Cloud Computing Using Elasticity

[1]**Karimella Vikram**, [2]**Dr. Soundara Rajan**, [3]**CH.Venkateshwara Rao**, [4]**B.Rajani**

[1,2,3,4]Dept. of CSE, TKR Engineeriong College, Medbowli, Meerpet, Saroor Nagar, Hyderabad, India

Abstract

Originating from the field of physics and economics, the term elasticity is nowadays heavily used in the context of cloud computing. In this context, elasticity is commonly understood as the ability of a system to automatically provision and deprovision computing resources on demand as workloads change. However, elasticity still lacks a precise definition as well as representative metrics coupled with a benchmarking methodology to enable comparability of systems. Existing definitions of elasticity are largely inconsistent and unspecific, which leads to confusion in the use of the term and its differentiation from related terms such as scalability and efficiency; the proposed measurement methodologies do not provide means to quantify elasticity without mixing it with efficiency or scalability aspects. In this short paper, we propose a precise definition of elasticity and analyze its core properties and requirements explicitly distinguishing om related terms such as scalability and efficiency Furthermore, we present a set of appropriate elasticity metrics and sketch a new elasticity tailored benchmarking methodology addressing the special requirements on workload design and calibration.

Keywords

Elasticity, Scalability, Cloud computing , Provision, De-provisional state

I. Introduction

Elasticity has originally been defined in physics as a material property capturing the capability of returning to its original state after a deformation. In economical theory, informally, elasticity denotes the sensitivity of a dependent variable to changes in one or more other variables [1]. In both cases, elasticity is an intuitive concept and can be precisely described using mathematical formulas. The concept of elasticity has been transferred to the context of cloud computing and is commonly considered as one of the central attributes of the cloud paradigm [10]. For marketing purposes, the term elasticity is heavily used in cloud providers' advertisements and even in the naming of specific products or services. Even though tremendous efforts are invested to enable cloud systems to behave in an elastic manner, no common and precise understanding of this term in the context of cloud computing has been established so far, and no ways havebeen proposed to quantify and compare elastic behavior. To underline this observation, we cite five definitions of elasticity demonstrating the inconsistent use and understanding of the term:

1. ODCA, Compute Infrastructure-as-a-Service [9] "[. . .] defines elasticity as the configurability and expandability of the solution [. . .] Centrally, it is the ability to scale up and scale down capacity based on subscriber workload."

2. NIST Definition of Cloud Computing [8] "Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time."

3. IBM, Thoughts on Cloud, Edwin Schouten, 2012 [11] "Elasticity is basically a 'rename' ofscalability [. . .]" and "removes any manual laborneeded to increase or reduce capacity."

4. Rich Wolski, CTO, Eucalyptus, 2011 [12] "Elasticity measures the ability of the cloud to map a single user request to different resources."

5. Reuven Cohen, 2009 [2] Elasticity is "the quantifiableability to manage, measure, predict and adapt responsiveness of an application based on real time demands placed on an infrastructure using a combination of local and remote computing resources."

Definitions (1), (2), and (3) in common describe elasticity as the scaling of system resources to increase ordecrease capacity, whereby definitions (1), (2) and (5) specifically state that the amount of provisioned resources is somehow connected to the recent demand orworkload. In these two points there appears to be some consent. Definitions (4) and (5) try to capture elasticity ina generic way as a 'quantifiable' system ability to handlerequests using different resources. Both of these definitions, however, neither give concrete details on the coreaspects of elasticity, nor provide any hints on how elasticitycan be measured. Definition (3) assumes that nomanual work at all is needed, whereas in the NIST definition(2), the processes enabling elasticity do not need tobe fully automatic. In addition, the NIST definition addsthe adjective 'rapid' to elasticity and draws the idealisticpicture of 'perfect' elasticity where endless resources areavailable with an appropriate provisioning at any point intime, in a way that the end-user does not experience anyperformance variability.

We argue that existing definitions of elasticity fail tocapture the core aspects of this term in a clear and unambiguousmanner and are even contradictory in someparts. To address this issue, in this short paper, we proposea new refined definition of elasticity considering indetail its core aspects and the prerequisites of elastic systembehavior (Section II). Thereby, we clearly differentiate elasticity from its related terms scalability and efficiency. In Section IV, we present metrics that are ableto capture elasticity, followed by Section V, in whichwe outline a benchmarking methodology for quantifying elasticity discussing the issues of representativeness, reproducibilityand fairness of the measurement approach.

II. Elasticity

In this section, we first describe some importantprerequisites in order to be able to speak of elasticity,present a new refined and comprehensive definition, andthen analyse its core aspects and dimensions. Finally,we differentiate between elasticity and its related termsscalability and efficiency.

A. Prerequisites

The scalability of a system including all hardware, virtualization,and software layers within its boundaries isa prerequisite in order to be able to speak of elasticity.

Scalability is the ability of a system to sustain increasingworkloads with adequate performance provided thathardware resources are added. Scalability in the context of distributed systems has been defined in [6], as wellas more recently in [3-4], where also a measurementmethodology is proposed.Given that elasticity is related to the ability of a systemto adapt to changes in workloads and resource demands,the existence of at least one specific adaptation processis assumed. The latter is normally automated, however,in

a broader sense, it could also contain manual steps. Without a defined adaptation process, a scalable system cannot behave in an elastic manner, as scalability on itsown does not include temporal aspects. When evaluating elasticity, the following points need to be checked beforehand:

1. Autonomic Scaling
What adaptation process is used for autonomic scaling?

2. Elasticity Dimensions
What is the set of resource types scaled as part ofthe adaptation process?

3. Resource Scaling Units
For each resource type, in what unit is the amountof allocated resources varied?

4. Scalability Bounds
For each resource type, what is the upper bound onthe amount of resources that can be allocated?

B. Definition
Elasticity is the degree to which a system is able toadapt to workload changes by provisioning and deprovisioningresources in an autonomic manner,such that at each point in time the available resourcesmatch the current demand as closely as possible.

C. Dimensions and Core Aspects
Any given adaptation process is defined in the context of at least one or possibly multiple types of resources thatcan be scaled up or down as part of the adaptation. Eachresource type can be seen as a separate dimension of theadaptation process with its own elasticity properties. If aresource type is a container of other resources types, like the case of a virtual machine having assigned CPUcores and RAM, elasticity can be considered at multiplelevels. Normally, resources of a given resource typecan only be provisioned in discrete units like CPU cores, Virtual Machines (VMs), or physical nodes. For each dimensionof the adaptation process with respect to a specificresource type, elasticity captures the following coreaspects of the adaptation: Speed The speed of scaling up is defined as the timeit takes to switch from an underprovisioned stateto an optimal or overprovisioned state. The speedof scaling down is defined as the time it takes toswitch from an overprovisioned state to an optimalor underprovisioned state. The speed of scalingup/down does not correspond directly to the technicalresource provisioning/deprovisioning time. Precision The precision of scaling is defined as the absolutedeviation of the current amount of allocatedresources from the actual resource demand.As discussed above, elasticity is always consideredwith respect to one or more resource types. Thus, a directcomparison between two systems in terms of elasticityis only possible if the same resource types (measured inidentical units) are scaled.

To evaluate the actual observable elasticity in a givenscenario, as a first step, one must define the criterionbased on which the amount of provisioned resources isconsidered to match the actual current demand neededto satisfy the system's given performance requirements. Based on such a matching criterion, specific metrics that quantify the above mentioned core aspects, as discussedin more detail in Section IV, can be defined to quantifythe practically achieved elasticity in comparison to thehypothetical optimal elasticity. The latter corresponds to the hypothetical case where the system is scalable with respect to all considered elasticity

dimensions without any upper bounds on the amount of resources that can be provisioned and where resources are provisioned and deprovisioned immediately as they are needed exactly matching the actual demand at any point in time. Optimal elasticity, as defined here, would only be limitedby the resource scaling units.

D. Differentiation
In this section, we highlight the conceptual differencesbetween elasticity and the related terms scalability andefficiency.
Scalability is a prerequisite for elasticity, but it does notconsider temporal aspects of how fast, how often,and at what granularity scaling actions can be performed.Scalability is the ability of the system tosustain increasing workloads by making use of additionalresources, and therefore, in contrast to elasticity,it is not directly related to how well the actualresource demands are matched by the provision edresources at any point in time.Efficiency expresses the amount of resources consumedfor processing a given amount of work. In contrastto elasticity, efficiency is not limited to resourcetypes that are scaled as part of the system's adaptation mechanisms. Normally, better elasticity results in higher efficiency. The other way round, this implication is not given, as efficiency can be influenced by other factors independent of the system's elasticity mechanisms (e.g., different implementationsof the same operation).

III. Derivation of the Matching Function
To capture the criterion based on which the amount ofprovisioned resources is considered to match the actualcurrent demand, we define a matching function m(w)=ras a system specific function that returns the minimalamount of resources r for a given resource type neededto satisfy the system's performance requirements at aspecified workload intensity. The workload intensity wcan be specified either as the number of workload units(e.g., user requests) present at the system at the sametime (concurrency level), or as the number of work loadunits that arrive per unit of time (arrival rate). A matchingfunction is needed for both directions of scaling(up/down), as it cannot be assumed that the optimal resource allocation level when transitioning from an under provisioned state (upwards) are the same as when transitioning from an overprovisioned state (downwards).
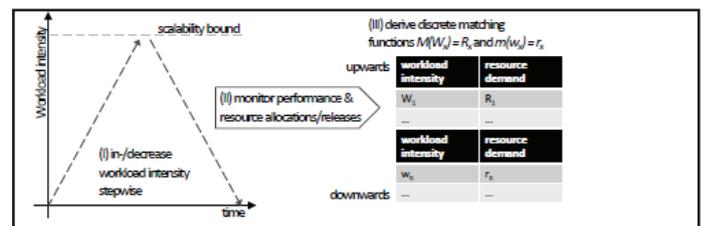

Fig. 1: Illustration of a Masurement-based Derivation of Matching Functions

The matching functions can be derived based on measurements,as illustrated in fig. 1, by increasing theworkload intensity w stepwise, while measuring the resourceconsumption r, and tracking resource allocationchanges. The process is then repeated for decreasing w.After each change in the workload intensity, the systemå A is the accumulated time in underprovisionedstate.
_ U is the average amount of underprovisioned resourcesduring an underprovisioned period.
_åU is the accumulated amount of under provisioned resources.
_ B, åB, O, and åO are defined similarly for overprovisionedstates.
shouldbegivenenoughtimetoadaptitsresourceallocationsreaching

a stable state for the respective workloadintensity. As a rule of thumb, at least two times the technicalresource provisioning time is recommended to useas a minimum. As a result of this step, a system specifictable is derived that maps workload intensity levelsto resource demands, and the other way round, for both scaling directions within the scaling bounds.

IV. Elasticity Metrics

To capture the core elasticity aspects speed and precision, we propose the following definitions and metrics asillustrated in Fig. 2:

_A is the average time to switch from an underprovisionedstate to an optimal or overprovisioned stateand corresponds to the average speed of scaling up.
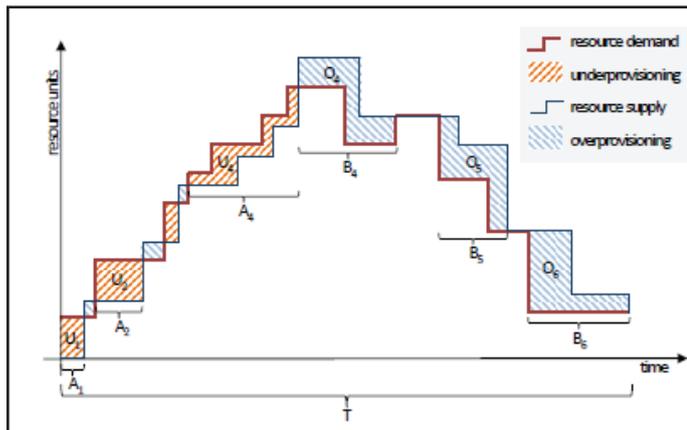

Fig. 2: Capturing Core Elasticity Metrics

We define the average precision of scaling up Pu asPu = åUT where T is the total duration of the evaluationperiod, and accordingly Pd = åOT is defined as the averageprecision of scaling down. Based on the above definedquantities, one could define an elasticity metric forscaling up Eu as inversely proportional to A and U, e.g.Eu = 1A_U , and accordingly elasticity for scaling down Ed = 1 B_O. The elasticity of a system under test (SUT) scan then be captured in a matrix Ms where each vector vdrepresents an elasticity dimension d and contains the valuesof the elasticity core metrics Eu, A, Pu for scaling upand Ed, B, Pd for scaling down. As an alternative to these metrics, the dynamic timewarping (DTW) distance [7] can be used as a robust distancemetric to capture the similarity between the demand and supply curves as well as to approximate thetechnical reaction time of the adaptation mechanism. Acase study demonstrating this approach can be foundin [5].

V. Towards Benchmarking Elasticity

Characterizing the elasticity of a single system is not aimple task on its own and it becomes even more complicatedwhen comparing different systems. An elasticitybenchmark is expected to deliver reproducible resultsand generate a consistent order of the different systemsunder test (SUTs) reflecting their potential and observedelasticity, while not mixing this with general system efficiencyand scalability aspects. Traditional benchmarkingapproaches induce identical workloads on differentSUTs to provide a basis for fair comparisons, whereasan elasticity benchmark is required to induce identicaldemand curves. If two elastic systems exhibit significant differences in efficiency (the amount of resources requiredfor meeting performance requirements at a givenworkload intensity level), it might well be that when processing an identical workload, their adaptation

mechanismsare exercised in a significantly different manner.As illustrated in Figure 3, in that case, deriving the elasticitymetrics for the same workload would result in unfaircomparison since the more efficient system wouldappear to exhibit better elasticity given that its adaptationmechanisms were not stressed to the same extent.
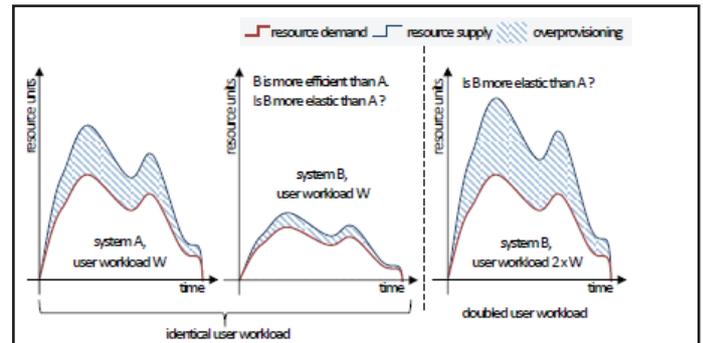

Fig. 3: Elasticity Vs. Efficiency

Therefore, the first step towards portability of an elasticitybenchmark and comparability of its results wouldbe the specification of a representative set of demandcurves and common performance goals in terms of responsiveness,throughput or utilisation for the consideredresource types. The demand curves themselvesshould contain bursts of different intensity, upward anddownward scaling trends and seasonal patterns of differentshapes, concerning amplitude, duration and baselevel capturing the most representative real-life scenarios.Further challenges include the automated derivation of the mapping functions as well as the generation of aworkload that induces the targeted demand curves as accuratelyas possible on the evaluated SUTs.

VI. Conclusion

In this short paper, we proposed a refined definition ofelasticity to contribute in establishing a common understandingof this term in the context of cloud computing.Furthermore, we examined the core aspects of elasticityexplicitly differentiating it conceptually from the classicalnotions of scalability and efficiency. Finally, we proposemetrics to capture the core elasticity aspects as wellas an elasticity benchmarking approach focusing on thespecial requirements on workload design and its implementation.

References

[1] CHIANG, A. C., AND WAINWRIGHT, K. Fundamental methodsof mathematical economics, 4. ed., internat. ed., [repr.] ed. McGraw-Hill [u.a.], Boston, Mass. [u.a.], 2009.

[2] COHEN, R. Defining Elastic Computing, September2009. [Online] Available: http://www.elasticvapor.com/2009/09/defining-elastic-computing.html, last consulted Feb.2013.

[3] DUBOC, L. A Framework for the Characterization and Analysis of Software Systems Scalability. PhD thesis, Department ofComputer Science, University College London, 2009. [Online] Available: http://discovery.ucl.ac.uk/19413/1/19413.pdf.

[4] DUBOC, L., ROSENBLUM, D., AND WICKS, T. A Framework for Characterization and Analysis of Software System Scalability.In Proceedings of the 6th joint meeting of the European SoftwareEngineering Conference and the ACM SIGSOFT Symposiumon The Foundations of Software Engineering (ESEC-FSE'07) (2007), ACM, pp. 375–384.

[5] HERBST, N. R. Quantifying the Impact of Configuration Space for Elasticity Benchmarking. Study thesis, Faculty of

Computer Science, Karlsruhe Institute of Technology (KIT), Germany, 2011. [Online] Available: http://sdqweb.ipd.kit.edu/publications/pdfs/Herbst2011a.pdf.

[6] JOGALEKAR, P., AND WOODSIDE, M. Evaluating the scalability of distributed systems. IEEE Transactions on Parallel andDistributed Systems 11 (2000), pp. 589–603.

[7] KEOGH, E., AND RATANAMAHATANA, C. A. Exact indexingof dynamic time warping. Knowl. Inf. Syst. 7, 3 (Mar. 2005), pp. 358–386.

[8] MELL, P., AND GRANCE, T. The NIST Definition of Cloud Computing. Tech. rep., U.S. National Institute ofStandards and Technology (NIST), 2011. Special Publication 800-145, [Online] Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[9] OCDA. Master Usage Model: Compute Infratructure as aService. Tech. rep., Open Data Center Alliance (OCDA), 2012.[Online]Available: http://www.opendatacenteralliance.org/docs/ODCA_Compute_IaaS_MasterUM_v1.0_Nov2012.pdf.

[10] PLUMMER, D. C., SMITH, D. M., BITTMAN, T. J., CEARLEY, D. W., CAPPUCCIO, D. J., SCOTT, D., KUMAR, R., ANDROBERTSON, B. Study: Five Refining Attributes of Public andPrivate Cloud Computing. Tech. rep., Gartner, 2009. http://ww.gartner.com/DisplayDocument?doc_cd=167182, lastconsulted Feb. 2013.

[11] SCHOUTEN, E. Rapid Elasticity and the Cloud, September 2012. [Online] Available: http://thoughtsoncloud.com/index.php/2012/09/rapid-elasticity-and-the-cloud/, last consultedFeb. 2013.

[12] WOLSKI, R. Cloud Computing and Open Source: WatchingHype meet Reality, May 2011. [Online] Available: http://www.ics.uci.edu/ccgrid11/files/ccgrid-11_Rich_Wolsky.pdf, last consultedFeb. 2013.5

Mr. CH.VenkateshwaraRao is working as a Asst.Professor in CSE Dept at TKR Engineering College Meerpet, Hyd. He has published number of National and International research papers. His interested research areas are DataMining BigData, Computer Networks.

Mrs. B.Rajani is working as a Asst. Professor in CSE Dept at TKR Engineering College Meerpet, Hyd. She has published number of National and International research papers. Her interested research areas are Data Mining, Big Data., Computer Networks.

Mr. K.Vikram is working as a Professor in TKR Engineering College, Hyderabad, India. He has published and presented good number of Technical papers in National and International Conferences. His main research interest areas are Software Engineering, Data Mining, and Networks. Image Processing He Worked as a I/C Principal in in various Engineering colleges. He is life member of IETE, CSI. He published 30 International Papers and 05 International Conferences and published 03 Text Books.

Dr. Sounadara Rajan Working as a Professor and Director of R & D. in TKREC, Hyderabad. He receive B.E From S.V University, M.Tech From JNTUH, Ph.D. From IIT Roorkee. His area specialization is Microcomputers Electronic Instrumentation, Computer Operating Systems Image Processing, Lower power VLSI Technology. He Published Many Technical papersNational and International Level. Best Principal Best Principal Award by International Institute of Education and Management, New Delhi in 2011. Best Teacher Award by Lead India -2020 in 2005