Clustering - Initial discussion

{scrollbar}

top

Geronimo Clustering Overview - 10,000 feet...

Clustering is one of the key features that distinguishes an enterprise JEE implementation from the rest of the pack. As such it is an important requirement for Apache Geronimo.

By using clustering technology to provide a scalable and highly available platform for JEE deployment, Apache Geronimo will be able to compete on equal terms with existing commercial offerings.

This document will begin the task of :

- enumerating clustering requirements in Geronimo's tiers
- abstracting out commonality from these
- cataloguing software available to us that may be used to implement them
- suggesting a clustering architecture

A 'Cluster' is an architecture that achieves scalability and high availability through the arrangement of multiple smaller, cheaper, less reliable, resources, rather than single large, expensive, extremely reliable ones.

Scalability is ensured by decoupling dependencies on shared resources so that related tasks may be run concurrently on many machines (nodes) in the cluster without interfering with each other. If your architecture achieves this, you can scale to service more users by just adding more nodes.

High Availability is achieved through redundancy. If one less-reliable node fails, you fail-over to the next. In this way, the availability of your system becomes not the sum, but the product of the availability of its constituent nodes.

The presence of State in a cluster, frustrates the achievement of both of these goals, through being a point of shared contention (many tasks may need to read/write the same piece of state at the same time) and a point of failure (if state is held in a fragile resource e.g. memory, that is lost, then so is the state i,e, it is no longer
available).

Partitioning state can help restore scalability and making and maintaining multiple copies of state (replication) can be used to circumvent contention issues. Both solutions lead to further smaller problems such as ensuring that processes are run in the same partition as their state and ensuring consistancy of view across multiple copies of state etc. Other solutions and further problems abound.

The number of different uses of state within a cluster precludes the possibility of a single effective solution. So we need to devise solutions for each usecase within Geronimo.

We will enumerate and examine these usecases.

Back to Top

Web

URL: http://java.sun.com/products/servlet/download.html#specs
Interest Group: Jules Gosnell, Jan Bartel, Jeff Genender, David Jencks,...

Clustering in the web-tier has two points of implementation :

The HTTP Load Balancer:

- our solution should work with any load-balancer
- we can expect the LB to support some form of affinity (sticky or persistant sessions).

The HttpSession:

- large numbers of potentially large objects (more data in the tier than can comfortably be held in one node) - needs partitioned cache
- typically frequently written
- typically frequently read
- typically a single consumer/client
- transactionless
- only one copy of an HttpSession may be 'active' at any one time
- multi-threaded
- pessimistic
- suggested impl - WADI

N.B.: Some other solutions allow the clustering of more than just the HttpSession. The spec only requires that data stored in the HttpSession is distributable.

Dev Threads
WADI and Network Partitions: http://www.mail-archive.com/dev%40geronimo.apache.org/msg15855.html
WADI/AS merger - http://www.mail-archive.com/dev@geronimo.apache.org/msg14749.html

Back to Top

EJB

URL: http://java.sun.com/products/ejb/docs.html#specs
Interest group: David Blevins, Gianny Damour

Clustering in the EJB tier (like the web) has two points of implementation :

The Client/Proxy (equiv to Web Load-balancer):

- cluster-aware proxies needed
- can same proxies work with both OpenEJB and IIOP protocols ?

The Server:

MDB

- stateless

SLSB

- stateless

SFSB

- large numbers of potentially large objects (more data in the tier than can comfortably be held in one node) - needs partitioned cache
- transactional
- typically frequently written
- typically a single consumer/client
- typically frequently read
- single threaded
- pessimistic
- suggested impl - WADI

Entity

- potentially multiple consumers
- mapped to shared, persistant store
- transactional
- read/write frequency variable, pluggable impls required (e.g. RO Beans etc..)
- distributed db backed cache needed
- since mapped to db, does not need partitioned cache, data can just be unloaded
- distributed caching of entity beans: http://www.mail-archive.com/dev%40geronimo.apache.org/msg15697.html
- suggested impl - ActiveSpace

Dev Threads:
client stubs and load-balancing: http://www.mail-archive.com/dev@geronimo.apache.org/msg13533.html
AC in client stub: http://www.mail-archive.com/dev@geronimo.apache.org/msg14756.html
Entity invalidation...

Back to Top

JNDI

Interest Group: Rajith Attapattu http://java.sun.com/products/jndi/1.2/javadoc/

Apache Directory - http://directory.apache.org/ (may use this?)

- small amounts of small objects
- typically seldom written
- typically frequently read
- typically multiple consumers/clients
- multi-threaded
- transactionless
- prime candidate for straight forward 1->all replication
- suggested impl - ActiveSpace

Dev Threads:
ActiveSpace in JNDI: http://www.mail-archive.com/dev@geronimo.apache.org/msg14743.html
WADI/AS merger - http://www.mail-archive.com/dev@geronimo.apache.org/msg14749.html

Back to Top

JMS

http://java.sun.com/products/jms/docs.html

- impl - ActiveMQ

The ActiveMQ team will happily look after the clustering of JMS. For more information see the current Clustering Support in ActiveMQ

Dev Threads:
AMQ Clustering: http://www.mail-archive.com/dev%40geronimo.apache.org/msg15717.html

Deployment

http://jcp.org/en/jsr/detail?id=088

- potentially large objects (perhaps we could just send references)
- union of tier content smaller than single node capacity
- seldom written (undeployed)
- frequently read (invoked)
- transactionless
- no uniqueness constraint ? (what should we do about creating heterogeneous deployments?)
- prime candidate for straight forward 1->all replication
- suggested impl - ActiveSpace

suggestion - 'deployments sets' - a groups of nodes that share/implement a homogeneous deployment. See "Homogeneous vs Heterogeneous Deployments" section.

suggestion - app is deployed on one node, it forms a url to the app via its http server and replicates this link to other servers within its deployment-set. Each member of this set, receives the link, pulls down the app and deploys it.

suggestion - could replication be done synchronously and serially, so that as the app is deployed on each node it can be sanity checked before the distribution continues to the next node ?

WADI,AC,AS & Deployment: http://www.mail-archive.com/dev@geronimo.apache.org/msg15214.html

Back to Top

Management/Monitoring

http://jcp.org/en/jsr/detail?id=077

- still little discussion here
- may need to be centralised
- history - perhaps selected statistics snapshots should be dropped into JMS every few secs by each server, then stashed in an rrdb ? (WHEN, WHERE, WHAT, VALUE)

Back to Top

POJO

JCache - http://jcp.org/en/jsr/detail?id=107

- all of the above - whatever is available to Geronimo should be available to application-space pojos.
- JCache, and related technologies...
- suggested impls - ActiveSpace/WADI

Back to Top

DB

any takers ?

Back to Top

Trans-Tier issues

Network Partitions: check out 'Totem' thread on dev - not yet archived...
Application Session: http://www.mail-archive.com/dev@geronimo.apache.org/msg12072.html
clustering shopping list: http://www.mail-archive.com/dev@geronimo.apache.org/msg10561.html

Back to Top

Web Services

Interest Group: Rajith Attapattu

As WS moves towards more transport independance, session management needs to be decoupled from Http.

Suggested impl - WADI/ActiveSpace - more discussion needed here

Back to Top

Clustering Substrate

- suggested impl - ActiveCluster on top of various protocols

Back to Top

Auto-wiring

- clients need to autolocate nodes (ejb-client->jndi-port, http-load-balancer->http=port etc...)...
- nodes on same box need to negotiate ownership of per-host resources (e.g. ports)

Back to Top

Homogeneous vs Heterogeneous deployments

- homogeneous cluster is one set of nodes - the universal set
- heterogeneous cluster is one set (the universal set), with subsets defining internal scopes

Perhaps ActiveCluster could be extended to allow subclusters via the same Cluster connection, or we could use multiple cluster connections, or an AMQ MessageGroups...?

Back to Top

WADI/ActiveSpace synergy

WADI and ActiveSpace are complimentary technologies. There is a scope for API convergence and code reuse here. If the two could coexist behind the same API, we might be able to completely abstract the Cache impl from the consumer, giving more flexibility and allowing us to tailor solutions more closely to particular problems.

Back to Top

Suggested Building Blocks

ActiveMQ

URL: http://incubator.apache.org/activemq/
Interest Group: James Strachan, Hiram Chirino, ...

Geronimo's JMS implementation with pluggable transports including a peer: // protocol which allows peers in a cluster to message each other directly without the need for a central broker.

Back to Top

ActiveCluster

URL: http://activecluster.codehaus.org/
Interest Group: James Strachan, Hiram Chirino, ...

An API providing basic clustering fn-ality (specifically membership change notification) along with 1->all and 1->1 messaging. Also, various impls of this API, the most notable using ActiveMQ.

Back to Top

ActiveSpace

URL: http://activespace.codehaus.org/
Interest Group: James Strachan, Hiram Chirino, ...

Provides two abstractions, a JCache style Cache (replicated and/or transactional) and a JavaSpaces style Space, for distributed computing.

Back to Top

WADI

URL: http://wadi.codehaus.org/
Interest Group: Jules Gosnell,...

Provides Jetty and Tomcat compatible SessionManagers, a portable HttpSession impl and a partitioned distributed cache, for support of distributable webapps. OpenEJB SFSB support, also underway.

Back to Top

Tomcat Clustering

URL: http://tomcat.apache.org/tomcat-5.0-doc/cluster-howto.html
Interest Group: Jeff Genender, ...

Clustered HttpSessions for Tomcat (not Jetty).
1->All replication, so clusters constrained to a few nodes

Back to Top

EVS4J/TOTEM

URL: http://www.bway.net/~lichtner/evs4j.html
Interest Group: Guglielmo Lichtner

Extended Virtual Synchrony for Java.
Potentially very useful in cases where we need fast, strictly ordered 1->all messaging.
Might be integrated via ActiveCluster or ActiveMQ

Introduction: http://www.mail-archive.com/dev%40geronimo.apache.org/msg15644.html
Totem and ActiveCluster: http://www.mail-archive.com/dev%40geronimo.apache.org/msg15695.html
Infiniband: http://www.mail-archive.com/dev%40geronimo.apache.org/msg15743.html

Back to Top

Further Discussion

Back to Top

Further Reading

http://portal.acm.org/citation.cfm?id=326136 http://scholar.google.com/url?sa=U&q=http://historical.ncstrl.org/tr/ps/cornellcs/TR99-1726.ps http://www.jgroups.org/javagroupsnew/docs/index.html http://citeseer.csail.mit.edu/amir95totem.html http://docs.codehaus.org/display/WADI/Library http://citeseer.ist.psu.edu/32449.html

"Fault Tolerance in Distributed Systems"
Pankaj Jalote, 1994
Chapter 7, Section 5, "Degree of Replication"