Service Placement in Network-aware Cloud Infrastructures

Andreas Reifert
Institute of Communication Networks and Computer Systems (IKR)
Universität Stuttgart, Pfaffenwaldring 47, D–70569 Stuttgart, Germany
andreas.reifert@ikr.uni-stuttgart.de

I. Extended Abstract

Cloud computing induces a major trend towards pushing application services into large runtime environments hosted at computing centers. While a clear definition of "the Cloud" still remains fuzzy at best, several Cloud service models emerge from different sides which include *Infrastructure as a Service*, *Platform as a Service*, and *Software as a Service* models [1]. They all build on virtualization technologies; they differ in the level at which virtualization is applied [2].

Virtualization decouples resource provision from resource usage. The *Cloud provider* profits from economy of scale effects by providing the resources to its many *Cloud customers*. They in turn profit from not having to invest in infrastructure anymore. Consequently, the hurdle for developing and instantiating new Internet services is lowered. Furthermore, the customers can become more easily *service providers* themselves with an own dynamic *service user* base.

Typical application services are request/response or transactions based server applications following an n-tier architecture. Current Cloud offerings (like EC2, S3, AppEngine) focus on hosting these kind of IT-services on servers within one of several geographically distributed computing centers. Inside the computing centers communication delays can be assumed small, bandwidth abundant enough. Consequently, Cloud providers consider network aspects like bandwidth on an ingress/egress basis only. And they cannot give guarantees on reachability and communication delays outside their centers.

This limits new Internet services in their form. Service providers cannot let the Cloud distribute the service across several computing centers. Distributed services with hard real-time requirements or services with a distributed service user base that require short response times have this requirement, though. Integrating an accurate network view into the Cloud management in order to support these types of services would make the Cloud more flexible.

It also increases the efficient use of bandwidth resources. We study the benefits of such an extended Cloud within a detailed service and network model: a service consists of different separable parts (*components*) with communication requirements on the *connections* between them. Some of the components should have fixed locations within the Cloud; the remaining ones can have arbitrary ones. Intelligent placement of the components alone can save bandwidth resources as back and forth communication is avoided. We can achieve this without modifying any routing mechanism and route the connections using standard methods.

The same considerations also apply for services within a large testbed like the GENI project [3]. The authors of [4] study such a scenario but focus on optimizing connections through reconfigurable multipaths. Our work focuses on the components' optimal placement of a distributed service.

We are currently comparing different placement algorithms for placing components on heterogeneous network nodes with limited resources and heterogeneous links between them. In particular, we quantify through Monte-Carlo simulations how the node/link infrastructure profits from good placement heuristics. Current results indicate the fraction of new services not admitted can drop by at least one magnitude, and resource consumption can drop by about one magnitude. Thus the utilization of the infrastructure increases significantly by just choosing good locations for components.

Taking the service's topology into account is the major contributing factor. At the current stage we can only give rough guidelines what aspects to pay attention to. To the best of our knowledge no detailed studies have been conducted on this what we call the *Topology Placement Problem*. Closest comes a study [5] where the authors call it *Network Testbed Problem* and restrict it to LAN specifics. In our opinion it should be extended to Internet scale networks. Our model and evaluation framework is a good starting point to fill the missing pieces relevant for placing Future Internet services.

References

[1] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: Towards a cloud definition," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, pp. 50–55, Jan. 2009.

[2] S. K. Nanda and T.-c. Chiueh, "A survey on virtualization technologies," Department of Computer Science, State University of New York, Stony Brook, Stony Brook, NY 11794-4400, TR 179, Feb. 2005.

[3] "The Global Environment for Network Innovations (GENI)," Apr. 2009. [Online]. Available: http://www.geni.net/wp-content/uploads/2009/04/geni-at-a-glance-final.pdf

[4] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: Substrate support for path spliting and migration," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 17–29, Apr. 2008.

[5] R. Ricci, C. Alfeld, and J. Lepreau, "A solver for the network testbed mapping problem," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 2, pp. 65–81, Apr. 2003.