

QoS Measurement and Control in Web and E-commerce Services

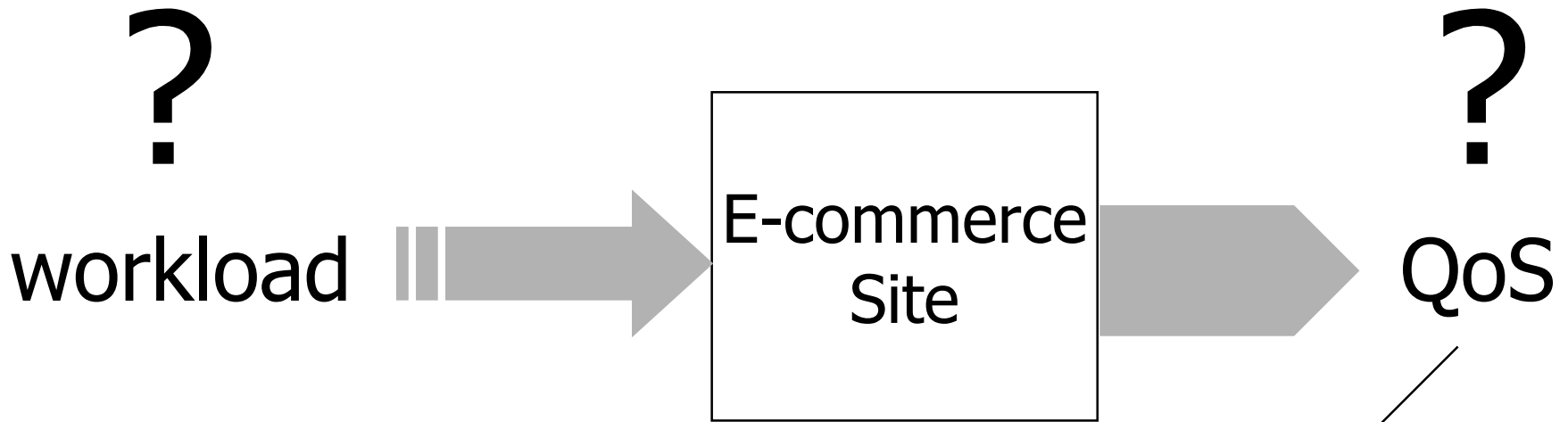
Daniel A. Menascé, Ph.D.

Dept. of Computer Science

George Mason University

Menasce@cs.gmu.edu

These slides are copyrighted by Daniel A. Menascé and cannot be copied, in part or in its entirety, into any medium, including electronic media, nor displayed on any Web site without the written authorization of the authors. The author hereby authorizes the attendees of the February 8, 2002 meeting of New York's Computer Measurement Group to download a copy of these slides and make a single printed copy.



response time
throughput
probability of rejection
availability

A blue-bordered rectangular box is positioned below the "QoS" label. An arrow points from the "QoS" label to the top-left corner of this box. Inside the box, four terms are listed vertically, each on a new line and rendered in an italicized black font: "response time", "throughput", "probability of rejection", and "availability".

QoS for E-commerce

❑ Conventional QoS metrics:

- response time, availability, and throughput

❑ Typical Web site metrics:

- Hits/sec
- Page views/sec
- Unique visitors/day

❑ New metrics for e-commerce:

- revenue/sec
- potential lost revenue/sec

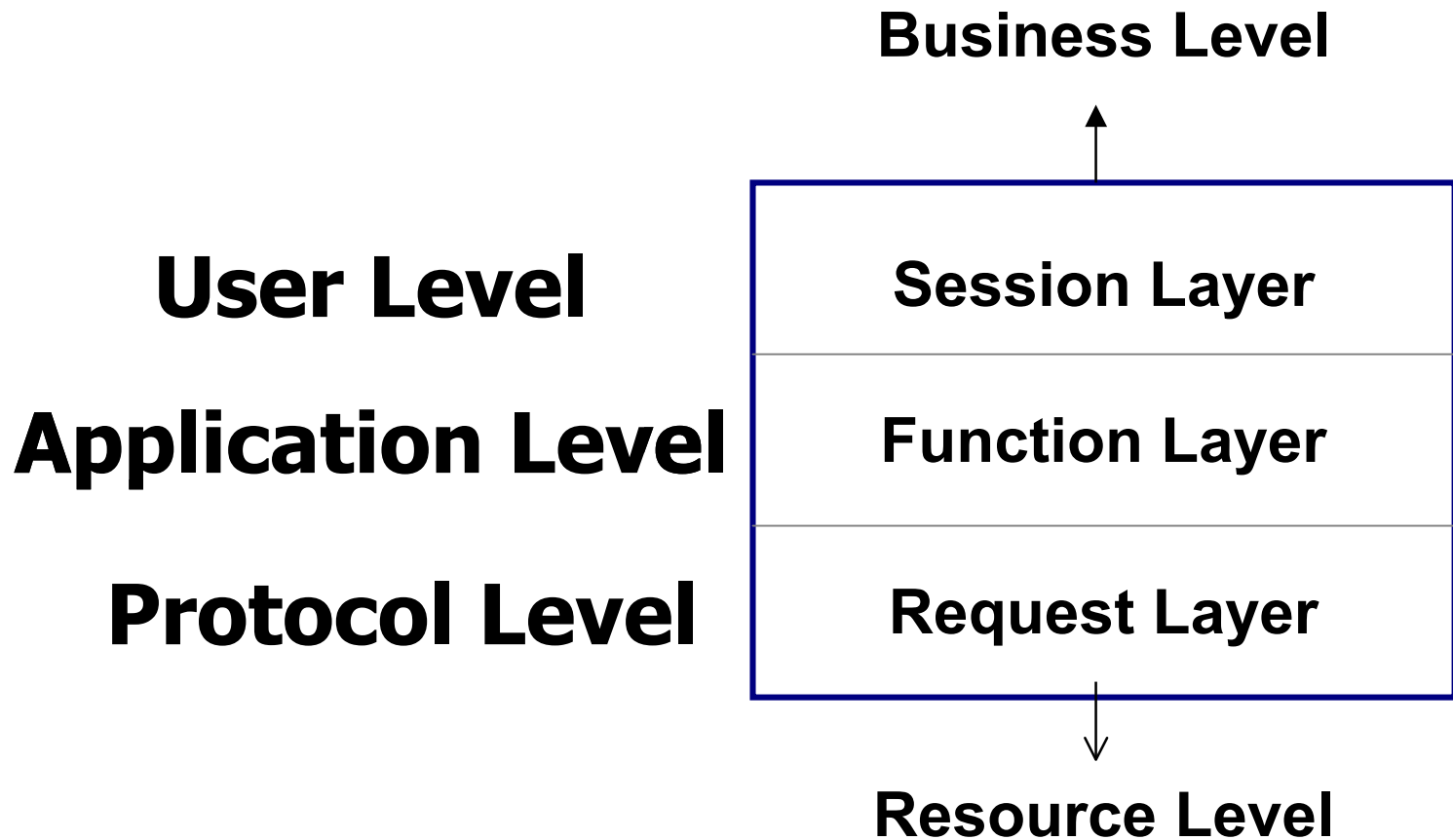
Performance Problems in E-commerce tend to get worse!

- ❑ Proliferation of mobile devices
- ❑ Easier to use interfaces (VUI, novel browsing paradigms)
- ❑ Increasing load placed by software agents
- ❑ Impacts of authentication protocols (e.g., TLS) on e-commerce site performance.

The Importance of Understanding E-Business Workloads

- ❑ Past studies have concentrated on information provider sites.
- ❑ Main studies reflect the Internet of several years ago. Some characteristics have changed:
 - clients have larger bandwidth
 - number of users has grown significantly
 - e-commerce is a major WWW application
 - proliferation of dynamic pages
 - authentication protocols

Workload Characterization Approach



Data Collection: access logs from two e-business sites

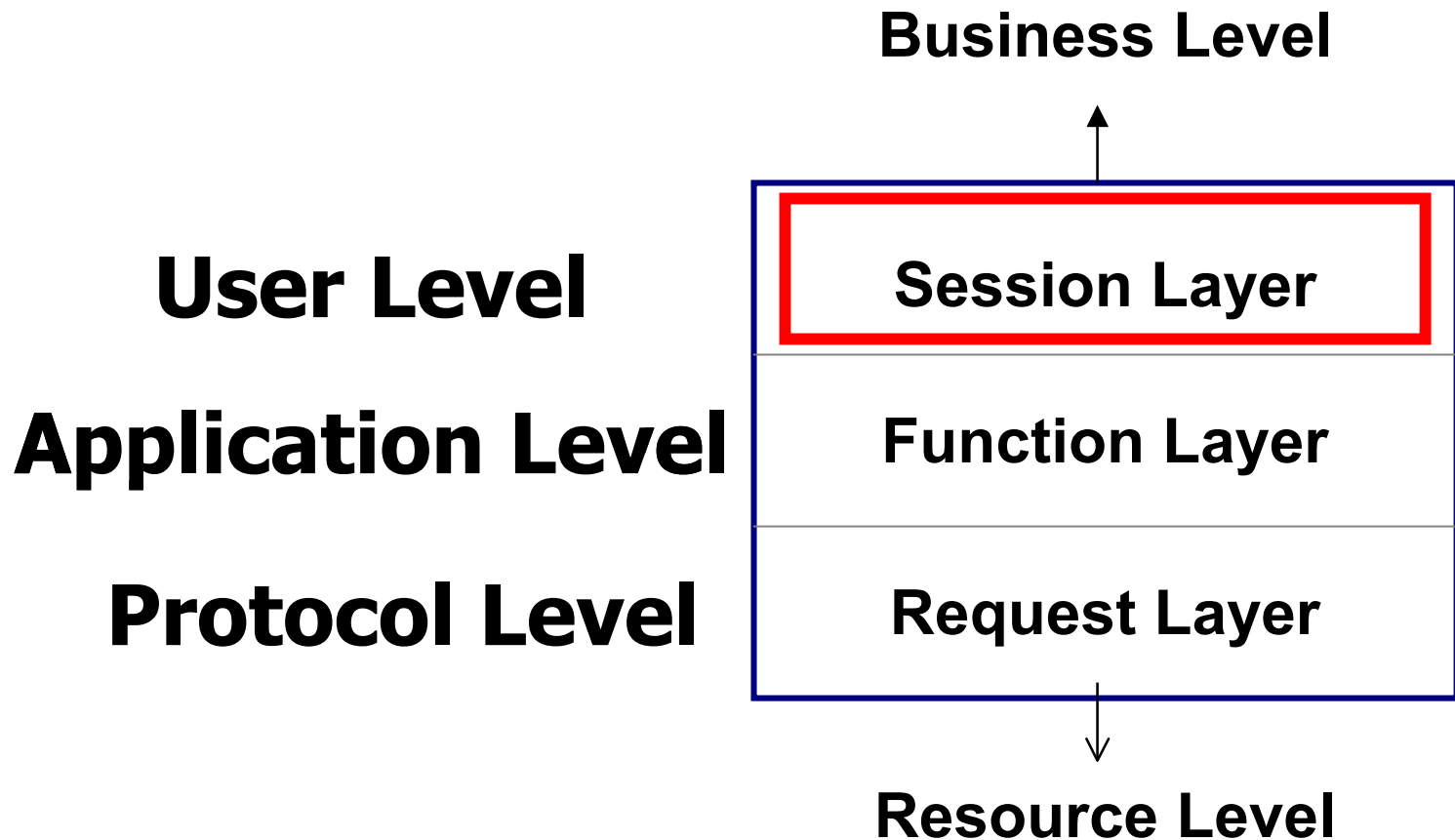
□ E-tailer: bookstore

- Logs from two weeks in August 99
- 3,630,964 requests
- images \cong 71% of the requests

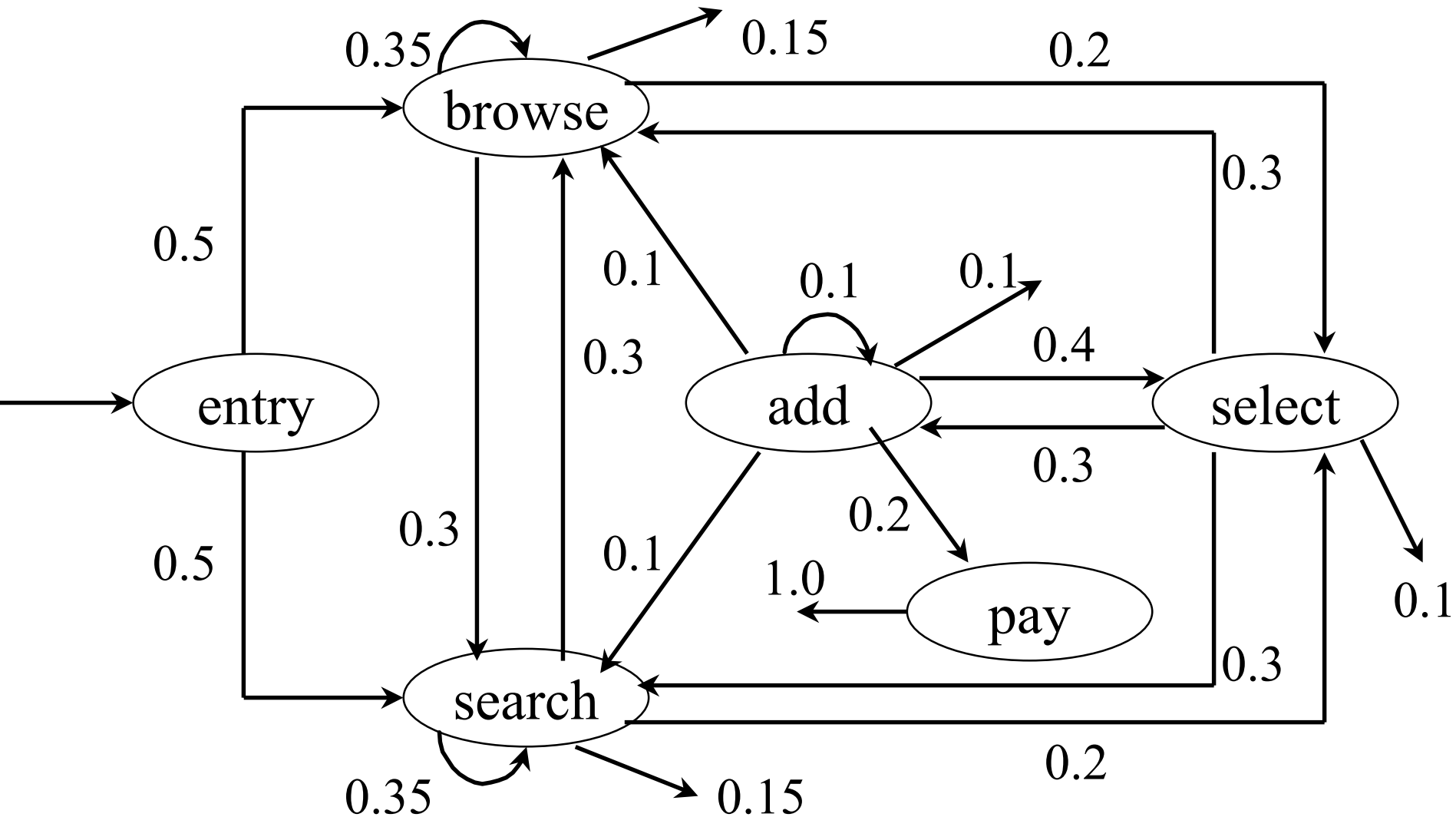
□ Online auction

- Logs from two weeks in March 00
- 466,058 requests
- images \cong 85% of the requests

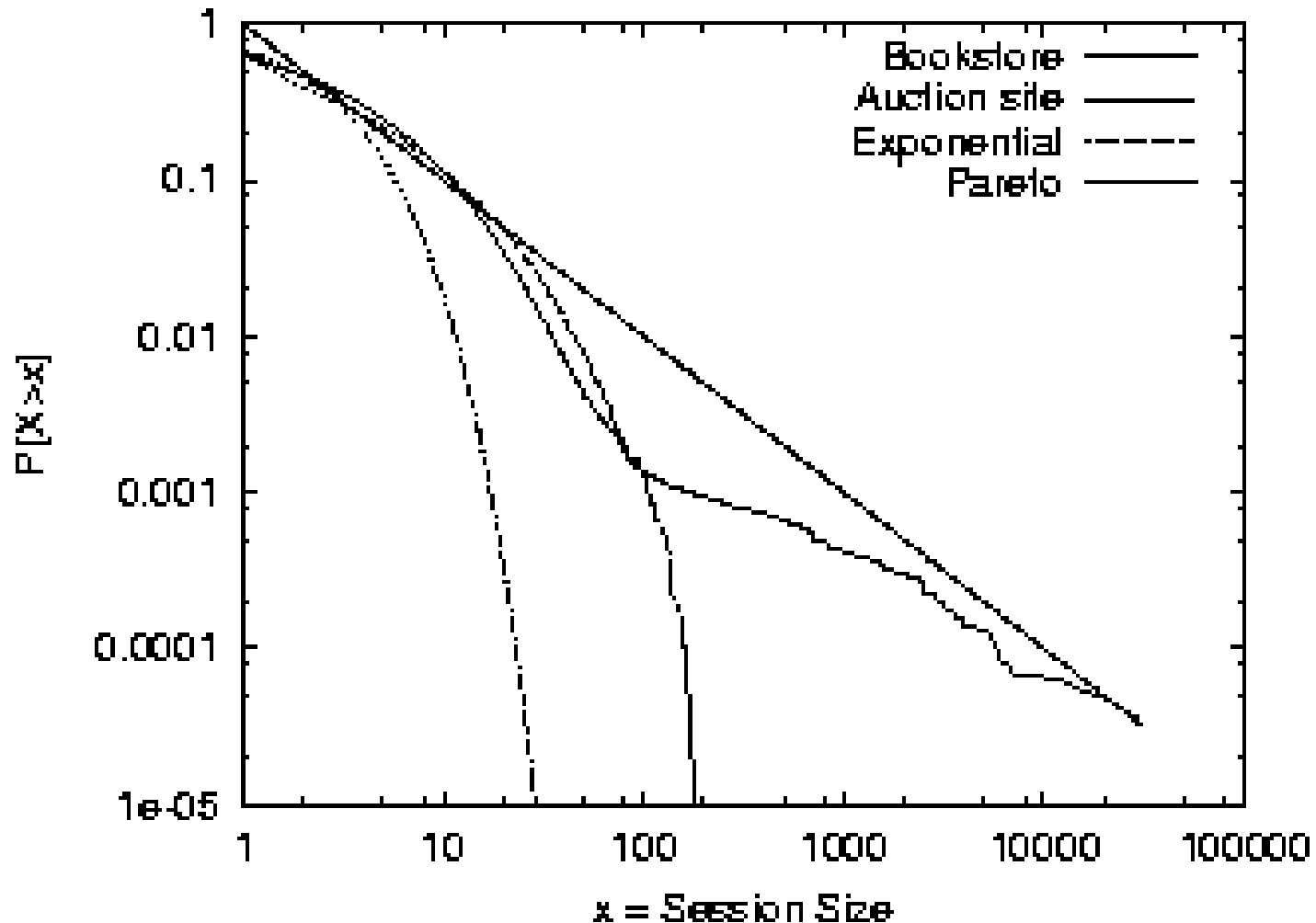
Workload Characterization Approach



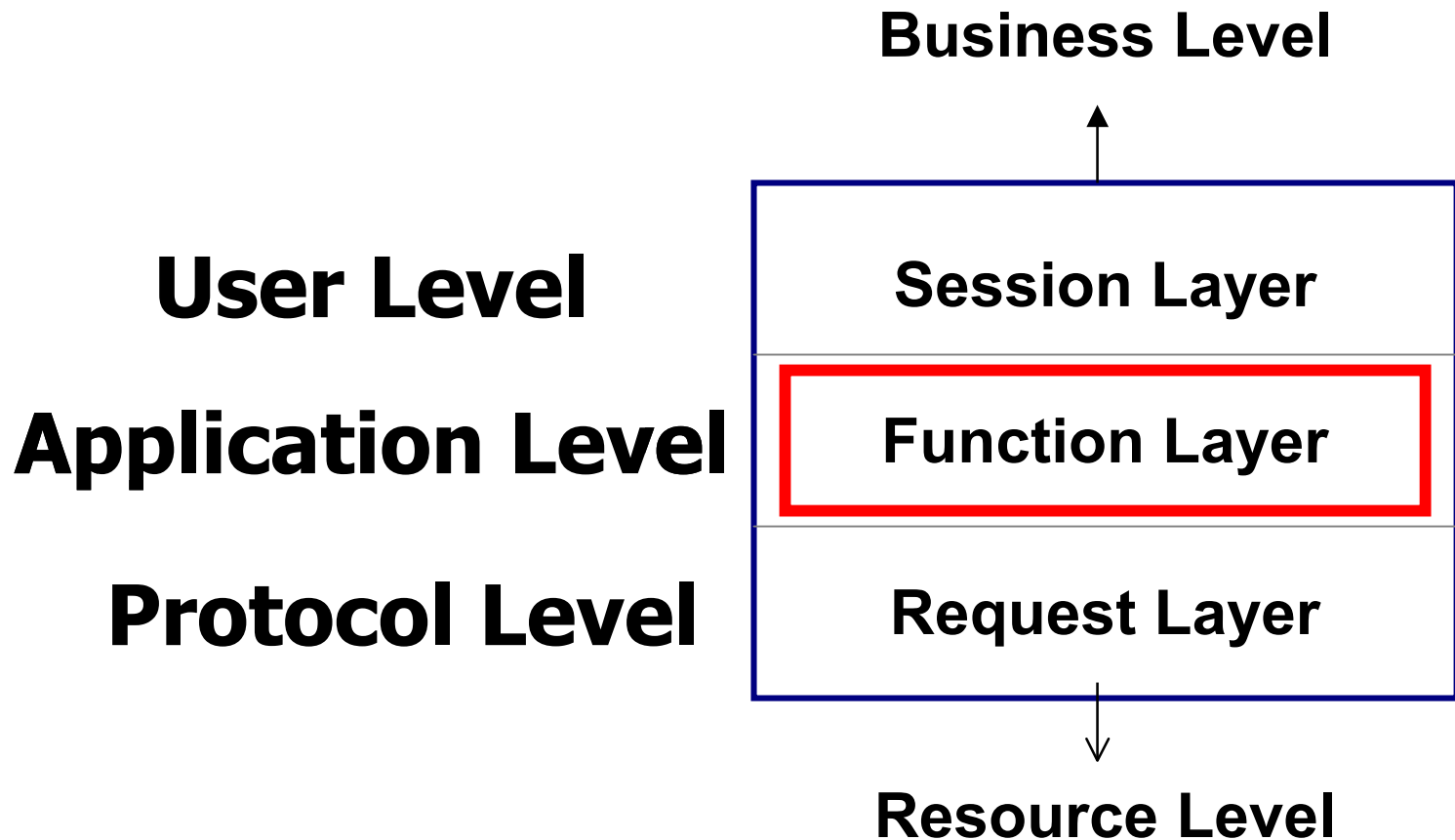
Customer Behavior Model Graph (CBMG)



Session Length (in requests to execute e-business functions)



Workload Characterization Approach



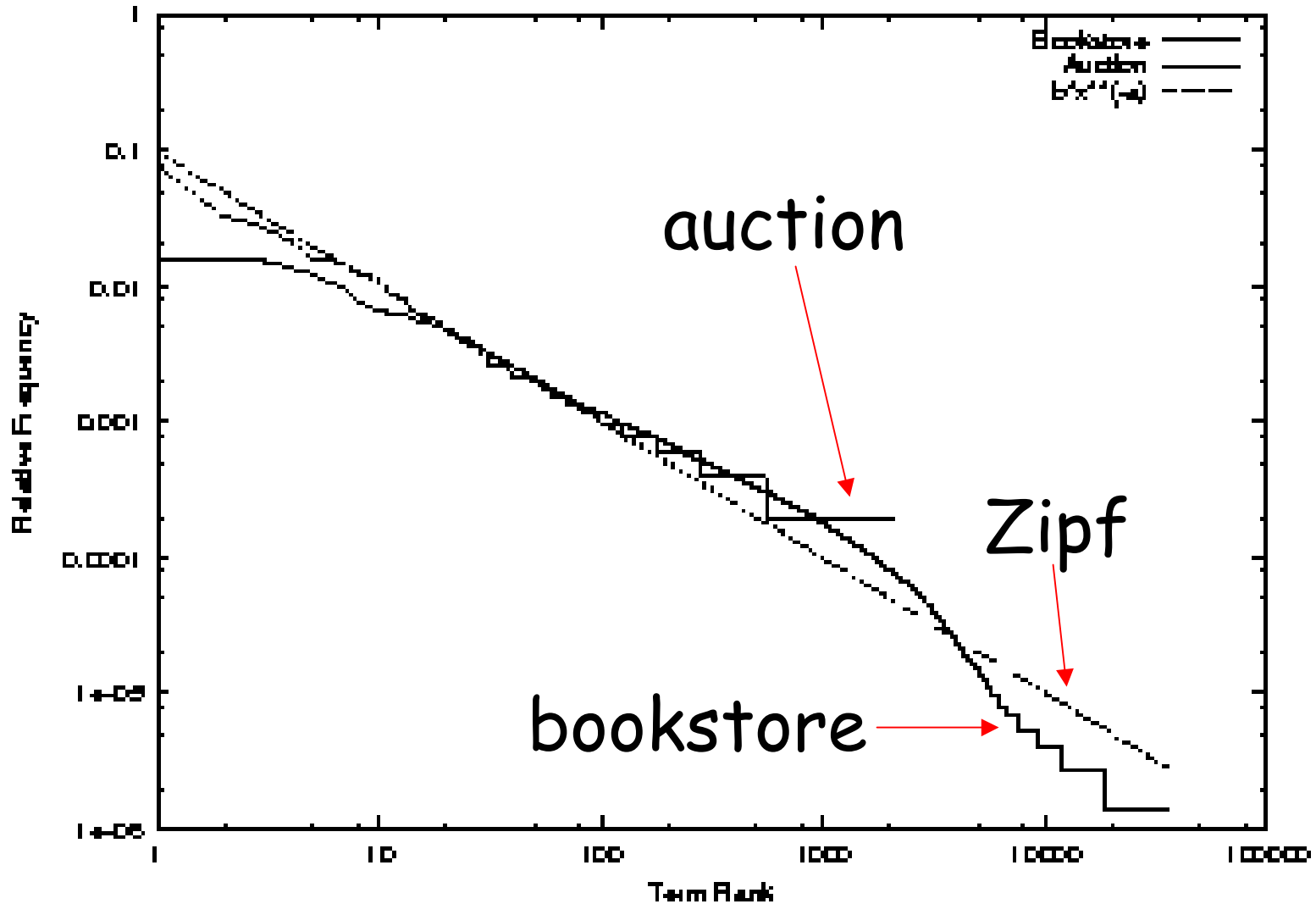
Function Characterization

- ❑ Functions: search, view, add to cart, pay, register, etc.
- ❑ About 70% of the functions are related to product/service selection
- ❑ Buying functions exhibit a very low frequency.
- ❑ Popularity of search terms

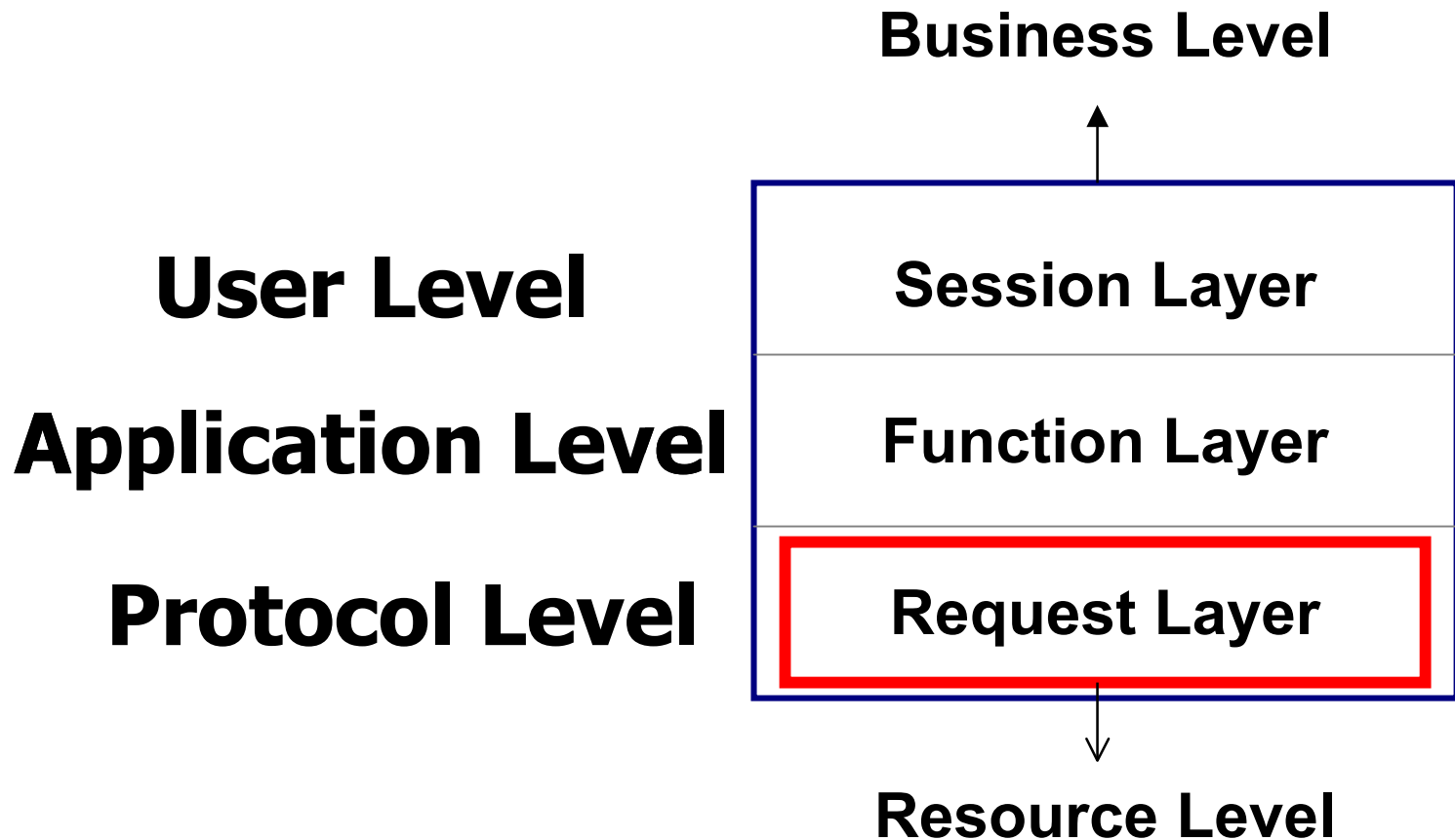
Function Characterization

| Bookstore | | Auction | |
|-----------|----------------|----------|----------------|
| Function | Frequency % | Function | Frequency % |
| Home | 11.92 | Home | 20.70 |
| Browse | 17.72 | Browse | 14.66 |
| Search | 36.30 | Search | 16.74 |
| View | 19.99 | View | 4.87 |
| Add | 5.44 | Bid | 0.08 |
| Pay | 1.19 | Sell | 7.99 |
| Account | 2.44 | Account | 5.99 |
| Robot | 0.04 | Robot | 0.06 |
| Info | 3.66 | Info | 9.44 |
| Other | 1.31 | Other | 19.89 |

Popularity of Search Terms: Zipf's Law!

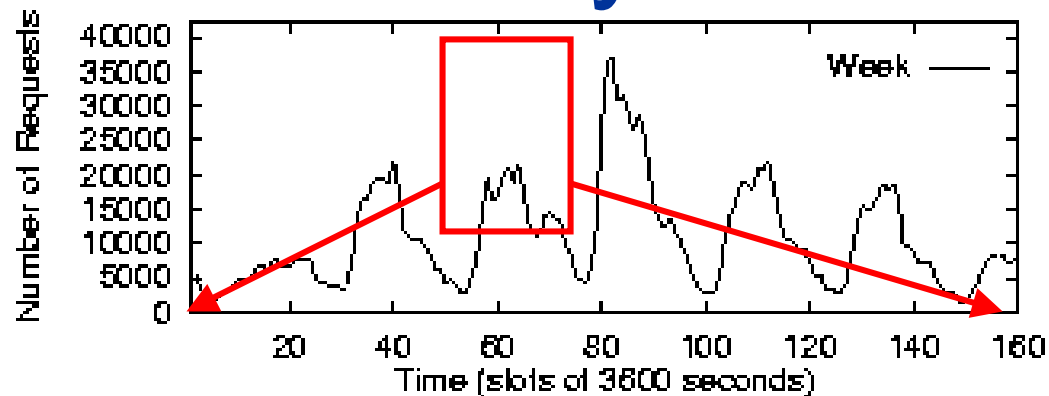


Workload Characterization Approach

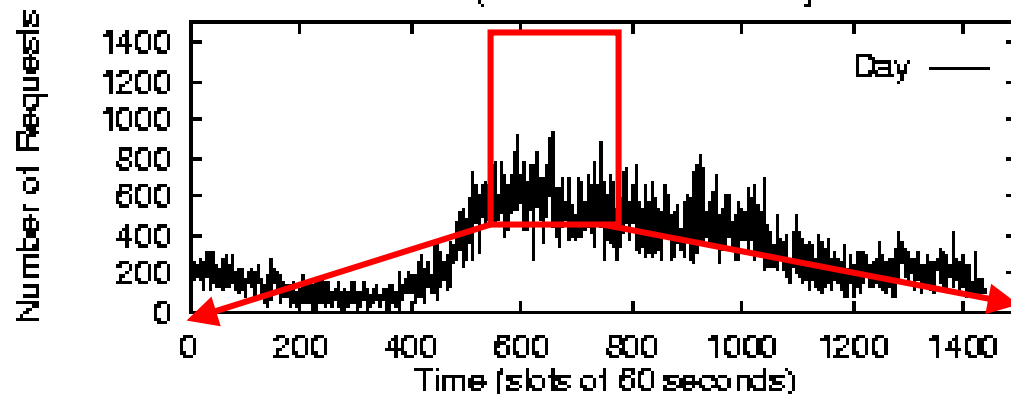


Request Characterization: Multi-scale time analysis

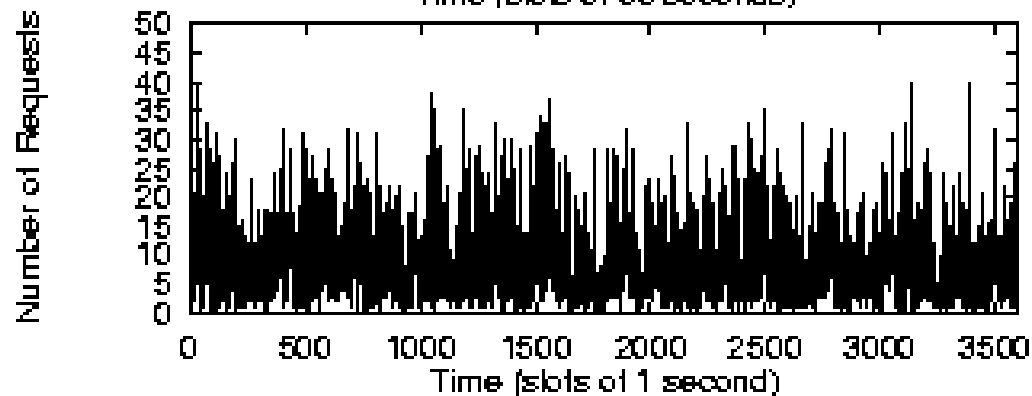
3600 sec



60 sec



1 sec



Agent Characterization: multi-layer criteria

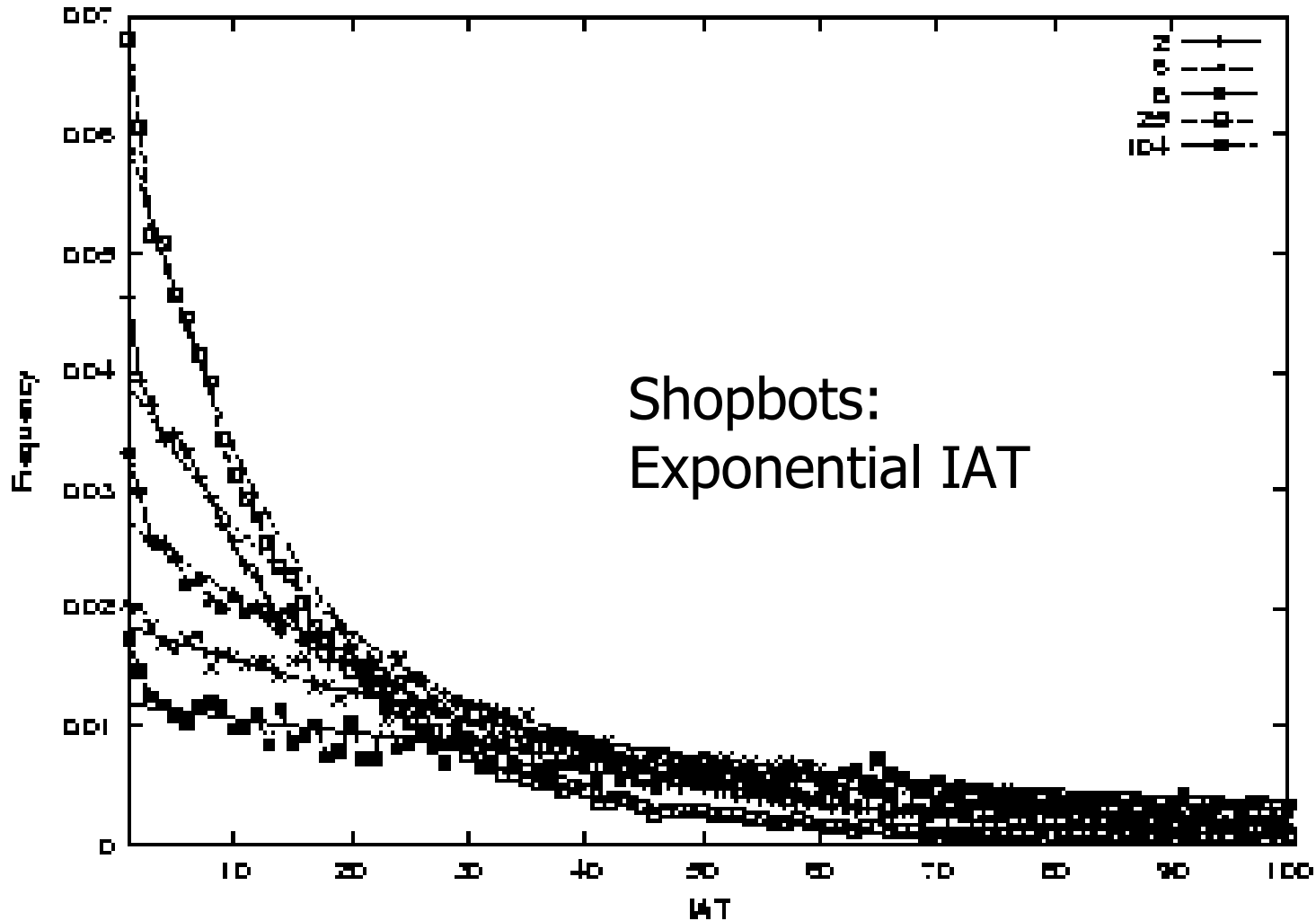
- ❑ Session layer
 - type of functions in a session
 - session length
- ❑ Function layer
 - entry point
 - unlikely functions
 - embedded files
- ❑ Request layer
 - inter-arrival time
 - self-identification



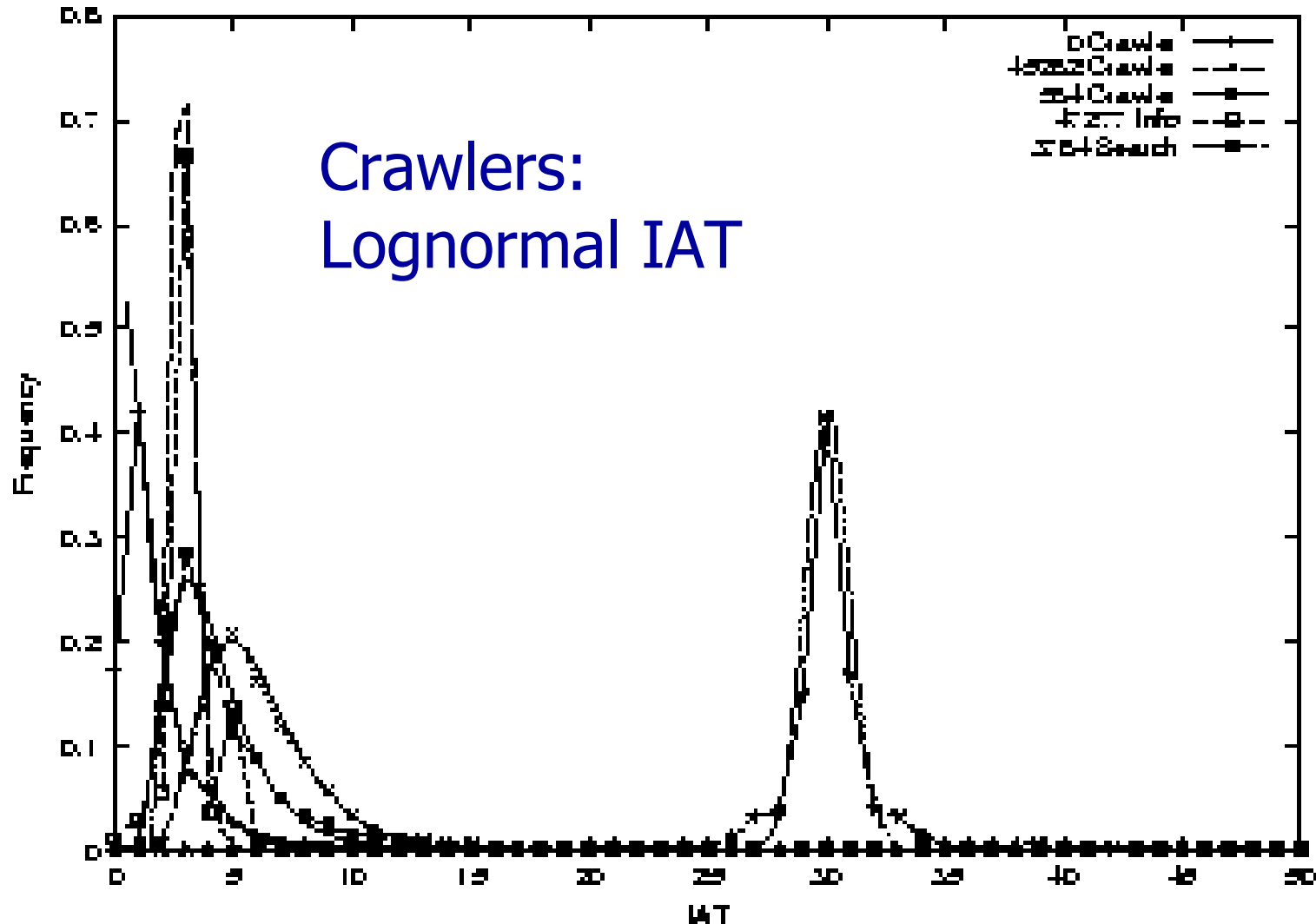
Characteristics of the Log Files for Robot Analysis

| | Bookstore | Berkeley | World Cup |
|----------------------------------|---------------|----------------|-----------|
| Interval | 1-15 Aug 1999 | 1-30 June 2000 | 23-May-98 |
| No. Requests | 3,630,964 | 3,643,208 | 2,225,475 |
| % Images | 74 | 44 | 84 |
| No. Functions | 955,818 | 2,038,249 | 340,719 |
| % Robot Functions | 33.5 | 17 | 6.5 |
| No. Sessions | 130,314 | 371,242 | 33,995 |
| Avg. Robot Session Length | 2,410 | 1,325 | 1,398 |


IAT for Human Generated Robots



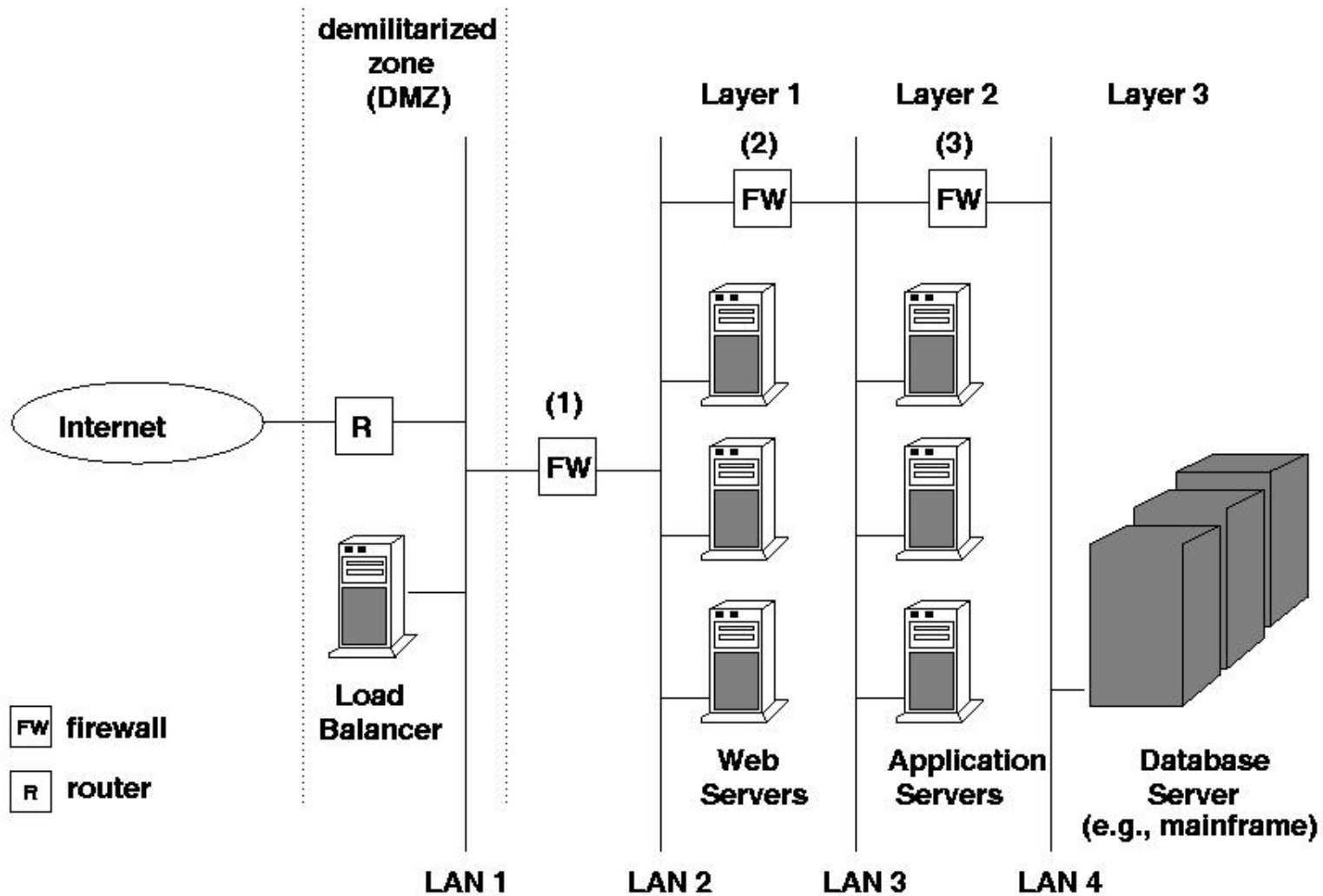
IAT for Automatically Generated Robots



Dynamic QoS Control: Motivation

- ❑ E-commerce sites are complex and composed of multiple tiers.
 - ❑ The workload presents short-term variations with high peak-to-average ratios.
 - ❑ Many software and hardware parameters influence the performance of e-commerce sites.
-  Manual reconfiguration is not an option!

Multi-tier Architecture



Dynamic QoS Control for E-commerce

- ❑ Definition of a combined QoS metric.

QoS Deviation

□ Relative difference between the observed QoS value and the QoS goal.

Response time deviation:

$$\Delta QoS_R = \frac{R_{\max} - R_{\text{measured}}}{R_{\max}}$$

Throughput deviation:

$$\Delta QoS_X = \frac{X_{\text{measured}} - X_{\min}}{X_{\min}}$$

Probability of rejection deviation:

$$\Delta QoS_{P_{\text{rej}}} = \frac{P_{\text{rej}}^{\text{MAX}} - P_{\text{rej}}^{\text{measured}}}{P_{\text{rej}}^{\text{MAX}}}$$

A negative deviation means that the QoS level for the metric has not been met.

QoS Metric

- ❑ The QoS metric is defined as a linear combination of QoS deviations.
- ❑ The weights are determined by management based on the relative importance of each metric.

$$QoS_v = (\Delta QoS_R \times W_R) + (\Delta QoS_{Prej} \times W_{Prej}) + (\Delta QoS_X \times W_X)$$

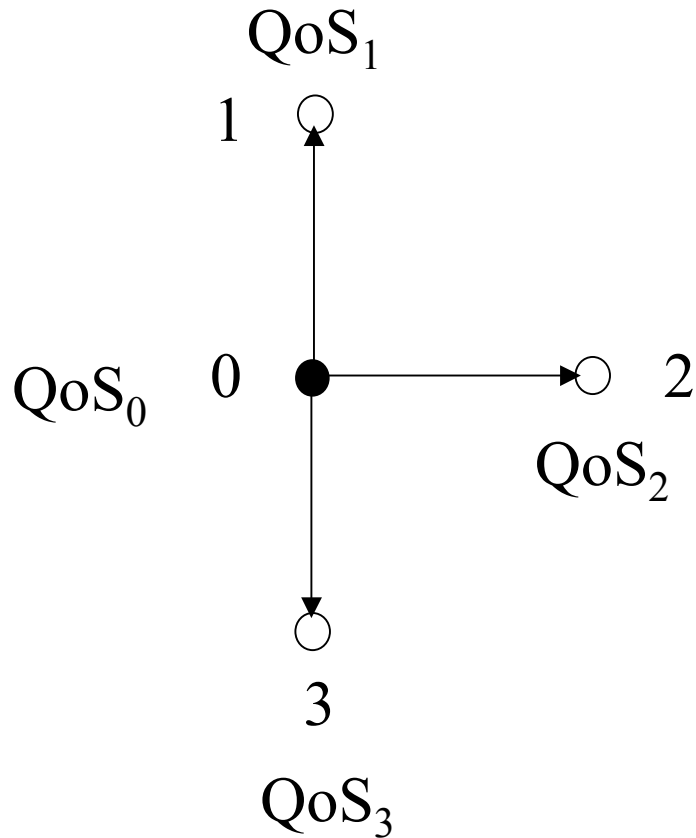
determined by management



Dynamic QoS Control for E-commerce

- ❑ Definition of a combined QoS metric.
- ❑ Use of hill-climbing techniques combined with predictive queuing models.

Heuristic Optimization Approach

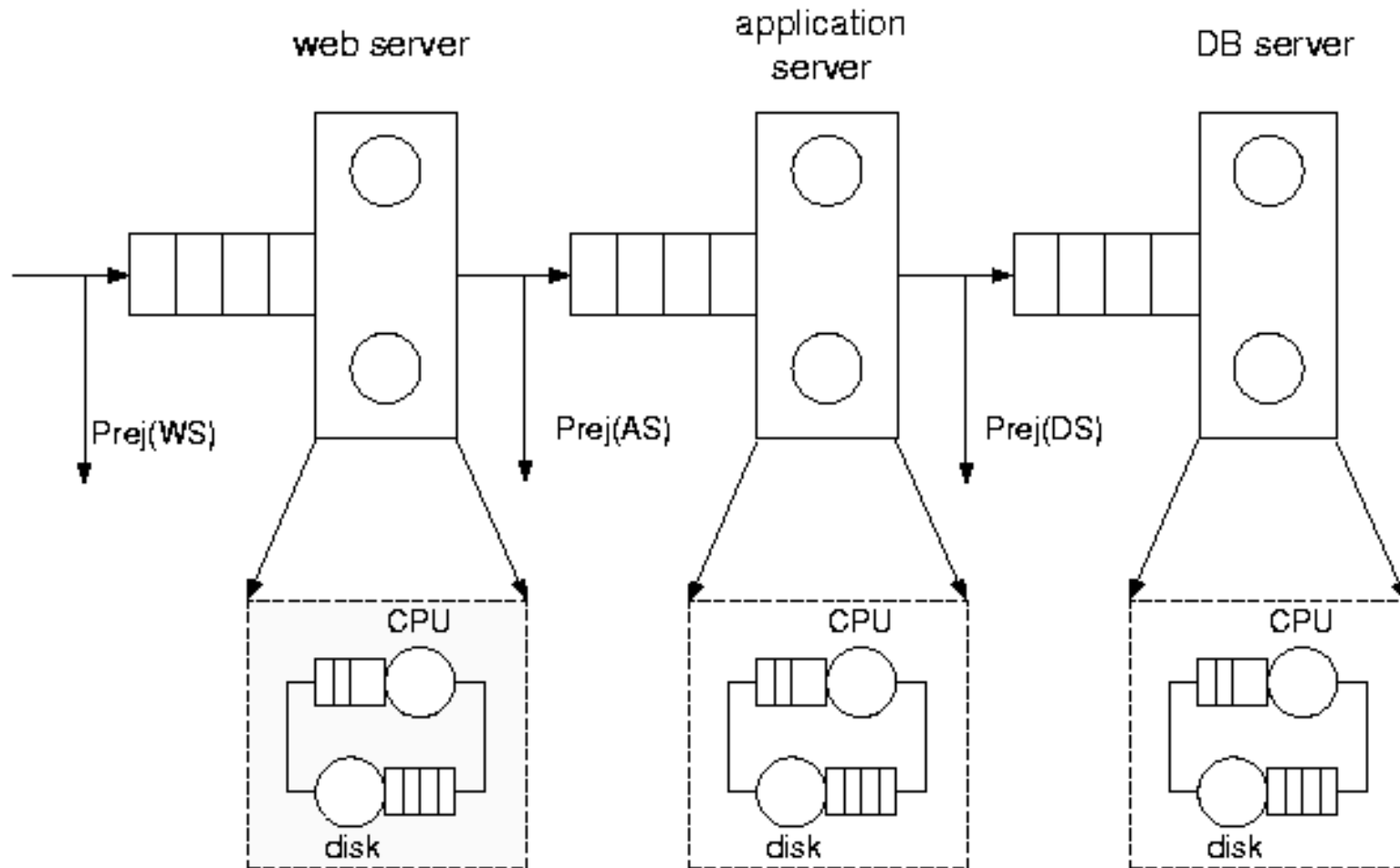


QoS_0 : observed QoS value for current configuration

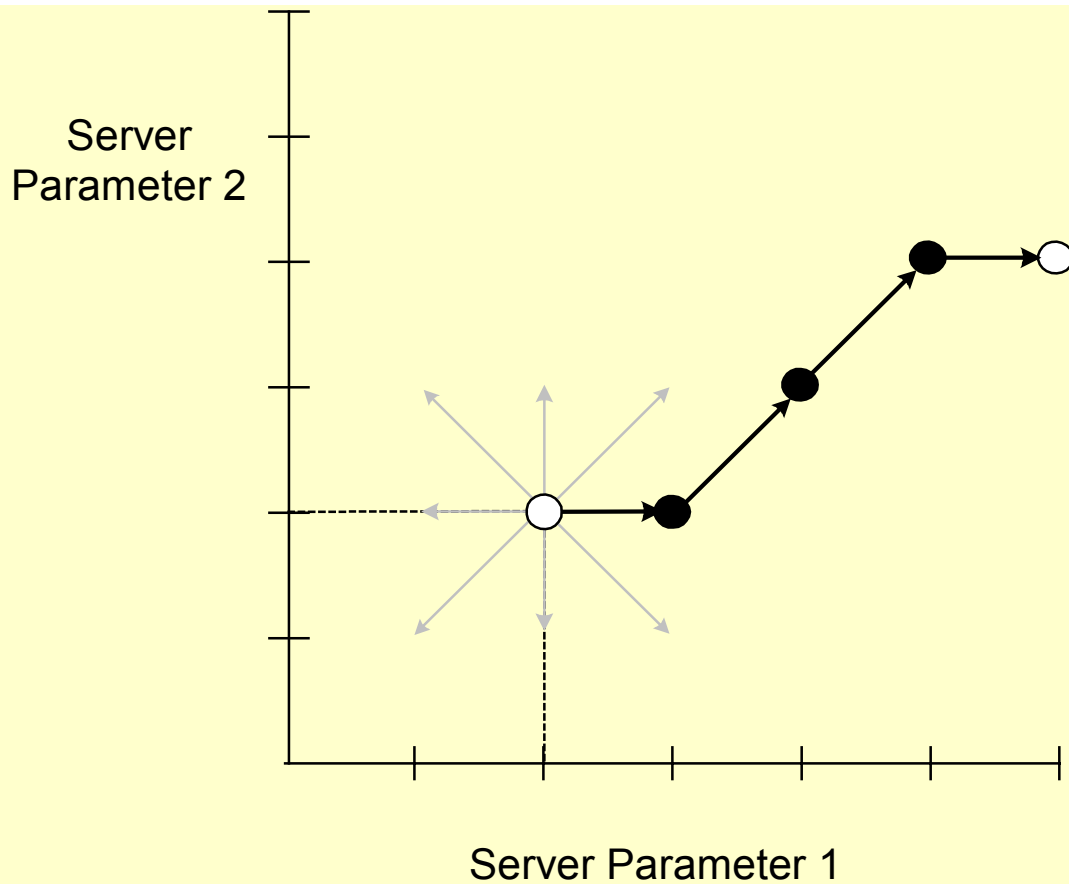
QoS_1 , QoS_2 , and QoS_3 are determined by a **predictive queuing model** of the site.

$$QoS = f(\vec{W}, c_1, c_2, \dots, c_m)$$

E-commerce Site Queuing Model



Heuristic Optimization Approach



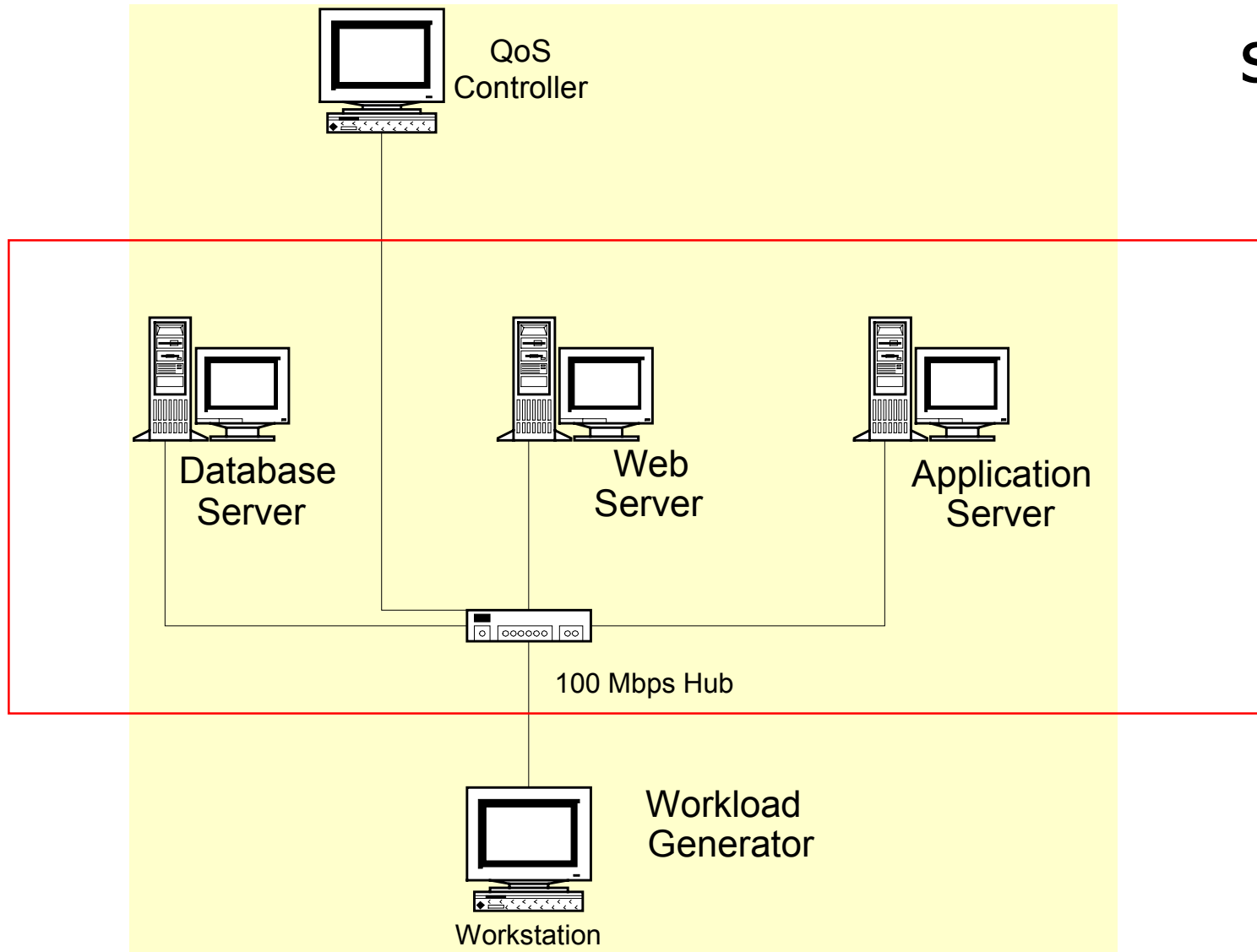
- Parameters for the queuing model are collected dynamically from the site.
- The QoS values for surrounding points are calculated
- The path with the greatest QoS gain is chosen. If no configuration improves the QoS or a limit is reached, the search ends

Dynamic QoS Control for E-commerce

- ❑ Definition of a combined QoS metric.
- ❑ Use of hill-climbing techniques combined with predictive queuing models.
- ❑ Implemented a TPC-W site in a multi-tier architecture.

Prototype Configuration

TPC-W
site



TPC-W: an E-commerce benchmark by the TPC

- ❑ www.tpc.org
- ❑ Designed to mimic operation of an e-commerce site (e-tailer).
- ❑ Scalable in number of concurrent users and in the database size.
- ❑ Transactions generated by TPC-W include:
 - Browsing activities (e.g., browse, search, select, view product detail)
 - Product order activities (e.g., shopping cart, login, register, buy request, and buy confirm)
- ❑ Database transactions must be ACID.
- ❑ Security through SSL is used for authentication.

TPC-W Metrics: throughput and cost/throughput

- ❑ WIPS (Web Interactions Per Second) during shopping mix sessions. Specified as WIPS@number_items
- ❑ WIPSp – Web Interactions Per Second during browsing mix sessions.
- ❑ WIPSo – Web Interactions Per Second during ordering mix sessions.
- ❑ Cost/Performance

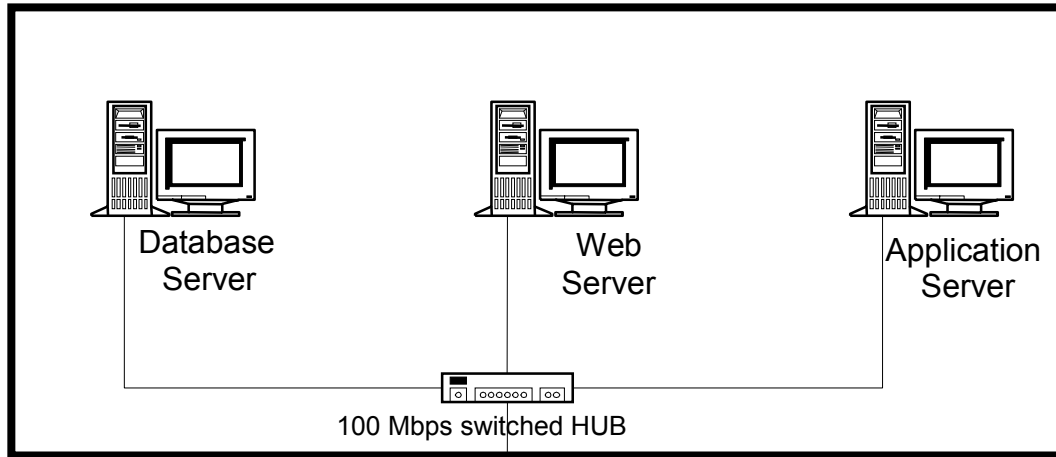
$$\frac{\text{Hdw Cost} + \text{Softw Cost} + \text{Maint. Cost}}{\text{WIPS}}$$

Example of TPC-W for 10,000 Items in the Catalog

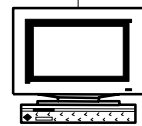
| Rank | System | WIPS | \$/WIPS |
|------|--------|-------|----------|
| 1 | A | 5,745 | \$ 69.00 |
| 2 | B | 3,130 | \$ 67.50 |
| 3 | C | 3,008 | \$ 81.77 |
| 4 | D | 1,262 | \$277.08 |

- ❑ the total price of System A is \$396,405, i.e., $5,745 \times \$69.00$.
- ❑ system D costs almost the same, i.e., \$349,675 but can only deliver 22% of the maximum throughput measured in WIPS.

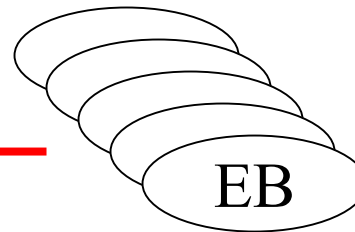
TPC Emulated Browsers (EBs)



E-commerce site:
System Under Test (SUT)



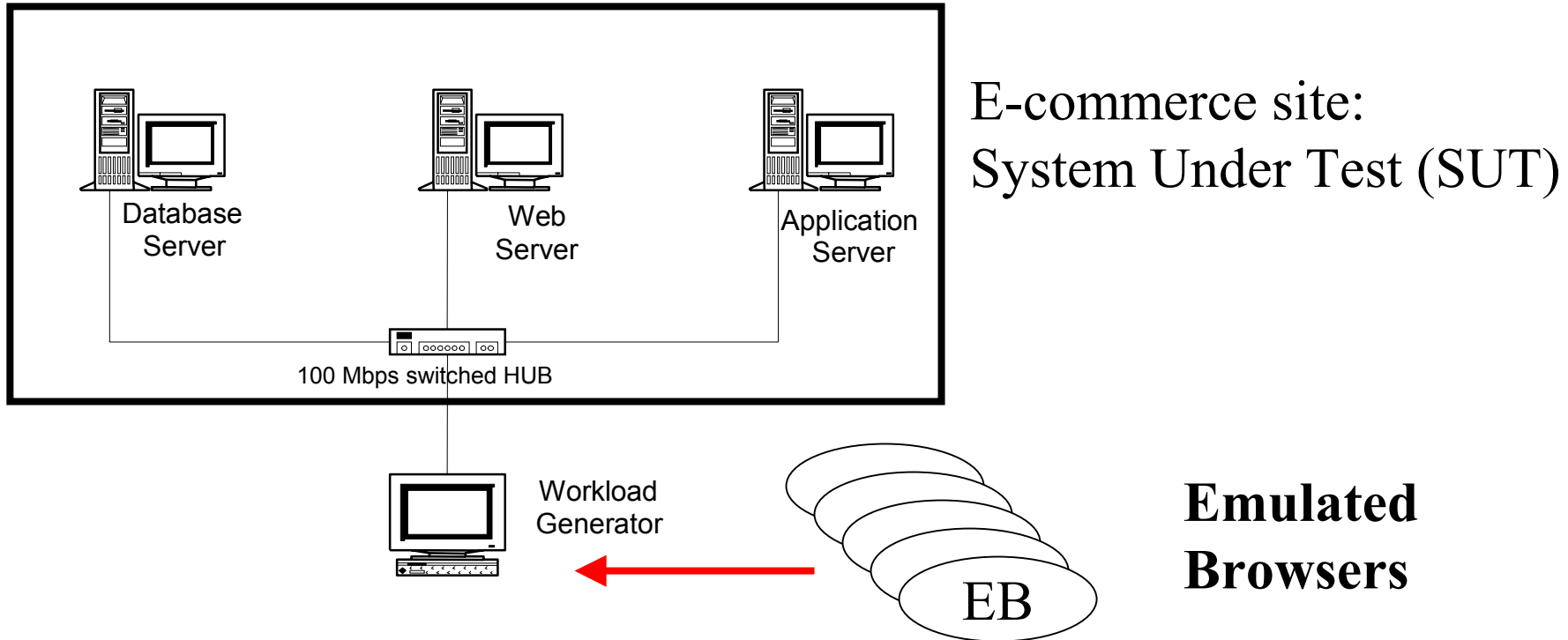
Workload
Generator



**Emulated
Browsers
(multithreaded)**

- ❑ Each EB starts a session and generates all requests of that session.
- ❑ The minimum duration of a session (USMD) is exponentially distributed with mean 15 minutes, truncated at 60 minutes.

TPC Emulated Browsers (EBs)



- ❑ Requests are separated by user think times (Z), which are exponentially distributed with mean 7 sec truncated at 70 sec.
- ❑ Response Time Law:

$$R = (\text{No. EBs}) / \text{WIPS} - Z$$

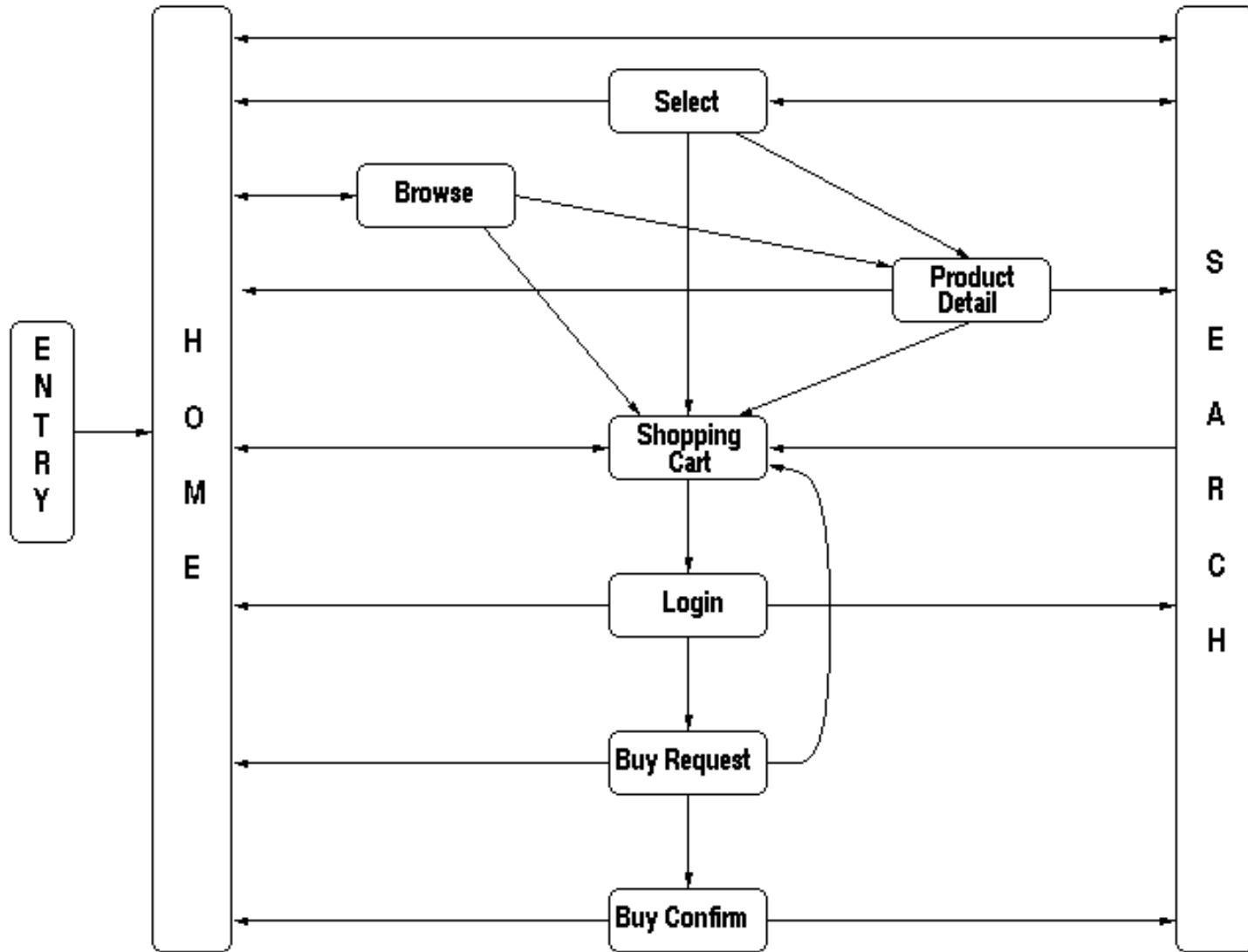
Use of Response Time Law to TPC-W Results

| Rank | System | WIPS | \$/WIPS |
|------|--------|-------|----------|
| 1 | A | 5,745 | \$ 69.00 |
| 2 | B | 3,130 | \$ 67.50 |
| 3 | C | 3,008 | \$ 81.77 |
| 4 | D | 1,262 | \$277.08 |

$$R = \frac{\text{No. EBs}}{\text{WIPS}} - Z$$

- ❑ Assume 50,000 concurrent users.
- ❑ System A: $R = 50,000/5,745 - 7 = 1.7$ sec.
- ❑ System D: $R = 50,000/1,262 - 7 = 32.6$ sec.

TPC-W CBMG



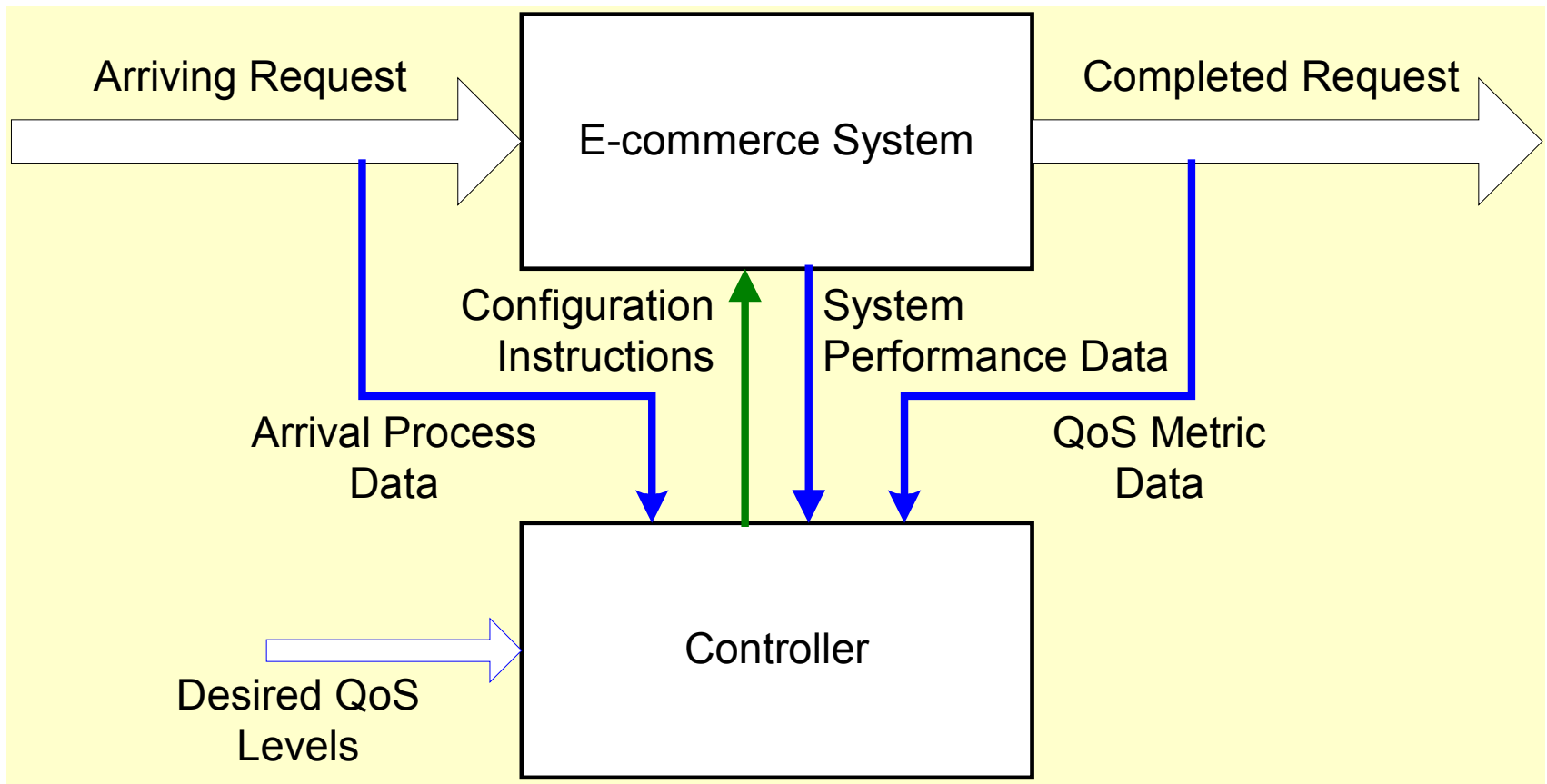
TPC-W Types of Sessions

- ❑ Browsing mix: 95% browse interactions and 5% ordering interactions – 0.69% buy to visit ratio.
- ❑ Shopping mix: 80% browse interactions and 20% ordering interactions – 1.2% buy to visit ratio.
- ❑ Ordering mix: 50% browse interactions and 50% ordering interactions – 10.18% buy to visit ratio.

Dynamic QoS Control for E-commerce

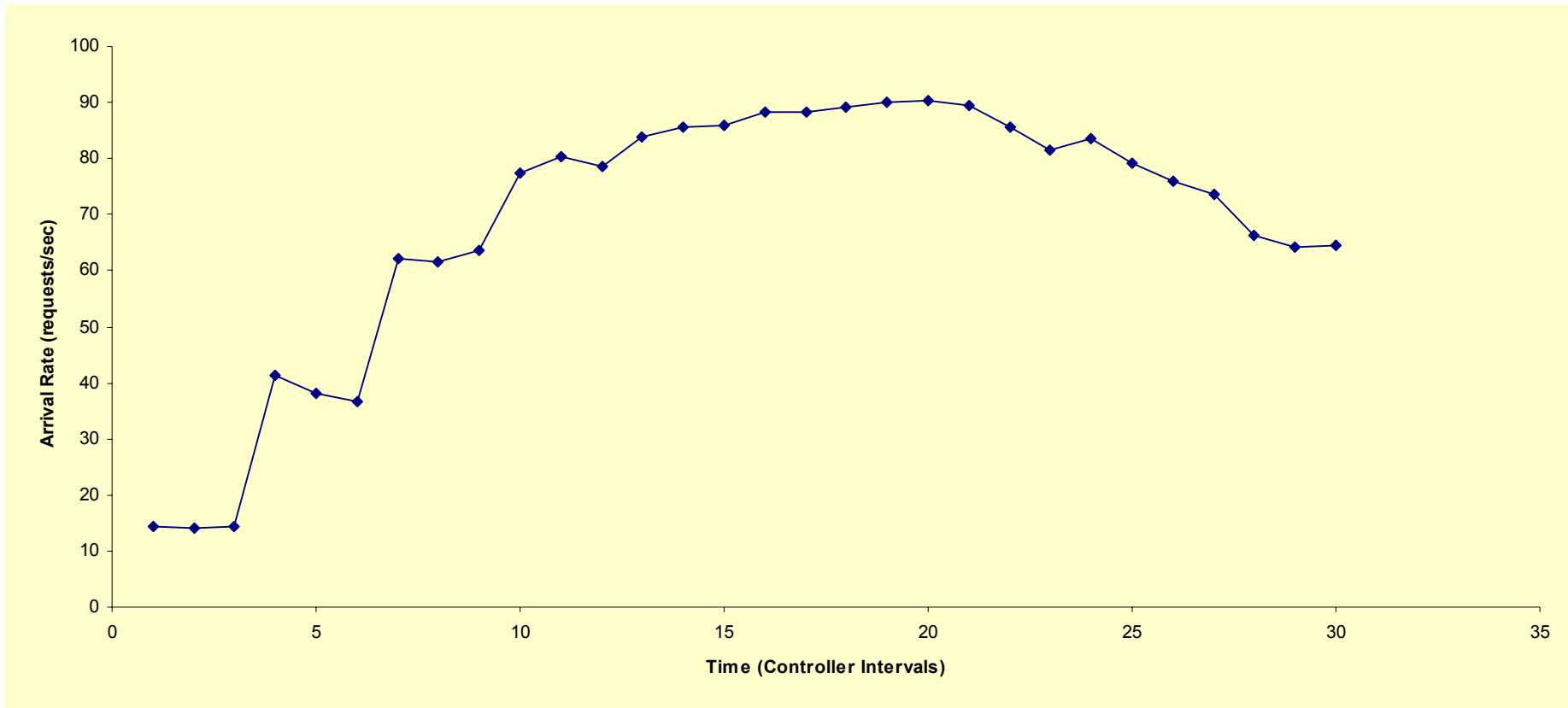
- ❑ Definition of a combined QoS metric.
- ❑ Use of hill-climbing techniques combined with predictive queuing models.
- ❑ Implemented a TPC-W site in a multi-tier architecture.
- ❑ Implemented a QoS Controller on a dedicated machine and evaluated the approach.

Dynamic QoS Controller

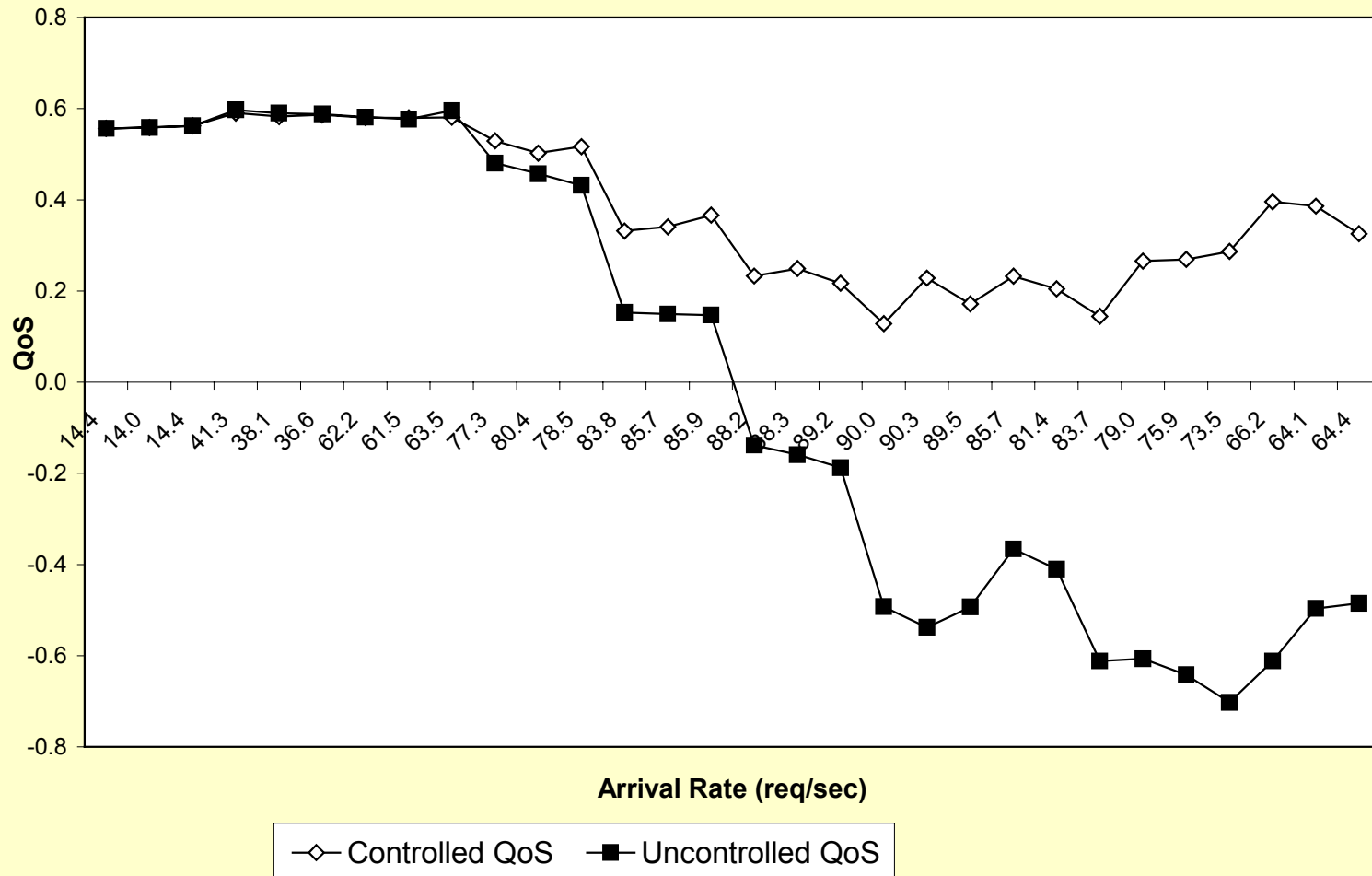


Experiment Results

Arrival rate

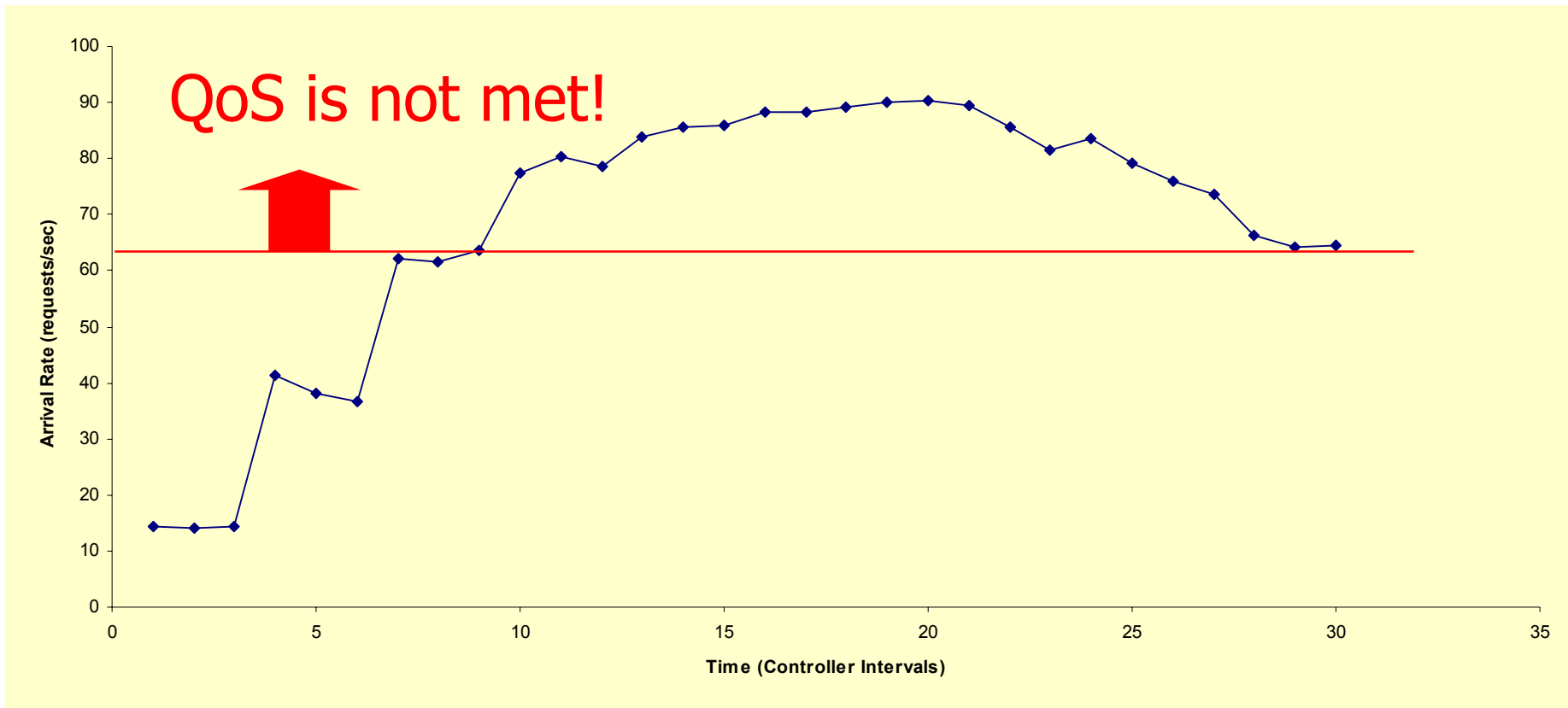


Results of QoS Controller



Experiment Results

Arrival rate



Concluding Remarks

- ❑ Hierarchical approach to workload characterization
- ❑ Some findings:
 - Session lengths are heavy-tailed
 - most sessions last less than 1000 sec
 - 88% of the sessions have less than 10 requests
 - Growing agent activity:
 - 33% of the requests generated by robots
 - Power law distributions:
 - popularity of search terms follows Zipf's Law
 - Predictability on fine time scales:
 - long range dependency between 4 and 4096 sec

Concluding Remarks (cont'd)

□ Some findings (cont'd):

- Robots can consume considerable resources.
- Crawlers consume more resources than shopbots.
- Utilization peaks increase in intensity with finer time scales.
- Robots significantly increase the miss ratio of server-side caches.
- Crawlers have a reference pattern that completely disrupt reference locality assumptions.
- Caches and servers should treat human- and robot-generated requests differently.

Concluding Remarks (cont'd)

- ❑ Performance models are useful to guide optimization techniques to dynamically control the QoS of e-commerce sites.

Bibliography

- ❑ “Capacity Planning for Web Services; models, methods, and metrics,” Menascé and Almeida, Prentice Hall, 2002.
- ❑ “Scaling for E-Business: technologies, models, performance, and capacity planning,” Menascé and Almeida, Prentice Hall, 2000.
- ❑ “In Search of Invariants in E-commerce Workloads,” Menascé, Almeida, Riedi, Ribeiro, Fonseca, and Meira, 2000 ACM Conference on Electronic Commerce, Minneapolis, MN, Oct. 17-20, 2000.
- ❑ "Preserving QoS of E-commerce Sites Through Self-Tuning: A Performance Model Approach," (with R. Dodge and D. Barbara), *Proc. 2001 ACM Conference on E-commerce*, Tampa, FL, October 14-17, 2001.