POLITECNICO DI MILANO

Scuola di Ingegneria dei Sistemi
Corso di Laurea Magistrale in Ingegneria Matematica
Dipartimento di Matematica

Service Provisioning Problem in Cloud and Multi-Cloud Systems: a Generalized Nash Equilibrium model

Relatore:
Prof. Danilo ARDAGNA - Politecnico di Milano
Correlatore:
Prof. Mauro PASSACANTANDO - Università di Pisa

Tesi magistrale di:
Anna SAVI - Matricola 762075

Anno Accademico 2011-2012

Contents

List of Figures

List of Tables

Abstract

In recent years the evolution and the widespread adoption of virtualization, service-oriented architectures, autonomic and utility computing have converged letting a new paradigm to emerge: The Cloud Computing. Cloud Computing aims at streamlining the on-demand provisioning of software, hardware, and data as services, providing end-user with flexible and scalable services accessible through the Internet. Since the Cloud offer is currently becoming wider and more attractive to business owners, the development of efficient resource provisioning policies for Cloud-based services becomes increasingly challenging. Indeed, modern Cloud services operate in an open and dynamic world characterized by continuous changes where strategic interaction among different economic agents takes place.

This thesis aims to study the hourly basis service provisioning and capacity allocation problem through the formulation of a mathematical model based on noncooperative-game-theoretic approach. We take the perspective of Software as a Service (SaaS) providers which want to minimize the costs associated with the virtual machine instances allocated in a multi-IaaSs (Infrastructure as a Service) scenario, while avoiding incurring in penalties for requests execution failures and providing quality of service guarantees. SaaS providers compete and bid for the use of infrastructural resources, while the IaaSs want to maximize their revenues obtained providing virtualized resources.

The problem has been modeled as a Generalized Nash Equilibrium Problem (GNEP). Thanks to a deep analysis of the game under study, we demonstrate the social equilibria existence for the corresponding generalized potential game. The best-reply solution is pursued heuristically with the implementation of two different algorithms suitable for a distributed implementation.

We demonstrate the effectiveness of our approach by performing numerical analyses, considering multiple workloads and system configurations. Results show that our algorithms are scalable. Equilibria efficiency is quantitatively analysed in terms of Price of Anarchy (PoA) and Individual Worst

Case (IWC) for both SaaSs and IaaSs, indicating the advantage of solving the equilibrium problem with the second approach (improvements by up tp 40%). Furthermore, a multiple IaaS analysis points out the SaaS benefits in exploiting multiple IaaS deployment of applications and redistribution of traffic simultaneously: in terms of efficiency, performance improves by 40%, while, in terms of costs, SaaSs average saving ranges from 10% up to 40% compared to single IaaS architectures.

Sommario

Negli ultimi anni l'evoluzione e la diffusa adozione di virtualizzazione, architetture orientate ai servizi, autonomic e utility computing sono confluiti in un nuovo paradigma emergente: il Cloud Computing. Il Cloud Computing mira a semplificare la fornitura on-demand di software e hardware, fornendo agli utenti finali servizi flessibili e scalabili accessibili tramite Internet. Poiché ultimamente l'offerta Cloud si sta ampliando, diventando più attrattiva per le aziende, lo sviluppo di politiche efficienti per la distribuzione delle risorse per servizi basati sul Cloud diventa sempre più complesso. Infatti, le moderne infrastrutture Cloud operano in un mondo dinamico e caratterizzato da continui cambiamenti e interazioni tra diversi concorrenti.

L'obiettivo di questa tesi è lo studio del problema della fornitura oraria di servizi e l'allocazione di risorse attraverso la formulazione di un modello matematico basato sui concetti della teoria dei giochi non cooperativi. Viene considerata la prospettiva dei fornitori di Sofware as a Service (SaaS) che vogliono minimizzare i costi associati alle macchine virtuali allocate in uno scenario multi-IaaS (Infrastructure as a Service), evitando sanzioni causate dalla mancata esecuzione delle richieste e garantendo la qualità del servizio. I SaaS competono fra loro facendo offerte per l'utilizzo delle infrastrutture, mentre gli IaaS vogliono massimizzare i propri introiti ottenuti fornendo risorse virtualizzate.

Il modello è costituito da un problema di equilibrio di Nash generalizzato. Grazie ad un'accurata analisi del problema studiato, dimostriamo l'esistenza di equilibri sociali per il corrispondente gioco generalizzato a potenziale. La soluzione del best-reply è ottenuta euristicamente con l'implementazione di due differenti algoritmi distribuiti.

L'efficacia del nostro approccio è stata dimostrata compiendo test numerici, considerando differenti configurazioni di carico e di sistema. I risultati mostrano che gli algoritmi sono scalabili. L'efficienza degli equilibri è quantitativamente analizzata in termini di Price of Anarchy (PoA) e Individual Worst Case (IWC) di SaaS e IaaS, indicando il vantaggio nella risoluzione del problema con il secondo approccio (ottenendo miglioramenti fino al 40%).

Inoltre, l'analisi multi-IaaS evidenzia i benefici dei SaaS nello sfruttare il dislocamento delle applicazioni e la distribuzione del traffico su molteplici IaaS: in termini di efficienza, le performance migliorano del 40%, in termini di costi, il risparmio medio per i SaaS oscilla tra il 10% e il 40% rispetto al caso mono IaaS.

Chapter 1

Introduction

Cloud computing has been a dominant ICT news topic over the past few years. It is essentially a way for ICT companies to deliver software/hardware on-demand as services through the Internet. Cloud computing applications are generally priced on a subscription model, so end-users may pay a yearly usage fee, for example, rather than the more familiar model of purchasing software licenses. The Cloud-based services are not only restricted to software applications (Software as a Service – *SaaS*), but could also be the platform for the deployment and execution of applications developed in house (Platform as a Service – *PaaS*) and the hardware infrastructure (Infrastructure as a Service – *IaaS*).

In the SaaS paradigm, applications are available over the Web and provide Quality of Service (*QoS*) guarantees to end-users. The SaaS provider hosts both the application and the data, hence the end-user is able to use and access the service from all over the world. With PaaS, applications are developed and deployed on platforms transparently managed by the Cloud provider. The platform typically includes databases, middleware, and also development tools. In IaaS systems, virtual computer environments are provided as services and servers, storage, and network equipment can be outsourced by customers without the expertise to operate them.

Many companies, e.g. Google, Amazon, and Microsoft are offering Cloud computing services such as Google's App Engine, Amazon's Elastic Compute Cloud (EC2) or Microsoft Windows Azure. Large data centers provide the infrastructure behind the Cloud and virtualization technology makes Cloud computing resources more efficient and cost-effective both for providers and customers. Indeed, end-users obtain the benefits of the infrastructure without the need to implement and administer it directly adding or removing capacity almost instantaneously on a "pay-as-you-use" basis. Cloud providers can, on the other hand, maximize the utilization of their physical resources also

obtaining economies of scale.

The development of efficient *service provisioning* policies is among the major issues in Cloud research. Indeed, modern Clouds operate in an open and dynamic world characterized by continuous changes which occur autonomously and unpredictably. Moreover, the rapid growth of the Internet and the late problems arising in the ICT industry, such as resource or quality of service, pricing, and load shedding, has led to a very complex interaction between all the involved competitors.

In this context, *Non Cooperative Game Theory models* and approaches allow to gain an in-depth analytical understanding of the service provisioning problem. Game Theory has been successfully applied to diverse problems such as Internet pricing, flow and congestion control, routing, and networking. One of the most widely used "solution concept" in Game Theory is the *Nash Equilibrium* approach: a set of strategies for the players constitute a Nash Equilibrium (NE) if no player can benefit by changing his/her strategy unilaterally or, in other words, every player is playing a *best response* to the strategy choices of his/her opponents.

In this thesis we take the perspective of SaaS providers which host their applications at multiple IaaS providers. Each SaaS provider wants to minimize the cost of use of Cloud resources and penalties for requests execution failures. The cost minimization is challenging since on-line services receive dynamic workloads that fluctuate over multiple time scales. Resources have to be allocated flexibly at run-time according to workload fluctuations. Furthermore, each SaaS behaves selfishly and competes with others SaaS for the use of infrastructural resources supplied by the IaaS. Each IaaS, in his turn, wants to maximize the revenues obtained providing the resources.

To capture the behavior of SaaSs and IaaSs in this conflicting situation in which the best choice for one depends on the choices of the others, we recur to the *Generalized Nash Equilibrium* (GNE) concept, which is an extension of the classical Nash equilibrium [75]. We then use Non Cooperative Game Theory results to develop efficient algorithms for the run-time management and allocation of IaaS resources to competing SaaSs, suitable also for a fully distributed implementation. Different solutions achieving generalized equilibria are proposed and evaluated in terms of their *efficiency* with respect to the *social optimum* of the Cloud. We demonstrate the effectiveness of our approach by performing a large set of numerical analyses.

The thesis is organized as follows:

- **Chapter 2**:
  We firstly give a general overview on Cloud Computing providing defini-

tions, identifying the main characteristics and illustrating the different structures models available. A discussion on today's run-time research challenges is followed by the analysis and the classification of the literature approaches. The state of the art is given in terms of type of problem, solution found and discipline adopted, as well as according to the approach used.

Secondly, we discuss the Game Theory notions and tools necessary to study and analyse our service provisioning problem. We focus the attention on the solution notion of GNE discussing the equilibria existence problem and providing the definition of potential games.

- **Chapter 3**:
  We present the multi-Cloud service provisioning problem by means of the formulation of a mathematical model. The discussion starts introducing the problem and the design assumptions, defining the variables and parameters. Then the formulated generalized Nash game is studied. Thanks to an accurate analysis of the constraints and to the illustration of important equilibria properties, we identify a potential function of the game and provide a solution existence theorem. In order to find an equilibrium of the game under study we make use of heuristic techniques which are formalized with two algorithms based on best reply approach. Finally, an extension of the model which includes an additional type of VMs is considered.

- **Chapter 4**:
  After specificating the tools used to model our problem (the modeling language for mathematical programming AMPL and the solver CPLEX), the experimental results obtained implementing and testing the two algorithms for a variety of system and workload configurations are presented in order to assess the quality of our solutions.

  The design of experiments (DoE) describes the simulation data used and the main features of our numerical experiments. The scalability analysis is evaluated considering the execution time required by our algorithms to solve the problem for instances of different sizes. The equilibria efficiency is measured in terms of two indicators: the Price of Anarchy (PoA) and the Individual Worst Case (IWC). Then, we analyse and compare the results which can be achieved by using single and multiple IaaSs; the evaluation of the advantages/disadvantages of the different cases is given both in terms of efficiency (PoA and IWC) and of average fee to be paid by SaaSs.

- **Chapter 5**:
  Finally, we draw the work's conclusions summarizing the steps of our study, pointing out the aims of our analysis and underling the acheived results. Future research directions are also presented.

Chapter 5

Conclusions and future works

In this work we discussed a multi-Cloud service provisioning problem through the formulation and the study of a Generalized Nash Equilibrium (GNE) model.

Since in all the most significant fields of application of today's society, dynamic systems are required to provide services and applications that are more competitive, more scalable and more responsive with respect to the classical systems, the new Cloud Computing paradigm is getting more and more popular, and Cloud-based services and resource provisioning become more and more challenging. In particular, in any time instant resources have to be allocated to handle effectively workload fluctuations of great diversity and enormous scale, while providing quality of service (QoS) guarantees to the end users.

In such a context, the costs and resources allocation optimization problem is a central topic from both customer's and provider's perspective and its social, economic and strategic structure can be perfectly reproduced with Noncooperative Game Theory tools such as the GNE Problem.

Within our work we proposed a GNE model, taking the perspective of SaaS providers. The overall goal we addressed is the minimization of the costs associated with the virtual machine instances allocated in a multi-IaaSs scenario (IaaS providers maximizing their revenues), while guaranteeing QoS constraints. To achieve this purpose we considered the problem of run-time management of IaaSs provider capacities among multiple competing SaaSs. The cost model consists of objective functions which includes revenues and penalties incurred depending on the achieved performance level and infrastructural costs associated with IaaSs resources.

Thanks to the deep game and constraints analysis, the formalization of an optimization method for IaaSs game and the formulation of some equilibria

properties, we provided a better understanding of the mathematical formulation. The most important analytical result we reached was the demonstration of the existence of social equilibria for the game under study. The theorem we proposed classified the problem under study as a generalized potential game and an accurate analysis of the potential function and of the feasible set was illustrated in the proof.

Therefore, we presented the solution techniques used to solve the problem under analysis. The best reply for each player was found by means of an heuristic method that was formalized with the implementation of two different algorithms.

We discussed the numerical results of our proposed solutions - achieved with AMPL language and the CPLEX solver - in order to assess the effectiveness of our approach, performing a wide set of analyses which considered multiple workloads and system configurations. Realistic workloads created from a large website statistics and performance parameters estimated on an industrial benchmarking application deployed in the Cloud were used.

Scalability analysis showed that systems up to thousands of applications can be managed very efficiently in a fully distributed manner. Since the computation times required to solve problem instances of the largest size ranged from few tenths of a second to 2-5 seconds, we could state that the execution time requested is compatible with the time scales considered: the algorithms for finding a solution with the two proposed methods were suitable to determine the resource provisioning of very large Cloud infrastructures at run-time on a hourly basis without introducing any system overhead.

Furthermore, the efficiency of the equilibria achieved by our approaches was evaluated qualitatively in terms of an upper bound of the Price of Anarchy (PoA) and of the Individual Worst Case (IWC) for SaaSs and IaaSs. Indeed, since the nature of our problem did not allow us to calculate the value of the social equilibria, we compared the equilibrium solution found using the best-reply method with the optimal solution of the relaxed social problem. On average the percentage difference of the sum of the payoff functions with respect to the social optimum was lower than 70%. In terms of efficiency the second algorithm proposed guaranteed better performance than the first one. SaaS providers improvement by 37% indicated the advantage of solving the equilibrium problem with the second approach. Moreover, the independence of the results of the three indicators from the size of the problem, indicated the robustness of the two implemented methods.

Finally, we performed a multiple IaaS analysis, comparing the results of the experiments that considered SaaSs hosting applications on 1, 2, or 3 IaaSs, and fixing the fraction of reserved VMs as 60-70% of the total number

available. The evaluation of the advantages/disadvantages of the different scenarios was given both in terms of efficiency (PoA and IWC) and of the average fee to be paid by SaaSs. The analysis demonstrated the benefit from the SaaS point of view in exploiting multiple IaaS deployment of applications and redistribution of traffic simultaneously. Indeed, in terms of efficiency the multi-IaaS scenario can improve performance by 40%, while, in terms of costs, SaaSs can have significant average savings ranging from 10% up to 40% compared to single IaaS architectures.

An extension of the model which considered an additional type of VMs was considered. Since it was possible to guarantee the equilibria existence only adding simplifying assumptions, the general solution of this new problem formulation can be object of study for future works.

Moreover, since with the tools and approaches we developed we could calculate only the optimal solution of the relaxed social problem, another direction for future study can be testing the model for solvers that handle non linear problem with continuous and binary variables in order to reach the real value of the social optima. The efficiency analysis of the results can be performed again, leading to more quantitatively precise conclusions.

Future work will also be devoted to a deeper investigation of the time scales which can be adopted to govern the behavior of Cloud systems, performing resource allocation also every few minutes.

Bibliography

[1] Amazon Elastic Cloud Computing. http://aws.amazon.com/ec2/.

[2] Amazon Web Services. http://aws.amazon.com/.

[3] AMPL. Ampl modeling language for mathematical programming. http://www.ampl.com/.

[4] Flexyscale. http://www.flexiscale.com/.

[5] GoGrid. http://www.gogrid.com/.

[6] Google App Engine. https://developers.google.com/appengine/.

[7] Google Apps for Business. http://www.google.com/enterprise/apps/business/.

[8] Google Compute Engine. https://cloud.google.com/products/compute-engine.

[9] Google Inc. http://www.google.com/about/company/.

[10] IBM ILOG CPLEX Optimizer. http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

[11] Kernel Based Virtual Machine. http://www.linux-kvm.org/.

[12] Microsoft Corporation. http://www.microsoft.com/.

[13] Microsoft Office 365. http://office365.microsoft.com/.

[14] Microsoft Windows Azure. http://www.windowsazure.com/.

[15] Microsoft Windows Azure Virtual Machines. http://www.windowsazure.com/en-us/home/features/virtual-machines/.

[16] MODAClouds. http://www.modaclouds.eu/.

[17] Onlive. http://www.onlive.com/.

[18] Rackspace. http://www.rackspace.com/.

[19] Salesforce. http://www.force.com/.

[20] SAP. http://www.sap.com/.

[21] VMware Inc. http://www.vmware.com/.

[22] Xen Hypervisor. http://www.xen.org/.

[23] *2011 International Conference on Distributed Computing Systems, ICDCS 2011, Minneapolis, Minnesota, USA, June 20-24, 2011*. IEEE Computer Society, 2011.

[24] *INFOCOM 2011. 30th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 10-15 April 2011, Shanghai, China*. IEEE, 2011.

[25] *2012 IEEE 32nd International Conference on Distributed Computing Systems, Macau, China, June 18-21, 2012*. IEEE, 2012.

[26] *26th IEEE International Parallel and Distributed Processing Symposium Workshops & PhD Forum, IPDPS 2012, Shanghai, China, May 21-25, 2012*. IEEE Computer Society, 2012.

[27] Vineet Abhishek, Ian A. Kash, and Peter Key. Fixed and market pricing for cloud services. *CoRR*, abs/1201.5621, 2012.

[28] A. Aleti, B. Buhnova, L. Grunske, A. Koziolek, and I. Meedeniya. Software architecture optimization methods: A systematic literature review.

[29] Jussara Almeida, Virgílio Almeida, Danilo Ardagna, ítalo Cunha, Chiara Francalanci, and Marco Trubian. Joint admission control and resource allocation in virtualized servers. *J. Parallel Distrib. Comput.*, 70(4):344–362, April 2010.

[30] Eitan Altman, Thomas Boulogne, Rachid El Azouzi, Tania Jiménez, and Laura Wynter. A survey on networking games in telecommunications. *Computers & OR*, 33:286–311, 2006.

[31] Anatoly Antipin. Differential equations for equilibrium problems with coupled constraints. *Nonlinear Analysis-Theory Methods and Applications*, 47(3):1833–1844, 2001.

[32] D. Ardagna, E. di Nitto, P. Mohagheghi, S. Mosser, C. Ballagny, F. D"Andria, G. Casale, P. Matthews, C.-S. Nechifor, D. Petcu, A. Gericke, and C. Sheridan. Modaclouds: A model-driven approach for the design and execution of applications on multiple clouds. In *Modeling in Software Engineering (MISE), 2012 ICSE Workshop on*, pages 50–56, 2012.

[33] Danilo Ardagna, Sara Casolari, Michele Colajanni, and Barbara Panicucci. Dual time-scale distributed capacity allocation and load redirect algorithms for cloud systems. *J. Parallel Distrib. Comput.*, 72(6):796–808, 2012.

[34] Danilo Ardagna, Sara Casolari, and Barbara Panicucci. Flexible distributed capacity allocation and load redirect algorithms for cloud systems. In Liu and Parashar [68], pages 163–170.

[35] Danilo Ardagna, Barbara Panicucci, and Mauro Passacantando. A game theoretic formulation of the service provisioning problem in cloud systems. In Srinivasan et al. [84], pages 177–186.

[36] Danilo Ardagna, Barbara Panicucci, and Mauro Passacantando. Generalized nash equilibria for the service provisioning problem in cloud systems. *Services Computing, IEEE Transactions on*, PP(99):1, 2012.

[37] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski, Gunho Lee, David A. Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. Above the clouds: A berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, Feb 2009.

[38] Kenneth Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3):265–290, 1954.

[39] Jeff Barr. *Host Your Web Site In The Cloud: Amazon Web Services Made Easy Amazon EC2 Made Easy*. Sitepoint, 1st edition, 2010.

[40] Michael R Baye, Guoqiang Tian, and Jianxin Zhou. Characterizations of the existence of equilibria in games with discontinuous and non-quasiconcave payoffs. *Review of Economic Studies*, 60(4):935–48, October 1993.

[41] Robert Birke, Lydia Y. Chen, and Evgenia Smirni. Data centers in the cloud: A large scale performance study. In Chang [44], pages 336–343.

[42] Mathias Björkqvist, Lydia Y. Chen, and Walter Binder. Opportunistic service provisioning in the cloud. In Chang [44], pages 237–244.

[43] Marco Caldirola. Tecniche di resource allocation per sistemi virtualizzati di larga scala. Master's thesis, Politecnico di Milano, 2010.

[44] Rong Chang, editor. *2012 IEEE Fifth International Conference on Cloud Computing, Honolulu, HI, USA, June 24-29, 2012*. IEEE, 2012.

[45] L. Cherkasova and P. Phaal. Session-based admission control: a mechanism for peak load management of commercial web sites. *Computers, IEEE Transactions on*, 51(6):669–685, Jun.

[46] Gerard Debreu. A social equilibrium existence theorem. *Nat. Acad. Science*, 38:886–893, 1952.

[47] Brian Dougherty, Jules White, and Douglas C. Schmidt. Model-driven auto-scaling of green cloud computing infrastructure. *Future Generation Comp. Syst.*, 28(2):371–378, 2012.

[48] Axel Dreves and Christian Kanzow. Nonsmooth optimization reformulations characterizing all solutions of jointly convex generalized nash equilibrium problems. *Computational Optimization and Applications*, 50(1):23–48, 2011.

[49] Parijat Dube, Zhen Liu, Laura Wynter, and Cathy H. Xia. Competitive equilibrium in e-commerce: Pricing and outsourcing. *Computers & OR*, 34(12):3541–3559, 2007.

[50] Parijat Dube, Corinne Touati, and Laura Wynter. Capacity planning, quality of service and price wars. *SIGMETRICS Performance Evaluation Review*, 35(3):31–33, 2007.

[51] Sourav Dutta, Sankalp Gera, Akshat Verma, and Balaji Viswanathan. Smartscale: Automatic application scaling in enterprise clouds. In Chang [44], pages 221–228.

[52] Yu.M. Ermoliev and S.D. Flaam. Repeated play of potential games. *Cybernetics and Systems Analysis*, 38:355–367, 2002.

[53] Francisco Facchinei, Andreas Fischer, and Veronica Piccialli. Generalized nash equilibrium problems and newton methods. *Math. Program.*, 117(1-2):163–194, 2009.

[54] Francisco Facchinei and Christian Kanzow. Generalized nash equilibrium problems. *Annals OR*, 175(1):177–211, 2010.

[55] Francisco Facchinei, Veronica Piccialli, and Marco Sciandrone. Decomposition algorithms for generalized potential games. *Computational Optimization and Applications*, 50:237–262, 2011.

[56] Yuan Feng, Baochun Li, and Bo Li. Price competition in an oligopoly cloud market. 2011.

[57] Hadi Goudarzi and Massoud Pedram. Multi-dimensional sla-based resource allocation for multi-tier cloud computing systems. In Liu and Parashar [68], pages 324–331.

[58] Albert G. Greenberg and Kazem Sohraby, editors. *Proceedings of the IEEE INFOCOM 2012, Orlando, FL, USA, March 25-30, 2012*. IEEE, 2012.

[59] Makhlouf Hadji and Djamal Zeghlache. Minimum cost maximum flow algorithm for dynamic resource allocation in clouds. In Chang [44], pages 876–882.

[60] Mohammad Mehedi Hassan, Biao Song, and Eui nam Huh. Distributed resource allocation games in horizontal dynamic cloud federation platform. In Thulasiraman et al. [85], pages 822–827.

[61] MohammadMehedi Hassan, M.Shamim Hossain, A.M.Jehad Sarkar, and Eui-Nam Huh. Cooperative game-based distributed resource allocation in horizontal dynamic cloud federation platform. *Information Systems Frontiers*, pages 1–20, 2012.

[62] Ting He, Shiyao Chen, Hyoil Kim, Lang Tong, and Kang-Won Lee. Scheduling parallel tasks onto opportunistically available cloud resources. In Chang [44], pages 180–187.

[63] Tatsuro Ichiishi. *Game Theory for Economic Analysis*. New York: Academic Press, 1983.

[64] Ganesh Neelakanta Iyer and Bharadwaj Veeravalli. On the resource allocation and pricing strategies in compute clouds using bargaining approaches. In Veeravalli and Foster [89], pages 147–152.

[65] Kleopatra Konstanteli, Tommaso Cucinotta, Konstantinos Psychas, and Theodora A. Varvarigou. Admission control for elastic cloud services. In Chang [44], pages 41–48.

[66] Dinesh Kumar, Asser N. Tantawi, and Li Zhang. Real-time performance modeling for adaptive software systems with multi-class workload. In *MASCOTS*, pages 1–4, 2009.

[67] Yi-Kuei Lin and Ping-Chen Chang. Reliability evaluation of a computer network in cloud computing environment subject to maintenance budget. *Applied Mathematics and Computation*, 219(8):3893–3902, 2012.

[68] Ling Liu and Manish Parashar, editors. *IEEE International Conference on Cloud Computing, CLOUD 2011, Washington, DC, USA, 4-9 July, 2011*. IEEE, 2011.

[69] Tieming Liu, Chinnatat Methapatara, and Laura Wynter. Revenue management model for on-demand it services. *European Journal of Operational Research*, 207(1):401–408, 2010.

[70] Michele Mazzucco and Dmytro Dyachuk. Optimizing cloud providers revenues via energy efficient server allocation. *Sustainable Computing: Informatics and Systems*, 2(1):1 – 12, 2012.

[71] Peter Mell and Tim Grance. The NIST Definition of Cloud Computing. Technical report, July 2009.

[72] Ishai Menache, Asuman Ozdaglar, and Nahum Shimkin. Socially optimal pricing of cloud computing resources. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, VALUETOOLS '11, pages 322–331, ICST, Brussels, Belgium, Belgium, 2011. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[73] Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, May 1996.

[74] Jacqueline Morgan and Vincenzo Scalzo. Pseudocontinuous functions and existence of nash equilibria. *Journal of Mathematical Economics*, 43(2):174 – 183, 2007.

[75] John Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, 1951.

[76] C.G.A.M. van den Nouweland, Peter Borm, W. van Golstein Brouwers, R. Groot Bruinderink, and S.H. Tijs. A game theoretic approach to problems in telecommunication. Open access publications from tilburg university, Tilburg University, 1996.

[77] Jong-Shi Pang and Jen-Chih Yao. On a generalization of a normal map and equation. *SIAM J. Control Optim.*, 33(1):168–184, January 1995.

[78] N.S.V. Rao, S.W. Poole, Fei He, Jun Zhuang, C.Y.T. Ma, and D.K.Y. Yau. Cloud computing infrastructure robustness: A game theory approach. In *Computing, Networking and Communications (ICNC), 2012 International Conference on*, pages 34 –38, 30 2012-feb. 2 2012.

[79] Philip J. Reny. On the existence of pure and mixed strategy nash equilibria in discontinuous games. *Econometrica*, 67(5):1029–1056, 1999.

[80] U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh. A cost-aware elasticity provisioning system for the cloud. In *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, pages 559 –570, june 2011.

[81] Upendra Sharma, Prashant J. Shenoy, Sambit Sahu, and Anees Shaikh. Kingfisher: Cost-aware elasticity in the cloud. In *INFOCOM* [24], pages 206–210.

[82] Yang Song, Murtaza Zafer, and Kang-Won Lee. Optimal bidding in spot instance market. In Greenberg and Sohraby [58], pages 190–198.

[83] Shekhar Srikantaiah, Aman Kansal, and Feng Zhao. Energy aware consolidation for cloud computing. In *Proceedings of the 2008 conference on Power aware computing and systems*, HotPower'08, pages 10–10, Berkeley, CA, USA, 2008. USENIX Association.

[84] Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors. *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*. ACM, 2011.

[85] Parimala Thulasiraman, Laurence Tianruo Yang, Qiwen Pan, Xingang Liu, Yaw-Chung Chen, Yo-Ping Huang, Lin-Huang Chang, Che-Lun Hung, Che-Rung Lee, Justin Y. Shi, and Ying Zhang, editors. *13th IEEE International Conference on High Performance Computing & Communication, HPCC 2011, Banff, Alberta, Canada, September 2-4, 2011*. IEEE, 2011.

[86] Fengguang Tian and Keke Chen. Towards optimal resource provisioning for running mapreduce programs in public clouds. In Liu and Parashar [68], pages 155–162.

[87] Guoqiang Tian and Jianxin Zhou. Transfer continuities, generalizations of the weierstrass and maximum theorems: A full characterization. *Journal of Mathematical Economics*, 24(3):281 – 303, 1995.

[88] Luis M. Vaquero, Luis Rodero-Merino, Juan Caceres, and Maik Lindner. A break in the clouds: towards a cloud definition. *SIGCOMM Comput. Commun. Rev.*, 39(1):50–55, December 2008.

[89] Bharadwaj Veeravalli and Ian T. Foster, editors. *Proceedings of the 17th IEEE International Conference on Networks, ICON 2011, Singapore, December 14-16, 2011*. IEEE, 2011.

[90] Anna von Heusinger and Christian Kanzow. Optimization reformulations of the generalized nash equilibrium problem using nikaido-isoda-type functions. *Computational Optimization and Applications*, 43(3):353–377, 2009.

[91] Mark Voorneveld. Equilibria and approximate equilibria in infinite potential games. *Economics Letters*, 56(2):163 – 169, 1997.

[92] Jian Wan, Dechuan Deng, and Congfeng Jiang. Non-cooperative gaming and bidding model based resource allocation in virtual machine environment. In *IPDPS Workshops* [26], pages 2183–2188.

[93] Hongyi Wang, Qingfeng Jing, Rishan Chen, Bingsheng He, Zhengping Qian, and Lidong Zhou. Distributed systems meet economics: pricing in the cloud. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, HotCloud'10, pages 6–6, Berkeley, CA, USA, 2010. USENIX Association.

[94] J Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II*, 1(36):352–362, 1952.

[95] Guiyi Wei, Athanasios V. Vasilakos, Yao Zheng, and Naixue Xiong. A game-theoretic method of fair resource allocation for cloud computing services. *The Journal of Supercomputing*, 54(2):252–269, 2010.

[96] Andreas Wolke and Gerhard Meixner. Twospot: A cloud platform for scaling out web applications dynamically. In Elisabetta Nitto and

Ramin Yahyapour, editors, *Towards a Service-Based Internet*, volume 6481 of *Lecture Notes in Computer Science*, pages 13–24. Springer Berlin Heidelberg, 2010.

[97] L.A. Wolsey. *Integer Programming*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 1998.

[98] Linlin Wu, Saurabh Kumar Garg, and Rajkumar Buyya. Sla-based admission control for a software-as-a-service provider in cloud computing environments. *J. Comput. Syst. Sci.*, 78(5):1280–1299, 2012.

[99] Z. Xiao, Q. Chen, and H. Luo. Automatic scaling of internet applications for cloud computing services. *Computers, IEEE Transactions on*, PP(99):1, 2012.

[100] Z. Xiao, W. Song, and Q. Chen. Dynamic resource allocation using virtual machines for cloud computing environment. *Parallel and Distributed Systems, IEEE Transactions on*, PP(99):1, 2012.

[101] PengCheng Xiong, Zhikui Wang, Simon Malkowski, Qingyang Wang, Deepal Jayasinghe, and Calton Pu. Economical and robust provisioning of n-tier cloud workloads: A multi-level control approach. In *ICDCS* [23], pages 571–580.

[102] Murtaza Zafer, Yang Song, and Kang-Won Lee. Optimal bids for spot vms in a cloud for deadline constrained jobs. In Chang [44], pages 75–82.

[103] Sharrukh Zaman and Daniel Grosu. An online mechanism for dynamic vm provisioning and allocation in clouds. In Chang [44], pages 253–260.

[104] Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: state-of-the-art and research challenges. *J. Internet Services and Applications*, 1(1):7–18, 2010.

[105] Qi Zhang, Quanyan Zhu, Mohamed Faten Zhani, and Raouf Boutaba. Dynamic service placement in geographically distributed clouds. In *ICDCS* [25], pages 526–535.

[106] Xiaoyun Zhu, Donald Young, Brian J. Watson, Zhikui Wang, Jerry Rolia, Sharad Singhal, Bret Mckee, Chris Hyser, Daniel Gmach, Robert Gardner, Tom Christian, and Ludmila Cherkasova. 1000 islands: an integrated approach to resource management for virtualized data centers. *Cluster Computing*, 12(1):45–57, March 2009.