vrije Universiteit amsterdam

Resource Provisioning of Web Applications in Heterogeneous Cloud

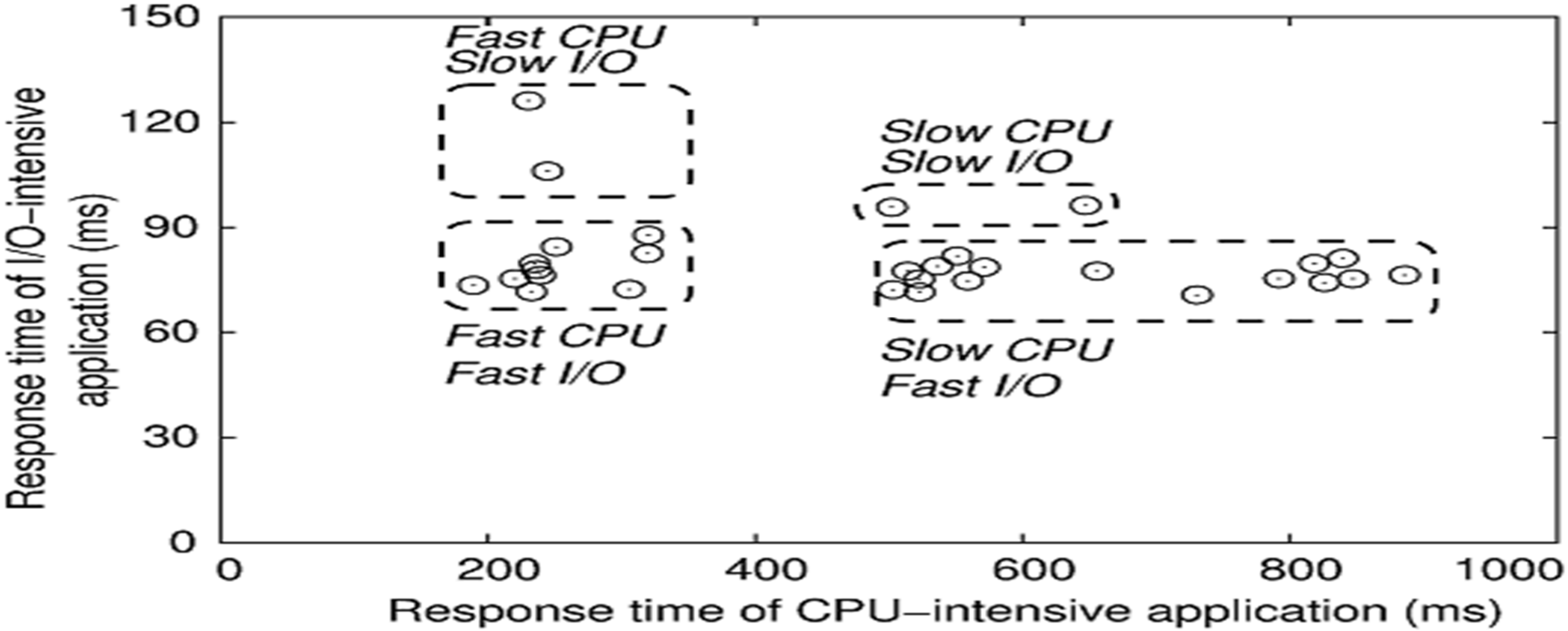
Jiang Dejun
Supervisor: Guillaume Pierre
2011-04-10

Background

- Cloud is an attractive hosting platform for startup Web applications
  - On demand resource provisioning
  - Pay-as-you-go model
- Web application performance is one primary concern when moving to Cloud
  - 83% of respondents worried about cloud performance (a survey conducted by IDC 2010)
  - Dynamic resource provisioning helps to guarantee Web application performance

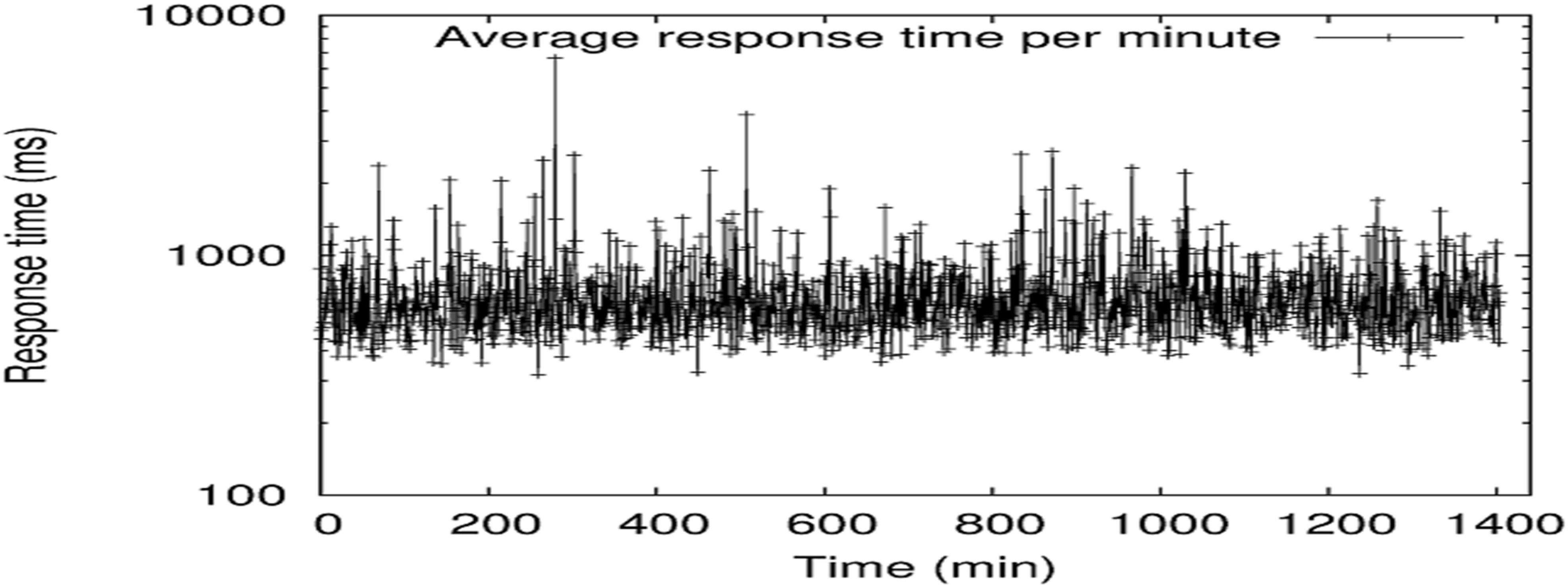
Motivation

- Cloud resource is heterogeneous
  - Heterogeneous virtual machine types
  - Heterogeneous performance of same type



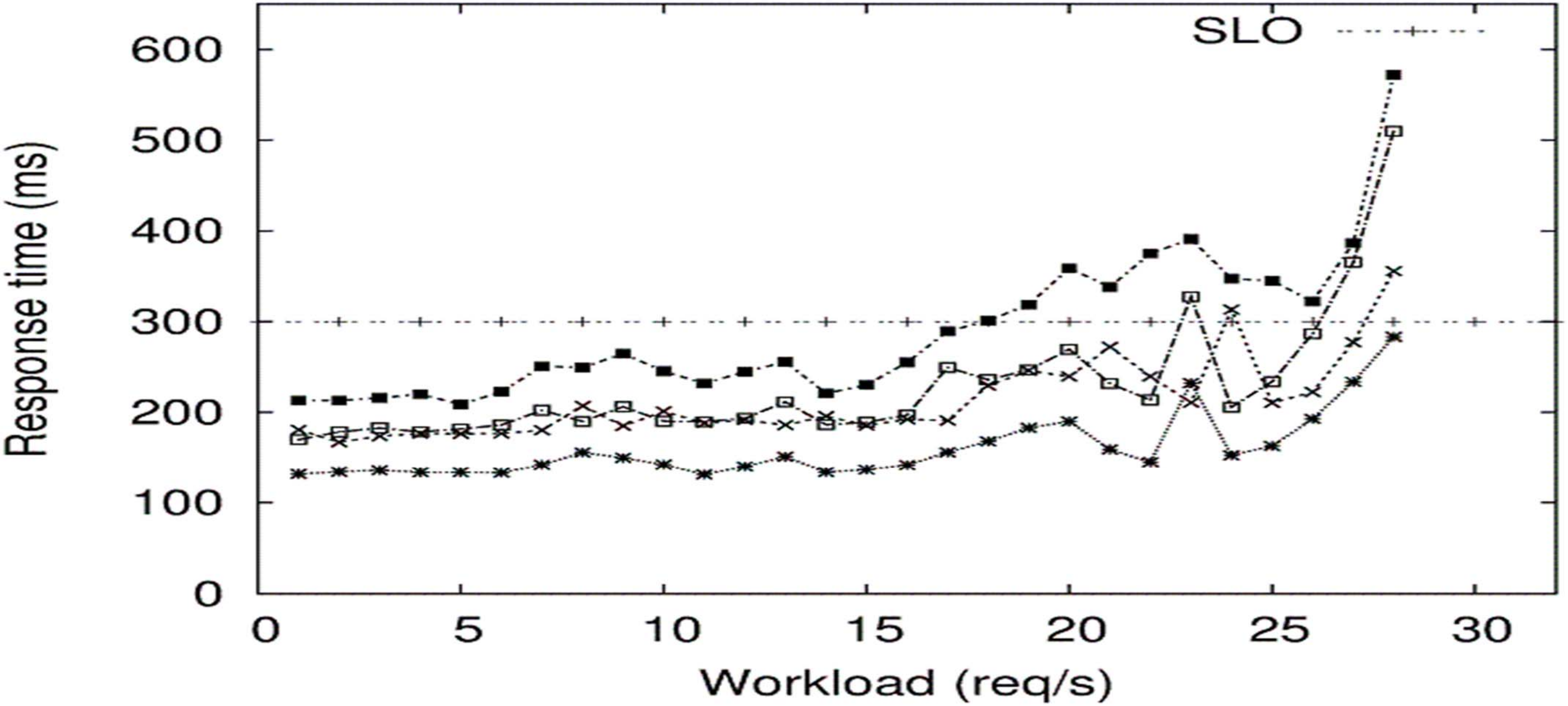
Motivation(cont.)

- Cloud resource is heterogeneous
  - Resource heterogeneity is a long-term observation
  - Resource heterogeneity is observed cross Clouds (e.g. EC2, Rackspace)



Motivation(cont.)

- Cloud resource is heterogeneous
  - Current resource provisioning in Clouds (e.g. EC2)



Problem statement

- How to provision Web applications in Clouds
  - If an instance with fast CPU, it may be better to use it as an application server
  - If an instance with fast IO, it may be better to use it as a database server
  - We do not know how to use the new instance but we need to make a decision
- Difficulties
  - Unpredictable performance of new instances
  - Different performance benefits on different tiers of a new instance

Intuitive solutions

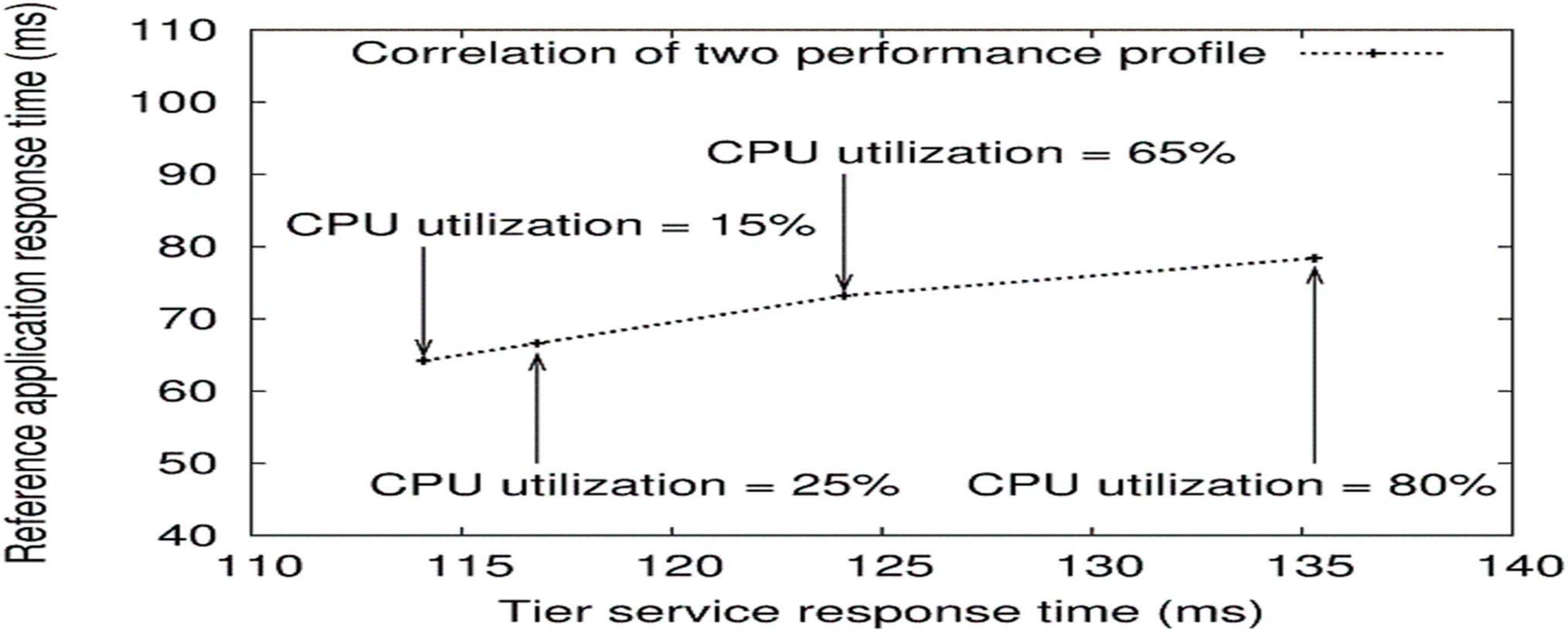
- Ignore the heterogeneous resource feature
  - Apply current resource provisioning algorithm to make decision
- Profile new instances at each tier to make decision
  - Deploy new instance as application server is fast
  - Deploy new instance as database server costs. e.g. DB size: 1.6GB. Dump: 190s; Transfer: 64s; Import 1530s. Total 30 min
  - This approach is inefficient and time-consuming

Outline

- Background
- Motivation
- Problem statement
- Intuitive solutions
- Our proposal
- Experimental evaluation
- Conclusion

Our proposal

- Performance correlation
  - Performance profile of a given tier is related to its resource utilization
  - Performance profiles of two different tiers (with same type resource demand) can be highly correlated



Our proposal(cont.)

- Performance prediction
  - Step 1: Employ reference applications as the calibration base
  - Step 2: Correlate resource demands of reference applications and tier services on the calibration instance
  - Step 3: Profile new instances with reference applications
  - Step 4: Derive performance of tier services on new instance

Our proposal (cont.)

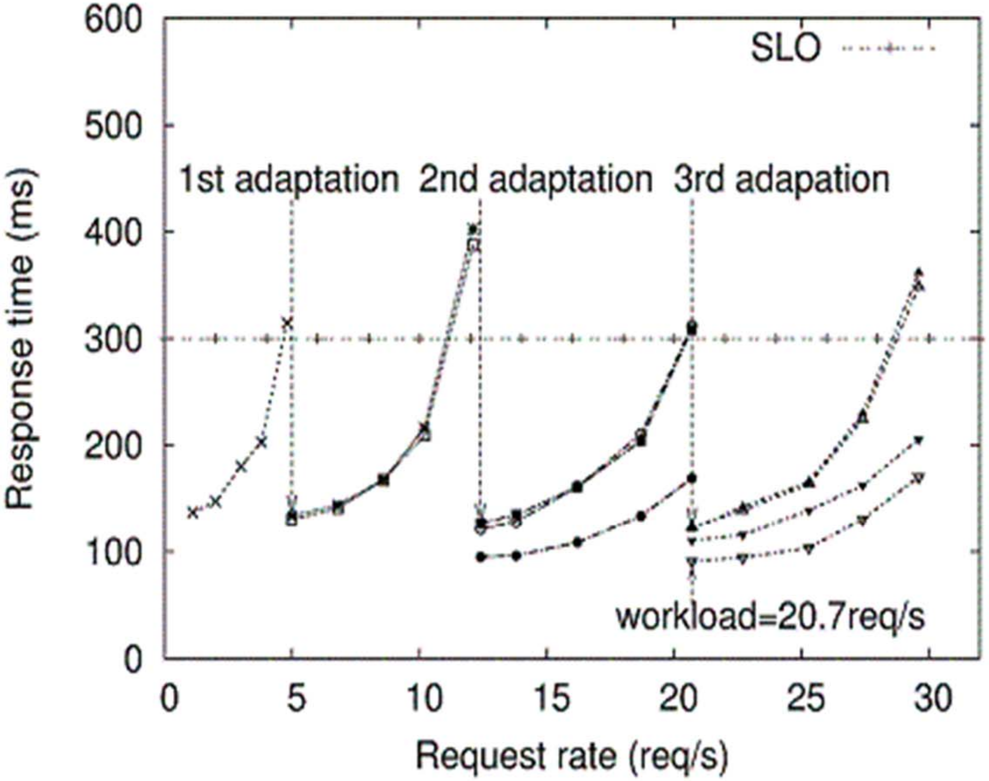
- Resource provisioning
  - Obtain performance profiles of new instances
  - Apply "what-if" analysis to predict the performance of the whole application if a new instance is added to a tier

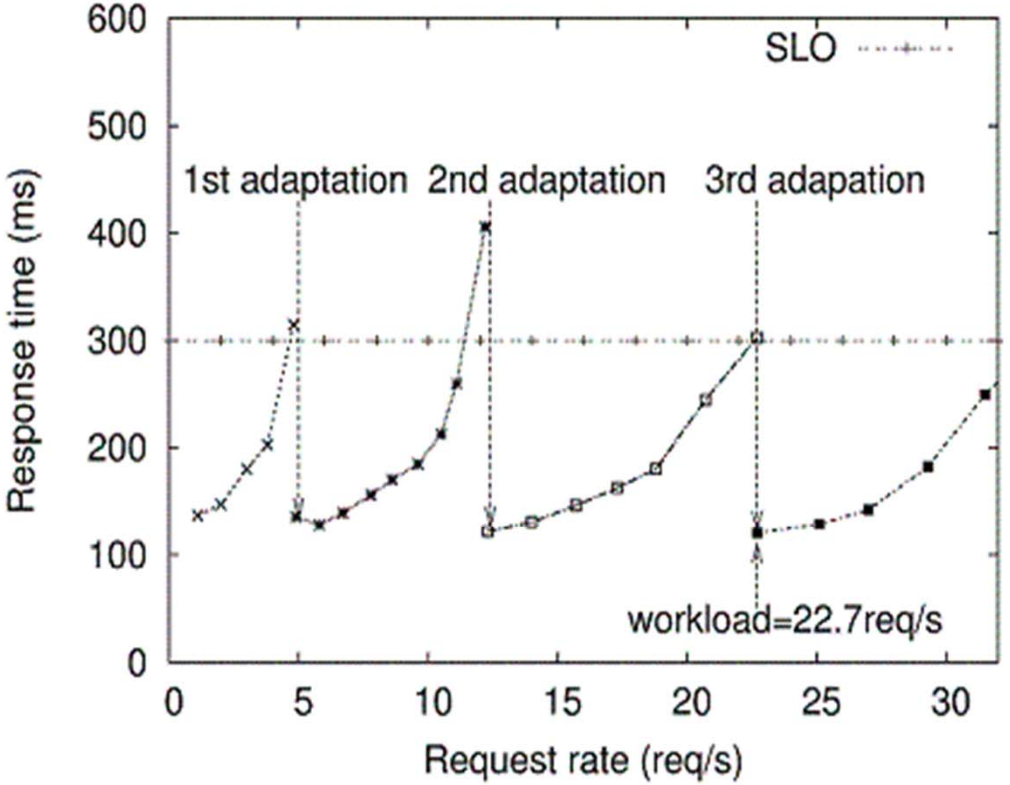
Experimental evaluation

- Experiment setup
  - Reference applications
    - a CPU-intensive application: CPU(ref)
    - an IO-intensive application: IO(ref)
  - Tested application: TPC-W (a benchmark modeling the online bookstore)
  - All experiments run on Amazon EC2

Experimental evaluation

- Importance of adaptive load balance adapting to capacities of backend instances

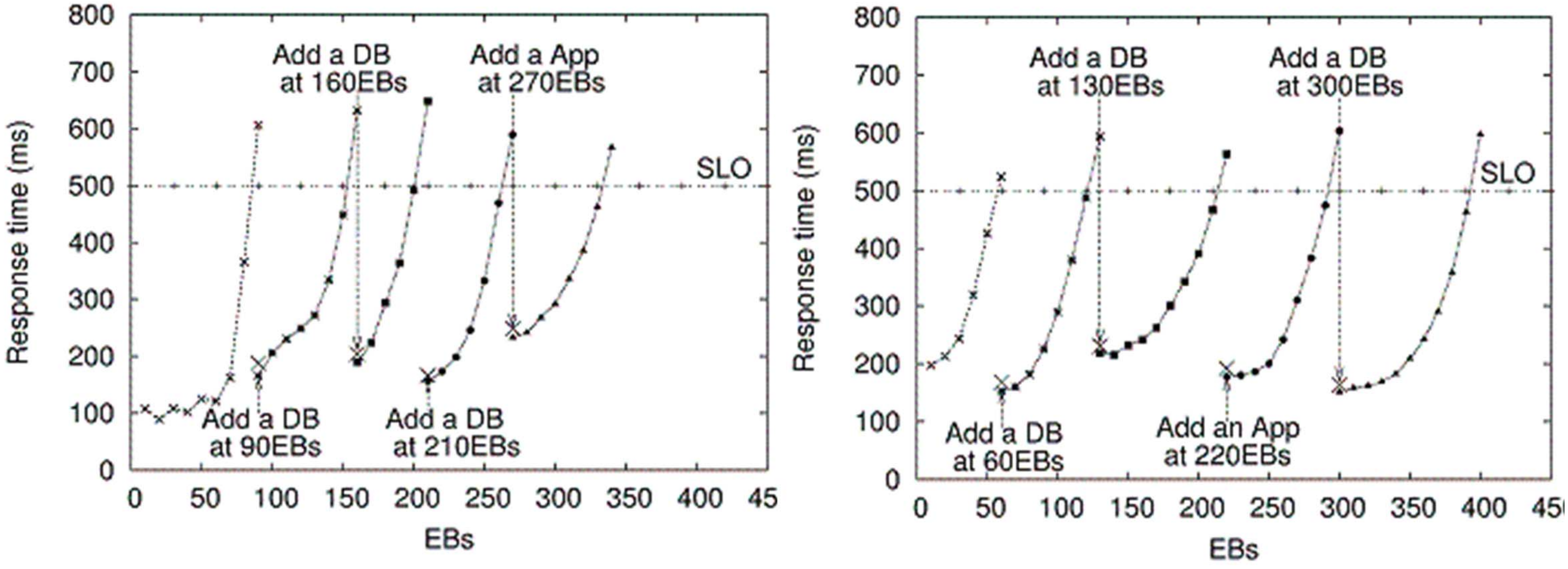




Our system has equal response times from each application server running CPU(ref) under increasing workload

Experimental evaluation

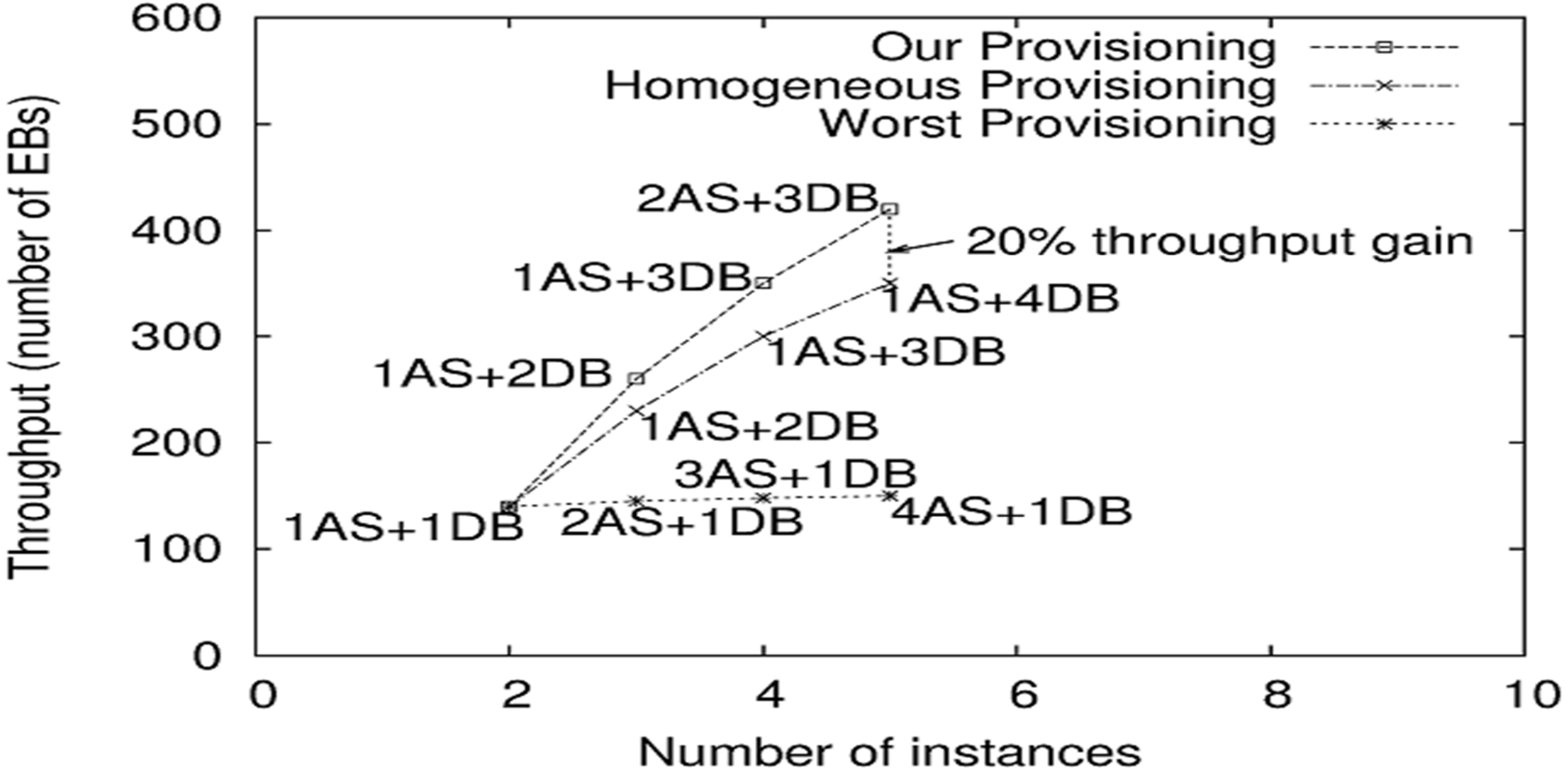
- Effectiveness of our system to provision TPC-W in Cloud



We have different adaptions in two groups of experiments when provisioning TPC-W on EC2 due to resource heterogeneity

Experimental evaluation

- Comparison with other provisioning techniques



Our system achieves 20% more throughput using the same instances compared with the homogeneous provisioning technique

Conclusion

- Guarantee performance of Web applications in Clouds is important
- Cloud is heterogeneous to make current resource provisioning difficult in it
- We propose to correlate resource demands of hosted applications with reference applications.
- One can derive the performance of Web application on new instances by just profiling new ones with reference applications.

Thank you!

Questions?