Business Agility in Action.

Innovate. Transform. Grow

WebSphere Elastic Caching Solutions

Shaun Lee – WebSphere Technical Specialist

03/13/2012

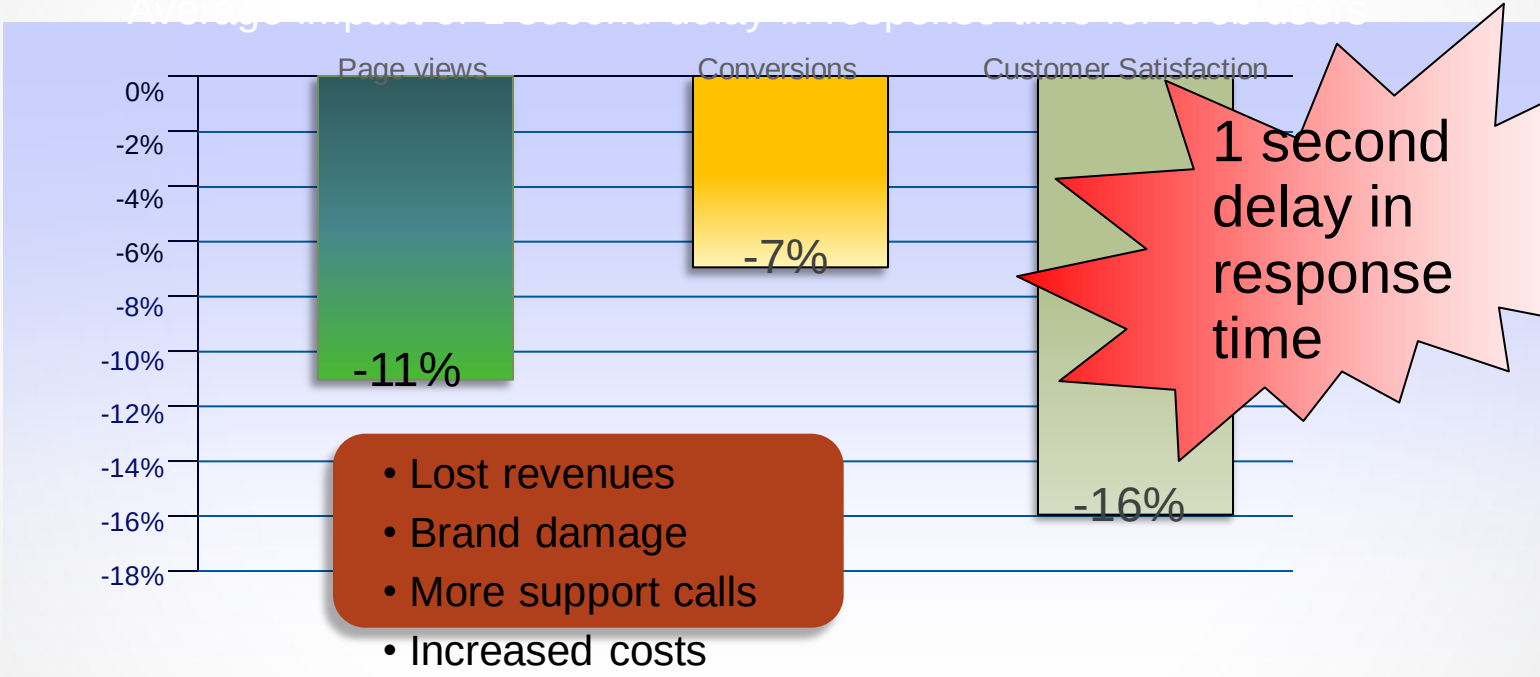
Slide Heading

- Market Drivers and Scaling Challenges
- Elastic Caching Solutions
- Caching Scenarios and Patterns
- Customer references
- Summary

Market Drivers

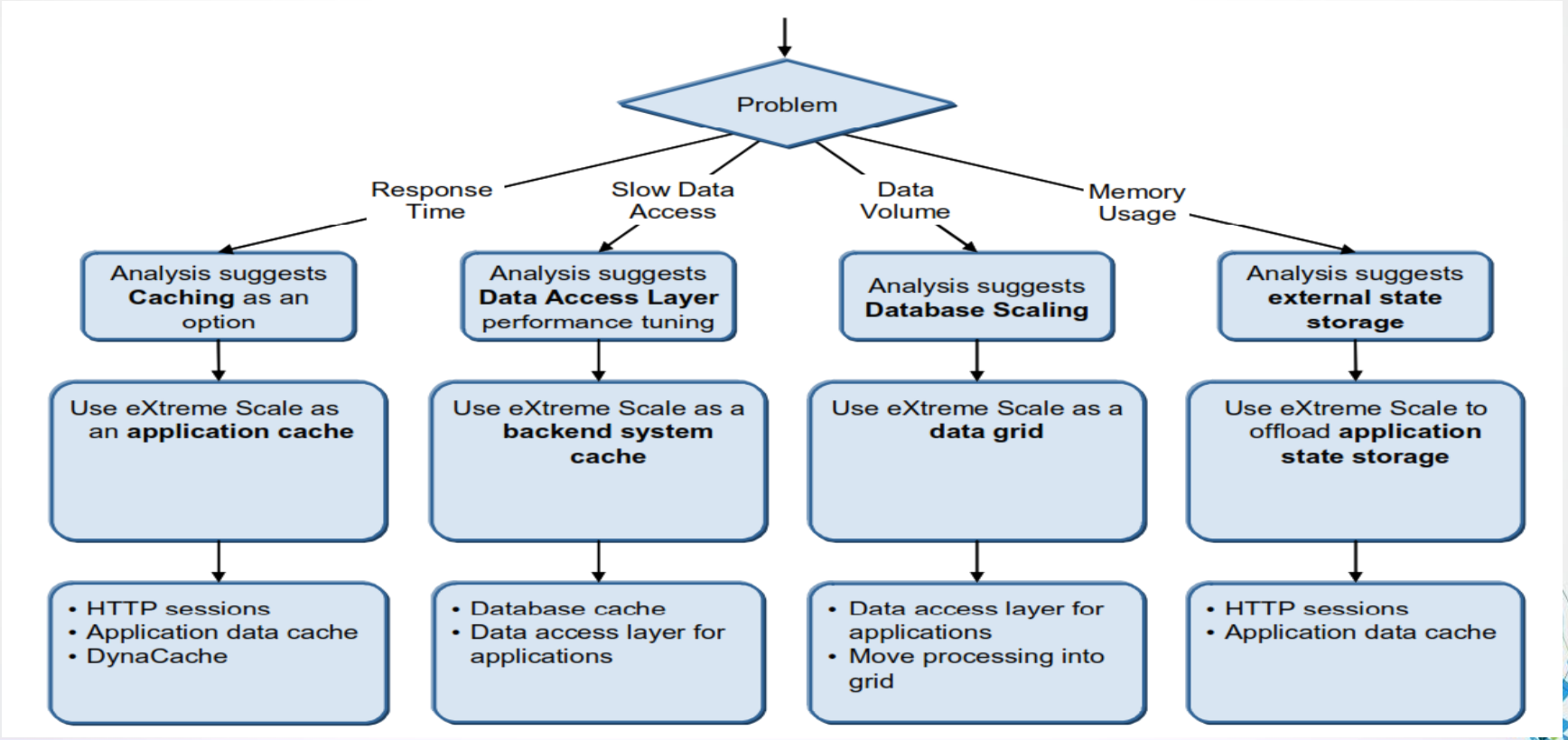
- Online access and interaction volumes on high growth trajectory
- Response times are critical to giving customers a good experience and generating revenue.
- Customer sessions are becoming more critical.
- Losing the data that they have entered will likely create a negative impression and result much higher abandonment rates
- Customers are looking to control the growth of their enterprise systems.
  - A caching tier in front of it can allow more growth without expanding the existing enterprise systems.
- Mobile is a game changer and will further increase transactions

Internet response time challenges negatively impact revenue and customer satisfactions



1. "The Performance of Web Applications: Customers Are Won or Lost in One Second," Bojan Simic, Aberdeen Group, November 2008
2. Source: Internet World Stats, Usage and Population Statistics, www.internetworldstats.com/stats.htm, December 22, 2010

When do I need an Elastic Grid?



Modern Application Infrastructure Topology

Web Server Tier

IBM HTTP Server

App Server Tier

WebSphere software

WebSphere Application Server

Elastic Data Grid

1

DataPower XC10 for simple data oriented scenarios:

- HTTP Session Replication
- Elastic Dynacache
- Web Side Cache

2 WebSphere software

eXtreme Scale for maximum flexibility covering data and application oriented scenarios

Back-end Systems Database Tier

Information Management

DB2 UDB

Innovative Elastic Caching Solutions

DataPower 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"

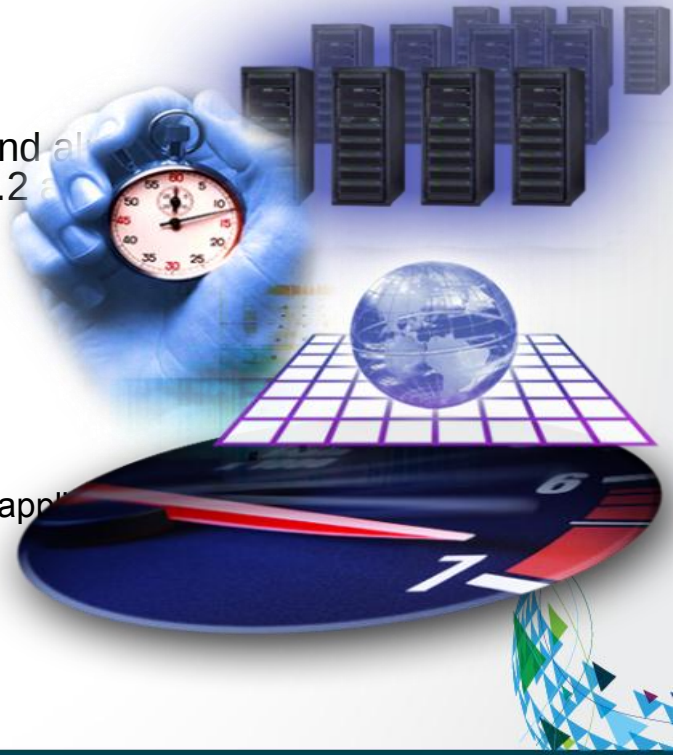
WebSphere. software

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

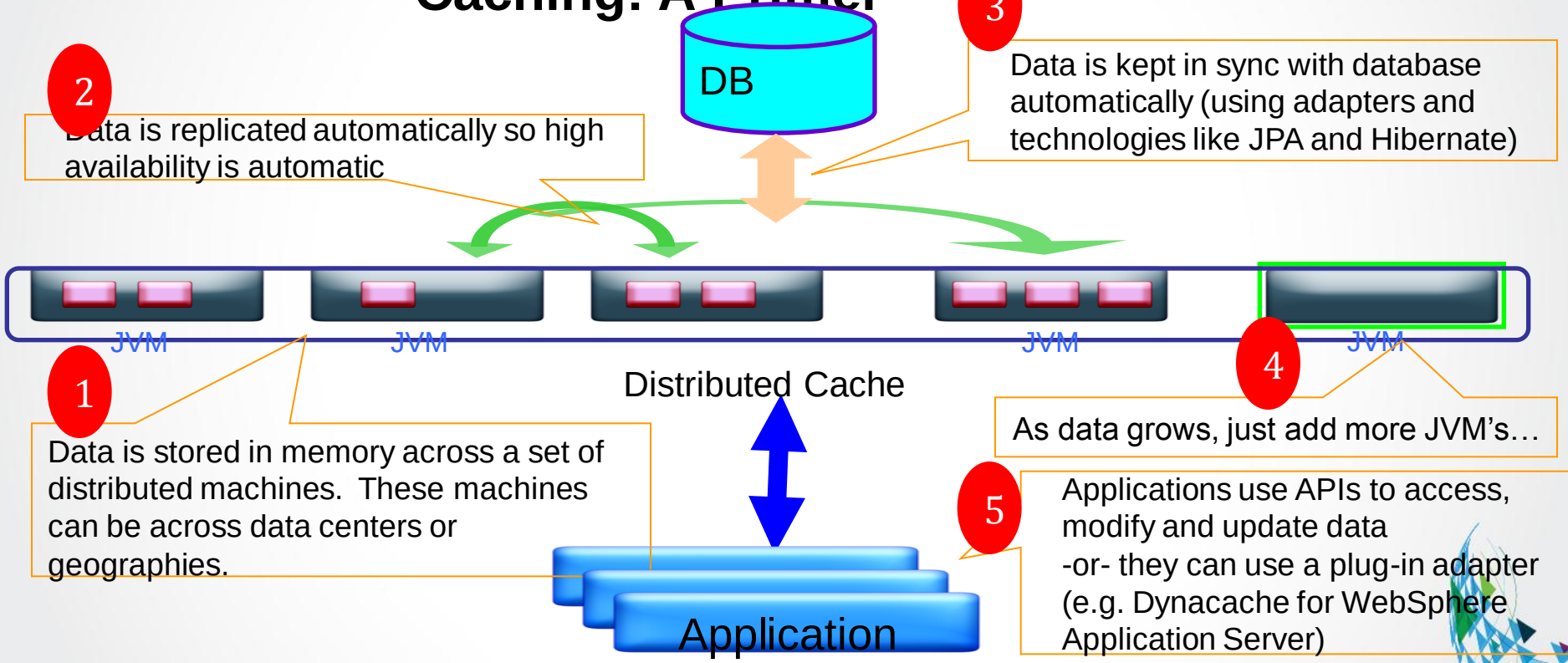
Elastic caching for linear scalability
High availability data replication
Simplified management, monitoring and administration

IBM WebSphere eXtreme Scale

- Proven mature product:
  - Fourth major release of product with V7.1 in 2010
  - Used at some of the largest web sites in the world
- Lightweight runtime footprint (15MB jar)
- Integrates with all versions of WebSphere Application Server and al
  Java-based application container or Java Virtual Machine (1.4.2
- Only requires Java SE run-time environment
  - Exploits WAS-ND environment when available
- Meets the needs of a variety of application environments:
  - Java SE, Java EE (WAS, WebLogic, Tomcat…), as well as for .Net appl
    REST APIs*
- Proven multi-data center capabilities
- Proven low-latency access to data



WebSphere eXtreme Scale and Distributed Caching: A Primer



IBM WebSphere DataPower XC10 Appliance V2

- Scale out with ease
  - **Large, 240GB elastic cache** allows you to scale more economically while providing high Quality of Service
  - Scales **elastically** without application downtime
  - Linear, **predictable** scaling at predictable cost
- Easy drop in use for common scenarios
  - Support for **data-oriented caching scenarios** without rip & replace
  - Unbinds cache from application server memory constraints
- Fault tolerance
  - Lower risk of data loss while providing **continuous availability**
- Flexible and simple user management
  - Simple solution for **real world management and monitoring**

Simple Caching Scenarios

Challenges
- Application makes redundant calls, doing something over and over again, on expensive back-end systems
- Generally, to access data that does not change much (e.g., user profiles)

→ Offload Redundant Processing

Benefits
- Free up expensive back-end systems for critical tasks
- Reduce costs of system cycles for repetitive data retrieval
- Increase performance through in-memory, network cache

Challenges
- Web sites that need better management and automatic fail-over of Web sessions – usually WebSphere Commerce, WebSphere Portal or retail-related sites

→ "Drop-In" HTTP Session Replication

Benefits
- Automatic "drop-in" IBM elastic cache without invasive coding changes
- Higher availability and performance for revenue-producing applications

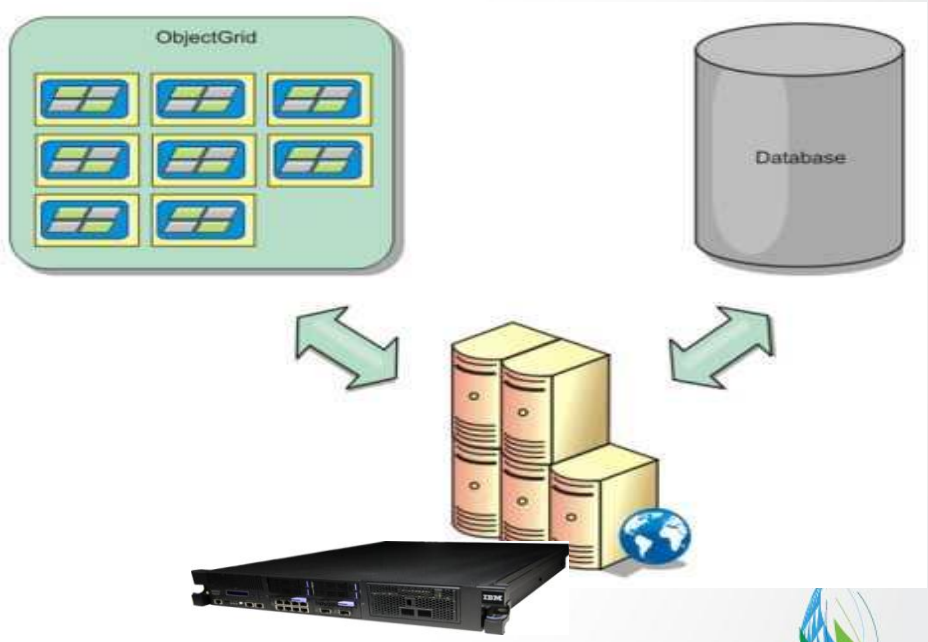
Challenges
- Web applications that use (WebSphere Application Server) DynaCache and need better performance and scalability of their caching investment

→ "Drop-In" extension for DynaCache

Benefits
- Better performance: turbo-charge WebSphere Application Server caching layer via IBM elastic cache "drop-In" cache with no coding changes

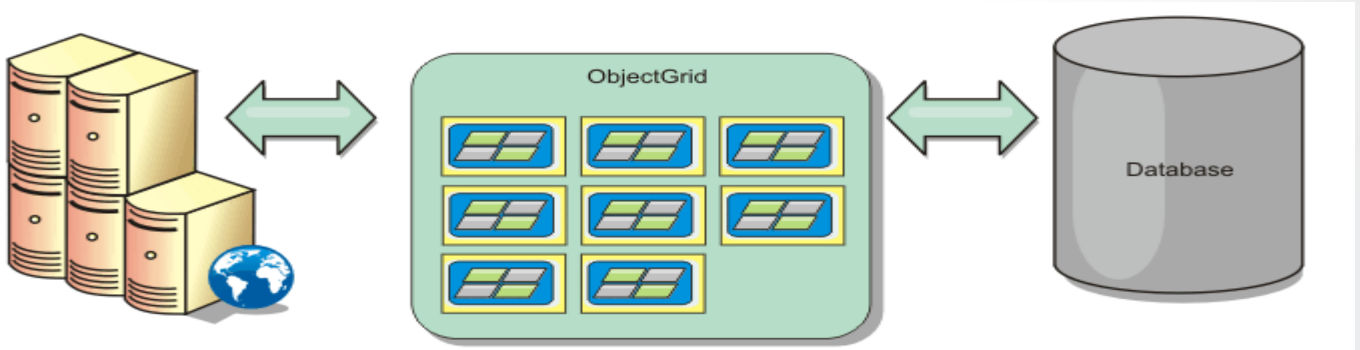
Offload Redundant Processing : Side Cache

- IBM elastic cache solution is used to temporarily store objects that would normally be retrieved from a back-end database.
- Applications check to see if the elastic cache Scale contains the desired data.
- If the data is there, the data is returned to the caller. If the data is not there, the data is retrieved from the back-end and inserted into the elastic cache so that the next request can use the cached copy.



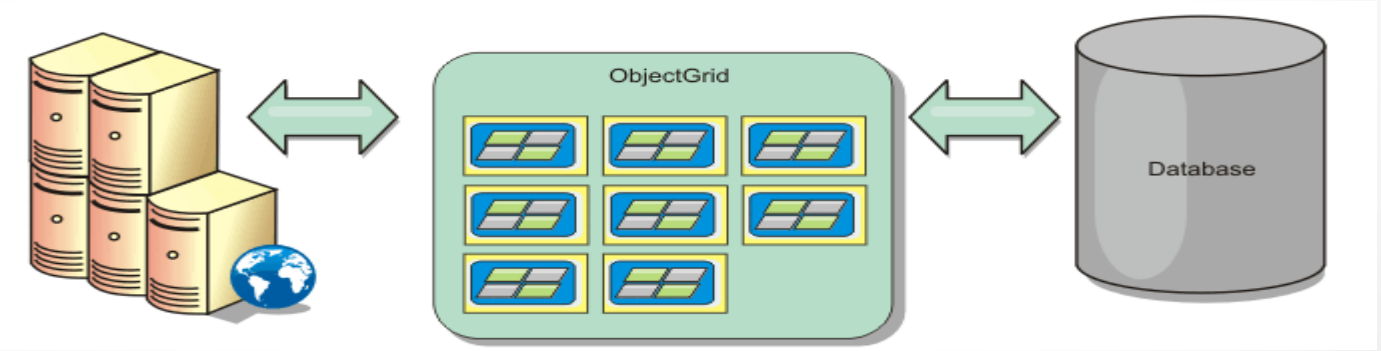
Offload Redundant Processing : In-line cache

- Applications check to see if WebSphere eXtreme Scale contains the desired data.
- If the data is there, the data is returned to the caller. If the data is not there, the data is retrieved from the back-end by WebSphere eXtreme Scale so that the next request can use the cached copy.
- Changes are written to the cache and back-end synchronously. A *write-through* cache.



Offload Redundant Processing : In-line cache with Write-Behind

- Variation of previous scenario. Changes are written to the back-end asynchronously. A *write-behind* cache.
- Back-end load is significantly reduced as there are fewer but larger transactions
- Back-end availability has no impact on application availability.



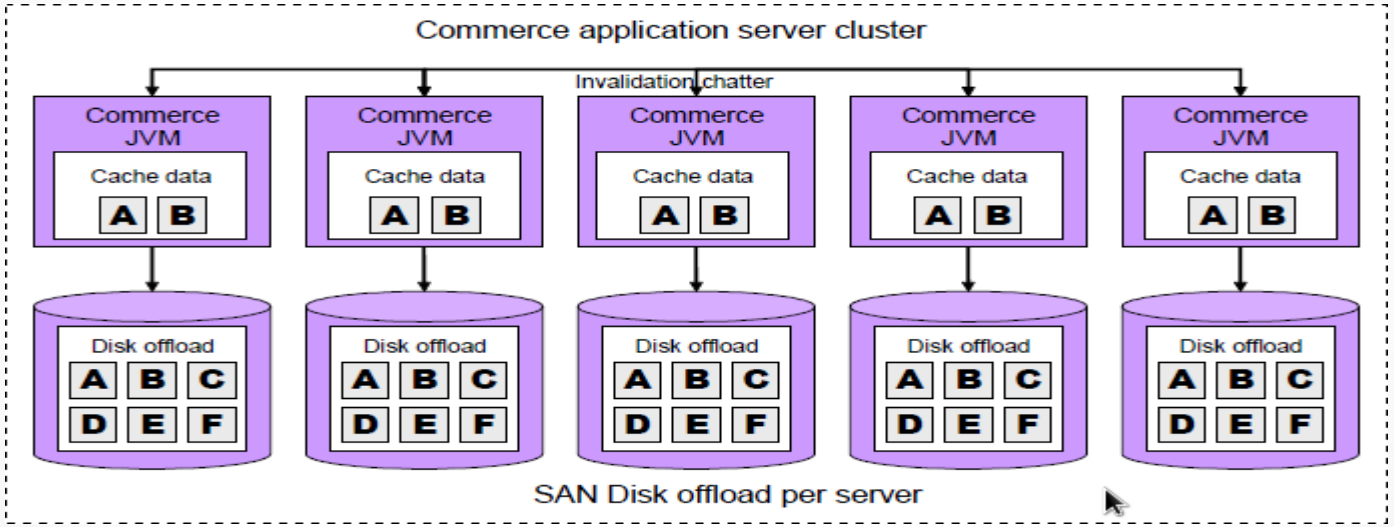
“Drop-In” HTTP Session Replication

- HTTP sessions can be replicated across servers using IBM elastic cache
  - A servlet filter that enables session replication can be inserted into any Web application
  - Provides a session persistence approach that is independent of the WebSphere cell infrastructure
- WebSphere products can use IBM elastic cache as an upgraded session persistence mechanism
- Non-WebSphere servers (such as Geronimo or JBoss) can also use this servlet filter

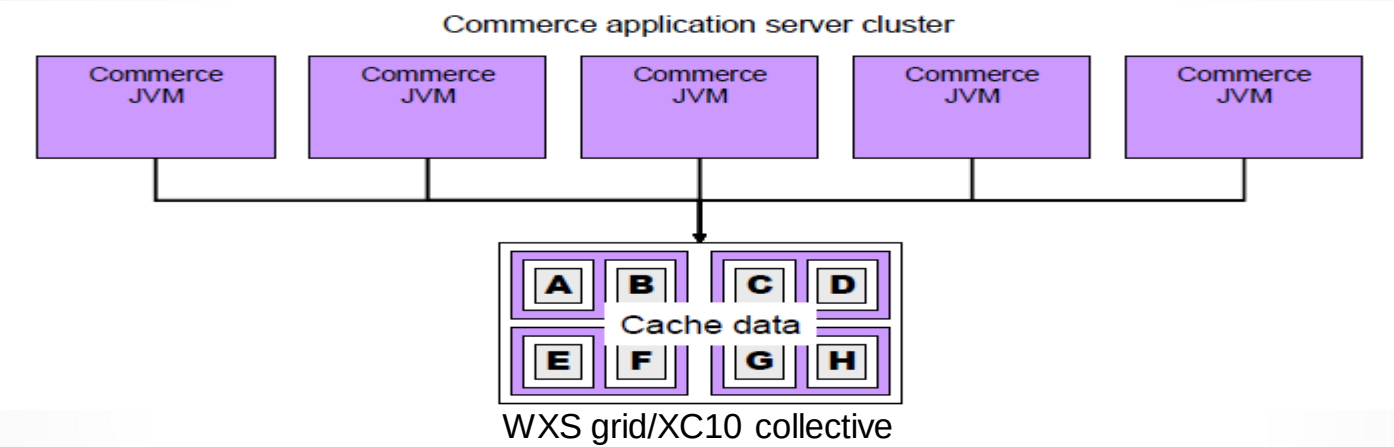
“Drop-In” Dynamic cache service support

- Allows applications using the WebSphere dynamic cache service to leverage the advanced features and performance improvements of IBM elastic cache
- Supports WebSphere Application Server V6.1 or higher
- Dynamic cache evictors, dependency-based invalidation functions, and event listeners can be used on the IBM elastic cache
- Dynamic cache can keep statistics for each grid instance

WebSphere dynamic cache provider



IBM Elastic Cache as Dynamic cache provider



- Much larger cache capacity
- Commerce JVMs run more efficiently
  - Lower local memory requirements
- Improved consistency of performance
  - Improved cache and environment stability
  - High availability of cached data

Elastic Caching for Commerce

Benefits/Value Proposition:

- Better performance:
  - Improved response time - up to 30% in internal tests
  - Faster startup – up to 40% in internal tests
  - More consistent response time
- Better scalability
- Less costly solution
- Rapid Time to Value: configuration, not a coding effort

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. Actual performance in your environment may vary.

Elastic Caching for WebSphere Portal

Problem:

•HTTP session replication required for high availability

•HTTP session data creates large memory requirements in the Portal Server (application) tier

•Costly

Solution:

•Use WebSphere eXtreme Scale / XC10 HTTP session manager for session persistence and replication of customer portlets

Key benefits:

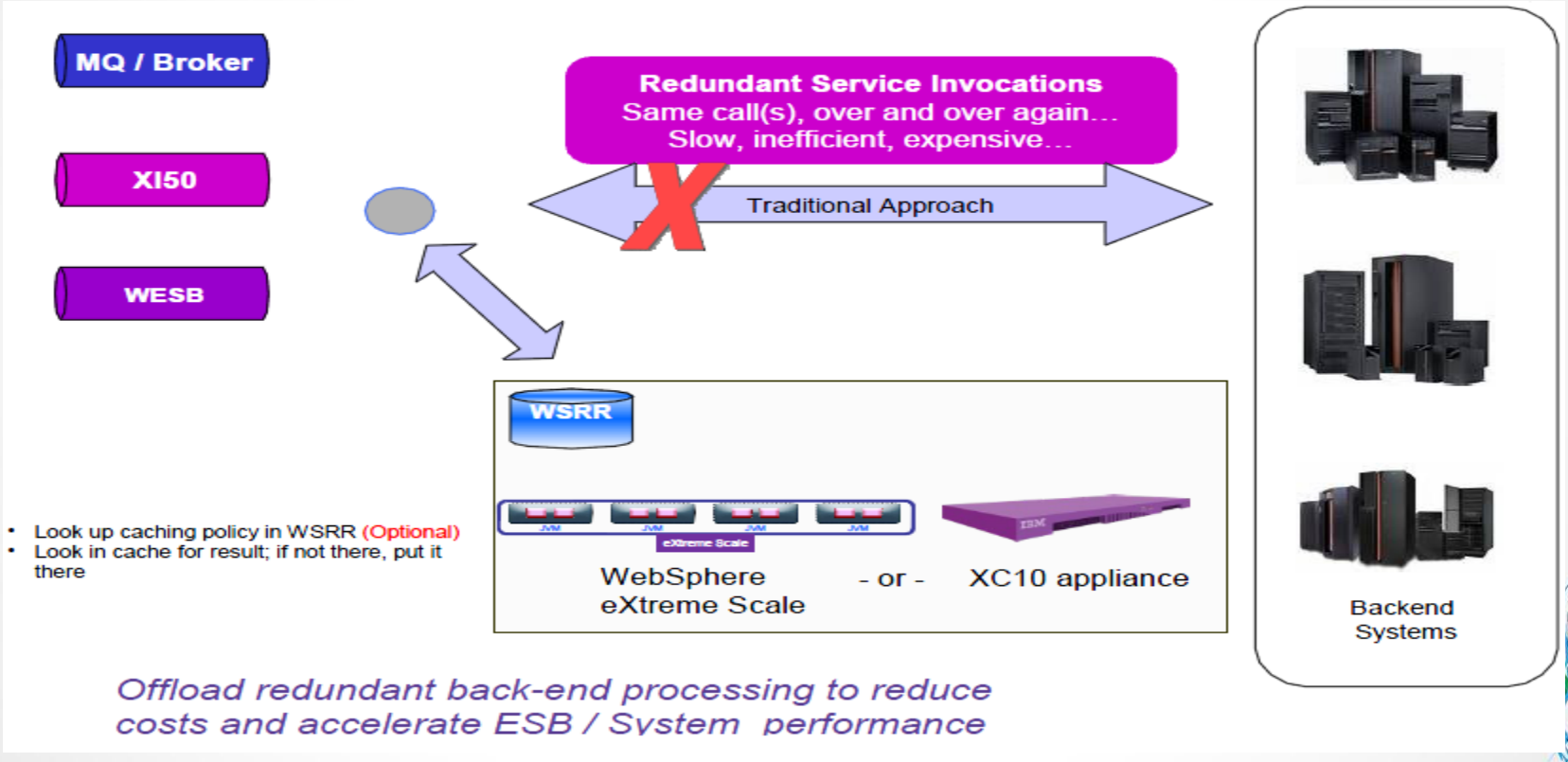
•Reduced HW resources and improved session QoS

- Fewer servers, less memory needed

•Multi-data center session replication enhances disaster recover solutions

•Rapid Time to Value: configuration, not a coding effort

Elastic Caching for Connectivity



- Look up caching policy in WSRR (Optional)
- Look in cache for result; if not there, put it there

Offload redundant back-end processing to reduce costs and accelerate ESB / System performance

Elastic Caching for Connectivity

Benefits/Value Proposition:

- Better performance: turbo-charge services
  - Potential reduction in response time of 50% - 400%
- Better scalability
- Less costly solution
- Rapid Time to Value
- Customer POC
  - 100x or 9,900% faster response time with cache hits
    - Backend takes 3-5 seconds to respond whereas XC10 responds in 0.01-0.05 sec

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. Actual performance in your environment may vary.

Client Usage: On-Line Banking

Retail Banking & Investments

22 Million
online banking users

35x
reduced response times

US$500k
reduced costs per month

20x
reduction in “FCIs”

Next-generation Online Banking

- Before: 700ms to login with 2 backend calls
- After: 20ms to login with profile cache access
- US$6M cost savings in Millions of Instructions Per Second (MIPS) reduction
- 700k transactions per hour across 3 data centers
- 8Gb of data transfer per hour between data centers

Provide seamless cache infrastructure across applications

Deliver high performance & consistent response times

Ensure high availability of critical online applications

Scale with simplicity and lower total cost of ownership (TCO)

© 2009 IBM Corporation

Client Usage: eCommerce: Find product pick list

Refine pick list as they type

2,500,000
Possible matches

3ms
Response time

Linear Scaling for more throughput

Next-generation eCommerce Site

- **Before:** Database table scan for finding all records containing substring (limited to prefix). Performance best described as “as fast as they hit ENTER…”, hundreds of milliseconds
- **After:** <3ms response time and linearly scalable
- Response time is critical when you want the pick list refined literally between key strokes.

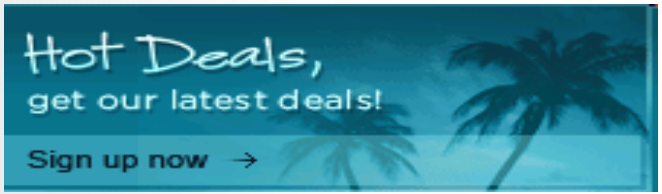
Client Usage: Travel and Transportation

Online Reservations

100x
performance
improvement

Reservations System

- **Before:** 3-5 sec response time
- **After:** .01 -.05 sec response time
- Caching service requests
- Improved the average response time of the Global Distribution System requests for Fare Availability and Category Availability
- 52% caching rate
- Maintained high data integrity. Faster responses were also accurate
- POC in 3.5 hrs



- Improved reliability and scalability of reservation channels
- Reduced traffic to backend systems
- Deliver high performance & consistent response times
- Scale with simplicity and lower TCO

Client Usage: Investment Bank

Stock price service

200,000

Prices updates/sec

reduced response times

reduced costs per month

better availability

Handle dramatic volume increases

- Provides latest and 20 minute old prices for each stock
- **Before:**
- Z/390 application using VSAM kept latest and 20 minute old price for each stock
- **After:**
- WXS Grid maintains the same data.
- 200k stocks
- 200k price update/sec
- 20k price lookups/day

Summary

- IBM's Elastic Caching solution provides a high performance, scalable cache system capable of performing massive volumes of transaction processing
- IBM's Elastic Caching solution integrates with existing environments to save money while improving response time and scalability
- Datapower XC10 appliance provides accelerated time to value
  - *'drop-in' use for side cache scenarios, HTTP Session replication, and WebSphere Application Server dynamic cache service*

We appreciate your feedback.
Please fill out the survey form in order to improve this educational event.

© 2009 IBM Corporation