White Paper: Assessing Performance & Response Time Requirements

Mark Houghtlin
IBM Cloud Advisory Services
& Application Performance Optimization Consulting Practice
Agenda

- Workload Transformation Analysis
- Workload Performance & Response time requirements
- End-to-End transaction tracking
- Application Performance Management technology for Cloud
A robust analysis of your workloads can help you identify candidates for your target cloud environment(s) and can help you gain an understanding of the viability, operational cost changes and migration impact.

- An essential factor for any cloud strategy, workloads represent collections of your key IT system components and the relationships among them.
- These components can include web servers and application servers, databases and behavioral policies such as availability, security and performance.
- A comprehensive analysis of your workloads enables you to answer key questions such as:
  - Can the workload run in the target cloud environment?
  - Is it compatible with my infrastructure, middleware and operating system image?
  - Can the target cloud environment satisfy my performance, availability and other nonfunctional requirements such as response time objectives?
  - Can the target cloud environment comply with applicable security, privacy and regulatory requirements?
  - What benefits can we realize from migration? (For example, can we lower overall operating costs? Improve service levels?)
  - How challenging might it be to get to cloud given my current workloads?
  - How would they support our business and IT objectives?
Workload transformation analysis for Cloud should provide detailed analysis of IT and business workloads to help determine the best fit, delivering the insights to make better strategic decisions about cloud migration and where it makes the most sense for workloads and data to reside.

**The challenges the analysis can address:**
- Now that I am ready for cloud, which workloads fit my target cloud(s)?
- What is the migration impact?
- What are the cost savings of moving those workloads to the cloud?
- What application integrations are of most concern?
- Will performance requirements be achievable?

**What an analysis approach should do:**
- Analyze workloads of both business applications and infrastructure components for fit for target cloud(s)
- Analyze and map application integration points to business requirements
- Uncover the true nature of each integration point and its suitability for cloud
- Determine which multi-connection application(s) can move to the cloud
- Determine which workloads will move to cloud

**The value the analysis can provide:**
- Identification and prioritization of workloads for possible delivery from a target cloud while quantifying the benefit
- Help ensure performance objectives will be met
- Analytics needed to facilitate better decision making when moving to cloud
- Help accelerate adoption of your cloud strategy
One of the identified barriers to cloud adoption is client concerns around performance of their workloads in the new environment

- It is common for an application being migrated to a cloud service to have connections of various kinds with other applications and systems
- Application owners need to understand and address the impact of these connections on overall transaction performance and availability
- Consider what happens if an application is migrated to a cloud service and the other applications it has connections to remain in-house. The following questions arise and must be properly addressed:
  - What protocols do these applications use to talk to each other?
  - Can the same connection continue working assuming the protocol is supported by the cloud service?
  - Does the response time remain adequate, given the frequency of transactions, the size of the payload, the bandwidth and latency of the connection to and from the cloud service?
- Failure to properly migrate applications to cloud computing could ultimately result in higher costs and potential loss of business, thus canceling any of the potential benefits of cloud computing

To answer these questions accurately and with confidence requires a detailed assessment of the performance and response time requirements the connection must support
A thorough analysis of workload connections helps to mitigate risk associated with data transfer and latency requirements which are critical elements in evaluating overall Cloud Affinity for specific workloads.

**Fit for Cloud**  
Not big deal if it takes a long time to transfer big data  
Need to transfer big data at high speed  
Response quickly

**Not a good Fit for Cloud**  
(Fit for private Cloud)

Workloads which require small amounts of data transfer to complete a transaction or where the transfer can be done asynchronously have more affinity to running in the cloud.

Physical separation of servers (App, database, web server, and so on) which can occur when part of the workload transaction resides in the cloud can result in poor performance due to large objects, latency, and the nature of the application integration techniques and protocols employed.

Included in Overall Cloud Affinity Workshops
Consider the following - Synchronous -vs- Asynchronous or combination of both, and determining whether or not a transaction requires further analysis

• The application depicted below may have an implementation of "Check out" which performs request validation, stores the order data into a database, and then starts some asynchronous processing relating to the order by means of publishing a request to a topic in a messaging system, before responding to the user that the order is placed, without waiting for the asynchronous processing to complete

• For the sake of our discussion, we would be concerned with understanding the effect of separating any elements of the business transaction due to migration – for example, the eCommerce front-end application and the messaging system

• There is a strong likelihood that in this case, a longer latency in the processing of the published message by the asynchronous back-end process is likely to have almost no effect on the end users of the “Check out” operation
Managing and tracking transactions end-to-end are necessary to link underlying infrastructure components to actual user experience and business requirements – this is where BTM comes in

• As the IT environment has evolved into a shared infrastructure resource, monitoring just infrastructure components does not readily identify performance and availability issues

• Applications themselves become services to other applications creating a inherited architecture of dependencies with little understanding between support teams
  – In today’s highly distributed environments, each user request can flow through a large number of components. This makes it extremely difficult to understand how changes to individual elements affect performance and availability

• The Business transaction map identifies the end to end support of your primary IT asset – the business service or transaction

• The BTM methodology is designed to protect the business by:
  – documenting the relationship between the critical business assets, such as selling an airline ticket, to the technical complexities of the supporting IT architecture required to complete that job within the allotted time
  – applying programmed quantitative analysis to access changes to the business or architecture, defined as risks, in order to make calculated financial decisions to protect these business assets
  – quantifying business value and enabling ITIL based business and services level management
Predictive modeling Example 1

All business functions tested were projected to experience an increase in response time ranging from $\frac{1}{2}$ second for Redeem Loyalty Points to over 4 minutes for Supply Chain Productivity Reporting.

<table>
<thead>
<tr>
<th>Test</th>
<th>Baseline Response Time (sec)</th>
<th>Projected Response Time (sec)</th>
<th>Increase in Seconds</th>
<th>Percentage Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store Systems - Redeem Loyalty Points Step 5</td>
<td>0.059</td>
<td>0.316</td>
<td>0.257</td>
<td>439.3%</td>
</tr>
<tr>
<td>Store Systems - Redeem Loyalty Points Step 8</td>
<td>0.089</td>
<td>0.347</td>
<td>0.258</td>
<td>290.5%</td>
</tr>
<tr>
<td>Supply Chain - ProRep Step 3</td>
<td>69.433</td>
<td>104.094</td>
<td>34.661</td>
<td>49.9%</td>
</tr>
<tr>
<td>Supply Chain - ProRep Step 2</td>
<td>145.972</td>
<td>201.69</td>
<td>55.718</td>
<td>38.2%</td>
</tr>
<tr>
<td>BI - SAS Exports Step 2</td>
<td>197.836</td>
<td>254.976</td>
<td>57.140</td>
<td>28.9%</td>
</tr>
<tr>
<td>BI - SAS Exports Step 1</td>
<td>1842.209</td>
<td>1937.099</td>
<td>94.890</td>
<td>5.2%</td>
</tr>
<tr>
<td>Supply Chain - ProRep Step 5</td>
<td>88.201</td>
<td>229.2</td>
<td>140.999</td>
<td>159.9%</td>
</tr>
<tr>
<td>Supply Chain - Status Reporting</td>
<td>513.541</td>
<td>711.067</td>
<td>197.526</td>
<td>38.5%</td>
</tr>
</tbody>
</table>

- Client applications utilizing HTTP protocol such as WebSphere applications, are expected to experience the least impact (sub-second) from increased latency.
- Keystroke applications such as rshell, telnet, TN5250 and Citrix are others that experience little impact from increased latency.
- Impact to direct database access is dependent on the quantity and quality of queries but are likely to experience an impact to performance.
- Applications with dependencies on file shares are expected to have the highest impact to end user response time.

- Store Systems – Redeem Loyalty Points increased $\frac{1}{2}$ second across the 2 steps that moved data.
- BI – Daily SAS Exports increased 2 ½ minutes across the 2 steps that moved data.
- Supply Chain – ProRep increased 3 minutes 51 seconds across 3 of the 16 steps that moved data.
- Supply Chain – Status Reporting increased 3 minutes 18 seconds across 1 of the 7 steps that moved data.
Although rated as unacceptable to any change in response time, the “<task 3>” step of the <business function> transaction is expected to be unnoticeable with a projected increase of ¼ of a second.

- SOAP over HTTP is used across all tier pairs.
- The average application message size between WAS and the AS400 was 199 bytes.
- The maximum application message size between WAS and the AS400 was 840 bytes (AS400 -> WAS).

<table>
<thead>
<tr>
<th>Test</th>
<th>Response Time (sec)</th>
<th>Application Turns</th>
<th>Application Messages</th>
<th>Application Data (bytes)</th>
<th>Application Request/Response Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store Systems 1.5_b (Total)</td>
<td>0.059</td>
<td>18</td>
<td>21</td>
<td>5,209</td>
<td>289</td>
</tr>
<tr>
<td>POS Client</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTTP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS400UAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POS Client &lt;-&gt; HTTP</td>
<td>0.059</td>
<td>5</td>
<td>6</td>
<td>1,593</td>
<td>319</td>
</tr>
<tr>
<td>HTTP &lt;-&gt; WAS</td>
<td>0.055</td>
<td>5</td>
<td>6</td>
<td>1,822</td>
<td>364</td>
</tr>
<tr>
<td>WAS &lt;-&gt; AS400UAT</td>
<td>0.019</td>
<td>8</td>
<td>9</td>
<td>1,794</td>
<td>224</td>
</tr>
</tbody>
</table>

Baseline with measured 1/2ms latency between client and web server

Typical <site> location with projected 33ms latency between client and web server
The Case for APM - APM use cases should focus on cloud suitability and workload readiness, and workload life-cycle management

<table>
<thead>
<tr>
<th>Service</th>
<th>Consumer</th>
<th>Provider</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload Suitability</td>
<td>X</td>
<td>X</td>
<td>Assessment of cloud suitability based on workload performance and response time requirements</td>
</tr>
<tr>
<td>End-to-End Application Performance Testing &amp;</td>
<td>X</td>
<td>X</td>
<td>Application Load testing with emphasis on end-to-end performance optimization; provide code-level visibility into components such as Java/.NET applications, web servers, middleware, portals, and commercial application components - ensure performance and response time objectives will be met</td>
</tr>
<tr>
<td>Scalability Testing</td>
<td>X</td>
<td>X</td>
<td>Identifies resource constraints and bottlenecks – determine and ensure application scalability requirements will be met</td>
</tr>
<tr>
<td>Application Response time SLA establishment,</td>
<td>X</td>
<td>X</td>
<td>Includes establishment, end user experience monitoring, and end-to-end transaction tracking – ensures SLA objectives are quantifiable and attainable, and are being met</td>
</tr>
<tr>
<td>reporting; may include usage tracking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Troubleshooting: Problem Determination /</td>
<td>X</td>
<td>X</td>
<td>combines end user experience monitoring with end-to-end transaction tracking and big data analytics; provides code-level visibility into components such as Java/.NET applications, web servers, middleware, portals, and commercial application components - potential to substantially improve Mean Time To Resolution (MTTR)</td>
</tr>
<tr>
<td>Problem Source Identification (PD/PSI)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APM use cases:

**Scenario - Cloud Migration**
- Migrating PRODUCTION workload(s) to cloud
- Concerned about whether performance and response time requirements will be achievable when the workload(s) is moved to the cloud
- Implementing a hybrid cloud deployment and unaware of and/or not sure which cloud to on-premise dependencies / connections are of most concern
- Concerned whether response time will meet required service levels and how to measure and report in cloud environment
- Systems of engagement will be WAN post-migration (cloud) versus LAN pre-migration (in-house)

**Scenario - New cloud application deployment**
- Development / Test environment
- Concerned if response time requirements will be achievable
- Assist with application development and pre-production performance optimization, load and scalability testing
- Need to establish ‘business aligned’ SLA’s that are obtainable, and can be measured and reported back to the business / client
APM use cases:

**Scenario - ‘Release Preparedness’ Existing cloud application**
- Development / Test environment
- Isolate and resolve performance problems during development and pre-production testing
- Streamline / optimize development and pre-production testing process to help with aggressive release schedules
- Compare new release performance and impact of application change

**Scenario - Incident Management**
- Production environment
- Assist with isolating and resolving performance problem(s) for cloud based application(s)

**CSCC whitepaper: Migrating Applications to the Cloud: Assessing Performance and Response Time Requirements**