SPEC RG CLOUD WG Telco

George Kousiouris, Athanasia Evangelinou

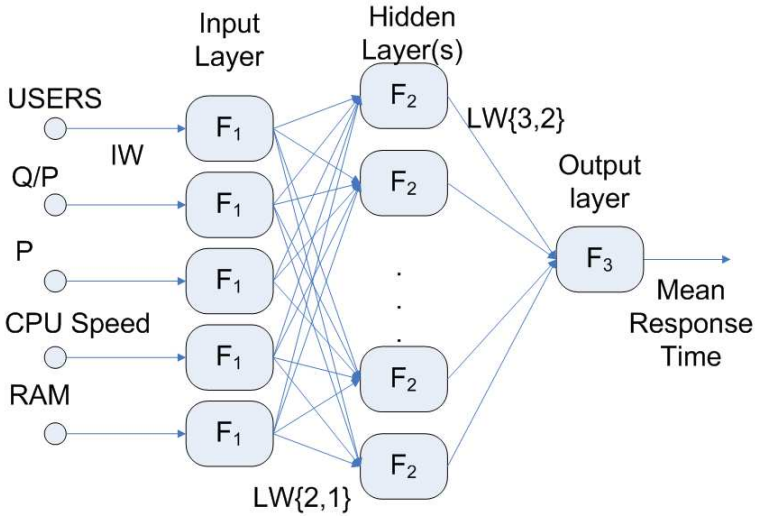
15/4/2014

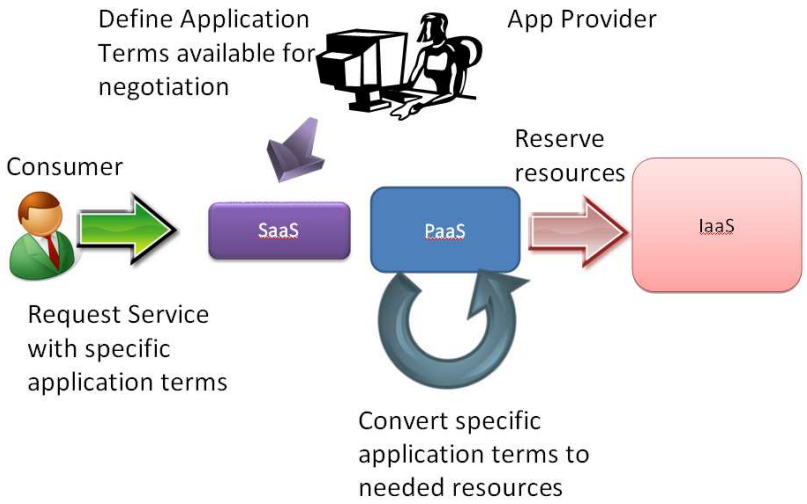
Research Group Info

- DKMS Lab of National Technical University of Athens, School of ECE
  - http://grid.ece.ntua.gr/
- Key research areas
  - Cloud Computing
  - SLAs
  - Social media and networks
- ~35 people, led by Prof. Theodora Varvarigou
- Very active in FP6 and FP7 research projects (mainly EC SSAI Unit)

Past efforts

- Application benchmarking on virtualized infrastructures to create models (based on ANNs) for SLA translation of application terms to resource level attributes





Current scope

- Performance isolation
- Enable performance guarantees on computation –based SLAs
  - Currently available for networks, storage but not computation
  - IaaS level research
- Select service offerings with fittest performance characteristics
  - External to IaaS

Motivation (1/2)

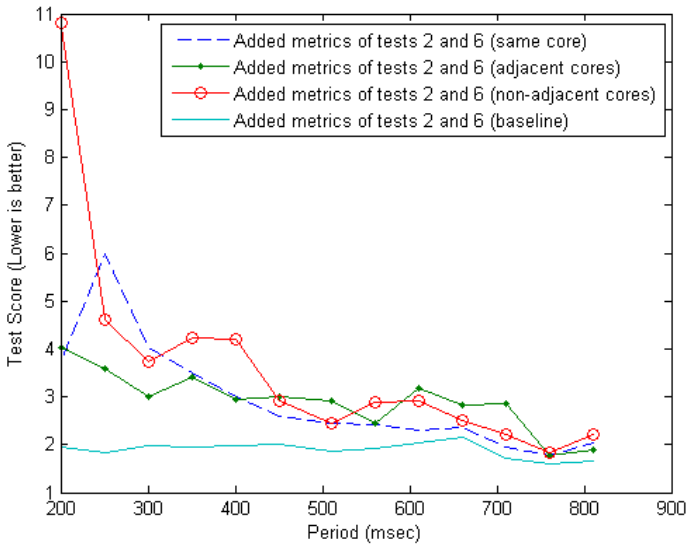
- Cloud Services
  - Innovative IT provisioning model
  - Promises for infinite resources and on-demand scalability
  - Performance?
- Varying performance and instability of Cloud Services because of:
  - Multitenancy and shared resources
    - Noisy neighbour
  - Hardware differences in Data Centers
  - Black box provider management
- Cloud Provider performance effects evident after application migration to the Cloud
  - A successful cloud migration means saving money and guaranteeing stability
- Must know in advance provider performance stability characteristics

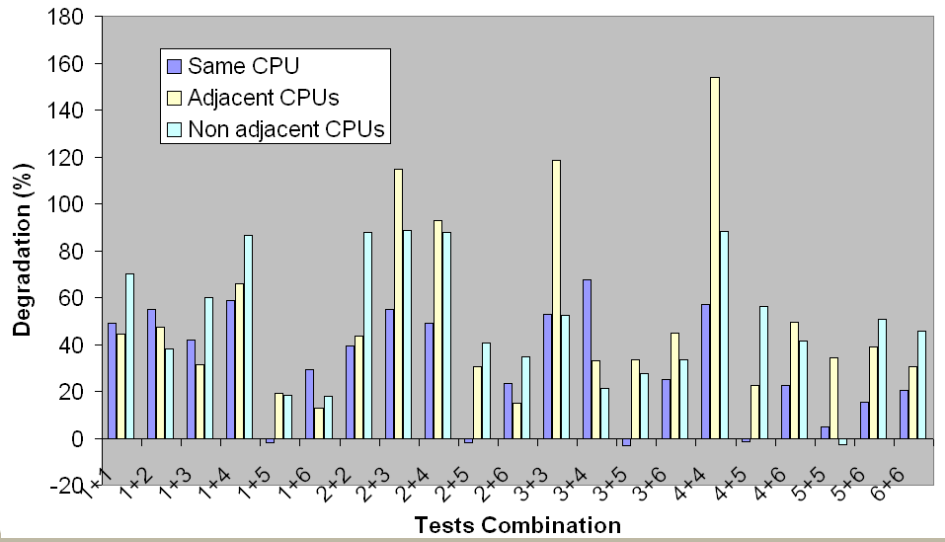
Motivation (2/2)

- Experiment with two VMs on a quad core CPU
- 6 benchmarks (Matlab benchmarks) scheduled in all possible combinations of 2
- Usage of real time scheduling to limit task usage of a core
- Severe degradation of VM performance

- Ability to predict degradation

- More info on
- George Kousiouris, Tommaso Cucinotta, Theodora Varvarigou, "The Effects of Scheduling, Workload Type and Consolidation Scenarios on Virtual Machine Performance and their Prediction through Optimized Artificial Neural Networks , The Journal of Systems and Software (2011),Volume 84, Issue 8, August 2011, pp. 1270-1291, Elsevier, doi:10.1016/j.jss.2011.04.013."





Placement optimization for minimizing degradation

- Multi-objective optimization to distribute VMs on physical nodes
  - Kleopatra Konstanteli, Tommaso Cucinotta, Konstantinos Psychas, Theodora A. Varvarigou: Admission Control for Elastic Cloud Services. IEEE CLOUD 2012: 41-48

What if we are not in the IaaS level?

- Macroscopic view is needed
- Provider service capabilities descriptions very limited and vague
  - E.g. Amazon ECU
- Mechanism for measuring externally the performance of various Cloud services(supports multiple Cloud Providers)
- Measuring of service performance by using abstracted and simple metrics (combination of cost, performance, deviation and workload)
- In the context of the FP7 ARTIST project
  - http://www.artist-project.eu/

Needs

- Different services may behave differently across various application domains
  - E.g. memory-optimized, graphics-optimized, computation-optimized
- More abstracted and common way should be found for identifying performance aspects of Cloud Environments
- Generic tools for multi-provider and multi-benchmark tests
- Key aspects of the benchmarking process
  - Iterated over time(different hardware/managements decisions included in the refreshed metric values)
  - Observe key characteristics(performance variation, standard deviation)
  - Cover a wide range of diverse application types

Related Work

- CloudHarmony.com
  - Vast number of benchmarks against various Cloud services
  - Offering results through API
  - No sufficient repetition of measurement process

- CloudSleuth.net
  - Focus on web-based applications and their response time/availability
  - Deploy and monitor across different cloud providers

Benchmarking approach in ARTIST

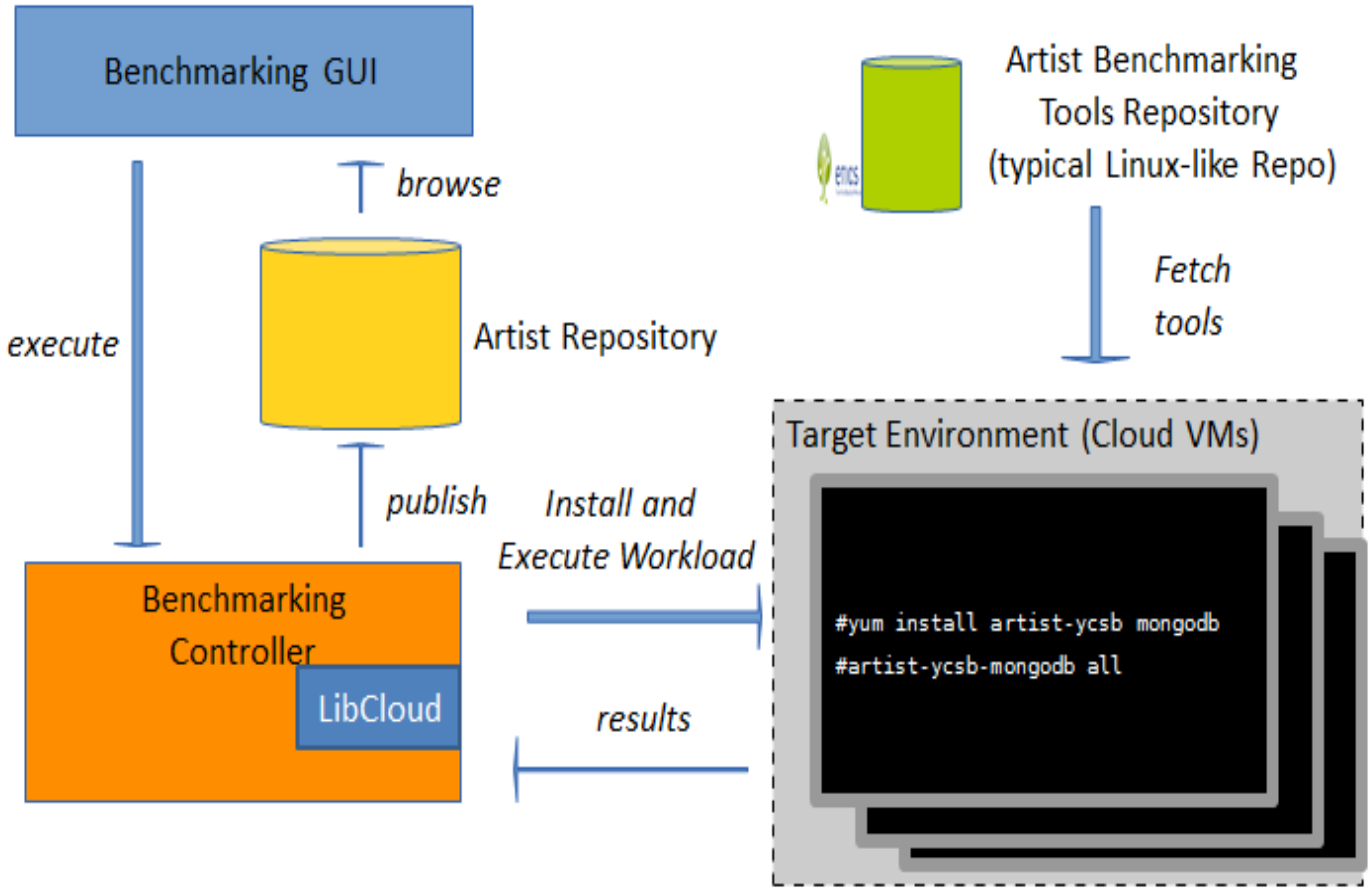
- Identification of a set of popular application types and the respective benchmarks
- Framework able to automatically install, execute and store benchmark results
- Multiprovider capabilities (through Apache LibCloud)
- Define comprehensible metrics

Application Benchmark Types

- Abstracts the question for best performance in the following format:
  - "**What is the best offering for my streaming application**"
  - Abstraction of question to non performance-aware individuals

Benchmark Test	Application Type
YCSB	Databases
Dwarfs	Generic Applications
Cloudsuite	Common web aps like streaming, web service etc.
Filebench	File System storage with specific workloads (e.g. mail servers etc.)
DaCapo	JVM aspects

Benchmarking Suite Architecture



Service Efficiency Metric Description

- We need to abstract further the user question and adapt it to specific user interests with relation to cost, performance, deviation etc.
- "What is the best offering to run my streaming application when I want a **cheap** service for **low** workload?"
- Related work: Service Measurement Index (SMI-Garg, 2012)
  - interesting features for the ranking (performance, sustainability, suitability, accuracy, interoperability, reliability , cost, usability etc.)
  - some factors are difficult and arbitrary to calculate (e.g. usability, interoperability) or need human intervention
  - required information is not provided by Cloud providers (e.g. Mean Time Between Failures for reliability)
  - performance is only considered in terms of response time, thus being applicable only in cases of web based offerings and not adapting to various application types
  - Too complex for the average user

Requirements and Formula of Service Efficiency Metric

- Workload aspects of a specific test
- Cost aspects of the selected offering
- Performance aspects for a given workload
- Weighted rankings based on user interests
- Intuitively higher values would be better
- Normalization for different value ranges of the parameters

$$SE = \frac{\sum_i s_i l_i}{\sum_j s_j w_j f_j}$$

Where s: scaling factor for normalization
l: workload metric
f: KPI or cost metric
w: weight factor

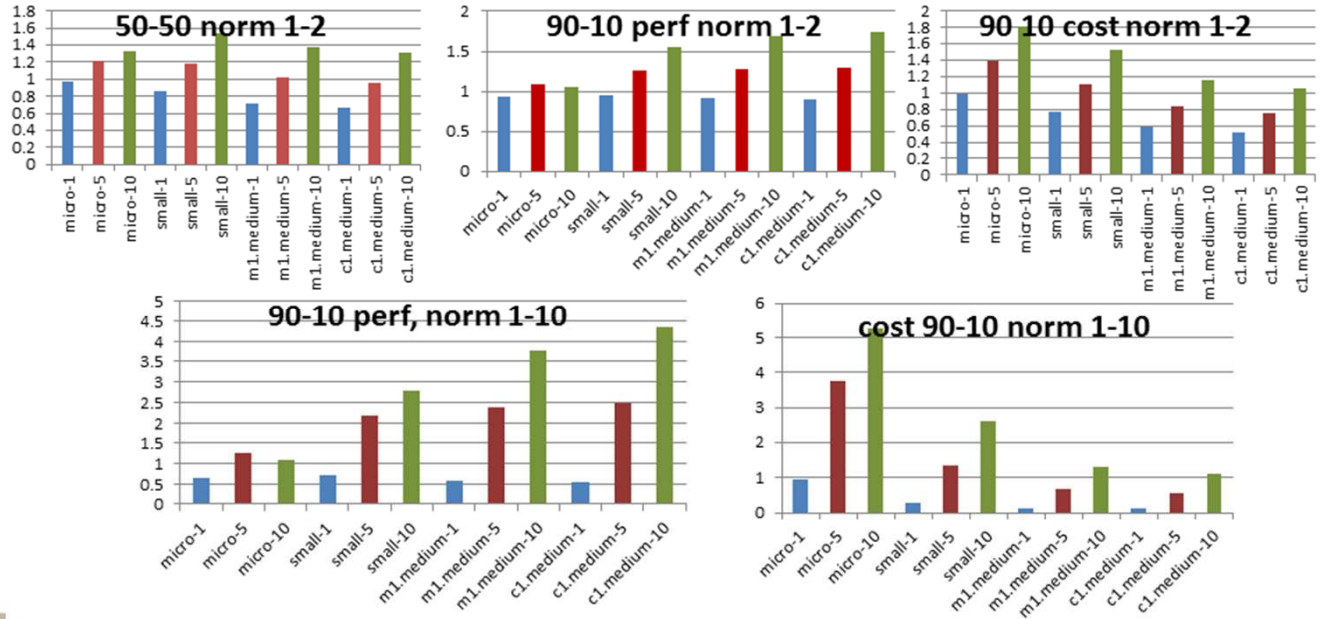
Metric Case study on Amazon EC2

- Application: Web Server for on-line time series prediction (Matlab back-end)
- Different VM sizes (micro, small,c1.medium,m1.medium)
- Number of concurrent clients(1,5,10-heavy workload)
- Different weights were given to the performance and cost aspects(50-50,90-10,10-90)
- Normalization intervals for metric's sensitivity
  - 1-2
  - 1-10
  - Avoided (0-1) due to infinite values

$$SE = \frac{\#Clients}{w_1 * \text{delay} + w_2 * Cost}$$

Results

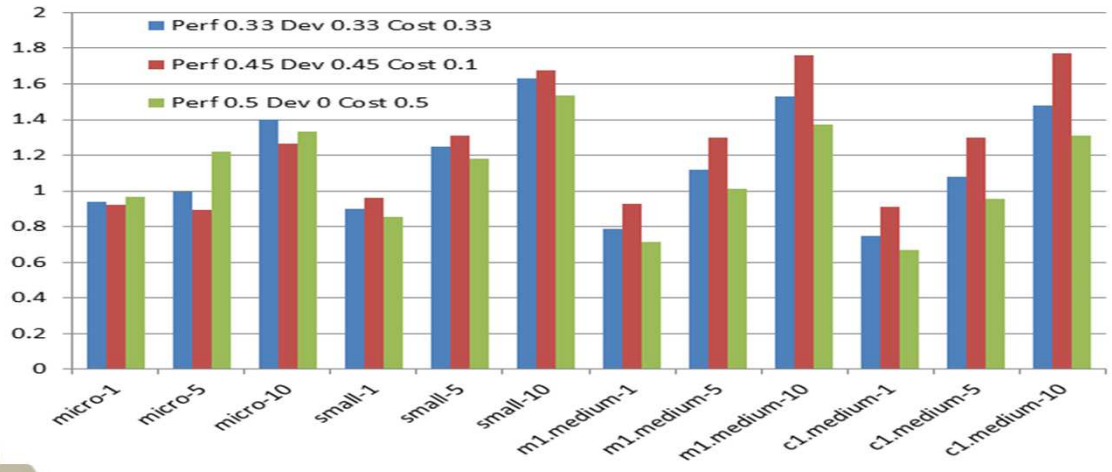
- In general there are cases in which selection is obvious without the usage of a metric (high interest in performance and high workload would direct us to largest instance)
- In other more borderline cases (e.g. 50-50 or 90-10 performance with low workload) it is not obvious which type to choose
  - For these cases the metric can aid us in selecting



Metric Case study on Amazon EC2

- Incorporation of the standard deviation
- Best selection changes
  - Green: small instance, Blue: small instance, Red: c1.medium
- not necessarily due to better stability, could be also lower cost importance

$$SE = \frac{\#Clients}{w_1 * \text{delay} + \text{w}\,2 * \text{deviation} + w_3 * Cost}$$

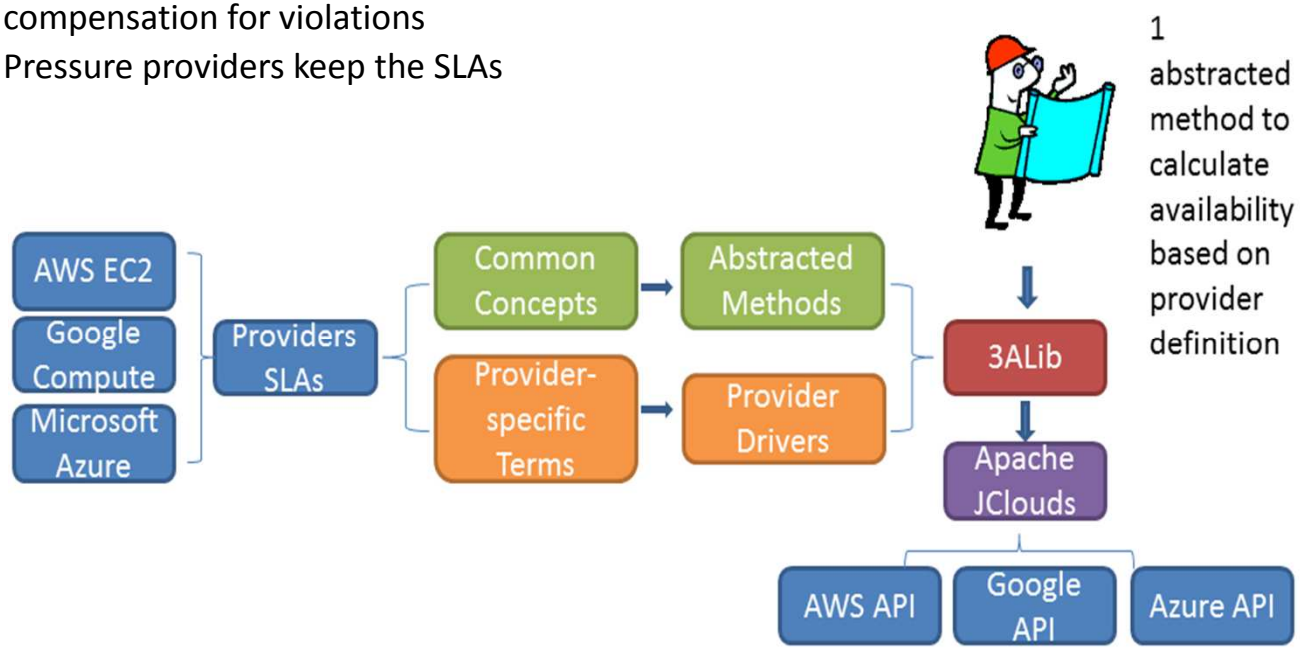


Future Work

- Complete the framework integration
  - Pending GUI and Controller integration
- Investigate addition of other measurable non functional properties
  - E.g. measured availability
- PaaS level metrics and options investigation
- Apply the metric based on the selected application types and relevant benchmarks
  - Not performed currently due to parallel work on the two topics (benchmark selection/controller implementation and metric form investigation)

3ALib (Availability Benchmark)

- Every provider states that the user must provide proofs of the violation
- Abstracted Availability Auditor Library (Java-based)
  - Each provider has their own SLA definition (availability formula and preconditions)
  - Implementation based on the conceptual abstractions of different providers SLAs
- Purpose
  - Align availability monitoring with specific provider definitions
  - Check preconditions of SLA applicability for a specific deployment and give feedback
  - Monitor and log SLA adherence levels in a consistent manner with the provider definitions and claim compensation for violations
  - Pressure providers keep the SLAs



Thank you!

- For more info:
- gkousiou@mail.ntua.gr
- aevang@mail.ntua.gr