructure located there is unavailable. Therefore, one of the most popular techniques that companies like X use is to deliver the services out of multiple locations in the geography [3,10,4]. Moreover, at each location, X ensures that there is sufficient residual capacities of different resource types, so that, during a crisis, some of the service delivered from the affected locations can be rerouted to unaffected locations. When X reroutes the services for a customer account from one location to another, it has to ensure that the right combination of the resources are available in the rerouted location. Such a reroute action is called "recourse". See [5,6] for elaboration on the importance and implications of "recourse aware" decision making in business operations resiliency.

Critical Services Contingency Planning Problem: As explained, for each customer account, X has to allocate an appropriate combination of resources, all co-located to enable service delivery. We consider the problem of contingency management plans for the critical services being delivered by the service provider. We assume that the organization has identified a set of challenging scenarios

(each scenario is defined by a set of unavailable locations) and a probability distribution on their likelihood. For each critical service, we are given a set of locations from where it can be delivered (based on the availability of the right combination of services). The contingency management of critical services has to : (i) compute a default assignment of the services to the locations for delivery during normal conditions and (ii) compute a scenario specific assignment of the services to the locations under each scenario. The goal is to compute the assignments in such a way that the expected total cost of the normal and contingency operations.

Broader Applicability: The concept of recourse based handling of contingency, as described here, can be applied in other settings like city governance, network design for internet service providers, utilities in the power sector, etc. In our setting, the effect of an incident is local; for example, flooding in an office building only affects the infrastructure situated in the building. In contrast, the effect of an incident could be global. A classic example of this phenomena is the way cascades spread in power grids. The contingency analysis in such networks need more global formulations than ours and can be seen in [1,8].

2 Critical Service Contingency Planning (CSCP)

Location Mapping: As mentioned in the introduction, each project has a requirement in terms of the resource types it requires. Each location in the service delivery infrastructure has a set of available resources. One way to formulate the problem would be to capture all the details of the resources in the problem definition itself, as done in [5,6]. Note that the resource requirements of the critical services and resource availability results in the mapping of each critical service to a possible set of locations that it can be assigned to (as done in [5,6]). But, for simplicity of presentation, we assume that the mapping is itself part of the input. Therefore, in our setting, there is a set of locations, a set of critical tasks, and associated with each critical service is a set of locations to which it can possibly be assigned.

No Capacity Constraints: Capacity constraints are an important consideration and have been modeled in [5,6]. However, it is a well known fact that the critical services form a small fraction of the overall services delivered by the organization. Often, it is in the range 5-10% of the overall work. But, from the point of view credibility of the business operations and client satisfaction it is the most important part of the work and always gets highest priority. Therefore, even when there are capacity constraints, non-critical services are de-prioritized and capacity is made available to the critical tasks. Therefore, we assume that there are no capacity limits at the locations for assigning the critical tasks.

Cost Considerations: When a location is assigned a set of critical services for normal operational setting, it incurs *set up cost*. The set up cost could cover special requirements of critical services, transportation of people, and other procurements. However, suppose a locations has not been assigned any critical service during normal operations and has to suddenly make arrangements for critical

services during a disruption, it incurs *recourse cost*. Typically, recourse cost is much higher than the set up cost as the required arrangements (and the implied procurements) have to be carried out at a short notice.

Scenarios: The uncertainty in the service delivery manifests in the form of disruptions in normal operations. We model disruption (also called scenario) as an event which disables the delivery of services from a set of affected locations. In most service delivery organizations, there are domain experts who can model the relevant set of scenarios for the organization. They could take into account aspects like bottlenecks in the infrastructure (example: suppose there is just one mail server for the entire organization, one must consider the disruption which disables the location of the mail server) or external parameters of the locations (example: if some locations are vulnerable to flooding, then, one must consider a scenario which affects such locations). We assume that a domain expert provides the set of scenarios for contingency planning. We further assume that a probability distribution on the likelihood of the scenarios is given.

Problem Formulation: Let the set of locations be specified by the set $L = \{S_1, S_2, \ldots, S_m\}$. Let the set of projects in the service delivery organization be given by $\mathcal{P} = \{P_1, P_2, \ldots, P_t\}$. Let the set of critical tasks across the different projects be given by the set $T = \{v_1, v_2, \ldots v_n\}$. Associated with a critical task v_i is a subset of locations, $\Gamma(v_i) \subseteq L$, which denotes the set of locations to which v_i can potentially be assigned. Given this, one can also define the set of services that can be defined from a location S_i as $CT(S_i) = \{v_i | S_i \in \Gamma(v_i)\}$. The set of scenarios modeled by the domain expert is given by $\mathcal{S} = \{E_1, E_2, \ldots E_k\}$ where $E_i \subseteq L$ is the set of locations affected in the ith scenario. The probability distribution on the likelihood of the scenarios is given by the mapping $\mu : \mathcal{S} \to [0, 1]$ such that $\sum_{E_i \in \mathcal{S}} \mu(E_i) = 1$. The set up cost and recourse cost of $S_i \in L$ is given by $c(S_i$ and $r(S_i)$ respectively. The CSCP problem requires us to compute an assignment R of critical services to the locations that is to be followed during normal operations and a set of recourse assignments F_{E_j} for each $E_j \in \mathcal{S}$. Let $Sites(R)$ denote the set of locations that are used in the assignment R and $Sites(F_{E_j})$ denote the set of locations that are used in the assignment F_{E_j}. The goal is to minimize the expected cost of the normal and contingency operations, i.e, minimize $\left(\left(\sum_{S_i \in Sites(R)} c(S_i)\right) + \left(\sum_{E_j \in \mathcal{S}} \mu(E_j) \cdot \left(\sum_{S_i \in Sites(F_{E_j})} r(S_i)\right)\right)\right)$.

3 Algorithm, Proof, and Generalization

We now present an algorithm for the CSCP problem with a provable approximation ratio. In other words, our algorithm always returns a solution whose cost with respect to the optimal solution is bounded by the approximation ratio.

We first begin by showing the equivalence of our problem to the *Stochastic Set Cover* problem. Set Cover is one of the fundamental problems in combinatorial optimization and approximation algorithms [9] and is the following: given a universe U and a family of sets, $\mathcal{X} = \{X_1, \ldots, X_q\}$ where all the X_is are subsets of U, and a weight function $w : \mathcal{X} \to \mathbb{R}$, the goal to pick a subset $R \subseteq \mathcal{X}$

such that $\cup_{X_i \in R} X_i = U$ and the weight $\sum_{X_i \in R} w(X_i)$ is minimized. Let us consider the problem of just assigning all the critical tasks to one of the sites. The universe is the set of all the critical tasks. Each site is a set which contains the critical tasks that can be assigned to it. The weight function associated with the sites is their set up cost. The task of computing an assignment is now equivalent to the minimum cost set cover of the universe of critical tasks by the sets corresponding to the sites weighted by their set up cost. Let us now consider the CSCP problem. Essentially, it is a two stage stochastic problem. We first want to pick an assignment which acts as the set cover during normal operations. At the second stage, one of the scenarios is revealed by the nature according to the probability distribution on $\mathcal{S}$. At that stage, we have to pick new sites for the critical services that were assigned to the affected locations. In the second stage, we have to incur the recourse cost at the new sites. Essentially, we want to pick a two stage set cover which minimizes the expected cost.

$$\text{Minimize} \sum_{S_j \in L} c(S_j) y_{S_j} + \sum_{E_l \in \mathcal{S}} \mu(E_l) \cdot \left(\sum_{S_j \in L} r(S_j) \cdot y^l_{S_j} \right)$$

$$\sum_{S_j \in \Gamma(v_i)} y_{S_j} \geq 1 \quad \forall v_i \in V \qquad \text{(Normal Covering)}$$

$$\sum_{j \in L \setminus E_l} y_{S_j} + y^l_{S_j} \geq 1 \quad \forall E_l \in \mathcal{S} \qquad \text{(Cover } E_l\text{)}$$

$$y_{S_j} \in \{0,1\} \quad \forall S_j \in L$$

$$y^l_{S_j} \in \{0,1\} \quad \forall S_j \in \mathcal{L} \text{ and } E_l \in \mathcal{S} \qquad \text{(LP)}$$

We present the integer linear programming formulation for the CSCP (See (LP)). Here, y_{S_i}s are decision variables that indicate whether set up cost is incurred in the first stage at S_is and $y^l_{S_i}$ are decision variables that indicate whether recourse cost is incurred in the second stage at S_i in the scenario E_l. The integer linear program is self-explanatory. We present it because it helps us to present generalizations of our result.

Our algorithm is as shown in Algorithm 1. The main idea is to not take any anticipatory decisions in the first stage and use any approximation algorithm for the set cover at both first and second stage. In the second stage, an appropriate set cover instance is created depending on the set of "failed" critical services corresponding to the scenario. Details are presented in the algorithm itself. The main result is the following theorem.

Theorem 1. *Suppose the deterministic algorithm $\mathcal{A}$ in Algorithm 1 has an approximation ratio of α for the Set Cover problem, then, the approximation ratio of the Algorithm 1 is $\left(1 + \max_{S_i \in L} \frac{r(S_i)}{c(S_i)}\right) \alpha$ for the CSCP problem.*

Proof. Due to space considerations, we present just the overview of the proof. Let $(R, F_{E_1}, \ldots, F_{E_k})$ denote the solution computed by the algorithm. Let $(R^*, F^*_{E_1}, \ldots F^*_{E_k})$ denote the optimal solution of cost:

input : Critical Tasks T, Locations L, Critical Task Mapping $\Gamma()$, set up costs $c(S_i)$s, recourse costs $r(S_i)$s, and scenarios $\mathcal{S}$
output: First stage assignment R and second stage assignments F_{E_i}s
1 Employ a deterministic algorithm $\mathcal{A}$ for the set cover problem with T as the universe, $CT(S_i)$s as the sets, and $c(S_i)$s as the weights. Assign tasks to any of the selected sites to which they can be mapped. This gives R and $Sites(R)$. ;
2 **for** $E_j \in \mathcal{S}$ **do**
3 Let $affected = \{v_i \in T | R(v_i) \in E_j\}$, i.e, tasks which are assigned by R to an affected location in the scenario E_j ;
4 If any of $affected$ nodes can be assigned to $R \setminus E_j$, then, assign them. Let $failed$ be the set of remaining nodes ;
5 Employ $\mathcal{A}$ on $failed$ as the universe, $CT(S_i)$ for $S_i \notin E_j$ as the sets, and $r(S_i)$ for $E_j \notin E_j$ as the costs. Assign tasks as in Step 1 to get F_{E_j} and $Sites(F_{E_j})$.
6 **end**
7 Output $R, F_{E_1}, \ldots, F_{E_k}$.

Algorithm 1. The Main Algorithm

$$OPT = \Big(\sum_{S_i \in Sites(R^*)} c(S_i) + \sum_{E_j \in \mathcal{S}} \mu(E_j) \cdot \Big(\sum_{S_i \in Sites(F^*_{E_j})} r(S_i) \Big) \Big)$$

Clearly, $\sum_{S_i \in Sites(R)} c(S_i) \leq \alpha \sum_{S_i \in Sites(R^*)} c(S_i)$ (due to the approximation property of $\mathcal{A}$). Now consider a scenario E_j. Let $R^*_{1,j}$ be the sites from R^* that are used by the optimal solution and $R^*_{2,j}$ be the sites with recourse cost used in the second stage. Note that $\sum_{S_i \in R^*_{1,j}} c(S_i) + \mu(E_j) \sum_{S_i \in R^*_{2,j}} r(S_i) \leq OPT$. Clearly, $R^*_{1,j} \cup R^*_{2,j}$ is candidate for F_{E_j} at step 5 of the algorithm and its recourse cost is at most $\max_{S_i \in R^*_{1,j} \cup R^*_{2,j}} \frac{r(S_i)}{c(S_i)} \cdot OPT$. Therefore, the algorithm $\mathcal{A}$ at step 5 picks a solution F_{E_j} of cost at most $(\alpha \max_{S_i \in R^*_{1,j} \cup R^*_{2,j}} \frac{r(S_i)}{c(S_i)} \cdot OPT)$ – this is due to the approximation property of $\mathcal{A}$. Since this applies for every $E_j \in calS$, we get the approximation ratio claimed in the theorem.

The greedy heuristic of picking the set with best coverage (least cost per element covered) at every step is an $O(\log n)$ approximation where n is the size of the universe and is asymptotically tight under the assumption of P $\neq$ NP [9]. Therefore, we get an approximation ratio of $(1 + \max_{S_i \in L} \frac{r(S_i)}{c(S_i)}) \log |T|$ and it is asymptotically the best possible.

3.1 Generalization

We present a generalization of our result to a family of stochastic planning problems (applicable to service delivery). The proofs will be included in longer version. Consider any planning problem which has to meet a fixed demand set T. The demand has to be met by setting up a required structure R by choosing from a universe options U. There is a set of scenarios $\mathcal{S}$ with a probability distribution

on them and each scenario mentions a subset of U that is not available. We want to build a stochastic solution for the structure R which minimizes the expected cost of building the structure. For example, an internet service provider might want to build Steiner Tree [9] on a set of demand points and the universe consists of the different edges (pairs of end-points) that can be included in the tree. Scenarios are sets of edges that can fail. We can prove the following. Suppose we can write an integer linear program for the problem of computing R and if there is an α approximation for the non-stochastic version of the problem, our algorithmic framework gives an approximation ratio of $(1 + \beta)\alpha$ where β is the worst ratio of the recourse cost to the set up cost of the elements in U.

Finally, we end with a comment on the complexity added by the stochastic part. Consider the problem of constructing Minimum Spanning Tree (MST) of a graph. This is solvable in polynomial time. But, the problem of computing the optimal two stage MST solution to the stochastic version as considered in this paper becomes NP-Complete via a reduction from the Hamiltonian Cycle Problem [2].

4 Experimental Evaluation

The insights obtained from this work have been used in contingency planning problems in service delivery organizations within IBM. However, data from real-life service delivery scenarios are sensitive and difficult to share in public.Therefore, we conduct our experiments on randomly generated instances of the CSCP problem. The random instances were generated by varying the number of sites, number of critical services, relative costs of $c(S_i)$ and $r(S_i)$. We generated up to 50 sites and 1000 services. These are realistic numbers as organizations typically have only tens of sites and hundreds of projects.

We consider following heuristics. P refers to the organizational procedure which the organization may follow for allocation of the critical service to various sites using business rules (state of the art today). We modeled P by simple business rules such as assign to the nearest site, load balance the allocations, etc. P is important because it can capture certain rules that are hard to encode into the CSCP formulation. Heuristic P+P refers to the case when the organizations procedure is used both for the initial allocation as well as drawing up the contingency plan. A refers to the greedy set cover heuristic, which provides α approximation where $\alpha = \log |T|$. Heuristic A+A refers to the case where the procedure A is used both for the initial allocation as well as the drawing up of the contingency plan. In heuristic P+A, we make the initial allocation using P and do the contingency planning using procedure A.

Due to space considerations, we present just a sample of the experimental results in Table 1. We experimented with different probability distributions. Since our theoretical results hold irrespective of the distribution, we present the results on the uniform distribution on scenarios, i.e, no special knowledge other than the list of scenarios is known. The (A+A) solution is consistently the solution of least cost. But, there could be a practical problem in using it since

Sites	Services	P+P	A+A	P+A
20	100	385	120	300
20	200	430	190	370
40	300	613.33	466.66	526.66
40	500	906.66	520	673.33
50	1000	1356.66	780	1326.66

Fig. 1. Comparison Of Heuristics

both the stage solutions are computed without contextual business knowledge. The (P+A) solution is much lower than the (P+P) solution and it may be easier to use effectively. This is because the first stage solution, which is used during normal operations is computed taking organizational constraints into account. The second stage solution computed using the greedy heurisitc can be easily adapted to fit into organizational constraints as the number of affected critical services is much lower than the total number of critical services.

References

1. Choi, J., Mount, T., Thomas, R.: Transmission expansion planning using contingency criteria. IEEE Transactions on Power Systems 22(4), 2249–2261 (2007)
2. Garey, M., Johnson, D.: Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman (1979)
3. Graham, J., Kaye, D.: A Risk Management Approach to Business Continuity. Rothstein Associates Inc. (2006)
4. Jalona, S., Chandrakar, A.: Evolution of IT services delivery model (2008) Infosys White Paper available at `http://www.infosys.com/global-sourcing/white-papers/pages/index.aspx`
5. Karthik, S., Kenkre, S., Narayanam, K., Pandit, V.: Recourse aware resource allocation for contingency planning in distributed service delivery. In: IEEE Conference on Services Operations, Logistics, and Information, SOLI (2012)
6. Karthik, S., Kenkre, S., Narayanam, K., Pandit, V.: Resiliency analytics framework for service delivery organizations. In: Proc. of the Global Conference of Service Research Innovation Institute, SRII (2012)
7. Keane White Paper. Going global with application outsourcing (2011), Report is available at `http://www.keane.com/resources/pdf/WhitePapers/WP-GGAO.pdf`
8. Street, A., Oliveira, F., Arroyo, J.M.: Contingency-constrained unit commitment with n-k security criterion: A robust optimization approach. IEEE Transactions on Power Systems 26(3), 1581–1590 (2011)
9. Vazirani, V.: Approximation Algorithms. Springer (2001)
10. Wipro Report. Wipro Business Continuity - "Plan B", Report is available `http://www.wipro.com/documents/Wipro_Business_Continuity.pdf`