Elasticity in Cloud Computing: What It Is, and What It Is Not

Nikolas Herbst, Samuel Kounev, Ralf Reussner
herbst@kit.edu

ICAC’13, San Jose, CA – 26th June 2013
What People Say Elasticity Is…

OCDA [1]
up & down scaling subscriber workload

IBM, Schouten [3]
scalability increase & reduce no manual labor

NIST [2]
rapid elasticity unlimited provision & release
sometimes automated with demand

Eukalyptus, Wolski [4]
measurable mapping of requests to resources

Cohen [5]
quantifyable real-time demands local & remote

Motivation Definition Metrics Benchmarking Conclusion
Elasticity

is the degree to which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible.
Elasticity Example

Service Level Agreement (SLA):
E.g.: resp. time ≤ 2 sec, 95%

Resource Demand:
Minimal amount of #VMs required to ensure SLAs.

Motivation
Definition
Metrics
Benchmarking
Conclusion

---

Motivation

Definition

Metrics

Benchmarking

Conclusion
Comparability

Motivation

Definition

Metrics

Benchmarking

Conclusion

<table>
<thead>
<tr>
<th>#req/sec.</th>
<th>workload intensity</th>
<th>curve i</th>
<th>#req/sec.</th>
<th>workload intensity</th>
<th>curve i</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>time</th>
<th>#VMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

system A

<table>
<thead>
<tr>
<th>time</th>
<th>#VMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

system B
Comparability

**Motivation**

- Definition: Comparability
- Metrics: Workload intensity
- Benchmarking: #req/sec.

**System A**
- #VMs: 8, 6, 4, 2
- Time: 2, 4, 6, 8

**System B**
- #VMs: 8, 6, 4, 2
- Time: 2, 4, 6, 8

**Curve i**
- Workload intensity

**Curve i***
- Workload intensity
Intuitive Elasticity?

Resource demand
Resource supply
Overprovisioning

Same user workload on system B
System B at a doubled user workload

Motivation
Definition
Metrics
Benchmarking
Conclusion
Elasticity Metrics

Motivation
Definition
Metrics
Benchmarking
Conclusion

resource demand
underprovisioning
resource supply
overprovisioning


### Elasticity Metrics

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
</table>
| $\bar{A}$ | Average time of switch from an underprovisioned to an optimal or overprovisioned state
| $\Sigma A$ | Accumulated time in underprovisioned state. |
| $\bar{U}$ | Average amount of underprovisioned resources during an underprovisioned period. |
| $\Sigma U$ | Accumulated amount of underprovisioned resources. |
| $\bar{B}, \Sigma B, \bar{O}, \Sigma O$ | Correspondingly for overprovisioned states |

\[
P_u = \frac{\Sigma U}{T}; P_d = \frac{\Sigma O}{T},\]

\[
T = \text{total evaluation duration}
\]

\[
E_u = \frac{1}{\bar{A}x\bar{U}}; E_d = \frac{1}{\bar{B}x\bar{O}}
\]

**Average precision of scaling up / down**

**Elasticity metric for scaling up / down**
Benchmarking Challenges I

→ Derivation of a matching function

(I) in-/decrease workload intensity stepwise
(II) monitor performance & resource allocations/releases
(III) derive discrete matching functions $M(W_x) = R_x$ and $m(w_x) = r_x$

<table>
<thead>
<tr>
<th>workload intensity</th>
<th>resource demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_1$</td>
<td>$R_1$</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>workload intensity</td>
<td>resource demand</td>
</tr>
<tr>
<td>$w_n$</td>
<td>$r_n$</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Motivation | Definition | Metrics | Benchmarking | Conclusion
---|---|---|---|---
Herbst, Kouniev, Reussner | | | | ICAC’13, San Jose, CA – 26th June 2013
1. **Derive the system specific matching function of workload intensity and resource demand**

2. **Define a representative set of workload intensity traces**

3. **Induce identical demand curves on different systems by parameterizing a workload intensity trace**

→ **Fair, consistent, reproducible ordering of elastic systems**
(same elasticity dimension & scaling units)
Conclusion

Elasticity

• Generic definition
• Core aspects
• Prerequisites
• Delineation from scalability and efficiency

Metrics

• Precision and speed of scaling up / down

Benchmarking Elasticity

• Derivation of a matching function
Literature

[1] COHEN, R.  
last consulted Feb. 2013.


[3] OCDA. Master Usage Model:  
Compute Infrastructure as a Service. Tech. rep., Open Data Center Alliance (OCDA), 2012.  

[4] SCHOUTEN, E.  
Rapid Elasticity and the Cloud, September 2012.  
last consulted Feb. 2013.

[5] WOLSKI, R.  
last consulted Feb. 2013.
Backup: Definitions

- ODCA, Compute Infrastructure-as-a-Service:
  "[...] 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." [1]

- NIST Definition of Cloud Computing
  "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 anytime.” [2]

- IBM, Thoughts on Cloud, Edwin Schouten:
  ”Elasticity is basically a ‘rename’ of scalability [...]” and ”removes any manual labor needed to increase or reduce capacity.” [3]

- Rich Wolski, CTO, Eucalyptus:
  ”Elasticity measures the ability of the cloud to map a single user request to different resources.” [4]

- Reuven Cohen:
  Elasticity is "the quantifiable ability to manage, measure, predict and adaptive responsiveness of an application based on real time demands placed on an infrastructure using a combination of local and remote computing resources.” [5]
Backup: Elasticity Prerequisites

**Autonomic Scaling**
- What adaptation process is used for autonomic scaling?

**Elasticity Dimension**
- What is the set of resource types scaled as part of the adaptation process?

**Resource Scaling Units**
- For each resource type, in what unit is the amount of allocated resources varied?

**Scalability Bounds**
- For each resource type, what is the upper bound on the amount of resources that can be allocated?
Speed

- The **speed of scaling up** is defined as the **time** it takes to **switch** from an under-provisioned state to an **optimal or overprovisioned state**. The **speed of scaling down** is defined as the **time** it takes to **switch** from an overprovisioned state to an **optimal or under-provisioned state**. The speed does not correspond directly to the technical resource provisioning / de-provisioning time.

Precision

- The **precision of scaling** is defined as the **absolute deviation** of the current amount of **allocated resources** from the actual **resource demand**
Scalability

- ... does not consider **temporal aspects** of how fast, how often, and at what granularity scaling actions can be performed.
- ... is not directly related to how well the actual resource demands are **matched** by the provisioned resources at any point in time.

Efficiency

- ... expresses the **amount of resources** consumed for processing a **given amount of work**.
- ... is not limited to resource types that are scaled as part of system’s adaptation mechanisms.
- Normally, better elasticity results in higher efficiency.