Introduction & Background
Resource Elasticity

- Utility Computing (Pay-Per-Use): Ability to dynamically adapt resource allocation based on the client’s demands.

- Existing benchmarking cloud offerings neglect certain aspect of elasticity.

- No good way to benchmark elasticity between
  - Cloud providers; and
  - Elasticity strategies
The existing approaches have limited perspective.

Take specialized approaches
- metrics are part of efficiency measurements, and not elasticity alone.
- Analyze scale out, but not scale down

Take business perspective
- helps make cost-based decisions, not performance-based
- Not scientific or accurate.
- Do not define penalty’s for SLA’s.
New Benchmark

- Quantify resource elasticity in the IaaS context.
- Compare changing resource demand with actual allocation.
- What sets it apart are
  - Strict calibration to match system behaviour.
  - Similar stress loads throughout.
  - Quantify metrics that capture elasticity.
  - Refine metrics iteratively.
Elastic Cloud Architecture

- Scalable Infrastructure
  - VMs with network access and storage
  - Ability to Scale-up or Scale-out

- Management System
  - Reconfiguration Management
    - Allows starting, stopping and reconfiguration of VMs
  - Load Balancer
  - Monitoring System
  - Elasticity Mechanism
Terms & Definitions
Efficiency

- Cost efficiency
- Energy efficiency
- Resource efficiency
Aspects of Scalability

- Fulfilled within a range across a specific quality
- Refers to input variables that are scaled
  - Problem size
  - Number of users
  - Commonly referred to as load intensity
- Measured by evaluating quality criteria
  - Performance measures/dependant variables
  - Ex. Memory consumption, Response time.
Dashed gray line: tolerable response time
Blue line: resources available
Black line: actual response time
Two kinds of Quality Criteria

Service Levels
- Described by measures like response time
- Specified by Service Level Objectives (SLOs)

Resource Amounts
- Measured at different levels of abstraction with varying granularity
  - Physical memory or CPU cores
  - CPU cycles per second
  - VM server images
  - Threads or locks
Scalability in the Cloud

- Cloud customers require constant service levels over variable load intensity to satisfy quality criteria defined in SLOs.
- Scaling behavior is characterized by resource amounts that vary over variable loads.
- Thus, Resource Scaling is the term used to emphasize that it is the underlying resources used that we are referring to.
“Elasticity is the degree to which a system is able to adapt to load changes by provisioning and deprovisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible.”

Scalability vs. Elasticity

- Scalability - degree to which a system can adopt to a varying load conditions.
- Resource Elasticity - the quality of the adaptation to a varying load conditions.
System A and B are equal except for their elasticity mechanisms. The elasticity mechanism of system B is superior compared to system A since it is able to match the demand closer.
Elasticity is the result of the adaptation process that scales the resources according to the load intensity.
- Automated mechanism
- Manual steps sometimes may be required
Resource Type and Scaling Unit

- **Base resources**
  - CPU, Memory or Disk Storage and Container Resources.

- **Scaling Unit**
  - Different resources have different measurement units
    - CPU time
  - Need to be the same across the various systems to be able to compare
Scalability Bounds

- Scalability is limited.
- Depends on
  - Maximum amount of physical resources
  - Service level constraints
- Can compare two systems if such system can scale within the same bounds.
Evaluation Aspects

- Accuracy
  - Whether the system can provision enough resources and match the demand precisely
  - Under-provisioning and over-provisioning hurts accuracy

- Timing
  - How quickly a system can react to the change in load conditions
Both systems adapt to the load changes instantly, but system C constantly over-provisions resources, while system D under-provisions and both systems do not match the demand very precisely.
Evaluation Aspects

Systems with imperfect timing. System E always lags behind the changes in demand, while system F changes the available resource too often without real change in demand. System F can be seen as having bad timing and bad accuracy.
Elasticity Mechanisms

- Resources are allocated according to changing demand
- Usage of different scaling methods
  - scaling vertically or horizontally
- Retractive elasticity
  - scalability happens after the change in demand
- Proactive elasticity
  - system can predict demand changes and scale accordingly before the demand increase or decreases
Proposed Benchmark
Load Intensity Variation

- Elastic behavior is triggered by a change in load intensity
- Realistic load is needed for proper benchmarking
- Load patterns or load profiles are used for benchmarking
  - bursts
  - linear trends
  - general variability over time interval
- LIMBO toolkit
Resource demand can be different under the same load
  - ex. System H has faster CPU than System G, as such it needs less CPU utilization for the same load.
  - creates a need for system calibration

- to test elasticity we need to trigger scalability under same load profile but not necessary the same load
Benchmark Calibration

- Iteratively analyze scalability potential of a system
  - small increment
  - elastic mechanism is turned off
  - scaling is controlled manually
    - provision / deprovision resources manually as load changes
  - helps establish load intensity levels needed to maintain load profile for benchmarking
Metrics
Metrics

- Average time to switch from an under-provisioned state to optimal or over-provisioned
- Average time to switch from an over-provisioned state to optimal or under-provisioned
- Accumulated time in under-provisioned / over-provisioned state state
- Average amount of under-provisioned / over-provisioned resources
- Total duration of tests
