An Autonomic Service Delivery Platform for Service-Oriented Network Environments

Robert D. Callaway*,†, Michael Devetsikiotis*,†, Yannis Viniotis*,†, and Adolfo Rodriguez†§
* Department of Electrical and Computer Engineering, NC State University, Raleigh, North Carolina 27695
† WebSphere Technology Institute, IBM, Research Triangle Park, North Carolina 27709
§ Department of Computer Science, Duke University, Durham, North Carolina 27708
rdcallaw@eos.ncsu.edu, mdevets@eos.ncsu.edu, candice@eos.ncsu.edu, adolfo@us.ibm.com

Abstract—In this paper, we propose a novel autonomic service delivery platform for service-oriented network environments. The platform enables a self-optimizing infrastructure that balances the goals of maximizing the business value derived from processing service requests and the optimal utilization of IT resources. We believe that our proposal is the first of its kind to integrate several well-established theoretical and practical techniques from networking, microeconomics, and service-oriented computing to form a fully-distributed service delivery platform. The principal component of the platform is a utility-based cooperative service routing protocol that disseminates congestion-based prices amongst intermediaries to enable the dynamic routing of service requests from consumers to providers. We provide the motivation for such a platform and formally present our proposed architecture. We discuss the underlying analytical framework for the service routing protocol, as well as key methodologies which together provide a robust framework for our service delivery platform that is applicable to the next-generation of middleware and telecommunications architectures.

I. INTRODUCTION

The overarching goal of adopting service-oriented architectures (SOA) is to allocate an organization’s computing resources such that they are directly aligned with core business processes. When implemented correctly, SOAs provide a framework that reuses existing elements of an IT infrastructure while reducing total cost of ownership and providing a more flexible and robust environment for the integration of IT and business processes. Services in SOAs are coarse-grained, discoverable software entities that exist as single instances and interact with consumers, applications and other services via a loosely coupled, message-based communication model. These properties enable the flexibility of SOAs because they remove dependencies on implementation specifics by relying on interactions between services through standardized interfaces.

The use of standardized interfaces also supports service virtualization, which allows entities to provide alternate interfaces to the same service instance. This further allows value-added functionality to be inserted into the flow of a service invocation in a manner transparent to the consumer; similar concepts are being adopted in next-generation IMS and telecommunication networks. Service virtualization can also provide overload protection and security benefits, as intermediaries are able to enforce admission control policies and prevent denial-of-service attacks from reaching an actual service instance.

Loose coupling and service virtualization enable a dynamic and flexible integration infrastructure where different service providers, each of which is a perfect substitute for another, can be chosen at runtime to fulfill service requests. The service selection problem has been well-addressed in service engineering literature and in dynamic supply chain management. In both of these research areas, transportation costs between the consumer and the provider should be considered because they may contribute substantially to the consumer’s perception of the overall performance of the service invocation. Dynamic service selection enables service-oriented supply chain environments to become more agile to changing economic and environmental conditions [1]. In general, service systems seek to gain efficiency by adapting autonomically to changes in the marketplace [2]. With these points in mind, we postulate that a mapping exists between the electronic services management required in SOAs and the more tangible supply chain management practices adopted by corporations today.

In this paper, we propose a novel service delivery platform that optimally routes service requests from consumers to providers through a network of cooperative intermediaries. The intermediaries will select the “best” service provider for the request, based on weighted criteria such as relative importance of requests (as defined by business policy) and current congestion observed in the intermediaries and in the providers. The platform seeks to provide optimal flow control and routing of service requests that adapts autonomically to current conditions observed in the service-oriented environment. This approach is novel in its goal to effectively maximize the value derived from the underlying IT resources in a manner proportional to the goals of the business [3]. An instantiation of such a service delivery platform delivers the promises of SOAs by enabling a dynamic and robust integration infrastructure that we believe is applicable to both middleware and next-generation telecommunication systems.

To build the platform, we apply a cross-disciplinary research approach, drawing insight from the diverse areas of dynamic supply chain management, service engineering, network economics, application-layer networking, and distributed systems to enable an autonomic service delivery platform based on the concept of a service-oriented network. Service-oriented networking, an emerging paradigm that enables network devices to operate at the application-layer with features such as
offloading, protocol integration, and content-based routing, is key to instantiating our service delivery platform [4].

The remainder of the paper is structured as follows: in the following section, we explicitly propose our service delivery platform and the function it enables. We also discuss how methodologies from diverse research areas can be integrated to create such a platform. In Section III, we present an overview of the analytic framework that is used to provide the optimal routing and flow control in the platform. In Section IV we provide a brief review of related literature in service-oriented brokered architectures, service selection algorithms, and dynamic supply chain management.

II. ARCHITECTURE OF SERVICE DELIVERY PLATFORM

A. Overview
In this section, we propose our autonomic service delivery platform that explicitly links the value extracted from IT resources to the business processes they support within an enterprise. The platform is composed of service consumers, service-oriented intermediaries, and service providers. The platform provides:

• A fully distributed, content-based, and optimal routing infrastructure
• Flexible and optimal selection of service providers that can be based on various system-level goals (e.g. end-to-end delay, proximity, etc.)
• Optimal flow control of service requests

The novelty of our proposal arises from the integration of several well-established theoretical and practical techniques from networking, microeconomics, and service-oriented computing that, together, form a fully-distributed service delivery platform. The core component that enables the service delivery platform is a utility-based cooperative service routing protocol. The objective of this protocol is to route requests such that the weighted “social welfare” of the system is maximized. It disseminates current pricing and utility information amongst service intermediaries in the service delivery platform to cause the system to optimally forward and rate limit service requests. The system administrator defines the requisite utility functions on a per class-of-consumer basis, rather than inferring them from consumers who can be untruthful in their appraisal of services. In this way, we avoid the selfish nature of consumers and subsequently the “tragedy of the commons” that can result from such a situation.

B. Key Assumptions
To build our service delivery platform, we make several key assumptions:

• We reuse a graph-based formulation proposed in [5], as illustrated in Figure 1. In this model, we add a logical destination node to the topology that is connected to all possible providers of a semantically equivalent service over zero-cost virtual links. We also assume that a semantic matching algorithm exists a priori that can be used to select available paths through the network topology to fulfill a consumer’s request. These assumptions allow us to directly apply existing optimal multipath routing algorithms to our architecture and use pricing information as the final decision variable to make a forwarding decision for a given request.

• We assume that consumers only submit their service request to a single intermediary. This delegates the service selection decision to an intermediary with current system state to make an optimal forwarding decision.

• Service providers advertise relevant metrics to all intermediaries that act as a “last hop” in the service-oriented network before the provider. The intermediaries that receive metrics from a provider will determine the current price for the service and propagate that price throughout the network. This limits the scope for distribution of metrics from service providers to the delivery platform.

• Since the platform assumes global knowledge of per-service utility functions and trusted relationships between intermediaries such that all nodes cooperate to optimally achieve common goals, it is assumed that the delivery platform exists within a single autonomous system.

C. Methodologies Integrated in the Platform
The service delivery platform is based on the integration of several key methodologies: content-based routing, optimal routing and flow control theory, network economics, and
congestion pricing. In the subsections below, we give a brief overview of relevant issues related to each the methodologies in our service delivery platform.

1) Content-Based Routing: While previously discouraged because it violates the networking end-to-end principle, the idea of using network intermediaries to provide value-added application-aware function in the network fabric has recently been embraced [6]. Similar to active and overlay networks in its objective, service-oriented networking challenges the previous assumption that implementing application-awareness in the network fabric is too costly and complex [4]. Due to advances in hardware, software, and networking technologies, intermediaries are able to understand data encoded in XML and legacy formats, act upon that content to enforce QoS or security policies, transform the data into an alternate representation, and/or make content-based routing decisions.

We directly leverage the content-based routing function provided by a service-oriented network to enable request forwarding in our service delivery platform. Content-based routing algorithms typically apply rules against some portion of a service request (header or content) to extract attributes. These attributes are used to semantically match the service request to possible providers in the service-oriented network topology.

2) Optimal Routing & Flow Control: In addition to considering the content of requests, our service delivery platform also incorporates the observed state of the system into its optimal routing algorithm. In the seminal paper [7], a distributed algorithm to an optimal minimum delay routing problem is presented. The algorithm populates routing tables with weights that represent the fraction of incoming traffic that should be forwarded to the neighboring nodes in the network. The solution reveals that these weights are a function of the measured marginal delay on the link to each neighbor. An extension of this work is presented in [8], where the restrictions in [7] of quasi-stationary traffic, synchronization of nodes, and knowledge of the aggregate traffic demand at each node are removed. It is also shown how a near-optimal multipath routing algorithm can be implemented in a distance-vector framework while maintaining loop-free routes at every instant.

In addition to using optimal routing, we must ensure that the rate of incoming requests to a particular node in our service delivery platform is throttled appropriately. We can achieve this by integrating optimal flow control into our architecture. A proposed method for integrating a utility-maximization problem and optimal flow control is presented in [9], where the optimal routing and flow control problems are solved simultaneously while observing capacity constraints.

The integration of a distributed, loop-free, and optimal multipath routing and flow control algorithm is essential to the robustness and scalability of our service delivery platform. Since forwarding costs are determined by the sum of the congestion price of the intermediary in question and the price as advertised by the next hop (an intermediary or a provider), we exploit the additive path cost property of the underlying economic framework to build the requisite service routing protocol.

3) Network Economics: Microeconomics offers a well-developed theory on the subject of rational choice in multi-agent environments; utility functions and price are natural ways to express the common tradeoffs in such systems. Microeconomic models have been extensively applied to various engineering problems; for example, network economics are used in [5] as a method to solve dynamic supply chain management problems. The solutions that are yielded from these methods have many desirable properties, for example provable convergence to a Pareto-optimal equilibrium, in which no alternate solution exists that could increase the benefit of a user without reducing the benefit of another user. A comprehensive review of how economic theory can be applied to various networking problems is found in [10].

In our architecture, we incorporate the economic concept of social welfare maximization when formulating our optimization problem for the platform, as seen in (1) in Section III. A key distinction of our work, as compared to prior attempts in the literature, is that our formulation does not rely on the perceived or advertised utility from consumers; rather, we explicitly link the utility of services to the benefit that a corporation derives from providing the IT infrastructure. The benefits of this distinction are two-fold; first, it allows us to avoid restrictive assumptions about the explicit knowledge and/or validity of the utility functions for the system. Second, it delivers a link between IT resources and the benefits that are derived from them, which is the premise for adopting SOAs.

4) Congestion Pricing: Congestion pricing was first proposed in [11] as a basis for welfare economics and has been subsequently been applied to many engineering disciplines [12], [13]. The use of congestion-pricing resources has been investigated extensively in the networking literature in an attempt to address resource allocation problems [14]. We apply the concept of congestion pricing to balance the current state of the underlying network conditions and the performance characteristics of service providers and network intermediaries in order to optimally route requests [10]. This is represented by the term \( f(x,s,\gamma_f,z_f) \) in (1), shown in Section III.

The notion of “split-edge” pricing was proposed in [15]. In this model, prices are determined locally and only reflect prices from onward networks and providers in providing the service; however, pricing information is consolidated at each step, whether it be an intermediate broker or the actual provider. Split-edge pricing is analogous to additive path cost in next-hop routing algorithms, such as a distributed Bellman-Ford algorithm, where knowledge of the full topology and paths through the network are not required in order to make minimum cost routing decisions. We leverage split-edge pricing in the distributed solution to the optimization problem described in the next section.

We believe that the combination of “split-edge” and congestion pricing provides an intuitive and scalable method to provide congestion control in our service delivery platform. Our architecture is flexible in such a way that it is configurable
for administrators to set congestion-based prices for invoking transport services, the services of an intermediary, and the desired service at a particular provider, or any subset of the prices therein. A description of how to set a congestion price for networked applications is presented in [16], and a realistic system built on this premise is proposed in [17].

III. ANALYTIC FRAMEWORK OF SERVICE DELIVERY PLATFORM

The analytic foundation for our service delivery platform comes from the merger of the key methodologies described in the previous section and the concept of network utility maximization [18].

Consider a service-oriented network with resources that consist of intermediaries and providers, denoted by \( J = 1, 2, \ldots, J \). Let \( c_j \) be the capacity of resource \( j \in J \) and \( c = [c_1, c_2, \ldots, c_J]^T \). Let \( S = 1, 2, \ldots, S \) be the set of sources (consumers). Each source \( s \) has \( K^s \) available loop-free paths from the source to the logical destination node corresponding to the semantic service that is being consumed by a source. Let \( H^s \) be a \( J \times K^s \) \( 0 \times 1 \) matrix that describes the mapping of resources on paths for particular sources; that is,

\[
H^s_{ji} = \begin{cases} 1, & \text{if path } i \text{ of source } s \text{ uses resource } j \\ 0, & \text{otherwise} \end{cases}
\]

Let \( \mathcal{H}^s \) be the set of all columns of \( H^s \) that represent all available paths to source \( s \) under single-path routing. Define the \( J \times K \) matrix \( H \) as

\[
H = [H^1, H^2, \ldots, H^S]
\]

where \( K = \sum_s K^s \). \( H \) defines the topology of the service-oriented network.

Let \( w^s \) be a \( K^s \times 1 \) vector where the \( i \)th entry represents the fraction of \( s \)'s flow on its \( i \)th path such that

\[
w^s_i \geq 0 \quad \forall i, \quad \text{and} \quad 1^T w^s = 1
\]

where \( 1 \) is a vector of an appropriate dimension with the value 1 in every entry. We allow \( w^s \in [0, 1] \) for multipath routing. Collect the vectors \( w^s, s = 1, \ldots, S \) into a \( K \times S \) diagonal matrix \( W \). Let \( \mathcal{W} \) be the set of all such matrices corresponding to multipath routing as

\[
\mathcal{W} = \{ W \mid W = \text{diag}(w^1, \ldots, w^S) \in [0, 1]^{K \times S}, 1^T w^s = 1 \}
\]

As mentioned above, \( H \) defines the set of loop-free paths available to each source, and also represents the network topology. \( W \) defines how the sources split the load across the multiple paths. Their product defines a \( J \times S \) routing matrix \( R = HW \) that specifies the fraction of \( s \)'s flow at each resource \( j \). The set of all multipath routing matrices is:

\[
\mathcal{R} = \{ R \mid R = HW, W \in \mathcal{W} \}
\]

A multipath routing matrix in \( \mathcal{R} \) is one with entries in the range \([0, 1]\):

\[
R_{js} = \begin{cases} > 0, & \text{if resource } j \text{ is in a path of source } s \\ = 0, & \text{otherwise} \end{cases}
\]

The path of source \( s \) is denoted by \( r^s = [R_{1s}, \ldots, R_{Js}]^T \), the \( s \)th column of the routing matrix \( R \).

We wish to consider the following optimization problem:

\[
\max_{R \in \mathcal{R}} \max_{x \geq 0} \sum_{s \in \mathcal{S}} U_s(x_s) - \sum_{f \in F_s} f(x_s, \gamma_f, z_f)
\]

s.t. \( Rx \leq c \) \hspace{1cm} (1)

(1) optimizes “social welfare” by maximizing utility over both source rates and routes. However, (1) is not a convex problem because the feasible set specified by \( Rx \leq c \) is generally not convex. We now transform the problem by defining the \( K^s \times 1 \) vectors \( y^s = x_s w^s \) and the problem becomes:

\[
\max_{y \geq 0} \sum_{s \in \mathcal{S}} U_s (1^T y^s) - \sum_{f \in F_s} f(1^T y^s, \gamma_f, z_f)
\]

s.t. \( Hy \leq c \) \hspace{1cm} (3)

Provided the functions \( U_s(\cdot) \) and \( f(\cdot) \) are strictly concave, this is a strictly concave problem with a linear constraint, and therefore has no duality gap [19]. While we do not directly solve or present an algorithm for solving (3) & (4) due to space restrictions, distributed solutions to network utility maximization problems are well known in the literature and should be easily adaptable to our problem [18], [20], [21].

IV. RELATED WORK IN SERVICE SYSTEMS

There are several previous attempts that have been made towards developing brokered architectures that connect service consumers to service providers. Several proposals have been made to create “service overlay networks” with the intent of applying advances in overlay network research to the services layer. In [22], an open service market architecture is presented that aims to balance load across multiple service providers by using a network of proxies configured by an external centralized “trader” that computes the optimal routes for service requests. This architecture does not consider the current state of the proxies when making routing decisions. The authors of [23] propose a management overlay for Web Services based on interconnected service intermediaries, but do not address the service selection or routing problems.

Several previous efforts have focused on using overlay methodologies to provide better end-to-end quality of service for requests in the network by provisioning bandwidth or selecting the best path through the network based on available bandwidth [24]–[26]. The integration of bandwidth and other QoS metrics into optimizations in a service overlay network is presented in [27]. There have also been attempts to develop a service overlay network based upon network economics [28]. While the overarching goals of the work in this area are similar to ours, the work assumes that nodes of the service overlay network are inherently selfish and non-cooperative; this distinction has a dramatic effect on the underlying economic framework they create, thus making their work inapplicable to the problems we address. A useful review of brokered service-oriented systems is shown in [29].
Service selection algorithms utilize rational decision making processes that are used to decide which service instance to invoke according to some predefined criteria. A common component of such algorithms is the concept of a QoS registry [30]. A multi-agent approach to distributed service selection is proposed in [31]; however, the underlying transportation costs of the network are not considered in the model. A network-sensitive service selection algorithm is proposed in [32], but it does not incorporate the current state of the service providers or the intermediaries in the selection decision.

The concepts of brokered architectures and service selection are also addressed in the supply chain management literature. There is an increasing amount of literature discussing the application of multi-agent systems to dynamic supply chain management problems [33]. Transportation and handling costs in a graph-theoretic framework are integrated with traditional supply chain analysis in [5] and the references therein. A combined service selection and service pricing framework for supply chain managers is discussed in [34]. Distributed pricing issues in supply chains are addressed in [35].

V. CONCLUSIONS & FUTURE WORK

In this paper, we proposed a novel autonomic service delivery platform for service-oriented network environments. The framework of the platform is based on the methodologies of content-based routing, network economics, congestion pricing, and optimal routing and flow control. With a direct link to the business value derived from a service, the service delivery platform maximizes the value derived from underlying IT resources. We believe that our architecture provides exciting new multidisciplinary research opportunities in service engineering.

Some future issues to address include investigating efficient methods to estimate the derivatives of the congestion prices $f(x, \gamma, \nu, \frac{\delta f}{\delta x})$ in (1). Another concern exists with the assumptions regarding the concavity of utility functions; currently this assumption is required in order for welfare economic systems to converge. Further work similar to [36] is necessary to relax this assumption and broaden the applicability of our work to real-time services. Finally, we believe that further investigation into the interactions between autonomous systems could have important effects in business-to-business interactions in such an instantiation of our distributed service delivery platform.

REFERENCES