# Data-Based Experimental Computer Systems Research

***Calton Pu***

and many collaborators in
academia and industry

# Scientific Method

- Theory produces testable hypotheses
- Experimentation confirms or disproves these hypotheses through measurements
  - *In vitro*: observations in controlled environments (laboratory)
  - *In vivo*: observations in live organisms (lab or real world)
  - *In silico*: observations through computer simulation

# Experimental CS Research

- Experimental research on computer systems through observations
  - *In vitro (in silico)* automated system management research based on measurements of realistic benchmarks
  - *In vivo*: spam research (email, web, social networks, ...) based on data collected from real world

# Comparison App & DB Servers

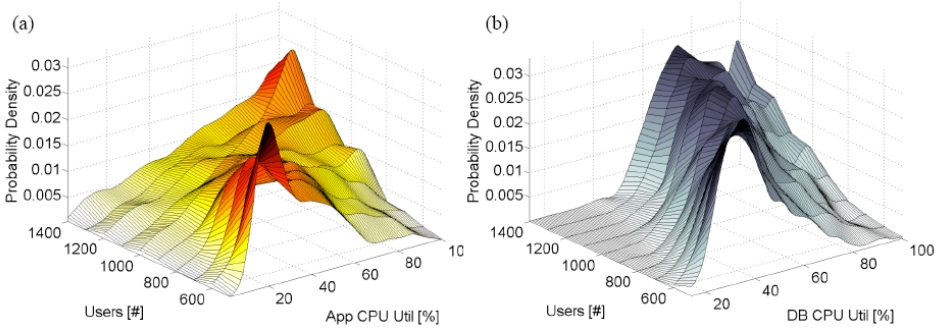


Figure 4: RUBiS on 1H/6H/1L configuration. Estimated probability density functions for CPU utilization in (a) application server for 70 % write ratio and (b) database server for 10 % write ratio.

# Non-Stationary Workload

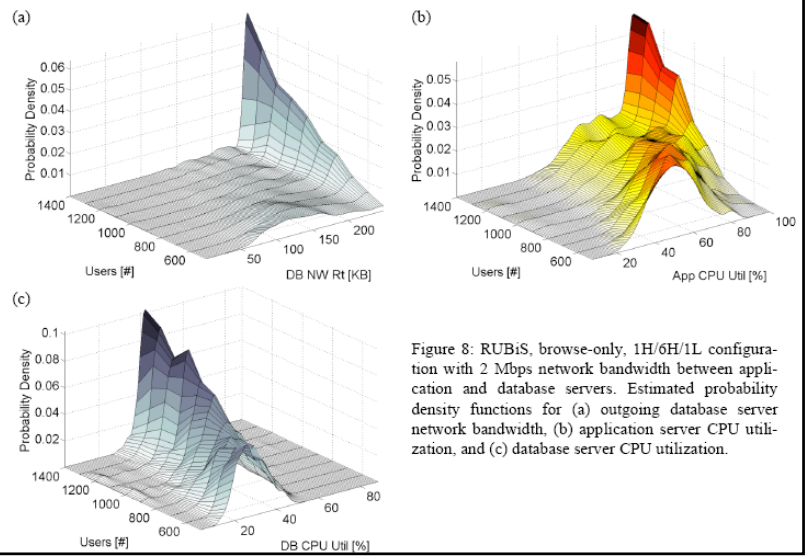


Figure 8: RUBiS, browse-only, 1H/6H/1L configuration with 2 Mbps network bandwidth between application and database servers. Estimated probability density functions for (a) outgoing database server network bandwidth, (b) application server CPU utilization, and (c) database server CPU utilization.

# Complexity of Experiments

| | **Experiment Scale (W-A-D)** | | | | |
|---|---|---|---|---|---|
| | 1-2-1 | 1-6-1 | 1-6-2 | 1-8-1 | 1-8-2 |
| Script lines (per exper. set) | 758 | 1298 | 1426 | 1564 | 1692 |
| Configuration lines (per experiment set) | 1168 | 1212 | 1212 | 1234 | 1234 |
| Number of experiment sets | 408 | 76 | 72 | 496 | 90 |
| Total script lines in all sets | 309K | 98K | 102K | 775K | 152K |
| Total configuration lines | 476K | 92K | 87K | 612K | 111K |
| Total nodes used | 3264 | 912 | 936 | 6944 | 1350 |
| Total data points (million) | 323M | 100M | 104M | 785M | 154M |

# Data-Based Spam Research

| | | |
|---|---|---|
| Email Spam | SpamArchive<br>Enron dataset | 1.4M messages<br>0.5M messages |
| Phishing | APWG archive | 382K messages |
| Web spam | Webb Corpus | 300K web docs |
| Social networks | MySpace crawl | 1M profiles |
| Other media | Being collected | |