Utility Driven Service Routing over Large Scale Infrastructures*

Pablo Chacin[1], Leando Navarro[1], and Pedro Garcia Lopez[2]

[1] Departament d'Arquitectura dels Computadors,
Universitat Politénica de Catalunya, Barcelona, Spain
pchacin@ac.upc.edu
[2] Department d'Enginyeria Informàtica i Matemàtiques,
Universitat Rovira i Virgili, Tarragona, Spain

Abstract. In this paper we present UDON, a novel Utility Driven Overlay Network framework for routing service requests in highly dynamic large scale shared infrastructures. UDON combines an application provided utility function to express the services's QoS in a compact way, with an epidemic protocol to disseminate this information in a scalable and robust way. Experimental analysis with a simulation model suggests that the proposed overlay allocates requests to service instances that match their QoS requirements with a high probability and low overhead, adapting well to a wide variety of conditions.

1 Introduction

Recent years have witnessed the emergence of service oriented applications deployed on large scale shared infrastructures such as computational clouds [15]. Under this paradigm, services are reused and combined, new services are introduced frequently, usage patterns varies continuously and diverse user populations may exist, with different QoS requirements in terms of attributes like response time, execution cost, security and others [11].

As service instances run on non dedicated servers, it results difficult to predict the QoS that they can provide due to the fluctuations on servers' workload. Additionally, the number of instances may vary over time in response to variations in the service demand.

In such infrastructures, the allocation (or routing) of service requests to instances that match their QoS requirements becomes a significant challenge in terms of how to: a) route requests accurately despite the dynamism of the environment; b) scale up to a very large number of service instances; c) handle the churn of instances as they are activated and deactivated or fail; d) accommodate different resource management policies, as nodes may belong to different administrative domains; and e) clearly separate the from application specific aspects from routing, to offer a generic infrastructure that supports multiple services.

* This work has been partly supported by Spanish MEC grant TIN2010-20140-C03-01.

E. Di Nitto and R. Yahyapour (Eds.): ServiceWave 2010, LNCS 6481, pp. 88–99, 2010.

The main contribution of this paper is the proposal of UDON, a novel Utility Driven Overlay Network framework for routing service requests to service instances in highly dynamic large scale environments. UDON combines an application provided utility function to express the service's QoS, with an epidemic protocol for information dissemination.

The utilization of utility functions allows a compact representation of the QoS requirements for services and facilitates the comparison of the QoS that different instances provide, even if they run on very heterogeneous nodes [9]. The utilization of an epidemic style dissemination algorithm (on which information is randomly propagated among nodes) makes UDON highly adaptable to changes in the infrastructure, scalable, and resilient to failures and churn, maintaining its routing capabilities with a modest maintenance cost.

In this paper we evaluate extensively the proposed overlay using simulation. Experiments show that UDON can route requests with a high probability of success and in a low number of hops under a wide variety of conditions. In particular, it adapts well to random fluctuations in the QoS provided by service instances, high churn rates due to instance activations/deactivations, and the scarcity of instances that provide an adequate QoS.

The rest of this paper is organized as follows. Section 2 introduces the system model and describes the algorithms used for overlay construction and routing. Section 3 presents an evaluation of UDON under diverse conditions and discusses the results. Section 4 reviews relevant related work. Finally, section 5 presents the conclusions, open issues and planned future work.

2 The Utility Driven Overlay

In UDON (see fig. 1) each service has an utility function that maps the attributes and execution state of a service instance (e.g response time, available resources, trustworthiness, reliability, execution cost) to a scalar value that represents the QoS it provides.

Requests coming from users are processed through a set of entry-points, which correspond to segments of users with similar QoS requirements, and must be routed to service instances that offer an adequate QoS. The QoS required by a request is defined as the minimum acceptable utility that must be provided by a service instance to process it.

Entry points and service instances are organized in a routing overlay on which each node maintains a local view with the utility of a subset of other nodes, and use this information to route the requests. The main objective of the overlay is to help each node to maintain fresh information about other nodes, while addressing the challenges of scalability, resilience and adaptability of the environment.

Service instances are continuously activated/deactivated from/to a pool of available servers in response to fluctuations in the service demand using one of different approaches proposed elsewhere [10] [13]. The scope of the present work concentrates on how requests are routed to the active service instances.

The next sections describe in detail how the overlay is constructed and how requests are routed.