A Model for Resource Sharing for Internet Data Center Providers within the Grid

Jörn Altmann
Department for Computer Networks and Distributed Systems
International University in Germany
76646 Bruchsal, Germany
jorn.altmann@acm.org

Abstract

Internet data center providers are still struggling to lower the operational costs of their data centers. One reason is the low utilization of servers over a long period of time during the day. The paper describes a system for optimizing the server resources within Internet data centers, which host different services such as web servers or enterprise resource planning systems. The system, called resource management system, allows Internet data center providers to allocate their resources in an economically efficient way. The results may indicate that there is free capacity or a lack of capacity. Based on the results, the resource management system can sell or purchase resources on the Grid. The idea behind this approach is to enable Internet data center providers to gradually transition from the current environment to an environment where utility computing is possible. Our approach separates between the local resource allocation and the external one (Grid).

Keywords

Market-Managed Resource Allocation, Internet Data Centers Management, Economics

1. Introduction

Data centers, which are managed by a variety of companies (e.g. IBM, SUN, AT&T, Sprint, HP, EDS, Verio (NTT), MCI, as well as smaller service provider), enable firms to outsource their IT services, such as web servers and ERP applications [2]. The compelling value proposition of Internet data centers (IDC) is the more efficient use of resources and flexible access to computational resources. IDC customers expect lower costs and better services [10].

The problems that data center operators are facing are to provide quality service at low cost. In order to solve this problem, they have to consider not only the technical requirements of applications [8] [9] but also the demand of applications for resources. However, the solutions proposed in literature assume only static demand of applications for computing and network resources if at all [5] [6] [7]. Apart from one work on network planning [1], the existing resource allocation approaches do not take into account the overall differences in demand between IDC customers. They also do not consider the change in demand of IDC customer in the course of a day.

If Internet data center provider would understand the demand of their customers, they would be able to optimize the allocation of servers with higher precision. If the optimization results in spare capacity, it could either be sold to other customers or be offered to other Internet data centers, which are in need for additional computational resources. Although this is currently not possible, the Grid could provide the environment for such a scenario. Internet data center providers could offer the spare resources for sale in the Grid. Other Internet data centers that are connected to the Grid would be able to easily purchase these resources. The Grid, therefore, will enable Internet data center provider to fully utilize their resources by providing an environment for selling spare resources.

2. Internet Data Center

2.1 Organization of Internet Data Centers

Currently, Internet data center comprise up to 10000 servers. In the future, the number of servers is expected to increase even further. The servers are hierarchically organized in order to ease management and operation of the Internet data center. The servers are grouped into clusters, which hold 100 servers (see Figure 1). Clusters are grouped into meta-clusters of 10 units currently [3] [4].

Although these attempts of simplifying the management of servers in the Internet data center, the management of each servers is still done manually. System administrators are observing different performance metrics of a server in order to find problems. Help is only provided in form of reporting tools, which show the behavior of all servers [9]. Based on this information, the capacity planning team of an Internet data center provider evaluates the capacity utilization and makes purchase decisions for new servers and network and storage components (considering the budget and financial situation). In general, current research only focuses on supporting system administrators in managing Internet data centers [3] [11].

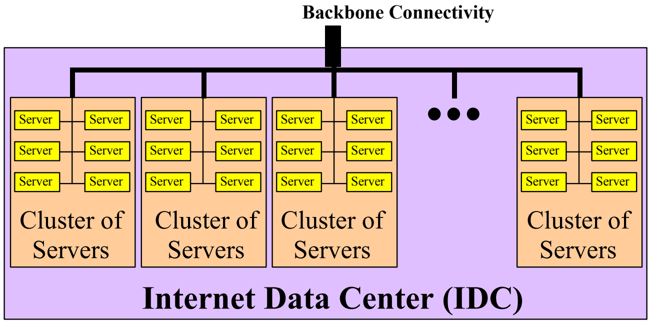

Figure 1. Clustering of Servers within an Internet Data Center

2.2 Freeing Resources within Internet Data Centers

Higher cluster utilization could be achieved by finding a group of IDC customers, whose total demand (i.e. sum of their separate demands) is almost the same in the course of a day. It reduces the overall demand fluctuations, avoiding under-utilization or over-utilization of servers at certain points in time. Figure 2 shows an example for this. The change in demand of three IDC customers in the course of the day is less than the change of demand of each IDC customer separately.

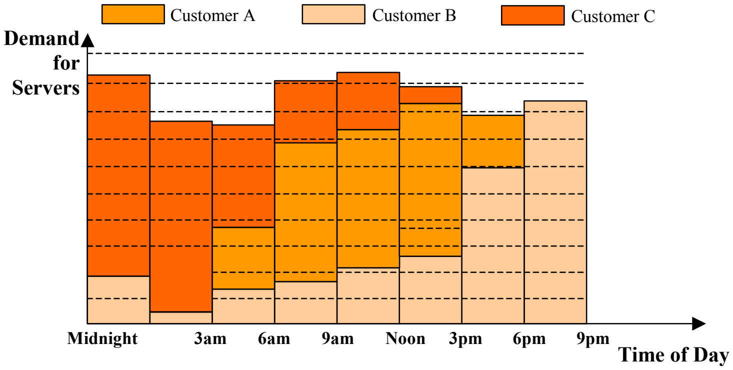

Figure 2. Example of three IDC customers' Demand over the Course of a Day

For example, assume a group of IDC customers, such as an online broker (e.g., Charles Schwab), an Internet retailer (e.g., Amazon.com), and an ERP application provider (e.g. SAP). While the peak demand of the online broker is during the opening hours of the stock market (9 a.m. to 4 p.m.), Internet retailing activity is mainly taking place in the after work hours. The ERP application provider could offer batch mode services to its customers during the night hours. Adding up these demands, the total demand could almost be the same during the entire day. If the IDC can find such customer groups, resources would be freed.

2.3 Market-Managed Resource Allocation

The example given in the previous section is implemented using a new resource allocation mechanism. This mechanism considers not only the individual demand of an application of a customer but also:

- the revenue generated by hosting the application of a certain customer,
- the long-term value of the customer, and
- the utility / demand function of the customer.

Since the optimization relies on these economic indicators, the allocation of server resources is market-managed [1].

The revenue parameter will help evaluating how business-critical the performance of the hosting service is. If the revenue generated is high, the application will have high priority to get further servers assigned under high utilization. If the revenue is low, the application has to tolerate lower performance under heavy load.

The long-term value of a customer to the company could be used to over-write the resource allocation decision based on the revenue parameter. If the customer is very important although the revenue from the hosting service of his application is low, the customer should get anyway high priority service.

Providing economic incentives to IDC customers can help to avoid times of heavy load. Customers that are offered to pay a lower price for the hosting service might be willing to run their calculations during off-peak times. Customers that are not flexible regarding the time have to pay the full price. This method helps to spread the total demand of all customers over the course of the day.

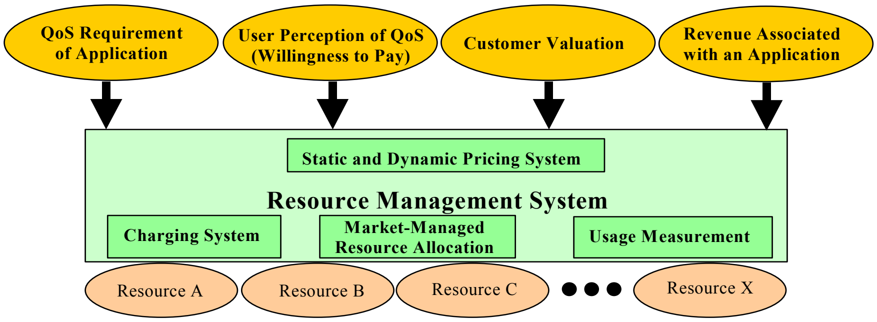
QoS Requirement of Application
User Perception of QoS (Willingness to Pay)
Customer Valuation
Revenue Associated with an Application
Static and Dynamic Pricing System
Resource Management System
Charging System
Market-Managed Resource Allocation
Usage Measurement
Resource A
Resource B
Resource C
Resource X

Figure 3. Resource Management System

The resource management system needs input about the QoS requirements of the applications, the revenue received from hosting the application, the customer valuation, and the customer's perception of QoS. The output is the allocation of applications on available physical servers (indicated as Resource in Figure 3). The resource management system itself consists of a module for accounting and charging, usage monitoring, and pricing. These modules will record the usage of resources and the revenue generated, and will calculate the price for services.

3. Interconnection between Internet Data Centers in the Grid

For selling and purchasing server resources on the Internet, the resource management system uses the component Grid Resource Agent. This component maintains a virtual resource. The virtual resource represents a collection of all local resources that are sold or are available for purchase in the Grid. The virtual resource also represents a resource that the resource management system can access, in order to meet the demand of local applications. An example of this system is shown in Figure 4.

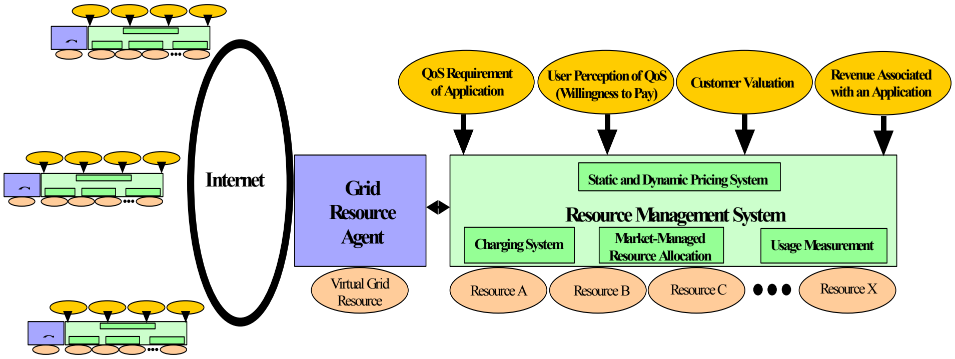
Internet
Grid Resource Agent
Virtual Grid Resource
QoS Requirement of Application
User Perception of QoS (Willingness to Pay)
Customer Valuation
Revenue Associated with an Application
Static and Dynamic Pricing System
Resource Management System
Charging System
Market-Managed Resource Allocation
Usage Measurement
Resource A
Resource B
Resource C
Resource X

Figure 4. Grid Resource Agent

For example, if the resource management system notices that additional resources are needed, it sends a request for purchasing the required resources on the Grid to the Grid resource agent. The Grid resource agent initiates request for services and prices to the Grid resource agents of other Internet data centers. After analyzing the incoming offers and choosing the most appropriate one, the Grid resource agent signs the service contract. The resource management system gets informed about the availability of the virtual resource and that it can access it via the Grid resource agents, that hides the actual location of the resource. In case the demand within a data center is lower than the available resources, the excess capacity can be made available to the Grid. For this purpose, the resource management system informs the Grid resource agent about the service parameters and price. The Grid Resource agent offers the service and waits for requests from other Internet data centers.

4. Conclusion

We presented an idea how demand analysis and market-managed principles can be used to allocate server resources economically efficient in Internet data centers. For this purpose, we suggested to consider economic indicators, such as the revenue received from hosting the application, the customer valuation, and the customer's perception of QoS. In a second step, we suggested to let Internet data center trade available resources within the Grid. Beside increasing the utilization of servers, it will be the first step towards pure utility computing.

5. References

[1] S. Jagannathan, J. Altmann, and L. Rhodes, "A Revenue-Based Model for Making Resource Investment Decisions in IP Networks," IFIP/IEEE IM2003, Integrated Symposium on Integrated Network Management, Colorado Springs, Colorado, USA, March 2003.

[2] J. Altmann, "A Reference Model of Internet Service Provider Businesses," ICTEC2000, 3rd International Conference on Telecommunication and Electronic Commerce, Dallas, Texas, USA, November 2000.

[3] S. Graupner, V. Kotov, H. Trinks, "A Framework for Analyzing and Organizing Complex Systems," ICECCS 2001, the Seventh IEEE International Conference on Engineering of Complex Computer Systems, Sweden, June 2001.

[4] Rolia, J. Singhal, S., Friedrich, R., "Adaptive Data Centers," Proceedings of SSGRR, 2000.

[5] Jose Renato Santos, Koustuv Dasgupta, G. (John) Janakiraman and Yoshio Turner, "Understanding Service Demand for Adaptive Allocation of Distributed Resources,". IEEE Globecom Global Internet Symposium, November 2002.

[6] Karen Appleby, Tamar Eilam, Liana L. Fong, German Goldszmidt, Michael H. Kalantar, "Resource Model for Self-Managing Computing Utility Services" in RC22377.

[7] Liana Fong, Michael Kalantar, Donald Pazel , Germán Goldszmidt, Karen Appleby, Tamar Eilam, Sameh Fakhouri, Srirama Krishnakumar, Sandra Miller, John Pershing, "Dynamic Resource Management in an eUtility" Proceedings of NOMS 2002 IEEE/IFIP Network Operations and Management Symposium.Piscataway, NJ, p.727-40, 2002.

[8] S. Ranjan and J. Rolia and H. Fu and E. Knightly, "QoS-Driven Server Migration for Internet Data Centers"

[9] Ted Buell, Matt Edwards, Nicole Forlano, Kevin Stansfield, Steve Patek, Adam Joseph, and Greg Gum, "Enhancing Data Center Asset Allocation and Information Management Processes,"

[10] Chris Kenyon, Giorgos Cheliotis, "Grid Resource Commercialization: Economic Engineering and Delivery Scenarios," in: Grid Resource Management: State of the Art and Research Issues, Editors: J. Nabrzyski, J. Schopf and J. Weglarz, Kluwer, 2003.

[11] Sven Graupner, Vijay Machiraju, Akhil Sahai, Aad van Moorsel, "Management += Grid," DSOM, Distributed Systems: Operations and Management Workshop, Heidelberg 2003.