August 2011

Elastic caching for scalability, dynamic growth and performance

Highlights:

- Improved system response time and enhanced throughput
- High availability and reliability
- Lower cost by reducing CPU load on back-end systems and reducing application memory consumption
- Ability to economically and efficiently scale your IT system as opportunities grow

Executive summary

As the world becomes more instrumented, interconnected and intelligent, Internet-based activities, online transactions and data volumes increase. Further, these increasing amounts of data, along with rising consumer expectations and the need to maintain a competitive edge require fast and reliable performance. Elastic caching is the answer.

Caching is the storing of data, closer to the application code to minimize response time and minimize redundant requests. Elastic caching is an in-memory data grid that moves the cache out of the application memory space into a fault-tolerant, highly scalable data grid. Elastic caching is a foundational technology that is highly reliable, can dramatically improve response times and enable enterprises to scale to more effectively serve a smarter planet.

Elastic caching can help minimize redundant transactions and improving response time. For example, one IBM customer observed a dramatic drop in response time—from more than 100 milliseconds when accessing the back-end systems—to less than one millisecond with elastic caching.

Traditional systems are expensive and complicated to scale up to meet increasing demands. By including elastic caching in your system architecture, you can more effectively and economically scale your IT systems to meet your growing business needs. IBM has invested in extending its elastic caching capabilities across the IBM software product line and integrating it into customer's solutions.

This paper discusses how elastic caching significantly improves our customer's Connectivity, Commerce and Portal solutions. It describes how the technology is used and the benefits it delivers to customers. Benefits include reduced memory requirements, better scalability, greater fault tolerance and improved system response time—leading to greater cost savings.

Introduction

No one likes to wait—especially potential customers. Customers (and your reputation) can be "lost in a second." Why is the performance benefit provided by techniques, such as elastic caching, so important?

Consider the results from an Aberdeen Group study on the effects of a one-second delay in website response time. (See Figure1.) Web commerce sites from a variety of industries were compared to their competitors' web sites. The study showed that one second slower response time resulted in:

- 11 percent less page views—meaning less opportunity for purchase.
- 7 percent less conversions of customers—meaning 7 percent of lost sales and lost revenue. These customers could have gone to a competitor's site!
- 16 percent lower customer satisfaction ratings. Dissatisfaction often results in more calls (usually, complaints) to customer service, meaning increased costs and can hurt the chances for a repeat customer.

Clearly, lower revenue and higher cost is not a sustainable business model. Moreover, customers' dissatisfaction with your web site also reflects poorly on your company's image, potentially impacting other customers. This study clearly makes the case for reduced response time for business sustainability. You can achieve that goal with elastic caching as part of the foundation for your IT systems.

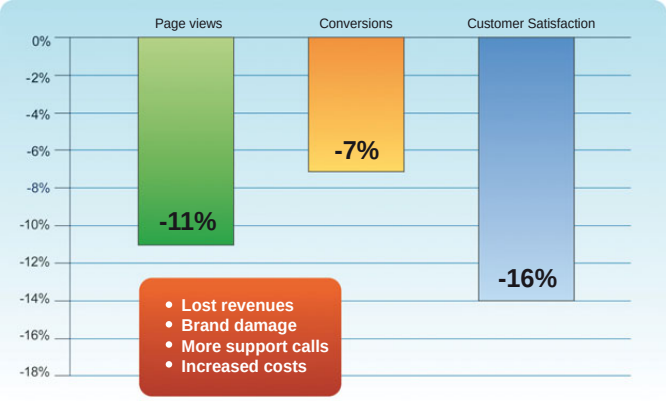

Figure 1: This graph shows the results of a study of web application performance and illustrates the extent to which Internet response time delays can negatively impact revenue and user satisfaction. "*The Performance of Web Applications: Customers Are Won or Lost in One Second*," Bojan Simic, Aberdeen Group, November 2008

Elastic Caching Overview

Caching is the storing of data, closer to the application code to minimize response time and minimize redundant requests. A traditional application cache occupies the same addressable memory space as the application. Therefore, there is a practical limit to the size of a traditional application cache.

If the cache occupies too much memory, it can actually degrade the performance of the application. When the cache is contained within the application, and if several applications are configured in a cluster, then each one contains a cache and these caches will eventually all contain the **same copies** of the cached application data.

The copies of the data are determined to be current (or out of date) based on communication of invalidation messages. The greater the number of application instances, the greater the amount of invalidation "chatter" (i.e., messages back-and-forth between applications) required to keep the cached data current among the application instances. Additionally, high-speed disks or databases (or both) are typically required to increase the size of the cache and to support high availability.

Elastic caching is an in-memory data grid that moves the cache out of the application memory space into a fault-tolerant, highly scalable data grid. An elastic data grid has the ability to expand and contract the size of the data grid based on the dynamic demand for capacity. You can easily expand the data grid, on demand, to increase or decrease capacity, for example, to remove servers from the system for maintenance. Both actions occur dynamically—while the elastic data grid is still running and handling requests.

The introduction of an elastic data grid into your application infrastructure can dramatically reduce the memory "footprint" required for each of the application instances. In a virtualized environment, this memory could be used to support additional virtualized servers, thus improving the utilization of the physical hardware. Elastic data grids provide a single, comprehensive cache shared by all of the application instances in a cluster. The data is always current because all application instances use the same, single copy in the cache. This removes **all** of the invalidation chatter in traditional clustered application architecture and can result in higher transactional performance.

Elastic data grids scale in a linear fashion, so there is virtually **no limit** to cache size. Elastic caching improves both performance and return on investment (ROI), and is a foundational element for elastic, scalable transaction processing. IBM has invested in extending the value of elastic caching to our customers by integrating it across our software portfolio and delivering turnkey solutions.

IBM offers two elastic caching options—IBM WebSphere® eXtreme Scale and the IBM WebSphere DataPower® XC10 Appliance. WebSphere eXtreme Scale provides the ultimate flexibility across a broad range of caching scenarios. The WebSphere DataPower XC10 appliance is built for simple, drop-in caching scenarios that require few application code changes. The XC10 appliance is a simple, cost-effective way to integrate elastic caching into your enterprise. WebSphere eXtreme Scale provides the ultimate flexibility across a broad range of caching scenarios. Both solutions provide linear scalability, high availability through data replications and simplified management and monitoring.

Benefits of elastic caching

Three-tier application infrastructure topologies are common today. Traditionally, to scale them, you would have to scale at all tiers. This means proliferation of the web server and application server tiers and increased complexity or manual work that would be needed to manage and maintain them. Scaling the back-end systems can be costly.

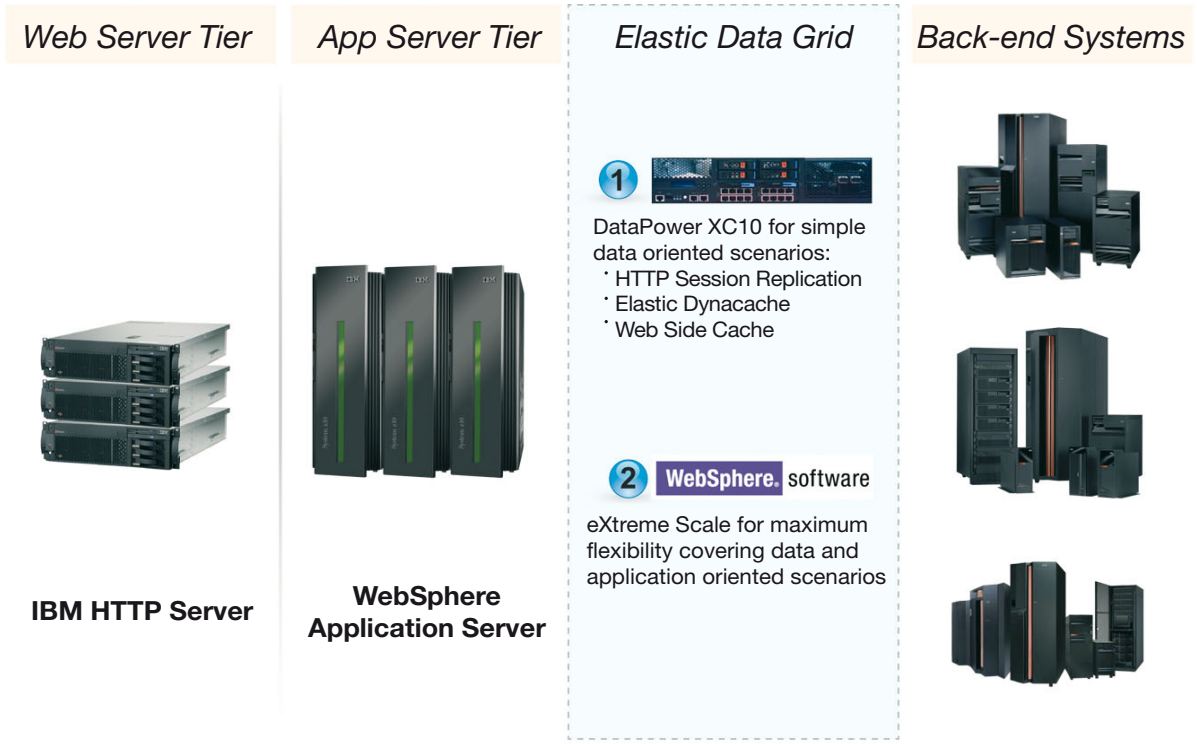

Figure 2: Elastic data grids based on WebSphere eXtreme Scale or WebSphere DataPower XC10 Appliance provide a simpler, more cost effective way to address the scaling needs of your applications.

By adding an elastic data grid into your architecture, you can very quickly and easily scale out, increase your cache capacity and your data transaction volumes with minimally invasive changes to your applications and architecture. You also drastically reduce the number of redundant transactions to the back-end systems, reducing those time- and resource-intensive calls that create bottlenecks in the traditional approach.

Elastic caching results in high availability, scalability and consistent response time as the transaction load increases. More specifically, the IT benefits of elastic caching include:

- Lowered CPU load for back-end systems by eliminating redundant transactions resulting in higher transaction throughput, and minimizing the need to scale costly back-end systems.
- Improved performance, quality and serviceability that have demonstrated up to triple the performance.
- Decreased application memory consumption by consolidating caching in one grid resulting in more efficient use of application memory.
- Improved system and transaction response times and enhanced throughput, as redundant calls are cached for rapid access.
- Reduced costs due to decreased CPU load on back-end systems and reduced application memory consumption.

These IT benefits, in turn, translate to reduced costs associated with scaling your application environment as well as a positive impact on your revenue, brand image and business operation costs outside of your IT environment.

Elastic Caching Scenarios

The WebSphere application server, along with other WebSphere family products, supports both dynamic cache-based optimizations and HTTP session persistence for performance enhancement, scalability and high availability. In particular, the WebSphere eXtreme Scale and the WebSphere DataPower XC10 appliance, when integrated with WebSphere application server, further enhance quality of service by offloading the application server cache memory requirements and disk usage for dynamic cache and HTTP session persistence.

A common use case for caching solutions is the **side cache**. The application is aware of both the cache and the back-end database or service. For each request, the application first checks the cache to see if the object is located in the cache. A cache "miss" occurs when the data is not there and the application must make a request to the back-end system. The application then inserts the data into the cache. The next time the same object is requested, it can be retrieved from the cache resulting in a cache "hit." The result is faster response time and reduced demand on the back-end system. Using WebSphere eXtreme Scale APIs, it is relatively simple to modify an existing application to use an elastic cache as a side cache.

When your database becomes the bottleneck in your business, the best solution is to add an elastic cache in front of the database. In this case, the applications now read and write to the elastic data grid. The elastic data grid is used as a data access layer and performs the reads and writes to the database as needed through a component called a **loader**. The elastic data grid acts as a "shock absorber" for the database. It can easily handle large fluctuations in transactional volume,

without overloading the database. In this scenario, even if the database is down for a period of time, applications can still read and write to the elastic data grid. When the database comes back online, subsequent changes to the elastic data grid will then be written to the database.

In a third elastic caching scenario, the elastic data grid is the system of record. In this case, all the applications read and write to the elastic data grid. This scenario is a natural extension of the data access layer scenario. However, in this case, a database is used simply as back-up storage for the elastic data grid.

Extending the Value of Elastic Caching

As previously mentioned, IBM has invested in extending elastic caching across the IBM software product line and delivering turnkey solutions to our customers. In the following sections, we will discuss three, major elastic caching solutions:

- Elastic caching for Connectivity
- Elastic caching for Commerce
- Elastic caching for Portal

Elastic caching for Connectivity

The side cache is a straightforward way to integrate caching with an enterprise service bus (ESB). An ESB is a critical component of a service-oriented architecture (SOA). The ESB connects and integrates applications, services and business process flows at the messaging layer. It performs protocol mediation, message transformation, routing, process choreography, and provides quality of service (i.e., security, reliable message delivery and transaction management).

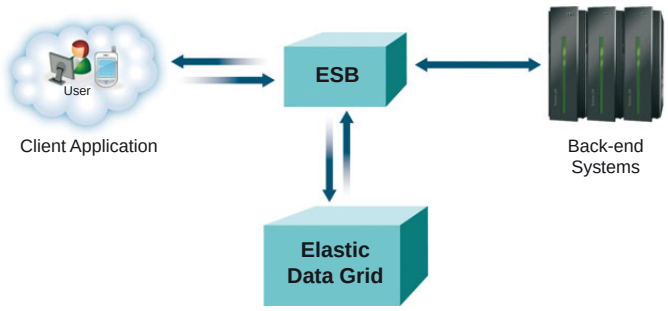

Figure 3: This diagram shows the elastic cache as a side cache for an Enterprise Service Bus.

In a SOA, application requests pass through the ESB **before** they are sent to the application. Therefore, if the result of an application request is retrieved from the elastic caching tier, the application processing and processing latency for that request are eliminated. The result is a significant decrease in response time and reduction of back-end application processing. In this case, the side cache operation is added into the ESB flow. Therefore, no changes to the application code are required.

Here are three scenarios where IBM WebSphere elastic caching solutions integrate with IBM ESB products to improve response time and increase total system throughput.

WebSphere DataPower XI50 Integration Appliance with WebSphere DataPower XC10
The WebSphere DataPower XI50 Integration Appliance is a secure, easy-to-deploy, hardware ESB. The current release of the DataPower XC10 firmware includes a Representational State Transfer (REST) gateway, allowing non-Java-based clients to access simple data grids using a set of HTTP-based operations. Using the REST gateway feature, the XC10 can be used as a side cache for the WebSphere DataPower XI50.

To use the XC10 as a side cache, an eXtensible Markup Language (XML) proxy is defined as the first component in the XI50 processing chain. It uses a set of "caching policy" rules to determine whether or not an incoming request can be cached. The rules are application-specific. In general, caching policy rules might trigger on the request Uniform Resource Identifier (URI), specific XML contents within the request body, or a combination of both. The rules are defined using a set of Extensible Stylesheet Languages (XSLs). The XSLs are then loaded into the XI50 memory. Additionally, this set of XSLs is used to generate the appropriately formatted REST requests to the XC10 REST gateway to store/retrieve data in the grid.

WebSphere Process Sever and WebSphere Enterprise Service Bus (ESB) with WebSphere Extreme Scale
WebSphere eXtreme Scale (as of v7.1 cumulative fix 1) includes two mediation primitives that allow you to insert and retrieve data from the cache with WebSphere Process Server and WebSphere Enterprise Service Bus. These products integrate with various back-end systems and WebSphere eXtreme Scale may be added to the configuration to cache the output of these systems, increasing the overall performance of your configuration.

You can integrate WebSphere eXtreme Scale into your configuration without changing the business process itself by using the mediation flows that are provided by the WebSphere Enterprise Service Bus. Read-only service requests can extract results from caches and be configured to load caches on misses. Due to the nature of mediations, this solution is both service- and binding-agnostic.

WebSphere Message Broker with WebSphere eXtreme Scale
In this scenario, a JavaCompute node is added to the message flow to check the cache for data. The code defines classes to represent the objects for caching and uses the eXtreme Scale ObjectGrid application programming interfaces (APIs) to interact with the cache.

In all three of these ESB scenarios, although there may be some coding involved such as a custom node or the addition of mediation into the messaging flow of the ESB, there are **no application code changes**. This simplifies the use of the WebSphere eXtreme Scale or WebSphere DataPower XC10 appliance as the caching solution for your ESB. Using these products as the side cache improves response times and enhances throughput, as redundant calls are cached for rapid access.

Example of elastic caching for Connectivity
One of the largest marketers of tires for the automotive replacement market in the United States was experiencing long delays in customer response time on their web site. The slow response was due to redundant data found in multiple data sources. The tire marketer has begun to implement a flexible, enterprise

connectivity infrastructure for integrating applications and services, built on a robust, platform-independent ESB. It allows the development of an enhanced SOA. The goal is to integrate existing and future business applications and provide a common framework for integrating and synchronizing distributed systems that would otherwise be incompatible. It must provide a robust, scalable caching mechanism.

Following a proof-of-concept trial, its decision to move the system to production was based on three factors:

- Implementation and integration of the ESB and the caching components was relatively fast and easy. The integration between WebSphere Message Broker and WebSphere eXtreme Scale was straightforward because of the Java support built into the WebSphere Message Broker.
- The response time observed by users was approximately 100 times better with cache hits. The response time observed when the data was in the cache was less than **one** millisecond, compared to more than **100** milliseconds when accessing the back-end systems.
- The simplified integration strategy provided by the ESB, versus a traditional, point-to-point approach, helped improve the productivity of the integration development team, reducing integration time and effort.

Overall, the client expects a 400 percent improvement in response time—especially for those objects included in the data grid.

Elastic caching for Commerce

Elastic caching allows your commerce systems to grow while improving customer satisfaction, due to improved response times, and improved user experiences. It also supports the rapid adoption of new business processes, such as the ability to change prices based on customer demand and new promotions dynamically, without restarting or resetting the current environment.

WebSphere Commerce is an industry-leading solution for web retail applications. It is a J2EE application deployed in the WebSphere Application Server. One of the key performance-related features of the WebSphere Application Server is DynaCache. DynaCache is a technique that helps improve performance by caching dynamically created data that contains the output results from the runtime program execution of code components, such as servlets and JavaServer Pages (JSPs). WebSphere Commerce sites use DynaCache to reduce database round trips, providing an important performance boost.

When planning for high volume sites, planners must consider the impact of invalidation traffic on network bandwidth and plan accordingly. WebSphere eXtreme Scale and the DataPower XC10 caching appliance are an elastic data grid technology. WebSphere Commerce can be configured to use either eXtreme Scale or the XC10 appliance as a DynaCache provider. The elastic data grid topology has a single, logical instance of the cache that is shared among WebSphere Commerce servers. Since this cache is shared across servers, multiple copies of the same pages and fragments are not needed for each. Instead, a single cache instance is created on the first request for that page or fragment, and is then available to all WebSphere Commerce servers sharing the cache.

What is cached in commerce applications?

- JSPs and web page fragments are stored in the servlet cache instance (baseCache).
- Commands are stored in the servlet cache instance (baseCache).
- Distributed maps are also stored in object cache instances.

Benefits of WebSphere eXtreme Scale and the DataPower XC10 Appliance

The benefits of deploying WebSphere eXtreme Scale or DataPower XC10 caching appliance, compared with DynaCache and disk offloading, include:

- Reduces the average response time by as much as 25 percent.
- Provides a more consistent user experience, due to less statistical variation in the response time.
- Improves the time needed to reach steady-state after full or partial site restart, or after full cache invalidation by as much as 40 percent. Since all WebSphere Commerce servers use the same grid, there is no "warm-up" time for the cache when additional WebSphere Commerce servers are added, or when servers are stopped and restarted.
- Simplifies tuning and operational maintenance.
- Eliminates the need for high-speed disk off-loading.
- Provides a consistent cache because the same version of the page is always shown as each client Java Virtual Machine (JVM) uses the same eXtreme Scale grid, rather than a separate cache for each. This adds additional capacity to the grid—seamlessly.

Example of elastic caching for Commerce

A United States retailer improved customer satisfaction with its key brands' web sites when it added IBM WebSphere eXtreme Scale software to its IBM WebSphere Commerce suite installations, speeding the site's response time and enhancing overall site performance. This retailer is committed to providing an exceptional online shopping experience. The last thing it wants to see on any of its retail sites is slow—or perceived as slow—response time, which can result in lost business. As an IBM WebSphere Commerce user, this customer is able to add new online services with relative ease. However, because online transaction volume has grown, content delivery speed and site availability have been affected.

Continuous web services growth has required the frequent addition of new JVM to sustain system scalability. However, too many decentralized JVMs, each running its own caching, resulted in performance challenges. The practice of regularly adding new JVMs to address volume/scalability issues was no longer serving the company well, with overall system maintenance also becoming an issue.

The solution was to use IBM WebSphere eXtreme Scale as a centralized memory grid to sustain scalability i.e., the ability to easily absorb growth without exponential increase in application infrastructure (commerce, in this case), and decrease the risk of outages. The customer is also realizing significant benefits to JVM startup cycles, since data does not need to be repopulated as it is instantly available in the grid.

Following a successful proof of concept trial, the customer moved to production in time for holiday shopping (aka "Black Friday" and "Cyber Monday") with significantly improved user experience due to reduced, and more consistent, response-times. The use of WebSphere eXtreme Scale software is expected to result in lower cost and increased revenue. It has allowed the customer to focus on the adoption of new business models, enabled by WebSphere eXtreme Scale. It also supports the ability to change prices based on customer demand and new promotions dynamically, without restarting or resetting the current environment.

Elastic caching for WebSphere Portal

Elastic caching allows your portal systems to support more users and, at the same time, maintain those user sessions through a server failure. The IBM WebSphere Portal server enables highly customized user experiences, in part, through the extensive use of session data. However, that session data consumes large amounts of memory on the application server which limits how many active users a given deployment can handle. Increasing the number of users requires increasing the number of portal servers—even if the portal servers are not overloaded from a CPU perspective. Also, increasing the number of portal servers increases the likelihood of a server failure, resulting in lost sessions for users on that server.

WebSphere eXtreme Scale and the DataPower XC10 caching appliance can each address both of these problems by offloading session data to an elastic data grid (cache) running on a separate set of machines which can be scaled up independently to address the volume of users. When the WebSphere Portal server is configured to store sessions in an elastic data grid, the portal server maintains a cache of active user sessions. Sessions that do not fit in this cache are stored in the data grid. Therefore, they do not consume resources on the portal server. This allows the portal server to manage an increased number of active users without requiring additional memory per server or an increased number of portal servers.

Further, active user sessions are not lost in the event of a portal server failure because all user sessions are stored in the highly available elastic data grid. If a portal server fails, and the user is directed to a different portal server (which will not contain their session data), the user's session will be automatically retrieved from the grid and cached in the new portal server.

As a result, using an elastic data grid to maintain session persistence has two key benefits: higher scalability (more users can be supported) and stronger failover support by maintaining user sessions through a server failure.

Session persistence is not a completely "free" operation. Session data must be copied to the grid and, occasionally, retrieved from the grid. This can increase CPU utilization on the portal server. How significant this increase is depends on the size of the local session cache relative to the total population of active sessions. However, this is usually negligible compared to the benefits of the solution.

Summary

Elastic caching helps IBM customers take advantage of the opportunities presented by a smarter planet by providing a technical foundation that helps deal with the huge increases in Internet usage, transactions and data. IBM offers two elastic caching solutions:

- IBM WebSphere eXtreme Scale for ultimate flexibility across a broad range of caching scenarios
- IBM WebSphere DataPower XC10 Appliance for simple, drop-in caching scenarios requiring few application code changes

IBM has invested in integrating elastic caching across the IBM software portfolio and delivering turnkey solutions that are providing tremendous benefit to our customers:

- Improved system response time and enhanced throughput
- High availability and reliability
- Lower cost by reducing CPU load on back-end systems and reducing application memory consumption
- Ability to economically and efficiently scale your IT system as opportunities grow

Explore the advantages of the IBM approach
Create competitive success with an approach to business agility that simplifies your environment and minimizes response time for your customers. The tools exist today.

Appendix A: Additional resources

IBM Global Asset Recovery Services help address environmental concerns with new, more energy-efficient solutions. To learn more, visit: **ibm.com**/financing/us/recovery/

Here is a list of resources to help you learn more about elastic caching:

For a fully functional J2SE trial download, visit: **ibm.com**/developerworks/downloads/ws/wsdg/learn.html

IBM eXtreme Scale web site: **ibm.com**/software/webservers/appserv/extremescale/#ibm-content

IBM WebSphere XC10 web site: **ibm.com**/software/webservers/appserv/xc10/

IBM WebSphere Elastic Caching – IBM WebSphere DataPower XC10 Appliance data sheet: **public.dhe.ibm.com**/common/ssi/ecm/en/wsd14088usen/WSD14088USEN.PDF and **public.dhe.ibm.com**/common/ssi/pm/sp/n/wsd14088usen/WSD14088USEN.PDF

Scalable, Integrated Solutions for Elastic Caching Using IBM WebSphere eXtreme Scale Redbook: **redbooks.ibm.com**/abstracts/sg247926.html?Open

IBM WebSphere eXtreme Scale and DataPower XC10 Appliance Wiki: **ibm.com**/developerworks/wikis/display/objectgrid/Getting+started

For more information

To learn more about IBM WebSphere eXtreme Scale and the WebSphere DataPower XC10 Appliance, contact your IBM sales representative or IBM Business Partner, or visit: **ibm.com**/extremescale and **ibm.com**/XC10

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. We'll partner with credit qualified clients to customize an IT financing solution to suit your business goals, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward. For more information, visit: **ibm.com**/financing

IBM Software
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
August 2011

Please Recycle

WSW14170-USEN-00