Elastic Scalability in the Application (XC10 Caching Appliance)

Charles Le Vay
Technical Evangelist
Smarter Planet Solutions Require a Dynamic Application Infrastructure

- Scale quickly and efficiently
- Optimize workload performance
- Flexibly flow resources
- Avoid downtime
- Save energy
- Automate management tasks
Dynamic Application Infrastructure
Builds on Smart SOA

**Business Needs**

- “Meet business objectives consistently, nimbly, cost-effectively”
- “Enable applications to adapt to changing market conditions”
- “Address extreme demands of clients & business models”

**Adoption Patterns**

- Application Foundation
- Intelligent Management
- Extreme Transaction Processing
“Internet Scale” is tightly interwoven with IBM’s vision for a Smarter Planet

Our World is Becoming:

- instrumented
- interconnected
- intelligent

1 billion people on the Internet

THINK

3 Billion Stock Trades

15 Petabytes of new information… per day!
Smarter Planet Workloads Drive the Need for Elastic Caching

“1 Billion People on the Internet”
Largest fantasy sports site in world

User Profile Management
7 billion requests per day
10x reduced response times

“3 Billion Stock Trades”
3rd largest mutual-fund company in the US

Order Matching / Routing
40x increase in number of transactions
4x increase in revenue

“15 Petabytes of new information per day”
World’s 6th largest property & casualty insurance company

Accelerate Business Processes
125,000 policies concurrently processed
Innovative Elastic Caching Solutions

**Data Power XC10 Appliance**
- Drop-in cache solution optimized and hardened for data oriented scenarios
- High density, low footprint improves datacenter efficiency

**“Data Oriented”**
- Session management
- Elastic DynaCache
- Web side cache
- Petabyte analytics
- Data buffer
- Event Processing
- Worldwide cache
- In-memory OLTP
- In-memory SOA

**“Application Oriented”**

**eXtreme Scale**
- Ultimate flexibility across a broad range of caching scenarios
- In-memory capabilities for application oriented scenarios

**Elastic caching for linear scalability**
- High availability data replication
- Simplified management, monitoring and administration
Caching concepts

Data cache

Can be used to hold data for a common set of clients

Data grid

Combined set of data caches which holds a collection of data from one or more client

IBM WebSphere DataPower XC10

Physical hardware device capable of running one or more grids

Collective

Set of IBM WebSphere DataPower XC10 appliances grouped together for scalability and management purposes
Modern Application Infrastructure Topology

Web Server Tier

App Server Tier

Elastic Data Grid

Database Tier

IBM HTTP Server

WebSphere Application Server

DataPower XC10 for simple data oriented scenarios:
- HTTP Session Replication
- Elastic Dynacache
- Web Side Cache

eXtreme Scale for maximum flexibility covering data and application oriented scenarios

DB2 UDB
IBM WebSphere DataPower XC10 Appliance

- **Scale out with ease**
  - *Large, elastic cache* allows you to scale more economically while providing high Quality of Service

- **Easy drop in use for common scenarios**
  - Support for *data-oriented caching scenarios* without rip & replace

- **Fault tolerance**
  - Lower risk of data loss while providing *continuous availability*

- **Flexible and simple user management**
  - Simple solution for *real world management and monitoring*
Scale out with Ease

- **160 GB** elastic cache for your business-critical applications
- Scales **elastically** without application downtime
- Linear, **predictable** scaling at predictable cost
- Quickly and easily increase cache capacity as needs grow
- Unbinds cache from application server memory constraints
Easy drop-in use for common scenarios

- Simple Cache
- HTTP Session Cache
- WAS Dynamic Cache Service ("Dynacache") support

Little or no code changes required!
Overview of data caches

- WebSphere DataPower XC10 extends its caching capabilities to these cache types:
  - Simple Cache
    - Basic create, read, update, and delete
  - Session Cache
    - Legacy session management caching mechanism in WebSphere Application Server
  - Dynamic Cache
    - Legacy caching mechanism in WebSphere Application Server using the Dynamic Cache API

- WebSphere DataPower XC10 provides client code and any necessary plug-ins for:
  - WebSphere Application Server to work with these three data cache types on the appliance
  - Java™ applications outside of WebSphere Application Server to work with simple data caches on the appliance

- WebSphere DataPower XC10 supports non-Java™ applications outside of WebSphere Application Server to work with simple data caches on the appliance using REST APIs
REST Gateway

- Enables non-Java™ based clients access to simple data grids using a set of HTTP based operations. (e.g. .NET, php)

- Simple HTTP methods
  - HTTP POST method to insert/update data in the grid
  - HTTP GET method to get data from the grid
  - HTTP DELTE method to delete data from the grid.

- Supports creation of dynamic maps with 3 Time to Live (TTL) templates
  - None (no time to live expiration)
  - Last Access Time
  - Last Update Time

- non-Java™ clients and Java™ clients can access the same data grids
Simple Cache

- Basic create, retrieve, update, and delete capability
- You provide an application that uses WebSphere eXtreme Scale API to manage your simple data cache on WebSphere DataPower XC10
- Application can be running within WebSphere Application Server or as a stand-alone Java application
  - Client code can be installed outside of WebSphere Application Server to enable stand-alone Java application access to simple data caches on WebSphere DataPower XC10
- WebSphere DataPower XC10 REST Gateway enables non-Java™ applications access to simple data caches on the appliance thru REST APIs
Simple Cache used as a Side Cache

- Used to store data for fast, lower-cost access than a database
- Uses ObjectMap APIs from WebSphere eXtreme Scale
- Every time data is needed, the side cache on the WebSphere DataPower XC10 Appliance is checked first
- If the value is not found (cache miss), then the data is retrieved from the backend system and inserted into the cache
- Client can run in a standard Java EE compliant server environment or in any Java Virtual Machine compliant with Java SE V1.4 or beyond
- Non-Java clients can use REST APIs
- Reduces load on the backend by eliminating redundant requests to backend systems improving response time and increase total system throughput
WebSphere DataPower XC10 used as a Side Cache for WebSphere Datapower XI50
HTTP Session data cache

- Extension of legacy session management caching mechanism in WebSphere Application Server
- Extensions to WebSphere Application Server administrative console to support WebSphere DataPower XC10 session management caching
- Two ways to create the cache
  - Create cache on the appliance and then point WebSphere Application Server to the cache
  - Create the cache on the appliance directly from the WebSphere Application Server administrative console
Offloaded session management for HTTP requests

- WebSphere Application Server connects seamlessly to the WebSphere DataPower XC10 appliance
  - Client code must be installed on WebSphere Application Server systems

- Easily configure WebSphere applications to store HTTP session data to a data cache on the WebSphere DataPower XC10 appliance through the WebSphere Application Sever administrative console

- Replaces other session replication mechanisms (memory-to-memory replication)

- Removes the need for Database traditionally used for persistence

- Enables HTTP session failover between WebSphere Application Server cells
Active/Active Datacenter HTTP Session Failover

Datacenter1
- Deployment Manager
- PlantsByWebSphere Cluster
- xPlantsByWebSphere Cluster
- memory to memory replication

Datacenter2
- Deployment Manager
- PlantsByWebSphere Cluster
- xPlantsByWebSphere Cluster
- memory to memory replication

Web Server
Elastic Dynamic Cache service support

- WebSphere DataPower XC10 provides client code and a plug-in for WebSphere Application Server applications to support Dynamic Cache API
- Allows applications deployed to WebSphere servers to use WebSphere DataPower XC10 as a “drop-in” cache, instead of storing cache data in local memory or multiple instances of a disk cache
- WebSphere DataPower XC10 can be used with WebSphere Commerce Suite as an alternative caching mechanism for Dynamic Cache to reduce local memory requirements
WebSphere Commerce using Dynacache

WebSphere Commerce heavily exploits for page caching

Each JVM has a private disk based cache to support caches much larger than possible with a memory only conventional cache

3 tier cache: JVM has a small local cache, then there is the file system cache and finally the disk itself.
WebSphere Commerce using WebSphere DataPower XC10

Clients can attach to the ‘cache’ using the network

No dependency on a large file system cache.

No disk dependency, no SAN required.

Cache is as large as the memory in the ‘grid’.

Each record is stored once in the grid and shared by all clients.
Fault Tolerance

- Create an appliance “collective” for high availability
- Lowers risk of data loss through automatic replication of data
- Continuous Availability!
- Self Healing - Failures are automatically detected
- Easy to configure
Collective

- A set of appliances which know about each other
- Allows for failover recovery if one of the appliances in the collective fails
- Default collective contains the appliance where you logged on
  - Add additional appliances using the console interface
- Add additional appliance by supplying:
  - Host name or IP address of the appliance to be added
  - Secret key of the appliance to be added
- Adding one appliance to collective that contains only 1 appliance:
  - Provides failover capability because 2nd appliance automatically gets replica of all data
Replication

- Collective replicates data between members
- Replication provides a backup of primary data
- Replicas act as drones
  - They are completely slaved to the primary
  - Modifications can only be made using the primary
- The replication mechanism uses a check-pointing technique to bring replicas up to date as fast as possible with minimum disruption to work occurring on the primary

- Synchronous replication
  - Data is replicated before transaction completes
  - Highly reliable
  - Can cause performance bottleneck.

- Asynchronous replication
  - Allows the transaction to complete before the data is replicated
  - Provides greater speed than synchronous replication
Adding a second appliance to a collective

- Adding one appliance to the initial collective (total of 2 in collective) provides failover capability
  - The 2nd appliance automatically gets a replica of all primary data cache data from the 1st appliance

- Partitions moved from 1st appliance to 2nd
  - Attempts to balance primaries and replicas

**Warning:** when you add an appliance to the collective, the appliance being added is stopped and restarted
  - This function takes several minutes to complete
    - Data is being replicated
    - Stop and restart of appliance takes several minutes as well
Replication using zones

- Zones are exclusive collections of appliances
- Each appliance can be tagged with a zone name and will belong to exactly one zone
- Zone rules allow placement rules for grids within zones
- Rules can specify that different grid types cannot be placed on machines within the same zone.
  - A replica of a grid cannot be placed in the same zone as its primary for example
- Zones enable advanced placement of grids across racks, floors, buildings or cities
Flexible and Simple User Management

- Native interfaces for administration and monitoring
- Relief to simple real world problems!
  - Real-time monitoring that dynamically updates with the number of entries in the grid
WebSphere Datapower XC10 Summary

- **Accelerated Time to Value**
  - Reduces the time necessary for install, setup and configuration through out-of-the-box, ‘drop-in’ use for simple side cache scenarios, HTTP Session replication and WebSphere Application Server dynamic cache service

- **Simplified management and administration**
  - Offers a built-in, simplified administration and monitoring console to enable efficient setup, configuration, and management of the appliance and transaction load within your datacenter

- Ensures high availability of data for mission-critical applications
- Scales with simplicity
- Delivers high performance and consistent response times
Learn More

For more on IBM’s XTP portfolio, visit: www.ibm.com/XTP

Download WebSphere eXtreme Scale for free and build a trial app


XC10 Web Site


Read Rob High’s Article on XTP and WebSphere eXtreme Scale


Additional resources

Weekly video podcasts covering customers questions and forum posts on the IBM WebSphere eXtreme Scale product.

http://www.youtube.com/user/ibmextremescale#p/a

WebSphere Extreme Transaction Processing for Developers Space will discuss various topics for developing and deploying XTP applications and will point out emerging trends, benefits, challenges, and features associated with it.

Learn More About Dynamic Application Infrastructure!

Application Foundation
ibm.com/appfoundation

Intelligent Management
ibm.com/intellmgmt

Extreme Transaction Processing
ibm.com/xtp

ibm.com/appinfrastructure
Questions?