Technically Speaking: Real World Elastic Scalability with WebSphere eXtreme Scale

The audio call numbers are listed below. Passcode 124020 Please dial in and stand by:

Country	Toll-free	Toll
Australia	1-800-24-2283	
Austria	0800-292-738	+43 179576264
Belgium	0800-7-3026	+32 22006114
Canada	877-421-0030 877-421-0033 877-421-0035 877-421-0038 877-421-0029 877-421-0528	770-615-1247 770-615-1250 770-615-1252 770-615-1254 770-615-1246 770-615-1258
China North	10800-711-0770	
China South	10800-110-0713	
Czech Republic	800-143-241	+420-272-133-001 or -527
Denmark	80-888377	+45 45245001
Finland	0800-914-630	+358 94596704
France	0800-902366	+33 157323040 or +33 157323041
Germany	0800-181-6323	+49 6951709081
Greece	00-800-11-004-9793	
Hungary	06-800-16-381	
India	Toll-free access no longer	
Ireland	1800-558728	+353 16569209

Country	Toll-free	Toll
Ireland	1800-558728	+353 16569209
Israel	1-809-317-098	-9164987
Italy	800-788634	+39 0269430413
Japan - KDD	00531-11-	
Japan - Cable & Wireless	0066-33-801263	
Japan - Softbank Telecom	0044-22-112668	
Japan – NTT	0034-800-900155	
Mexico	877-421-0030 877-421-0033 877-421-0035 877-421-0038 877-421-0029 877-421-0528	770-615-1247 770-615-1250 770-615-1252 770-615-1254 770-615-1246 770-615-1258
Netherlands	0800-022-8558	+31 20513 4100
Norway	800-18373	+47 66999780
Poland	0-0800111-1712 or IBM PlusGSM 85400	+48 22 366 5400 or +48 22 609 5400
Portugal	800-8-11111	
Russia	8-10-800-2-172-10-44	
Slovakia		+421-2-4954-6539 or -14480

Country	Toll-free	Toll
South Africa	0800-981-078	+27 11 302 7996
Spain	900-95-1089	+34 912754171
Sweden	020-799414	+46 87931800
Switzerland	0800-564-331	-7687
United Kingdom	0808-234-1969	+44 2070260533
USA	877-421-0528	770-615-1258

WebSphere software

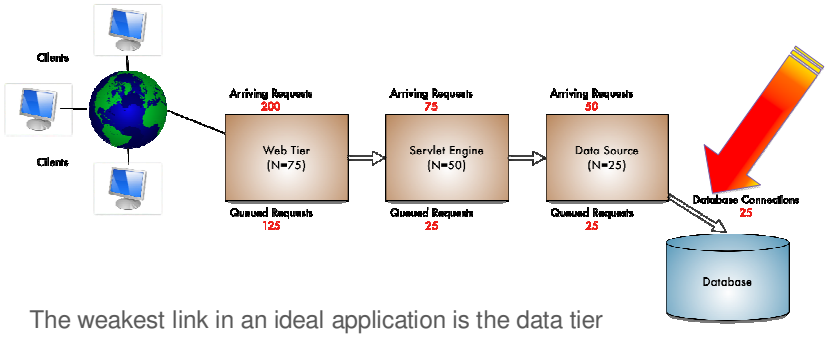
Speakers For Today's Webcast

Scott Lewis, Software IT Architect, MSI Systems Integrators

Joe Lea, WebSphere eXtreme Scale Product Manager

Why do we have a scalability problem?

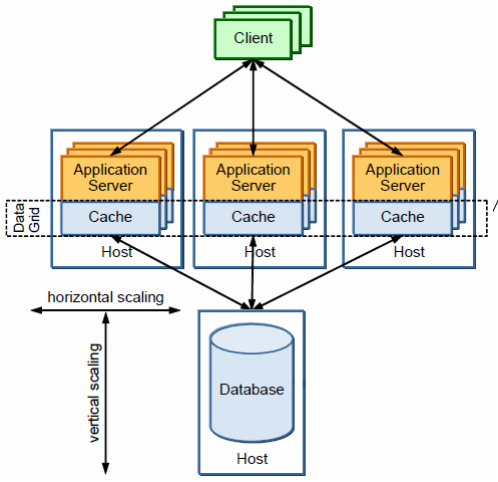
Queuing theory (and common sense) tells us that the weakest link limits the potential of a system.



The weakest link in an ideal application is the data tier and your application's interactions with it. It is what limits the absolute scalability of your application. It is also what constrains the performance and often the reliability of your applications.

Data Grid Relative to Traditional Data Stores

Think of a data grid as a more flexible, faster access data store that sits between the application and the more traditional hardened data stores like DB2 or IMS:



The data grid sits between the application layer and the database layer

What is Elastic Scalability?

Elastic computing, cloud computing, utility computing, and grids: these are essentially synonymous ideas.

An elastic environment leverages data grids and:

- *is cost efficient,*
- *offers incremental growth,*
- *offers linear scale with linear expense,*
- *eliminates single points of failure (SPOFs)*
- *eliminates choke points,*
- *has a predictable response time.*

Elastic Scalability Architecture

The architecture for elastically scalable applications is about balance.

"The Cray series of super-computers were an engineering marvel. The brilliance in the machine architecture wasn't only about speed. Cray's brilliance was about the balance within the machine. As they saw it, there is no point in having a superfast CPU if it was only going to be starved for work. So, much of the extreme engineering that took place was in making sure that the CPUs were never hung on wait conditions."

-Kirk Pepperdine

Five Rules for Elastic Scalability

1) Reduce the impact of commodity operations.

2) All components of the application system should add ***business*** value or they compromise reliability and add latency.

3) Consolidate systems of record and data models ***as perceived*** by the application tier.

4) Loosen the coupling between the front-end and back-end to eliminate serial dependencies.

5) Assume the probability of failure and compensate ***proactively***.

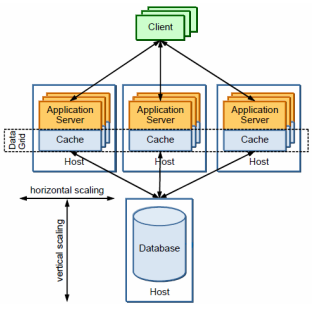
Real-World Elastic Scalability

Rule 1: Reduce the impact of commodity operations.

Real world use: Accelerate access to frequently used data by caching it in the application tier. Store static and mostly static data in an application-side cache.

Case study: Small non-profit organization was moving from the iSeries "everything-in-a-box" world to Windows and SQL Server. Performance on Windows and SQL Server was abysmal, and the complexity of the multi-tier environment caused great pains to the shoestring staff.

The new blade systems used for WAS had enough memory to deploy WXS to cache the entire data set used for the applications. (~24G on each box) Now, WAS talks to WXS, and WXS talks to SQL Server when necessary. Application response time is instantaneous, and the customer is now in position to move to a public cloud, with the database remaining in house.



Real-World Elastic Scalability

Rule 2: All components of the application system should add business value or they compromise reliability and add latency.

Real world use: Persist web sessions in the distributed cache to avoid bloating your workload JVMs. Handle grid replication asynchronously to avoid blocking during replication.

Case study: A customer was running lots of small JVMs that hosted an eCommerce workload. The application servers were stacked vertically (6 per node) across 4 mid-tier pSeries servers. The customer thought they had two choices for session persistence. M2M replication would clone every session object to every peer, causing massive waste of memory and a whole lot of work for the systems.

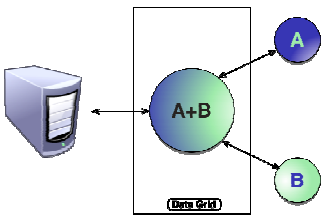
An alternative was DB2 and HACMP which was costly, far too slow for their requirements, unreliable from a failover perspective, and did nothing for site-level failover. A much simpler, far more capable, and dramatically less expensive option was to use WXS for session persistence. Secondary benefits are the use of the same grid to cache product catalogue data.

Real-World Elastic Scalability

Rule 3: Consolidate systems of record and data models as perceived by the application tier.

Real world use: Represent data in a format usable by the application rather than a format that is convenient to the data system – or worse, multiple data systems.

In the opposite direction, we can load multiple systems of record when an object in the cache changes. This can help avoid fragile transaction mechanisms like global transactions in the scope of the customer interaction.



Case study: Customer's primary OLTP system uses global transactions to pull and push data to Oracle and CICS in the scope of the customer's transaction. If either system slows down, everything crashes hard. An alternative implementation is to hydrate formatted objects in the distributed cache, and use write-behind caching to commit transactions to the back-end systems at a normalized (i.e. managed) rate.

Real-World Elastic Scalability

Rule 4: Loosen the coupling between the front-end and back-end to eliminate serial dependencies.

Real-world use: Write-behind caches isolate the application from the database. If we make changes to the database **through** the grid, all changes are buffered until the database is available. Think "store and forward" as applied to a database cache. We can also take advantage of batching and conflation to reduce DB transaction volume.

Bottom line: we need to move away from a queuing model (and the bad things that come with it) to move to the next level.

Case study: IBM's Fantasy Sports XS benchmark shows that using write-behind caching to loosen the coupling between the app and data tier improves response time for write events by 60-76% and throughput by 300-414%.

It does this by summarizing operations, including batching and conflation. For read operations, throughput increased by 366% and a whopping 560% improvement in response time.

Real-World Elastic Scalability

Rule 5: Assume the probability of failure and compensate proactively.

Real-world use: Reliability, availability, and scalability are sides of a triangle. To improve one, we need to improve all of them.

If we use a highly reliable, high performance, distributed cache to *augment* your persistence environment, it is possible to improve reliability, availability, and scalability to a degree that is prohibitively difficult and expensive to handle in any other way.

Case study: A customer moved from a monolithic Oracle database to RAC. The promise of "grid-like" scalability and reliability did not materialize. In fact, the customer had to run in active/passive mode with manual geo-failover, leaving them no better off than they were before the migration.

WebSphere eXtreme Scale does not need to replace the EIS platform. While it could in theory, WXS' job is to help improve the apparent quality of service of the "hardened" EIS platforms. This customer could have easily taken a lot of load off of Oracle. They could have let Oracle deal with more critical work, instead of data that never changes. WXS could have simplified the work that Oracle had to do when Oracle is online, and made it easier for WAS to handle the business logic. Finally, WXS could have compensated for those times when Oracle is offline for whatever reason.

Smarter Planet Solutions Require a Dynamic Application Infrastructure


Smart industries


Smart supply chains


Smart countries

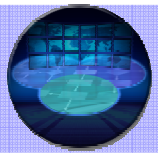
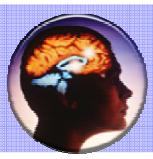

Smart weather


Smart regions


Smart cities

- 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		***Adoption Patterns***
"Meet business objectives consistently, nimbly, cost-effectively"		Application Foundation
"Enable applications to adapt to changing market conditions"		Intelligent Management
"Address extreme demands of clients & business models"		Extreme Transaction Processing

What is WebSphere eXtreme Scale?

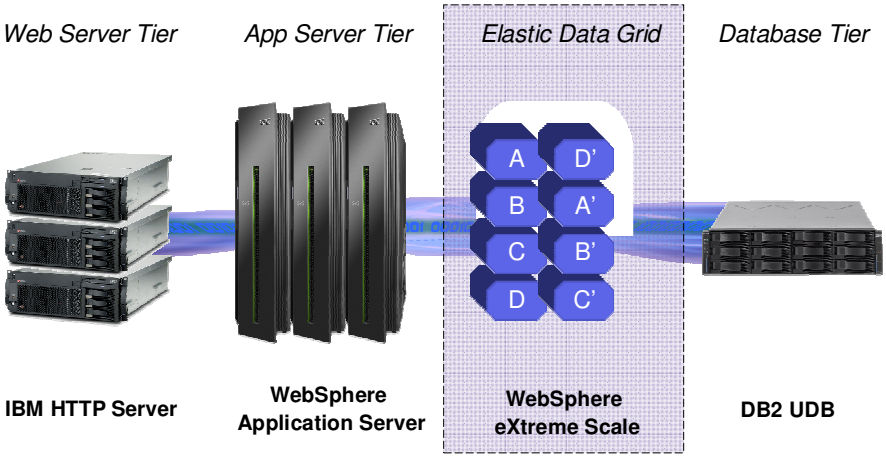
A flexible framework for realizing high performance, scalable and data-intensive applications

It can be used as a very powerful cache that scales from simple in-process topologies to powerful distributed topologies.

It can be used as a form of in memory database to manage application state (and it scales to 1000's of servers).

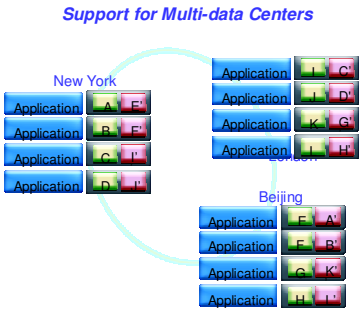
It can be used as a platform for building powerful XTP/Data Grid applications.

Modern Application Infrastructure Topology



IBM WebSphere eXtreme Scale

- **A powerful, elastic caching platform**
  - Very quick time-to-value and high ROI
- **Provides range of scaling strategies**
  - From simple, in-process cache to a distributed grid
- **Meets the needs of a variety of application environments:**
  - Java SE, Java EE (WAS, Tomcat, WebSphere Business Events, etc.), as well as for .Net applications via REST APIs
- **Offers automatic scaling and fail-over**
  - At predictable costs as transaction volumes grow
- **Provides flexibility to place data replicas**
  - At desired locations through support for Zones
- **Only requires Java SE run-time environment**
  - And, exploits WAS-ND environment when available



Spotlight on WebSphere eXtreme Scale V7

- Strategic Value
  - An essential IBM solution for elastic scalability and next-generation cloud environments
  - Fundamental part of IBM's approach to "Extreme Transaction Processing" (XTP)
- Key Product Differentiators
  - Ability to seamlessly use WebSphere eXtreme Scale with IBM middleware.
    - Improves application performance and time-to-value
  - Lightweight runtime footprint (15MB) that integrates with all versions of WebSphere and almost any java-based application container.
  - Access to data in the grid via REST broadens usage beyond Java including base ADO.Net support (Windows)
  - Ability to tie into monitoring tools such as Tivoli, Wily and Hyperic
  - Proven low latency access to data
  - Used at some of the biggest web sites in the world

Client Usage: Worldwide Fantasy Sports Web Site

Entertainment

5 Billion
requests per day

10x
reduced response times

Fantasy Sports Web Infrastructure

- **Before:** 60ms response time against database
- **After:** Improved to 6ms response time with WebSphere eXtreme Scale
- 450k concurrent users
- 80k requests per second up to 1M in 2011
- 8 weeks from concept to production

Support transaction-intensive services

Deliver consistent & predictable response times

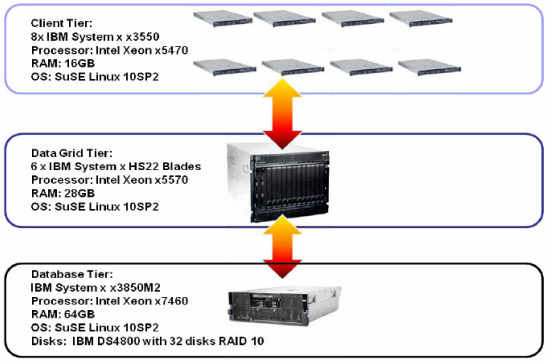
Take action on growing volumes of business events

Scale with simplicity and lower cost

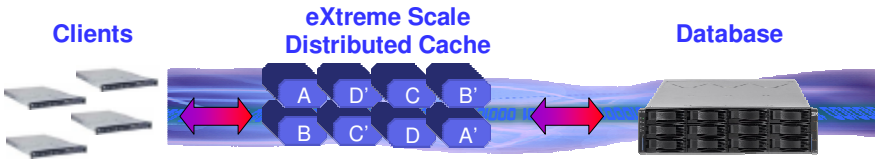
Performance and Economics of Caching

A worldwide Fantasy Sports XS application is used as a benchmark to demonstrate how eXtreme Scale can optimize database hardware needs while speeding up overall data access.

- This benchmark simulates the user management aspects of a worldwide fantasy sports Web site.
- The user session management is the key aspect here since game engines would handle the visualization aspects.
- Common user session operations such as retrieving and changing the core user's data, team members and athletes of interest form the workload.



Performance and Economics of Caching



	Read	Update	Insert	Delete	Database Load
Baseline Throughput	8971.4	2124.6	47.2	66.5	NA
Write-through Throughput	32843.7	7765.6	199.7	210.7	**-75%**
	+366%	**+366%**	**+423%**	**+317%**	
Write-behind Throughput	32086.2	7574.4	195.5	199.2	**-80%**
	+357%	+356%	+414%	+300%	
Baseline Response Time	56	79	77	95	**NA**
Write-through Response Time	10	57	47	55	**-75%**
	-82%	**-15%**	**-39%**	**-42%**	
Write-behind Response Time	19	25	31	23	**-80%**
	-84%	**-68%**	**-60%**	**-76%**	

ftp://ftp.software.ibm.com/software/webservers/appserv/WebSphere_eXtreme_Scale_WriteBehindPerformance.pdf

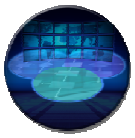
Real World Business Impact

- End-users' expectations for performance and availability are increasing across all constituents: prospects, customers, employees, partners.
- Organizations that are not able to optimize the performance and availability of applications are at risk of seeing significant declines in key business metrics such as customer satisfaction, conversion rates and revenue growth.
- Lost web sessions are not tolerated

Fortune 500 Call Center
• Response time improvement of 1 sec per database request
• 10 server requests per incident = 10 sec savings per incident
• 100 incident's per day per call center employee = 1000 seconds
• 100 call center employees = 100,000 seconds saved per day
• 27 idle hrs per day saved

Bottom Line: 27 hrs/ day = cost savings of 3 full time employees or a significant increase number of in incidents processed

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

Resources

- IBM Redbook: User's Guide to WebSphere eXtreme Scale
  - http://www.redbooks.ibm.com/abstracts/sg247683.html?Open&open&cm_mmc=4443-_-n-_-vrm_newsletter-_-10300_100294&cmibm_em=dm:0:13397061
- IBM Redpaper: WebSphere eXtreme Scale Mediator Pattern
  - http://www.redbooks.ibm.com/redpapers/abstracts/redp4398.html?Open
- WebSphere eXtreme Scale V7.0 Online Info Center
  - http://publib.boulder.ibm.com/infocenter/wxsinfo/v7r0/index.jsp
- WebSphere eXtreme Scale Wiki
  - https://www.ibm.com/developerworks/wikis/display/extremescale/Home
- Article: Leveraging WebSphere eXtreme Scale as an in-line database buffer
  - http://www.ibm.com/developerworks/websphere/library/techarticles/0906_vuong/0906_vuong.html
- WebSphere eXtreme Scale Trial Download
  - http://www.ibm.com/developerworks/downloads/ws/wsdg/learn.html
- WebSphere Extended Deployment Forum
  - http://www.ibm.com/developerworks/forums/forum.jspa?forumID=778
- WebSphere eXtreme Scale Videos
  - http://www.youtube.com/ibmextremescale

Additional resources

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

YouTube

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

developerWorks.

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.

http://www.ibm.com/developerworks/spaces/xtp

Attend IBM IMPACT 2010 to learn more

Top 5 Reasons To Attend to Impact 2010

1. Work Smarter and Achieve Business Agility Now!
2. Get the best technical education in the shortest time
3. Lower risk and optimize business performance
4. Access to the leading experts in Business & Technology
5. **Save Money! GWC and WUG members receive $125 off the standard rate**

Register Today

IBM Software
Impact2010
The Premier Conference for Business and IT Leaders
MAY 2-7 Las Vegas, NV
Discover. Interact. Optimize

Register now and save!

- Over **$4,200 worth of technical education** for only $21,50
  - Standard tech training rate: $725/day X 5 days = $3625 value
  - Two free certification tests = $400 value
  - 50% discount on additional certification testing
- Get a year's worth of World Class education and a lifetime of contacts in one week!

Q&A


Thank
You