Karlsruhe Institute of Technology

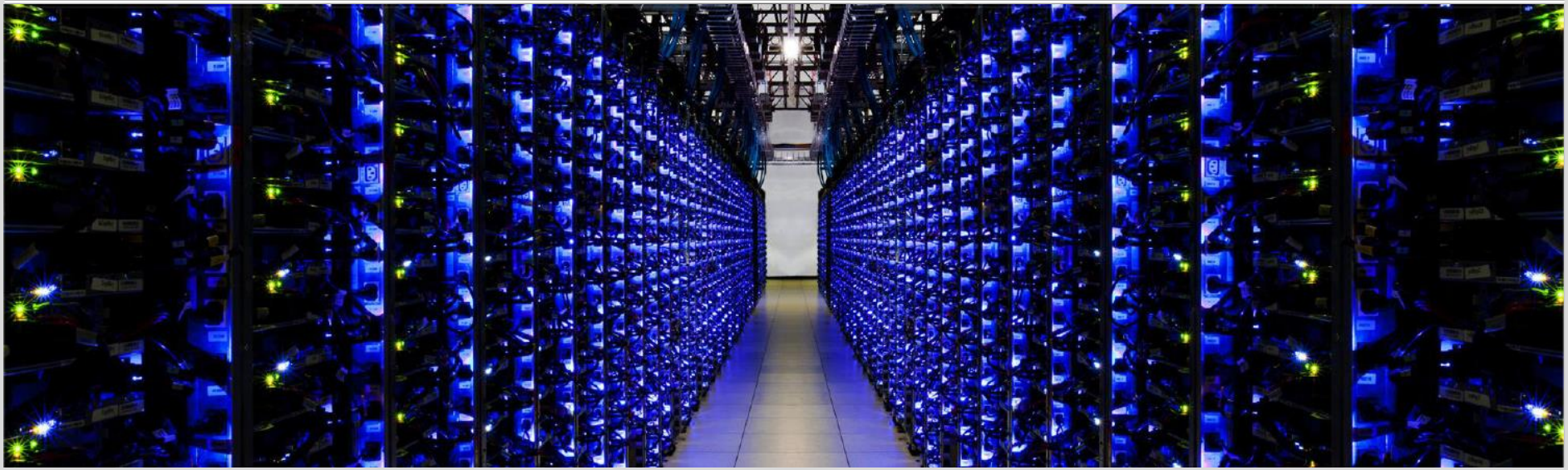
Descartes research

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

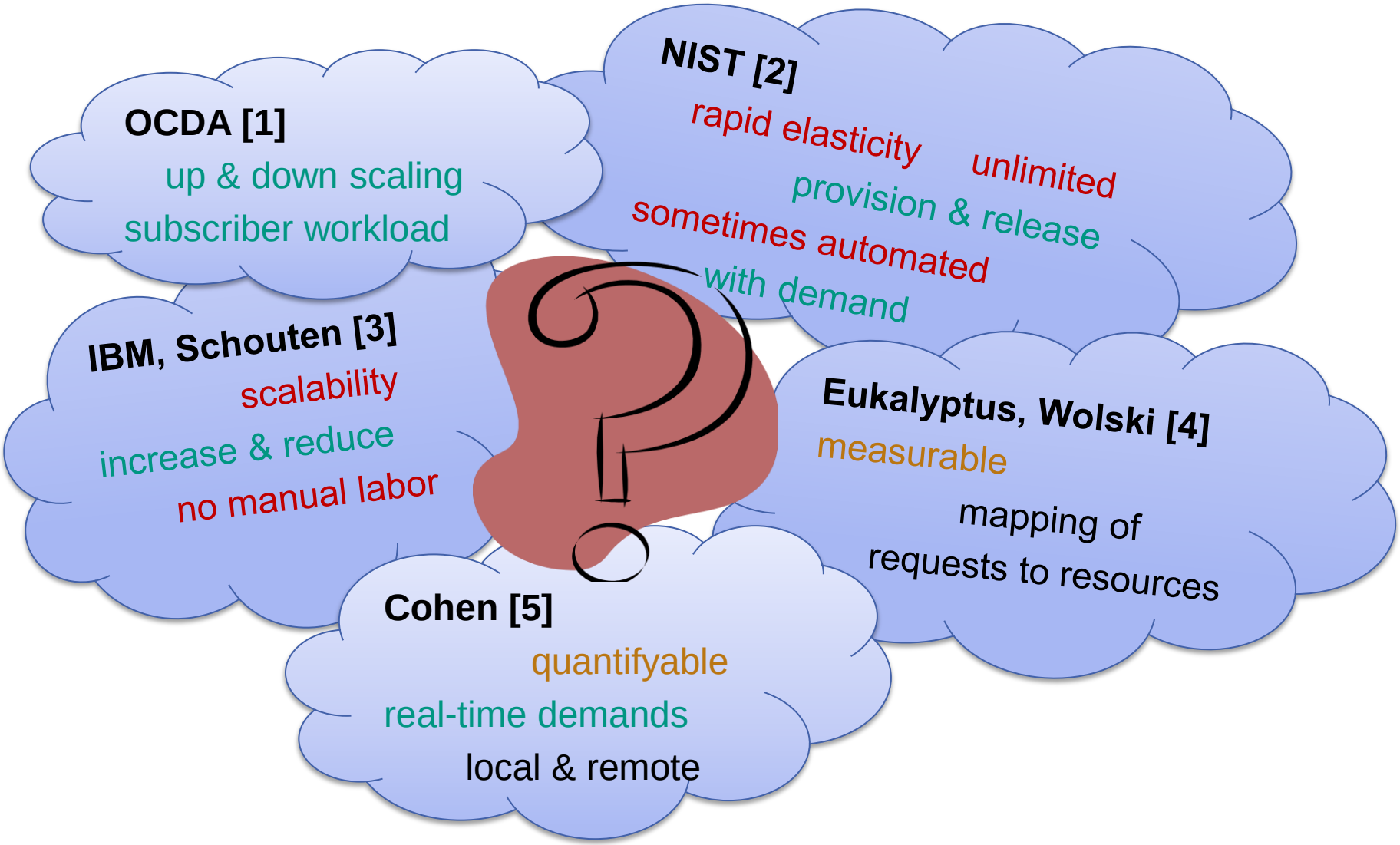
Nikolas Herbst, Samuel Kounev, Ralf Reussner
herbst@kit.edu

ICAC'13, San Jose, CA – 26th June 2013

SOFTWARE DESIGN AND QUALITY GROUP
INSTITUTE FOR PROGRAM STRUCTURES AND DATA ORGANIZATION, FACULTY OF INFORMATICS



What People Say Elasticity Is…

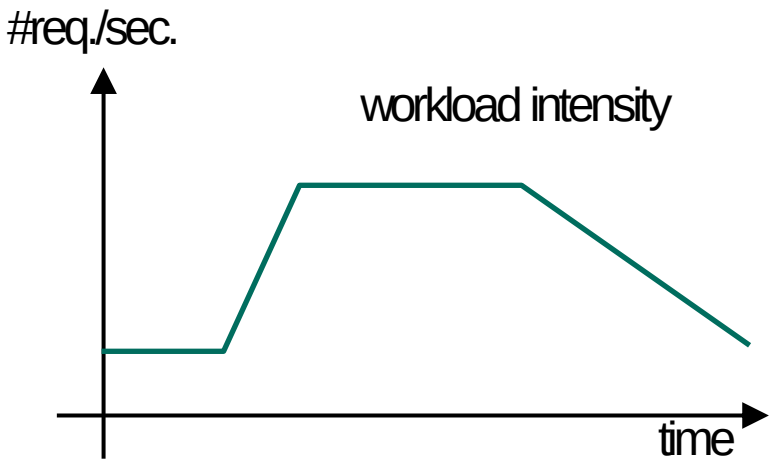


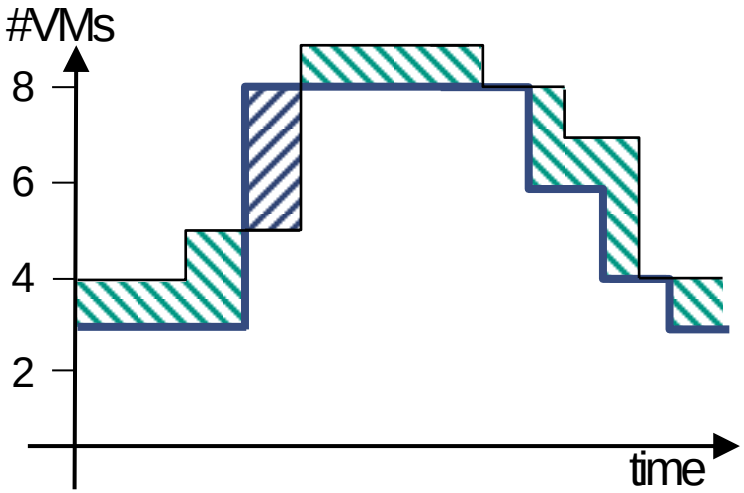
Elasticity Definition

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





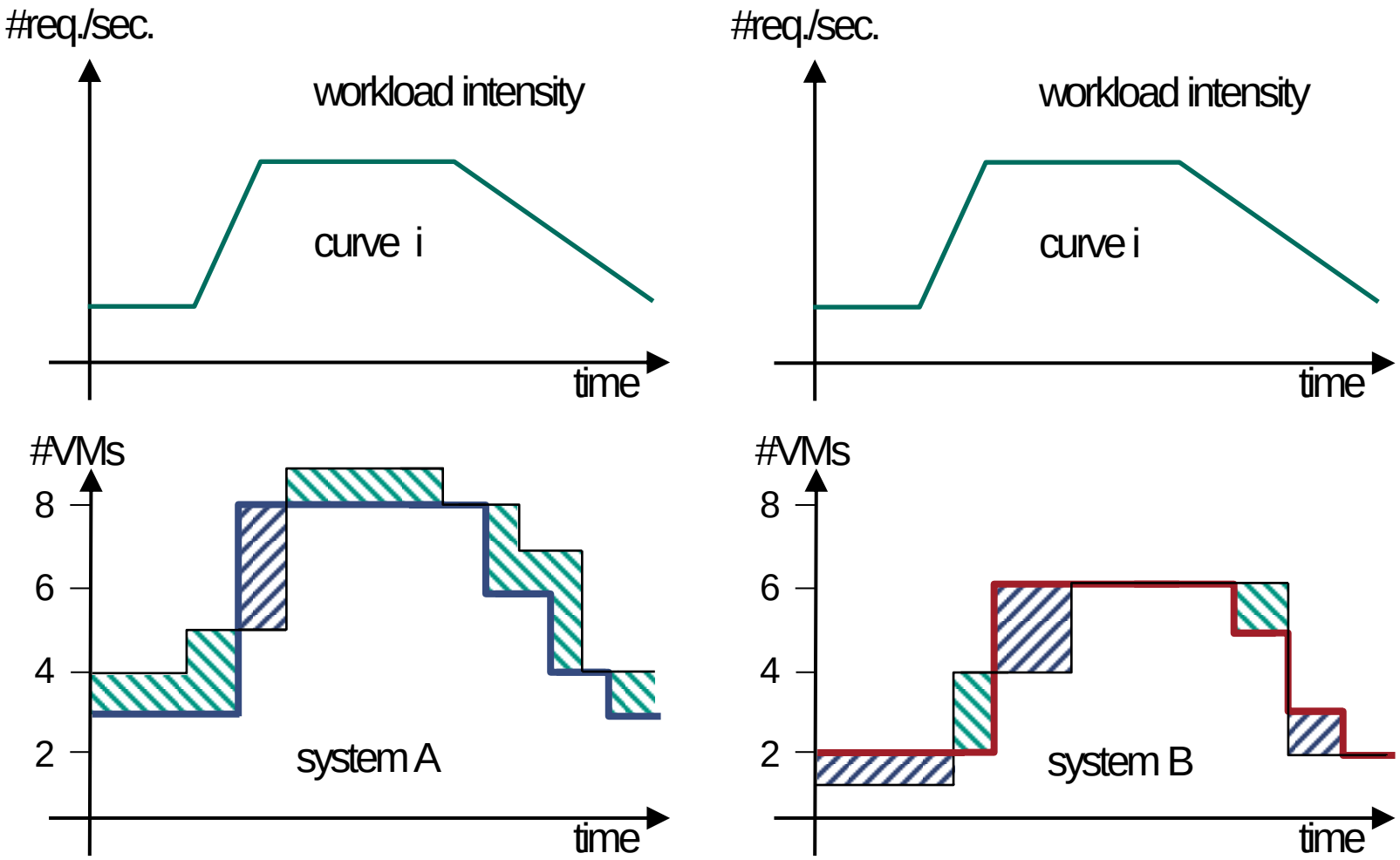
Service Level Agreement (SLA):

E.g.: resp. time ≤ 2 sec, 95%

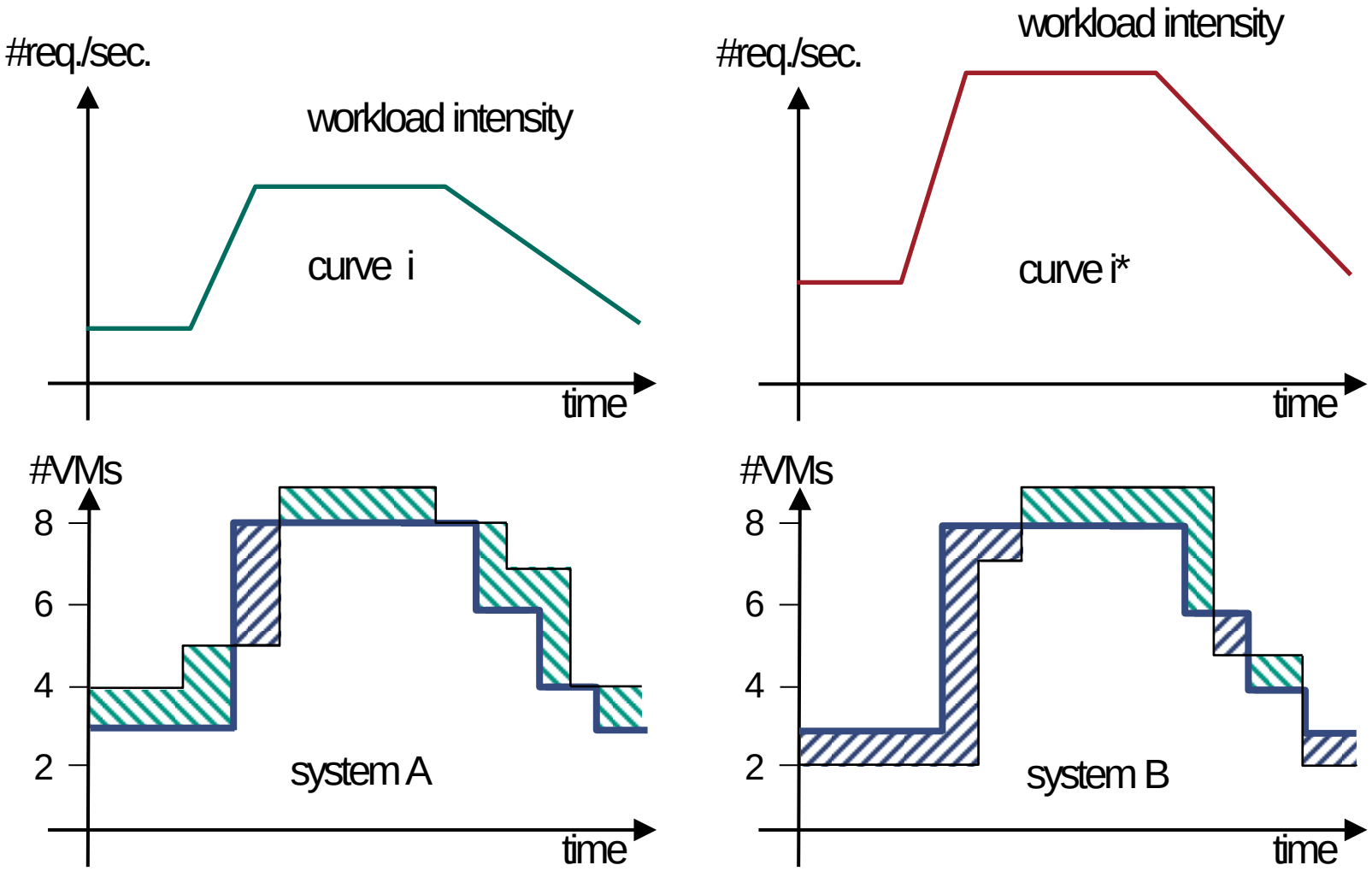
Resource Demand:

Minimal amount of #VMs required to ensure SLAs.

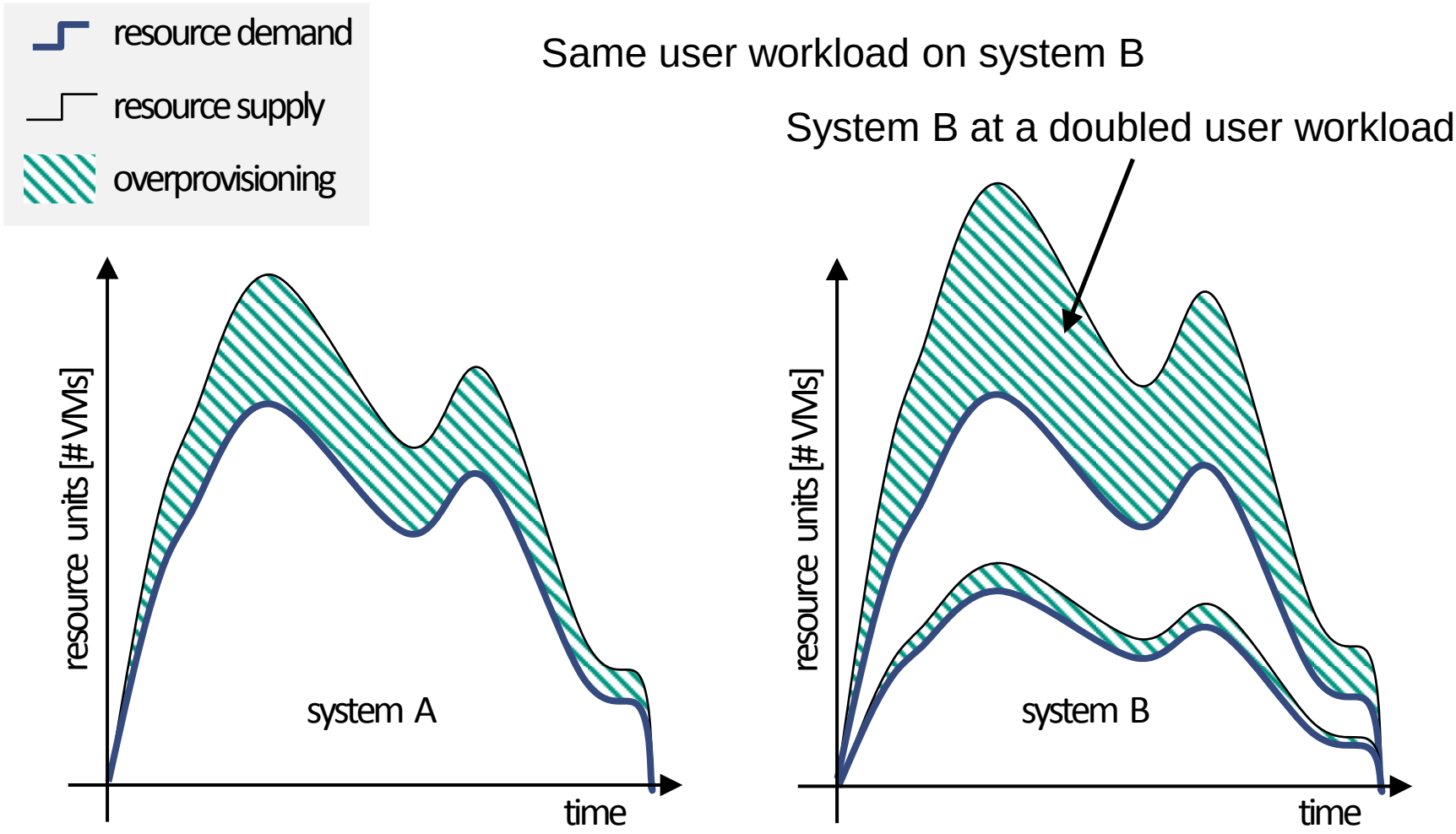
Comparability



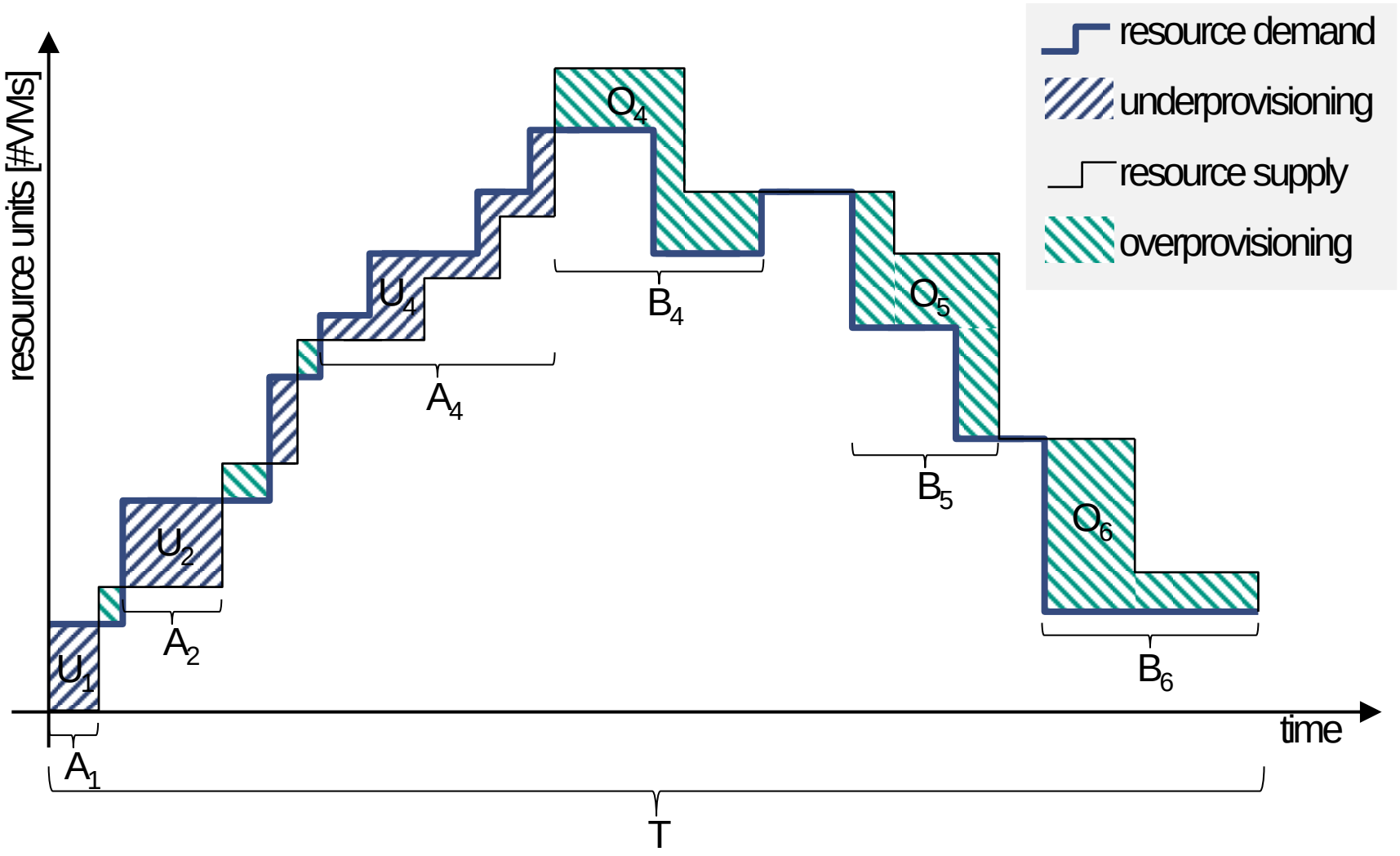
Comparability



Intuitive Elasticity ?



Elasticity Metrics



Elasticity 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

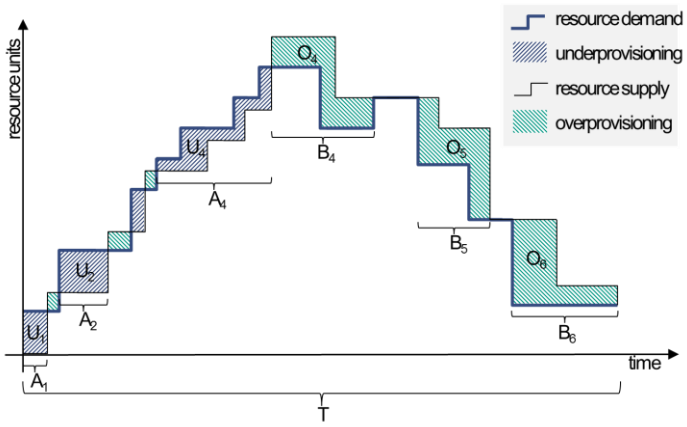
$$P_u = \frac{\sum U}{T}; P_d = \frac{\sum O}{T},$$

$$T = total\ evaluation\ duration$$

Average precision of scaling up / down

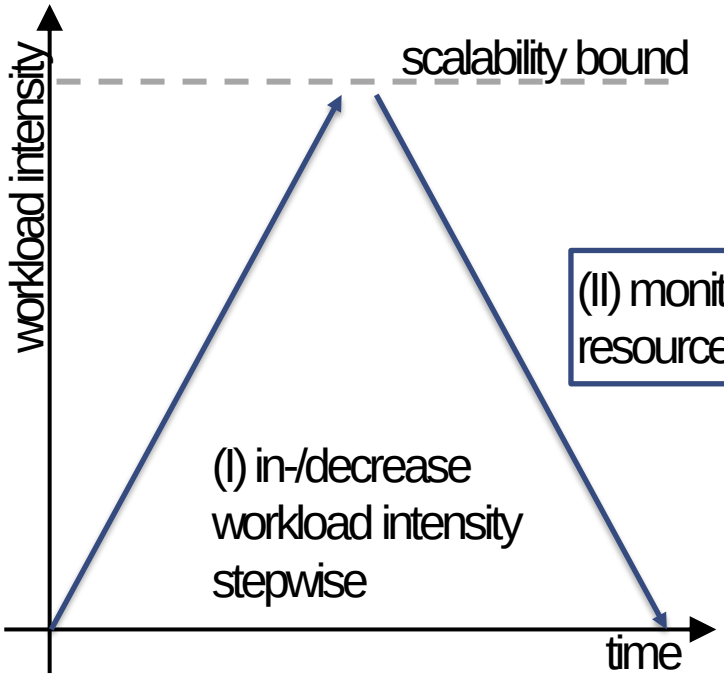
$$E_u = \frac{1}{\bar{A} x \bar{U}}; E_d = \frac{1}{\bar{B} x \bar{O}}$$

Elasticity metric for scaling up / down



Benchmarking Challenges I

→ Derivation of 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
…	…

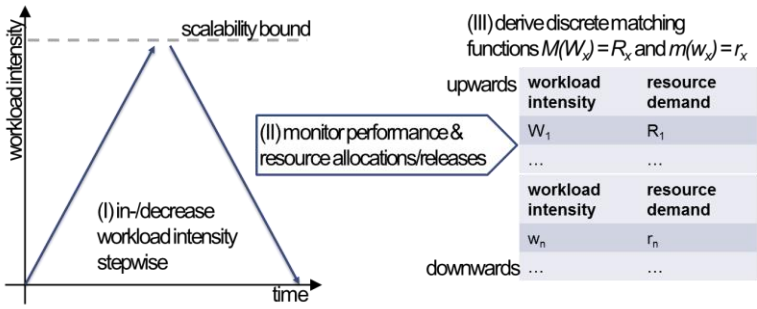
downwards

Benchmarking Challenges II

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.
Defining Elastic Computing, September 2009.
http://www.elasticvapor.com/2009/09/defining-elastic-computing.html,
last consulted Feb. 2013.

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

[3] OCDA. Master Usage Model:
Compute Infrastructure as a Service. Tech. rep., Open Data Center Alliance (OCDA), 2012.
http://www.opendatacenteralliance.org/docs/ODCA_Compute_IaaS_MasterUM_v1.0_Nov2012.pdf.

[4] SCHOUTEN, E.
Rapid Elasticity and the Cloud, September 2012.
http://thoughtsoncloud.com/index.php/2012/09/rapid-elasticity-and-the-cloud/
last consulted Feb. 2013.

[5] WOLSKI, R.
Cloud Computing and Open Source: Watching Hype meet Reality, May 2011.
http://www.ics.uci.edu/~ccgrid11/files/ccgrid-11_Rich_Wolsky.pdf
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?

Backup: Elasticity Core Aspects

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**

Backup: Scalability & Efficiency

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.