Regression Based Multi-tier Resource Provisioning For Session Slowdown Guarantees

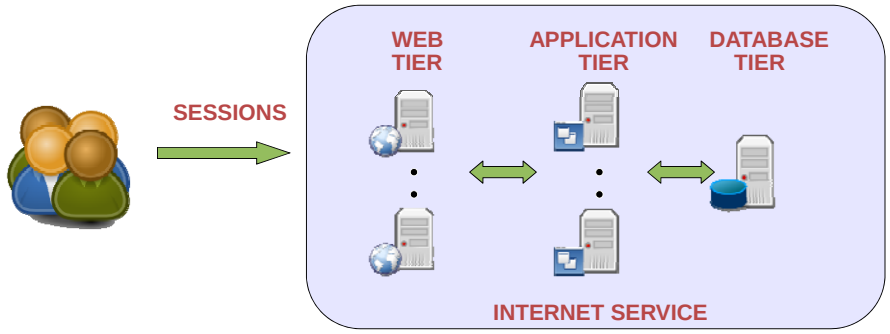
Sireesha Muppala
Xiaobo Zhou
Department of Computer Science
University of Colorado at Colorado Springs

Liqiang Zhang
Department of Information & Computer Sciences
Indiana University

Outline

- Multi-tier Internet service architecture
- Dynamic resource provisioning : Challenges
- Related work
- Novel session based performance metric
- Regression based dynamic resource provisioning
- Performance evaluation
- Conclusions

Typical 3-tier Internet Service Architecture



- Web Tier : User visible browser content
- Application Tier : Logic and functionality
- Database Tier : Data persistence and retrieval

Dynamic Resource Provisioning

- Dynamic Resource Provisioning
  - Timely addition of resources to satisfy increased user demand
  - Timely removal of resources to ensure cost/utilization efficiency
  - Resource : Abstract representation of a computing entity with a fixed capacity that processes session based workloads (For e.g., a virtual server)
  - Resource Utilization : Percentage of resource capacity that is utilized to serve sessions.

Dynamic Resource Provisioning - Challenges

- Two critical challenges for dynamic resource provisioning in multi-tier service
  - Understand the service dynamic behavior when subjected to dynamic workloads.
  - Adaptive management of the service resources to achieve performance guarantees
- Techniques exist for request based performance metrics in single tier and multi tier services, but none for session based metrics.

Multi-Tier Architecture Challenges

- Inter-tier dependencies
  - A resource's ability to serve a request depends on not only its but also on its neighbouring tier resource capabilities.
- Each tier has different type of servers with distinct characteristics.
- Different constraints at the tiers : Replication, Caching.
- Dynamic Internet workloads
- Dynamic bottleneck tier shift challenge
  - Bottleneck tier : A tier with utilization exceeding predefined threshold.

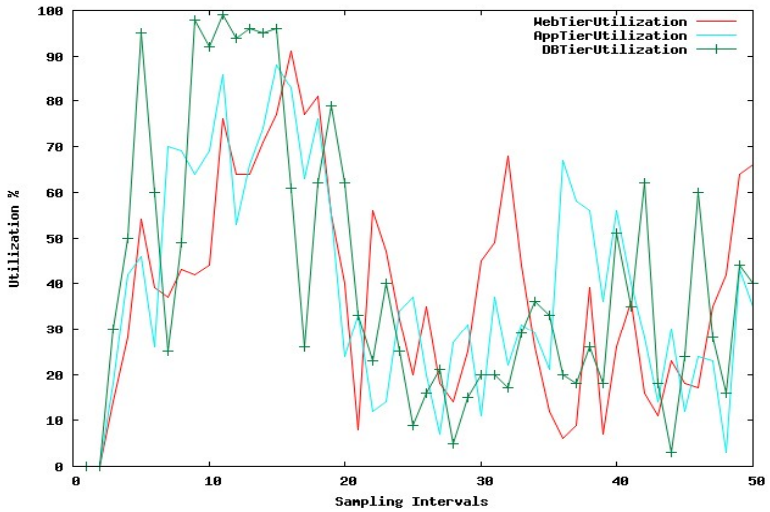
Multi-tier Architecture Challenges

- Extending single tier strategies to multiple tiers is non-trivial.
- Related Work
  - B.Urgaonkar, P.Shenoy, A.Chandra, P.Goyal Agile dynamic provisioning of multi-tier {Internet} applications. ACM Trans. on Autonomous and Adaptive Systems 3(1):1--39, 2008.
    - Independent per-tier provisioning of additional resources - Ineffective
    - Treating the multi-tier system as a black box - Ineffective
    - Queuing based analytical model of the multi-tier system
      - Fails to capture the session based workload dynamics
      - Does not reflect the dynamic bottleneck tier shift

Demonstrating Dynamic Bottleneck Tier shift

- TPC-W : Three different traffic mixes with varying navigational behavior patterns
  - Browsing, Shopping and Ordering
  - Varying Resource Demands
- Related Work
  - J.Rao and C.Xu. Online measurement of the capacity of multi-tier websites using hardware performance counters. Proc. IEEE Int'l Conf. on Distributed Computing Systems, 2008.
    - Browsing : Database tier intensive
    - Shopping : Application tier intensive
    - Ordering : Web tier intensive

Demonstrating Dynamic Bottleneck Tier Shift



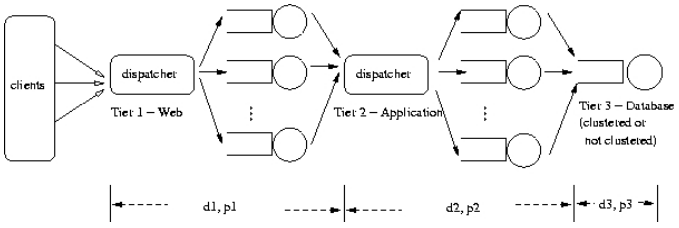
Contributions

- A new session oriented performance metric for multi-tier Internet services.
- Statistical regression models that capture the dynamic behavior of the multi-tier Internet service when subjected to dynamic workloads.
- A dynamic resource provisioning strategy that combines offline training and online monitoring to achieve session based performance guarantees.
  - Offline training learns the regression models
  - Online monitoring utilizes the learned regression models to control the upper and lower bounds of resources allocated to the multi-tier Internet service.

Session based performance metric

- Request based metrics measure system responsiveness and service quality
  - Absolute metrics
    - Response time and queuing delay
  - Relative metric
    - Slowdown : Relative ratio of request's queuing delay to its service time.
    - Measure user perceived relative performance
- Session based performance metric
  - Absolute metrics are not practical – dynamic session length
  - Need a metric that is independent of session length
  - Session slowdown
    - Ratio of the total queuing delay of the requests to the total processing time of the requests.
    - Measure user perceived relative performance at session level.

Session Slowdown



d_{ij}, p_{ij} - queuing delay and processing time of request j at tier i

m – session length

n tier application

$$s = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}}$$

Tier Session Slowdown Ratio

- Varying resource demands at the individual tiers leads to dynamic bottleneck tier shift challenge.
  - Capture the varying resource demands at the individual tiers
  - Tier session slowdown : Ratio of the total queuing delay of the requests of the session at a tier to the total processing time of the requests of a session at that tier.
    - Effected by the dynamic resource demand on the tier and the resources allocated to the tier.

$$s_i = \frac{\sum_{j=1}^{m} d_{ij}}{\sum_{j=1}^{m} p_{ij}}$$

Tier Session Slowdown Ratio

- The normalized tier session slowdown at a tier i

$$sr_i = \frac{s_i}{s}$$

- Tier session slowdown ratio of a *n*-tier service is the ratio of the normalized tier session slowdowns at the individual tiers

$$sr_1 : sr_2 : \ldots : sr_n$$

- Tier session slowdown : Reflects the resource demand at the tier
- Tier session slowdown ratio : Reflects the proportional resource demands on the individual tier

Related Work : Statistical Learning

- Gaining popularity in understanding the dynamics of large scale multi-tier systems
  - J.Rao and C.Xu. CoSL: A coordinated statistical learning approach to measuring the capacity of multi-tier Websites. Proc. IEEE Int'l Parallel and Distributed Processing Symp (IPDPS), 2008.
    - Bayesian network is used to correlate low level instrumentation data collected at run-time to high level system states of each tier.
    - A decision tree is induced over a set a Bayesian models to dynamically identify the bottleneck

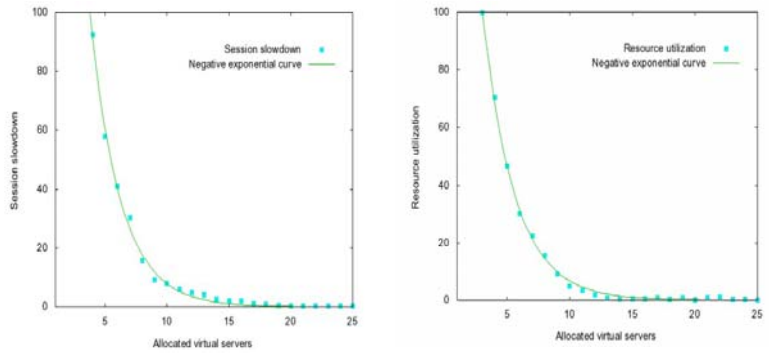
Related Work : Statistical Learning

- S.Muppala and X.Zhou. CoSAC: Coordinated Session-based Admission Control for multi-tier Internet Applications. Proc. IEEE Int'l Conf. on Computer Communications and Networks (ICCCN), 2009.
  - The multi-tier Internet application is modelled as a Bayesian network.
  - Observed performance is applied to the network as evidence.
  - Output is the probability with which an incoming session may be accepted.

Regression Based Dynamic Resource Provisioning

- Combination of offline training and online monitoring
- Offline training
  - Learn and model the multi-tier system 'behavior' dynamics when subjected to dynamic workloads
  - Two main regression relations : 'resources allocated – session slowdown' and 'resources allocated – resource utilization'
  - Behavior model : Capture the workload characteristics used for the training instance and the resulting regression models.
  - Training repeated with diverse workloads resulting in an extensive set of behavior models.
  - The set of behavior models collectively represent the multi-tier service behavior when subjected to dynamic workloads.

Single Training Instance (Relationship Trends)



Exponential Regression Analysis

Negative Exponential Relation : $y = ae^{-bx}$

Linearized form : $\ln y = \ln a - bx \ln e$

After linear regression analysis :

$$\ln a = \frac{\sum \ln y_i + b \sum x_i}{n}$$

$$b = \frac{n \sum x_i \ln y_i - \sum x_i \sum \ln y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

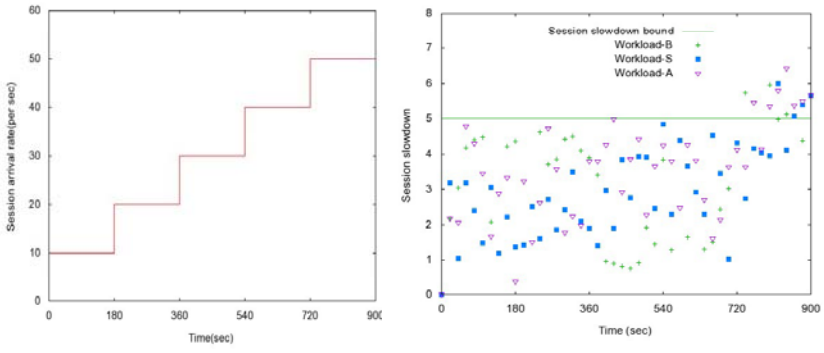
A Sample Behavior Model

Session Arrival Rate	20 sessions/sec
Session type	TPC-W Browsing
“resources - session slowdown" regression model	y=65.4833e^{-0.067x}
resources- session slowdown" correlation coefficient	0.9938
resources- resource utilization" regression model	y=76.2381e^{-0.0989x}
resources-resource utilization" correlation coefficient	0.9458
Tier session slowdown ratio	0.25 : 0.16 : 0.59

Online Monitoring

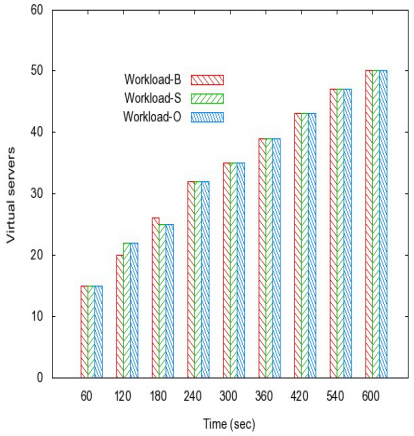
- At periodic intervals, session slowdown and resource utilization metrics are monitored and compared with predefined thresholds.
- On threshold violations :
  - A behavior model with the closest matching workload characteristics observed in the interval is selected.
  - Session slowdown threshold violation
    - Additional resources to be allocated are predicted using the 'resources allocated – session slowdown' regression model.
    - Resources are distributed to the tiers in proportion to the measured tier session slowdown ratio.
  - Resource utilization threshold violation
    - Resources to be removed are predicted using the 'resources allocated – resource utilization' regression model.
    - Resources are distributed to the tiers in inverse proportion to the measured tier session slowdown ratio.

Performance Evaluation (Session slowdown guarantees 1)

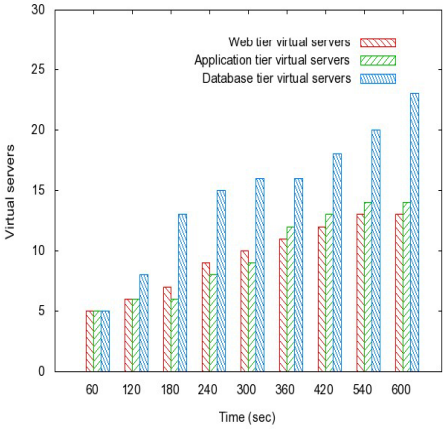


Session slowdown guarantees

Performance Evaluation (Session slowdown guarantees 2)



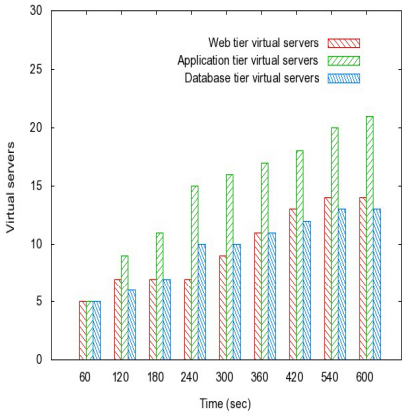
Total resources allocated for 3 workloads



Resources at individual tiers

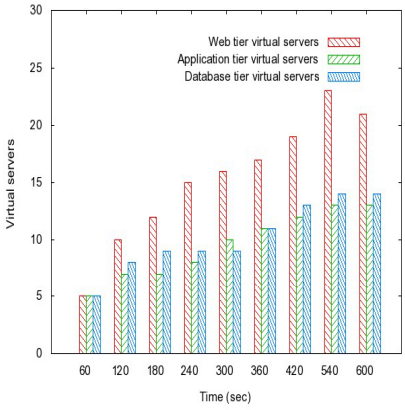
(Browsing workload)

Performance Evaluation (Session slowdown guarantees 3)



Resources at individual tiers

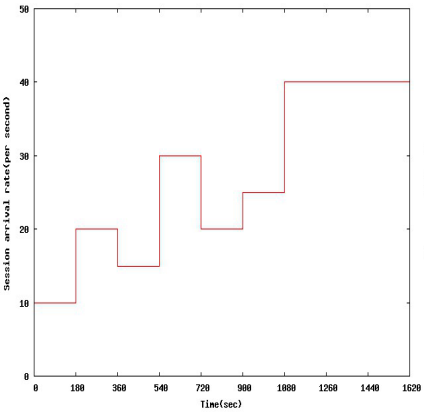
(Shopping workload)

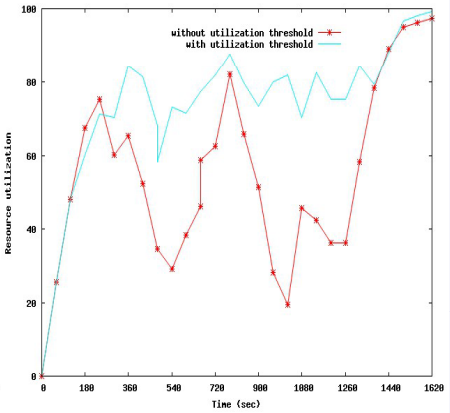


Resources at individual tiers

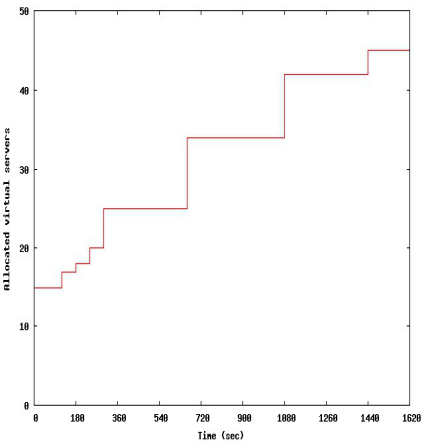
(Ordering workload)

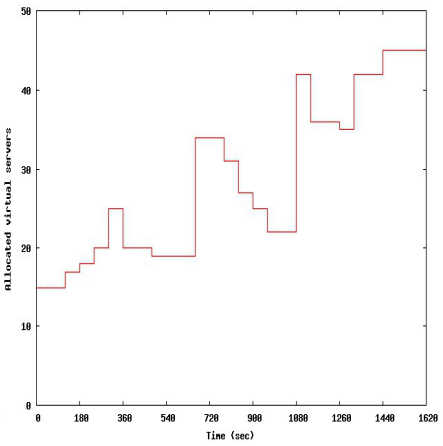
Performance Evaluation (Efficient Resource Utilization 1)





Performance Evaluation (Efficient Resource Utilization 2)





Without utilization threshold **With utilization threshold**

Conclusions

- A relative performance metric, session slowdown, which is independent of session length is appropriate for session based workloads.
- Two statistical regression models capture the dynamic behavior of the multi-tier Internet service with respect to the service parameters.
- The upper and lower bounds of the multi-tier Internet service resources are controlled using two distinct regression models.
- Simulation tests using TPC-W benchmark workload demonstrate the effectiveness of the proposed dynamic resource provisioning strategy in meeting the session slowdown guarantees while ensuring resource utilization.

Q & A

THANKS