A Model of a Web Server with Dynamic Content

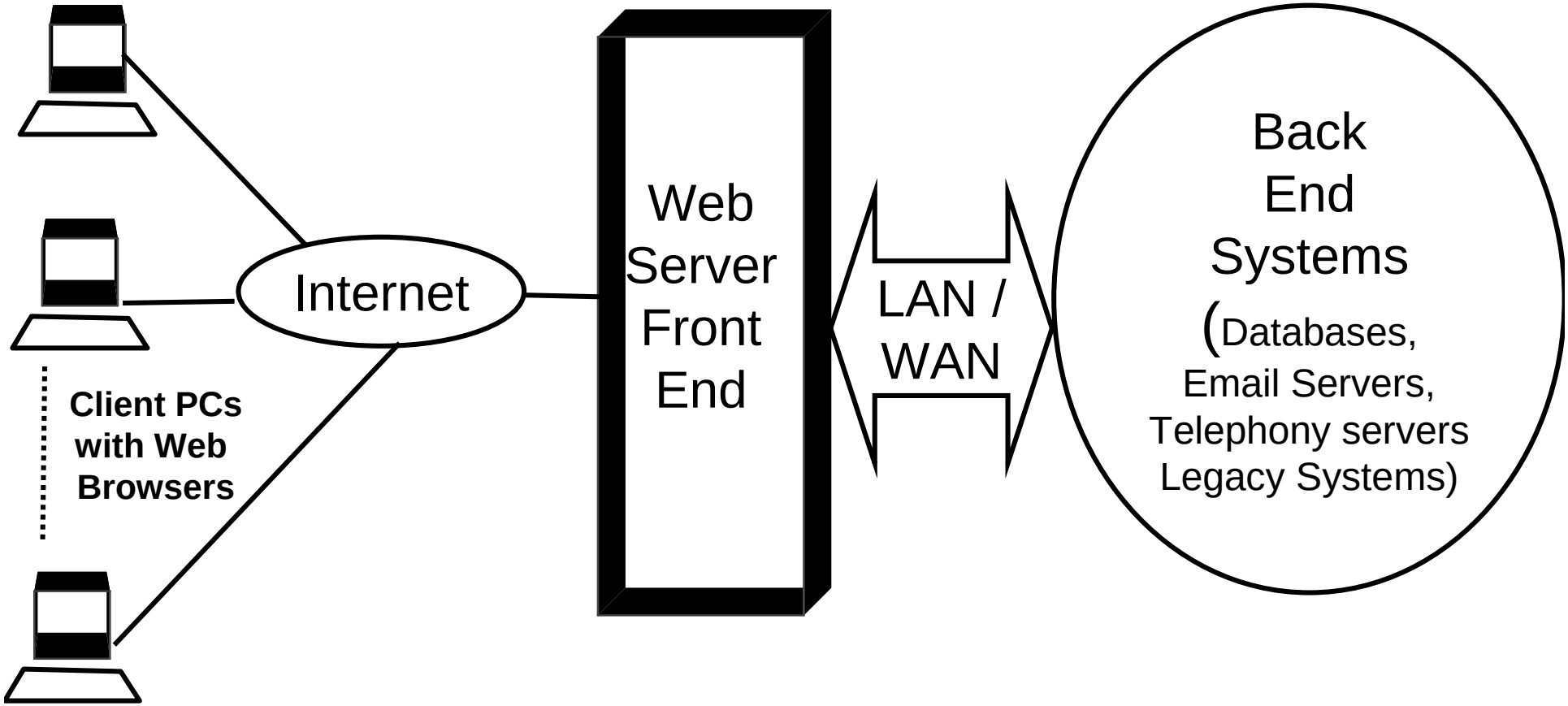
Varsha Mainkar

Network Design and Performance Analysis Department

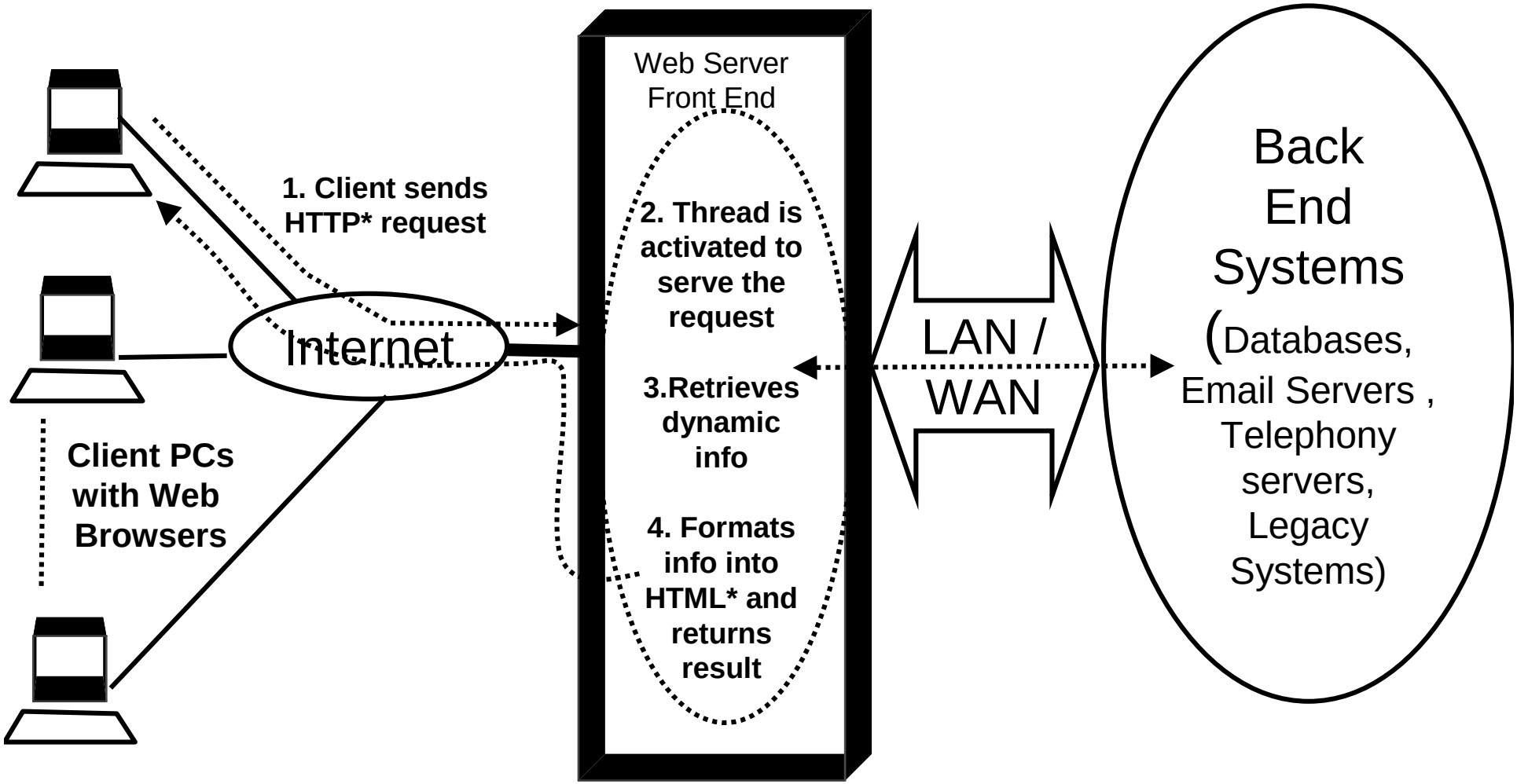
INFORMS Fall '99 meeting

Philadelphia, November 8, 1999

Typical Web-Based Service



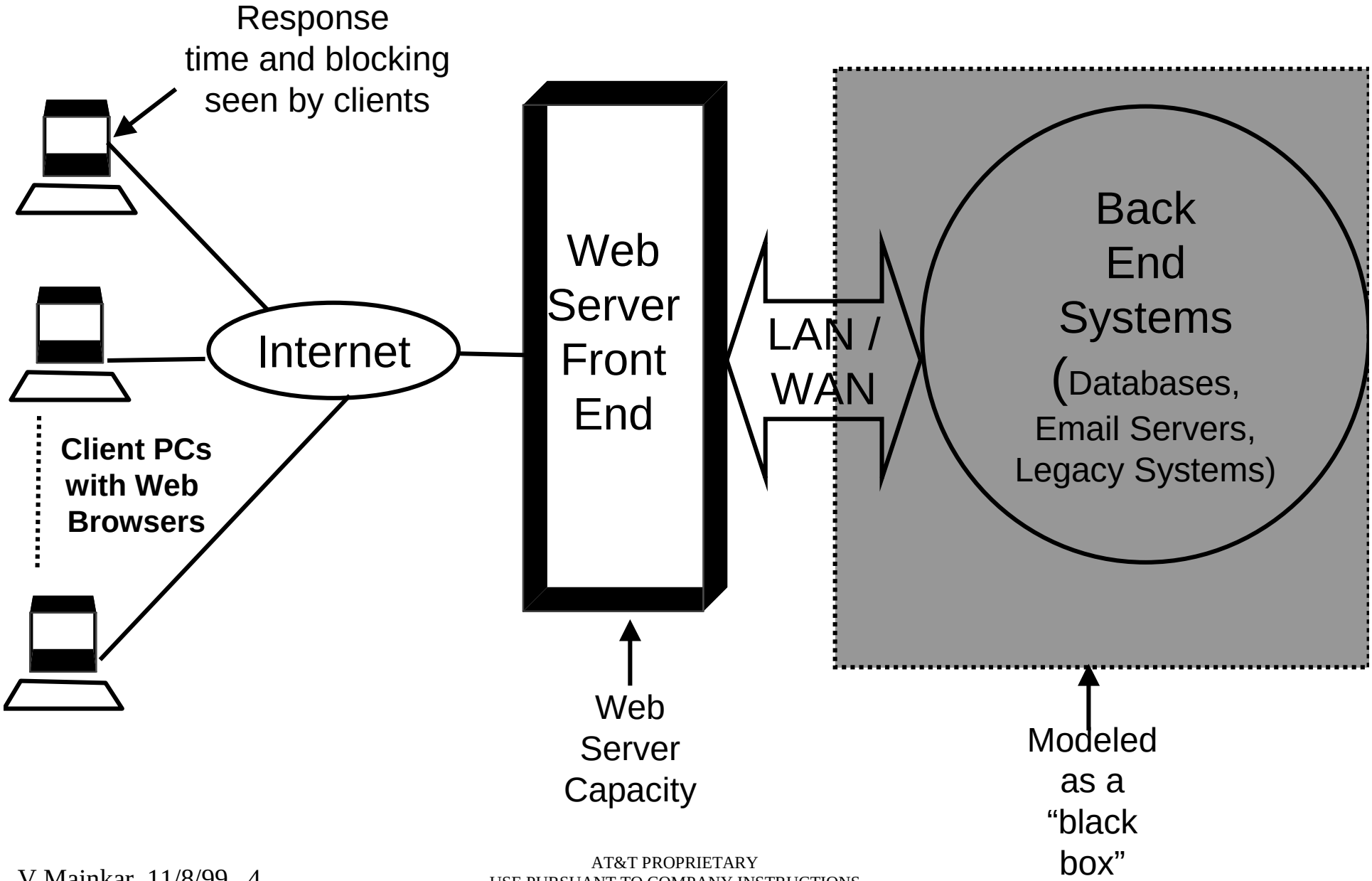
Typical Web-Based Service



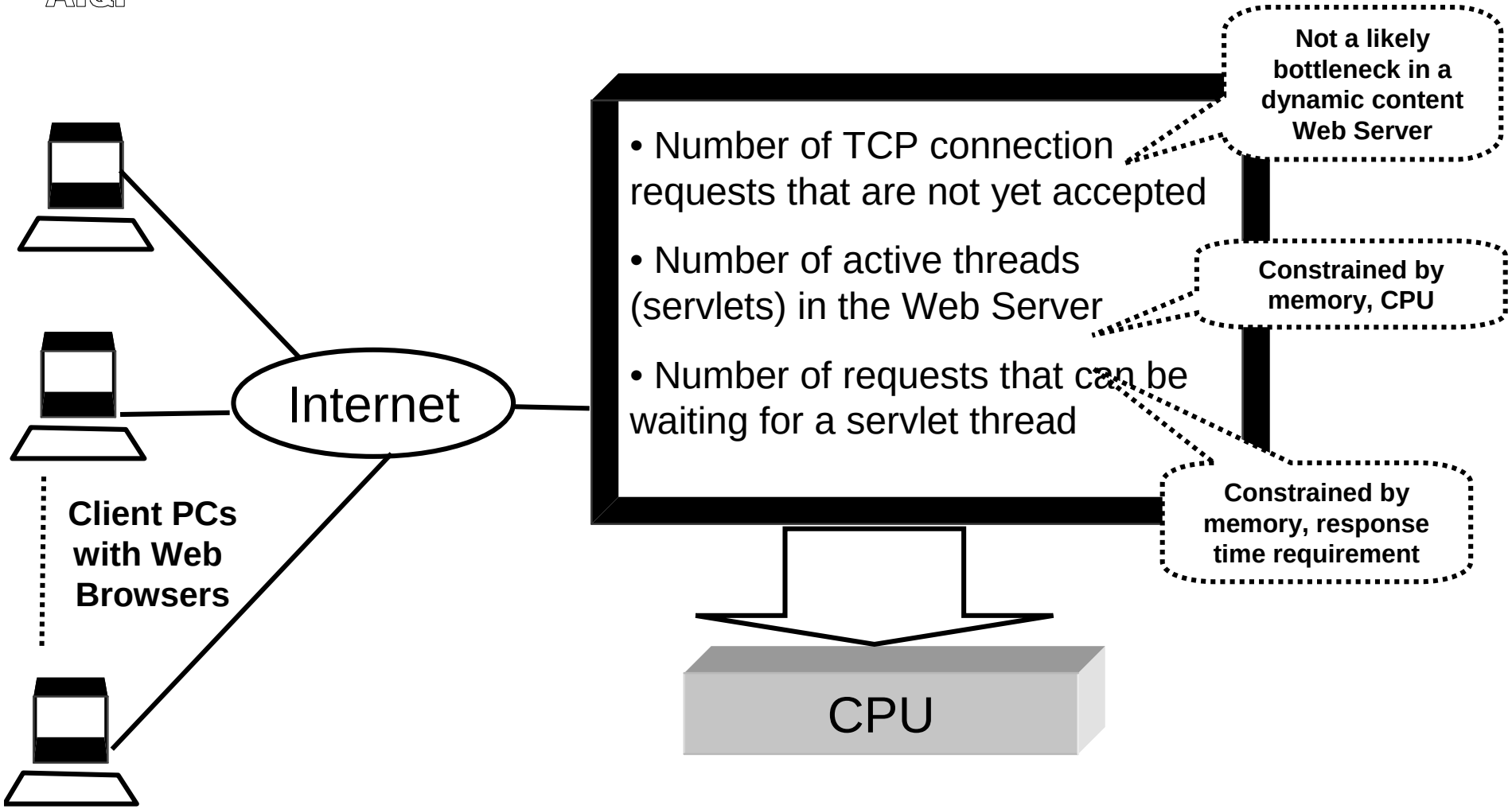
*HTTP = HyperText Transfer Protocol, the protocol used for Web transactions. HTML = HyperText Markup Language, the formatting language for Web pages

Performance Measures of a Web-Based Service

AT&T Labs

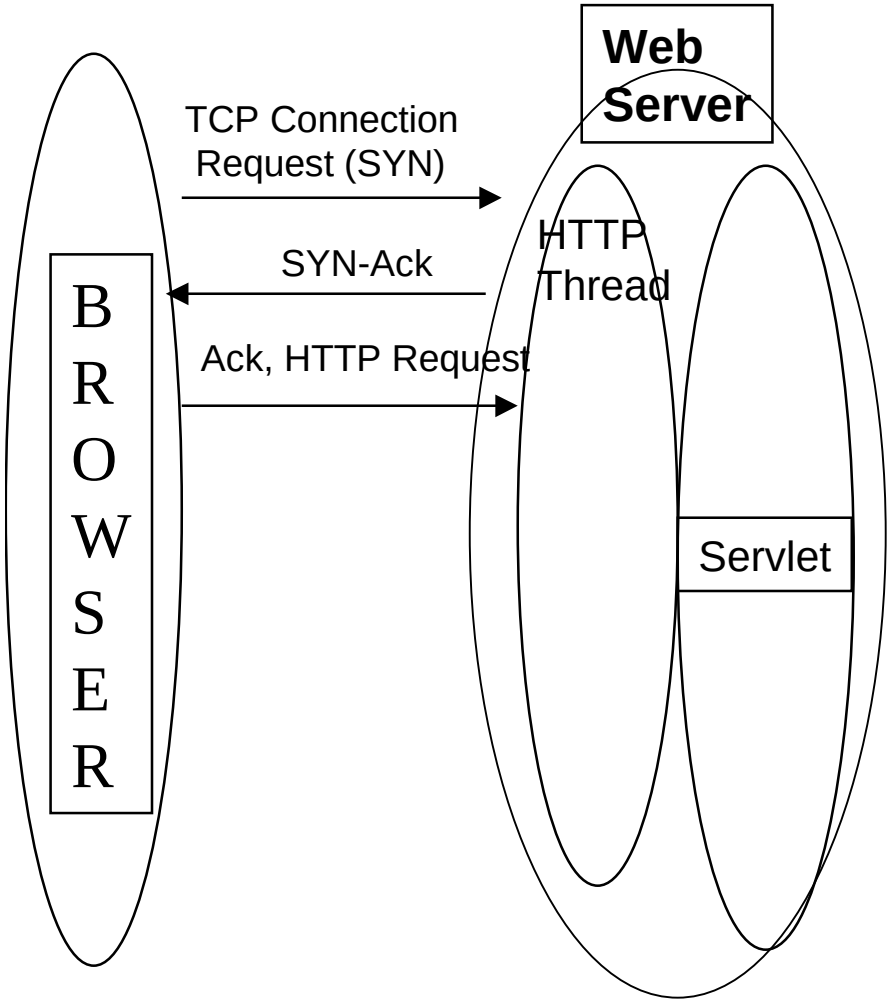


Server resources that are potential bottlenecks

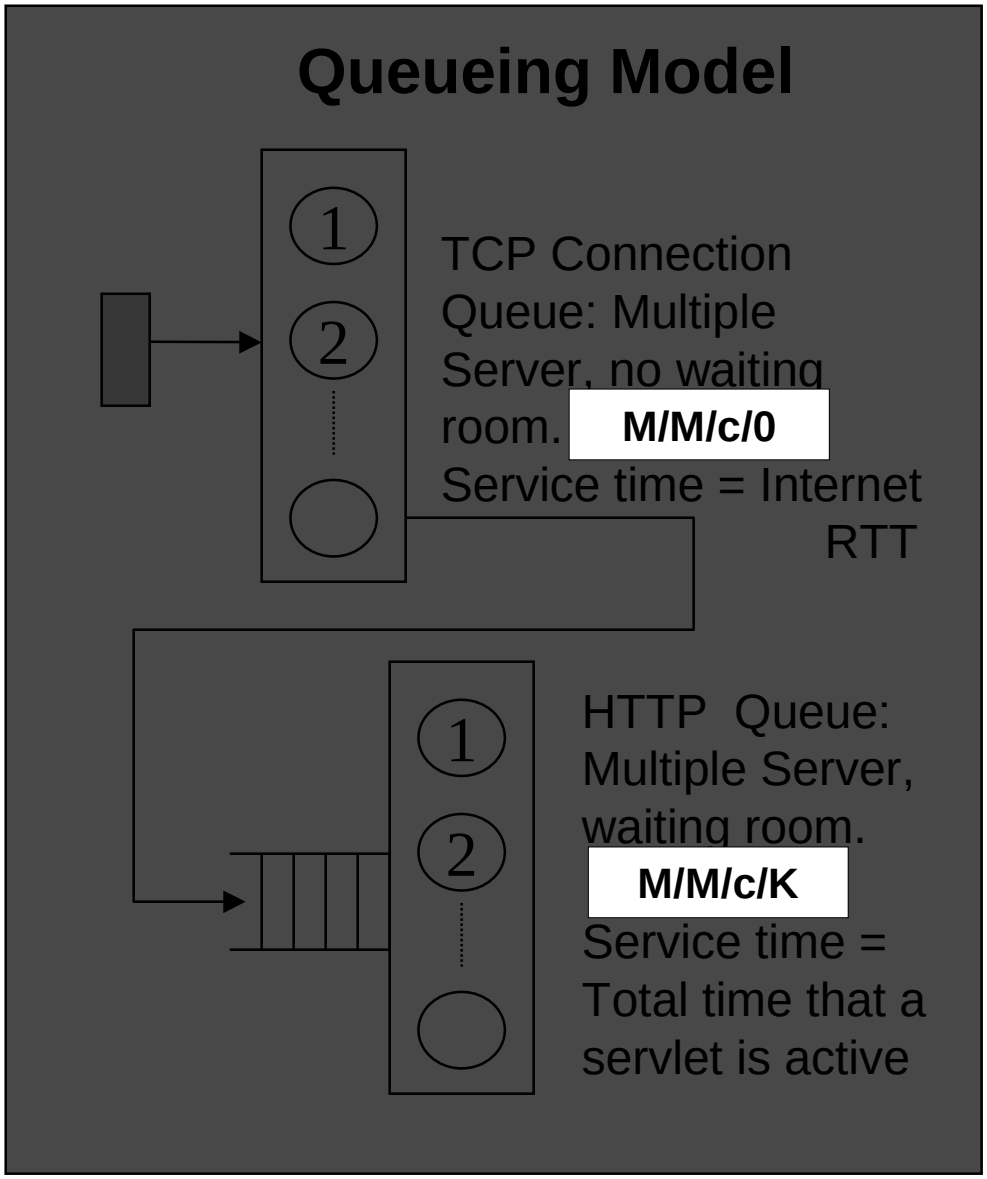


Network I/O subsystem may also be a potential bottleneck.
However, in this talk we focus on “CPU-bound” applications.

Web Transaction Flow & Queueing Model



Servlet = Thread spawned by a Netscape-type Web server, to handle dynamic processing
RTT = Round Trip Time



Queueing Model : CPU

Flow of typical servlet that generates dynamic content :

Request for CPU : t1 secs

Wait for I/O with back end system : w1 secs

Request for CPU : t2

Wait for I/O with back end system : w2

Request for CPU : t3

CPU modeled as a processor sharing queue
Arrival rate of requests to this queue = Web transaction throughput rate X number of CPU request segments in the servlet
Response time of a request segment of time *t* is *t/(1 - a)* where *a* is ...

Hierarchical Queueing Model

Then, holding time of servlet is =

w1 + w2 + ….

+ $R_{cpu}(t1) + R_{cpu}(t2) + R_{cpu}(t3) + \ldots$

where $R_{cpu}(t)$ is the response time of a request in the CPU queue

Finally, model variables are interdependent, so iterate until convergence is achieved.

Implemented in Mathematica.

Performance Measures

- Web Transaction Response time :
  - TCP connection set up time + HTTP queue waiting time + servlet holding time + 0.5 x Internet RTT
  - TCP connection set up time = 1.5 Internet RTT

- Blocking :
  - Blocking at TCP queue (B_{tcp}) and at HTTP queue (B_{http})
    - $B_{tcp} + (1 - B_{tcp}) B_{http}$
- Web Server Capacity : the transaction arrival rate at which a certain response time and blocking requirement is met

Model Validation

- Validation of this model was done against measurements on a simple test environment
- Test Environment :
  - Hardware : PC with 200 Mhz Pentium, 96 MB memory
  - OS : Windows NT 4.0 workstation
  - Web Server : Netscape Enterprise 3.6
- Web transaction :
  - A simple “test” servlet that uses the CPU for some time, then waits (sleeps), then uses CPU again, then waits…
  - Specifically : t1 = t2 = t3 = t4 = 2.1 seconds.
  - And w1 = 1 sec, w2 = 2 secs, w3 = 3 secs

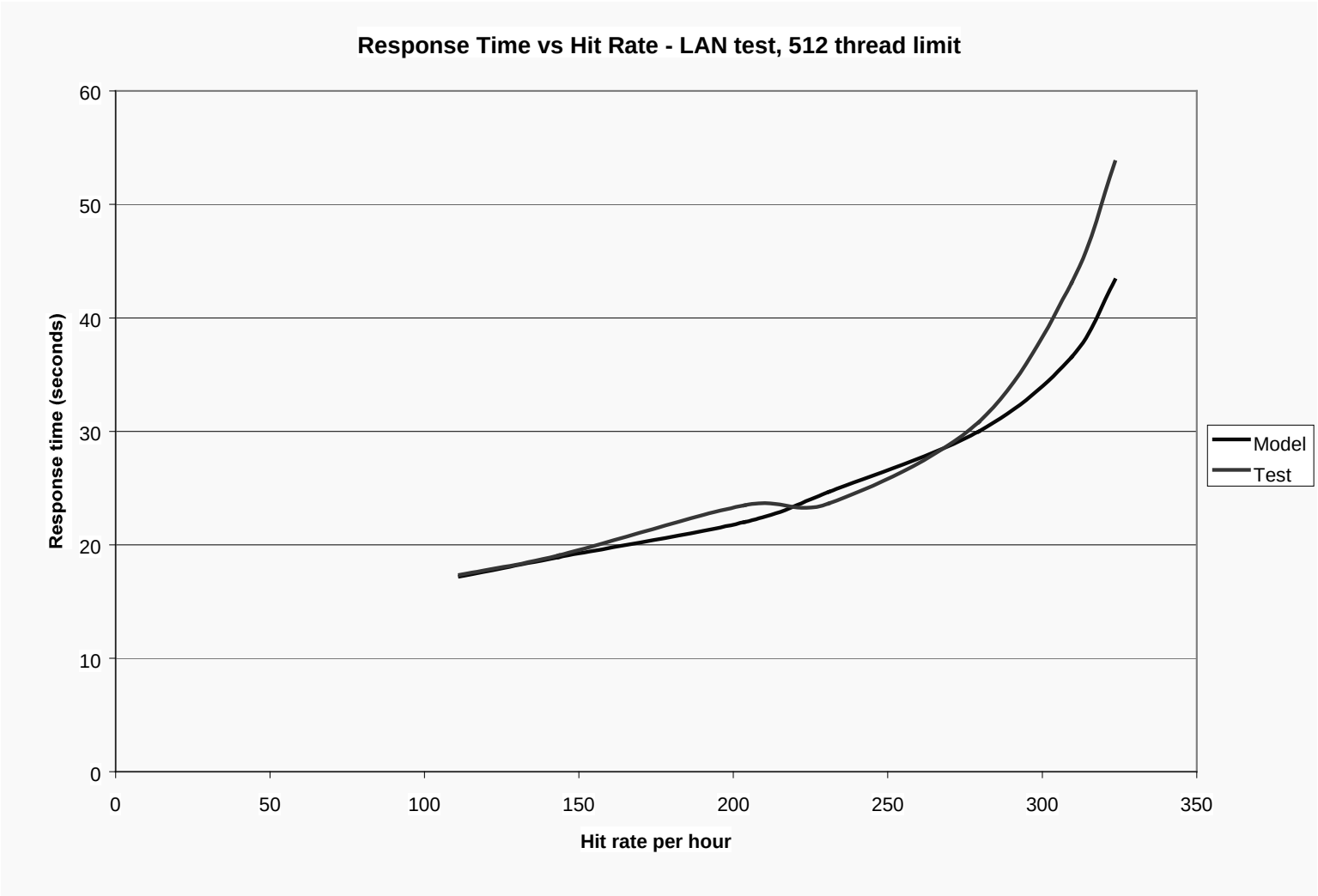
Model Validation

- Measurements were done using traffic generated by Silk Performer, using 1-11 users
- The following was measured
  - Average response time
  - Blocking percentage

Test vs Model : Scenario 1

AT&T Labs

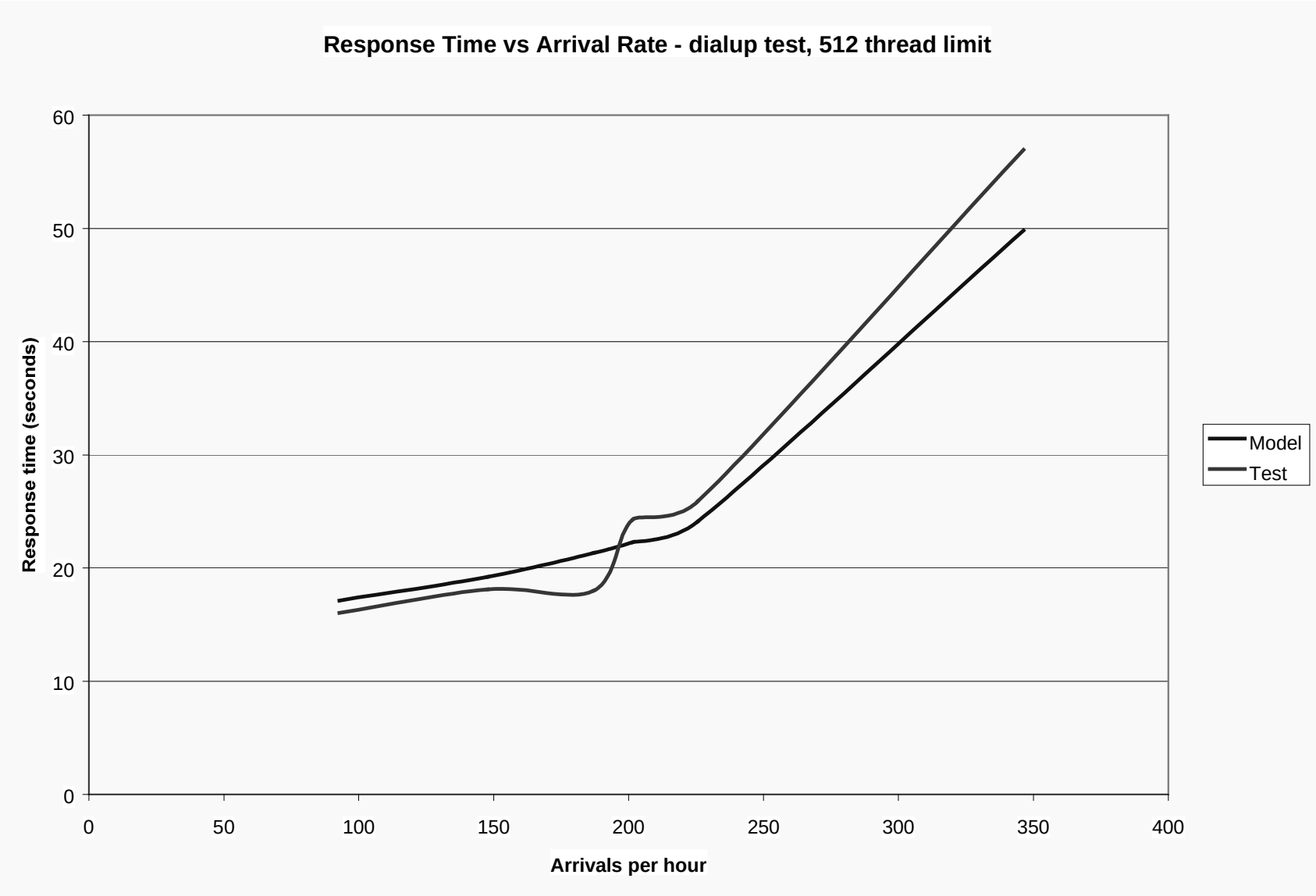
- Tests on a LAN, Web Server thread limit = 512
- Internet RTT ~ 0



Test vs Model : Scenario 2

AT&T Labs

Tests on a dial up line (Internet RTT ~ 140 ms), Web Server thread limit = 512

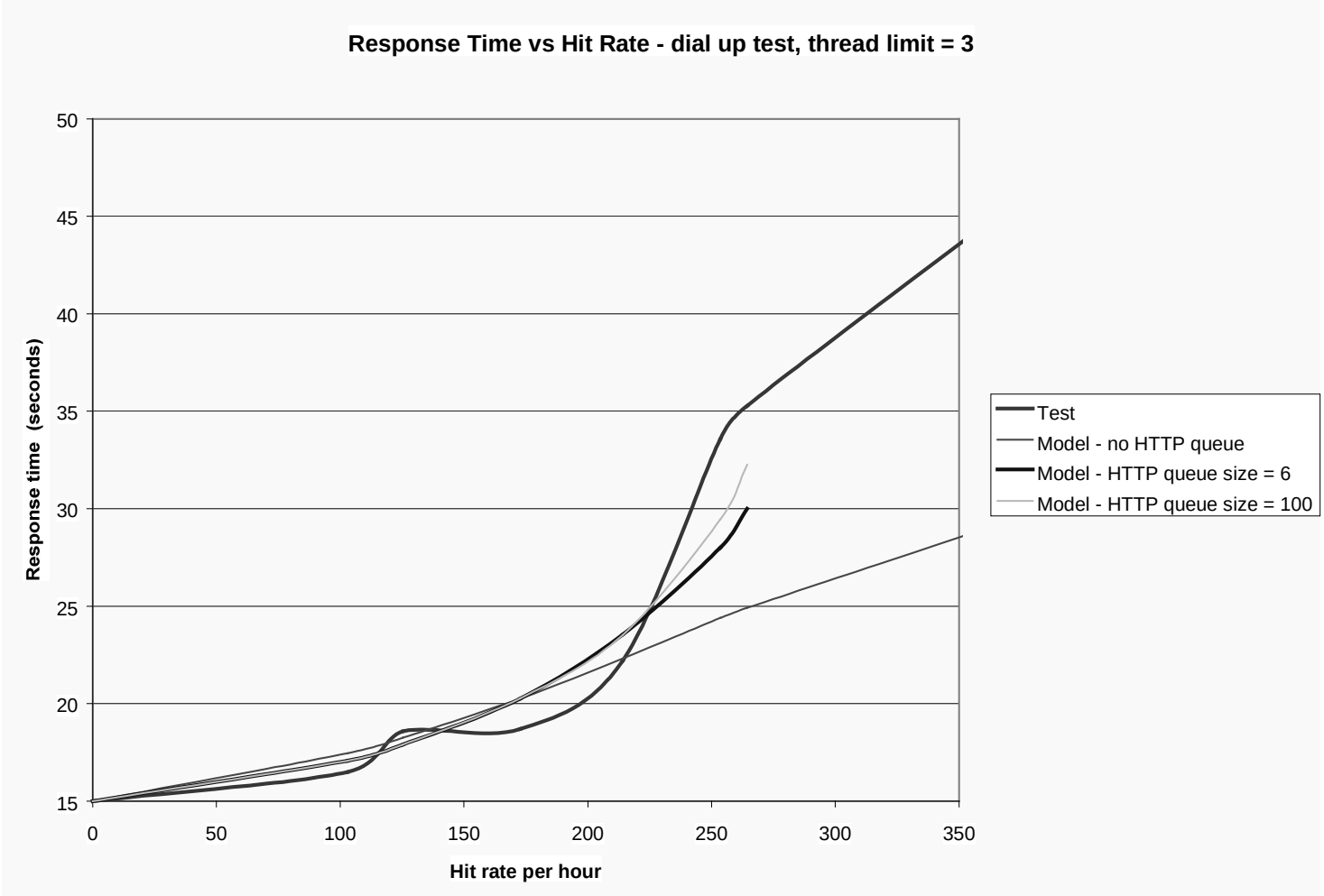


Test vs Model : Scenario 3

AT&T Labs

- Tests on a dial up line, Web Server thread limit = 3
- HTTP waiting room size : unknown
- Model can be used to estimate that size

Response TIme

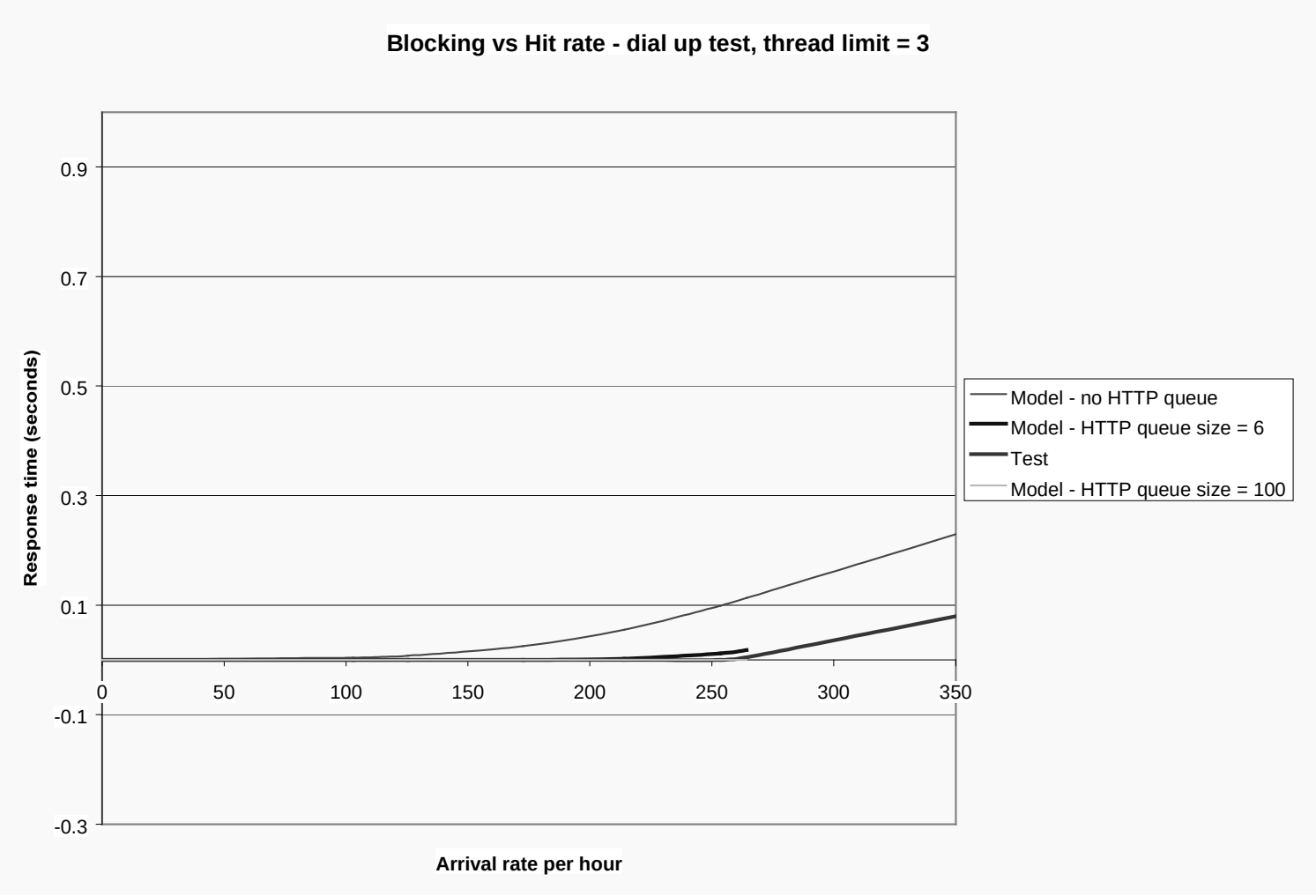


Test vs Model : Scenario 3

AT&T Labs

- Tests on a dial up line, Web Server thread limit = 3
- HTTP waiting room size : unknown
- Model can be used to estimate that size

Blocking



Conclusions

- Simple testing shows promising results -
  - Although model was simple, the model results were acceptably close to test results
  - There is a lot of room for improvement, which should result in closer estimation of measurements
- Modeling can help in quick prediction of performance even when parameters of Web Server or OS software are not known