Is Your Database Stressed? Load Testing for Peak Performance

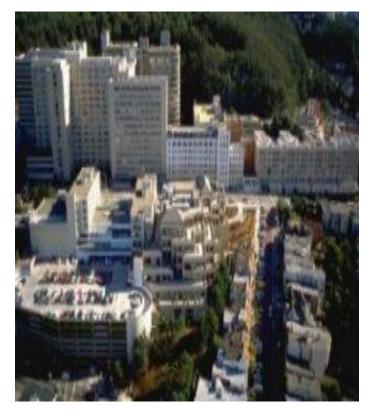
Noelle A. Stimely
Senior Performance Test Engineer
University of California, San Francisco
noelle.stimely@ucsf.edu

Who am I?

- Senior Oracle Database Administrator for over 13 years
- Senior Performance Test Engineer
- Past Oracle Instructor
- Past Lead Database Administrator
- Oracle Certified Professional (OCP) 8, 8i, 9i,10g
- Hewlett Packard AIS LoadRunner v11Certified

University of California, San Francisco

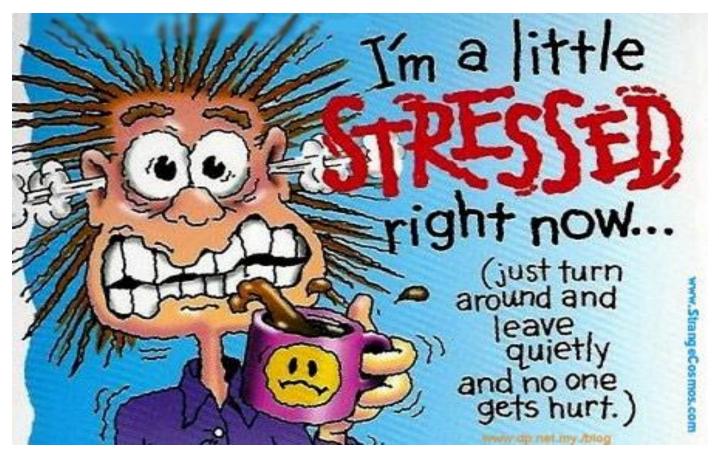
- University of California, San Francisco is a leading university dedicated to defining health worldwide through advanced biomedical research, graduate level education in the life sciences and health professions, and excellence in patient care
 - 3000+ Students
 - 20,000+ faculty, staff and post docs
 - Annual budget about \$2.4 billion
 - San Francisco's second largest employer





This is not going to be a presentation on database tuning, but on load testing.

Stressed out due to performance issues?



OVERVIEW

- What is load testing?
- Types of load tests
- Purpose of load testing
- Developing the load test
- Load test design
- Load testing tools
- The test itself
- Analyze and report
- Conclusions
- DEMO



What *is* Load Testing?



Types of Load Tests

- Transaction response testing
 - Order entry must complete within 8 seconds
 - AP query must return results within 5 seconds
 - PDF attachment must upload within 5 seconds

End user experience testing

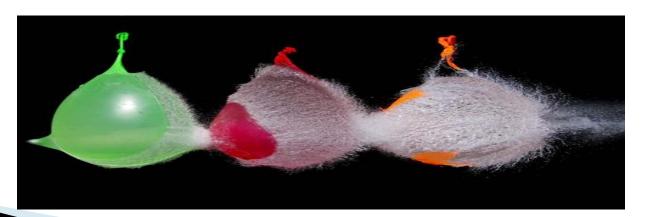
- Run full-scale web load test while a subset of users logs into system to conduct normal work activity
- Have a subset of end users log into the system in the middle of load test to gauge performance

Stress test

- Multiple users logging into the system at one time (100 users log in at one time)
- Users logging into the system very rapidly (e.g. 1 user every second)
- Extended concurrency times (25 users remain logged into system running heavy transactions for an extended period)

Why load test?

- After a software/hardware upgrade?
- Does the application perform acceptably with a minimum amount of users on the system?
- Does the application respond quickly with additional concurrent users?
- Are there hardware bottlenecks in the system?
- Can the application handle a growing number of users/data?
- What is the maximum amount of load the system can handle?

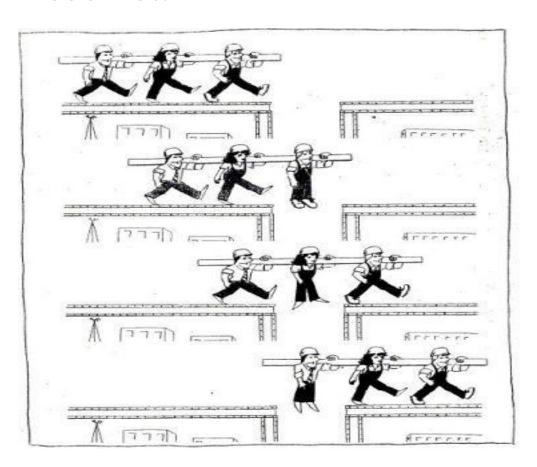


Developing the Load Test



Load Test Team

- Application owner(s)
- Database administrator(s)
- Web server administrator(s)
- Network administrator(s)
- System administrator(s)
- Developer(s)
- Project manager(s)

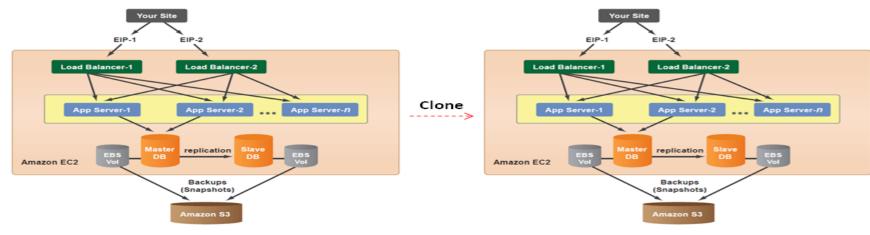


Load Test Criteria Identification



Test environment

- Close to production as possible (ideal)
- How to adjust for inconsistencies between test and production
- Building environment
- Testing the test environment



Production Deployment

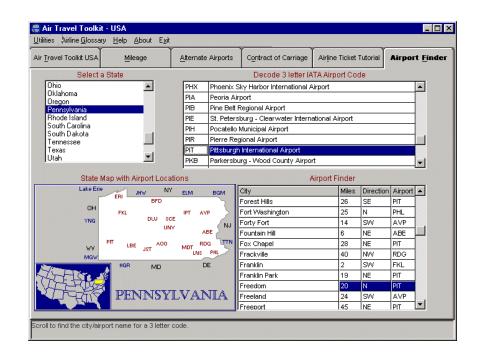
Staging Deployment (Clone of the Production Deployment)

Performance acceptance criteria

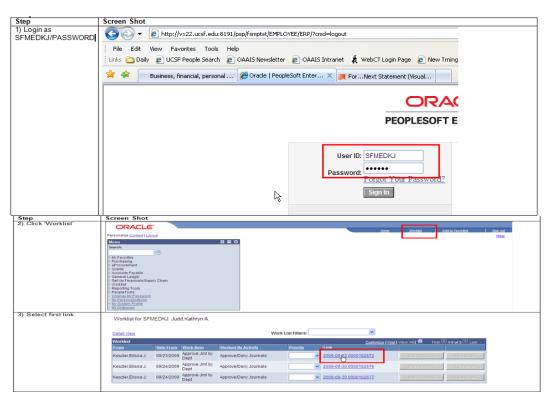
- Service Level Agreements (SLAs)
- End user expectations
- Business requirements
- Workload key scenarios representing reality
- Performance optimization objectives
- Scalability
- Range of accepted workload conditions
- Resource utilization objectives



- Application Scenarios
 - What aspect of application going to test?
 - Breakdown into separate scenarios or tests?
 - Actions to be performed?
 - Example Book a flight:
 - Log into application
 - Search airline flights
 - Book flight
 - Pay for flight with credit card
 - Logout of application
 - Scenario mix
 - Screenshots of scenario



- Application Scenarios (cont'd)
 - Screenshot example:



Load Test Profile

Performance Characteristics

- How many users have been concurrently logged into the system at the busiest time?
- How many concurrent users will be used for the test?
- How will users be divided among the script tasks to be tested, i.e. test 4 actions during test and divide users among scripts - 100 for script 1, 200 for script 2, 50 for script 3 and 150 for script 4.
- Number of users: the total number of concurrent and simultaneous users who access the application in a given time frame
- Rate of requests: the requests received from the concurrent load of users per unit time.
- Patterns of requests: A given load of concurrent users may be performing different tasks using the application. Patterns of requests identify the average load of users, and the rate of requests for a given functionality of an application.
- Will user pause before re-logging in again to perform action, i.e. wait 30 seconds before logging in again?
- How long will users spend on each web page of application during test (think time)
- How long will the test run?

Application Usage Patterns

Scenario

- Search and book flight.
- Search for flights without booking.
- Change an existing reservation.
- Ticket agent running a report.

Script	% of users	Total
Search and book flight	20	200
Search for flights without	60	600
booking		
Change an existing reservation	10	100
Ticket agent running a report	10	100
Total	100	1000

Load Test Profile Considerations

User activity

 Some transactions occur more often than others and therefore should be a larger part of the test data and scenarios.

Ramp-up time

A major component of stress on a site/application is how many and how quickly users log into the system.

Ramp-down time

Load test ramp down implies gradually stopping the threads during a load run in order to detect memory leaks and check system recovery after the application has reached a threshold.

Think Times

The time it takes for a user to respond to an application page has a significant impact on the number of users the application can support without performance issues.

Run Length

• The length of the actual load test can vary based on many factors. You should aim to run the test long enough to get a representative example of true system performance.

Workload/users

 Calculating the number of users for the load test is critical in order to have a valid test which accurately forecasts performance.

Load Profile

Scenario 1 - 20 Users		Load Profile						
Script	Action	# Iteration s	Think Time sec	User	Pacing	# Users Ramp Up	# Users Ramp Down	Duration (mins)
Script0	Search & Book Flight	1	Random of recorded think time: 50% - 150%. Limit 10 seconds.	10	0 Seconds	1 User every 10 seconds	1 User every 10 seconds	30
Script0 2	Search for flights without booking	1	Random of recorded think time: 50% - 150%. Limit 10 seconds.	7	0 Seconds	1 User every 10 seconds	1 User every 10 seconds	30
Script0	Change existing reservati on	1	Random of recorded think time: 50% – 150%. Limit 10 seconds.	3	0 Seconds	1 User every 10 seconds	1 User every 10 seconds	30

Load Test Design

- Application accounts are created for the test with appropriate privileges.
- Test data feeds are properly implemented for testing. Test data feeds are data which is inputted into the system during the load test to create variances in activity thus mimicking real life.
- Validation of scenario transaction on the webserver, database, application server and hosting server.
- Special data used in application which needs to be incremented for proper execution such as product ID.



Test IDs

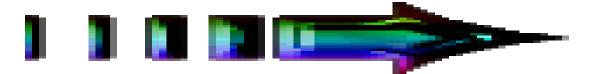
- Create test id(s)
- Application owner role
- Appropriate privileges
- ID data masking
- Unique ids
- Parameterization of values



Login ID	Password
Test040	pass01
Test041	pass01
Test042	pass01
Test043	pass01
Test044	pass01
Test045	pass01
Test046	pass01
Test047	pass01
Test048	pass01
Test049	pass01
Test050	pass01
Test051	pass01
Test052	pass01
Test053	pass01
Test054	pass01
Test055	pass01
Test056	pass01
Test057	pass01
Test058	pass01
Test059	pass01
Test060	pass01
Test061	pass01
Test062	pass01
Test063	pass01
Test064	pass01
Test065	pass01
Test066	pass01
Test067	pass01
Test068	pass01
Test069	pass01
Test070	pass01
Test071	pass01
Test072	pass01
Test073	pass01

Test Data Feeds

- What is test data?
- Where is test data used?
- How to use test data
- Creation of test data
- Parameterization of test data
- Unique/non-unique data
- Oracle Flashback
- Data restore for non-Oracle systems



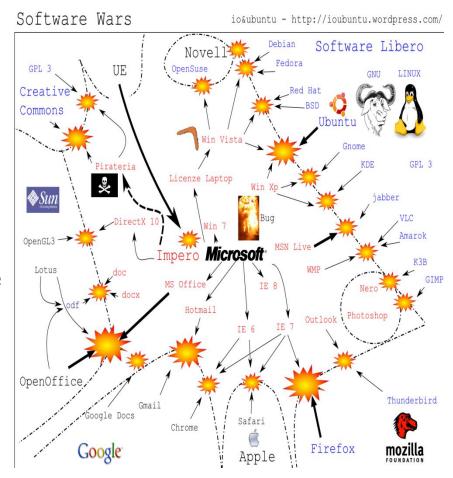
Load Testing Tools - Freeware vs. Commercial

Freeware

- Jmeter
- HammerOra
- Swingbench
- OpenSTA

Commercial

- HP Loadrunner
- Oracle Application Testing Suite
- SilkPerformer



Oracle Real Application Testing vs Load Testing

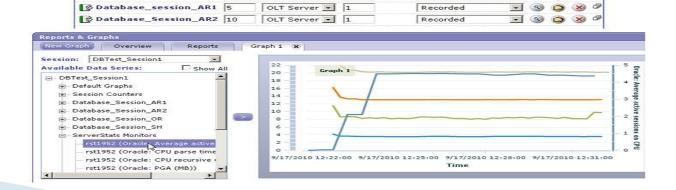
What is Oracle Real Application Testing (RAT)?

Database_Session_OR 15

- Why RAT should not be used by itself for formal load testing.
- How is RAT used?
- Oracle Real Application Testing Suite

How to use Real Application Testing Suite with RAT to create load

tests



OLT Server 💌 1

OLT Server 1

Iteration Delay VU Pacing (Think Time)

Recorded

Recorded

Monitoring the Load Test

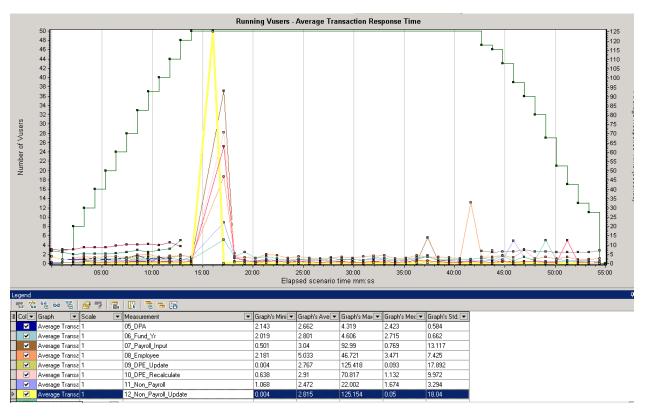
- Identify metrics to be used for the load test
 - OS metrics which could include CPU, memory, disk I/O and network I/O statistics.
 - Database metrics to watch for are waits, top SQL, Buffer gets, etc.
 - Network metrics would include the overall health of your routers, switches and gateways.
 - Webserver metrics would include hits per second, number of requests, etc.
- Identify team members who will monitor during the test
 - What areas will be monitored?
 - What tools will be used?
 - How will the tools be used to monitor (invasive/non-invasive?)
 - How will the results be reported at the conclusion of the test?



Analyze and Report Test Results

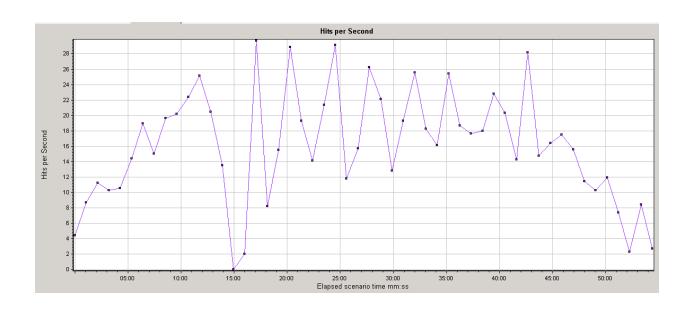
- Pinpoint bottlenecks found during load test
 - Examples:
 - Checkout should not take longer than 4 seconds under 15 hits per second workload
 - · CPU utilization for the test should not exceed 80% during any transaction
- Transactions in average response time graph breach the performance acceptance criteria.
- Response time that is worse than what the SLA for the customer allows.
- Correlate transactions with other graphs and measurements to find bottlenecks.
- Involve developers, system and network administrators in investigation.
- Overlay response time graph with monitoring graphs and look for performance trends.
- Check standard deviation of transactions if transactions behave inconsistently identify the cause.
- Transactions availability: (count of pass transactions/count of total transactions) x 100. Usually applications have around 99.9% availability. Lower values don't necessarily indicate an issue, but should be verified why lower.

Load Test Report Graphs



 LoadRunner graph showing a performance irregularity occurred between 14 and 18 minutes into the test

Hits per Second Graph - LoadRunner

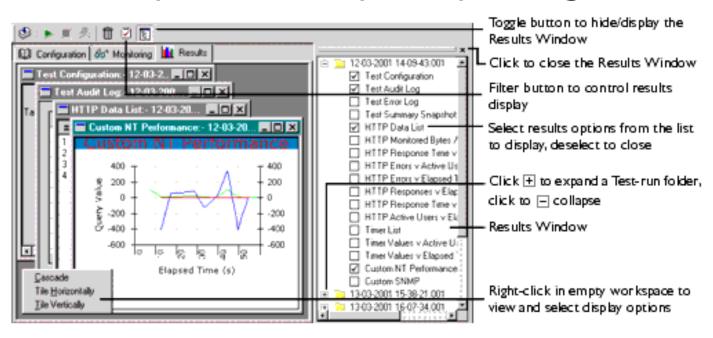


Same load test as previous slide - Note drop off in hits at 15 minute point in test

Reporting Audience

- Management/Project Managers/Team Leads?
 - Technical aptitude
 - Length of report
 - Type of report
- Suggested Graphs to include in report
 - Running Users Shows the impact of ramp-up and ramp-down as well as when errors begin to occur during the load test.
 - Hits per Second Shows the amount of load generated by virtual users in terms of hits to web servers.
 - Throughput Shows the amount of load virtual users create on network resources raw amount of bytes client receives each second as a result of hits on the web server.
 - Average Transaction Response Time Shows the actual load generated by each transaction throughout the entire load test.

OpenSTA Graph Reporting



- Graphs can be customized for presentation purposes
- Data from graphs can be exported to Excel for further analysis

Load Test Summary

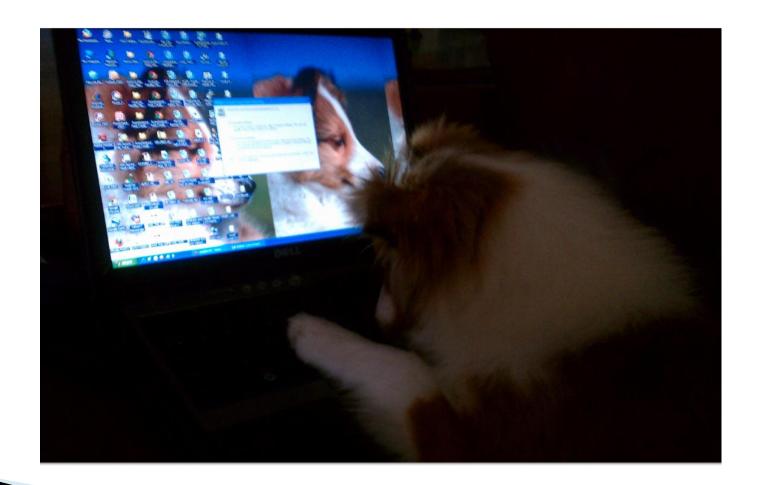
- Load test summary report
 - Top 5 Slowest Pages
 - Slowest pages encountered during load test
 - Top 5 Slowest Transactions
 - Do not meet the performance criteria defined at test inception
 - Top 5 Slowest SQL Statements
 - Slowest SQL which is typically 80% of performance problems



Find Bottlenecks, Tune System - Retest?



CONCLUSIONS



DEMONSTRATION