Karlsruhe Institute of Technology

Emmy Noether Research Group Descartes
Karlsruhe Institute of Technology
Germany

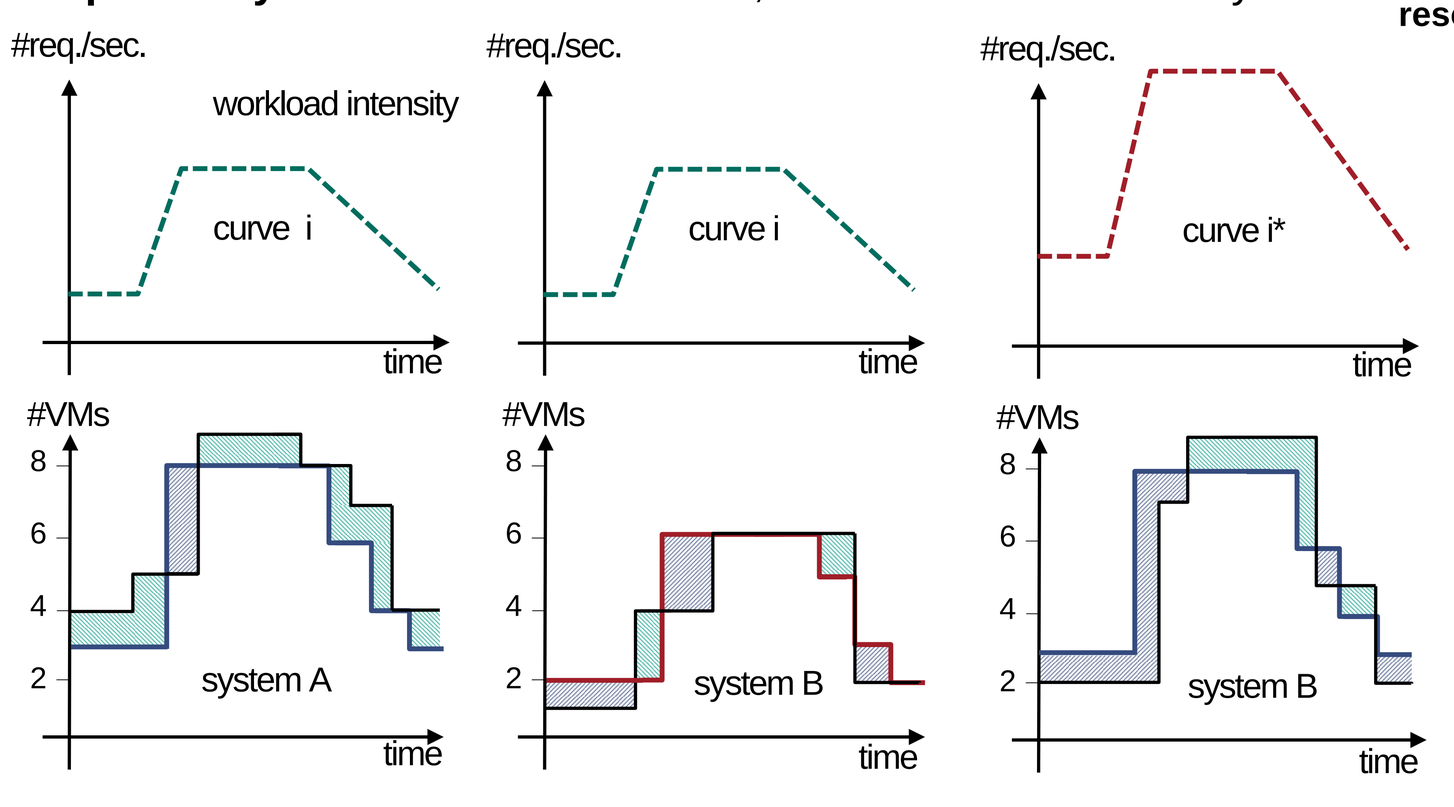
Elasticity in Cloud Computing: What It Is, and What It Is Not

Nikolas Herbst, herbst@kit.edu

Elasticity

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.

Comparability based on demand curve, not on workload intensity



Core Aspects of Elasticity

- **Speed**
  The **speed of scaling up** is the **time** it takes to **switch** from an under-provisioned state to **an optimal** or **overprovisioned state** and vice versa for the **speed of scaling down**.
- **Precision**
  The **precision of scaling** is the **absolute deviation** of the current amount of **allocated resources** from the actual **resource demand.**

Service Level Agreement (SLA)

E.g.: resp. time ≤ 2 sec, 95%

Resource Demand

Minimal amount of #VMs required to ensure SLAs.

workload intensity
resource demand
underprovisioning
resource supply
overprovisioning

Prerequisites

- Autonomic scaling
- Scalability bounds
- Elasticity dimensions
- Resource scaling units

Metrics

$\bar{A}$	Average time of switch from an underprovisioned to an optimal or overprovisioned state → **Average speed of scaling up**
$\sum A$	Accumulated time in underprovisioned state.
$\bar{U}$	Average amount of underprovisioned resources during an underprovisioned period.
$\sum U$	Accumulated amount of underprovisioned resources.
$\bar{B}, \sum B, \bar{O}, \sum O$	correspondingly for overprovisioned states

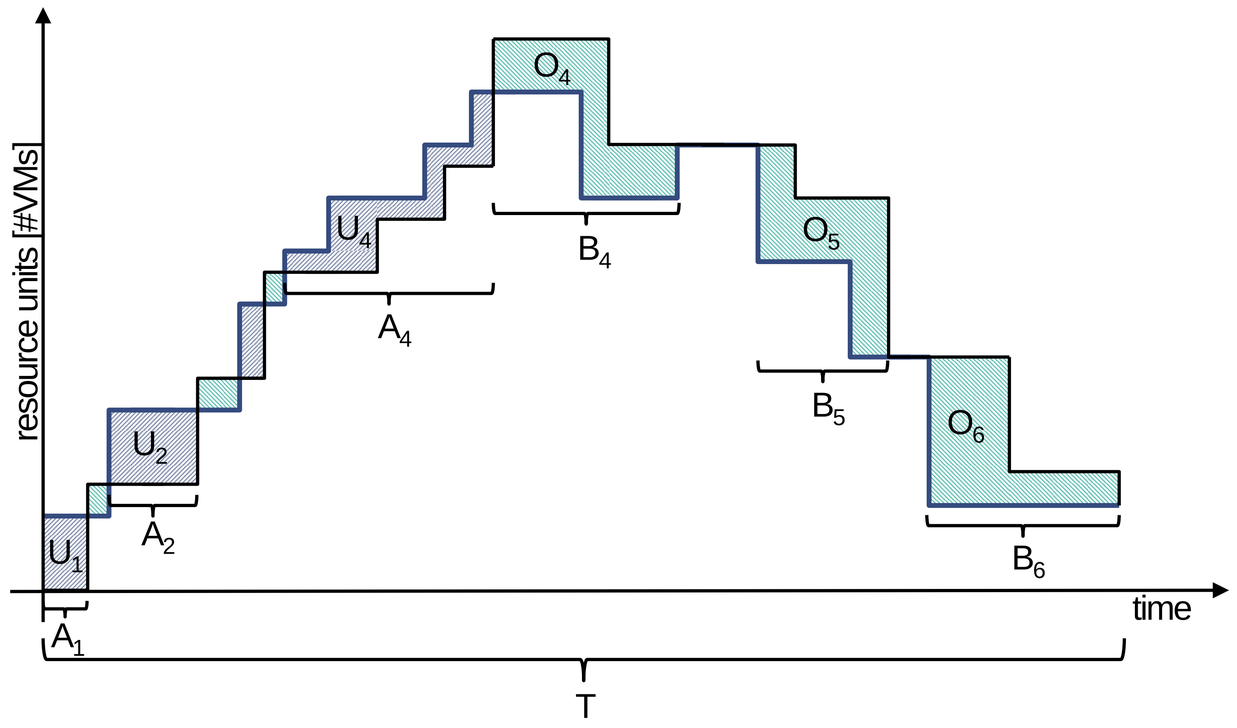
$$P_u = \frac{\sum U}{T}; P_d = \frac{\sum O}{T},$$

$T = total\ evaluation\ duration$

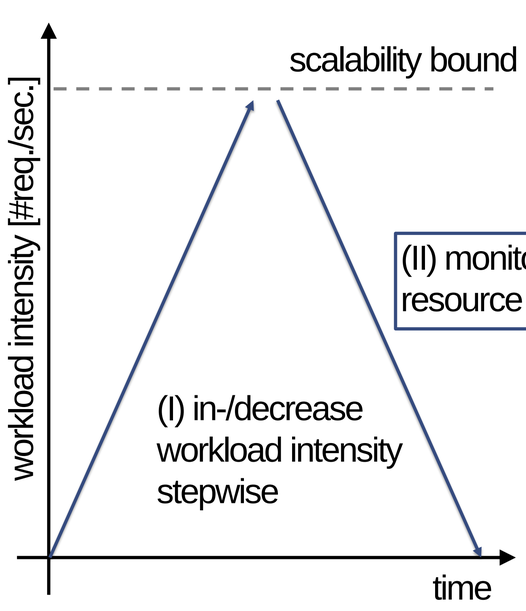
Average precision of scaling up / down

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

Elasticity metric for scaling up / down



Scalability Analysis to derive a matching function



(II) monitor performance & resource allocations/releases

(III) derive discrete matching functions $M(W_x) = R_x$ and $m(w_x) = r_x$

upwards

workload intensity	resource demand
W_1	R_1
…	…

workload intensity	resource demand
w_n	r_n
…	…

downwards

Scalability

- No temporal aspects of how fast, how often, and at what granularity
- Not related to the actual resource demand

Efficiency

- Amount of resources consumed for a given amount of work
- Not limited to resource types that are scaled
- Better elasticity results in higher efficiency

Towards Benchmarking Elasticity

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 in same elasticity dimension and same scaling units

Further details in:
Nikolas Roman Herbst, Samuel Kounev, and Ralf Reussner. *Elasticity in Cloud Computing: What it is, and What it is Not*.
In *Proceedings of the 10th International Conference on Autonomic Computing (ICAC 2013), San Jose, CA, June 24-28*. 2013.