<table>
<thead>
<tr>
<th>SESSION</th>
<th>&quot;Best Practices for High Volume Web Sites&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER</td>
<td>Presentation Number PE420511, E420521</td>
</tr>
<tr>
<td>SPEAKER</td>
<td>Dr. Willy Chiu</td>
</tr>
<tr>
<td>TITLE</td>
<td>VP, High Volume Web Sites</td>
</tr>
<tr>
<td>AFFILIATION</td>
<td>IBM</td>
</tr>
</tbody>
</table>

wchiu@us.ibm.com
I/T Professional's Survey
(Marketplace Needs to Improve the Performance of the Web Site?)

- Completely Agree: 15%
- Agree: 28%
- Neither Agree nor Disagree: 43%
- Disagree: 9%
- Completely Disagree: 5%

Total: 100

High Volume Web Sites
Best Practices

Life Cycle of a Web Site

Planning → Architecture → Design Build & Test → Deploy → Service delivery

Best Practices

- Design for scalability
- Manage end-to-end performance
- Plan for growth
- Design pages for performance
- Know your workload

Note: Material regarding the topics addressed in this presentation may be obtained at: "www.ibm.com/developerworks/library/hvws"
Patterns for e-Business & Workload Taxonomy

<table>
<thead>
<tr>
<th>Publish &amp; Subscribe (User-to-Data)</th>
<th>Customer Self-Service (User-to-Business)</th>
<th>Online Trading (User-to-Business)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Shopping (User-to-Online Buying)</td>
<td>Business-to-Business (Business-to-Business)</td>
<td></td>
</tr>
</tbody>
</table>


Work Site Loads
Usage Pattern Over One Day

When do peaks occur?

Retailer 1
Retailer 2
Gov.
Retailer 3

Hits/Hour (Thousands)

0 50 100 150 200 250

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Online Retailer
Monthly Trend in Peak Hits/Hour
Seasonal Effects

Peak hits/hour

Millions

0 1 2 3 4 5 6

Jan-99 Feb-99 Mar-99 Apr-99 May-99 Jun-99 Jul-99 Aug-99 Sep-99 Oct-99 Nov-99 Dec-99 Jan-00 Feb-00 Mar-00 Apr-00 May-00 Jun-00 Jul-00 Aug-00 Sep-00 Oct-00 Nov-00 Dec-00
Bursty Traffic Patterns
1998 Nagano Olympic Games

Contains very large bursts
Large traffic bursts - strong interest in Japan for events related to ski jumping

Sydney Olympics much less bursty due to 18 hour broadcast delay
Six steps to scaling:

- Understand the environment
- Categorize your workload
- Determine the components most impacted
- Select the scaling techniques to apply
- Apply the techniques
- Reevaluate
Highly Scalable
Online Stock Trading Web Site

- 12% signon
- 40% quote
- 3% trade

1/00 Secure site
110K Peak concurrent users
500K Peak trades/day
10.7M Peak transactions/day
80M Peak hits/day
4000 Peak trans/sec
Large Scale WebSphere EJB/Servlet Configuration

Workload Driver (Aptiva PC)

Workload Generators RS6000

Network Dispatchers RS6000

Web Servers (Servlets) RS6000

Business Logic (EJBs) RS6000

Legacy Code Simulators

Data Base Server

SERVLET THROUGHPUT

EJB THROUGHPUT

Transactions/Second

# of servlet nodes

0 50 100 150 200 250 300 350

3 6 9 12 15 18 21 24 27 30 33 36 39 42 45

Transactions/Second

# of Beans nodes

0 50 100 150 200 250 300 350

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
WebSphere Architecture Alternatives

Architectural choices typically driven by:
- Security, isolation from the open Internet
- Geographical considerations - separate data centers
- Separation of function (presentation and business logic)
Where do you focus your scalability and tuning?

End-to-end response time

Network Latency

Web Site Latency

Scope of performance management

<table>
<thead>
<tr>
<th>% Latency</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banking</td>
<td></td>
</tr>
<tr>
<td>Shopping</td>
<td></td>
</tr>
<tr>
<td>Trading</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service</th>
<th>ES</th>
<th>WS</th>
<th>AS</th>
<th>DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banking</td>
<td>5</td>
<td>11</td>
<td>23</td>
<td>61</td>
</tr>
<tr>
<td>Shopping</td>
<td>26</td>
<td>54</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Trading</td>
<td>5</td>
<td>27</td>
<td>53</td>
<td>15</td>
</tr>
</tbody>
</table>

Legend:
- **Blue** = Edge Server
- **Yellow** = Web Server
- **Green** = Application Server
- **Light Blue** = Data Base Server
# Workload Characteristics for Scaling

<table>
<thead>
<tr>
<th>System Workload Characteristics</th>
<th>Publish/Subscribe</th>
<th>Online Shopping</th>
<th>Customer Self-Service</th>
<th>Trading</th>
<th>Business-To-Business</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transaction Volumes</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Dynamic Content</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Dynamic Searches</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>User Specific Responses (Personalization)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Cross-session Information</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Transaction Complexity</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Legacy Integration</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Transaction Volume Swings</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Data Volatility</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Number of Content Publishers/Sources</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Page Content Volatility</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Number of Unique Items per page</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Number of Page Views</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Percentage of Secure Pages</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Security, Authentication etc.</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

- ○ Low
- ● Medium
- ● High

Refer to "Design for Scalability" White Paper for further information

"[www.ibm.com/developerworks/library/hvws]"
# How Do Workloads Dynamics Impact Solution Components?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Transaction volumes</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Dynamic content</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Dynamic searches</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>User specific responses (personalization)</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Cross-session information</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>Transaction complexity</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Legacy integration</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Transaction volume swings</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Data volatility</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Number of content publishers/sources</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Page content volatility</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>Number of unique items per page</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Number of page views</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Percentage of secure pages</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Security, authentication, etc.</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

○ Low  ○ Medium  ● High

Refer to "Design for Scalability" White Paper for further information
How Do Scaling Techniques Relate to Scaling Objectives?

<table>
<thead>
<tr>
<th>Scaling Technique</th>
<th>Increase Power</th>
<th>Improve Efficiency</th>
<th>Shift / Reduce Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster Machine</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Replicate Machines</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specialized Machines</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Segmented Workload</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Request Batching</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>User Data Aggregation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connection Management</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Caching</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

All of these techniques are in use somewhere, yet our top customers are not using all of the techniques that would be useful.

Refer to "Design for Scalability" White Paper for further information "www.ibm.com/developerworks/library/hvws"
Best practice: Follow a methodology to optimize for scalability

- Understand thoroughly your application environment
- Categorize your workload
- Determine the components most impacted
- Select the scaling techniques to apply
- Apply the techniques
Planning for Growth
Lessons Learned

- Determine your workload characteristics
  - Trends
  - Burstiness

- Plan your infrastructure from end to end
  - Leads to balanced solutions

- Integrate capacity planning into your I/T and business processes
Capacity Planning Process with the HVWS Simulator

Online shopping
- **Browse**
  - Home page
  - Choose department
  - Choose category
  - Choose subcategory
  - Choose product 1
  - etc.
- **Search**
  - Home page
  - Select product search
  - Submit keyword
  - etc.
- **Buy**
  - Home page
  - Select "AtHome" department
  - Select "Candles" category
  - Select "Scented" subcategory
  - etc.

Business Model
- Business and workload patterns

User Behavior Model
- Arrival patterns and transition matrices

Infrastructure Model
- Server structure, resource use

Configuration
## Web Page Measurement Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Criteria Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average server response time</td>
<td>&lt;0.5 sec</td>
</tr>
<tr>
<td>Number of items per page</td>
<td>&lt;20</td>
</tr>
<tr>
<td>Page load time</td>
<td>&lt;30 sec *</td>
</tr>
<tr>
<td>Page size in bytes</td>
<td>&lt;64K</td>
</tr>
</tbody>
</table>

### Seconds to load *

<table>
<thead>
<tr>
<th>Seconds to load</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 10</td>
<td>Excellent</td>
</tr>
<tr>
<td>10-15</td>
<td>Very good</td>
</tr>
<tr>
<td>15-20</td>
<td>Good</td>
</tr>
<tr>
<td>20-25</td>
<td>Adequate</td>
</tr>
<tr>
<td>25-30</td>
<td>Slow</td>
</tr>
<tr>
<td>Over 30</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>

* Based on 56K connection speed

- Minimize number, size, and complexity of page items -- consider the business value of each
- Store and retrieve items used more than once from the browser cache
- Avoid server redirection, use persistent connections, and use same code level across servers
- Avoid unnecessary use of SSL, banner ads, and dynamic pages
- Use preproduction utilities that remove extra white space from the source HTML (e.g., GZIP)
- Structure sequences to fully exploit parallelism - request items early

Refer to "Design Pages for Performance" White Paper for further information

WebSphere Studio PageDetailer Analysis

- Excessive page size of 127KB -- should be below 64KB (consider reduction of page size & use of GZIP to compress files)

- High server response times (0.5-2.5 seconds)
- Objective is 0.5 seconds

- 50 items requested -- should generally be <= 20
- Considerable overhead to request/respond/deliver
- Considerable overhead to establish/close 37 connections
- Consider consolidation of items where possible

[Image of chart with annotations]
Manage End-to-End Performance Lessons Learned

- Use proactive end-to-end measurements to identify constraints and bottlenecks

- Consider performance management methodology from the beginning
  - e-Business performance requirements
  - New Web site architectures

- Enforce processes that link performance objectives to planning, design and operations
e-Business Web Application System
End-To-End Performance Management Methodology

Methodology Overview

Establish Performance Budget  Monitor  Analyze  Predict & Plan

Focal points for performance management
HVWS White Papers

• Design for Scalability
  • Describes component selection and management techniques you can use to make your Web site ready to adapt to increasing traffic

• Planning for Growth
  • Introduces a methodology for capacity planning, identifying workload patterns and configuring site infrastructure

• Design Pages for Performance
  • Discusses what drives Web page download times
  • Introduces page design practices that can reduce download time and improve resource utilization

• Manage Web Site Performance
  • Discusses end to end performance management

• Web Site Personalization
  • Introduces current and future techniques for personalizing your Web site
"Best Practices for High Volume Web Sites"

Presentation Number  PE420511,  E420521

Dr. Willy Chiu

VP, High Volume Web Sites

IBM

wchiu@us.ibm.com