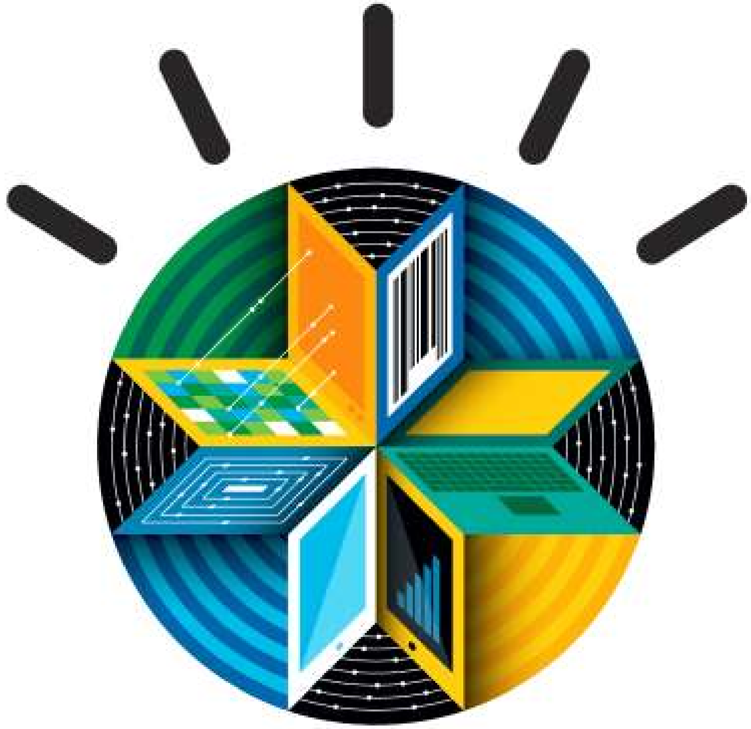
Charles Le Vay
Technical Evangelist, IBM WebSphere Foundation
ccl@us.ibm.com

Elastic Caching Patterns in the Enterprise

Elastic Cache is critical for performance, scalability & high availability



What is a Cache Anyways?

A cache allows you to get stuff faster and helps you avoid doing something over and over again (which may be redundant and may not make sense)

(far away)

(near)

(happy)

© 2012 IBM Corporation

Why do I need Elastic Caching?

Retail, Banking, Finance, Insurance, Telecom, Travel & Transportation …

Mobile Access

Laptops, Ultra-books, Tablets, Smartphones + **3G, 4G, free wi-fi** = **Transaction Overload**

Social Media

TV, Movie, Sports Personality mentions / endorses product + **YouTube, twitter, Facebook** = **Transaction Overload**

Targeted Advertising

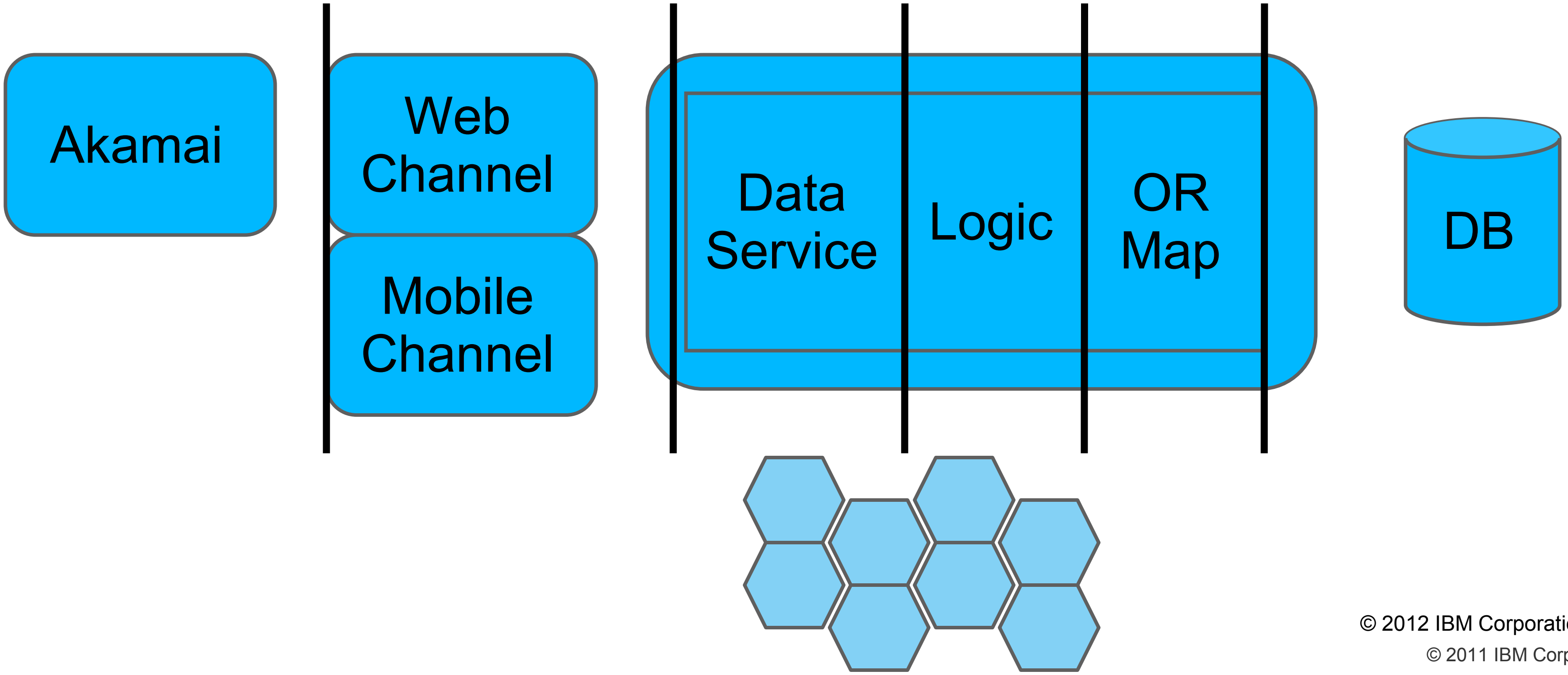
E-mail, SMS, Pop-up, Click-thru Promotions, Web Crawlers + **All of the above** = **Transaction Overload**

Where do we cache?

- A database cache? A page fragment cache? A service Cache?

TOO SPECIFIC!

- A cache is a tool for **reducing application path length**
- OR the **distance data has to travel** before it gets to the customer/   data sink

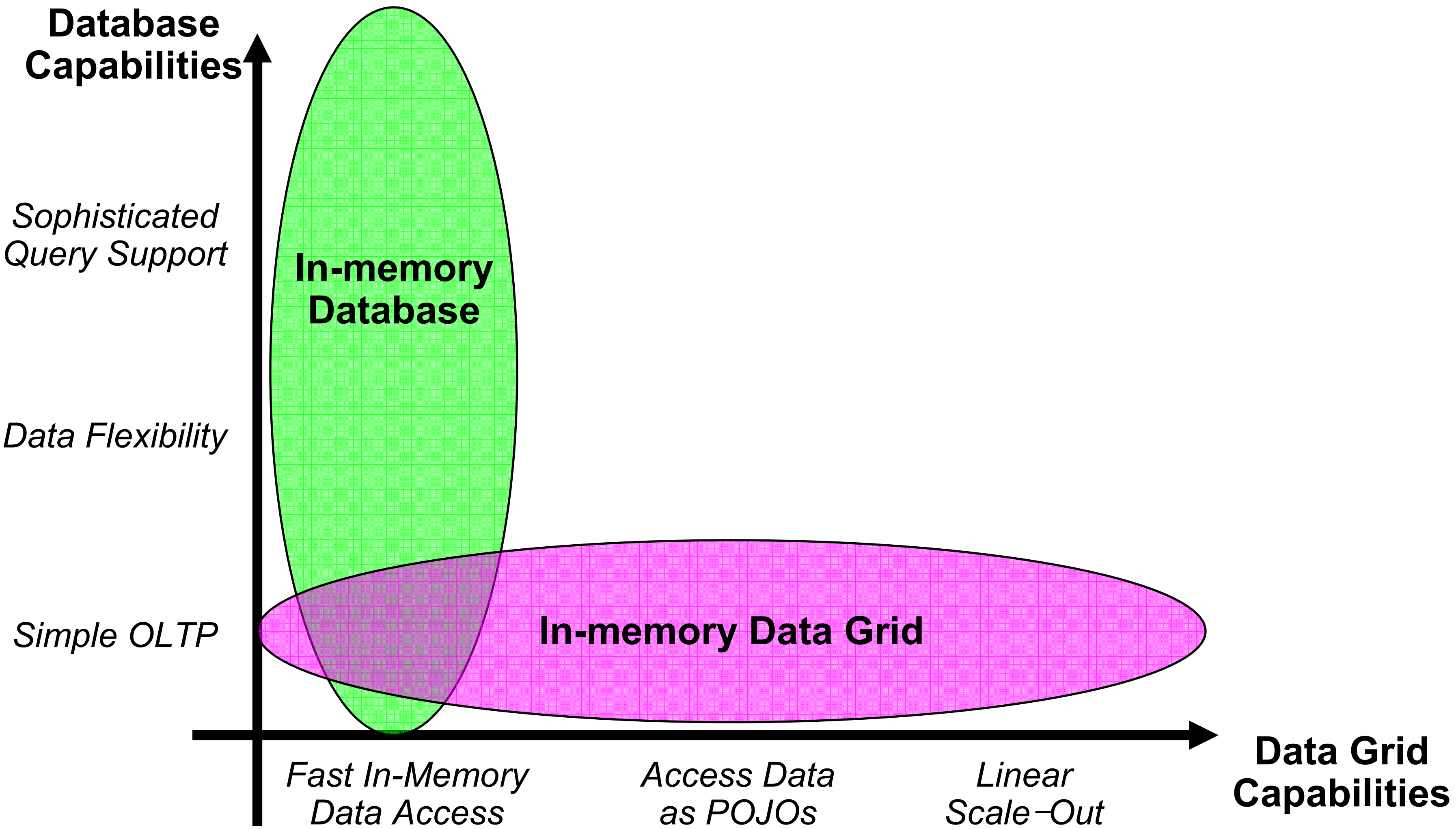


Gee, I already do caching…

Problems with local caching:

- Local cache doesn’t scale
- Local cache is not fault tolerant or highly available
- Need to handle invalidation across a cluster
- Local cache is typically single function or application specific
- Local cache memory requirements could actually degrade performance due to Garbage Collection cycles on large JVM heap sizes
- Resource contention for managing local cache (CPU, memory, I/O)

In-memory Database versus In-memory Data grid



Why a Data Grid?

Scalability issues with database servers

- Adding extra hardware is not easy
- Licensing costs

Large volume of data

- Ability to handle volumes of data without slowing down data access
- Handle data surges during product launches and live events

Fault tolerance and self-healing

- Need for automatic mechanisms to avert system failure affecting end-users
- Data integrity

Data redundancy and replication

- Maintain data reliability in case of failover

What is a Data Grid?

Distributed in-memory object cache

- Elastic, scalable, coherent in-memory cache
- Dynamically caches, partitions, replicates and manages application data and business logic across multiple servers

Capable of massive volumes of transactions

- Provides qualities of service such as transaction integrity, high availability, and predictable response times

Self-healing, allow scale-out / scale-in

- Automatic failure recovery
- on-the-fly addition / removal of memory capacity

Splits a given dataset into partitions

- Primary and Replica shards

Characteristics of a Data Grid

- ***Data stored as key-value pairs.*** *(Think: hash map of 'infinite' size)*

- ***Data is de-normalized***

-***Data Grid is transactional***

-***Simple APIs***
    - *Get, Insert, Update, Delete*
    - *SQL-like query language (map-reduce, grid based applications)*

- ***Can be horizontally partitioned.***
    - list-based, hash-based, range-banged partitioning schemes can be applied to the data

- ***Often transient or referential data***
    -HTTP sessions, user profile, etc
    -Mainframe DBMS offloading
    -Read-only or read-mostly
    -Can tolerate some staleness

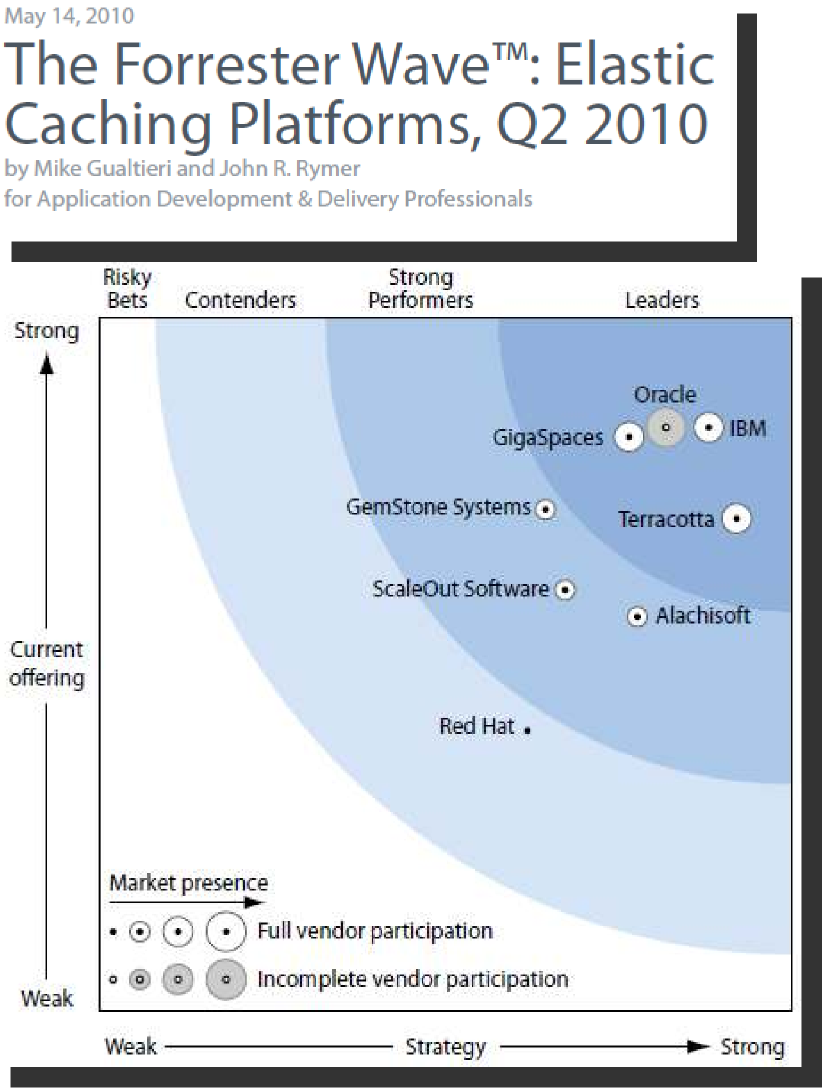
Analyst View

Forrester Research
MAKING LEADERS SUCCESSFUL EVERY DAY

*“IBM has a **strong strategy and execution plan** for WebSphere eXtreme Scale, as well as strong product features across the board.”*

*“eXtreme Scale is especially strong in **runtime architecture**, distributed caching features, **performance** and **scalability** features, and support of **standards**.”*

“By 2014, at least 40% of large organizations will have deployed one or more in-memory data grids. Today, we estimate that less than 10% of large user organizations have IMDG products deployed in production.”



May 14, 2010

The Forrester Wave™: Elastic Caching Platforms, Q2 2010

by Mike Gualtieri and John R. Rymer
for Application Development & Delivery Professionals

Findings

Forrester TEI of WebSphere eXtreme Scale

A Forrester Total Economic Impact™ Study Prepared For IBM

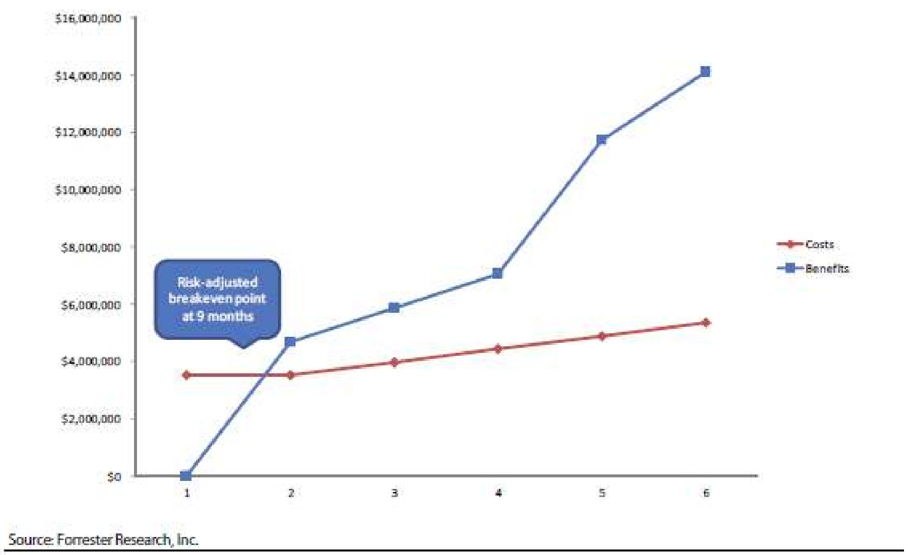
Total Economic Impact ™ Of IBM WebSphere eXtreme Scale

February 2012

The financial analysis found that the organization experienced:

- ROI of 123%
- Payback Period of 9 months
- Net Present Value of $5.9 million

Three-Year Risk Adjusted Analysis



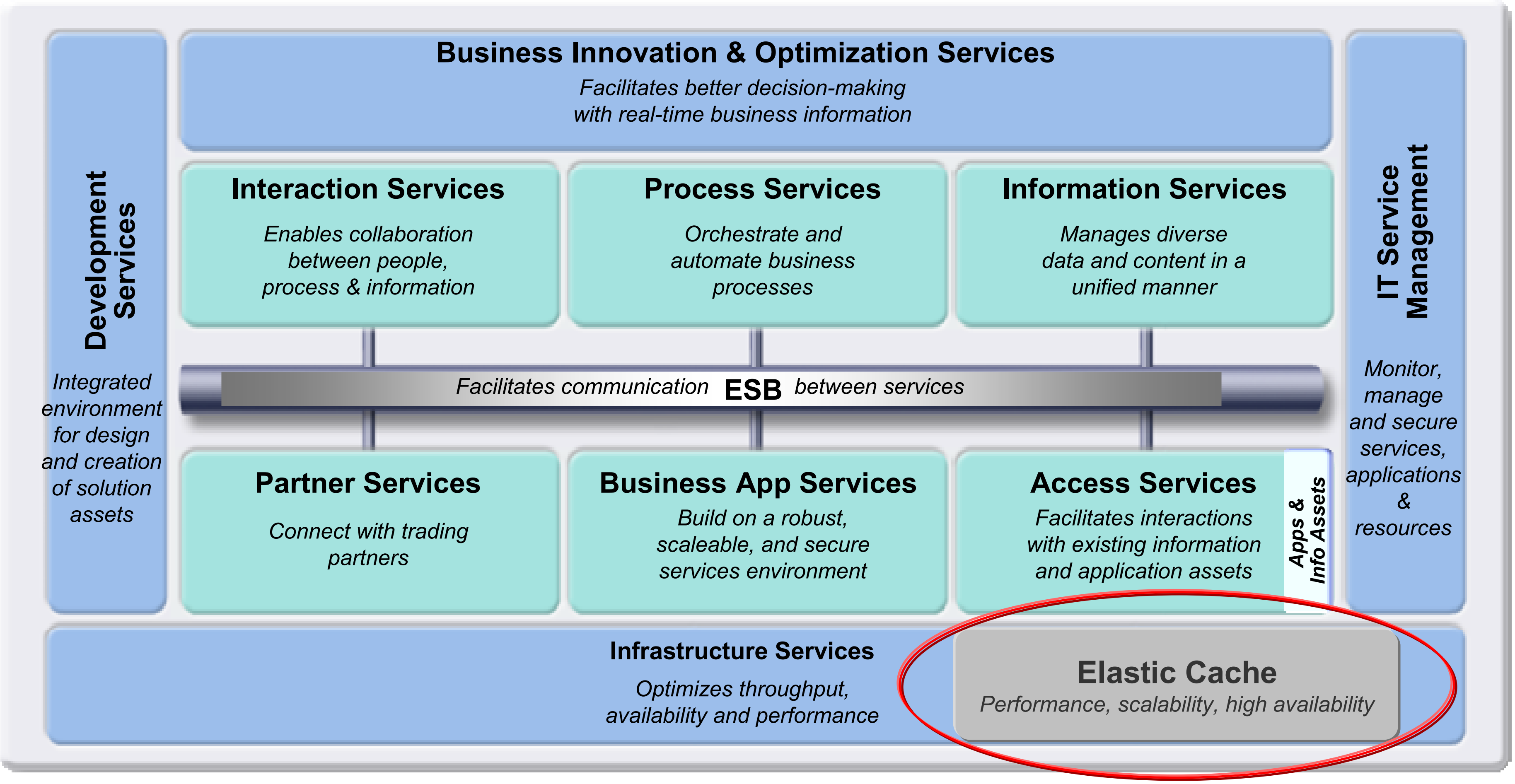
Source: Forrester Research, Inc.

- **Benefits.** The organization Forrester interviewed experienced the following benefits:
    - **Reduction in hardware and software costs.** This benefit represents the hardware and software savings associated with eliminating the need to expand to additional databases.
    - **Annual ongoing staffing costs**. This represents the savings from the ongoing maintenance of additional databases.
    - **Incremental gross revenue (Not quantified)** This benefit represents the incremental revenue associated with the mitigation in user drop-off when users experience slow response resulting from a large surge in traffic during live events and product launches

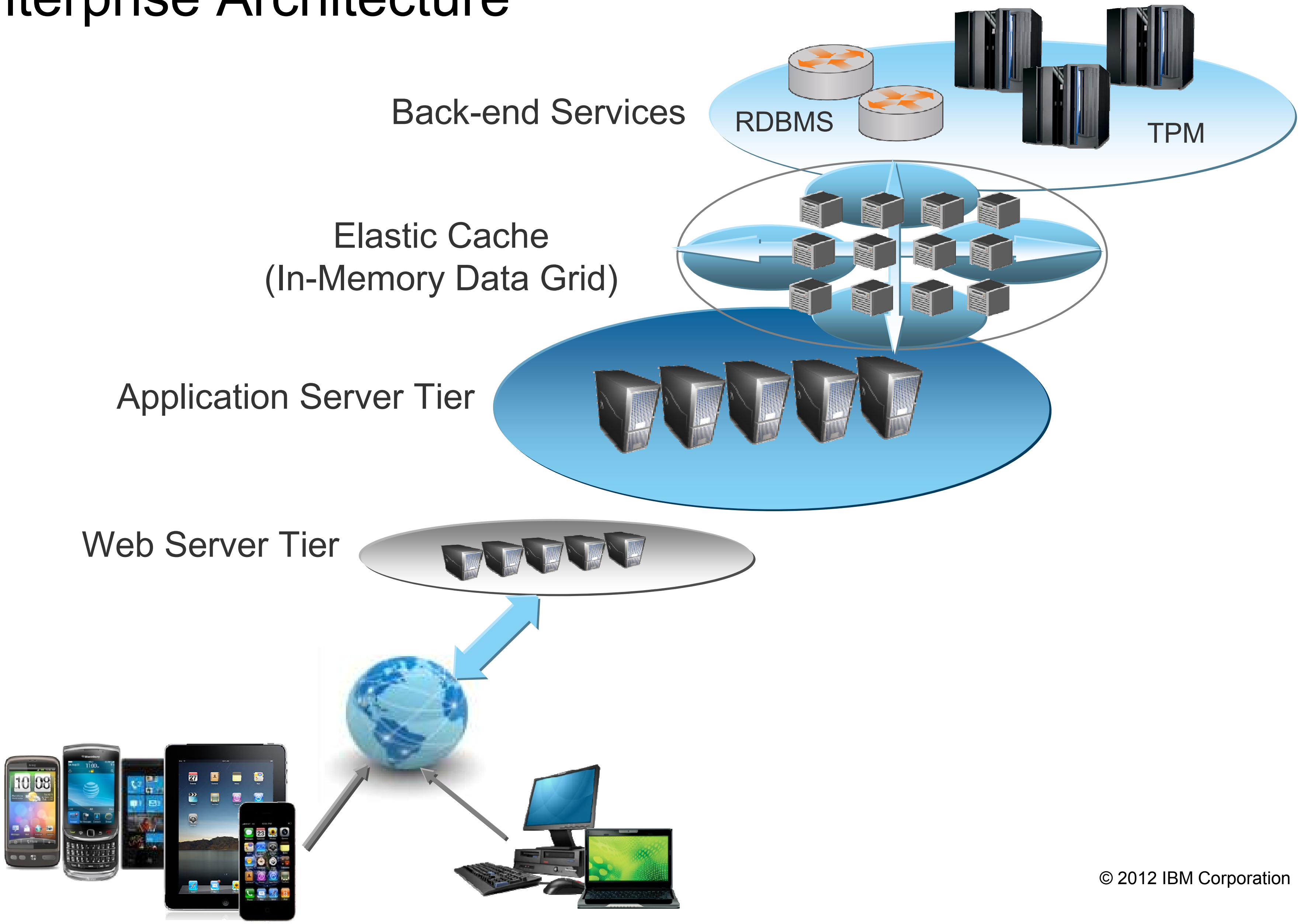
The following is taken from a commissioned study conducted by Forrester Consulting on behalf of IBM."

SOA Reference Architecture

Elastic Cache is critical for performance, scalability & high availability


Development Services
Integrated environment for design and creation of solution assets
Business Innovation & Optimization Services
Facilitates better decision-making with real-time business information
Interaction Services
Enables collaboration between people, process & information
Process Services
Orchestrate and automate business processes
Information Services
Manages diverse data and content in a unified manner
Facilitates communication ESB between services
Partner Services
Connect with trading partners
Business App Services
Build on a robust, scaleable, and secure services environment
Access Services
Facilitates interactions with existing information and application assets
Apps & Info Assets
IT Service Management
Monitor, manage and secure services, applications & resources
Infrastructure Services
Optimizes throughput, availability and performance
Elastic Cache
Performance, scalability, high availability

Enterprise Architecture



Back-end Services

RDBMS

TPM

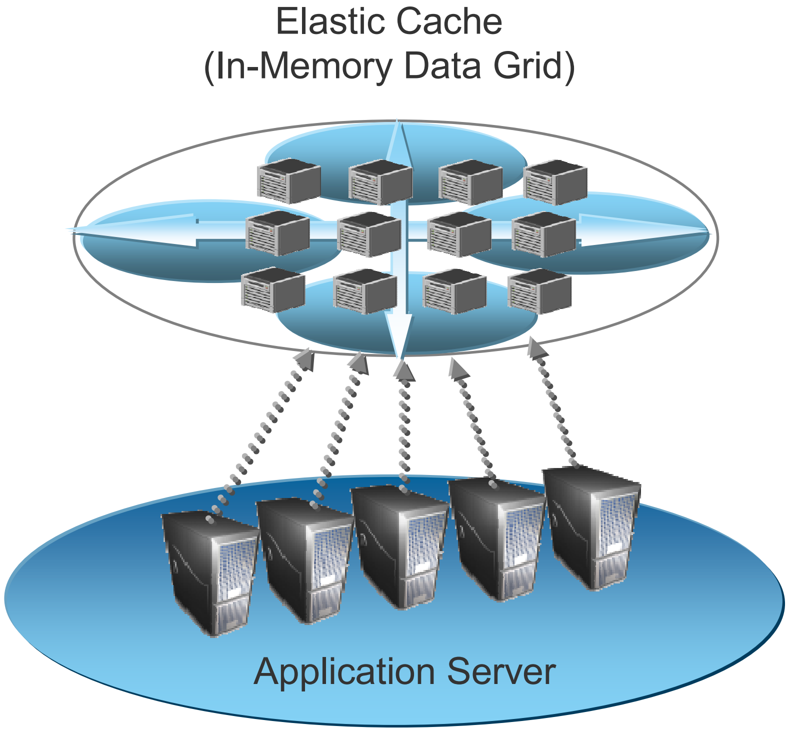
Elastic Cache
(In-Memory Data Grid)

Application Server Tier

Web Server Tier

1 Application State Store Pattern

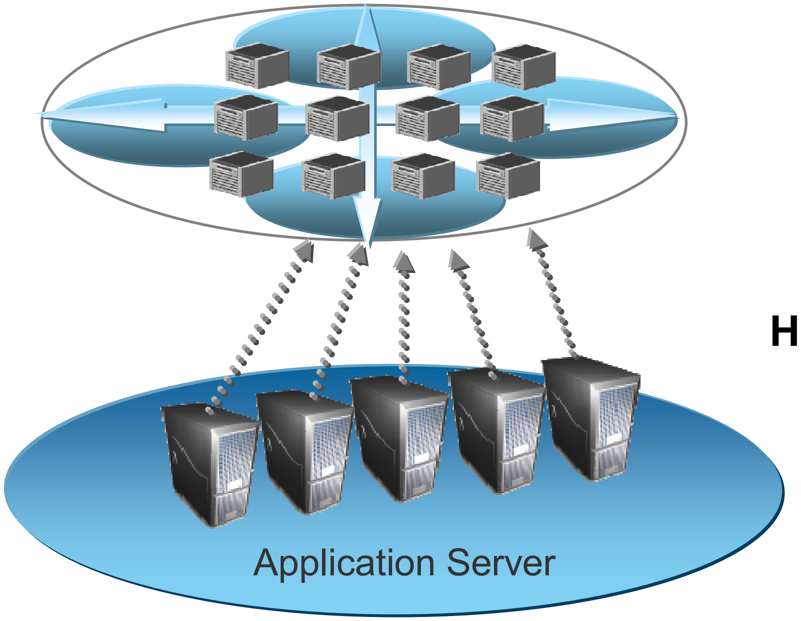
Applications use single coherent, highly- available, scalable cache



- Single replacement for multiple local caches
- Consistent response times
- Reduces Application Server JVM heap size
    - Improved memory utilization - more memory for applications
    - Faster Application Server start-up
- Removes invalidation chatter of local caches
- Applications move application state to grid
    - Stateless applications scale elastically
- Application state can be shared across data centers for high availability

HTTP Session Distribution

Elastic Cache
(In-Memory Data Grid)



Improved Performance

- Improves start-up time when bringing a new server on-line

Better Scalability (Less expensive)

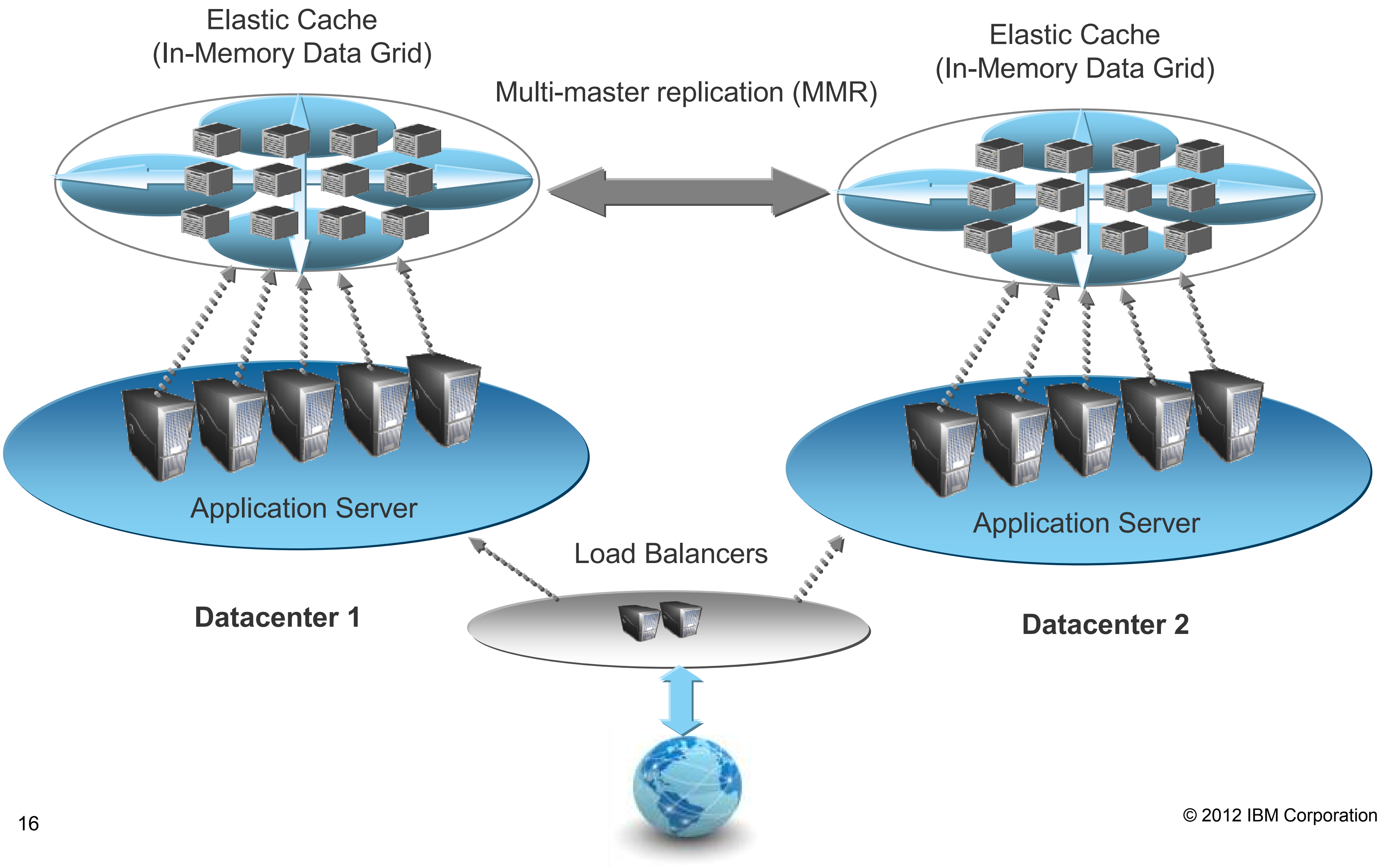
- Replaces memory to memory replication
- Replaces need for database persistence
- Less expensive than scaling the database
- Faster, more consistent response times
- Makes better use of system resources
- Larger cache capacity

Higher Availability

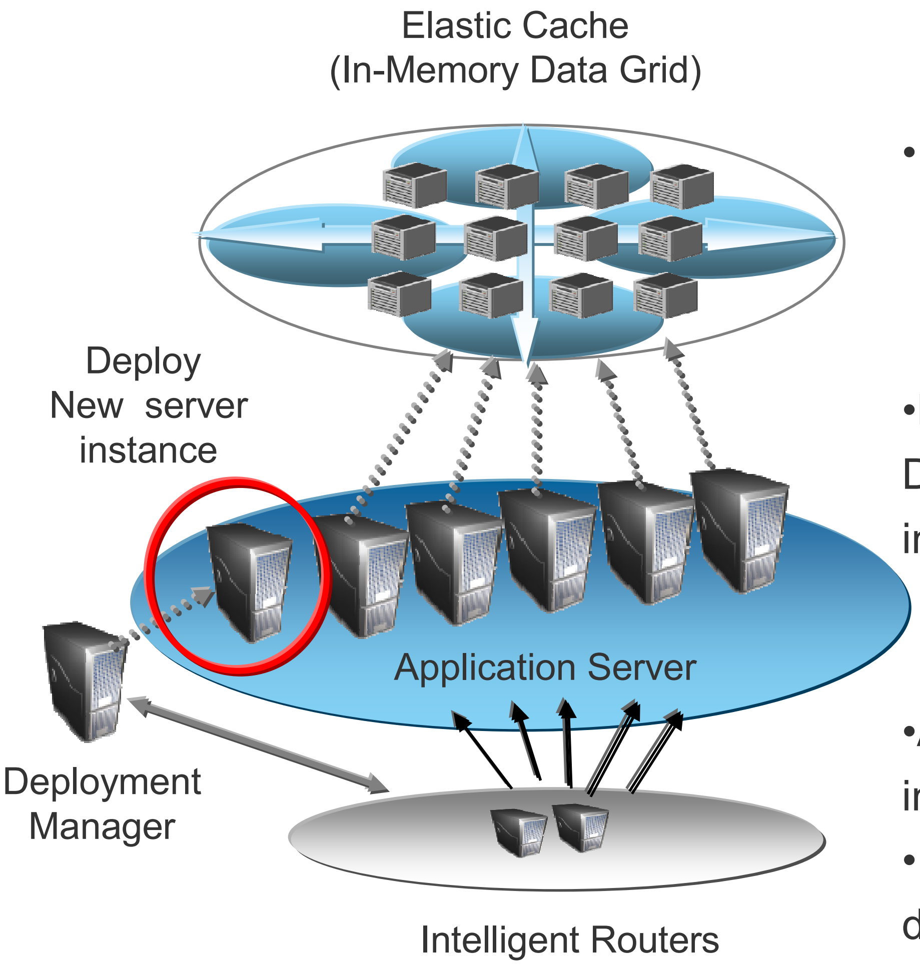
- Provides fault tolerance and high availability of session
- Not only within the datacenter, but across datacenters
- Session replication or distribution is crucial in highly available systems to provide uninterrupted user experience (e.g. Shopping Cart).

No new code required! Easy to configure

Active/Active Datacenter HTTP session failover

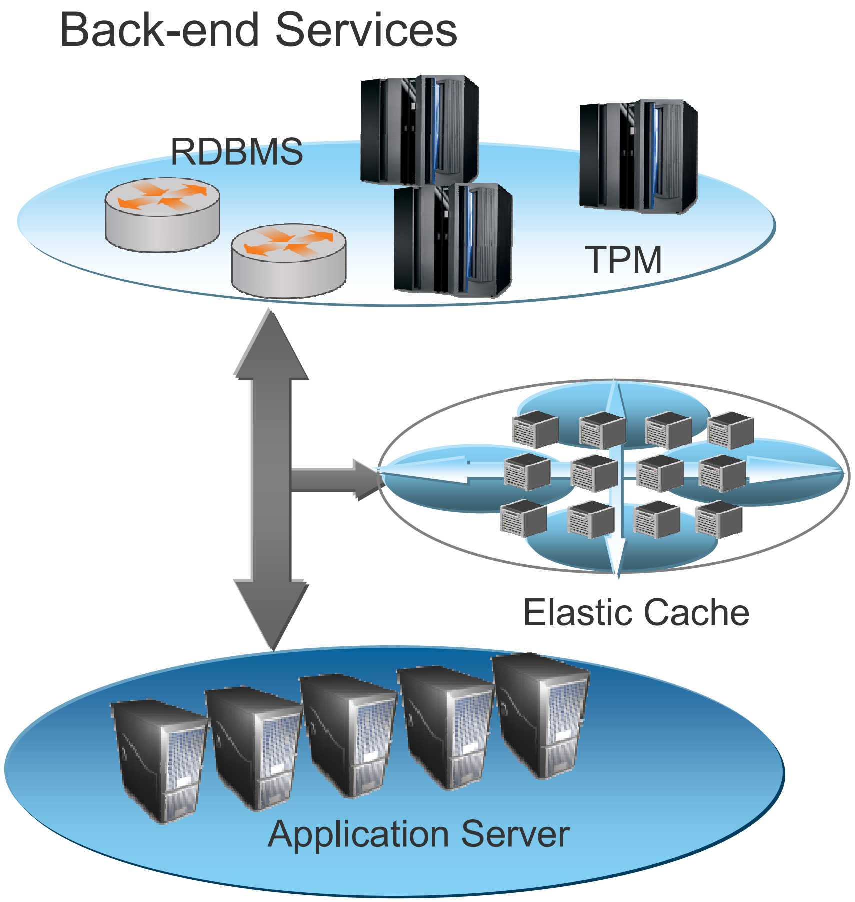


Application Server Elasticity – Dynamic Web App



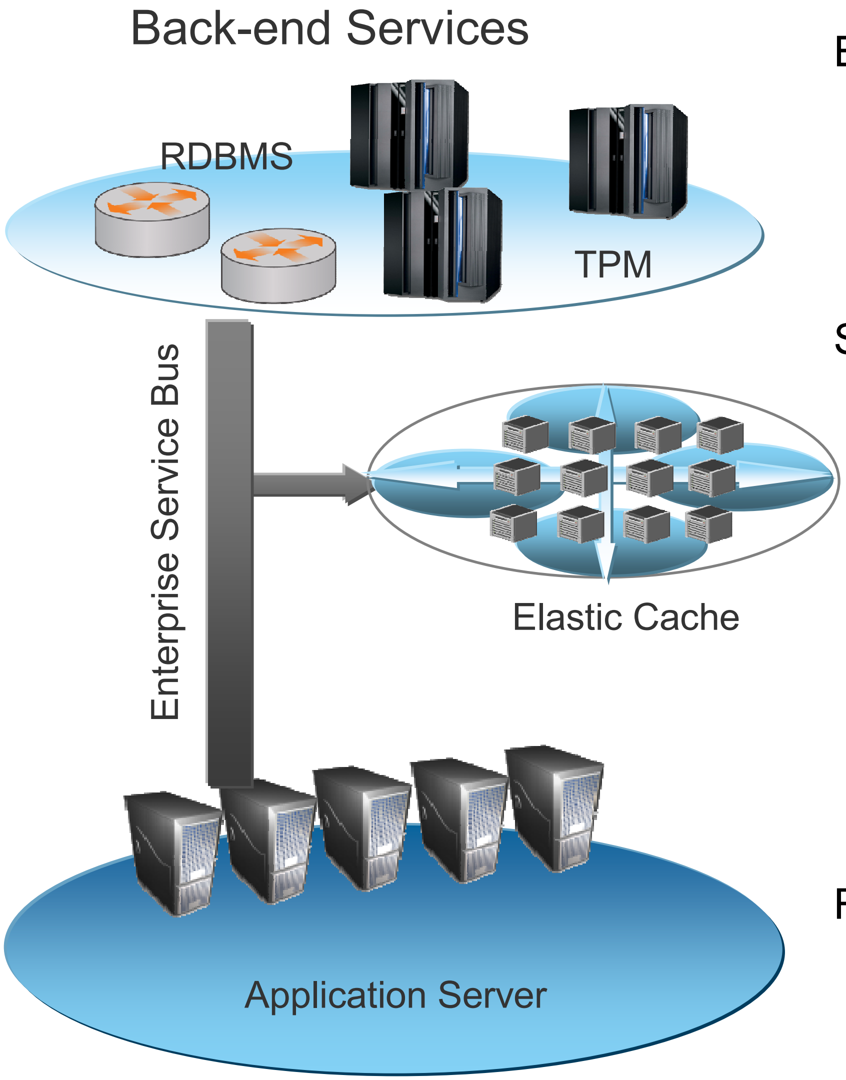
- Intelligent Routers monitor
  - Number of connections
  - Response Time
  - Application Server health
- Based on SLA, Intelligent Router coordinates with Deployment Manager to deploy another server instance to meet SLA
  - Hypervisor instance (App Server/Portal)
  - Dynamic cluster member
- Application state is immediately available to new instance via elastic cache
- During periods of reduced load, system can scale down and release resources for other purposes

2 Side Cache Pattern



- Client first checks the grid before using the data access layer to connect to a back end data store.
- If an object is not returned from the grid (a cache “miss”), the client uses the data access layer as usual to retrieve the data.
- The result is put into the grid to enable faster access the next time.
- The back end remains the system of record, and usually only a small amount of the data is cached in the grid.
- An object is stored only once in the cache, even if multiple clients use it. Thus, more memory is available for caching, more data can be cached, which increases the cache hit rate.
- Improve performance and offload unnecessary workload on backend systems.

Enterprise Service Bus – Side Cache



Easily integrates into the existing business process

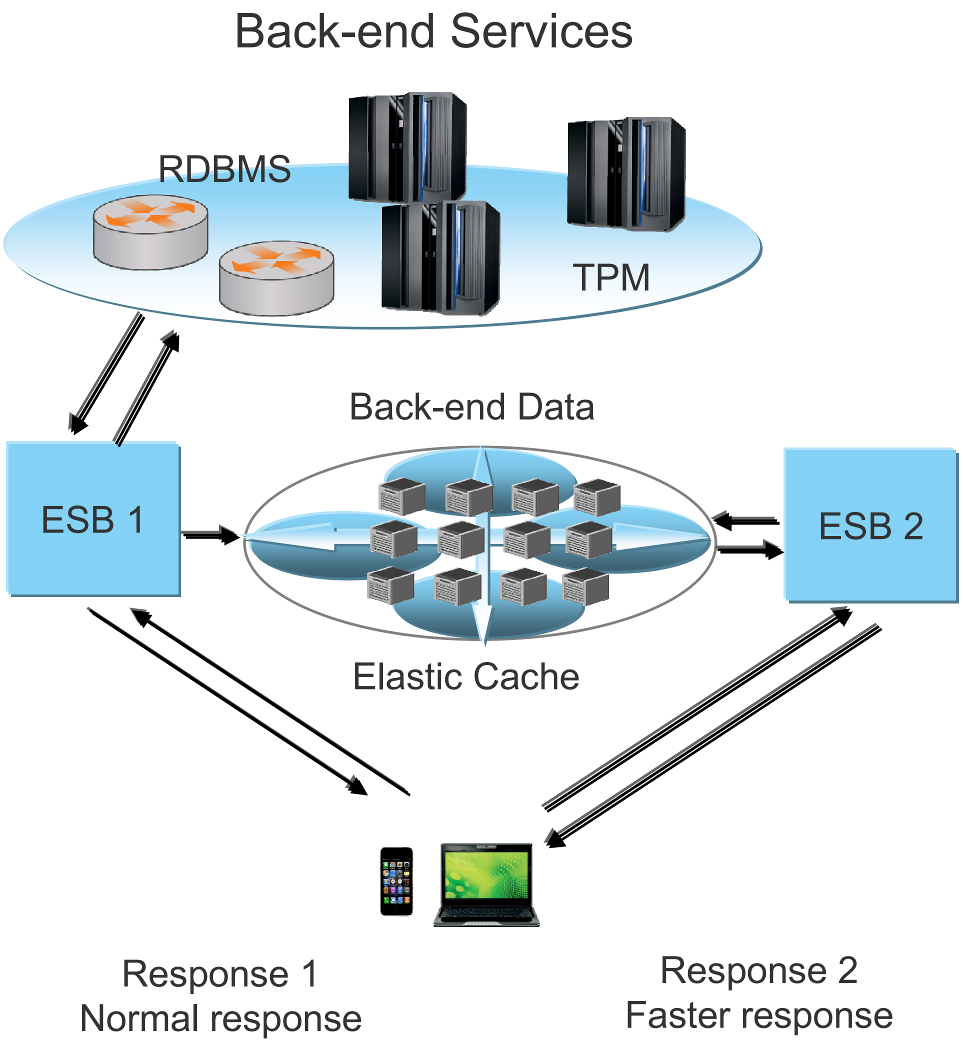
- No code changes to the client application or back-end application
- Simply add the side cache mediation at the ESB layer

Significantly reduces the load on the back-end system by eliminating redundant requests

- Eliminates costly MIPS by eliminating redundant request
- Allows for more “REAL” work to be performed
- Improves overall response time
- Minimizes the need to scale hardware to increase processing capacity since the back-end system no longer has to handle redundant requests.

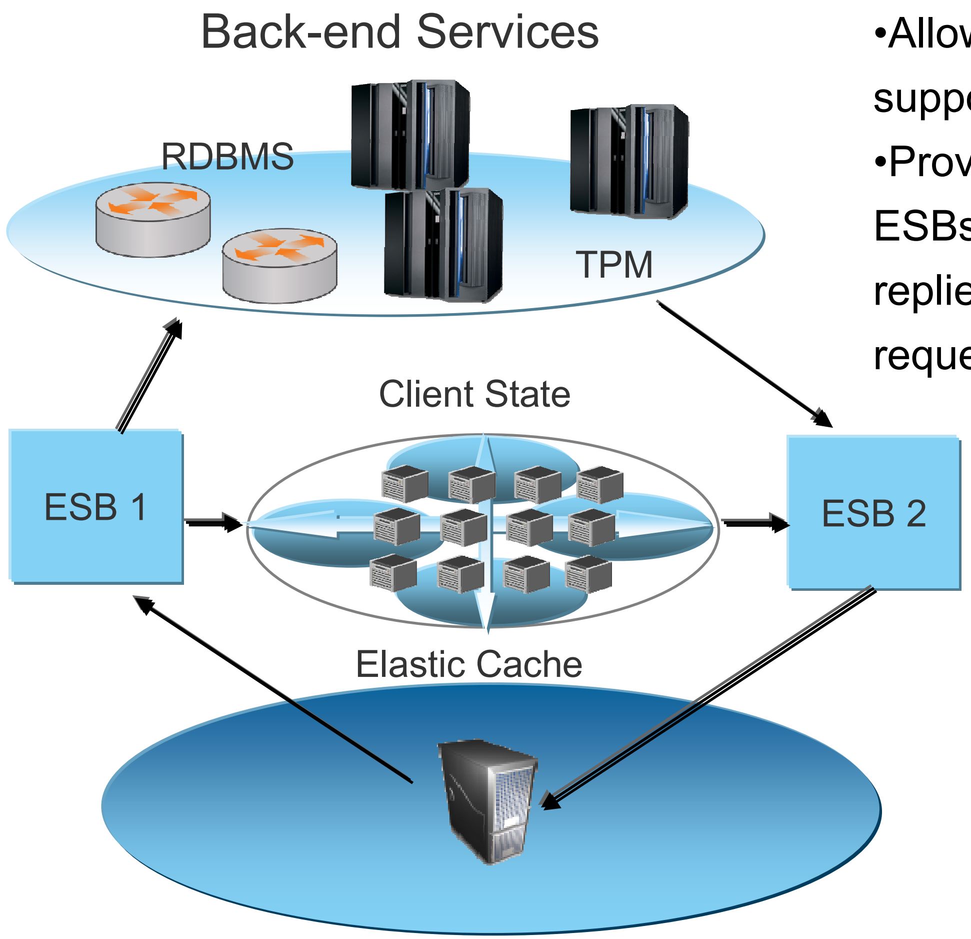
Response time from elastic cache is in milliseconds

Enterprise Service Bus – Global Cache for Performance



- •Lay foundation for low latency access to data required for real-time analytics
- •Response time is key driver in choice of infrastructure provider for mobile implementations
- •Fast access to cached data for common DB lookups
- •Reduce load on DB by eliminating redundant lookups

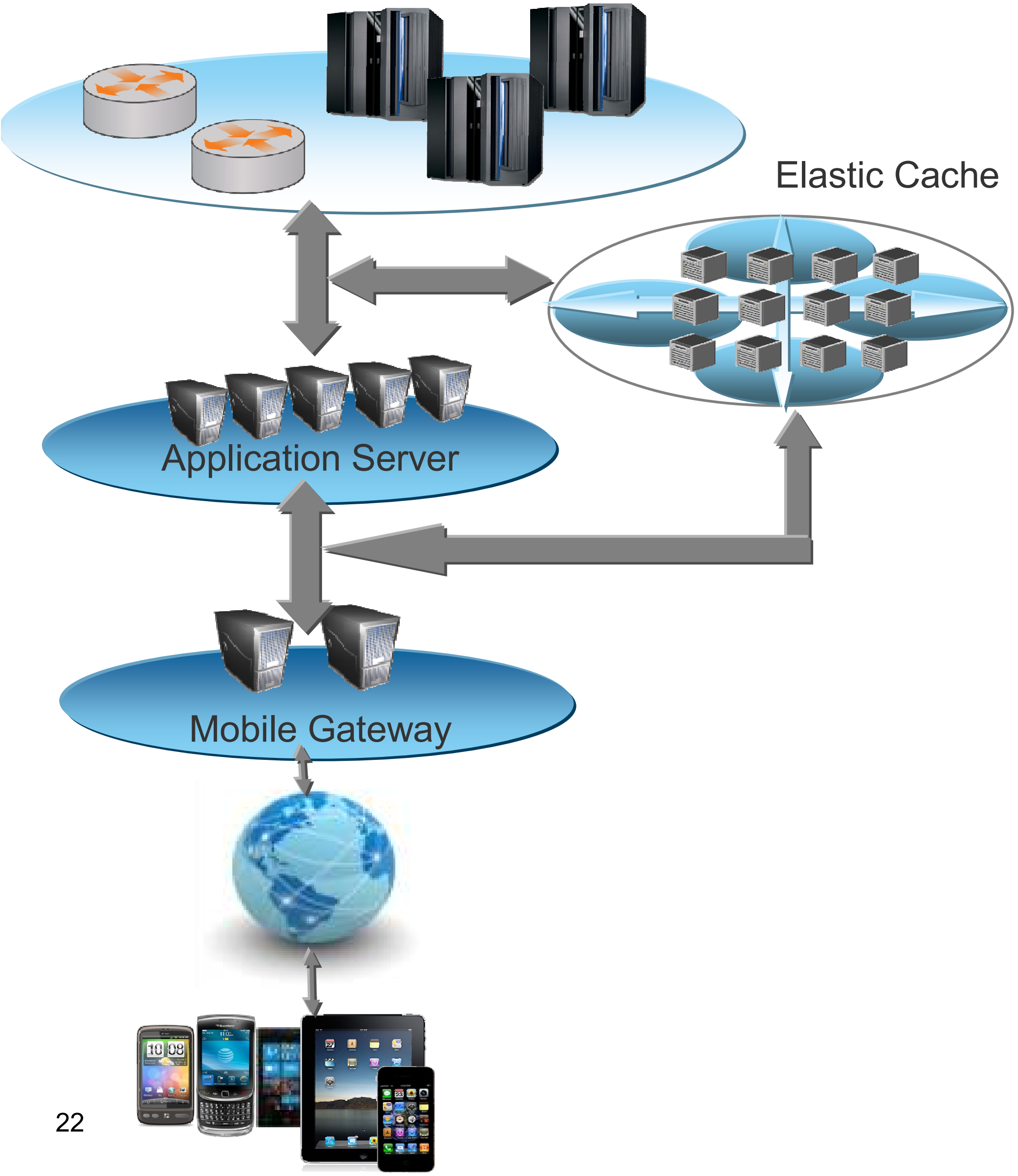
Enterprise Service Bus – Global Cache for Scalability



•Allow the building of scalable infrastructure to support growth in demand for services

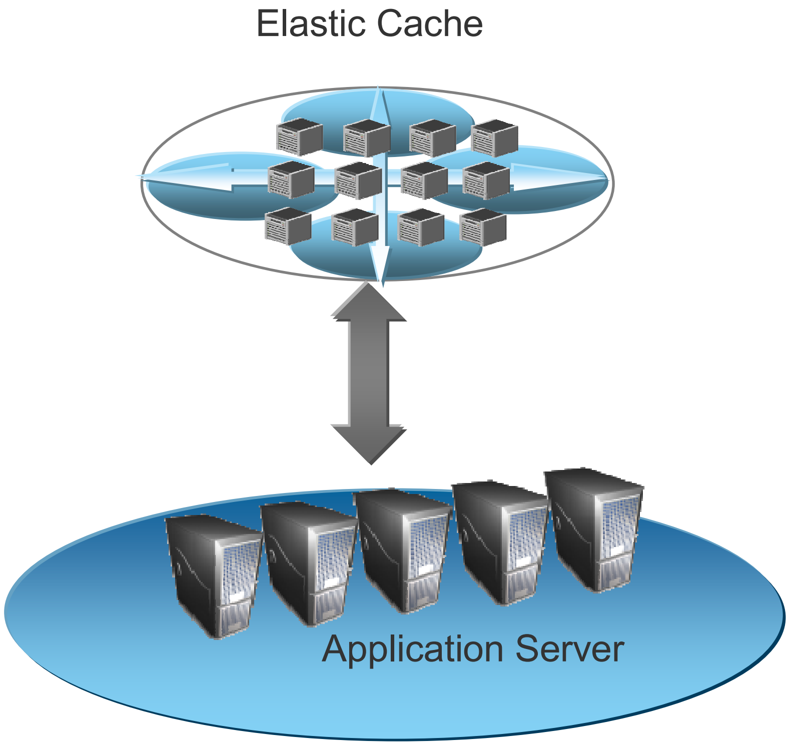
•Provide a cache for sharing data between ESBs, which enables...Ability to correlate replies to **share workload between ESBs** in request-response scenarios

Mobile Gateway Acceleration



- By integrating Elastic Cache with Mobile Gateway, users can seen improved performance without the penalty of having to scale to a large cluster of Mobile Gateways.
- Use Side cache to cache XSLT transforms
- Directly access the Elastic Cache to retrieve cached objects
- Use Elastic Cache to provide session state for stateless communication

Dynamic Cache provider



Primary benefits of deploying elastic caching, compared with DynaCache + disk offload

Measurable performance improvement is possible when compared to WAS dynacache with DRS or accessing disk offloaded data

Less statistical fluctuation in response-time = more consistent user experience

Potentially up to 40% improvement in time to reach steady-state after full or partial site restart, or after full cache invalidation

Extremely fast Commerce recovery time

Organic cache warm up

Simplified tuning and operational maintenance

Reduced I/O volume to high-speed disk

Elastic – SAN based disk offload systems are not elastic

Coherent and consistent cache

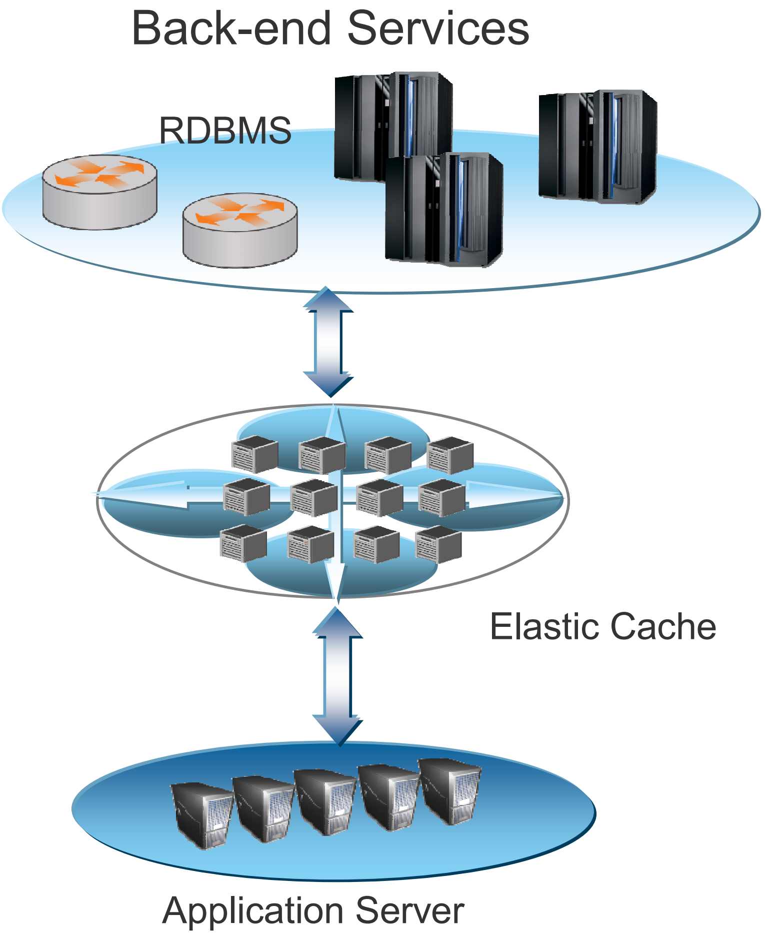
- Same version of the page is always shown
- Facilitate edge-caching
- No possibility of encountering stale data

No cold cache hit on the back-end

Customer results may vary due to differences in scenario and environment

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

3 In-line cache – Database shock absorber



The grid can be used as a special data access layer where it is configured to use a loader to get data from the back-end system.

- Read through cache
- Write through cache (Synchronous writes)
- Write-behind cache (Asynchronous writes)

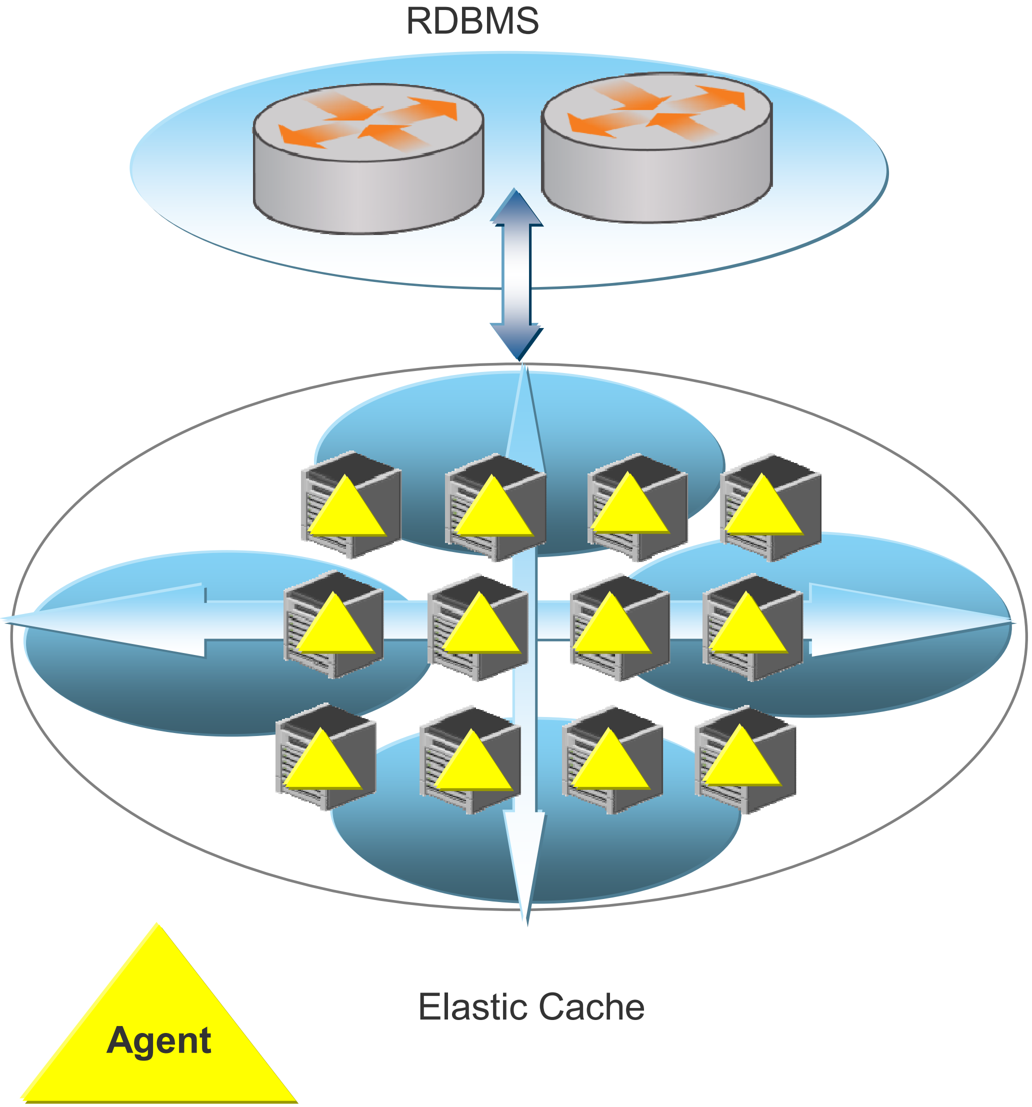
System of Record Data Store

- Cache is used as the system of record
- Write behind technology pushes changes asynchronously to the backend.
  Changes batched
  Only last change written
- Runs through backend outages!

Benefits

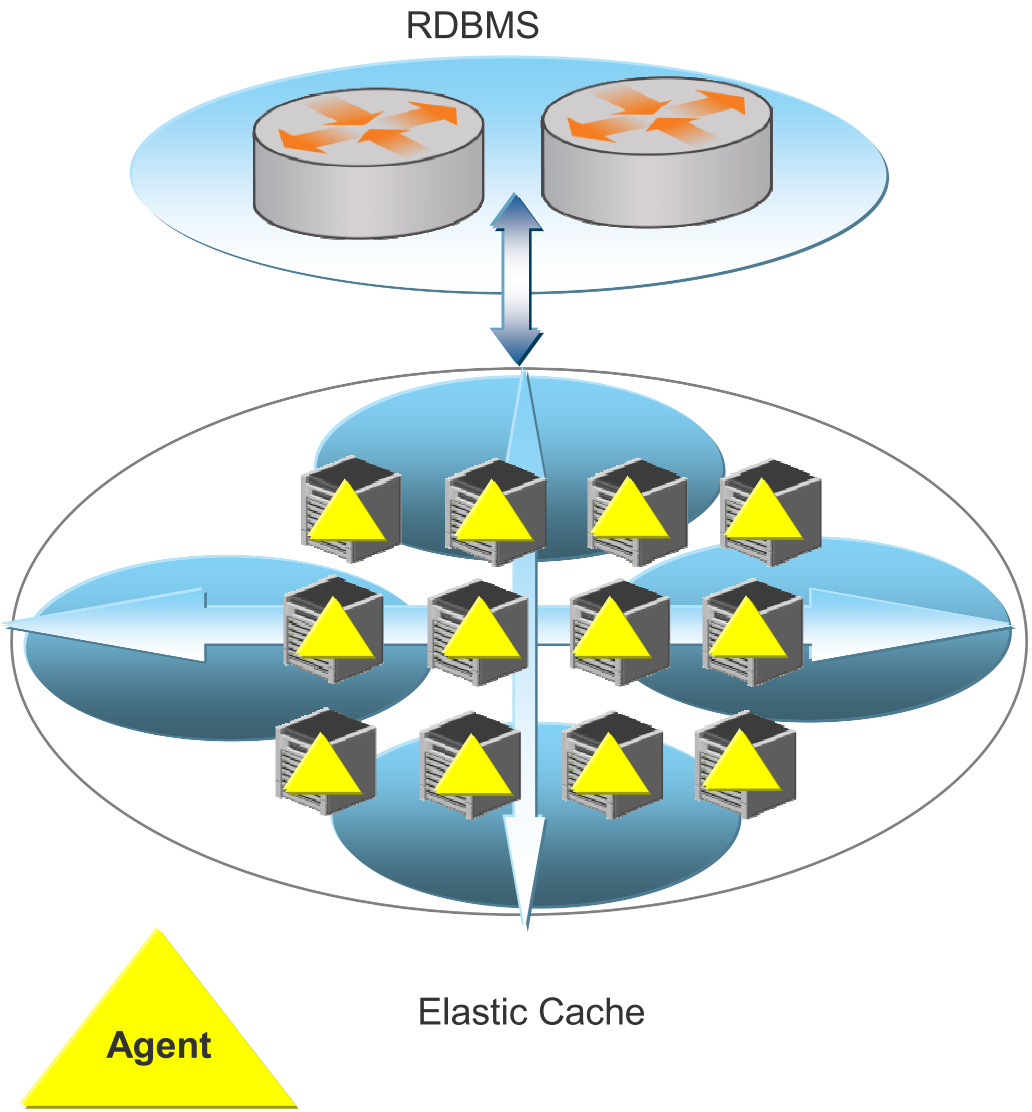
- Writes faster (memory vs. disk speed)
- Backend load reduced, throughput improved
- Increased availability and scalability

4 eXtreme Transaction Processing



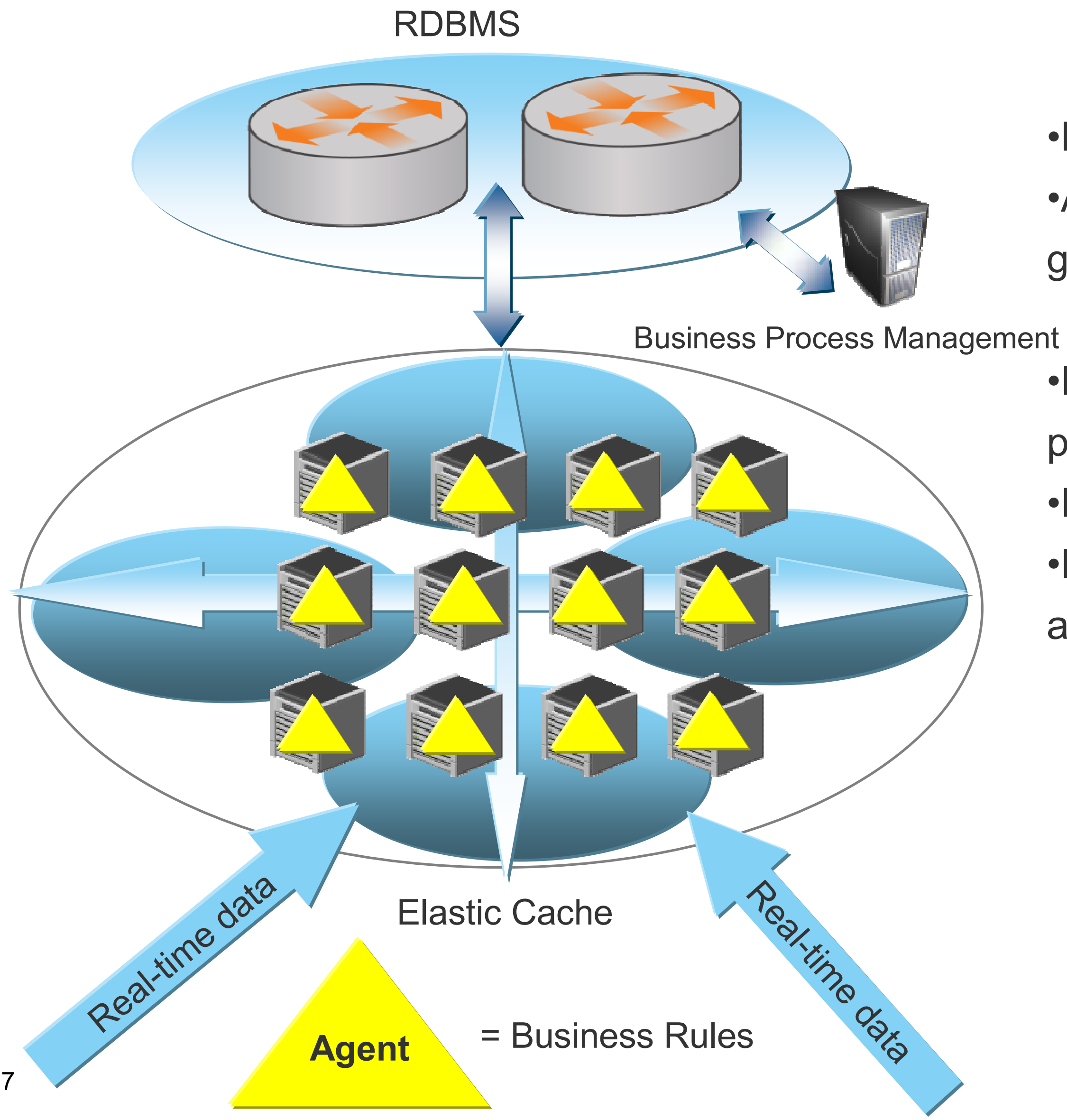
- Lowest possible latency
- Application code (Agent) runs in the grid itself
  - Map/Reduce API supported
- Events routed to correct partitions for processing
- Extension of Write behind scenario
- Databases relegated to durable log and reports

Map Reduce Parallel Processing



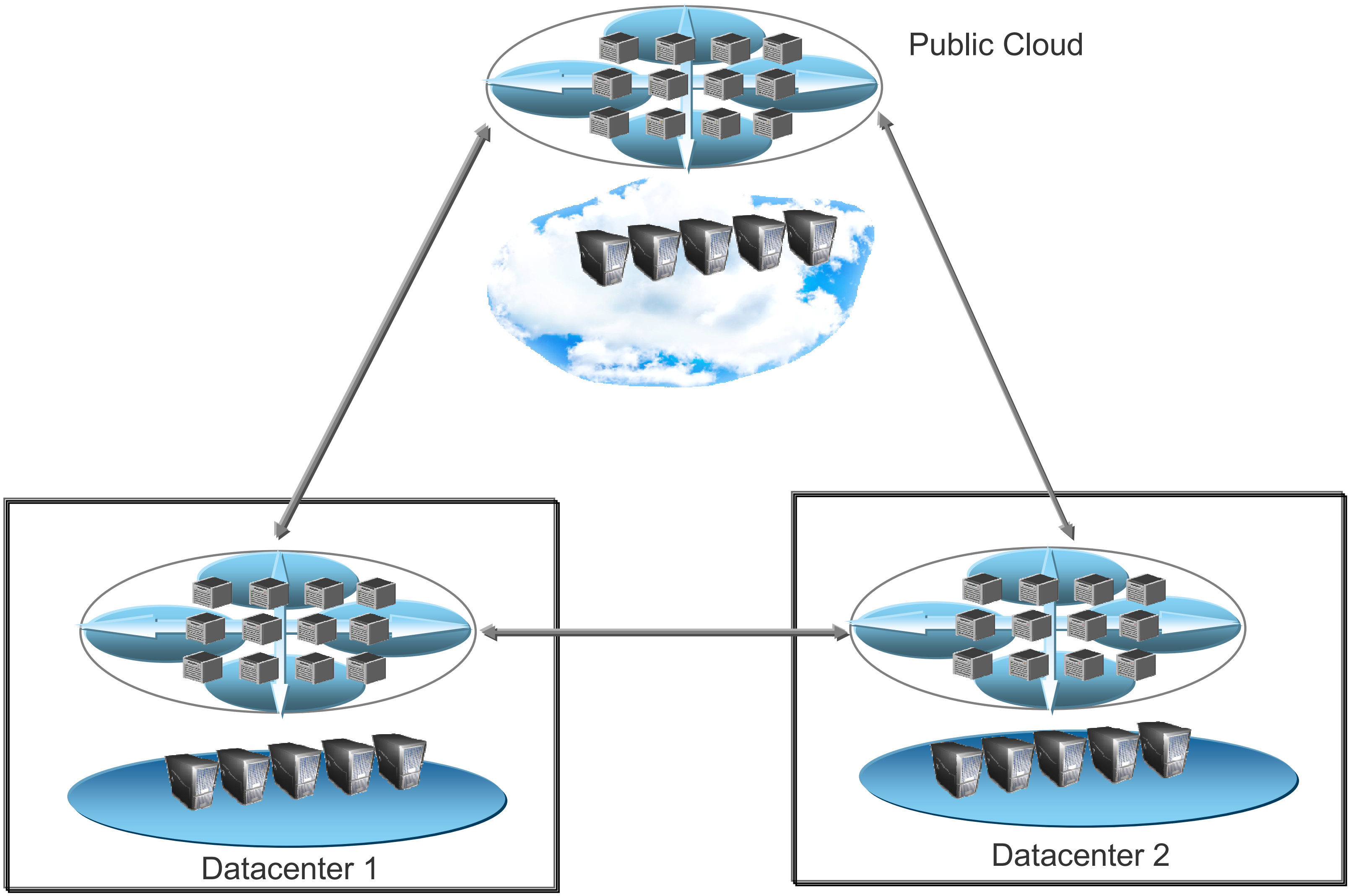
- Parallel Map
  - Allows the entries for a set of Entities or Objects to be processed and returns a result for each entry processed
- Parallel Reduction
  - Processes a subset of the entries and calculates a single result for the group of entries
- Since the Elastic Cache is the system of record, there is little to no load on the back-end data stores

Real-Time Business Rules / Event Processing

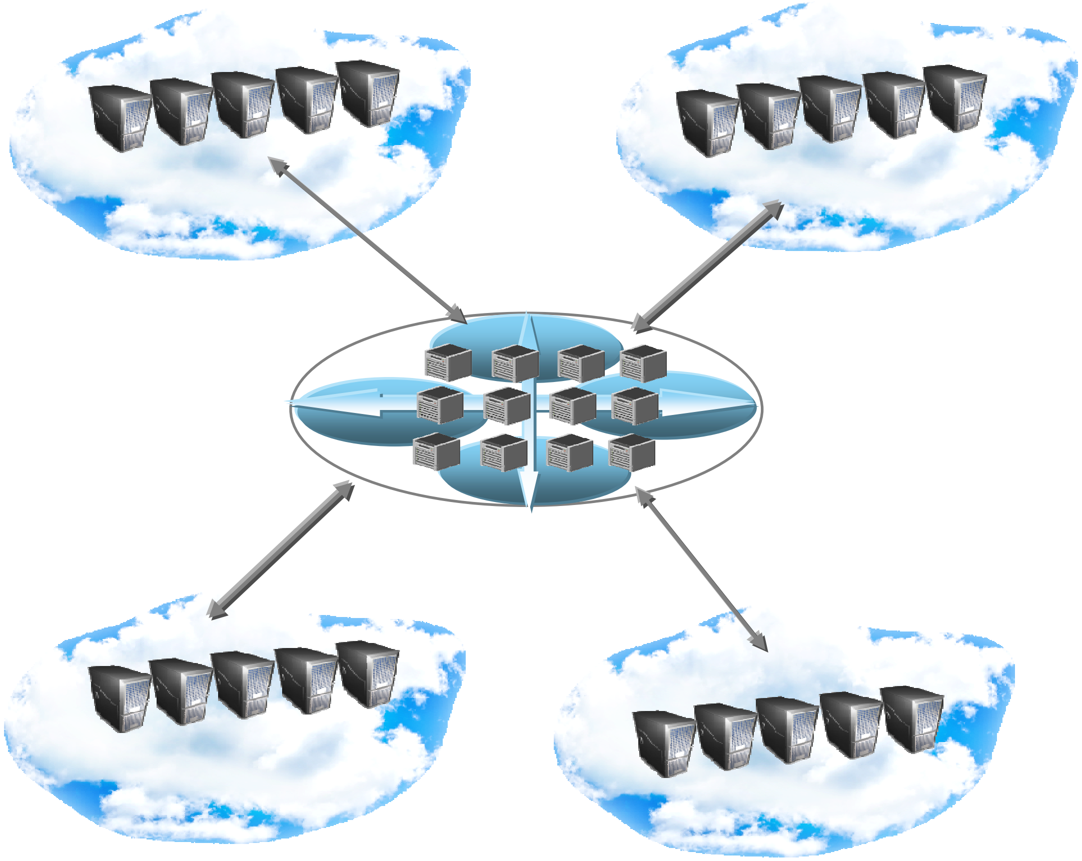


- Lowest possible latency
- Application code (Agent) runs in the grid itself
  - Map/Reduce API supported
- Events routed to correct partitions for processing
- Extension of Write behind scenario
- Databases relegated to durable log and reports

Multi-datacenter - High Availability/Distributed Computing



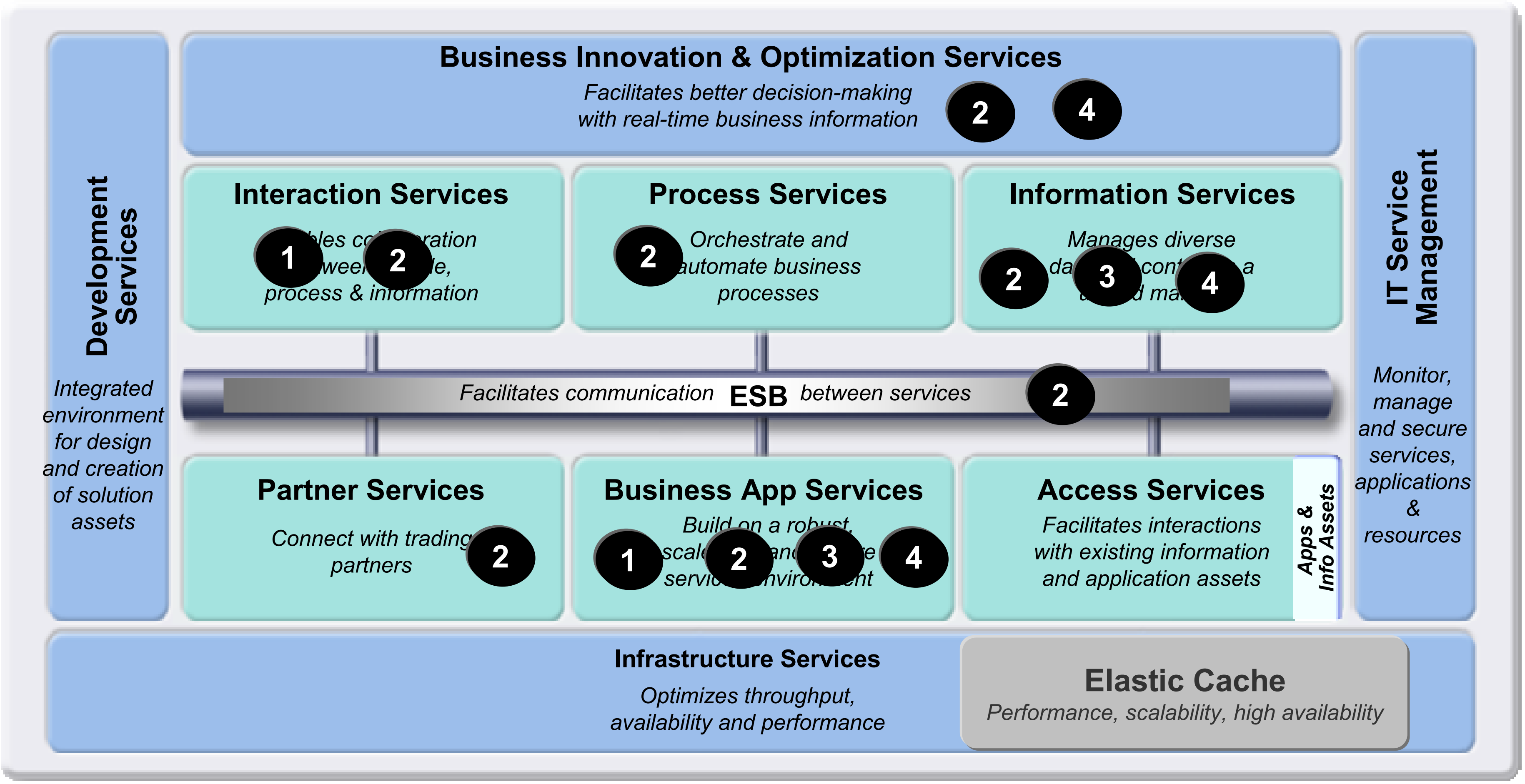
Elastic Cache Shared Service



- •Provides Elastic Caching resource for cloud based architectures
- •Elastic Cache service is multi-tenant
    - •Support grid capping
    - •Individual maps per cloud group
    - •Authentication/Authorization per map/grid
- •Used for
    - •Simple Cache
    - •HTTP session distribution
    - •Dynamic Cache provider

SOA Reference Architecture

Elastic Cache is critical for performance, scalability & high availability



Business Innovation & Optimization Services

Facilitates better decision-making with real-time business information

Development Services

Integrated environment for design and creation of solution assets

Interaction Services

bles co ration wee e, process & information

Process Services

Orchestrate and automate business processes

Information Services

Manages diverse da cont a d ma

Facilitates communication **ESB** *between services*

Partner Services

Connect with trading partners

Business App Services

Build on a robust, scal an e servic nvironment

Access Services

Facilitates interactions with existing information and application assets

Apps & Info Assets

IT Service Management

Monitor, manage and secure services, applications & resources

Infrastructure Services

Optimizes throughput, availability and performance

Elastic Cache

Performance, scalability, high availability

IBM Elastic Caching Delivers

Consistent Response Times, High Availability of Data & Linear Scalability for Enterprise-wide Data Grids

WebSphere eXtreme Scale V8.6

A powerful, scalable, elastic in-memory grid for your business-critical applications

Java and .NET applications can now interact natively with the same data in the same data grid, leading the way toward a true enterprise-wide data grid.

A new REST Gateway provides simple access from other languages.

WXS 8.6 delivers a faster, more compact serialization format called eXtreme Data Format (XDF), which is neutral to programming languages.

A new transport mechanism, eXtreme IO (XIO) removes the dependency on the IBM ORB, enabling easier integration with existing environments.

Built in pub/sub capabilities enable WXS 8.6 to update client "near caches" whenever data is updated, deleted, or invalidated on the server side.

API enhancements enable continuous query or data that is inserted and updated in the grid.

DataPower XC10 Appliance V2.1

Rapid, "drop-in" use with a broad range of Java and non-Java application environments

Multi-data center support allows customers to host data on XC10s in multiple locations with data kept in synch through multi-master replication

Support for elastic caching for WAS Liberty brings scalability, fault tolerance and high availability

Dynamic cache replacement for Web Content Manager running on WebSphere Portal & WebSphere Commerce

Monitoring enhancements ease administration and improve serviceability

Support for Simple Network Management Protocol (SNMP) traps for event notification

Capability to query grid contents

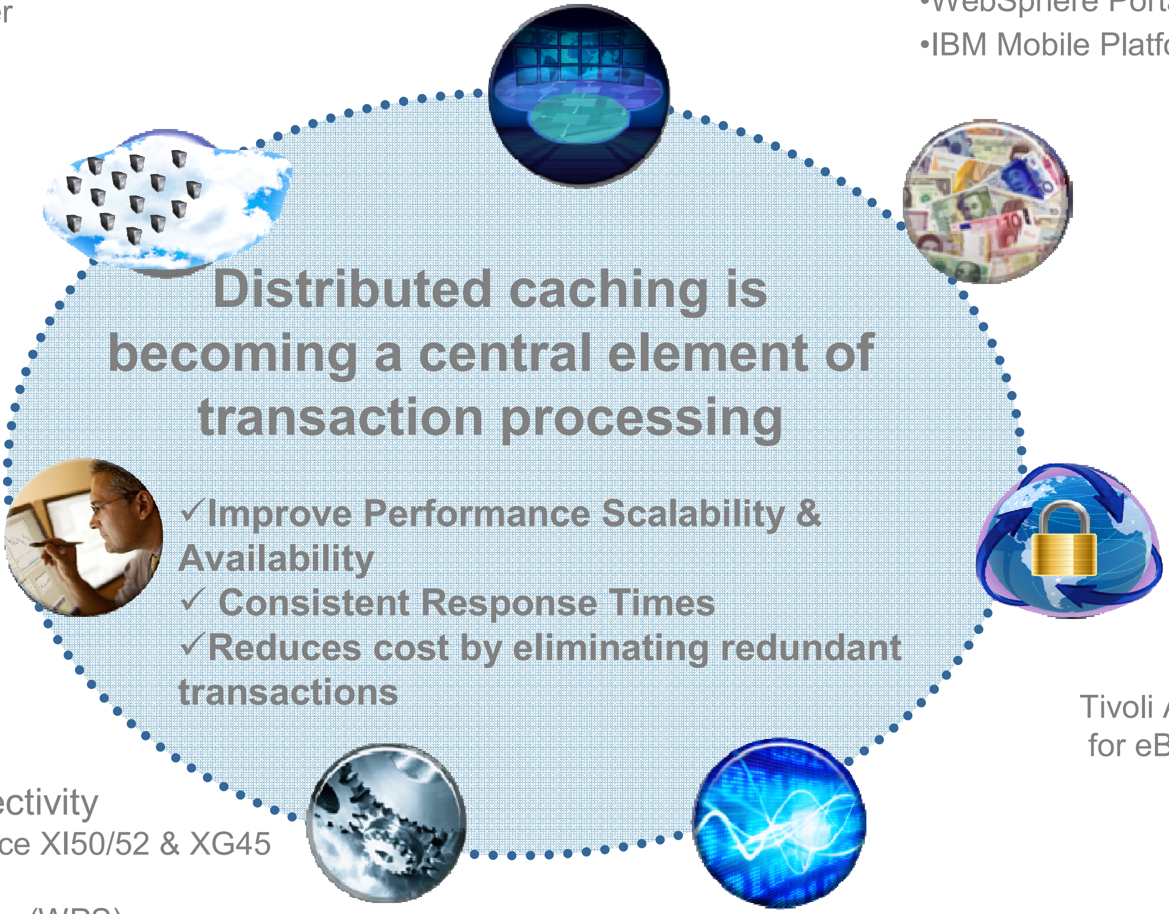
Spring Cache 3.1 Support - Enables Spring cache to develop or maintain Java Spring applications.

Need some Caching with That….?

Cloud
- IBM PureSystems
- IBM Workload Deployer
- Cast Iron Live (Saas)
- IBM Smart Cloud Application services

Infrastructure
- WebSphere Application Server
- Liberty
- Rational Team Concert

Stack Product Integration
- WebSphere Commerce
- WebSphere Portal
- IBM Mobile Platform / Worklight



Distributed caching is becoming a central element of transaction processing

✓Improve Performance Scalability & Availability
✓ Consistent Response Times
✓Reduces cost by eliminating redundant transactions

Tivoli Access Manager for eBusiness

Business Process & Connectivity
- DataPower Integration Appliance XI50/52 & XG45
- WebSphere Message Broker
- IBM Business Process Manager (WPS)
- WebSphere Registry and Repository
- IBM Operation Decision Management (iLOG JRules/WBE)

Getting Started

WebSphere eXtreme Scale Product Page
http://www-01.ibm.com/software/webservers/appserv/extremescale/

WebSphere DataPower XC10 Appliance Product Page
http://www-01.ibm.com/software/webservers/appserv/xc10/

WebSphere eXtreme Scale and WebSphere DataPower XC10 wiki
http://www.ibm.com/developerworks/connect/caching

WebSphere eXtreme Scale Free Trial
http://www.ibm.com/developerworks/downloads/ws/wsdg/

Contact your IBM Representative